New Photo-Crosslinking Mass Spectrometry Approaches for the Study of Intrinsically Disordered Proteins

Lindsey Danielle Ulmer

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Matthew F. Bush, Chair

Rachel E. Klevit

Robert E. Synovec

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

New Photo-Crosslinking Mass Spectrometry Approaches for the Study of Intrinsically Disordered Proteins

Lindsey Danielle Ulmer

Chair of the Supervisory Committee:

Matthew F. Bush

Department of Chemistry

Loss of proteostasis, the process of maintaining protein health and structure, is one of the hallmarks of aging. Loss of proteostasis can lead to the accumulation of aggregates and can lead to Parkinson's, Alzheimer's, and cataracts. Intrinsically disordered proteins play key roles in both the prevention of and aggregation in these aging-associated diseases. Small heat shock proteins (sHSPs) are one type of intrinsically disordered protein that help prevent protein aggregation. sHSPs are chaperones that prevent protein aggregation by binding aggregation-prone proteins before they aggregate. The disordered N-terminal region of sHSPs is known to be the chaperone active region. Still, not much is known about its interactions or how it prevents aggregation due to the difficulty in studying the disordered region. In this dissertation, photo-

crosslinking mass spectrometry approaches that were designed to study intrinsically disordered proteins are described.

Chapter 2 describes an informatic method for identifying residue-level crosslinks from the reagent benzolphenylalanine (BPA). BPA is a non-canonical amino acid that is incorporated site-specifically and reacts with any amino acids when treated with UV light. Because BPA can react with any amino acids, it can be used to identify residue-level crosslinks, but most informatic tools are not designed for residue-level crosslink identification. The informatic workflow in Chapter 2 makes it possible to identify residue-level BPA crosslinks on a large scale and could be applied to other photo-reactive amino acids. Chapter 3 describes a statistical workflow based on bootstrapping to compare qualitative data quantitatively. When applied to BPA crosslinking data, it determined that different sites of BPA incorporation yield significantly different crosslinks and that the crosslinks are not consistent solely with the reactivity of BPA. These findings establish that the crosslinks are probing structural features of the protein. The statistical method in Chapter 3 is highly versatile and could be applied to a wide variety of hypotheses. Chapter 4 describes the use of targeted-data-dependent acquisition to increase the number of identifications. This optimization of the mass spectrometry methods resulted in up to eight times more crosslinked peptide spectral matches from a single dataset. The increase in identifications from work in Chapter 4 increases the feasibility of applying the BPA crosslinking method to different systems, and similar techniques could be used to increase identifications for other crosslinking reagents. Chapter 5 describes the use of stable isotope labeling of amino acids in cell culture (SILAC) to quantify the depletion of peptides from crosslink formation. It describes best practices for sample preparation and data analysis to reduce the noise in the data, making it most feasible to quantify small degrees of change. Overall, the work in this dissertation

improves the ease and feasibility of photo-crosslinking mass spectrometry, which has great implications for the study of highly biologically relevant intrinsically disordered proteins.

Dedication

To Mom and Dad

Acknowledgements

I have lots of people to thank for my success in graduate school. First, I would like to thank my advisor, Matthew Bush. I appreciate Matt's support in pursuing a new research area in the lab, his openness to new ideas and directions, and his belief in me. He trusted me to lead a new collaborative project as a first-year graduate student. I also appreciate Matt's mentoring around science communication and the supportive lab culture that all the group meetings providing feedback on presentations have created.

Growing a new area of research in the Bush lab was challenging and exciting and would not have been possible without the support of my collaborators. I thank Rachel Klevit for supporting the project and me over the years. I learned so much from her not just about structural biology and small heat shock proteins but also about experimental design and planning. I am very fortunate to have had her and her lab as a second home base during graduate school. I worked with many of her lab members over the years, including Christoper Woods, Maria Janowska, Natalie Stone, Mia Cervantes, and Jasleen Kaur Sidhu. I always looked forward to subgroup meetings and even preparing samples in their lab space because of our conversations about science. Speaking with them about the project helped me to stay motivated throughout graduate school and helped me get excited about my work. The BPA crosslinking project would not have existed without Christopher Woods. He found literature using BPA and wanted to use it to study HSPB5, and his work troubleshooting the digestion and other aspects early in the project was crucial to its success and, more recently, its applications to different systems. I'd like to thank Chip Asbury and Lucas Murray for their support in studying interactions with tubulin. I'd like to thank Mike Hoopmann and Davidy Shteynberg for their support in using the Trans-Proteomic Pipeline.

I'd like to thank the rest of my committee members over the years as well, Rob Synovec, František Tureček, Dustin Maly, Nick Riley, and Michael MacCoss. Their feedback helped improve and shape the project. The work in Chapter 3 is a result of a discussion of a question Michael MacCoss asked during my general exam.

I want to thank members of the Bush lab for their support. Although their research was very different from mine, they were always very supportive and open to learning new things, and I also enjoyed learning from them. I want to thank Alice Martynova and Theresa Gozzo for being good friends and fellow teaching assistants. I want to thank Bruce Feng for being the best office buddy and Addison Roush for being a great fellow TA during the pandemic Zoom labs. I want to thank Daniele Canzani for training me as a new graduate student. I learned a lot about approaching life as a graduate student from him. I want to thank Beth Fawcett for her support and advice as I navigated the job search, and I'd like to thank Lucas Narisawa for his enthusiasm and interest in the project. Training Lucas over the past year has been one of the highlights of my graduate experience, and I know he will accomplish great things and continue to grow the crosslinking project.

I want to thank the wider UW community as well. The supportive environment across UW drew me to UW for my PhD, and that positively impacted my graduate school experience, from attending grad club events and making connections to receiving support from other students while teaching. I want to thank Brandon Bol, Martin Sadilek, and Ashley Dostie for their support throughout my teaching duties. Their support made teaching a much less stressful experience, and I don't know how I would have made it through without them. I'd also like to thank Priska Von Haller for her support at the University of Washington Proteomics Resource Facility. I want to thank Jessica Young, Rose Anderson, and members of the Biological Mechanisms of Healthy

Aging Training Grant. I learned so much from the other trainees during my time on the training grant and really appreciated being a part of such a supportive community during my last year as I was preparing to defend.

I'd also like to thank my advisors and mentors from my undergraduate at Georgia Institute of Technology. I was fortunate to perform undergraduate research in glycoproteomics in Ronghu Wu's lab. My research experience there sparked my interest in mass spectrometry and proteomics and led me to pursue graduate school. I'd especially like to thank Suttipong Suttapitugsakul for being a great mentor, as the graduate student with whom I worked most closely during my undergraduate research. Thanks to him trusting me with his cells, I was able to have a much more meaningful undergraduate research experience. His support throughout the graduate school application process led me to UW.

I want to thank my family for their support as well. My parents, sisters (Shannon and Kayley), and grandma were always there for me, even from the other side of the country. I appreciate their support of me moving so far away from Georgia to pursue my Ph.D. and their openness to learning about academia and science through my experiences. I'd also like to thank my partner, Jacob, for his support over the years, from driving me to the lab on a Saturday night to fix a clogged column, listening to practice talks, and celebrating my successes; I wouldn't be the person I am today without you. Our little family of us and our big, fuzzy dogs (Orson the Bouvier and Jack the Mioritic and Maremma sheepdog) has been my biggest source of joy.

Contents

Chapter 1. Introduction 14
1.1 Aging and Small Heat Shock Proteins 14
1.2 Crosslinking Mass Spectrometry 15
1.3 Photo-Reactive Amino Acids 17
1.3.1 Benzoylphenylalanine (BPA) 17
1.3.2 Other Photoactive Crosslinkers 18
1.3.3 Challenges in Identifying Crosslinks from Photo-Active Amino Acids 19
1.4 MS/MS Acquisition Methods 21
1.4.1 Data-Dependent Acquisition (DDA) 21
1.4.2 Data Independent Acquisition 23
1.4.3 Comparison of Different Acquisition Modes 24
1.5 Outline of Present Studies 26
1.6 References 27
Chapter 2. A High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid 43
2.1 Abstract 43
2.2 Introduction 45
2.3 Methods 49
2.3.1 Sample Preparation and Analysis 49
2.3.2 Identification of Crosslinks Using TPP 50
2.3.3 Access to Data and Software 51
2.4 Results and Discussion 51
2.4.1 Identifying the Proteins in a Sample 52
2.4.2 Effects of Protein Database Size. 53
2.4.3 Performance of Informatic Workflow 57
2.4.4 Validating Crosslinks 58
2.4.5 Visualizing Crosslinks from Photoreactive Amino Acids 61
2.4.6 Ambiguities in Assigning Residue-Specific Crosslinks 63
2.4.7 Reproducibility 64
2.4.8 Effects of Digestion 65
2.4.9 Further Evidence of Site-Specific Crosslinks 68
2.5 Conclusions 70

2.6 Acknowledgements 71
2.7 References 72
2.8 Supporting Information 82
2.8.1 LC-MS Methods 82
2.8.2 Comet Search Settings 83
2.8.3 Kojak Search Settings 83
2.8.4 PeptideProphet Settings 84
2.8.5 Identification of Crosslinks Using StavroX or MeroX 85
2.8.6 Effects of Software Versions 86
2.8.7 Effects of Gradient 87
2.8.8 Supporting Information References 100
Chapter 3: Bootstrapping for Quantitative Comparisons of Datasets 102
3.1 Introduction 102
3.2 Methodology 105
3.2.1 Source of Experimental Data 105
3.2.2 Terminology 105
3.2.3 Code Availability 107
3.3 Results and Discussion 107
3.3.1 Need for Method 107
3.3.2 Evaluation of Similarity Scores 108
3.3.3 Applying Bootstrapping to Hypotheses 111
3.3.4 Comparing to a Hypothetical Distribution 111
3.3.5 Comparing Datasets to Each Other 115
3.3.6 Comparing Replicates 118
3.4 Conclusion 121
3.5 Acknowledgements 121
3.6 References 122
3.7 Supporting Information 129
3.7.1 Calculating Rate Constants 129
3.7.2 Supporting Information References 141
Chapter 4. Discovering Crosslinks with Targeted DDA 142
4.1 Introduction 142

4.2 Methods 145
4.2.1 Sample Preparation 145
4.2.2 Inclusion List Creation 146
4.2.3 LC-MS 147
4.2.3 Data Analysis 148
4.3 Results and Discussion 148
4.3.1 PRM Methods Used 150
4.3.2 Effects of PRM or DDA 153
4.3.3 Effects of Dynamic Exclusion 153
4.3.4 Effects of Method Used to Generate Inclusion Lists 155
4.3.5 Reproducibility in Number of PSMs Identified 157
4.3.6 Sensitivity and Error 158
4.3.7 Crosslinks Identified across Methods 161
4.3.8 Number of Spectra and Precursors 164
4.3.9 Crosslink Precursor Intensity 171
4.4 Conclusion 173
4.5 Abbreviations 175
4.6 Acknowledgements 176
4.7 References 176
Chapter 5. Using SILAC to Identify Crosslinked Peptides through Quantifying Depletion of Unlinked Peptides 184
5.1 Introduction 184
5.2 Methods 188
5.2.1 Sample Preparation 188
5.2.2 LC-MS 188
5.2.3 Data Analysis 189
5.3 Results and Discussion 189
5.3.1 Heavy Amino Acid Incorporation 189
5.3.2 Effect of Peptides Considered for Quantification 192
5.3.3 Effect of Digestion Phase 198
5.3.4 Results for Crosslinked Samples 201
5.4 Challenges and Potential Next Steps 206
5.5 Conclusion 208

5.6 Acknowledgements 208
5.7 References 209

Chapter 1. Introduction

1.1 Aging and Small Heat Shock Proteins

Aging is a complicated process that results from a buildup of damage throughout time.[1] Protein protective networks deteriorate with aging, making it more difficult to maintain proteostasis. About 5% of proteins are misfolded throughout the life of a person, but damage to the proteostasis network from aging causes misfolded proteins to accumulate.[2] This buildup of misfolded proteins from failures in the proteostasis network is one factor that leads to the accumulation of protein aggregates, which causes diseases such as Alzheimer's and Parkinson's and the formation of cataracts.[3] Chaperones such as small heat shock proteins (sHSPs) play vital roles in preventing aging-associated protein aggregation.

sHSPs hold unfolded proteins until ATP-dependent chaperones take and refold them, but a lot is still unknown about their structure and mode of action. sHSPs exist as dynamic heterooligomers interacting with clients and other sHSPs through weak and transient interactions. All sHSPs include the following structural elements: a disordered N-terminal region (NTR), an α-crystallin domain (ACD) that consists of anti-parallel β-sheets, and a disordered C-terminal region (CTR).[4,5] Interactions between the ACDs of two sHSP monomers results in the formation of homodimers; interactions involving the NTR and CTR of different monomers then result in the formation of larger oligomers.[6] NMR data indicates that a wide variety of interactions in the NTR contributes to the heterogeneity.[7,8] This wide variety of interactions makes it challenging to propose models of higher structures of sHSPs as evidenced by the two very different structures for 24mers of HSPB5.[7,9] HSPB5, which is also called αB-crystallin, is a human sHSP that is expressed throughout the body so prevents the formation of aggregates associated with cataracts, Alzheimer's disease, and Parkinson's disease.[10] The large difference

between the HSPB5 24mer structures illustrates the need for additional structural methods—X-ray crystallography, NMR, and even cryo-EM have been unable to resolve the heterogeneity of human sHSP complexes, demonstrating the need for additional structural biology tools.

1.2 Crosslinking Mass Spectrometry

Crosslinking mass spectrometry (XL-MS) is a powerful method for characterizing protein-protein interactions, in part because it can capture transient interactions.[11] In traditional XL-MS, a chemical reagent reacts with proteins in samples prior to enzymatic digestion and liquid chromatography-mass spectrometry (LC-MS) analysis. Then, the binding sites of the crosslink are identified, and those binding sites pinpoint interacting partners and give distance constraints that can refine structural information. XL-MS is amendable to large, dynamic protein complexes because it focuses on identifying stable, newly formed covalent bonds, so the many different interactions present in heterogeneous systems such as sHSPs can be detected.[12,13] However, traditional XL-MS itself is limited by the available reagents, challenges in identifying crosslinked products, and challenges in refining structures based on the identified crosslinked products.[14]

In this study, we refer to experiments using crosslinkers such as disuccinimidyl suberate (DSS) as conventional crosslinking, due how commonly used reagents such as DSS are. However, reagents such as DSS have significant limitations because they are bifunctional crosslinkers with two reactive groups. DSS is shown in Figure 1. The reactive NHS-ester groups react with primary amines, so they can crosslink only with Lysine and the N-terminus of proteins, which could result in a low number of potential crosslinked products, depending on the sequence composition of the target proteins. DSS also has a long linker arm between the two reactive groups. DSS is reported to span 11.4 Å when fully extended,[15] and when taking into

account the dynamics of protein systems, the alpha carbons of lysine residues crosslinked with DSS can be up 26-30 Å apart.[16] In addition, DSS can only probe solvent-accessible interactions.

The limitations in DSS's reactivity and incorporation make it difficult to employ in certain systems. For example, the yeast sHSPs Hsp26 and Hsp42 have been studied with XL-MS.[14] The crosslinker DSS, which reacts only with primary amines, was used to crosslink Hsp26 and Hsp42 to the client protein malate dehydrogenase (MDH) prior to tryptic digestion and LC-MS analysis. Crosslinks identified interactions between Hsp26 and MDH, but crosslinks were not identified between Hsp42 and MDH even though Hsp42/MDH complex formation was confirmed by western blot analysis. The 171 N-terminal residues of Hsp42 do not include lysine residues, which is likely why client crosslinks could not be identified between Hsp42 and MDH using DSS. This study on Hsp42/MDH complex formation illustrates the limitations of XL-MS due to crosslinker reactivity. Lysine residues are low abundance in other sHSPs as well, so XL-MS studies of sHSPs have been limited.[14,17]

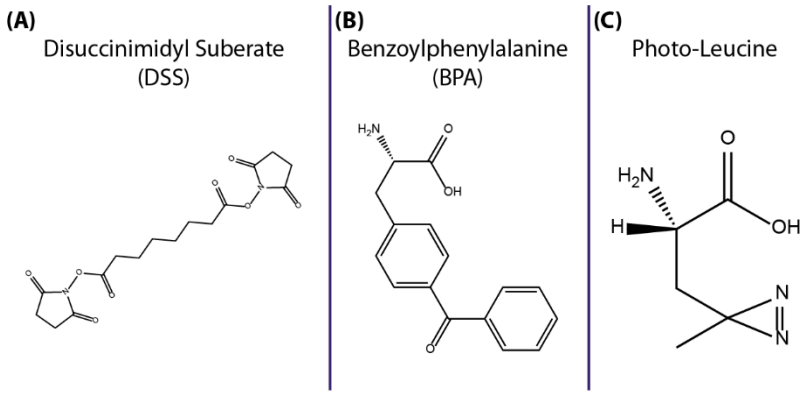

Figure 1. The crosslinking reagents DSS, BPA, and Photo-Leucine are depicted here. DSS (Panel A) is a conventional crosslinking reagent. BPA (Panel B) and Photo-Leucine (Panel C) are photoactive amino acids with different reactive groups.

1.3 Photo-Reactive Amino Acids

A strategy to overcome these challenges in XL-MS is to use photo-reactive crosslinking reagents such as benzoylphenylalanine (BPA), photo-methionine, or photo-leucine that are non-canonical amino acids that can potentially react with any amino acid when exposed to UV light.[18] These photo-reactive amino acids benefit from having short reactive distances and have been referred to as zero-length because they lack the linker arms common in conventional crosslinking.[19] Because they are incorporated within the protein sequence, they can analyze non-solvent accessible interactions in a more native environment than conventional crosslinkers.[20] However, photo-reactive crosslinkers differ in how they are incorporated into the protein sequence and in the chemistry of their crosslinking reaction.

1.3.1 Benzoylphenylalanine (BPA)

Benzoylphenylalanine (BPA) is a non-canonical amino acid crosslinker that has a benzophenone functional group as shown in Figure 1B. BPA reacts with CH groups when photoactivated, so can react with any amino acid. If no crosslink is formed, BPA relaxes back to the starting product, giving few side products and high crosslinking yields.[21] It has a 2.5-3.1 Å reactive distance and no mass change with the crosslinked reaction.[22] Prior studies of calmodulin and a BPA-variant of the calmodulin-binding peptide[23] and of crystallography of a BPA crosslinked complex of the liver oncoprotein gankyrin and the C-terminal domain of the S6 proteasomal protein exhibited little structural change with BPA incorporation and crosslink formation.[24] Whether BPA reacts to an equal extent with each amino acid has been debated. There is evidence that it will react with methionine to a greater extent than other amino acids,[25] and that whether the reaction is performed in an aqueous or organic solvent has a large effect on which amino acids BPA is most likely to react with.[26] It has also been proposed that the

geometry and steric hindrance of BPA's two aromatic rings may play a role and limit what can be crosslinked.[27,28] However, this has been viewed as a positive for BPA in that it results in fewer different products, so the resulting products should be more intense.[29]

BPA is incorporated site-specifically, even at non-solvent accessible sites, using an amber stop codon.[30] However, BPA crosslinking has been performed in vivo in yeast,[31–33], *E coli*,[34] and mammalian cells.[35] BPA's site-specific incorporation creates a targeted-crosslinking analysis because the single site of BPA incorporation defines one end of all crosslinks.

1.3.2 Other Photoactive Crosslinkers

Photo-methionine and photo-leucine (shown in Figure 1C) both have diazirine functional groups that form crosslinks via a carbene intermediate through the loss of two nitrogen atoms. However, diazirines can instead form electrophiles that yield side products.[27,36] If they are not in a hydrophobic environment, they are likely to react with water, which results in no crosslink formation and low yields.[37] They can also isomerize to linear diazo compounds that can react with carboxylic acids to yield esters. The MS-cleavability of ester products and the mass change from the elimination of two nitrogen atoms upon crosslink formation have been leveraged to assist in crosslink identification.[37]

Photo-methionine and photo-leucine are often incorporated in restricted media lacking the canonical amino acid, creating proteome-wide incorporation and an untargeted analysis.[38,39] When incorporated this way, the noncanonical amino acid uses the endogenous tRNA for incorporation, so it competes with the endogenous amino acid for incorporation.[40] The generation of specific tRNAs for non-canonical amino acids with smaller side chains similar to photo-leucine and photo-methionine is an ongoing field of research.[41] Longer linker chains have been added to diazirine-based photo crosslinkers to gain higher yields by increasing the reactive

distance. These longer crosslinkers have been incorporated site specifically in *E. coli*[42–44] and HEK 293 T cells.[43–45]

1.3.3 Challenges in Identifying Crosslinks from Photo-Active Amino Acids

The ability of photoactive amino acids to react with any amino acid makes it more difficult to identify crosslinked products. Crosslinked products are already very low-intensity compared to unlinked peptides due to inefficiencies in the crosslinking reactions and because crosslinkers can form multiple products, further dividing the intensity of the crosslinks.[12] Because BPA and other photoactive amino acids can react with all amino acids, they will form more different crosslinked products, which causes the crosslinks to be lower in intensity.[46] Different enrichment strategies have been used to overcome the difficulty of identifying low-intensity crosslinked products by aiming to purify crosslinked products. Size exclusion chromatography on digested peptides[47,48] and affinity tag-based purification for crosslinkers with alkyne,[49] phosphonic acid,[50] or biotin groups[51] has been used to purify and increase the intensity of crosslinked products. In the work presented here, we use an in-gel digestion of dimer species to enrich crosslinked products.

Identifying crosslinks is also a challenge in XL-MS in general because two peptides must be identified for each crosslinked spectrum match (PSM). This creates what is referred to as the n^2 problem, where n is the number of peptides in the search library, because all possible combinations of two peptides must be considered for each spectrum, increasing the search space exponentially and resulting in long search times.[12] The n^2 problem makes it difficult to identify crosslinked PSMs and has led to the wide variety of informatic tools for XL-MS with many labs using their own in-house developed tools.[52,53]

However, suppose the crosslinker used is a photo-active amino acid that was incorporated at a single site. In that case, all crosslinks originate from the single incorporation site, simplifying the identification of crosslinked peptides. A crosslink is defined as from the incorporated photo-active amino acid to any other amino acid. Therefore, every crosslink includes a peptide containing the incorporated photo-active amino acid. This increases the ability to confidently identify crosslinks because half of every crosslink is pre-defined by the site of the incorporated photo-active amino acid. The n^2 problem inherent to conventional crosslinking does not apply to these searches. However, even without the n^2 problem, the sensitivity to individual, residue-level crosslinks is still lowered because there are far fewer crosslinked than unlinked PSMs, so it is more difficult to generate models to validate crosslinked PSMs.

There are additional challenges in identifying crosslinks unique to photo-active amino acids. Even though many tools are available, they were developed with conventional crosslinkers in mind and do not account for the broad reactivity of photo-active amino acids. Because each amino acid can be a crosslink site when using a photo-active crosslinker, there can be ambiguity in the crosslink site assignment if there is a gap in the b- and y-ions identified in the crosslinked peptide. Figure 2 shows an example where b-ions containing the first four residues of the crosslinked peptide are identified and y-ions containing the last two residues of the crosslinked peptide are identified, which leaves the middle six residues as potential crosslink sites.

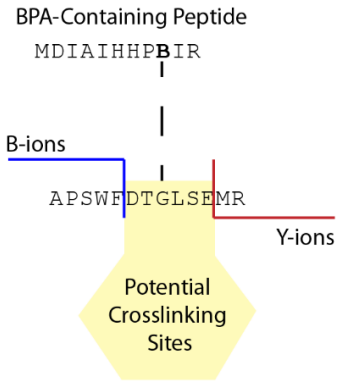

Figure 2. An example of ambiguity in the crosslink site assignment when using BPA (represented as B) as the crosslinker is shown.

1.4 MS/MS Acquisition Methods

In bottom-up proteomics experiments where proteins are enzymatically digested into peptides before LC-MS analysis, peptides form intact ions (precursors) that must be fragmented for sequence identification.[54] This analysis is performed in tandem MS experiments. In tandem MS, an MS1 spectra is collected of intact precursor ions. Then, precursors are fragmented, and an MS2 spectra is collected of the resulting fragments. Different acquisition methods differ in how precursors are selected for fragmentation as described in the following sections.

1.4.1 Data-Dependent Acquisition (DDA)

In untargeted data-dependent acquisition (DDA), precursor ions from the MS1 scan are automatically selected for fragmentation and MS2 analysis.[55] The instrument automatically selects precursor ions for MS2 based on their intensity and on other vendor-software-specific criteria, such as whether isotope patterns resemble those expected for peptides.[56] User-defined parameters for DDA include how many precursors from a given MS1 to consider for MS2 (topN) and dynamic exclusion. Dynamic exclusion sets a defined amount of time (typically 20

seconds up to two minutes, depending on the gradient used) for a precursor not to be selected for fragmentation again after initial selection and fragmentation. DDA's automated selection of precursors for MS2 results in little interference in MS2 spectra (a single m/z fragmented in each MS2) and requires little knowledge about the sample to set up. The relatively simple MS2 spectra make the resulting data easier to analyze.

Untargeted DDA's automated selection of precursor ions based primarily on precursor intensity leads to undersampling many peptides in a sample, especially low-intensity peptides. This results in a relatively small portion of the precursors in DDA being analyzed with MS2.[57] DDA has been referred to as stochastic in that it samples mostly the same (higher intensity) things each time.[56] In typical DDA, only 25-30% of MS2 spectra are identified,[58] and attempts to get more identifiable spectra led to the use of targeted acquisition methods.

Targeted DDA is not commonly used, but uses an inclusion list of precursor values to target for MS2 analysis preferentially.[59] It has also been referred to as directed mass spectrometry.[60] It has been used with inclusion lists of features detected in MS1 scans but not identified to increase the number of identifications,[61,62] to facilitate the identification of low-intensity PTMS,[63] and to generate targeted Data-Independent Acquisition assays. [64] Similar to untargeted DDA, targeted methods select a single precursor at a time for MS2 analysis, resulting in relatively simple MS2 spectra. Therefore, targeted DDA data can be analyzed using the same tools as untargeted DDA data.

A typical DDA data processing workflow involves database searching against a database of known protein sequences, validation of the database search results, and, depending on the goals of the study, MS1-based quantification using a program such as Skyline.[54,65] Many different programs exist for database searching such as MSFragger[66] and Comet[67] and for search

validation such as Philosopher[68] and PeptideProphet.[69] Which programs are used affects the obtained results, and these differences can be minimized by analyzing 2 or 3 replicates or using multiple search engines.[70]

1.4.2 Data Independent Acquisition

In untargeted Data-Independent Acquisition, the instrument scans a range of precursor m/z values in a set interval, fragmenting and collecting an MS2 on everything within that interval.[71] MS1 scans are collected at regular intervals throughout the method. The range is typically 500-900 m/z with a 10 m/z interval. However, multiplexing strategies have been introduced to facilitate narrower windows without increasing the duty cycle.[72] Untargeted DIA has also been called Sequential Window Acquisition of all Theoretical Fragment ion spectra (SWATH) DIA, and the intervals can be designed to overlap to ensure complete isotope distributions of any precursor are captured in a single MS1 scan.[73] Because DIA fragments and collects an MS2 on every precursor within an interval, it is highly reproducible and samples everything within the precursor range. However, the data is much more challenging to analyze because DIA MS2 spectra often contain fragments from multiple peptides (from whatever precursors were in the interval).[74]

In targeted DIA methods, desired m/z values are input, and the input values are selected and analyzed in MS2. The targeted methods differ in how many ions are analyzed. In selected reaction monitoring (SRM), quantitation is typically based on the intensity of one fragment ion from one precursor m/z.[75] In multiple reaction monitoring (MRM), multiple fragment ion values are monitored sequentially.[76,77] MRM and SRM assays are commonly performed on triple quadrupoles but can be performed on a wide variety of instruments. Parallel Reaction Monitoring (PRM) analyzes multiple fragment ions from a single precursor m/z simultaneously. PRM was

first implemented in an orbitrap instrument[78] but could be performed in a wide variety of instruments as long as there is a multiplexed analyzer capable of detecting multiple fragment ions at once.[79] Defining select precursor m/z values for analysis in these targeted methods dramatically increases the selectivity and removes the stochastic nature of DDA by improving the sampling of the select precursors.[75] However, unlike DDA and untargeted DIA methods, prior knowledge of the sample and species of interest is required to input the desired precursor values for analysis. Skyline can be used to design (pick target peptides) for targeted methods.[65]

DIA data can be analyzed based on spectral libraries created from previous DDA data. Skyline can create spectral libraries from DDA data and use those to analyze DIA data. Spectral libraries can also be predicted based on protein sequences so that no prior data is needed to create a spectral library. DIA-NN uses this approach.[80] Similar to the effect of the protein database used in proteomics, the specific spectral library used to search DIA data affects the results. Using sample-specific libraries for identification and smaller libraries for quantification is recommended.[81] Instead of using a spectral library for DIA data analysis, DIA data can be converted to pseudo-DDA files using a program such as DIA-Umpire[82] for use in database searching algorithms.

1.4.3 Comparison of Different Acquisition Modes

The different acquisition modes are commonly applied for different purposes. Since untargeted DDA does not require much prior information about the sample to set up the method and analyze the resulting data, it is often used for discovery studies on novel samples with little previous knowledge. However, DDA's automated selection of precursors leads to a relatively small portion of the precursors being analyzed with MS2.[57] Targeted DDA methods can be used to facilitate the identification of specific species of interest, as has been done for precursor ions

that were present in MS1 but not identified in MS2. [61,62] Targeted DDA has a lower dynamic range and sensitivity than targeted DIA methods because the precursor ions must be detected to collect an MS. Targeted DDA is also compatible with analyzing more precursor values than targeted DIA methods.[83] For both targeted and untargeted DDA studies, the data can be analyzed using the same database searching programs.[84] The inconsistent sampling also makes accurate quantification more difficult with DDA data.

When precursor ions of interest are identified (typically from untargeted DDA data), PRM or other targeted DIA methods can be set up to analyze select precursors in the samples. Limiting the analysis to select precursors with PRM improves the sampling of the specified precursors, significantly increasing the selectivity and reproducibility relative to DDA.[75] The increased sampling is also beneficial for improved quantification.[85,86] In untargeted DIA data, because everything within a set interval is fragmented for MS2, there is a comprehensive sampling of species in the sample, which results in high reproducibility and accuracy for quantification. Because DIA data fragments all precursors in a specified window, there is lots of interference in the MS2 spectra, so different methods need to be employed to analyze the data compared to DDA data. DIA data is commonly analyzed using spectral library searching. Although spectral library searching has largely become more popular because it is helpful for analyzing complicated DIA MS2 spectra, it has been demonstrated that using it on DDA data also improves quantification and reproducibility.[87]

In summary, DIA helps improve sampling relative to DDA. For PRM and other targeted DIA methods, target species must be input to create the method, so only select precursors are analyzed in a given run. Because only select precursors are analyzed, additional precursors cannot be detected later when analyzing the data. For untargeted DIA, many things are detected

and well-sampled in a single run, so additional precursors could be looked for when analyzing the data, but untargeted DIA data is more complicated to analyze because of the interference in the MS2 spectra.

1.5 Outline of Present Studies

In this work, I implement new photo-crosslinking mass spectrometry data collection and analysis methods. These methods aim to increase the feasibility of photo-crosslinking experiments to facilitate wide-scale identification of residue-level crosslinks.

In the work discussed in Chapter 2, I developed a bioinformatic workflow for identifying BPA crosslinks that adopts open-source bioinformatic tools from the Trans Proteomic Pipeline[69,88,89] and accounts for ambiguity in the crosslink site assignment. There are some available database searching tools (StavroX[90] and MeroX[91,92]) that account for ambiguity in the crosslink site assignment that I compare my workflow to in Chapter 2, but briefly, my workflow can identify up to 10 times more crosslink PSMs. Work in Chapter 2 uses untargeted DDA data. This work addresses the challenges in identifying crosslinks that result from the wide-variety of database searching programs by benchmarking different database searching tools.[52] By applying my workflow to HSPB5, I have been able to establish that there is order within the NTR and a network of NTR-NTR interactions.[93,94] This application of my method is the first time that NTR-NTR interactions have been identified with residue-level resolution. These residue-level interactions have characterized the selectivity of HSPB4/5[93] and long-range perturbations of HSPB5 accessibility.[94]

Comparing crosslinked datasets presents a unique challenge due to the relatively low number of observations. In Chapter 3, I described a bootstrapping-based statistical method for quantitatively comparing datasets that can account for differences in the number of observations

and determine if datasets or hypothetical probability distribution functions are consistent. I apply this method to untargeted DDA data to determine that different sites of BPA incorporation yield significantly different crosslinks, which confirms that we are obtaining site-specific information from BPA crosslinking.

In Chapter 4, I implement targeted DDA methods with large, comprehensive precursor lists that contain every potential crosslinked product to help target MS2 analysis toward crosslinked precursors. In doing so, up to three times more crosslink PSMs are identified. This work helps alleviate challenges in detecting crosslinked products due to their low intensity. Targeted DIA (PRM) has been used previously to improve crosslink identification for better quantification.[95] However, in Chapter 4, we use targeted DDA to identify crosslinks through using comprehensive precursor lists.

My efforts to date show that it is feasible to identify and assign residue-level crosslinks on a large scale (Chapter 2). While this information is highly informative, it is qualitative. To develop more quantitative methods, Chapter 5 discusses the feasibility of using stable isotope labeling of amino acids in cell culture to quantify non-crosslinked peptides and identify crosslinked peptides through non-crosslinked peptide depletion. This quantification is done using untargeted DIA data, and different aspects affecting the quantification results, such as the sample preparation and data analysis method, are described in detail in Chapter 5.

1.6 References

(1) Kirkwood, T. B. L. Understanding the Odd Science of Aging. *Cell* **2005**, *120* (4), 437–447. https://doi.org/10.1016/j.cell.2005.01.027.

(2) Magalhaes, S.; Goodfellow, B. J.; Nunes, A. Aging and Proteins: What Does Proteostasis Have to Do with Age? *Curr. Mol. Med.* **2018**, *18* (3), 178–189. https://doi.org/10.2174/1566524018666180907162955.

(3) Klaips, C. L.; Jayaraj, G. G.; Hartl, F. U. Pathways of Cellular Proteostasis in Aging and Disease. *J. Cell Biol.* **2018**, *217* (1), 51–63. https://doi.org/10.1083/jcb.201709072.

(4) Haslbeck, M.; Franzmann, T.; Weinfurtner, D.; Buchner, J. Some like It Hot: The Structure and Function of Small Heat-Shock Proteins. *Nat. Struct. Mol. Biol.* **2005**, *12* (10), 842–846. https://doi.org/10.1038/nsmb993.

(5) Aquilina, J. A.; Benesch, J. L. P.; Bateman, O. A.; Slingsby, C.; Robinson, C. V. Polydispersity of a Mammalian Chaperone: Mass Spectrometry Reveals the Population of Oligomers in B-Crystallin. *Proc. Natl. Acad. Sci.* **2003**, *100* (19), 10611–10616. https://doi.org/10.1073/pnas.1932958100.

(6) Delbecq, S. P.; Klevit, R. E. One Size Does Not Fit All: The Oligomeric States of αB Crystallin. *FEBS Lett.* **2013**, *587* (8), 1073–1080. https://doi.org/10.1016/j.febslet.2013.01.021.

(7) Jehle, S.; Vollmar, B. S.; Bardiaux, B.; Dove, K. K.; Rajagopal, P.; Gonen, T.; Oschkinat, H.; Klevit, R. E. N-Terminal Domain of B-Crystallin Provides a Conformational Switch for Multimerization and Structural Heterogeneity. *Proc. Natl. Acad. Sci.* **2011**, *108* (16), 6409–6414. https://doi.org/10.1073/pnas.1014656108.

(8) Jehle, S.; Rajagopal, P.; Bardiaux, B.; Markovic, S.; Kühne, R.; Stout, J. R.; Higman, V. A.; Klevit, R. E.; van Rossum, B.-J.; Oschkinat, H. Solid-State NMR and SAXS Studies Provide a Structural Basis for the Activation of αB-Crystallin Oligomers. *Nat. Struct. Mol. Biol.* **2010**, *17* (9), 1037–1042. https://doi.org/10.1038/nsmb.1891.

(9) Braun, N.; Zacharias, M.; Peschek, J.; Kastenmüller, A.; Zou, J.; Hanzlik, M.; Haslbeck, M.; Rappsilber, J.; Buchner, J.; Weinkauf, S. Multiple Molecular Architectures of the Eye Lens Chaperone αB-Crystallin Elucidated by a Triple Hybrid Approach. *Proc. Natl. Acad. Sci.* **2011**, *108* (51), 20491–20496. https://doi.org/10.1073/pnas.1111014108.

(10) Cox, D.; Selig, E.; Griffin, M. D. W.; Carver, J. A.; Ecroyd, H. Small Heat-Shock Proteins Prevent α-Synuclein Aggregation via Transient Interactions and Their Efficacy Is Affected by the Rate of Aggregation. *J. Biol. Chem.* **2016**, *291* (43), 22618–22629. https://doi.org/10.1074/jbc.M116.739250.

(11) Klykov, O.; Steigenberger, B.; Pektaş, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990. https://doi.org/10.1038/s41596-018-0074-x.

(12) Yu, C.; Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **2018**, *90* (1), 144–165. https://doi.org/10.1021/acs.analchem.7b04431.

(13) Singh, P.; Panchaud, A.; Goodlett, D. Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Anal. Chem.* **2010**, *82* (7), 2636–2642. https://doi.org/10.1021/ac1000724.

(14) Ungelenk, S.; Moayed, F.; Ho, C.-T.; Grousl, T.; Scharf, A.; Mashaghi, A.; Tans, S.; Mayer, M. P.; Mogk, A.; Bukau, B. Small Heat Shock Proteins Sequester Misfolding Proteins in Near-Native Conformation for Cellular Protection and Efficient Refolding. *Nat. Commun.* **2016**, *7* (1), 13673. https://doi.org/10.1038/ncomms13673.

(15) Thermo Fisher Scientific. https://www.thermofisher.com/order/catalog/product/21655.

(16) Merkley, E. D.; Rysavy, S.; Kahraman, A.; Hafen, R. P.; Daggett, V.; Adkins, J. N. Distance Restraints from Crosslinking Mass Spectrometry: Mining a Molecular Dynamics Simulation Database to Evaluate Lysine–Lysine Distances. *Protein Sci.* **2014**, *23* (6), 747–759. https://doi.org/10.1002/pro.2458.

(17) Lambert, W.; Rutsdottir, G.; Hussein, R.; Bernfur, K.; Kjellström, S.; Emanuelsson, C. Probing the Transient Interaction between the Small Heat-Shock Protein Hsp21 and a Model Substrate Protein Using Crosslinking Mass Spectrometry. *Cell Stress Chaperones* **2013**, *18* (1), 75–85. https://doi.org/10.1007/s12192-012-0360-4.

(18) Mishra, P. K.; Yoo, C.-M.; Hong, E.; Rhee, H. W. Photo-Crosslinking: An Emerging Chemical Tool for Investigating Molecular Networks in Live Cells. *ChemBioChem* **2020**, *21* (7), 924–932. https://doi.org/10.1002/cbic.201900600.

(19) Hage, C.; Iacobucci, C.; Rehkamp, A.; Arlt, C.; Sinz, A. The First Zero-Length Mass Spectrometry-Cleavable Cross-Linker for Protein Structure Analysis. *Angew. Chem. Int. Ed.* **2017**, *56* (46), 14551–14555. https://doi.org/10.1002/anie.201708273.

(20) Häupl, B.; Ihling, C. H.; Sinz, A. Combining Affinity Enrichment, Cross-Linking with Photo Amino Acids, and Mass Spectrometry for Probing Protein Kinase D2 Interactions. *PROTEOMICS* **2017**, *17* (10), 1600459. https://doi.org/10.1002/pmic.201600459.

(21) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(22) Prestwich, G. D.; Dormán, G.; Elliott, J. T.; Marecak, D. M.; Chaudhary, A. Benzophenone Photoprobes for Phosphoinositides, Peptides and Drugs. *Photochem. Photobiol.* **1997**, *65* (2), 222–234. https://doi.org/10.1111/j.1751-1097.1997.tb08548.x.

(23) Kauer, J. C.; Erickson-Viitanen, S.; Wolfe, H. R.; DeGrado, W. F. P-Benzoyl-L-Phenylalanine, a New Photoreactive Amino Acid. Photolabeling of Calmodulin with a Synthetic Calmodulin-Binding Peptide. *J. Biol. Chem.* **1986**, *261* (23), 10695–10700. https://doi.org/10.1016/S0021-9258(18)67441-1.

(24) Sato, S.; Mimasu, S.; Sato, A.; Hino, N.; Sakamoto, K.; Umehara, T.; Yokoyama, S. Crystallographic Study of a Site-Specifically Cross-Linked Protein Complex with a Genetically Incorporated Photoreactive Amino Acid,. *Biochemistry* **2011**, *50* (2), 250–257. https://doi.org/10.1021/bi1016183.

(25) Wittelsberger, A.; Thomas, B. E.; Mierke, D. F.; Rosenblatt, M. Methionine Acts as a "Magnet" in Photoaffinity Crosslinking Experiments. *FEBS Lett.* **2006**, *580* (7), 1872–1876. https://doi.org/10.1016/j.febslet.2006.02.050.

(26) Deseke, E.; Nakatani, Y.; Ourisson, G. Intrinsic Reactivities of Amino Acids towards Photoalkylation with Benzophenone − A Study Preliminary to Photolabelling of the Transmembrane Protein Glycophorin A. *Eur. J. Org. Chem.* **1998**, *1998* (2), 243–251. https://doi.org/10.1002/(SICI)1099-0690(199802)1998:2<243::AID-EJOC243>3.0.CO;2-I.

(27) Tanaka, Y.; R. Bond, M.; J. Kohler, J. Photocrosslinkers Illuminate Interactions in Living Cells. *Mol. Biosyst.* **2008**, *4* (6), 473–480. https://doi.org/10.1039/B803218A.

(28) Wittelsberger, A.; Mierke, D. F.; Rosenblatt, M. Mapping Ligand-Receptor Interfaces: Approaching the Resolution Limit of Benzophenone-Based Photoaffinity Scanning. *Chem. Biol. Drug Des.* **2008**, *71* (4), 380–383. https://doi.org/10.1111/j.1747-0285.2008.00646.x.

(29) Chu, F.; Thornton, D. T.; Nguyen, H. T. Chemical Cross-Linking in the Structural Analysis of Protein Assemblies. *Methods* **2018**, *144*, 53–63. https://doi.org/10.1016/j.ymeth.2018.05.023.

(30) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proc. Natl. Acad. Sci.* **2002**, *99* (17), 11020–11024. https://doi.org/10.1073/pnas.172226299.

(31) Kolhe, J. A.; Babu, N. L.; Freeman, B. C. The Hsp90 Molecular Chaperone Governs Client Proteins by Targeting Intrinsically Disordered Regions. *Mol. Cell* **2023**, *83* (12), 2035-2044.e7. https://doi.org/10.1016/j.molcel.2023.05.021.

(32) Kolhe, J. A.; Babu, N. L.; Freeman, B. C. Protocol for Establishing a Protein Interactome Based on Close Physical Proximity to a Target Protein within Live Budding Yeast. *STAR Protoc.* **2023**, *4* (4), 102663. https://doi.org/10.1016/j.xpro.2023.102663.

(33) Mohibullah, N.; Hahn, S. Site-Specific Cross-Linking of TBP in Vivo and in Vitro Reveals a Direct Functional Interaction with the SAGA Subunit Spt3. *Genes Dev.* **2008**, *22* (21), 2994–3006. https://doi.org/10.1101/gad.1724408.

(34) Mori, H.; Ito, K. Different Modes of SecY–SecA Interactions Revealed by Site-Directed in Vivo Photo-Cross-Linking. *Proc. Natl. Acad. Sci.* **2006**, *103* (44), 16159–16164. https://doi.org/10.1073/pnas.0606390103.

(35) Hino, N.; Okazaki, Y.; Kobayashi, T.; Hayashi, A.; Sakamoto, K.; Yokoyama, S. Protein Photo-Cross-Linking in Mammalian Cells by Site-Specific Incorporation of a Photoreactive Amino Acid. *Nat. Methods* **2005**, *2* (3), 201–206. https://doi.org/10.1038/nmeth739.

(36) Das, J. Aliphatic Diazirines as Photoaffinity Probes for Proteins: Recent Developments. *Chem. Rev.* **2011**, *111* (8), 4405–4417. https://doi.org/10.1021/cr1002722.

(37) Iacobucci, C.; Götze, M.; Piotrowski, C.; Arlt, C.; Rehkamp, A.; Ihling, C.; Hage, C.; Sinz, A. Carboxyl-Photo-Reactive MS-Cleavable Cross-Linkers: Unveiling a Hidden Aspect of Diazirine-Based Reagents. *Anal. Chem.* **2018**, *90* (4), 2805–2809. https://doi.org/10.1021/acs.analchem.7b04915.

(38) Suchanek, M.; Radzikowska, A.; Thiele, C. Photo-Leucine and Photo-Methionine Allow Identification of Protein-Protein Interactions in Living Cells. *Nat. Methods* **2005**, *2* (4), 261–268. https://doi.org/10.1038/nmeth752.

(39) Kohl, B.; Brüderlin, M.; Ritz, D.; Schmidt, A.; Hiller, S. Protocol for High-Yield Production of Photo-Leucine-Labeled Proteins in *Escherichia Coli*. *J. Proteome Res.* **2020**, *19* (8), 3100–3108. https://doi.org/10.1021/acs.jproteome.0c00105.

(40) Jecmen, T.; Tuzhilkin, R.; Sulc, M. Photo-Methionine, Azidohomoalanine and Homopropargylglycine Are Incorporated into Newly Synthesized Proteins at Different Rates and Differentially Affect the Growth and Protein Expression Levels of Auxotrophic and Prototrophic E. Coli in Minimal Medium. *Int. J. Mol. Sci.* **2023**, *24* (14), 11779. https://doi.org/10.3390/ijms241411779.

(41) Koch, N. G.; Goettig, P.; Rappsilber, J.; Budisa, N. Engineering Pyrrolysyl-tRNA Synthetase for the Incorporation of Non-Canonical Amino Acids with Smaller Side Chains. *Int. J. Mol. Sci.* **2021**, *22* (20), 11194. https://doi.org/10.3390/ijms222011194.

(42) Zhang, M.; Lin, S.; Song, X.; Liu, J.; Fu, Y.; Ge, X.; Fu, X.; Chang, Z.; Chen, P. R. A Genetically Incorporated Crosslinker Reveals Chaperone Cooperation in Acid Resistance. *Nat. Chem. Biol.* **2011**, *7* (10), 671–677. https://doi.org/10.1038/nchembio.644.

(43) Ai, H.; Shen, W.; Sagi, A.; Chen, P. R.; Schultz, P. G. Probing Protein–Protein Interactions with a Genetically Encoded Photo-Crosslinking Amino Acid. *ChemBioChem* **2011**, *12* (12), 1854–1857. https://doi.org/10.1002/cbic.201100194.

(44) Chou, C.; Uprety, R.; Davis, L.; Chin, J. W.; Deiters, A. Genetically Encoding an Aliphatic Diazirine for Protein Photocrosslinking. *Chem Sci* **2011**, *2* (3), 480–483. https://doi.org/10.1039/C0SC00373E.

(45) Nguyen, T.; Gronauer, T. F.; Nast-Kolb, T.; Sieber, S. A.; Lang, K. Substrate Profiling of Mitochondrial Caseinolytic Protease P via a Site-Specific Photocrosslinking Approach. *Angew. Chem.* **2022**, *134* (10), e202111085. https://doi.org/10.1002/ange.202111085.

(46) Kluger, R.; Alagic, A. Chemical Cross-Linking and Protein–Protein Interactions—a Review with Illustrative Protocols. *Bioorganic Chem.* **2004**, *32* (6), 451–472. https://doi.org/10.1016/j.bioorg.2004.08.002.

(47) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. Expanding the Chemical Cross-Linking Toolbox by the Use of Multiple Proteases and Enrichment by Size Exclusion Chromatography. *Mol. Cell. Proteomics* **2012**, *11* (3), M111.014126. https://doi.org/10.1074/mcp.M111.014126.

(48) Mendes, M. L.; Fischer, L.; Chen, Z. A.; Barbon, M.; O'Reilly, F. J.; Giese, S. H.; Bohlke-Schneider, M.; Belsom, A.; Dau, T.; Combe, C. W.; Graham, M.; Eisele, M. R.; Baumeister, W.; Speck, C.; Rappsilber, J. An Integrated Workflow for Crosslinking Mass Spectrometry. *Mol. Syst. Biol.* **2019**, *15* (9). https://doi.org/10.15252/msb.20198994.

(49) Yang, T.; Li, X.; Li, X. D. A Bifunctional Amino Acid to Study Protein–Protein Interactions. *RSC Adv.* **2020**, *10* (69), 42076–42083. https://doi.org/10.1039/D0RA09110C.

(50) Steigenberger, B.; Van Den Toorn, H. W. P.; Bijl, E.; Greisch, J.-F.; Räther, O.; Lubeck, M.; Pieters, R. J.; Heck, A. J. R.; Scheltema, R. A. Benefits of Collisional Cross Section Assisted Precursor Selection (Caps-PASEF) for Cross-Linking Mass Spectrometry. *Mol. Cell. Proteomics* **2020**, *19* (10), 1677–1687. https://doi.org/10.1074/mcp.RA120.002094.

(51) Tang, X.; Bruce, J. E. A New Cross-Linking Strategy: Protein Interaction Reporter (PIR) Technology for Protein–Protein Interaction Studies. *Mol. Biosyst.* **2010**, *6* (6), 939. https://doi.org/10.1039/b920876c.

(52) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; Chaignepain, S.; Chavez, J. D.; Claverol, S.; Cox, J.; Davis, T.; Degliesposti, G.; Dong, M.-Q.; Edinger, N.; Emanuelsson, C.; Gay, M.; Götze, M.; Gomes-Neto, F.; Gozzo, F. C.; Gutierrez, C.; Haupt, C.; Heck, A. J. R.; Herzog, F.; Huang, L.; Hoopmann, M. R.; Kalisman, N.; Klykov, O.; Kukačka, Z.; Liu, F.; MacCoss, M. J.; Mechtler, K.; Mesika, R.; Moritz, R. L.; Nagaraj, N.; Nesati, V.; Neves-Ferreira, A. G. C.; Ninnis, R.; Novák, P.; O'Reilly, F. J.; Pelzing, M.; Petrotchenko, E.; Piersimoni, L.; Plasencia, M.; Pukala, T.; Rand, K. D.; Rappsilber, J.; Reichmann, D.; Sailer, C.; Sarnowski, C. P.; Scheltema, R. A.; Schmidt, C.; Schriemer, D. C.; Shi, Y.; Skehel, J. M.; Slavin, M.; Sobott, F.; Solis-Mezarino, V.; Stephanowitz, H.; Stengel, F.; Stieger, C. E.; Trabjerg, E.; Trnka, M.; Vilaseca, M.; Viner, R.; Xiang, Y.; Yilmaz, S.; Zelter, A.; Ziemianowicz, D.; Leitner, A.; Sinz, A. First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961. https://doi.org/10.1021/acs.analchem.9b00658.

(53) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and

Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 742. https://doi.org/10.1038/s41467-020-14608-2.

(54) Meyer, J. G. Qualitative and Quantitative Shotgun Proteomics Data Analysis from Data-Dependent Acquisition Mass Spectrometry. In *Shotgun Proteomics*; Carrera, M., Mateos, J., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2021; Vol. 2259, pp 297–308. https://doi.org/10.1007/978-1-0716-1178-4_19.

(55) Guo, J.; Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **2020**, *92* (12), 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.

(56) Krasny, L.; Huang, P. H. Data-Independent Acquisition Mass Spectrometry (DIA-MS) for Proteomic Applications in Oncology. *Mol. Omics* **2021**, *17* (1), 29–42. https://doi.org/10.1039/D0MO00072H.

(57) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC−MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793. https://doi.org/10.1021/pr101060v.

(58) Bourmaud, A.; Gallien, S.; Domon, B. Parallel Reaction Monitoring Using quadrupole-Orbitrap Mass Spectrometer: Principle and Applications. *PROTEOMICS* **2016**, *16* (15–16), 2146–2159. https://doi.org/10.1002/pmic.201500543.

(59) Jiang, Y.; Rex, D. A. B.; Schuster, D.; Neely, B. A.; Rosano, G. L.; Volkmar, N.; Momenzadeh, A.; Peters-Clarke, T. M.; Egbert, S. B.; Kreimer, S.; Doud, E. H.; Crook, O. M.; Yadav, A. K.; Vanuopadath, M.; Hegeman, A. D.; Mayta, M. L.; Duboff, A. G.;

Riley, N. M.; Moritz, R. L.; Meyer, J. G. Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry. *ACS Meas. Sci. Au* **2024**, acsmeasuresciau.3c00068. https://doi.org/10.1021/acsmeasuresciau.3c00068.

(60) Schmidt, A.; Claassen, M.; Aebersold, R. Directed Mass Spectrometry: Towards Hypothesis-Driven Proteomics. *Curr. Opin. Chem. Biol.* **2009**, *13* (5–6), 510–517. https://doi.org/10.1016/j.cbpa.2009.08.016.

(61) Hoopmann, M. R.; Merrihew, G. E.; Von Haller, P. D.; MacCoss, M. J. Post Analysis Data Acquisition for the Iterative MS/MS Sampling of Proteomics Mixtures. *J. Proteome Res.* **2009**, *8* (4), 1870–1875. https://doi.org/10.1021/pr800828p.

(62) Picotti, P.; Aebersold, R.; Domon, B. The Implications of Proteolytic Background for Shotgun Proteomics. *Mol. Cell. Proteomics* **2007**, *6* (9), 1589–1598. https://doi.org/10.1074/mcp.M700029-MCP200.

(63) Schmidt, R.; Böhme, D.; Singer, D.; Frolov, A. Specific Tandem Mass Spectrometric Detection of AGE-Modified Arginine Residues in Peptides: Tandem Mass Spectrometry for AGE Detection. *J. Mass Spectrom.* **2015**, *50* (3), 613–624. https://doi.org/10.1002/jms.3569.

(64) Jaffe, J. D.; Keshishian, H.; Chang, B.; Addona, T. A.; Gillette, M. A.; Carr, S. A. Accurate Inclusion Mass Screening. *Mol. Cell. Proteomics* **2008**, *7* (10), 1952–1962. https://doi.org/10.1074/mcp.M800218-MCP200.

(65) Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; MacLean, B.; MacCoss, M. J. The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. *Mass Spectrom. Rev.* **2020**, *39* (3), 229–244. https://doi.org/10.1002/mas.21540.

(66) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(67) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(68) Da Veiga Leprevost, F.; Haynes, S. E.; Avtonomov, D. M.; Chang, H.-Y.; Shanmugam, A. K.; Mellacheruvu, D.; Kong, A. T.; Nesvizhskii, A. I. Philosopher: A Versatile Toolkit for Shotgun Proteomics Data Analysis. *Nat. Methods* **2020**, *17* (9), 869–870. https://doi.org/10.1038/s41592-020-0912-y.

(69) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392. https://doi.org/10.1021/ac025747h.

(70) Paulo, J. A. Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. *WebmedCentral* **2013**, *4* (10), WMCPLS0052. https://doi.org/10.9754/journal.wplus.2013.0052.

(71) Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. Automated Approach for Quantitative Analysis of Complex Peptide Mixtures from Tandem Mass Spectra. *Nat. Methods* **2004**, *1* (1), 39–45. https://doi.org/10.1038/nmeth705.

(72) Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V.; Wu, C. C.; MacCoss,

M. J. Multiplexed MS/MS for Improved Data-Independent Acquisition. *Nat. Methods* **2013**, *10* (8), 744–746. https://doi.org/10.1038/nmeth.2528.

(73) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717. https://doi.org/10.1074/mcp.O111.016717.

(74) Hu, A.; Noble, W. S.; Wolf-Yadlin, A. Technical Advances in Proteomics: New Developments in Data-Independent Acquisition. *F1000Research* **2016**, *5*, 419. https://doi.org/10.12688/f1000research.7042.1.

(75) Picotti, P.; Bodenmiller, B.; Aebersold, R. Proteomics Meets the Scientific Method. *Nat. Methods* **2013**, *10* (1), 24–27. https://doi.org/10.1038/nmeth.2291.

(76) Yocum, A. K.; Chinnaiyan, A. M. Current Affairs in Quantitative Targeted Proteomics: Multiple Reaction Monitoring-Mass Spectrometry. *Brief. Funct. Genomic. Proteomic.* **2009**, *8* (2), 145–157. https://doi.org/10.1093/bfgp/eln056.

(77) Murray, K. K.; Boyd, R. K.; Eberlin, M. N.; Langley, G. J.; Li, L.; Naito, Y. Definitions of Terms Relating to Mass Spectrometry (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (7), 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06.

(78) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol. Cell. Proteomics* **2012**, *11* (11), 1475–1488. https://doi.org/10.1074/mcp.O112.020131.

(79) Shi, T.; Song, E.; Nie, S.; Rodland, K. D.; Liu, T.; Qian, W.; Smith, R. D. Advances in Targeted Proteomics and Applications to Biomedical Research. *PROTEOMICS* **2016**, *16* (15–16), 2160–2182. https://doi.org/10.1002/pmic.201500449.

(80) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput. *Nat. Methods* **2020**, *17* (1), 41–44. https://doi.org/10.1038/s41592-019-0638-x.

(81) Barkovits, K.; Pacharra, S.; Pfeiffer, K.; Steinbach, S.; Eisenacher, M.; Marcus, K.; Uszkoreit, J. Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-Based Data-Independent Acquisition. *Mol. Cell. Proteomics* **2020**, *19* (1), 181–197. https://doi.org/10.1074/mcp.RA119.001714.

(82) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258–264. https://doi.org/10.1038/nmeth.3255.

(83) Schmidt, A.; Beck, M.; Malmström, J.; Lam, H.; Claassen, M.; Campbell, D.; Aebersold, R. Absolute Quantification of Microbial Proteomes at Different States by Directed Mass Spectrometry. *Mol. Syst. Biol.* **2011**, *7* (1), 510. https://doi.org/10.1038/msb.2011.37.

(84) Vidova, V.; Spacil, Z. A Review on Mass Spectrometry-Based Quantitative Proteomics: Targeted and Data Independent Acquisition. *Anal. Chim. Acta* **2017**, *964*, 7–23. https://doi.org/10.1016/j.aca.2017.01.059.

(85) Barkovits, K.; Chen, W.; Kohl, M.; Bracht, T. Targeted Protein Quantification Using Parallel Reaction Monitoring (PRM). In *Quantitative Methods in Proteomics*; Marcus, K.,

Eisenacher, M., Sitek, B., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2021; Vol. 2228, pp 145–157. https://doi.org/10.1007/978-1-0716-1024-4_11.

(86) Canessa, E. H.; Goswami, M. V.; Alayi, T. D.; Hoffman, E. P.; Hathout, Y. Absolute Quantification of Dystrophin Protein in Human Muscle Biopsies Using Parallel Reaction Monitoring (PRM). *J. Mass Spectrom.* **2020**, *55* (2), e4437. https://doi.org/10.1002/jms.4437.

(87) Fernández-Costa, C.; Martínez-Bartolomé, S.; McClatchy, D. B.; Saviola, A. J.; Yu, N.-K.; Yates, J. R. Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *J. Proteome Res.* **2020**, *19* (8), 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153.

(88) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *J. Proteome Res.* **2015**, *14* (5), 2190–2198. https://doi.org/10.1021/pr501321h.

(89) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics. *Proteomics Clin. Appl.* **2015**, *9* (7–8), 745–754. https://doi.org/10.1002/prca.201400164.

(90) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (1), 76–87. https://doi.org/10.1007/s13361-011-0261-2.

(91) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated Assignment of MS/MS Cleavable Cross-Links in Protein 3D-Structure Analysis. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 83–97. https://doi.org/10.1007/s13361-014-1001-1.

(92) Iacobucci, C.; Götze, M.; Ihling, C. H.; Piotrowski, C.; Arlt, C.; Schäfer, M.; Hage, C.; Schmidt, R.; Sinz, A. A Cross-Linking/Mass Spectrometry Workflow Based on MS-Cleavable Cross-Linkers and the MeroX Software for Studying Protein Structures and Protein–Protein Interactions. *Nat. Protoc.* **2018**, *13* (12), 2864–2889. https://doi.org/10.1038/s41596-018-0068-8.

(93) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(94) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

(95) Yu, C.; Wang, X.; Huang, L. Developing a Targeted Quantitative Strategy for Sulfoxide-Containing MS-Cleavable Cross-Linked Peptides to Probe Conformational Dynamics of Protein Complexes. *Anal. Chem.* **2022**, *94* (10), 4390–4398. https://doi.org/10.1021/acs.analchem.1c05298.

Chapter 2. A High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid

This chapter is reproduced with permission from Lindsey D. Ulmer, Daniele Canzani, Christopher N. Woods, Natalie L. Stone, Maria K. Janowska, Rachel E. Klevit, Matthew F. Bush "High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid" *J. Proteome Res.* **2024**. *23* (8), 3560-3570 https://doi.org/10.1021/acs.jproteome.4c00194

2.1 Abstract

In conventional crosslinking mass spectrometry, proteins are crosslinked using a highly selective, bifunctional chemical reagent, which limits crosslinks to residues that are accessible and reactive to the reagent. Genetically incorporating a photoreactive amino acid offers two key advantages: any site can be targeted, including those that are inaccessible to conventional crosslinking reagents, and photoreactive amino acids can potentially react with a broad range of interaction partners. However, broad reactivity imposes additional challenges for crosslink identification. In this study, we incorporate benzoylphenylalanine (BPA), a photoreactive amino acid, at selected sites in an intrinsically disordered region of the human protein HSPB5. We report and characterize a workflow for identifying and visualizing residue-level interactions originating from BPA. We routinely identify 30 to 300 crosslinked peptide spectral matches with this workflow, which is up to ten times more than existing tools for residue-level BPA crosslink identification. Most identified crosslinks are assigned to a precision of one or two residues, which is supported by a high degree of overlap between replicate analyses. Based on these results, we anticipate that this workflow will support the more general use of genetically

incorporated, photoreactive amino acids for characterizing the structures of proteins that have resisted high-resolution characterization.

2.2 Introduction

Crosslinking mass spectrometry (MS) is a powerful method for identifying protein-protein interactions and characterizing the spatial relationships within macromolecular assemblies.[1] Crosslinking MS is especially useful for studying protein dynamics and transient interactions, thereby capturing information that is often challenging to obtain through standard structural biology methods.[2,3] In conventional crosslinking MS, proteins in a sample are reacted with a bifunctional chemical reagent, digested enzymatically, and then analyzed by liquid chromatography (LC) MS. Crosslinked peptides identified through this process are used to infer protein-protein interactions and distance constraints between protein residues. Traditional crosslinking MS is challenged by the chemistry of crosslinking reagents and the identification of crosslinked products,[2] which has led to many lab-specific bioinformatic workflows.[4,5] There are additional challenges related to refining structures based on the inferred distance constraints.

The most-used crosslinking reagents, e.g., disuccinimidyl suberate (DSS) and bis(sulfosuccinimidyl)suberate (BS3), react with primary amines.[6] These conventional crosslinkers can only detect interactions that involve solvent accessible, primary amines, i.e., the sidechain of lysine and the N-terminal amine. However, the limited number of reactive amino acids makes it easier to identify crosslink sites because only a subset of residues can participate in crosslinks. Conventional crosslinkers may react with any exposed residue with the correct functional group in the sample, i.e., the reaction is untargeted. This typically results in a wide variety of low-intensity products that are difficult to detect, especially for the low abundance and transient interactions that are expected in heterogenous protein systems.[7] Fragmentation spectra of ions of crosslinked peptides can include contributions from both peptides, which contributes to the challenges of making confident peptide spectral matches (PSMs).[4,5]

An alternative strategy to conventional crosslinking is to incorporate photoreactive amino acids into the protein sequence that can potentially react with any amino acid when exposed to UV light,[8,9] albeit with different efficiencies.[10] For example, photo-methionine and photo-leucine both have diazirine functional groups that react to form crosslinks via a carbene intermediate.[11,12] These amino acids are commonly incorporated by including those artificial amino acids in restricted media lacking the canonical amino acid, resulting in proteome-wide, albeit incomplete, incorporation.[13] However, photo-methionine and photo-leucine can also form electrophiles that generate side products.[11,12] Benzoylphenylalanine (BPA) can also form crosslinks[14] and photoactivated BPA typically relaxes back to the ground state if no crosslink is formed, enabling few side products and high crosslinking yields.[15] BPA is amenable to site-specific incorporation using amber codon suppression, enabling highly targeted and complete incorporation.[16,17] Prior studies of calmodulin and a BPA-containing variant of the calmodulin-binding peptide suggest that calmodulin binds the peptide variant with the same affininty.[14] X-ray crystallography of a BPA-crosslinked complex of the liver oncoprotein gankyrin and the C-terminal domain of the S6 proteasomal protein indicates that the presence of the BPA crosslink at the interface did not otherwise affect the structures of the complex.[18] This approach enables the targeting of sites in proteins, even those that are solvent inaccessible, because all crosslinks will include the selected site of incorporation.

Crosslinks originating from photo-methionine, photo-leucine, and BPA have been identified using database searching with programs including StavroX,[19–23] MeroX,[24,25] and Crossfinder.[20] Relative to methods for conventional crosslinking reagents, it can be challenging to unambiguously identify the specific residue involved in crosslinks for photoreactive amino acids. A larger variety of crosslink sites can form because the crosslinker can potentially react

with any residue; incomplete sequence coverage in the fragmentation spectrum can result in ambiguities of the specific residue participating in the crosslink. Because of these challenges, many studies utilizing photoreactive amino acids only analyze products at the protein level, e.g., using gel-based assays to determine whether crosslinks were formed to a target[26–28] or using quantitative proteomics to characterize the preferential co-isolation of interacting proteins after exposure to UV light.[29,30] Some studies have localized the site of crosslinking, but often only a single crosslink is reported.[31,32]

To overcome the challenges that have hindered the broader use of photoreactive amino acids to identify residue-specific interactions, we developed experimental methods and an informatics workflow (Figure 1) and then characterized the performance of that workflow for variants of the human small heat shock protein (sHSP) HSPB5 that each contain BPA incorporated at a single site. The informatics workflow benefits from the use of msconvert,[33] Comet,[34] Kojak,[35,36] PeptideProphet,[37] and other tools from the Trans Proteomic Pipeline (TPP),[38] which is a suite of open-source tools for MS data analysis. Specifically, the speed and transparency of Kojak and PeptideProphet supported our development of tools for representing ambiguities in the residue-level assignment of crosslinks originating from photoreactive amino acids. We also compared this new workflow with ones reported previously using StavroX[39] and MeroX.[40,41] Using the workflow described in Figure 1, we identified residue-level interactions originating from the disordered N-terminal region (NTR) of the human sHSP HSPB5. Despite the importance of human sHSPs as chaperones,[42,43] this class of proteins remains under characterized due to their intractability to conventional structural biology approaches; up to half of the protein sequence is disordered and these proteins assemble into large, polydisperse, dynamic oligomers.[44–47] Here, we report and characterize the workflow that we applied recently

to identify novel, residue-level interactions that were key factors in characterizing NTR-NTR interactions[48,49] and the origin of selectivity[48] of HSPB5. These results indicate that this strategy offers great potential for the more general use of genetically incorporated, photoreactive amino acids to study protein targets that include elements of heterogeneity and intrinsic disorder.

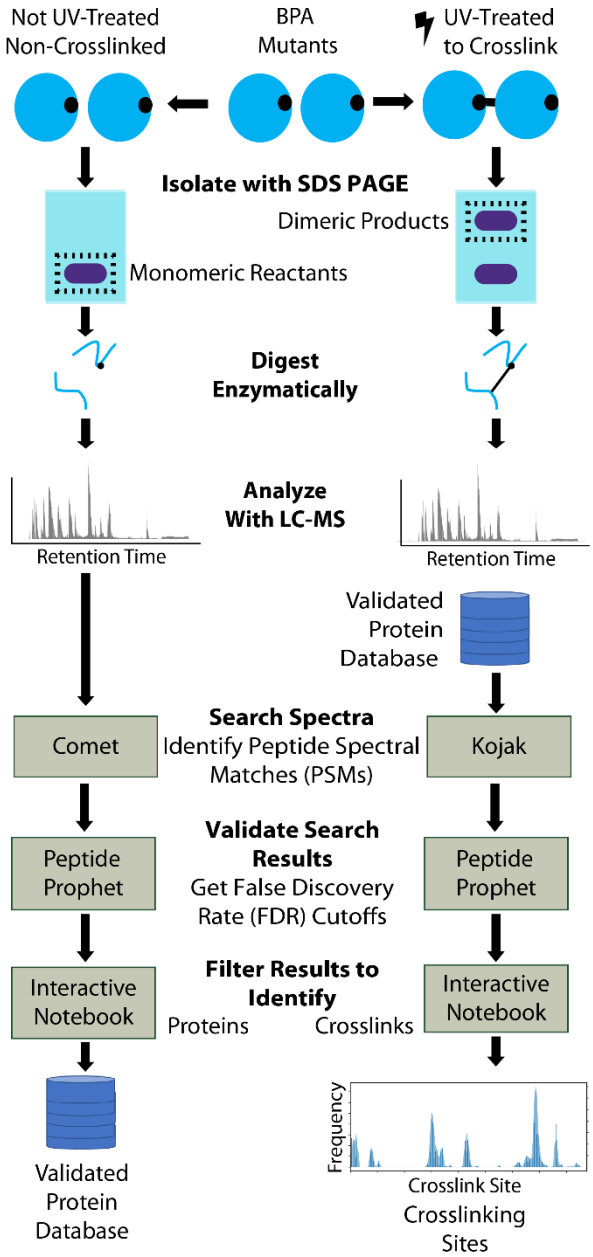

Figure 1. The BPA-containing variant proteins were first purified without the use of antibodies or purification tags, as described previously.[48] Then, samples are

divided into two fractions prior to the crosslinking reaction to have a non-crosslinked control. The left illustrates the workflow for the analysis of the monomeric reactants to create the validated-protein database. Alternatively, a validated-protein database can be created using the LC-MS data from the dimeric products. The right describes the workflow for the analysis of the dimeric products, which was used to identify site-specific crosslinks.

2.3 Methods

2.3.1 Sample Preparation and Analysis

The BPA-containing HSPB5 variants W9B, F17B, F24B, L33B, F47B, and F61B were prepared in BL21 *E. coli* cells using amber codon suppression.[17] Details of the cell growth, cell lysis, and protein purification for these variants are described elsewhere.[48,49] Purified BPA-containing variants were UV-treated to form crosslinks, and the product mixtures were subjected to SDS-PAGE on precast 4–20% acrylamide gradient gels (Bio-Rad, 4561096) as described elsewhere.[48,49] The monomeric reactants (proteins in the monomer band from the non-UV treated sample) and dimeric products (proteins in the dimer band from the UV-treated sample) were excised and were each digested in-gel following the procedure for the Thermo In-Gel Digestion Kit.[50] The weight (μg) of protein loaded onto the gel and the relative color intensity of the band excised from the gel prior to digestion was used to estimate the weight of the digested peptides loaded on the column. For example, when 7.6 μg of protein was loaded onto a gel lane and the dimeric product band was roughly 30% of the total color intensity, we estimated that the excised dimeric products contained about 2.3 μg of protein. In that case, about 20% (500 ug) of the digested peptides were loaded on the column. For trypsin-GluC-digested samples, GluC and

trypsin were each added to the digestion buffer at 0.004 $mg \cdot mL^{-1}$. Samples were digested overnight at 37 ºC, and then prepared for LC-MS using C18 spin columns (Thermo Scientific Pierce, 89870). Samples were analyzed using an Easy Nano LC system coupled to a Thermo Orbitrap Fusion Lumos Tribrid and data-dependent acquisition, as described in the Supporting Information. Either a 30- or 85-minute gradient was used. Effects of the gradient length are discussed in the Supporting Information and shown in Figure S1.

2.3.2 Identification of Crosslinks Using TPP

The left column of Figure 1 schematically shows the process of identifying the proteins that are present in the monomeric reactant sample. TPP version 6.3.2 was used; some effects of the software version are discussed in the Supporting Information. First, Comet[34] was used to search for non-crosslinked peptides in the non-UV-treated control to construct the validated-protein database. The search database contained the BL21 *E.coli* database from UniProt (UP000431028), the cRAP database from the Global Proteome Machine with all 5 levels of proteins,[51] the pertinent BPA-containing variant of HSPB5, peptides used for quality control (AngioNeuro), and reverse-sequence decoys. Samples were searched using Comet and validated using PeptideProphet through using a non-enzyme constrained search as described in the Supporting Information. After filtering using a 1% False Discovery Rate at the PSM level (FDR) and a minimum of 2 peptide spectral matches (PSMs), this process yields a validated protein database for the sample. Alternatively, this workflow can be used to identify the proteins that are present in the dimeric product sample.

The right column of Figure 1 schematically shows the process of identifying the crosslinks that are present in the UV-treated samples using Kojak[35,36] and the validated-protein database. Kojak version 2.0.3 was used; some effects of the software version are discussed in the

Supporting Information. The search settings used are described in detail in the Supporting Information, but mimic those used for Comet except a narrower precursor tolerance and enzyme selection rules were used. For histograms, each PSM was associated with the residue that was assigned the highest probability of participating in a crosslink with BPA. When more than one residue was assigned the same probability, an equal fraction of that PSM was assigned to each of those residues.

2.3.3 Access to Data and Software

The mass spectrometry data, FASTA files, search parameters, and PeptideProphet results have been deposited to the ProteomeXchange Consortium via the PRIDE[52] partner repository with the dataset identifier PXD050493. That repository also contains a data summary that relates all data to the corresponding Figures and Tables reported here. An interactive notebook and sample files for generating residue-specific crosslinking distributions and error-sensitivity plots have been deposited to https://github.com/bushgroup/Identifying-Site-Specific-Crosslinks.

2.4 Results and Discussion

The objective of this research was to develop a high-performance workflow for identifying residue-specific crosslinks originating from photoreactive amino acids. Towards that end, we developed experimental methods and an informatics workflow, which uses Comet,[34] Kojak,[35,36] PeptideProphet,[37] and other open-source tools from the Trans Proteomic Pipeline (TPP).[38] Our workflow is described in detail in the Methods section and is shown schematically in Figure 1. We then characterized the performance of this workflow for variants of human HSPB5 that each contain BPA incorporated at a single site. sHSPs form large, polydisperse, and dynamic oligomers. Each individual protein includes the following structural elements: a disordered N-terminal region (NTR), an ordered α-crystallin domain (ACD) that folds into two

anti-parallel β-sheets, and a disordered C-terminal region (CTR).[44–47] Tertiary structure is maintained through inter-molecular interactions between HSPB5 subunits. The wide variety of interactions within the oligomer has made it challenging to achieve consensus models for higher-order structures of sHSPs, which is evidenced by the significant differences between the structures proposed for 24mers of HSPB5.[53,54] The large difference between those structures illustrates the need for a structural method that is more tolerant of disorder and heterogeneity. Here, we study HSPB5 variants with BPA incorporated into the disordered NTR to characterize a highly heterogeneous protein that has resisted characterization by conventional structural biology approaches.

2.4.1 Identifying the Proteins in a Sample

As shown in the left column of Figure 1, we analyzed non-UV-treated control samples to generate validated-protein databases. Although the BPA-containing variants used in this study were purified without the use of antibodies or purification tags,[48] chaperones such as HSPB5 often co-purify with other proteins. To gain the broadest understanding of the proteins present in the sample, we consider all proteins from the *E. coli* expression system and potential contaminants (e.g., keratins and common proteins associated with molecular biology)[51] that could be introduced during sample handling. We used a non-enzyme specific search and then filtered the results to all proteins with at least 2 PSMs at a 1% FDR. This list of proteins is used as the validated-protein database for the sample. Across all samples analyzed, the size of the validated-protein databases ranged from 11 to 13 for samples analyzed using an 85-minute gradient (Table S1) and from 17 to 21 for samples analyzed using a 30-minute gradient (Table S2). Experiments using a shorter gradient identified a larger number of proteins; this may be

attributable to narrower peak widths and greater peak heights during experiments with shorter gradients.

2.4.2 Effects of Protein Database Size.

As shown in the right column of Figure 1, we then analyzed dimeric products to identify crosslinks originating from the incorporated BPA. We first considered a sample with BPA incorporated at site 9 (W9B) that was digested with trypsin; that sample yielded a validated-protein database containing 12 proteins. Figure 2A shows the number of crosslink peptide-spectra matches (PSMs) identified as a function of the number of proteins in the search database. Interestingly, the validated-protein database yielded the most crosslink PSMs; the crosslinks are shown in Figure S2. In general for target-decoy searches, increasing the number of proteins in the database would be expected to decrease sensitivity (how many PSMs are identified) because experimental spectra will be scored against additional targets and decoys, some of which may yield competitive scores. This trend is observed for searches using databases with more proteins than the validated-protein database and has been reported previously,[5] but using a single-protein database resulted in the identification of fewer crosslink PSMs than the validated-protein database (Figure 2A). The origin of the latter observation is not fully understood, but we hypothesize that using a single-protein database, which considers far fewer targets and decoys, may lower the quality of the validation models. In general, for target-decoy searches, increasing database size to consider more of the proteins that may be in the sample would allow for the identification of crosslinks with those additional proteins, which would be expected to increase specificity by decreasing the likelihood of a spectrum being assigned to the incorrect candidate. Therefore, using a single-protein database also increases the risk of false positives because spectra that correspond best to other proteins may be misidentified. These results suggest that the

use of the validated-protein database minimizes the loss of sensitivity from considering additional proteins, while still providing the benefits of considering signals that may originate from other proteins that are in the sample.[5]

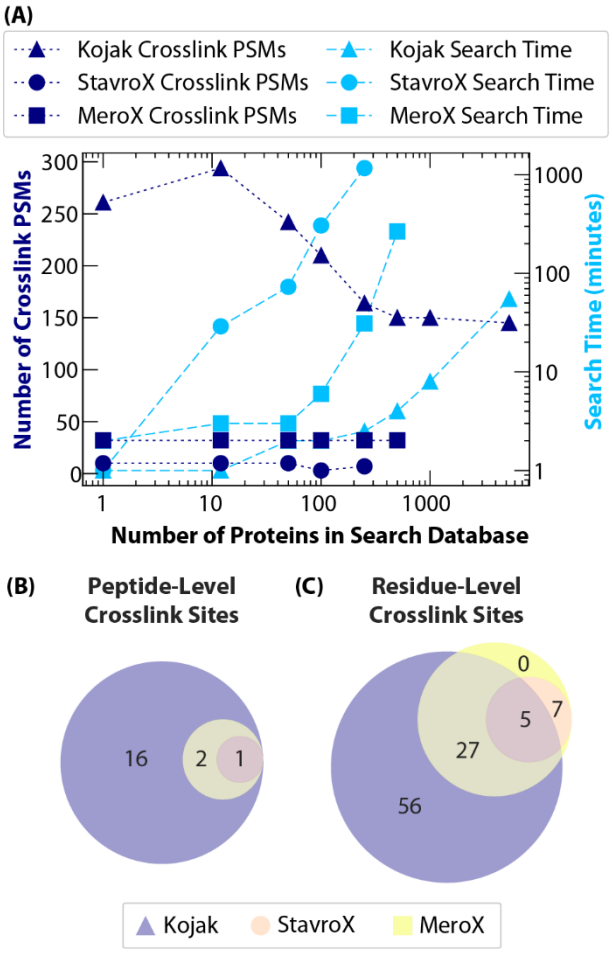

Figure 2. (A) The number of crosslink PSMs found and the corresponding search times are shown as a function of database size and using Kojak, StavroX, or MeroX. These searches were all performed on the same LC-MS data of trypsin-digested, dimeric products of W9B. StavroX identified 10 or fewer crosslink PSMs across database sizes. StavroX results are only reported for up to the 250-protein database because searches timed out for larger database sizes. MeroX identifies 32 crosslink PSMs across database sizes. MeroX results are only

reported for up to the 500-protein database because the search timed out at larger database sizes. We used a 2-PSM minimum at a 1% FDR as the criteria for the validated-protein database. In this plot, the 12-protein database is the validated-protein database. (B) At the peptide level, this Venn Diagram shows that using StavroX or MeroX resulted in the identification of a small subset of the crosslinked peptides identified using Kojak. (C) At the residue level, this Venn diagram shows that using MeroX or StavroX resulted in the identification of a smaller number of residue-level crosslink sites than Kojak, many of which were assigned with greater precision using Kojak.

The protein databases used in Figure 2A were all created using results for monomeric reactants. We have also created protein databases using results for dimeric products. For the W9B variant that was digested with trypsin, analysis of the monomeric reactants yielded a database with 12 proteins (Table S3), whereas analysis of the dimeric products yielded a database with 11 proteins (Table S3). Eight proteins were common to both: the target (the HSPB5 variant), trypsin, and other common contaminants such as human keratins. The unique proteins were mostly from the expression system (*E. coli*). Similar numbers of crosslink PSMs were identified with the two databases: 294 using the monomeric reactant database and 298 with the dimeric product database. The crosslinks identified are very similar and are shown in Figures S2A and S2B. 86 of the total of 89 unique crosslink sites are identified with both databases (Figure S2C), and the crosslink sites that are identified with just one database are very low intensity (1 PSM or less). Because similar crosslinks were identified, using either sample to create the validated protein database appears to be sufficient. In most of the results presented

here, we used monomeric reactants to create the validated-protein database. We used analysis of the dimeric products to create search databases for the 9 replicates of trypsin-digested W9B (Figure 3, Figure S3, and Table S2). Although crosslinks were identified using each method, creating the validated-protein database using data from the dimeric products does not require any additional LC-MS analysis (that sample is already analyzed to identify crosslinks).

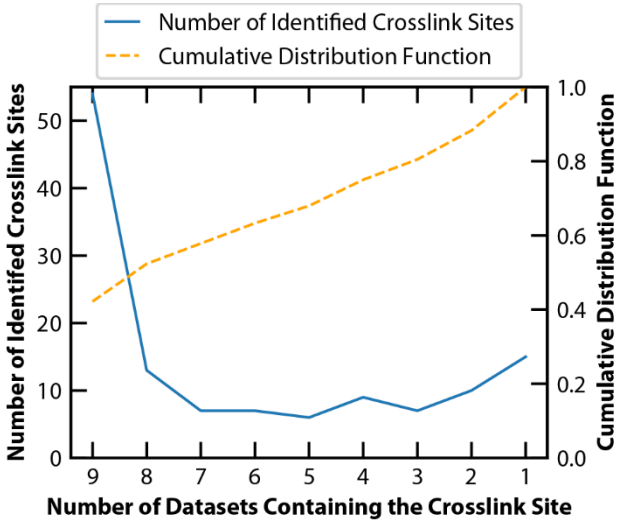

Figure 3. The results of nine replicates of trypsin-digested, dimeric products of W9B at pH 6.5 (Figure S3 and Table S3) are compared here. The number of crosslink sites identified (defined as the number of residues assigned a frequency value greater than zero in Figure S3) across differing numbers of replicates is indicated. A value of nine indicates that a given crosslink was identified in all nine datasets, whereas a value of one indicates that a given crosslink was only identified in a single dataset. Of the 128 total crosslink sites identified 54 are identified in all nine replicates.

2.4.3 Performance of Informatic Workflow

Figure 2A also shows the number of crosslink PSMs identified and the corresponding search times for our workflow, StavroX,[39] and MeroX.[40,41] Using the same LC-MS data for W9B, our informatics workflow identified numbers of crosslink PSMs ranging from 145 (when considering the 5315 proteins in the full BL21 *E.coli* and cRAP databases) to 294 (when considering the 12 proteins in the validated-protein database). For each of the smaller databases, StavroX identified 10 or fewer crosslink PSMs and MeroX identified 32 crosslink PSMs. The small number of PSMs was surprising given that StavroX[19–23] and MeroX[24,25] have been used to identify site-specific BPA crosslinks in previous studies and because we enriched crosslinks by only excising the band for the crosslinked HSPB5-HSPB5 dimeric product prior to in-gel digestion. For our workflow, the search time increased with the size of the database; searches using the 12 and 5315 protein databases finished within 1 and 55 minutes, respectively. In contrast, searches using StavroX and MeroX required significantly more time. StavroX took 29 minutes to complete the 12-protein database search and over 19 hours to complete the search using a 250-protein database. That was the largest protein database used for StavroX because searches using larger databases timed out and did not finish successfully. MeroX finished the search using a 12-protein database in 3 minutes, but the search using a 500-protein database required over 4 hours and those using larger protein databases timed out and did not finish successfully. These results demonstrate the excellent performance of our workflow, in terms of the large number of crosslink PSMs identified, the fast search times, and the scaling of those figures of merit with respect to database size

Figure 2B compares the crosslinks identified by the different programs. MeroX and StavroX identified subsets of the 19 crosslinked peptides identified by Kojak. This illustrates that

the identified crosslinks are consistent with each other at the peptide level. At the residue-level, there is also a high degree of similarity in the crosslink sites identified across methods. Of the 95 total residue-level crosslink sites in Figure 2C, only 7 were identified by StavroX and MeroX and not identified by Kojak. These 7 sites are attributed to additional ambiguity in the crosslink assignments with StavroX and MeroX. SI Tables 4 and 5 list the StavroX and MeroX results in more detail, but briefly, these differences result from every residue within the crosslinked peptide being considered as a potential crosslink site. Overall, the crosslinks identified by MeroX and StavroX are consistent with crosslinks identified with Kojak, but Kojak identifies more crosslinks.

2.4.4 Validating Crosslinks

To summarize and visualize the results from the searches for crosslinks, we developed a Python-based, interactive notebook to integrate results from Kojak and PeptideProphet. The identified crosslinks are filtered to a target FDR value (typically 1%), which is based on probabilities from PeptideProphet. Each probability is based on the validation model that is generated for ions with that charge state. The vast majority of the crosslink PSMs that met these criteria only include peptides originating from HSPB5, i.e., HSPB5-HSPB5 crosslink PSMs. The number of crosslink PSMs that included a peptide originating from a different protein, i.e., HSPB5-nontarget crosslink PSMs, are reported for 13 samples containing one of six variants in Table S1 and for 9 replicates of W9B in Table S2. When considering all 22 of those analyses, the number of HSPB5-nontarget crosslink PSMs ranged from zero (for one sample with 34 HSPB5-HSPB5 crosslink PMSs) to 9 (for one sample with 375 HSPB5-HSPB5 crosslink PSMs); the mean was 3.6 and the median was 3. These nontarget peptides originated from *E. coli*, contaminant, and decoy proteins. Figure 4 shows that the number of crosslink PSMs identified

depends on FDR. Results are separated for PSMs that include a peptide from the target (HSPB5), another protein (an *E. coli* or contaminant protein), or a decoy protein. The number of HSPB5-HSPB5 crosslinks increases sharply at low FDR and levels off by about 2.5%. The number of crosslinks that include another protein or a decoy also increases sharply to an FDR of 2.5%, but then continues to increase with FDR. At lower FDR values, similar numbers of PSMs are identified that include a peptide from another protein or a decoy. At higher FDR values, the number of PSMs identified that include a decoy are greater than those that include another protein.

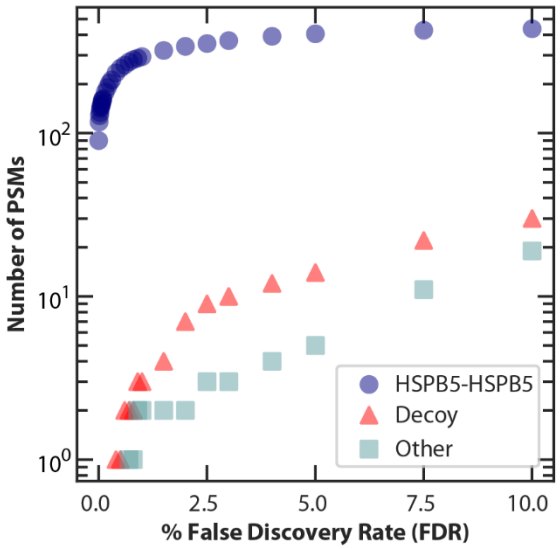

Figure 4. The number of HSPB5-HSPB5 crosslinks, decoy crosslinks, and other crosslinks identified as a function of FDR. The search represented in this data uses the validated-protein database (12 proteins) on the same dataset of trypsin-digested, dimeric products of W9B represented in Figure 2.

These FDR values are estimated from the validation models, which consider many factors including matches to decoys that are not in the sample.[55] The shapes of the curves in

Figure 4 are consistent with those that would be obtained for other proteomics experiments. That is, with increasing FDR the number of matches to decoys increases monotonically and the number of matches to targets (HSPB5-HSPB5 crosslinks) levels off as the sensitivity of the search approaches an asymptote. The curves for matches to HSPB5-nontarget crosslinks and matches to decoys have very similar shapes, suggesting that the former may be predominantly false positives. As for the analysis of other proteomics data, the objectives of high sensitivity and low error must be balanced. Here, we report all results at a 1% FDR at the PSM level. However, it would be reasonable to select a different FDR depending on the response of the sensitivity, the response of the error rate, and the objectives of the analysis.

StavroX and MeroX also use decoy-based validation to estimate FDR values at the PSM level. StavroX reports the scores associated with FDR cutoffs as histograms; Figure S4 shows a histogram generated for the search using a 12-protein database that was generated for the analysis in Figure 2. In that histogram, StavroX reports only six PSMs identified at scores where no decoys are identified. MeroX reports the scores that determine FDR cutoffs as histograms and as a plot relating the spectrum score to the FDR. Figure S5 shows the plots generated for the search using a 12-protein database that was generated for the analysis in Figure 2. In the histogram, MeroX reports 32 PSMs identified at scores where no decoys are identified. Because of the low number of PSMs at high scores, both StravoX and MeroX would have to be operated at a very high FDR rate to obtain the 294 HSPB5-HSPB5 crosslink PSMs that our workflow identified at a 1% FDR. A detailed FDR analysis, like that shown for our workflow in Figure 4, was not performed on StavroX or MeroX because of the smaller number of high-confidence PSMs.

2.4.5 Visualizing Crosslinks from Photoreactive Amino Acids

To illustrate the information content of these experiments, Figure 5 shows a visual representation of the crosslinks originating from BPA in the W9B variant. Throughout the crosslink results reported here, we indicate the structural elements common to all sHSPs: a disordered N-terminal domain (NTR), an α-crystallin domain (ACD) that folds into two anti-parallel β-sheets, and a disordered C-terminal domain (CTR).[44–47] Figure 5A depicts a peptide-level interpretation of our crosslinking results: for each PSM, a frequency of one was assigned to every residue within the crosslinked peptide. This representation mimics the information content of conventional crosslinking experiments, which use crosslinking reagents that can react with a limited subset of amino acids. At the peptide level, this visualization only enables a coarse understanding of the ensemble of interactions originating from the ninth residue of this protein (i.e., the BPA residue in the W9B variant).

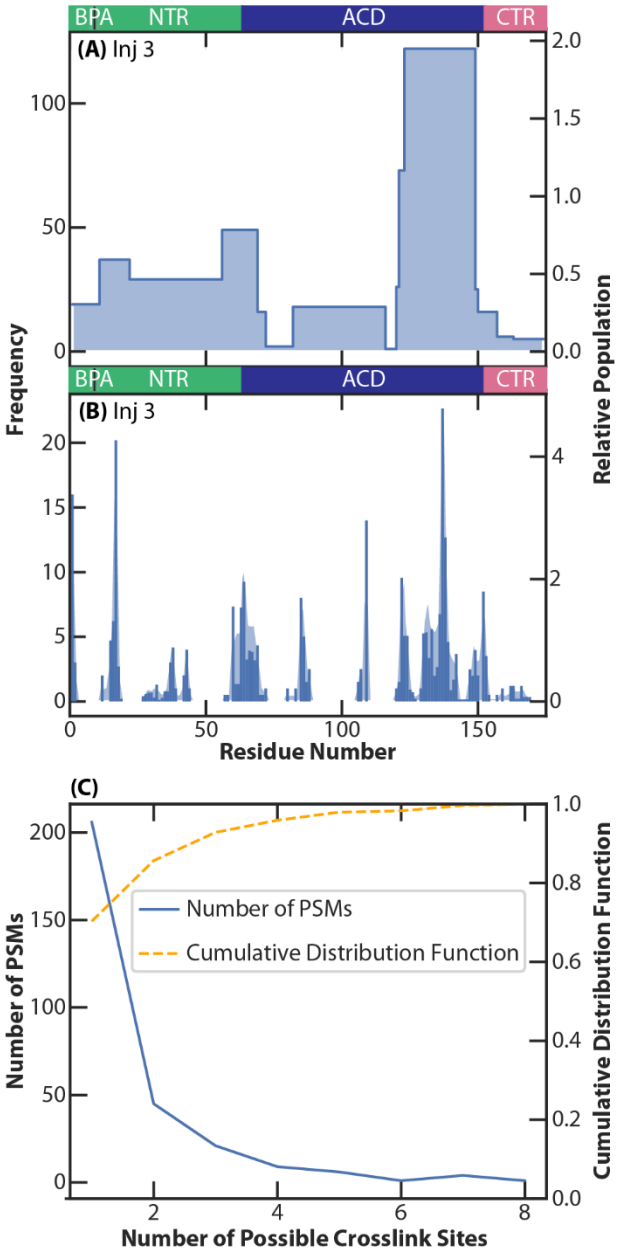

Figure 5. Above panels A and B, the BPA position and domain boundaries are shown for the structural regions of HSPB5: a disordered N-terminal region (NTR), an α-crystallin domain (ACD), and a disordered C-terminal region (CTR). All panels represent the same analysis of trypsin-digested W9B at pH 6.5, which is injection 3 in Figure S3 and Table S3. Panel A shows a peptide-level representation of the results, in which every residue in a crosslinked peptide received a frequency of one. Panel B shows a residue-level representation of the results, in which the PSM for the crosslinked peptide was distributed among the potential crosslinking sites to account for ambiguity as described in the text. Crosslink results are reported as both a histogram and a rolling average of three because of ambiguity in the crosslinking site. The frequency axis corresponds to the number of PSMs, and the relative population axis corresponds to the percent of total crosslink PSMs. Panel B has 293 crosslink PSMs. In Panel C, the number of possible crosslink sites (x-axis) indicates how many potential equivalent crosslinking sites a PSM has. Over 70% of the PSMs have no ambiguity in the crosslink site assignment, and over 80% of the crosslink PSMs have two possible crosslink sites or fewer.

Figure 5B depicts a residue-level interpretation of those same crosslinking results. For each PSM, we first found the number (*n*) of residues that were assigned the highest probability of participating in the crosslink and then assigned a frequency of 1/*n* to each of those residues. For example, a single residue that had the single highest probability was assigned a frequency of one (an unambiguous assignment) and two residues that had the same highest probability were each assigned a frequency of one half (an ambiguous assignment). At the residue level, this visualization enables a far more detailed understanding of the ensemble of interactions originating from the ninth residue of this protein. The contributions of ambiguous assignments will be discussed further in the following section.

BPA crosslinking reveals a dense network of residue-level interactions through detecting many different crosslinks from a single site of incorporation. Because of the density of information, these results are visualized as a histogram of single-residue crosslink sites detected from a single site of incorporation. In contrast to conventional crosslinkers, which provide coarse peptide-level results that are often visualized with lines connecting sites of crosslinking,[56] our results only require one end of the crosslink to be visualized because all crosslinks originate from the single BPA residue in each variant. The histograms convey the depth of residue-level information from BPA crosslinking without overcomplicating the visualization by indicating the site of BPA incorporation for each crosslink.

2.4.6 Ambiguities in Assigning Residue-Specific Crosslinks

When there are gaps in the coverage of b and y ions in the fragmentation spectrum of the crosslinked peptide ion, there can be ambiguity in the crosslink site assignment. As described above, **f**or each PSM, we first found the number (*n*) of residues that were assigned the highest

probability of participating in the crosslink (as determined from the Kojak results) and then assigned a frequency of 1/n to each of those residues. Figure 5C shows the frequency of ambiguities underlying the data in Figure 5B. Of the 293 crosslink PSMs, over 70% are assigned to a single residue (no ambiguity) and over 80% are assigned to one or two residues (no or some ambiguity). Therefore, using this process, the vast majority of crosslinks are assigned to a very precise region of the sequence. However, because the fragmentation spectra may include contributions from mixtures of crosslinks, it is possible that this workflow underestimates the heterogeneity of the crosslinking in the sample. To help account for that possibility, we also include depictions of rolling averages over a window of 3 residues (e.g., Figure 5B).

2.4.7 Reproducibility

A total of nine replicates were performed of the preceding analysis; one replicate is shown in Figure 5B and all are shown in Figure S3. These nine replicates analyzed three samples derived from two photo-crosslinking reactions, as described in Table S3. Across the nine replicates, the number of crosslink PSMs identified ranged from 293 to 375 (Table S3) and have remarkably similar crosslinking patterns. For example, Figure S3 shows that all replicates exhibit the highest number of crosslinks around residue 137 with less prominent clusters of crosslinks around residues 1, 17, 37, 64, 86, 109, 124, and 152. Figure 3 shows how many crosslink sites are identified in multiple replicates; a crosslink site identification is defined as a residue in the histogram that has a frequency value greater than zero. A total of 128 unique crosslink sites were identified across all 9 replicates; 54 were identified in all replicates and an additional 13 were identified in only 8 replicates. Only 15 crosslinks were identified in only a single replicate, but those crosslinks were also low in frequency when observed. The replicates reproducibly identify

the same predominant crosslinking sites, and the majority of the crosslinking sites are identified in at least 8 of the 9 replicates.

Our replicates of W9B crosslinks were not only highly reproducible with each other, but both corroborate results from complementary experiments and identify previously unreported interactions. For example, the most frequent crosslink from W9B is to site 137, which is within the edge groove of the ACD. Titration of a peptide containing residues 1−13 of HSPB5 against the isolated ACD caused chemical shifts in NMR signals assigned to residues in the edge grove, which suggested that the N-terminal residues represented by the peptide may bind to the edge groove in the full-length protein.[57] Most of the other identified crosslinks represent interactions originating from W9B that have not been reported previously; notably these crosslinks helped reveal a network of NTR-NTR interactions that had resisted characterization by other structural methods. The context and mechanistic implications of those results are reported elsewhere.[48,49] More generally, these results illustrate the promising potential of this workflow to robustly identify residue-level interactions in heterogenous systems with intrinsically disordered regions.

2.4.8 Effects of Digestion

The previous sections only considered samples isolated from dimeric products and then in-gel digested with trypsin. We will first discuss some factors related to the selection of the enzyme and then some factors related to the gel-based isolation. Figure 6 shows crosslinks identified from dimeric products of W9B using in-gel digestions with only trypsin or with a trypsin and GluC parallel digestion. HSPB5 has a 4 kDa tryptic peptide that spans from sites 23 to 56 (LFDQFFGEHLLESDLFPTSTSLSPFYLRPPSFLR). Based on trypsin-only digestion, Figure 6A (trypsin only) does not exhibit crosslinks to the region spanning the large tryptic peptide. However, Figure 6B (a parallel digestion) exhibits crosslinks to that region,

predominantly clustered near residue 43. The Venn diagram in Figure 6C illustrates that 28 crosslink sites are identified from both digestion conditions, 30 are identified with just the trypsin digestion, and 21 are identified with just the parallel digestion. The similarity of results across digestion conditions illustrates that this workflow is highly targeted and can reliably identify many interactions across different experimental conditions. For HSPB5, the selection of the digestion enzyme(s) does affect the sensitivity of the method to specific regions. The effect of digestion on the identification of specific crosslinks has been reported previously and has been attributed to factors including the mass of crosslinked peptides and to crosslinks hindering access to specific cleavage sites.[58–60]

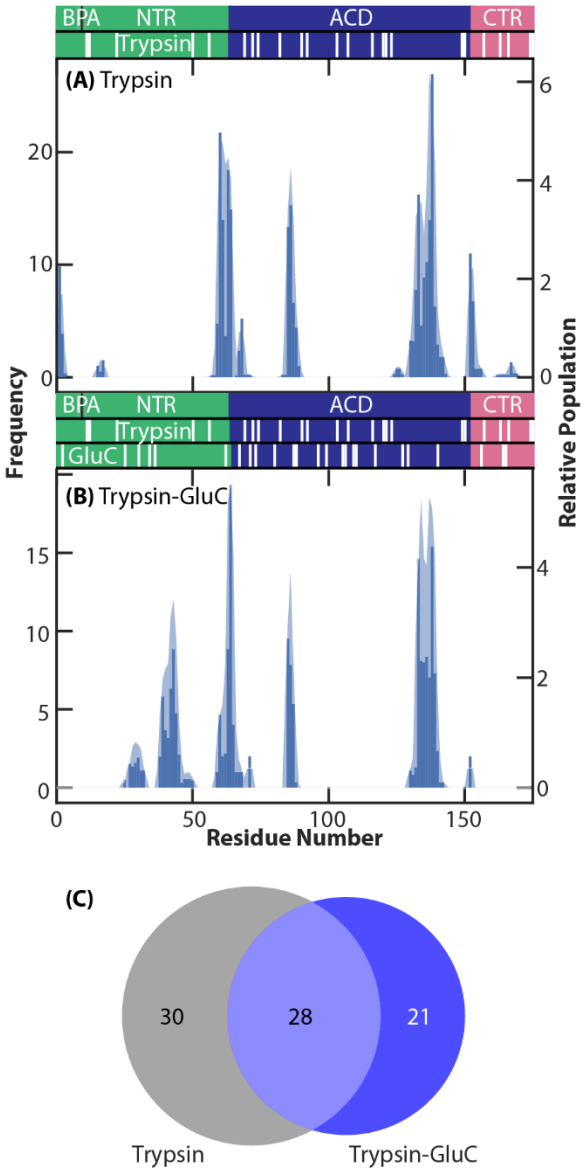

Figure 6. Above panels A and B, the top row shows the BPA position and domain boundaries. The following rows show the expected cleavage sites for trypsin or GluC. Panels A and B depict W9B samples with varying digestion enzymes. Panel A has 277 crosslink PSMs from trypsin-digested W9B at pH 6.5. Panel B has 195 crosslink PSMs from the trypsin-GluC-digested W9B at pH 6.5. The large tryptic peptide spans from sites 23 to 56, and we observe crosslinks to that region when using a trypsin-GluC, parallel digestion. The raw file used for the analysis in panel B has been reported previously.[48] In panel C, the Venn diagram illustrates the overlap in crosslink sites identified in trypsin-digested W9B at pH 6.5 (panel A) and trypsin-GluC-digested W9B at pH 6.5 (panel B).

To enrich crosslinked products, we used SDS-PAGE prior to in-gel digestion. After performing SDS-PAGE on denatured, crosslinked samples, we observe distinct monomer and dimer bands and trace amounts of higher-order products. We then excised and digested the dimeric products. In-gel digestion of this excised band using trypsin and GluC resulted in the identification of 195 crosslink PSMs. In contrast, in-solution digestion using trypsin and GluC of crosslinked samples without SDS PAGE results in the identification of only 12 crosslink PSMs. Therefore, in-gel digestion resulted in a ten-fold increase in the number of identifications relative to the in-solution digestion of the original mixture. In-gel digestion enriches inter-protein crosslinks because all dimeric products have at least one inter-protein crosslink. Excluding proteins in the monomer band removes non-crosslinked proteins from analysis. Because we are analyzing dimeric products, in which each variant only contains one BPA, at least half of the crosslinks are inter-protein. However, up to half of the dimeric products may have one intra-protein crosslink and one inter-protein crosslink.

2.4.9 Further Evidence of Site-Specific Crosslinks

This work demonstrates the performance and potential of this workflow using W9B; we have also analyzed F17B, F24B, L33B, F47B, and F61B using this workflow as reported elsewhere.[48] After trypsin-GluC parallel digestion, sites 24 and 33 are in the same peptide, LFDQFFGEHLLE (sites 23-34). Therefore, we will focus on comparing F24B and L33B, as shown in Figures 7A and 7B. Both sites show clusters of crosslinks around positions 1 and 15. However in F24B, the most frequent crosslinking site is at position 60, whereas in L33B the most frequent crosslink site is at position 3. Similar numbers of crosslink PSMs were detected in these samples (66 for F24B and 62 for L33B). Because these crosslinks originate from the same peptides after digestion and similar products should have similar ionization efficiencies, these

results suggest that site 24 likely interacts with residue 60 to a greater extent than site 33. Fourteen crosslink sites were identified in only the F24B sample, and 13 crosslink sites were identified in only the L33B sample. Only 14 of the crosslink sites identified were identified in both samples, so about half of the crosslink sites identified vary between the sites of incorporation. The large differences exhibited by variants with BPA located in the same peptide after digestion provides further supports that this workflow yields residue-level interactions.

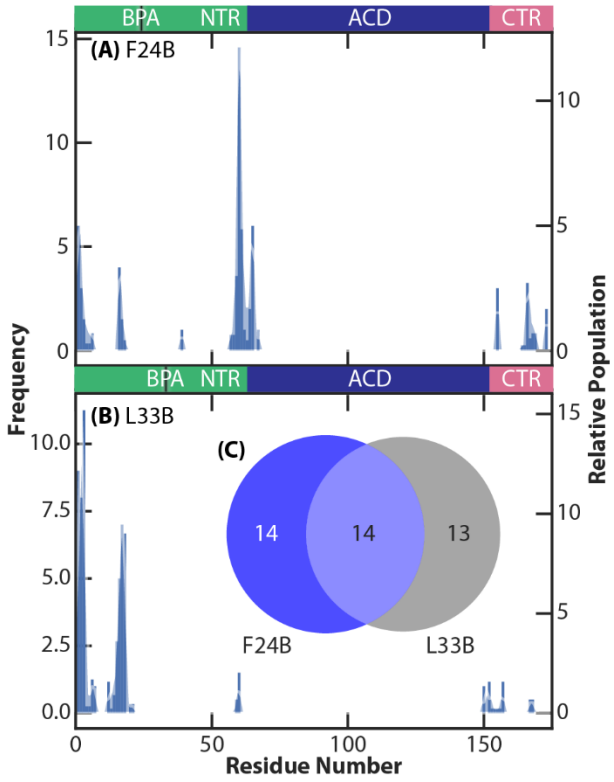

Figure 7. Panels A and B depict trypsin-GluC-digested samples with differing sites of BPA incorporation. Panel A has 66 crosslink PSMs from F24B. Panel B has 62 crosslink PSMs from L33B. The Venn diagram (panel C) illustrates the overlap in crosslink sites identified in F24B (panel A) and L33B (panel B). About half of the crosslink sites identified are found in both samples. The raw files used for this analysis have been reported previously.[49]

2.5 Conclusions

We developed a robust, high-performance workflow for identifying residue-level crosslinks originating from a genetically incorporated, photoreactive amino acid (Figure 1). The informatics workflow uses TPP tools that are free, open source, and updated regularly. We developed an interactive notebook that integrates results from Kojak and PeptideProphet in order to identify BPA crosslinks and account for ambiguities in the specific sites of crosslinking. Relative to existing methods for identifying BPA crosslinks, this workflow exhibits excellent performance in terms of the large number of crosslink PSMs identified, the fast search times, and the scaling of those figures of merit with respect to database size (Figure 2). This enables the routine identification of 30 to 300 crosslink PSMs (Table S1 and Table S2) for variants with a single BPA residue. The vast majority of the crosslinks identified are HSPB5-HSPB5 crosslinks, with comparatively small numbers of crosslinks to other proteins (Figure 4). This analysis suggests that the FDR estimate is conservative. The crosslinks we identify have low amounts of ambiguity and most are assigned with a precision of one or two residues (Figure 5C). The crosslinks we identify are highly similar across replicates (Figure 3) and exhibit key similarities under different digestion conditions as well (Figure 6). Crosslinks also differ when BPA is incorporated at different sites within the same peptide, further corroborating the residue-specific nature of the results (Figure 7). Using this workflow, we identified novel, distinct, and reproducible interactions of the highly disordered NTR.

The strategy described here differs substantially from conventional crosslinking. In conventional crosslinking, only solvent-accessible interactions can be detected and the limited range of amino acids that conventional crosslinkers can react with leads to results that are most often interpreted at the peptide or even protein level. Therefore, conventional crosslinking yields

a broad, albeit sparse, coverage of the potential interactions that could be formed in the sample. Here, a photoreactive amino acid was incorporated at specified sites at the time of protein expression, enabling the targeting of any site, including those that are inaccessible to conventional crosslinking reagents. BPA reacts with all amino acids, and with our informatics workflow, the resulting crosslinks are identified at the residue level. This combination of targeted analysis towards a region of interest and residue-level interactions enables a deep coverage of interactions of a narrow region of interest. The results in this study were generated from samples in which all intentionally introduced proteins are the same BPA-containing variant. Based on the outcomes of this study, we propose that this strategy can be extended to samples containing a BPA-containing variant that is (a) diluted into similar proteins that do not contain BPA and/or (b) combined with candidate interaction partners. Both cases will benefit from the isolation of dimeric products (as demonstrated here), but the latter may require additional optimization in terms of product isolation and digestion. Therefore, we anticipate that this workflow will be a powerful, general strategy for characterizing the structures of proteins that have resisted high-resolution characterization, including disordered and heterogeneous proteins.

2.6 Acknowledgements

We thank Mike Hoopman and David Shteynberg (Institute for Systems Biology) for useful discussions and technical assistance related to the Trans-Proteomic Pipeline. We thank Lucas Narisawa (University of Washington) for critical evaluation of the manuscript. This material is based upon work supported by the National Eye Institute through R01 EY017370 to REK, the National Institute of General Medical Sciences through T32 GM008268 to CNW, the National Institute of Aging through T32 AG066574 to LDU, and the University of Washington's Proteomics Resource (UWPR95794).

2.7 References

(1) Klykov, O.; Steigenberger, B.; Pektaş, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990. https://doi.org/10.1038/s41596-018-0074-x.

(2) Yu, C.; Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **2018**, *90* (1), 144–165. https://doi.org/10.1021/acs.analchem.7b04431.

(3) Singh, P.; Panchaud, A.; Goodlett, D. Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Anal. Chem.* **2010**, *82* (7), 2636–2642. https://doi.org/10.1021/ac1000724.

(4) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; Chaignepain, S.; Chavez, J. D.; Claverol, S.; Cox, J.; Davis, T.; Degliesposti, G.; Dong, M.-Q.; Edinger, N.; Emanuelsson, C.; Gay, M.; Götze, M.; Gomes-Neto, F.; Gozzo, F. C.; Gutierrez, C.; Haupt, C.; Heck, A. J. R.; Herzog, F.; Huang, L.; Hoopmann, M. R.; Kalisman, N.; Klykov, O.; Kukačka, Z.; Liu, F.; MacCoss, M. J.; Mechtler, K.; Mesika, R.; Moritz, R. L.; Nagaraj, N.; Nesati, V.; Neves-Ferreira, A. G. C.; Ninnis, R.; Novák, P.; O'Reilly, F. J.; Pelzing, M.; Petrotchenko, E.; Piersimoni, L.; Plasencia, M.; Pukala, T.; Rand, K. D.; Rappsilber, J.; Reichmann, D.; Sailer, C.; Sarnowski, C. P.; Scheltema, R. A.; Schmidt, C.; Schriemer, D. C.; Shi, Y.; Skehel, J. M.; Slavin, M.; Sobott, F.; Solis-Mezarino, V.; Stephanowitz, H.; Stengel, F.; Stieger, C. E.; Trabjerg, E.; Trnka, M.; Vilaseca, M.; Viner, R.; Xiang, Y.; Yilmaz, S.; Zelter, A.; Ziemianowicz, D.; Leitner, A.; Sinz, A. First Community-Wide, Comparative

Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961. https://doi.org/10.1021/acs.analchem.9b00658.

(5) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 742. https://doi.org/10.1038/s41467-020-14608-2.

(6) Petrotchenko, E. V.; Borchers, C. H. Crosslinking Combined with Mass Spectrometry for Structural Proteomics. *Mass Spectrom. Rev.* **2010**, *29* (6), 862–876. https://doi.org/10.1002/mas.20293.

(7) Kluger, R.; Alagic, A. Chemical Cross-Linking and Protein–Protein Interactions—a Review with Illustrative Protocols. *Bioorganic Chem.* **2004**, *32* (6), 451–472. https://doi.org/10.1016/j.bioorg.2004.08.002.

(8) Chen, Y.; Topp, E. M. Photolytic Labeling and Its Applications in Protein Drug Discovery and Development. *J. Pharm. Sci.* **2019**, *108* (2), 791–797. https://doi.org/10.1016/j.xphs.2018.10.017.

(9) Mishra, P. K.; Yoo, C.-M.; Hong, E.; Rhee, H. W. Photo-Crosslinking: An Emerging Chemical Tool for Investigating Molecular Networks in Live Cells. *ChemBioChem* **2020**, *21* (7), 924–932. https://doi.org/10.1002/cbic.201900600.

(10) Deseke, E.; Nakatani, Y.; Ourisson, G. Intrinsic Reactivities of Amino Acids towards Photoalkylation with Benzophenone − A Study Preliminary to Photolabelling of the Transmembrane Protein Glycophorin A. *Eur. J. Org. Chem.* **1998**, *1998* (2), 243–251. https://doi.org/10.1002/(SICI)1099-0690(199802)1998:2<243::AID-EJOC243>3.0.CO;2-I.

(11) Tanaka, Y.; Bond, M. R.; Kohler, J. J. Photocrosslinkers Illuminate Interactions in Living Cells. *Mol. Biosyst.* **2008**, *4* (6), 473–480. https://doi.org/10.1039/B803218A.

(12) Das, J. Aliphatic Diazirines as Photoaffinity Probes for Proteins: Recent Developments. *Chem. Rev.* **2011**, *111* (8), 4405–4417. https://doi.org/10.1021/cr1002722.

(13) Suchanek, M.; Radzikowska, A.; Thiele, C. Photo-Leucine and Photo-Methionine Allow Identification of Protein-Protein Interactions in Living Cells. *Nat. Methods* **2005**, *2* (4), 261–268. https://doi.org/10.1038/nmeth752.

(14) Kauer, J. C.; Erickson-Viitanen, S.; Wolfe, H. R.; DeGrado, W. F. P-Benzoyl-L-Phenylalanine, a New Photoreactive Amino Acid. Photolabeling of Calmodulin with a Synthetic Calmodulin-Binding Peptide. *J. Biol. Chem.* **1986**, *261* (23), 10695–10700. https://doi.org/10.1016/S0021-9258(18)67441-1.

(15) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(16) Liu, C. C.; Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **2010**, *79* (1), 413–444. https://doi.org/10.1146/annurev.biochem.052308.105824.

(17) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proc. Natl. Acad. Sci.* **2002**, *99* (17), 11020–11024. https://doi.org/10.1073/pnas.172226299.

(18) Sato, S.; Mimasu, S.; Sato, A.; Hino, N.; Sakamoto, K.; Umehara, T.; Yokoyama, S. Crystallographic Study of a Site-Specifically Cross-Linked Protein Complex with a Genetically Incorporated Photoreactive Amino Acid,. *Biochemistry* **2011**, *50* (2), 250–257. https://doi.org/10.1021/bi1016183.

(19) Pettelkau, J.; Ihling, C. H.; Frohberg, P.; van Werven, L.; Jahn, O.; Sinz, A. Reliable Identification of Cross-Linked Products in Protein Interaction Studies by 13C-Labeled p-Benzoylphenylalanine. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (9), 1628–1641. https://doi.org/10.1007/s13361-014-0944-6.

(20) Hauser, M.; Qian, C.; King, S. T.; Kauffman, S.; Naider, F.; Hettich, R. L.; Becker, J. M. Identification of Peptide-Binding Sites within BSA Using Rapid, Laser-Induced Covalent Cross-Linking Combined with High-Performance Mass Spectrometry. *J. Mol. Recognit.* **2018**, *31* (2), e2680. https://doi.org/10.1002/jmr.2680.

(21) Schwarz, R.; Tänzler, D.; Ihling, C. H.; Müller, M. Q.; Kölbel, K.; Sinz, A. Monitoring Conformational Changes in Peroxisome Proliferator-Activated Receptor α by a Genetically Encoded Photoamino Acid, Cross-Linking, and Mass Spectrometry. *J. Med. Chem.* **2013**, *56* (11), 4252–4263. https://doi.org/10.1021/jm400446b.

(22) Nguyen, T. T.; Sabat, G.; Sussman, M. R. In Vivo Cross-Linking Supports a Head-to-Tail Mechanism for Regulation of the Plant Plasma Membrane P-Type H+-ATPase. *J. Biol. Chem.* **2018**, *293* (44), 17095–17106. https://doi.org/10.1074/jbc.RA118.003528.

(23) Piotrowski, C.; Moretti, R.; Ihling, C. H.; Haedicke, A.; Liepold, T.; Lipstein, N.; Meiler, J.; Jahn, O.; Sinz, A. Delineating the Molecular Basis of the Calmodulin–bMunc13-2 Interaction by Cross-Linking/Mass Spectrometry—Evidence for a Novel CaM Binding Motif in bMunc13-2. *Cells* **2020**, *9* (1), 136. https://doi.org/10.3390/cells9010136.

(24) Kolhe, J. A.; Babu, N. L.; Freeman, B. C. The Hsp90 Molecular Chaperone Governs Client Proteins by Targeting Intrinsically Disordered Regions. *Mol. Cell* **2023**, *83* (12), 2035-2044.e7. https://doi.org/10.1016/j.molcel.2023.05.021.

(25) Kolhe, J. A.; Babu, N. L.; Freeman, B. C. Protocol for Establishing a Protein Interactome Based on Close Physical Proximity to a Target Protein within Live Budding Yeast. *STAR Protoc.* **2023**, *4* (4), 102663. https://doi.org/10.1016/j.xpro.2023.102663.

(26) Jaya, N.; Garcia, V.; Vierling, E. Substrate Binding Site Flexibility of the Small Heat Shock Protein Molecular Chaperones. *Proc. Natl. Acad. Sci.* **2009**, *106* (37), 15604–15609. https://doi.org/10.1073/pnas.0902177106.

(27) Zhang, M.; Lin, S.; Song, X.; Liu, J.; Fu, Y.; Ge, X.; Fu, X.; Chang, Z.; Chen, P. R. A Genetically Incorporated Crosslinker Reveals Chaperone Cooperation in Acid Resistance. *Nat. Chem. Biol.* **2011**, *7* (10), 671–677. https://doi.org/10.1038/nchembio.644.

(28) Wang, R. Y.-R.; Noddings, C. M.; Kirschke, E.; Myasnikov, A. G.; Johnson, J. L.; Agard, D. A. Structure of Hsp90–Hsp70–Hop–GR Reveals the Hsp90 Client-Loading Mechanism. *Nature* **2022**, *601* (7893), 460–464. https://doi.org/10.1038/s41586-021-04252-1.

(29) Kleiner, R. E.; Hang, L. E.; Molloy, K. R.; Chait, B. T.; Kapoor, T. M. A Chemical Proteomics Approach to Reveal Direct Protein-Protein Interactions in Living Cells. *Cell Chem. Biol.* **2018**, *25* (1), 110-120.e3. https://doi.org/10.1016/j.chembiol.2017.10.001.

(30) McKenna, M. J.; Sim, S. I.; Ordureau, A.; Wei, L.; Harper, J. W.; Shao, S.; Park, E. The Endoplasmic Reticulum P5A-ATPase Is a Transmembrane Helix Dislocase. *Science* **2020**, *369* (6511), eabc5809. https://doi.org/10.1126/science.abc5809.

(31) Chu, N.; Salguero, A. L.; Liu, A. Z.; Chen, Z.; Dempsey, D. R.; Ficarro, S. B.; Alexander, W. M.; Marto, J. A.; Li, Y.; Amzel, L. M.; Gabelli, S. B.; Cole, P. A. Akt Kinase Activation Mechanisms Revealed Using Protein Semisynthesis. *Cell* **2018**, *174* (4), 897-907.e14. https://doi.org/10.1016/j.cell.2018.07.003.

(32) Ji, Z.; Li, H.; Peterle, D.; Paulo, J. A.; Ficarro, S. B.; Wales, T. E.; Marto, J. A.; Gygi, S. P.; Engen, J. R.; Rapoport, T. A. Translocation of Polyubiquitinated Protein Substrates by the Hexameric Cdc48 ATPase. *Mol. Cell* **2022**, *82* (3), 570-584.e8. https://doi.org/10.1016/j.molcel.2021.11.033.

(33) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(34) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(35) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *J. Proteome Res.* **2015**, *14* (5), 2190–2198. https://doi.org/10.1021/pr501321h.

(36) Hoopmann, M. R.; Shteynberg, D. D.; Zelter, A.; Riffle, M.; Lyon, A. S.; Agard, D. A.; Luan, Q.; Nolen, B. J.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Improved Analysis of Cross-Linking Mass Spectrometry Data with Kojak 2.0, Advanced by Integration into the

Trans-Proteomic Pipeline. *J. Proteome Res.* **2023**, *22* (2), 647–655. https://doi.org/10.1021/acs.jproteome.2c00670.

(37) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392. https://doi.org/10.1021/ac025747h.

(38) Deutsch, E. W.; Mendoza, L.; Shteynberg, D. D.; Hoopmann, M. R.; Sun, Z.; Eng, J. K.; Moritz, R. L. Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **2023**, *22* (2), 615–624. https://doi.org/10.1021/acs.jproteome.2c00624.

(39) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (1), 76–87. https://doi.org/10.1007/s13361-011-0261-2.

(40) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated Assignment of MS/MS Cleavable Cross-Links in Protein 3D-Structure Analysis. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 83–97. https://doi.org/10.1007/s13361-014-1001-1.

(41) Iacobucci, C.; Götze, M.; Ihling, C. H.; Piotrowski, C.; Arlt, C.; Schäfer, M.; Hage, C.; Schmidt, R.; Sinz, A. A Cross-Linking/Mass Spectrometry Workflow Based on MS-Cleavable Cross-Linkers and the MeroX Software for Studying Protein Structures and Protein–Protein Interactions. *Nat. Protoc.* **2018**, *13* (12), 2864–2889. https://doi.org/10.1038/s41596-018-0068-8.

(42) Sun, Y.; MacRae, T. H. The Small Heat Shock Proteins and Their Role in Human Disease. *FEBS J.* **2005**, *272* (11), 2613–2627. https://doi.org/10.1111/j.1742-4658.2005.04708.x.

(43) Magalhaes, S.; Goodfellow, B. J.; Nunes, A. Aging and Proteins: What Does Proteostasis Have to Do with Age? *Curr. Mol. Med.* **2018**, *18* (3), 178–189. https://doi.org/10.2174/1566524018666180907162955.

(44) Delbecq, S. P.; Klevit, R. E. One Size Does Not Fit All: The Oligomeric States of αB Crystallin. *FEBS Lett.* **2013**, *587* (8), 1073–1080. https://doi.org/10.1016/j.febslet.2013.01.021.

(45) Clouser, A. F.; Baughman, H. E.; Basanta, B.; Guttman, M.; Nath, A.; Klevit, R. E. Interplay of Disordered and Ordered Regions of a Human Small Heat Shock Protein Yields an Ensemble of 'Quasi-Ordered' States. *eLife* **2019**, *8*, e50259. https://doi.org/10.7554/eLife.50259.

(46) Haslbeck, M.; Weinkauf, S.; Buchner, J. Small Heat Shock Proteins: Simplicity Meets Complexity. *J. Biol. Chem.* **2019**, *294* (6), 2121–2132. https://doi.org/10.1074/jbc.REV118.002809.

(47) Collier, M. P.; Benesch, J. L. P. Small Heat-Shock Proteins and Their Role in Mechanical Stress. *Cell Stress Chaperones* **2020**, *25* (4), 601–613. https://doi.org/10.1007/s12192-020-01095-z.

(48) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(49) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

(50) *In-Gel Tryptic Digestion Kit*. https://www.thermofisher.com/order/catalog/product/89871 (accessed 2022-03-01).

(51) *cRAP protein sequences*. https://www.thegpm.org/crap/ (accessed 2022-04-19).

(52) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543–D552. https://doi.org/10.1093/nar/gkab1038.

(53) Braun, N.; Zacharias, M.; Peschek, J.; Kastenmüller, A.; Zou, J.; Hanzlik, M.; Haslbeck, M.; Rappsilber, J.; Buchner, J.; Weinkauf, S. Multiple Molecular Architectures of the Eye Lens Chaperone αB-Crystallin Elucidated by a Triple Hybrid Approach. *Proc. Natl. Acad. Sci.* **2011**, *108* (51), 20491–20496. https://doi.org/10.1073/pnas.1111014108.

(54) Jehle, S.; Vollmar, B. S.; Bardiaux, B.; Dove, K. K.; Rajagopal, P.; Gonen, T.; Oschkinat, H.; Klevit, R. E. N-Terminal Domain of B-Crystallin Provides a Conformational Switch for Multimerization and Structural Heterogeneity. *Proc. Natl. Acad. Sci.* **2011**, *108* (16), 6409–6414. https://doi.org/10.1073/pnas.1014656108.

(55) Ma, K.; Vitek, O.; Nesvizhskii, A. I. A Statistical Model-Building Perspective to Identification of MS/MS Spectra with PeptideProphet. *BMC Bioinformatics* **2012**, *13* (S16), S1. https://doi.org/10.1186/1471-2105-13-S16-S1.

(56) Riffle, M.; Jaschob, D.; Zelter, A.; Davis, T. N. ProXL (Protein Cross-Linking Database): A Platform for Analysis, Visualization, and Sharing of Protein Cross-Linking Mass Spectrometry Data. *J. Proteome Res.* **2016**, *15* (8), 2863–2870. https://doi.org/10.1021/acs.jproteome.6b00274.

(57) Klevit, R. E. Peeking from behind the Veil of Enigma: Emerging Insights on Small Heat Shock Protein Structure and Function. *Cell Stress Chaperones* **2020**, *25* (4), 573–580. https://doi.org/10.1007/s12192-020-01092-2.

(58) Mendes, M. L.; Fischer, L.; Chen, Z. A.; Barbon, M.; O'Reilly, F. J.; Giese, S. H.; Bohlke-Schneider, M.; Belsom, A.; Dau, T.; Combe, C. W.; Graham, M.; Eisele, M. R.; Baumeister, W.; Speck, C.; Rappsilber, J. An Integrated Workflow for Crosslinking Mass Spectrometry. *Mol. Syst. Biol.* **2019**, *15* (9). https://doi.org/10.15252/msb.20198994.

(59) Ser, Z.; Cifani, P.; Kentsis, A. Optimized Cross-Linking Mass Spectrometry for in Situ Interaction Proteomics. *J. Proteome Res.* **2019**, *18* (6), 2545–2558. https://doi.org/10.1021/acs.jproteome.9b00085.

(60) Dau, T.; Gupta, K.; Berger, I.; Rappsilber, J. Sequential Digestion with Trypsin and Elastase in Cross-Linking Mass Spectrometry. *Anal. Chem.* **2019**, *91* (7), 4472–4478. https://doi.org/10.1021/acs.analchem.8b05222.

2.8 Supporting Information

2.8.1 LC-MS Methods

Portions of the raw LC-MS data analyzed here have been reported previously.[1,2] This work includes updated analyses (newer software versions were used) of previously reported raw data. Relevant figure captions and Table S1 indicate which data sets were reported previously. Data were acquired using an EASY NanoLC coupled to Thermo Orbitrap Fusion Lumos Tribrid (Lumos). Samples were resuspended in 95% water 5% ACN with 0.1% FA The weight (μg) of protein loaded onto the gel and the relative color intensity of the band excised from the gel prior to digestion was used to estimate the weight of the digested peptides loaded on the column. For example, when 7.6 μg of protein was loaded onto a gel lane and the dimeric product band was roughly 30% of the total color intensity, we estimated that the excised dimeric products contained about 2.3 μg of protein. In that case, about 20% (500 ug) of the digested peptides were loaded on the column. Data was collected with an Easy Nano LC coupled to a Thermo Orbitrap Fusion Lumos Tribrid. Roughly 500 ng of peptide was loaded onto an 8-cm trap column. The sample was then separated on a 25-cm analytical column with a 75 μm inner diameter using an 85-minute or 30-minute gradient from 6% B to 45% B, where A was water and B was 80% acetonitrile, at 300 $nL \cdot min^{-1}$. The column was then flushed and regenerated for 35 minutes. Most samples presented here were separated using an 85-minute gradient, but those reported in Table S2, Figure 3, and Figure S3 used a 30-minute gradient. Comparable crosslinks are identified with either gradient (see Effect of Gradient section in the Supporting Information). Spectra were acquired across the entire LC method using data-dependent acquisition with dynamic exclusion after one time for a duration of 30 seconds and an intensity threshold of $2.0x10^4$. Orbitrap detection and higher-energy collisional dissociation (HCD) fragmentation (30% normalized

collision energy) were used with a target value of $1.00x10^5$, maximum injection time of 22 ms, top N of 20, and isolation width of 1.6. MS1 were acquired at a resolving power of 120,000 over the range of 400-2000 *m*/*z*, and MS2 were acquired with a resolution of 15,000. Mass spectra were acquired using profile mode and the Thermo RAW format. Those files were centroided and saved as mzml files using the ProteoWizard[3] msconvert tool as implemented in the Trans-Proteomic Pipeline version 6.3.2.[4]

2.8.2 Comet Search Settings

The searches were enzyme nonspecific using a peptide mass tolerance of 20.0 ppm. The isotope error offset was 3, and BPA was defined as an additional amino acid, B, that has a mass of 251.09462859 Da. Methionine oxidation and cysteine iodoacetamide alkylation were variable modifications.

2.8.3 Kojak Search Settings

The search settings matched those described for the Comet searches except that the precursor tolerance was 15 ppm and enzyme selection rules were used. For the trypsin digested samples, the preexisting trypsin setting was used. For the trypsin-GluC digested samples, the cleavage sites of D and E were added to the trypsin settings. Crosslinks were defined as from BPA to any residue, with no change in mass.[5]

2.8.4 PeptideProphet Settings

PeptideProphet[6] was used to validate both Comet and Kojak results with the following options: ppm for accurate mass binning, only use Expect Score as the discriminant, decoys to pin down the negative distribution, use non-parametric model, and report decoy hits with computed probability. For trypsin-GluC-digested samples, the enzyme also was defined in the additional options line as -e"TrypGluC:specific:true:KRDE|P".

For Comet searches, after filtering using a 1% False Discovery Rate (FDR) and a minimum of 2 Peptide Spectral Matches (PSMs), this yields a protein database for the sample. A 1% FDR is defined as the error rate of 0.0100 that PeptideProphet reports in the error table. This is a PSM level FDR. PSMs that meet the 1% FDR threshold are found by filtering to spectra that have a PeptideProphet probability greater than or equal to that for an error rate of 0.0100. For Comet, overall (considering all charge states) error rates, FDRs, and PeptideProphet probabilities were used. For Kojak searches, the same PeptideProphet settings were used, but for ions of each charge state, the corresponding error rate, FDR, and PeptideProphet probability were used.

2.8.5 Identification of Crosslinks Using StavroX or MeroX

StavroX version 3.6.6.6 was used because it is the most recent version with a built option for BPA.[7] StavroX settings included methionine oxidation and cysteine iodoacetamide alkylation as variable modifications, inverse decoys, and the default settings for trypsin digestion and BPA as a crosslinker. MeroX version 2.0.1.4 was used with BPA as the crosslinker and other default settings for BPA that matched those used for StavroX.[8,9] A trypsin-digested and UV-treated sample of the W9B-BPA variant of HSPB5 was used to compare search databases and informatic workflows. Results for the 1% FDR at the PSM level are reported here.

Databases used for the comparison searches included: only the HSPB5 variant (1 protein), the corresponding validated-protein database (12 proteins), the full *E.coli* BL21 and cRAP database with (5315 proteins), and a series of databases constructed based on Protein Probability values (50, 100, 250, 500, and 1000 proteins). Protein probability values are from the ProteinProphet tool in TPP, and these results from the monomeric reactant sample were used. When constructing search databases at varying sizes, the proteins with the highest protein probability values were included. If the number of proteins desired exceeded the number of proteins with ProteinProphet probability values, additional proteins were randomly selected from the full *E.coli* BL21 and cRAP database to meet the desired number of proteins.

2.8.6 Effects of Software Versions

Data analysis presented in this manuscript used TPP version 6.3.2 and Kojak version 2.0.3. Data that has been published in other manuscripts used different software versions (TPP 6.0.0 and Kojak 2.0.0 alpha 16).[1,2] Results from different software versions differ slightly in the number of crosslink PSMs identified at a 1% FDR but give similar crosslink identifications.

Some potential errors can occur if using older software versions. PeptideProphet constructs models for unlinked peptides, crosslinked peptides, and loop-link peptides separately when validating Kojak crosslink search results. Crosslinked peptides are inter-peptide crosslinks, and loop-link peptides are intra-peptide crosslinks. Our method's compatibility with PeptideProphet's target-decoy based validation was non-trivial due to BPA crosslinks' unique properties. Because BPA is an amino acid and reacts with no mass change, it is very difficult to differentiate between a spectrum that corresponds to unreacted BPA (unlinked peptide) or an intra-peptide crosslink (loop link). For TPP v5.2.0, PeptideProphet did not export models for crosslinked and unlinked peptides when there was not enough data to construct loop link models. For TPP v6.0.0 and later versions, changes were made to PeptideProphet so that models are produced for crosslinked and unlinked peptides, even when there is inadequate data to generate models of loop-linked peptides.

Since PeptideProphet v6.0.0, PeptideProphet has reported overall probabilities for crosslinks as well as reporting probabilities for each charge state. Using an overall probability value gives slightly different results than using separate probability values for each charge state because of the discrepancies between the overall and charge state probability values, so we have continued to split crosslink results by charge state because the validation models are created separately by charge state.

2.8.7 Effects of Gradient

Figure S1A and S1B illustrates results for trypsin-digested W9B from both a 30- and 85-minute gradient. 250 crosslink PSMs and 2042 total PSMs (crosslinked and unlinked PSMs) are identified with the 85-minute gradient, which results in 12% of identified PSMs being crosslinks. 155 crosslink PSMs and 1148 total PSMs are identified with the 30-minute gradient, which results in 14% of the identified PSMs being crosslinks. The number of PSMs differs significantly between the two gradients because data is acquired for twice as long when using the longer gradient. However, a similar proportion of crosslink PSMs is identified between both gradients.

As shown in Figure S1C, 60 crosslinking sites are identified with both gradients. Only 14 crosslinking sites are identified with only the 85-minute gradient, and only 12 crosslinking sites are identified with only the 30-minuite gradient. Both gradients also identify the same high frequency crosslinking sites at residues 1, 38, 64, 85, and 137. These high-frequency crosslink sites are identified with residue-level specificity in both datasets, even with the ambiguity evidenced by the lower frequency crosslink sites at surrounding residues. Both datasets also have a high frequency cluster of crosslinks around residue 16. With the 85-minute gradient, site 16 is the maxima of that cluster, and with the 30-minute gradient, site 17 is the maxima of that cluster. A similar range of crosslink sites is detected around residue 17, and the maxima differ by only a single residue between the datasets. The shorter gradient identifies more crosslink PSMs around site 109 (4 PSMs distributed across 2 potential crosslinking sites) than the longer gradient does (1 PSM distributed across 5 potential crosslinking sites). This slight difference is one of the main differences between the two. It’s likely that using a shorter gradient results in less peak broadening and more intense chromatographic peaks, which may increase sensitivity towards certain crosslink sites. Overall, results from the different gradients are highly similar and identify

the same predominate crosslinking sites with residue-level specificity, which suggests that a shorter gradient can be used to reduce the instrument time and cost needed to analyze each sample.

Table S1. Summary of results for different sites of BPA incorporation**.** The LC-MS experiments used 85-minute gradients and the raw data was sourced from this work or other reports.[1,2] The validated-protein databases were determined using monomeric reactants, as described schematically in the left column of Figure 1. Non-Target crosslinks include a peptide that is not from HSPB5, which is either a decoy or another contaminant protein.

BPA site	**9**	**9**	**17**	**17**	**24**	**24**	**33**	**47**	**47**	**61**	**61**	**9**	**9**
Raw Data Source	2	1	2	1	2	1	1	This Work	1	2	1	This Work	This Work
Enzyme[a]	TG	TG	TG	TG	TG	TG	TG	TG	TG	TG	TG	T	T
pH	6.5	7.5	6.5	7.5	6.5	7.5	7.5	6.5	7.5	6.5	7.5	6.5	6.5
# Proteins in Validated-Protein Database	11	11	12	12	12	12	11	13	13	12	12	12	12
# HSPB5-HSPB5 Crosslink PSMs	195	203	169	45	136	66	62	34	61	179	281	277	294
# Non-Target Crosslink PSMs	1	2	2	2	1	2	1	0	1	3	3	7	5

[a] Enzymes used for digestion, where T is only trypsin and TG is a trypsin and GluC parallel digestion.

Table S2. Summary of results from 9 replicate analyses of dimeric products from W9B-HSPB5 that were acquired for this study. All samples originate from the same expression and purification of the variant, but two different photo-crosslinking reactions were performed in parallel in separate wells of the same 96-well plate. An aliquot from the first reaction was loaded onto a single lane of a gel, the dimeric products were isolated, excised, and digested, and the resulting sample was used for Injections 1−3. This process was repeated for two aliquots from the second reaction, which resulted in two samples that were used for Injections 4−6 and Injections 7−9, respectively. These samples were analyzed independently using a 30-minute gradient. Monomeric reactant samples were not generated for these samples; instead, validated-protein databases were determined using the LC-MS data from the dimeric products.

Photo-Crosslinking Reaction	1						2		
Sample	1			2			3		
Injection	1	2	3	4	5	6	7	8	9
Number of Proteins in Validated Protein- Database	17	17	18	18	21	18	18	18	19
Number of HSPB5-HSPB5 Crosslink PSMs	342	325	293	330	301	317	375	352	310
Number of Non-Target Crosslink PSMs	7	4	4	6	3	4	9	7	6

Table S3. These validated-protein databases were obtained using either a monomeric-reactant or dimeric-product sample of trypsin-digested W9B, as described in the caption for Figure S1. Proteins in the database had at least 2 PSMs at a 1% FDR at the peptide level.

Non-UV-Treated Monomer	**UV-Treated Dimer**
ALBU_HUMAN, AngioNeuro, A0A3Y3V6H6_ECOLX, A0A024L2A1_ECOLX, A0A0Q3IA67_ECOLX, HSPB5_B9, K1C15_SHEEP, K1C10_HUMAN, K1C9_HUMAN, K22E_HUMAN, K2C1_HUMAN, TRYP_PIG	ALBU_BOVIN, ALBU_HUMAN, A0A3Y3V2U8_ECOLX, C3SM83_ECOLX, HSPB5_B9, K1C15_SHEEP, K1C10_HUMAN, K1C9_HUMAN, K22E_HUMAN, K2C1_HUMAN, TRYP_PIG

Table S4. The crosslinks identified from the validated-protein database search using StavroX are described in detail here. Score is StavroX's reported score. The protein columns are omitted because the protein is W9B-HSPB5 for all represented sites. Site #1 is the first end of the crosslink. Site #2 is the second end of the crosslink. Multiple sites being listed indicates ambiguity in the crosslink site assignment.

Score	Site #1	Site #2	Number of PSMs
147	8/9/10/11	8/9/10/11	3
120	0/1/2/3/4/5/6/7/8/9/10/11	0/2/3/4/5/6/7/8/9/10/11	2
117	2/3/4/5/6/7/8/9/10/11	2/3/4/5/6/7/8/9/10/11	1
105	0/1/2/3/4/5/6/7/8/9/10/11	8/9/10/11	2
104	7/8/9/10/11	7/8/9/10/11	1
94	0/1/2/3/4/5/6/7/8/9/10/11	0/1/2/3/4/5/6/7/8/9/10/11	1

Table S5. The crosslinks identified from the validated-protein database search using MeroX are described in detail here. Score is MeroX's reported score. The protein columns are omitted because the protein is W9B-HSPB5 for all represented sites. Site #1 is the first end of the crosslink. Site #2 is the second end of the crosslink. Multiple sites being listed indicates ambiguity in the crosslink site assignment.

Score	Site #1	Site #2	Number of PSMs
120	8/9/10/11	8/9/10/11	14
104	0/1/2/3/4/5/6/7/8/9/10/11	0/1/2/3/4/5/6/7/8/9/10/11	7
93	0/2/3/4/5/6/7	9	2
77	0/2/3/4/5/6/7/8/9/10/11	0/1/2/3/4/5/6/7/8/9/10/11	2
58	9	135/136/137/138/139/140/141/142/143/144/145/147/148/149	1
57	9	57/58/59/60/61	1
50	9	57/58/59/60/61/62/63/64/65/66/67/68/69	1
45	2/3/4/5/6/7/8/9/10/11	2/3/4/5/6/7/8/9/10/11	1
39	7/8/9/10/11	7/8/9/10/11	1
30	9	58/59/60/61	1
29	9	63/64/65/66/67/68/69	1

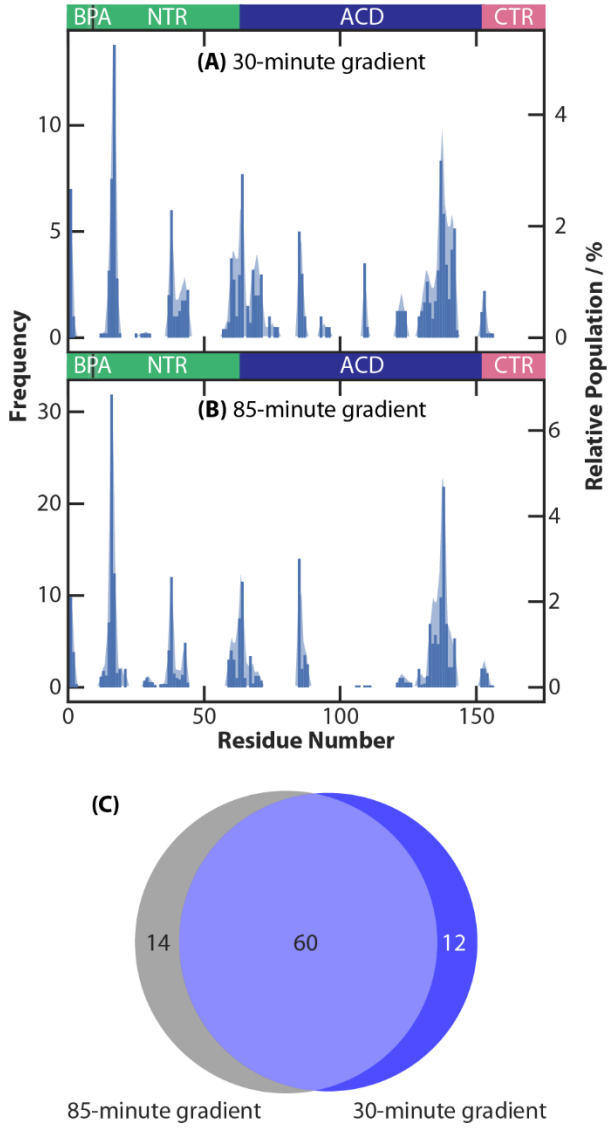

Figure S1. Panels A and B depict trypsin-digested W9B at pH 6.5 analyzed with a 30- or an 85-minute gradient. Both analyses are separate injections from the sample solution on the same date. Panel A has 155 crosslink PSMs from a 30-minute gradient. Panel B has 250 crosslink PSMs from an 85-minute gradient. Panel C illustrates overlap in the crosslink sites identified (sites with a PSM value greater than zero) in panels B and C. Despite the difference in the number of crosslink PSMs, 60 of the 86 total crosslink sites are identified with both gradients.

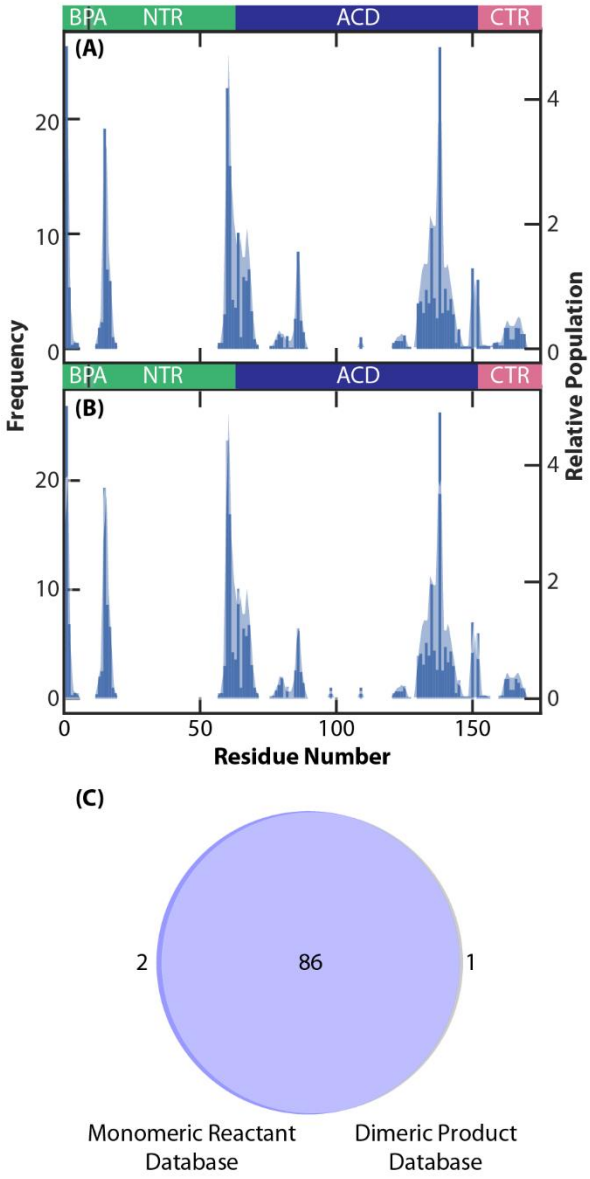

Figure S2. Crosslinks identified using a monomeric-reactant or a dimeric-product database**.** The dataset shown here in Panel A has 294 crosslink PSMs from trypsin-digested W9B at pH 6 .5. This data is not previously published and used an 85-minute gradient and monomeric reactants to create the validated-protein database. Panel B shows the results when using the dimeric products to create the validated-protein database and has 298 crosslink PSMs. Panel C compares the number of crosslink sites identified when using either database. Table S3 and Figures 2, 4, S4, and S5 are based on this same dataset.

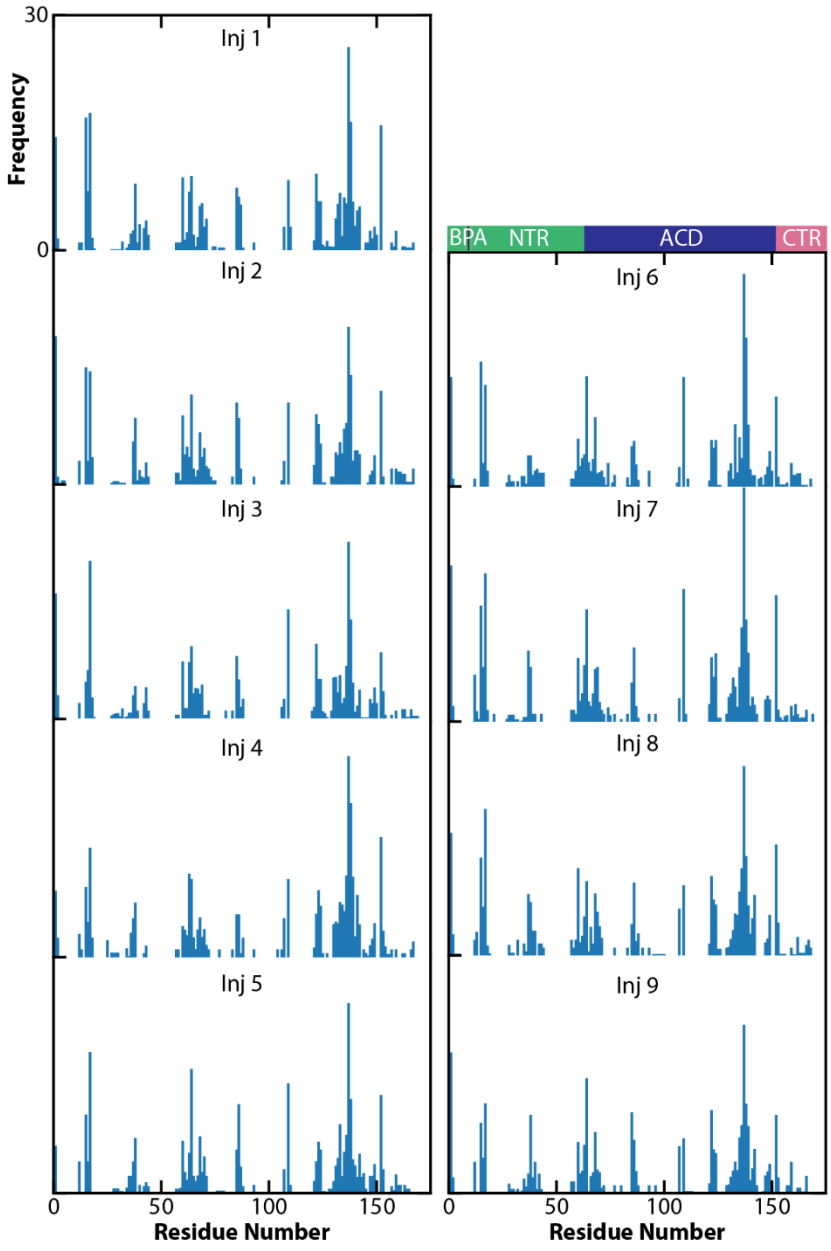

Figure S3. Crosslinks identified for replicates. The results of 9 replicates of trypsin-digested W9B at pH 6.5 are illustrated here. The x-axes depict the residue number of the crosslink site ranging from 0-175 in each plot. The y-axes depict the frequency of the crosslink site ranging from 0-30 in each plot. Table S3 indicates the number of crosslink PSMs identified in each replicate. All samples originate from the same expression and purification of the variant, but two

different photo-crosslinking reactions were performed as described in Table S2. All data represented in this figure is not previously published. For injection 7, the number of crosslinks to site 137 is 31.9, outside the y-axis range shown.

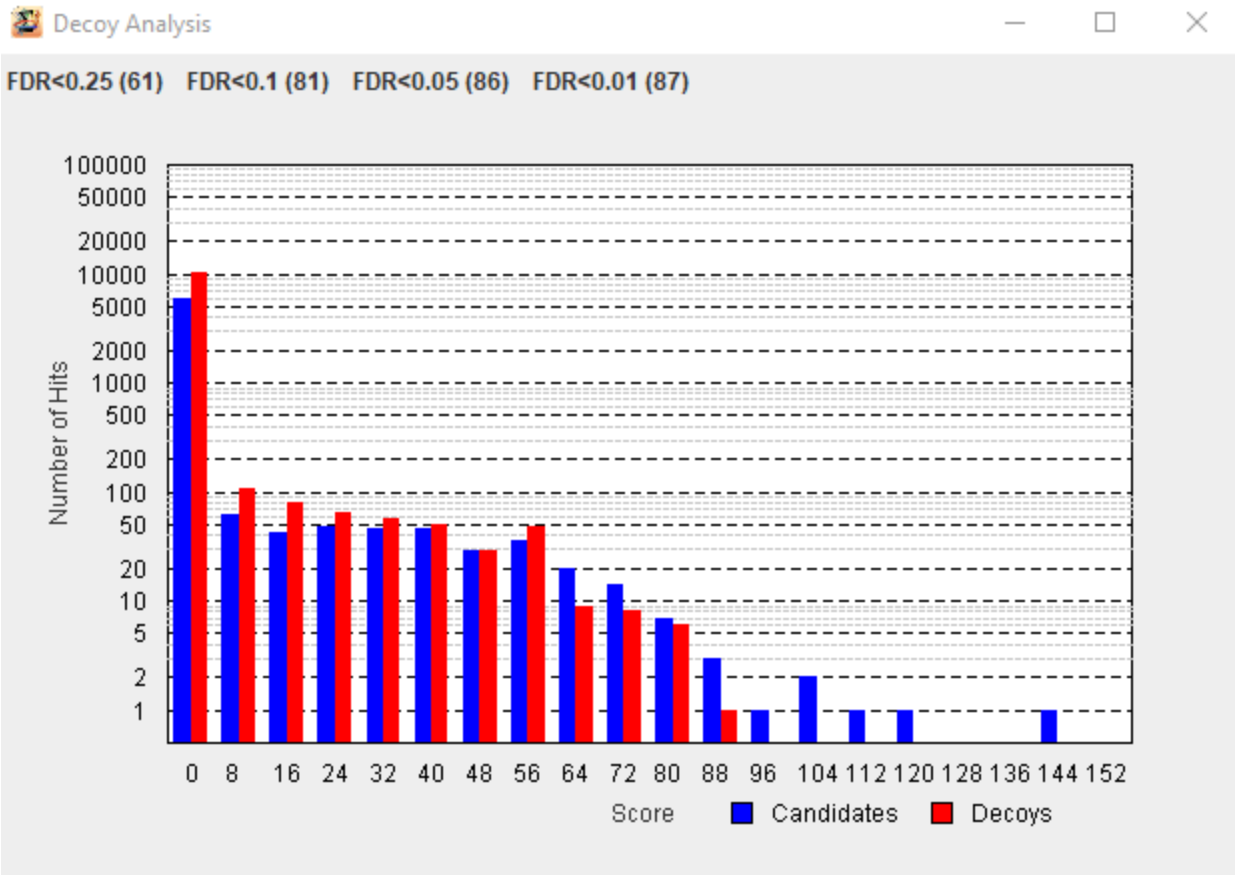

Figure S4. StavroX FDR calculation**.** This figure illustrates StavroX's scores that are used to calculate FDR cutoffs. This is from a search with the 12-protein control database on the same W9B-trypsin digested sample represented in Figure S2.

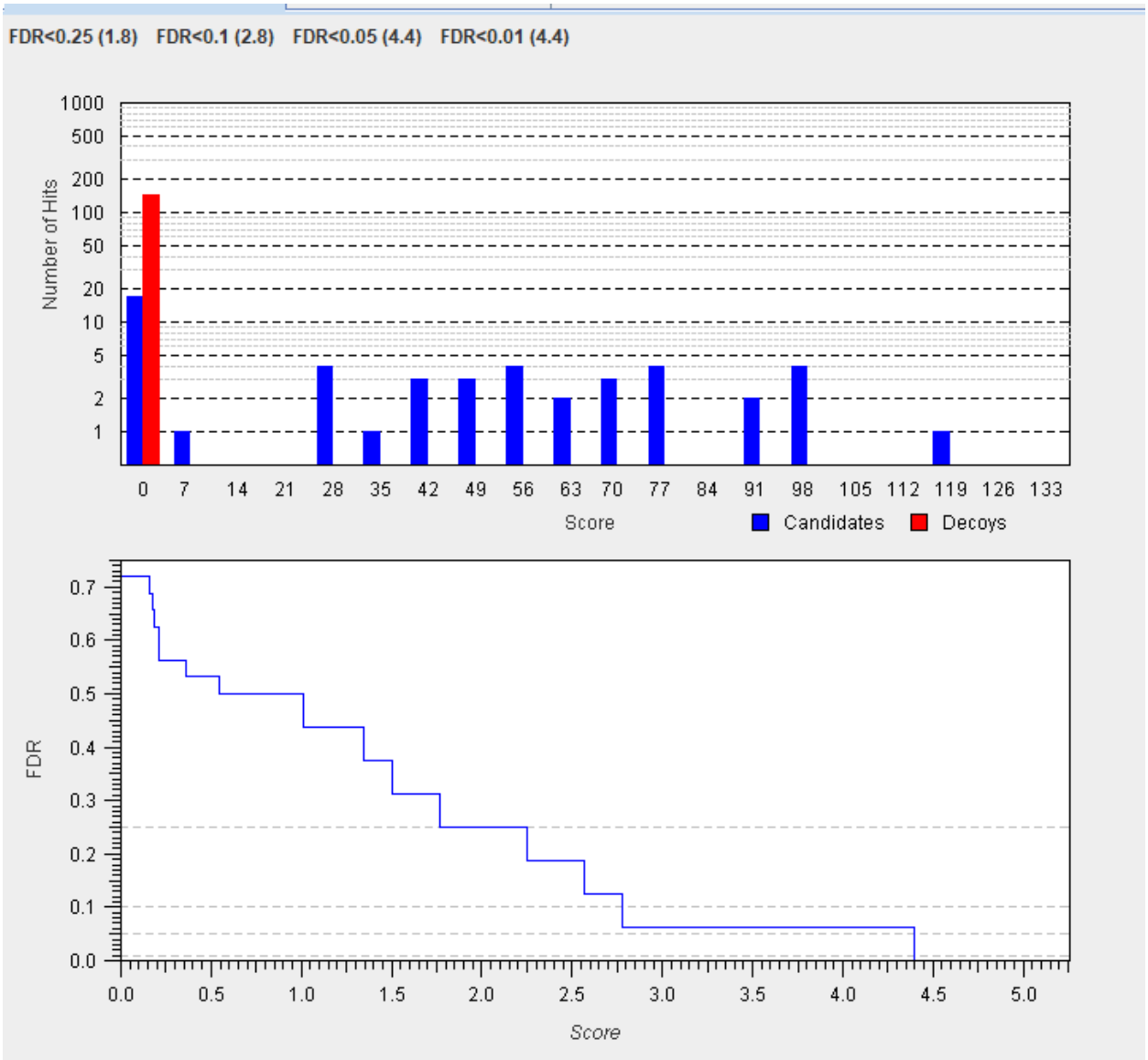

Figure S5. MeroX FDR calculation**.** This figure illustrates MeroX's scores that are used to calculate FDR cutoffs. This is from a search with the 12-protein control database on the same W9B-trypsin digested sample represented in Figure S2.

2.8.8 Supporting Information References

(1) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions https://doi.org/10.1101/2022.05.30.493970.

(2) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(3) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(4) Deutsch, E. W.; Mendoza, L.; Shteynberg, D. D.; Hoopmann, M. R.; Sun, Z.; Eng, J. K.; Moritz, R. L. Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **2023**, *22* (2), 615–624. https://doi.org/10.1021/acs.jproteome.2c00624.

(5) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(6) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392. https://doi.org/10.1021/ac025747h.

(7) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (1), 76–87. https://doi.org/10.1007/s13361-011-0261-2.

(8) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated Assignment of MS/MS Cleavable Cross-Links in Protein 3D-Structure Analysis. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 83–97. https://doi.org/10.1007/s13361-014-1001-1.

(9) Iacobucci, C.; Götze, M.; Ihling, C. H.; Piotrowski, C.; Arlt, C.; Schäfer, M.; Hage, C.; Schmidt, R.; Sinz, A. A Cross-Linking/Mass Spectrometry Workflow Based on MS-Cleavable Cross-Linkers and the MeroX Software for Studying Protein Structures and Protein–Protein Interactions. *Nat. Protoc.* **2018**, *13* (12), 2864–2889. https://doi.org/10.1038/s41596-018-0068-8.

Chapter 3: Bootstrapping for Quantitative Comparisons of Datasets

3.1 Introduction

Crosslinking mass spectrometry (XL-MS) is a powerful method for identifying protein-protein interactions due to its ability to capture transient interactions.[1] In most applications of XL-MS, a chemical reagent reacts with samples prior to enzymatic digestion and LC-MS analysis; the reagent binding sites give distance constraints that can be used to refine structural information and identify interactions' partners.[2] However, XL-MS itself is limited by the challenges in identifying crosslinks and in refining structures.[3,4]

XL-MS results can be used to refine structures through the use of software packages such as the Integrative Modeling Platform that use many different types of data including crosslinks to create a single model.[5] A model is created by translating data into spatial restraints, sampling models that score well, and analyzing and assessing the ensemble.[5] However, there are many issues that arise when converting crosslinks into structures and many software packages to help alleviate those issues. Inter- and intra-subunits crosslinks must be treated differently to gain structural information for larger complexes.[6] Because crosslinks form along the surface of proteins, how likely residues are to be at surface can factor into the calculated distance.[7] In addition, because proteins are dynamic, multiple conformations could be represented by crosslinks from a single experiment. Accounting for sidechain flexibility allows for multiple structures to be considered.[8] Despite the difficulties in transforming XL-MS data into structures, XL-MS data and integrative modeling have been pivotal for determining structures for proteins such as photoreceptor phosphodiesterase (PDE6).[9]

XL-MS data is often used to compare structures or protein interactions across states such as apo/holo,[10] drug-treated/not,[11,12] or drug-sensitive/drug-resistant.[13] In comparing across states,

analyses focusses on identifying differences between datasets, which could lead to identifying different structures. However, because of the difficulties in obtaining structures from crosslinks, it's possible to focus on identifying differences between datasets without or before trying to solve structures. Quantitative XL-MS focuses on identifying changes in protein structure or interactions through applying quantitative proteomics techniques to XL-MS.[14] Techniques for quantitative XL-MS vary significantly and include label free quantification,[15] isotopically labeled crosslinkers,[16,17] TMT labeling,[18] and SILAC labeling.[11,13]

Advancements have made these quantitative XL-MS strategies more accessible. However, because crosslinks are lower in intensity, they are more difficult to quantify and fewer features in the data can be used to determine if there is a change in abundance.[19] Because of these challenges, many questions remain in the field about what degree of crosslinking change indicates a structural change, when the absence of a signal counts as a change, and how ambiguous changes (differences that could correspond to sample preparation) can be interpreted.[19] Here, we develop a statistical method to address these questions and determine what constitutes a significant difference between datasets.

Bootstrapping is a statistical method based on resampling with replacement that is used to calculate the distribution of a related test statistic.[20] Bootstrapping makes it possible to avoid making distributional assumptions.[21] To apply bootstrapping to compare datasets, we use similarity scores as the test statistic. Similarity scores are scores from 0-1 with 0 being the most different and 1 being the most similar. There are many different types of similarity scores that differ by the mathematical measure for distance including Cosine score,[22] Bray-Curtis,[23,24] and Jenson-Shannon.[25,26]

In this work, we establish a statistical method that determines if observations could have occurred by chance and compares datasets to establish if they are significantly different from each other as described in Figure 1. In this analysis, the population is all crosslinks formed, and the sample is all crosslinks that were correctly identified. Using the residue numbers crosslinked together as coordinates, similarity scores can be applied to determine the similarity of datasets. This results in a quantitative comparison that can be used on previously analyzed, qualitative data and does not depend on isotopic labeling. We apply this method to crosslinking data here, but this method could be used with many other types of data.

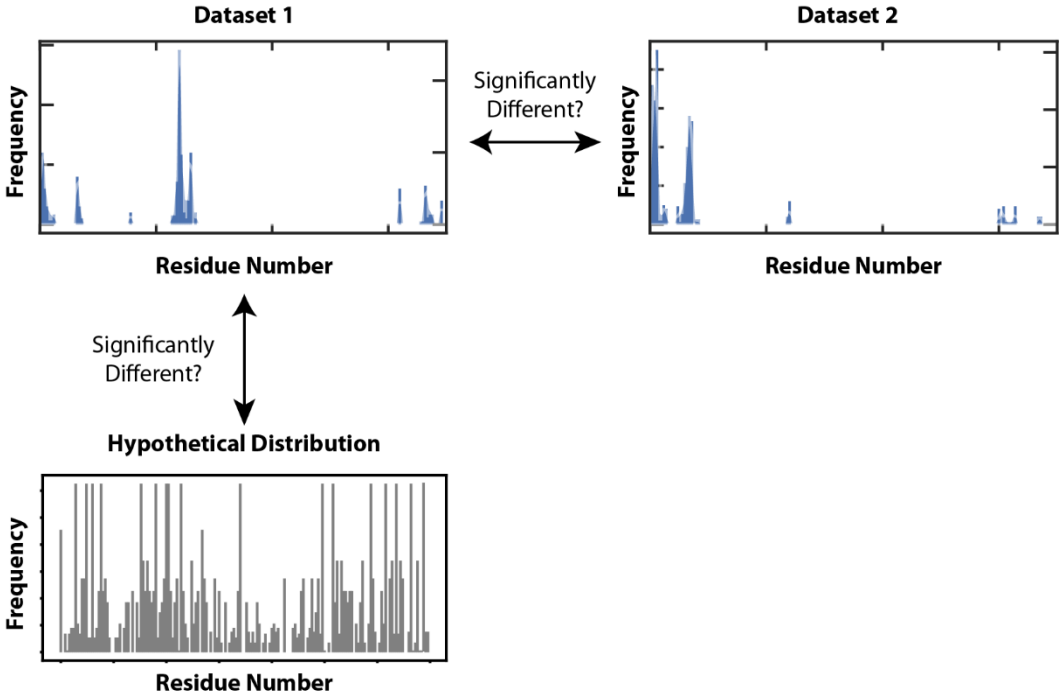

Figure 1. The goals of this study are to establish if datasets are significantly different from each other and if datasets are consistent with hypothetical probability distribution functions. The crosslinking results in this figure have been described previously (dataset 1 BPA site 24 and dataset 2 BPA site 33),[27] and the hypothetical distribution is described in more detail in Figure 3.

3.2 Methodology

3.2.1 Source of Experimental Data

The data analyzed here uses the non-canonical amino acid benzoylphenylalanine (BPA) as a crosslinking reagent. BPA is unique as a crosslinker in that it can be incorporated site-specifically[28] and reacts with all amino acids when exposed to UV light.[29] In the work described here, we incorporate BPA within the human small heat shock protein, HSPB5. HSPB5 is a highly dynamic, heterogenous system, and the BPA residue was incorporated at different positions in the highly flexible N-terminus. This results in many different crosslinks identified from a single BPA position. Experimental details have been described previously.[27,30,31] This work describes an additional statistical analysis of previously reported crosslinking results. Publications containing the raw data or crosslinks that are further statistically analyzed here are cited in the relevant figure captions.

3.2.2 Terminology

Here, we define a crosslink as a number of peptide spectral matches (PSMs) that correspond to crosslinks between particular residues. The crosslink density is the normalized value of the frequency of crosslinks (the value for each residue divided by the total sum of values, resulting in values that add up to 1). In previous work,[27,30,31] the crosslink density was indicated in histograms with the relative population axis. A matrix of the crosslink density to each residue is used to calculate either similarity scores, maximum density, or variance (standard deviation squared).

The general bootstrapping procedure is illustrated in Figure 2. The probability distribution function (PDF) is what is resampled with replacement. In resampling with replacement, a selection is made from the starting distribution and that selection is put back

before the next selection. It results in each selection coming from the starting distribution, creating the differences between the resamples. Based on each resample, a test statistic is calculated. In Figure 2, the test statistic is the Jensen-Shannon similarity score. Resampling and calculating a test statistic is repeated many times to generate a distribution of test statistics, which can be used to establish thresholds of statistical significance.

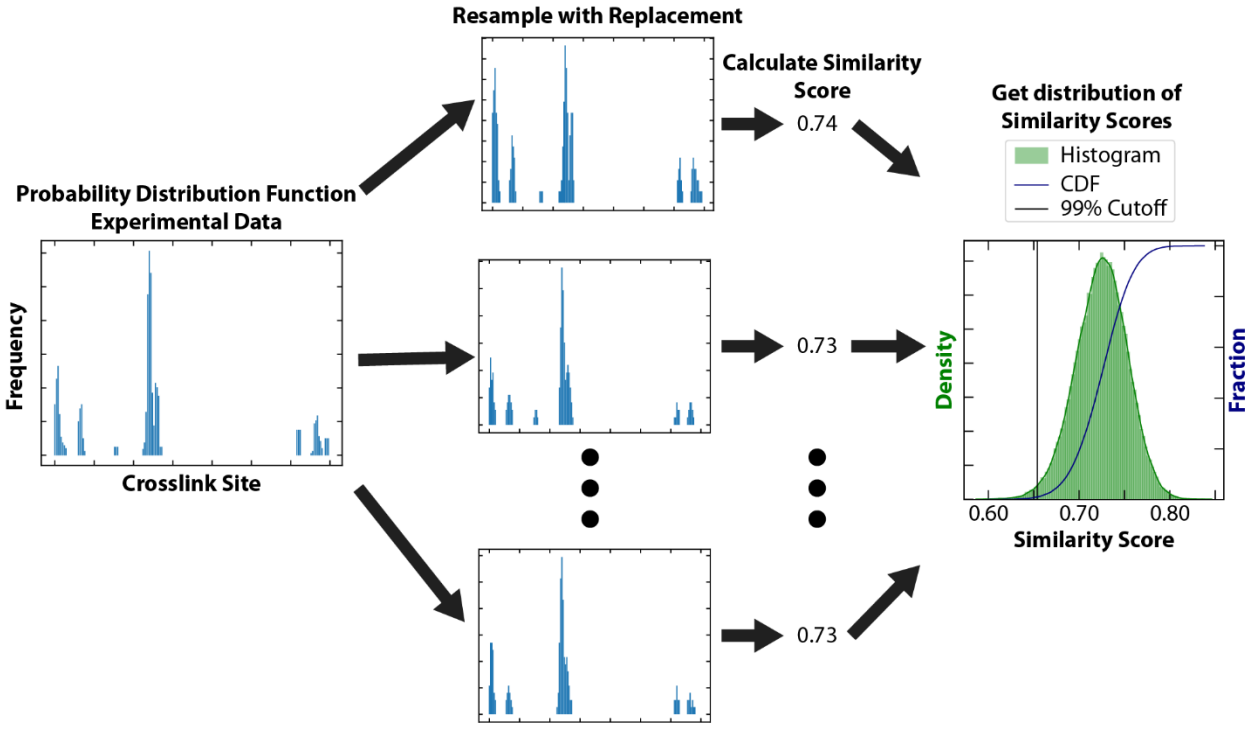

Figure 2. The bootstrapping workflow for determining if datasets are similar to each other is illustrated here. A probability distribution function (PDF) is resampled with replacement, and based on the resampled distributions, a test statistic is calculated. Resampling and calculating a test statistic is repeated many times to generate a distribution of the test statistic that can be used to get thresholds of statistical significance. In this example, an experimental dataset is the PDF, and the Jensen-Shannon similarity score is the test statistic. Similarity

score distributions for the Bray-Curtis and Cosine similarity scores are shown in Figure S1.

3.2.3 Code Availability

The code used in this analysis will be published on GitHub. For each analysis, 100,000 resamples were performed. The number of selections made in each resample was determined by the number of PSMs in the comparison dataset.

3.3 Results and Discussion

3.3.1 Need for Method

Often, experiments are designed to compare and look for differences across two states such as diseased/healthy and bound/apo. However, there is an unmet need for a quantitative way to compare datasets with different numbers of observations.

Additional validation is difficult for datasets that deal with distributions of integers such as the PSMs and residue number histograms discussed here. Poisson statistics is relevant for rare events and counting-based datasets.[32,33] For the datasets analyzed here, the low number of PSMs observed compared to the number of possible residues makes the observations rare occurrences, and many other crosslinking datasets also fit this criteria. Because Poisson statistics apply to datasets that are based on relatively low numbers of PSMs, random number generation is not sufficient. Random number generation does not mimic datasets based on relatively low numbers of PSMs because of the low number of observations compared to the number of possibilities. Bootstrapping is a promising alternative because it captures Poisson-like behavior, is simpler than other existing mathematical methods for analyzing Poisson datasets,[32,33] and is often

recommended for the analysis of small samples and samples with unknown or non-normal dsitributions.[34] Bootstrapping has also been applied to large datasets with many observations.[35,36]

In this work, qualitative data of crosslink identifications are interpreted and compared quantitively with bootstrapping. This approach, sampling with replacement, assumes that each observation has the same probability. Factors such as ionization efficiency and difficulties in identifying PSMs limit the feasibility of these assumptions. More quantitative MS methods using techniques such as SILAC[11,13] or isotopically labeled crosslinkers[16,17] could help avoid making these assumptions. However, this method allows for a more thorough interpretation of existing qualitative data without additional experiments.

3.3.2 Evaluation of Similarity Scores

Similarity scores are scores from 0 to 1, with 0 being the most different and 1 being the most similar. There are many different types of bivariate similarity scores that differ by the mathematical measure for distance including Cosine score,[22] Bray-Curtis,[23,24] and Jenson-Shannon.[25,26] Cosine score has most often been used with fuzzy databases (databases that can handle incomplete or uncertain information) such as text analysis.[37,38] The Bray-Curtis score originated in ecology, but there is a fair amount of literature critiquing its use.[24,39] The Jensen-Shannon divergence score has been used in genomics before.[40,41] Figure 3 compares values obtained from these different similarity scores when comparing the same datasets of either 9 replicates (datasets expected to be the same), bootstrap distributions of a single replicate, and crosslinks from different sites of BPA incorporation (datasets expected to be different). Figure 3 illustrates that the cosine score has less variance in the similarity scores for different replicates and is higher in value than the Jensen-Shannon and Bray-Curtis scores. Figures S2, S3, and S4 illustrate the similarity scores for the replicates represented in Figure 3. The Jensen-Shannon and

Bray-Curtis scores have more variance in the scores between replicates, and the Jensen-Shannon score distribution is centered slightly lower, but the two distributions overlap significantly. In Figure 3, the values from comparing bootstrap distributions overlap significantly with the values from comparing multiple replicates and exhibit the same trend with the cosine score being higher than both the Jensen-Shannon and Bray-Curtis scores. This illustrates that bootstrapping a single replicate leads to a similar degree of variability in similarity score as taking multiple replicates, which suggests that bootstrapping approximates the differences expected between replicates.

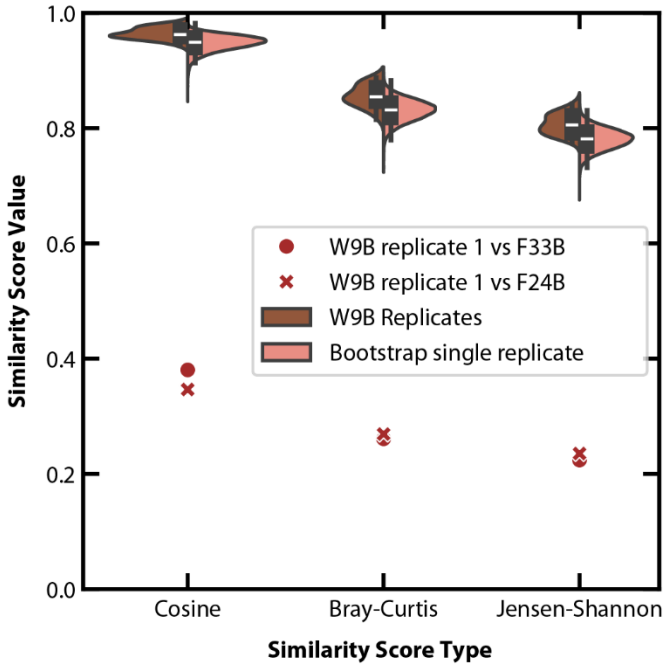

Figure 3. Here, the similarity scores of different types for replicates, bootstrapped distributions of one replicate, and different sites of BPA incorporation are compared. The Cosine, Bray-Curtis, and Jensen-Shannon similarity scores for the 9 replicates of the W9B variant of HSPB5 are shown as the left half of the violin plot in orange. This corresponds to 36 similarity scores from comparing each replicate with every other replicate. Figures S2, S3, and S4 illustrate these individual scores and the underlying histogram. The similarity scores from

bootstrapped distributions of a single replicate are shown on the right half of the violin plot in pink. To generate this, replicate one was resampled with replacement 100,000 times, and each resampled distribution was compared to another resampled distribution to calculate a similarity score without repeating any distributions, resulting in 50,000 similarity scores. Figure S5 depicts the histogram of similarity scores from bootstrapping analysis. The similarity scores comparing a single replicate to both the F24B and L33B variants are shown as red points. The experimental data and identified crosslinks represented has been reported previously.[27]

When looking at the similarity score values comparing crosslinks from the first W9B replicate to both the F24B- and L33B-varaints, the Cosine scores of 0.35 and 0.38 (comparing to F24B and L33B, respectively) are higher than the Jensen-Shannon scores of 0.24 and 0.22 and Bray-Curtis scores of 0.27 and 0.26. Similarly to the replicates, the Jensen-Shannon and Bray-Curtis scores are very close in value, and the cosine score is higher. Notably, these values are much lower than when comparing replicates, and the different score types still yield different values.

The different values obtained from different similarity scores illustrates that different similarity scores are not equivalent. For example, a Cosine score of 0.5 means something different than a Jensen-Shannon or a Bray-Curtis score of 0.5. The difference in scores suggests that similarity scores themselves are not necessarily sufficient to establish if datasets are similar or different. Here, we implement a bootstrapping-based statistical method to contextualize the results of different similarity scores. Because of the similar previous applications, we focus on

the Jensen-Shannon divergence score in the main text, but results from the Cosine and Bray-Curtis score are shown in the SI.

Although similarity scores enable quantitative comparisons between pairs of data, on their own, they do not address the significance of the similarity. Could the results have occurred by chance? Are the two distributions similar or different? Similarity scores do provide an idea of what is more similar (higher score) or less similar (lower score), but they need additional validation to determine what values are significant and represent a significant difference.

3.3.3 Applying Bootstrapping to Hypotheses

Here, we applied the same general bootstrapping procedure to two different questions: Are these observations consistent with a hypothesis? Are pairs of observations consistent with each other, or significantly different? To determine if observations could have occurred by chance, we used the null hypothesis of crosslinks are consistent with an alternative PDF, PDFs based on reactivity (see results for more information), and the test statistics of maximum density and variance. To determine if distributions are similar or different, we used the null hypothesis of datasets probe the same interaction, a PDF of an experimental dataset, and the test statistic of the Jensen-Shannon similarity score.

3.3.4 Comparing to a Hypothetical Distribution

Our goal here is to test a hypothesis by representing the hypothesis as a comparison distribution. Here, we determine if crosslinks were formed randomly based on comparing to a reactivity-based probability distribution function. The probability distributions used to assess if datasets could have occurred by random chance are shown in Figure 4A. Both a uniform distribution, where the probability to each residue is equal, and a rate constant distribution, where the probability to each residue corresponds to the rate of BPA's reaction with each free

amino were used. The rate constant probability distribution function was calculated based on previous data as described in the SI.[42] The uniform PDF represents if each residue in the sequence is equally likely to form crosslinks, and the rate constant PDF represents if the chemistry of BPA's reaction is the sole driver of crosslink formation. 4B and 4C illustrate results from the rate constant PDF. Similar results from the uniform PDF are shown in Figure S6.

Figure 4B shows distributions of the maximum density and variance test statistics that were calculated based on resamples from the rate constant PDF. Maximum density is the maximum crosslink density. Variance is the squared standard deviation of the crosslink density values. The 99 and 99.99% thresholds represent values that are unlikely to result from the probability distribution function. Upper cutoffs are used for maximum density and variance because those values should be lower if results are nonrandom. The number of PSMs represents how many selections are made with replacement during each resample. For example, in the distributions labeled as 100 PSMs, 100 selections with replacement were made from the PDF to create each resampled distribution. The maximum density distribution has discrete values at lower numbers of PSMs (100) and becomes more continuous at higher numbers of PSMs (1000). The variance distributions were continuous across PSM values. The discrete values of maximum density at lower PSM values are because with less PSMs the PDF is not sampled as thoroughly, so it is more likely that the same values are obtained multiple times.

Figure 4C illustrates the thresholds for statistical significance at 99 and 99.99% confidence levels from resampling the rate constant PDF and the values from experimental datasets. The number of PSMs considered in Figure 4C ranges from 10 to 310 with an interval of one. The threshold for statistical significance increases quite sharply at lower number of PSMs and level off at about 80 PSMs. The experimental data considered here identified between 45 and

375 PSMs. For each experimental dataset, the maximum density and variance values lie well above the 99.99% threshold for the corresponding number of PSMs, indicating that the results are not consistent with the rate constant probability distribution function.

With fewer PSMs, not all features of the data are observable because of the reduced sampling. Not observing all of the features results in increases in the maximum density and variance and higher thresholds for statistical significance, and indicates how results with less PSMs need to be interpreted more strictly. Even the experimental datasets considered here with lower numbers of PSMs (45) were well above the thresholds for statistical significance. This suggests that by mimicking the effect of the number of PSMs observed with the number of selections made during each resample, our bootstrapping method can account for differences in confidence that results from differing numbers of identified PSMs. Varying the number of PSMs through varying how many selections are made in each resample mimics the experimental variance from how many PSMs are identified.

That each experimental value in 4C is above the thresholds of statistical significance indicates that we can reject the null hypothesis and conclude that the experimental data is not consistent with the rate constant PDF. Experimental data being inconsistent with the rate constant PDF indicates that BPA's reactivity is not the sole driving factor behind crosslink formation, so the crosslinks formed due to the interactions and structure of the target protein.

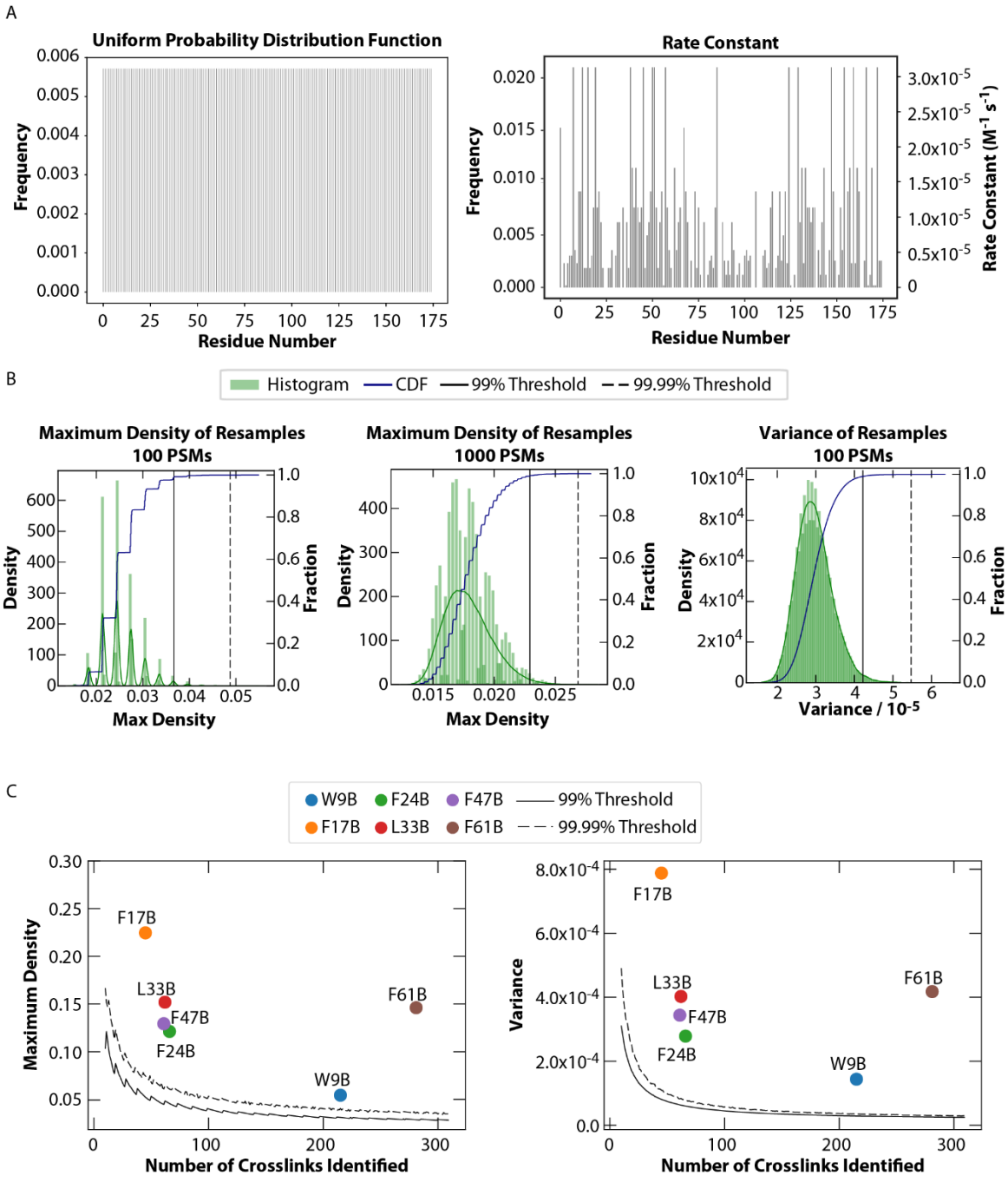

Figure 4. Panel A shows the uniform and rate constant PDFs. Panel B shows distributions of maximum density and variance from resampling the rate constant PDF. The PSMs value indicates how many selections were made to

generate each resampled distribution. Panel C illustrates the 99 and 99.99% thresholds for variance and maximum density over a range of PSM values and compares them with values from experimental data. All experimental data points lie above the thresholds, indicating that results are not consistent with the rate constant PDF. The experimental data and identified crosslinks represented in panel C have been reported previously.[27,30]

3.3.5 Comparing Datasets to Each Other

Our goal here is to determine if crosslinks from different datasets are significantly different. Here, we determine if crosslinks from different sites of BPA incorporation are significantly different. However, this could be applied to many situations to determine if crosslinks vary due to any change in the sample such as introduction of a ligand or a diseased state or drug treatment. The method used to evaluate if two distributions are similar to each other is illustrated in Figure 2. An experimental dataset is used as the PDF that is resampled with replacement many times. This experimental dataset is called the PDF dataset. The similarity score between each resampled distribution and the PDF dataset is calculated. Resampling and calculating the similarity score is repeated many times (100,000) to get a distribution of similarity scores. The dataset that is compared to the PDF dataset is the comparison dataset. The number of PSMs in the comparison dataset determines how many selections are made when making each resample of the PDF dataset. The similarity score between the PDF dataset and the comparison dataset is the sample score. The sample score is compared to the similarity score distribution to determine if the null hypothesis of the datasets probing the same interactions can

be rejected. The lower end of the similarity score distribution is used to determine if there is statistical significance because if the datasets are different, the similarity score will be lower.

For example, in Figure 2, the PDF dataset is from F24B. The comparison dataset is from L33B. L33B has 62 PSMs. For each resample of the PDF, 62 selections are made, and the similarity score between the resample and F24B is calculated. Resampling and calculating the similarity score is repeated 100,000 times to yield the similarity score distribution shown in Figure 2. The sample score (similarity score between F24B and L33B) is 0.34. This sample score of 0.34 is below all values of the similarity score distribution, so the null hypothesis of datasets probing the same interactions can be rejected, and we can conclude that the F24B and L33B datasets probe different interactions. F24B and L33B are observed in the same peptide under the digestion conditions used in this study,[27] so that they probe different interactions is very strong evidence that BPA crosslinking reveals site specific information.

Figure 5 shows the results from comparing 6 different BPA sites. The x-axis indicates the comparison dataset, and the y-axis indicates the PDF dataset. The diagonal contains values for the median of the similarity score distribution. Values are represented as the probability of obtaining a lower score. The probability of obtaining a lower score is the intersection of the sample score and the cumulative distribution function of the similarity score distribution. This analysis is based on 100,000 resamples. Therefore, for sample scores outside of the similarity score distribution, the probability is less than 1 in 100,000 (i.e., 10^{-5}). The diagonal in Figure 5 is all values of 0.5 which is as expected for the median of the distribution because half of the values in the distribution should be below the median. All the comparisons represented in Figure 5, except for one, had sample scores outside of the distribution and probabilities of obtaining a lower similarity score of less than 10^{-5}. The sample score that did intersect the distribution

intersected at the lower end and had only a $4.5x10^{-4}$ probability of obtaining a lower score. The low probability of $4.5x10^{-4}$ means there is only a 0.045% chance of the distribution accounting for the sample score, so the null hypothesis is rejected.

Because all the probability values in Figure 4 are $4.5x10^{-4}$ or lower, we can reject the null hypothesis and determine that different sites of BPA incorporation probe different interactions. That different sites of BPA incorporation probe significantly different interactions establishes that the site of BPA incorporation changes what interactions are detected. This supports the idea that the BPA crosslinking method is site-specific. The bootstrapping method of comparing datasets could also be applied to many data types to compare datasets obtained from different biological conditions.

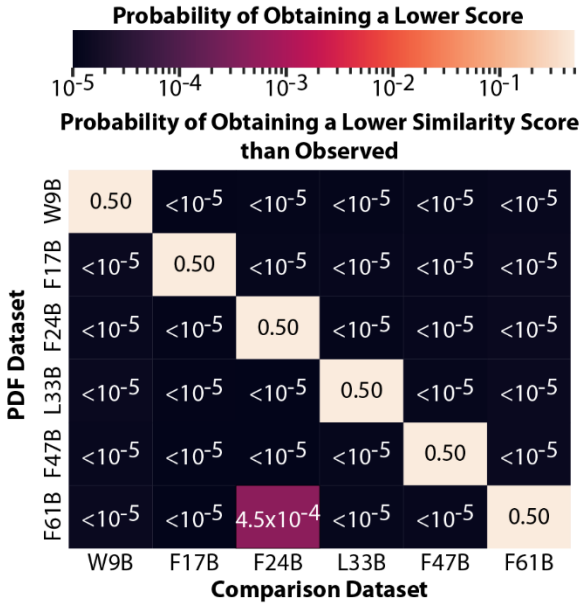

Figure 5. Results from comparing datasets are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. This analysis is based on 100,000 resamples. Therefore, for sample scores outside of the similarity score distribution, the probability is less than 1 in 100,0000 (10^{-5}).

Datasets are indicated by BPA mutation site, and the diagonal corresponds to the median value of the distribution. Results here are from the Jensen-Shannon similarity score. Figure S7 shows results from the Bray-Curtis similarity score, and Figure S8 shows results from the cosine similarity score. The experimental data and identified crosslinks represented in panel C have been reported previously.[27,30]

3.3.6 Comparing Replicates

Figure 6 shows results for comparing 9 replicates to each other. In this analysis, the null hypothesis is that datasets probe the same interaction. This analysis is the same as comparing different sites of BPA incorporation in Figure 5, except here we are evaluating replicates that we expect to be the same. About 30 of the 72 comparisons reflected in Figure 6 lie above the similarity score distribution that results from bootstrapping. The remaining scores all lie somewhere within the distribution to varying degrees with the single lowest value being 5.7×10^{-2}, but most scores are much higher. Because sample scores lie within or above the similarity score distributions, we fail to reject the null hypothesis and conclude that datasets do not probe different interactions. This illustrates that we do not meet thresholds for differences when comparing replicates.

The results in Figure 6 are in stark contrast to Figure 5 which compares different BPA sites where almost every sample score was below the similarity score distribution, and the null hypothesis was rejected. In many cases when analyzing replicates, we obtain the opposite result where the sample score lies above the similarity score distribution instead of below. That we obtain many values above the similarity score distribution strengthens the conclusions from

comparing different BPA sites by illustrating that is possible to obtain the opposite result where sample scores are above the similarity score distribution. Because we obtained many values above the similarity score distribution, we may be overestimating in the similarity score distributions (obtaining higher sample scores and higher probabilities of obtaining lower scores) with this method. This overestimation would represent a Type II error (false negative) in that it is more difficult to reject the null hypothesis because values below the similarity score distribution are needed to reject it in this one-tailed test, and values above the distribution are also possible.

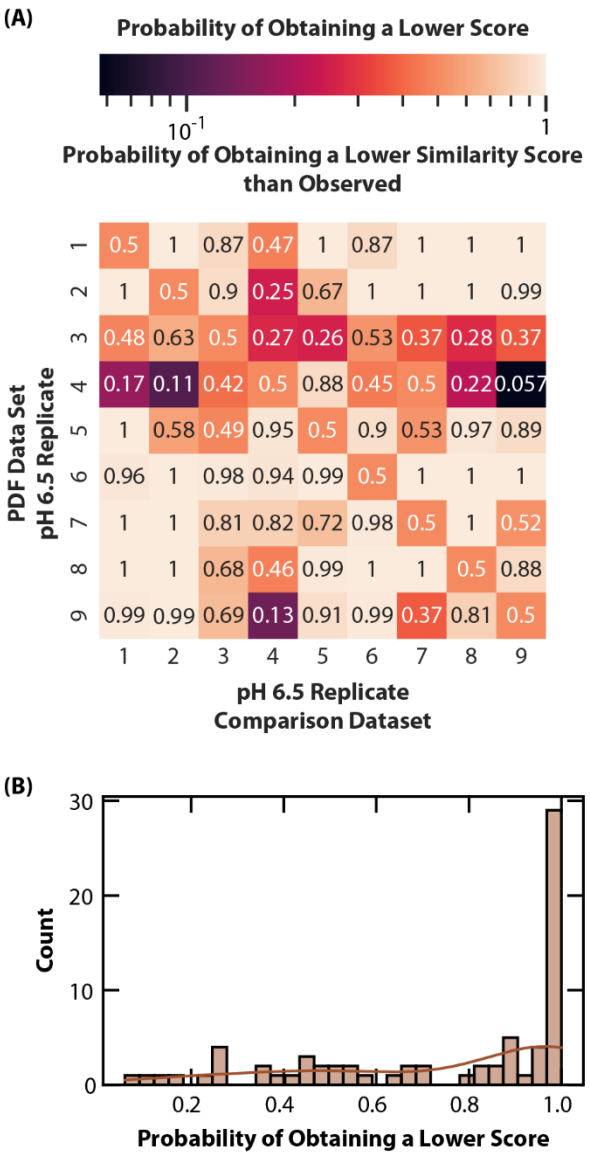

Figure 6. Results from comparing replicates are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. This analysis is based on the Jensen-Shannon similarity score. Figures S9 and S10 show the results from using Cosine and Bray-Curtis similarity scores. A value of 1 corresponds to sample scores that lie above the similarity score distribution. This analysis is based on 100,000 resamples. In Panel A, Datasets are indicated by a number corresponding to the order in which replicates were collected. The diagonal corresponds to the median value of the distribution. Panel B shows a distribution of the values in panel A, excluding the diagonal. The experimental data and identified crosslinks represented in panel here have been reported previously.[27]

3.4 Conclusion

Here we describe a highly versatile, bootstrapping based statistical method to quantitatively compare datasets that is illustrated in Figure 2. We applied this method to BPA crosslinking to assess the hypotheses that datasets are not consistent with an alternative distribution (Figure 4) and that datasets probe different interactions (Figure 5) through using the test statistics of maximum density, variance, or similarity score and the PDFs of uniform, rate constant, and experimental data. These test statistics and PDFs could be changed to any other metric to assess a wide variety of hypotheses about a wide range of data. Through applying this method to BPA crosslinking, we determined that results are not consistent with random chance or BPA's reactivity (Figure 4) and that different sites of incorporation yield significantly different results (Figure 5). These conclusions suggest that BPA crosslinking does effectively measure aspects of protein structure and interaction that are specific to the residue of incorporation.

The bootstrapping method presented here provides a way to quantitatively compare datasets that are typically more difficult to analyze further. Data such as the BPA crosslinking data presented here which consists of integer values or counts of observations follow Poisson statistics and become more difficult to analyze. Other types of crosslinking and covalent labeling experiments and non-MS based experiments that are distributions of integer values would also follow Poisson statistics and benefit from this analysis.

3.5 Acknowledgements

I thank Lucas Narisawa, Christopher N. Woods, Natalie L, Stone, Maria Janowska, Rachel E. Klevit, and Matthew F. Bush for their contributions to this work. This material is based upon work supported by the National Eye Institute through R01EY017370 to R.E.K., the

National Institute of General Medical Sciences through T32 GM008268 to C.N.W., the National Institute of Aging through T32 AG066574 to L.D.U., and the University of Washington's Proteomics Resource (UWPR95794).

3.6 References

(1) Klykov, O.; Steigenberger, B.; Pektaş, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990. https://doi.org/10.1038/s41596-018-0074-x.

(2) Yu, C.; Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **2018**, *90* (1), 144–165. https://doi.org/10.1021/acs.analchem.7b04431.

(3) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; Chaignepain, S.; Chavez, J. D.; Claverol, S.; Cox, J.; Davis, T.; Degliesposti, G.; Dong, M.-Q.; Edinger, N.; Emanuelsson, C.; Gay, M.; Götze, M.; Gomes-Neto, F.; Gozzo, F. C.; Gutierrez, C.; Haupt, C.; Heck, A. J. R.; Herzog, F.; Huang, L.; Hoopmann, M. R.; Kalisman, N.; Klykov, O.; Kukačka, Z.; Liu, F.; MacCoss, M. J.; Mechtler, K.; Mesika, R.; Moritz, R. L.; Nagaraj, N.; Nesati, V.; Neves-Ferreira, A. G. C.; Ninnis, R.; Novák, P.; O'Reilly, F. J.; Pelzing, M.; Petrotchenko, E.; Piersimoni, L.; Plasencia, M.; Pukala, T.; Rand, K. D.; Rappsilber, J.; Reichmann, D.; Sailer, C.; Sarnowski, C. P.; Scheltema, R. A.; Schmidt, C.; Schriemer, D. C.; Shi, Y.; Skehel, J. M.; Slavin, M.; Sobott, F.; Solis-Mezarino, V.; Stephanowitz, H.; Stengel, F.; Stieger, C. E.; Trabjerg, E.; Trnka, M.; Vilaseca, M.; Viner, R.; Xiang, Y.; Yilmaz, S.; Zelter, A.; Ziemianowicz, D.; Leitner, A.; Sinz, A. First Community-Wide, Comparative Cross-Linking

Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961. https://doi.org/10.1021/acs.analchem.9b00658.

(4) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 742. https://doi.org/10.1038/s41467-020-14608-2.

(5) Webb, B.; Viswanath, S.; Bonomi, M.; Pellarin, R.; Greenberg, C. H.; Saltzberg, D.; Sali, A. Integrative Structure Modeling with the Integrative Modeling Platform: Integrative Structure Modeling with IMP. *Protein Sci.* **2018**, *27* (1), 245–258. https://doi.org/10.1002/pro.3311.

(6) Bullock, J. M. A.; Sen, N.; Thalassinos, K.; Topf, M. Modeling Protein Complexes Using Restraints from Crosslinking Mass Spectrometry. *Structure* **2018**, *26* (7), 1015-1024.e2. https://doi.org/10.1016/j.str.2018.04.016.

(7) Ferrari, A. J. R.; Gozzo, F. C.; Martínez, L. Statistical Force-Field for Structural Modeling Using Chemical Cross-Linking/Mass Spectrometry Distance Constraints. *Bioinformatics* **2019**, *35* (17), 3005–3012. https://doi.org/10.1093/bioinformatics/btz013.

(8) Degiacomi, M. T.; Schmidt, C.; Baldwin, A. J.; Benesch, J. L. P. Accommodating Protein Dynamics in the Modeling of Chemical Crosslinks. *Structure* **2017**, *25* (11), 1751-1757.e5. https://doi.org/10.1016/j.str.2017.08.015.

(9) Zeng-Elmore, X.; Gao, X.-Z.; Pellarin, R.; Schneidman-Duhovny, D.; Zhang, X.-J.; Kozacka, K. A.; Tang, Y.; Sali, A.; Chalkley, R. J.; Cote, R. H.; Chu, F. Molecular Architecture of Photoreceptor Phosphodiesterase Elucidated by Chemical Cross-Linking and Integrative Modeling. *J. Mol. Biol.* **2014**, *426* (22), 3713–3728. https://doi.org/10.1016/j.jmb.2014.07.033.

(10) Abe, R.; Caaveiro, J. M. M.; Kozuka-Hata, H.; Oyama, M.; Tsumoto, K. Mapping Ultra-Weak Protein-Protein Interactions between Heme Transporters of Staphylococcus Aureus. *J. Biol. Chem.* **2012**, *287* (20), 16477–16487. https://doi.org/10.1074/jbc.M112.346700.

(11) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Bruce, J. E. In Vivo Conformational Dynamics of Hsp90 and Its Interactors. *Cell Chem. Biol.* **2016**, *23* (6), 716–726. https://doi.org/10.1016/j.chembiol.2016.05.012.

(12) Chavez, J. D.; Keller, A.; Zhou, B.; Tian, R.; Bruce, J. E. Cellular Interactome Dynamics during Paclitaxel Treatment. *Cell Rep.* **2019**, *29* (8), 2371-2383.e5. https://doi.org/10.1016/j.celrep.2019.10.063.

(13) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Zheng, C.; Taipale, A.; Zhang, Y.; Takara, K.; Bruce, J. E. Quantitative Interactome Analysis Reveals a Chemoresistant Edgotype. *Nat. Commun.* **2015**, *6* (1), 7928. https://doi.org/10.1038/ncomms8928.

(14) Wippel, H. H.; Chavez, J. D.; Tang, X.; Bruce, J. E. Quantitative Interactome Analysis with Chemical Cross-Linking and Mass Spectrometry. *Curr. Opin. Chem. Biol.* **2022**, *66*, 102076. https://doi.org/10.1016/j.cbpa.2021.06.011.

(15) Chen, Z. A.; Rappsilber, J. Quantitative Cross-Linking/Mass Spectrometry to Elucidate Structural Changes in Proteins and Their Complexes. *Nat. Protoc.* **2019**, *14* (1), 171–201. https://doi.org/10.1038/s41596-018-0089-3.

(16) Schmidt, C.; Zhou, M.; Marriott, H.; Morgner, N.; Politis, A.; Robinson, C. V. Comparative Cross-Linking and Mass Spectrometry of an Intact F-Type ATPase Suggest a Role for Phosphorylation. *Nat. Commun.* **2013**, *4* (1), 1985. https://doi.org/10.1038/ncomms2985.

(17) Zhong, X.; Navare, A. T.; Chavez, J. D.; Eng, J. K.; Schweppe, D. K.; Bruce, J. E. Large-Scale and Targeted Quantitative Cross-Linking MS Using Isotope-Labeled Protein Interaction Reporter (PIR) Cross-Linkers. *J. Proteome Res.* **2017**, *16* (2), 720–727. https://doi.org/10.1021/acs.jproteome.6b00752.

(18) Yu, C.; Huszagh, A.; Viner, R.; Novitsky, E. J.; Rychnovsky, S. D.; Huang, L. Developing a Multiplexed Quantitative Cross-Linking Mass Spectrometry Platform for Comparative Structural Analysis of Protein Complexes. *Anal. Chem.* **2016**, *88* (20), 10301–10308. https://doi.org/10.1021/acs.analchem.6b03148.

(19) Chen, Z. A.; Rappsilber, J. Protein Dynamics in Solution by Quantitative Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43* (11), 908–920. https://doi.org/10.1016/j.tibs.2018.09.003.

(20) Mooney, C. Z.; Duval, R. D. *Bootstrapping: A Nonparametric Approach to Statistical Inference*; Sage University papers series; Sage Publications: Newbury Park, Calif, 1993.

(21) Haukoos, J. S. Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with "Difficult" Distributions. *Acad. Emerg. Med.* **2005**, *12* (4), 360–365. https://doi.org/10.1197/j.aem.2004.11.018.

(22) Li, B.; Han, L. Distance Weighted Cosine Similarity Measure for Text Classification. In *Intelligent Data Engineering and Automated Learning – IDEAL 2013*; Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., Yao, X., Eds.; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Series Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; Vol. 8206, pp 611–618. https://doi.org/10.1007/978-3-642-41278-3_74.

(23) Bray, J. R.; Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **1957**, *27* (4), 325–349. https://doi.org/10.2307/1942268.

(24) Gauch, H. G. A Quantitative Evaluation of the Bray-Curtis Ordination. *Ecology* **1973**, *54* (4), 829–836. https://doi.org/10.2307/1935677.

(25) Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37* (1), 145–151. https://doi.org/10.1109/18.61115.

(26) Menéndez, M. L.; Pardo, J. A.; Pardo, L.; Pardo, M. C. The Jensen-Shannon Divergence. *J. Frankl. Inst.* **1997**, *334* (2), 307–318. https://doi.org/10.1016/S0016-0032(96)00063-4.

(27) Ulmer, L.; Canzani, D.; Woods, C.; Stone, N.; Janowska, M.; Klevit, R.; Bush, M. High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid. *J. Proteome Res.* **2024**. https://doi.org/10.1021/acs.jproteome.4c00194.

(28) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proc. Natl. Acad. Sci.* **2002**, *99* (17), 11020–11024. https://doi.org/10.1073/pnas.172226299.

(29) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(30) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

(31) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(32) Thompson, W. J. Poisson Distributions. *Comput. Sci. Eng.* **2001**, *3* (3), 78–82. https://doi.org/10.1109/5992.919271.

(33) Coxe, S.; West, S. G.; Aiken, L. S. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *J. Pers. Assess.* **2009**, *91* (2), 121–136. https://doi.org/10.1080/00223890802634175.

(34) Egbert, J.; Plonsky, L. Bootstrapping Techniques. In *A Practical Handbook of Corpus Linguistics*; Paquot, M., Gries, S. Th., Eds.; Springer International Publishing: Cham, 2020; pp 593–610. https://doi.org/10.1007/978-3-030-46216-1_24.

(35) Fang, K.; Ma, S. Analyzing Large Datasets with Bootstrap Penalization. *Biom. J.* **2017**, *59* (2), 358–376. https://doi.org/10.1002/bimj.201600052.

(36) Kleiner, A.; Talwalkar, A.; Sarkar, P.; Jordan, M. I. A Scalable Bootstrap for Massive Data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76* (4), 795–816. https://doi.org/10.1111/rssb.12050.

(37) Lahitani, A. R.; Permanasari, A. E.; Setiawan, N. A. Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment. In *2016 4th International Conference on Cyber and IT Service Management*; IEEE: Bandung, Indonesia, 2016; pp 1–6. https://doi.org/10.1109/CITSM.2016.7577578.

(38) Liu, D.; Chen, X.; Peng, D. Interval-Valued Intuitionistic Fuzzy Ordered Weighted Cosine Similarity Measure and Its Application in Investment Decision-Making. *Complexity* **2017**, *2017*, 1–11. https://doi.org/10.1155/2017/1891923.

(39) Yoshioka, P. Misidentification of the Bray-Curtis Similarity Index. *Mar. Ecol. Prog. Ser.* **2008**, *368*, 309–310. https://doi.org/10.3354/meps07728.

(40) Sims, G. E.; Jun, S.-R.; Wu, G. A.; Kim, S.-H. Alignment-Free Genome Comparison with Feature Frequency Profiles (FFP) and Optimal Resolutions. *Proc. Natl. Acad. Sci.* **2009**, *106* (8), 2677–2682. https://doi.org/10.1073/pnas.0813249106.

(41) Itzkovitz, S.; Hodis, E.; Segal, E. Overlapping Codes within Protein-Coding Sequences. *Genome Res.* **2010**, *20* (11), 1582–1589. https://doi.org/10.1101/gr.105072.110.

(42) Deseke, E.; Nakatani, Y.; Ourisson, G. Intrinsic Reactivities of Amino Acids towards Photoalkylation with Benzophenone − A Study Preliminary to Photolabelling of the Transmembrane Protein Glycophorin A. *Eur. J. Org. Chem.* **1998**, *1998* (2), 243–251. https://doi.org/10.1002/(SICI)1099-0690(199802)1998:2<243::AID-EJOC243>3.0.CO;2-I.

3.7 Supporting Information

3.7.1 Calculating Rate Constants

Previous work reports the amount the yield from different amino acids reacting with BPA.[1] In that previous work, equal amounts of BPA were reacted with equal amounts of different amino acids and the percent of each amino acid converted is reported. We used the results reported from pyridine/water because our reaction was performed in an aqueous solution. We used the percent reacted reported to calculate the final concentrations of BPA and the amino acid. Because it is a dimerization reaction, we assumed that the reaction is first order in both BPA and the corresponding amino acid, so we used the second-order integrated rate law along with the reported reaction times and final concentration values to calculate the rate constants.

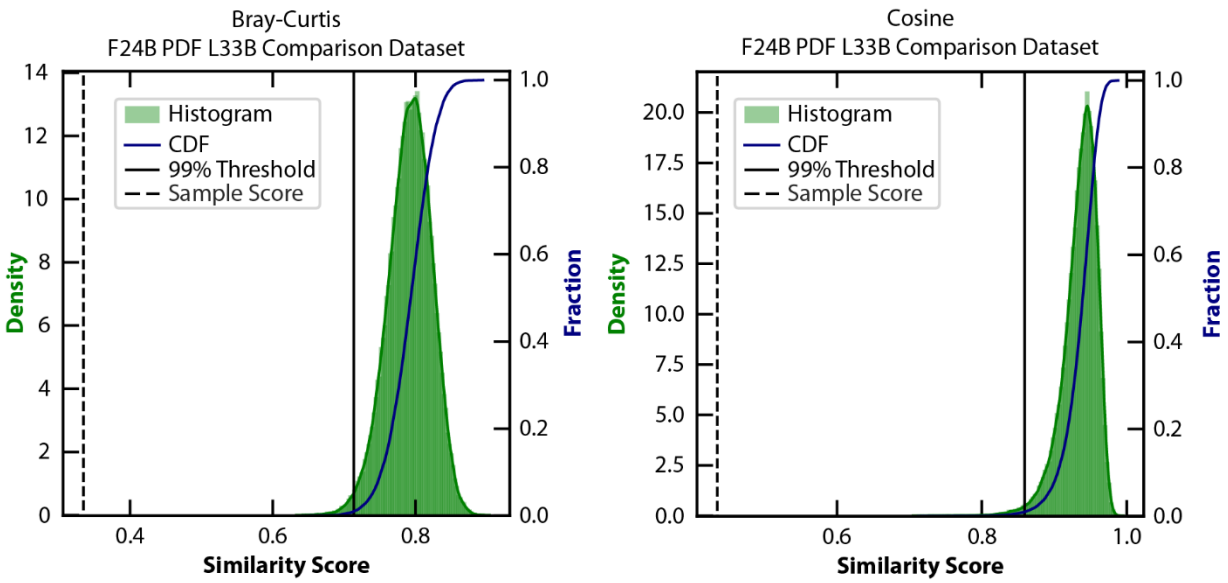

Figure S1. Example similarity score distributions from using the Bray-Curtis and cosine similarity scores with F24B as the PDF dataset and L33B as the comparison dataset are shown here. With Bray-Curtis, the sample score is 0.33, and with cosine, the sample score is 0.43. In both cases, the sample score is outside of the distribution, so we can reject the null hypothesis and conclude that the datasets are significantly different from each other. However, in comparison with main text Figure 2, it appears that the Bray-Curtis and cosine distribution are less gaussian looking and symmetrical than the Jensen-Shannon distributions. Both the Bray-Curtis and cosine similarity score distributions have some peak trailing at the lower end with the cosine distributions exhibiting more peak trailing.

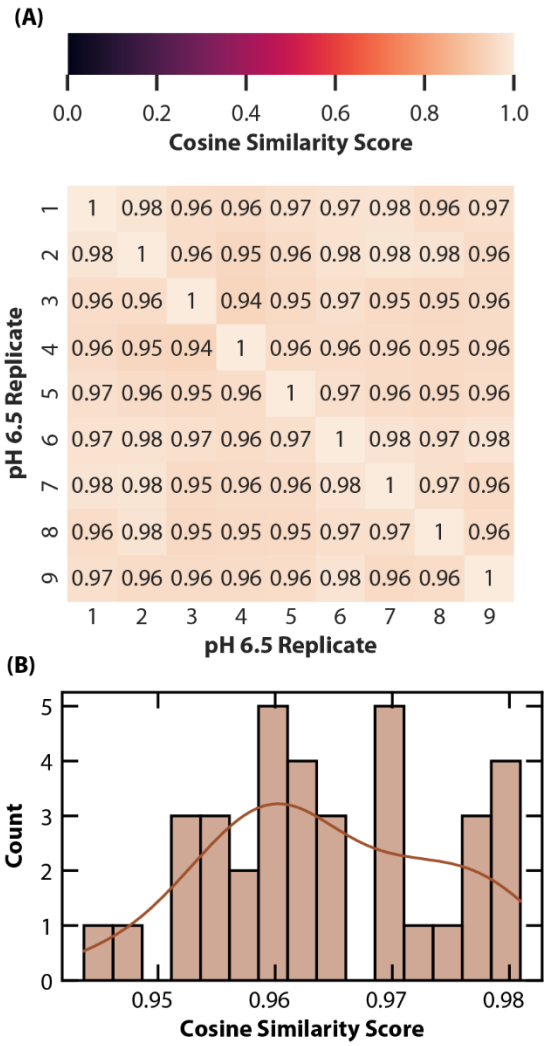

Figure S2. The Cosine similarity scores of 9 replicates of W9B-HSPB5 are shown. In Panel A, the diagonal represents comparing a given replicate to itself, and the plot is symmetric across the diagonal. In Panel B, the diagonal and repeat values are excluded so that the 36 values representing one comparison between each pair of datasets (i.e. all values above the diagonal) are represented in the distribution.

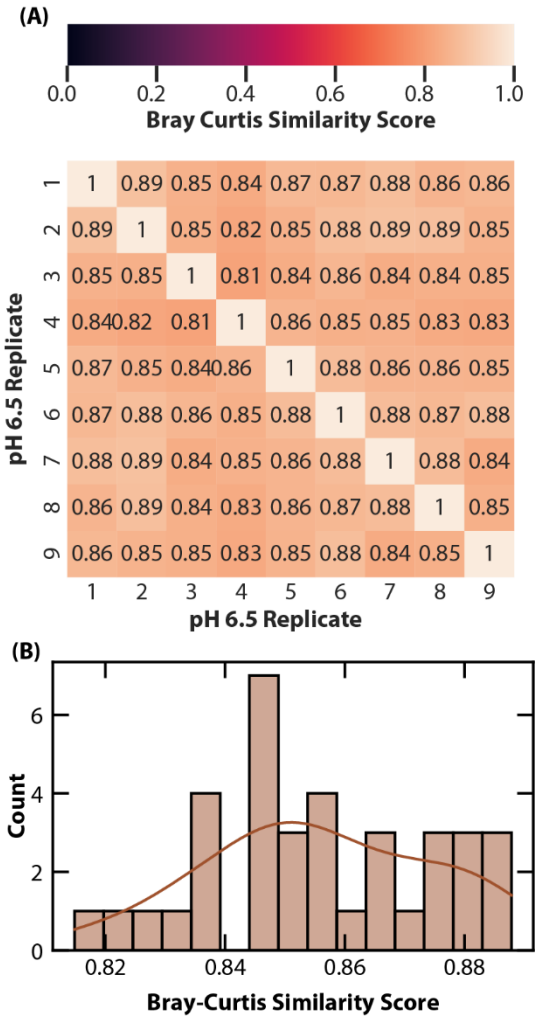

Figure S3. The Bray-Curtis similarity scores of 9 replicates of W9B-HSPB5 are shown. In Panel A, the diagonal represents comparing a given replicate to itself, and the plot is symmetric across the diagonal. In Panel B, the diagonal and repeat values are excluded so that the 36 values representing one comparison between each pair of datasets (i.e. all values above the diagonal) are represented in the distribution.

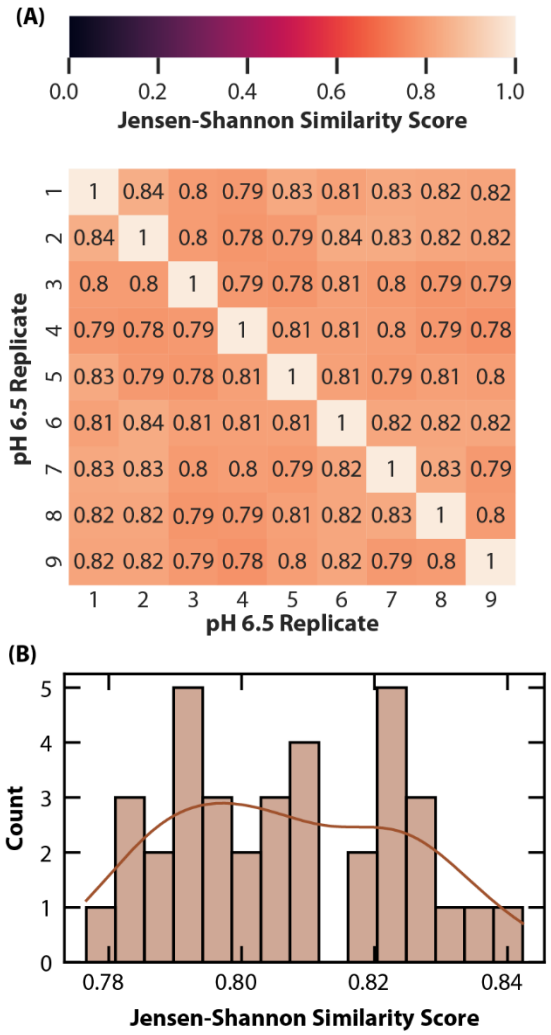

Figure S4. The Jensen-Shannon similarity scores of 9 replicates of W9B-HSPB5 are shown. In Panel A, the diagonal represents comparing a given replicate to itself, and the plot is symmetric across the diagonal. In Panel B, the diagonal and repeat values are excluded so that the 36 values representing one comparison between each pair of datasets (i.e. all values above the diagonal) are represented in the distribution.

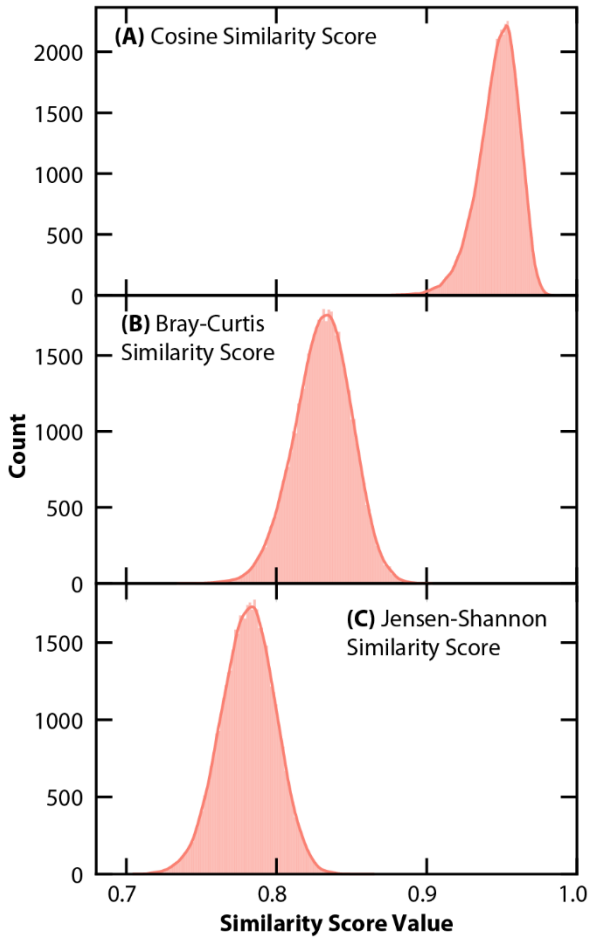

Figure S5. The distributions here reflect the similarity scores obtained when comparing bootstrapped distributions of a single replicate of trypsin digested W9B. Panel A contains the cosine similarity score, pane1 B contains the Bray-Curtis similarity score, and panel C contains the Jensen-Shannon similarity score. The scores here are from 100,000 resamples and from comparing each resample with the following without repeating any values. For example, resample 1 was scored with resample 2 and resample 3 with resample 4 and so on. This yielded the 50,000 similarity scores reflected in this each panel of this chart.

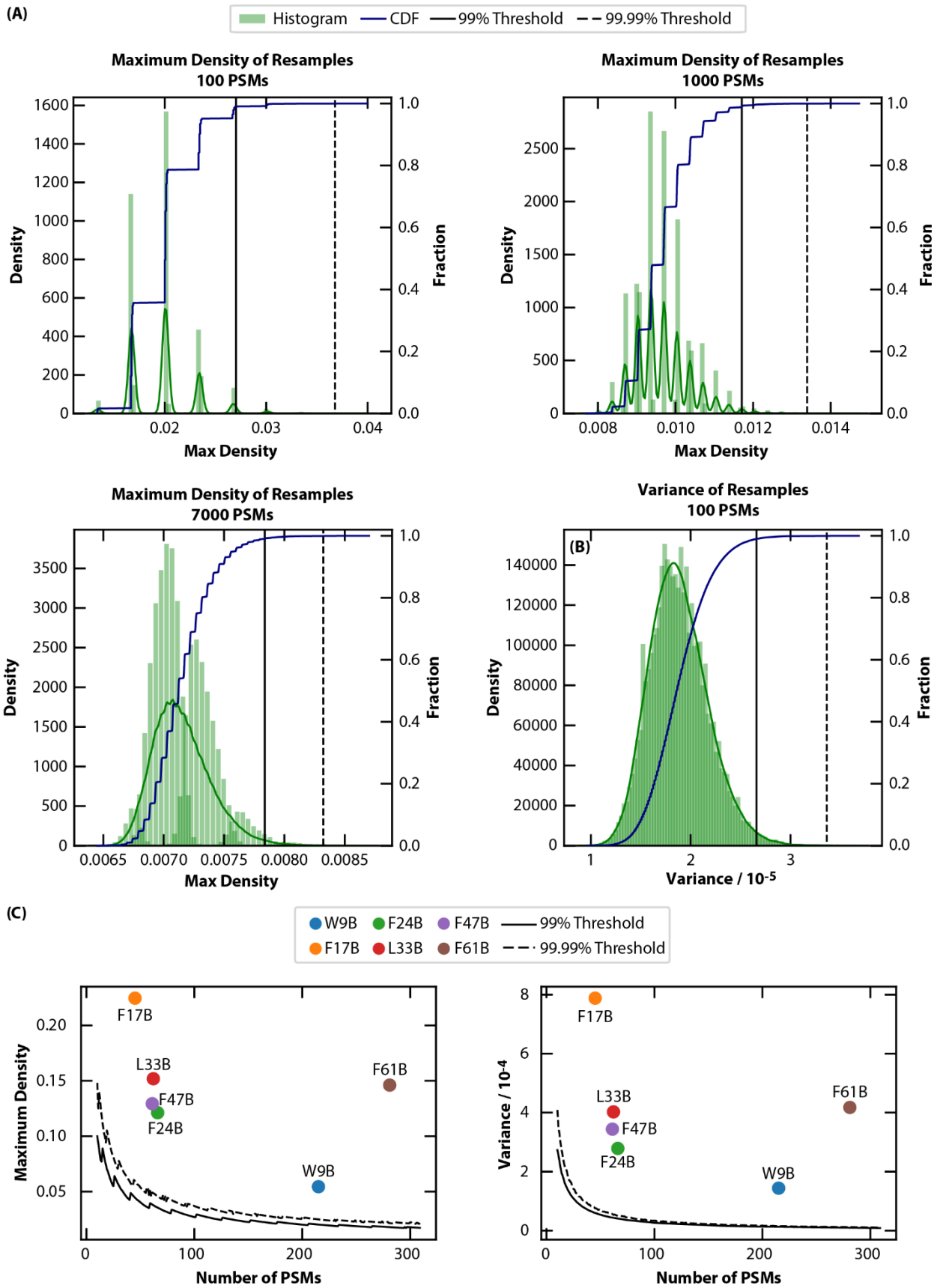
(A)
Histogram
CDF
99% Threshold
99.99% Threshold
Maximum Density of Resamples
100 PSMs
Density
Fraction
Max Density
Maximum Density of Resamples
1000 PSMs
Maximum Density of Resamples
7000 PSMs
Variance of Resamples
100 PSMs
(B)
Variance / 10-5
(C)
W9B
F24B
F47B
F17B
L33B
F61B
Maximum Density
Number of PSMs
Variance / 10-4

Figure S6. The results from resampling the uniform PDF (Figure 3A) are shown here. As shown in Panel A, The effect of having discrete values of maximum density at lower numbers of PSMs is observed for the uniform PDF as well and is more pronounced in that it takes a higher number of PSMs (7000 vs 1000) for the distribution to become more smooth an continous. It is unsurprising that a uniform distribution where every residue number has an equal frequency takes more sampling to smooth out. The variance distributions were continous across PSM values (Panel B) as was observed for the rate constant PDF (main text). In Panel C, the thresholds from the variance and maximum denisty distributions are compared to experimental data. Similarly to the rate constant PDF, the thersholds increase sharply at low numbers of PSMs and level off at around 80 PSMs. The value for each dataset is well above the threshold, so we can reject the null hypothesis and conclude that experimental data is not consistent with the uniform PDF. That experimental data is not consistent with the uniform PDF illustrates that results are not random.

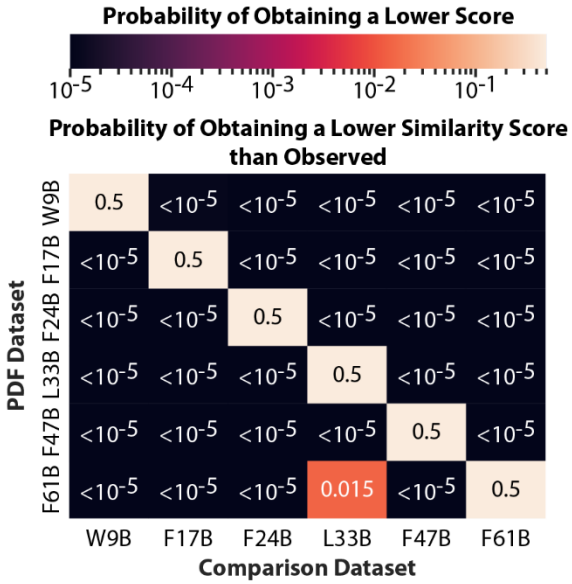

Figure S7. Results from comparing datasets using the Bray-Curtis similarity score are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. This analysis is based on 100,000 resamples. Therefore, for sample scores outside of the similarity score distribution, the probability is less than 1 in 100,0000 (10^{-5}). Datasets are indicated by BPA mutation site, and the diagonal corresponds to the median value of the distribution. The experimental data and identified crosslinks represented in panel C have been reported previously.[2,3] For each comparison except for one the sample score was outside of the distribution, so we can reject the null hypothesis and determine that the datasets probe different interactions. When F61B is the PDF dataset and L33B is the comparison dataset, the sample score is higher than 1.5% of the values in the similarity score distribution, so if operating at a 98.5% or lower confidence level, the null hypothesis can be rejected.

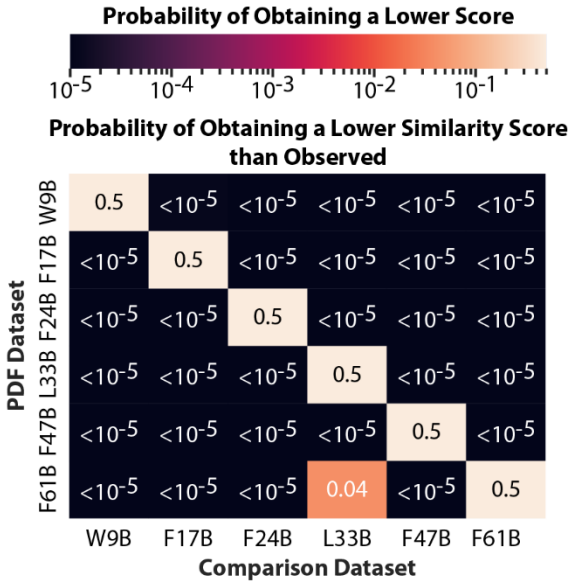

Figure S8. Results from comparing datasets using the Cosine similarity score are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. This analysis is based on 100,000 resamples. Therefore, for sample scores outside of the similarity score distribution, the probability is less than 1 in 100,0000 (10-5). Datasets are indicated by BPA mutation site, and the diagonal corresponds to the median value of the distribution. The experimental data and identified crosslinks represented in panel C have been reported previously.[2,3] For each comparison except for one the sample score was outside of the distribution, so we can reject the null hypothesis and determine that the datasets probe different interactions. When F61B is the PDF dataset and L33B is the comparison dataset, the sample score is higher than 4% of the values in the similarity score distribution, so if operating at a 96% or lower confidence level, the null hypothesis can be rejected.

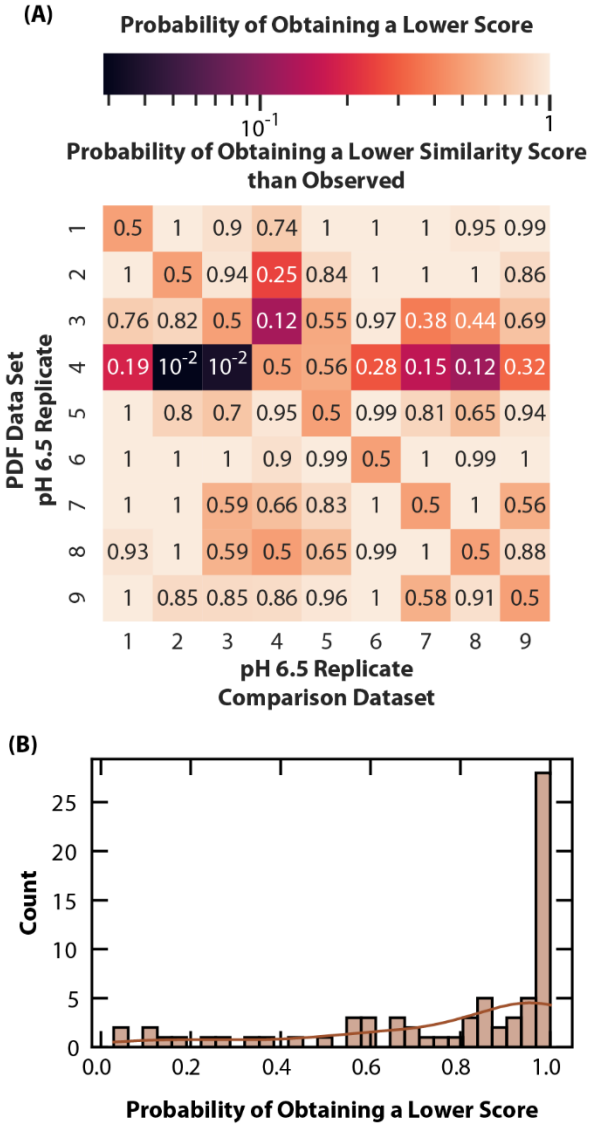

Figure S9. Results from comparing replicates using cosine similarity score are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. A value of 1 corresponds to sample scores that lie above the similarity score distribution. The intersection of comparison dataset 2 and PDF dataset 4 displayed as 10^{-2} due to the space is 2.9×10^{-2}, and the following intersection of comparison dataset 3 and PDF dataset 4 is 3.6×10^{-2}. This analysis is based on 100,000 resamples. In Panel A, Datasets are indicated by a number corresponding to the order in which replicates were collected. The diagonal corresponds to the median value of the distribution. Panel B shows a distribution of the values in panel A, excluding the diagonal. The experimental data and identified crosslinks represented in here have been reported previously.[2]

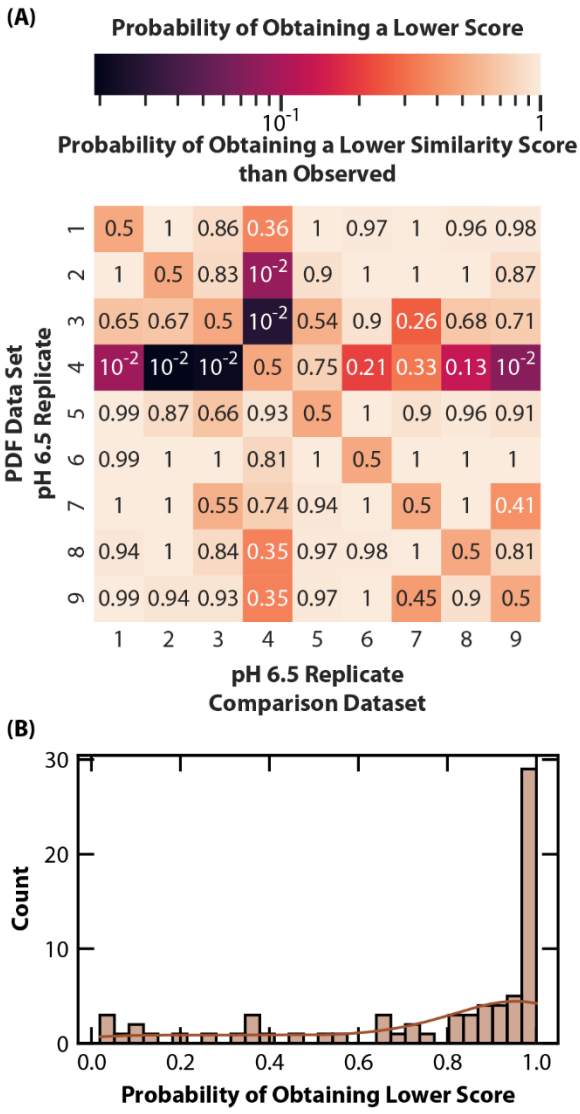

Figure S10. Results from comparing replicates using the Bray-Curtis similarity score are shown here. The probability of obtaining a lower score is determined by the intersection of the sample score and the cumulative distribution function of the similarity score distribution. A value of 1 corresponds to sample scores that lie above the similarity score distribution. The exact values of intersections of comparison datasets and PDF dataset displayed as 10^{-2} due to the space is as follows: 9.1×10^{-2} for comparison dataset 1 and PDF dataset 4, 1.9×10^{-2} for comparison dataset 2 and PDF dataset 4, 2.1×10^{-2} for comparison dataset 3 and PDF dataset 4, 2.8×10^{-2} for comparison dataset 4 and PDF dataset 3, 8.5×10^{-2} for comparison dataset 4 and PDF dataset 2, and 7.3×10^{-2} for comparison dataset 9 and PDF dataset 4. This analysis is based on 100,000 resamples. In Panel A, Datasets are indicated by a number corresponding to the order in which replicates were collected. The diagonal corresponds to the median value of the distribution. Panel B shows a distribution of the values in panel A, excluding the diagonal. The experimental data and identified crosslinks represented in here have been reported previously.[2]

3.7.2 Supporting Information References

(1) Deseke, E.; Nakatani, Y.; Ourisson, G. Intrinsic Reactivities of Amino Acids towards Photoalkylation with Benzophenone − A Study Preliminary to Photolabelling of the Transmembrane Protein Glycophorin A. *Eur. J. Org. Chem.* **1998**, *1998* (2), 243–251. https://doi.org/10.1002/(SICI)1099-0690(199802)1998:2<243::AID-EJOC243>3.0.CO;2-I.

(2) Ulmer, L.; Canzani, D.; Woods, C.; Stone, N.; Janowska, M.; Klevit, R.; Bush, M. High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid. *J. Proteome Res.* **2024**. https://doi.org/10.1021/acs.jproteome.4c00194.

(3) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

Chapter 4. Discovering Crosslinks with Targeted DDA

4.1 Introduction

Crosslinking mass spectrometry (XL-MS) is a powerful method for identifying protein-protein interactions and characterizing the spatial relationships within macromolecular assemblies.[1] XL-MS is especially useful for studying protein dynamics and transient interactions, thereby capturing information that is often challenging to obtain through standard structural biology methods.[2,3] In most XL-MS applications, a bifunctional chemical reagent reacts with samples prior to enzymatic digestion and liquid chromatography-mass spectrometry (LC-MS) analysis; the reagent binding sites give distance constraints that can be used to refine structural information and identify interacting partners.[2] However, XL-MS itself is limited by the challenges in identifying crosslinks, largely due to their low intensity relative to other components of the sample.[4–6]

The effects of different tandem MS acquisition methods on proteomics experiments have been described previously.[7,8] Here, we focus on aspects of these acquisition modes that are most relevant to crosslinking mass spectrometry. In untargeted data-dependent acquisition (DDA), precursor ions from the MS1 scan are automatically selected for fragmentation and MS2 analysis based on factors such as how intense the precursor ions are and when the precursor *m/z* value was last fragmented if dynamic exclusion is enabled.[9,10] Untargeted DDA's automatic selection of single precursor *m/z* values for MS2 results in little interference in the MS2 spectra because only a single *m/z* value is represented. The little interference in the MS2 spectra is especially important in XL-MS because crosslinked products contain two peptides so are already more complicated.[2] However, untargeted DDA's automated selection of precursors based primarily on

precursor intensity leads to the undersampling of many precursors in a sample, especially low-intensity precursors, such as crosslinks.[11]

In targeted DDA, a list of desired precursors *m/z* values is input for preferential MS2 analysis.[12] Similar to untargeted DDA, single precursor *m/z* values are selected for MS2 analysis, so MS2 spectra have little interference. Data can be analyzed using the same database search algorithms as untargeted DDA data. Defining precursors for analysis in targeted DDA can significantly increase the number of identifications.[13,14] Targeted DDA has been used to facilitate the identification of low-intensity peptides with rare PTMs.[15]

In untargeted data-independent acquisition (DIA), the instrument scans across a range of precursor values in a set interval, fragmenting and collecting MS2 spectra on everything within that interval.[16] Because untargeted DIA fragments and collects an MS2 spectrum on everything within a range, it has high reproducibility and samples everything within the precursor range. However, the data is much more challenging to analyze because untargeted DIA MS2 spectra often contain fragments from multiple peptides (whatever precursors were in the window). Because of this challenge, it would be exceedingly difficult to identify crosslinks from untargeted DIA data without previous crosslink identifications, so we do not focus on DIA here, even though it has been used to quantify previously identified crosslinks.[17–19]

Almost all crosslink discovery experiments are performed using untargeted DDA.[20,21] Targeted DDA has been used with crosslinking mass spectrometry to increase the number of identifications but with minimal precursor lists, considering only up to 4 precursor values at a given retention time.[22] The targeted DIA method, PRM, has been used with previously identified crosslinks as input precursors to improve identifications and aid in quantification.[23–25] This application for quantification requires prior knowledge of the crosslinks to be quantified.

Recently, a program that creates exhaustive potential crosslink lists for SRM analysis (similar to PRM but analyzing a single fragment ion from a precursor)[26,27] based on databases of protein-protein interactions has been described.[28] The methods created by this program are informed by prior protein-protein interactions, not sample-specific LC-MS, meaning that previous DDA data on a specific sample is not needed to create the targeted SRM method.

Here, we aim to create similarly exhaustive targeted DDA lists based on the target protein sequence to force additional crosslink identifications and compare this targeted DDA method with corresponding untargeted DDA methods. We considered the effect of dynamic exclusion as well. Dynamic exclusion sets a defined amount of time (typically 20 seconds up to two minutes, depending on the gradient used) for a precursor not to be selected for fragmentation again after initial selection and fragmentation. Dynamic exclusion is commonly used for untargeted and targeted DDA.[15,29] Dynamic exclusion is meant to help identify more species by facilitating the identification of low-intensity species, and when enabled, it is known to reduce the total spectra count.[30] Because dynamic exclusion helps identify low-intensity species, DDA crosslinking studies often use dynamic exclusion and sometimes use longer exclusion windows to favor the MS2 analysis of crosslinks.[29] By comparing targeted and untargeted DDA with and without dynamic exclusion, we will determine if any change is due to the use of a targeted method or dynamic exclusion. Previous work has been inconsistent in its use of dynamic exclusion when comparing methods, making it unclear what change is causing any increases in identifications.[31] Here, we analyze samples using targeted and untargeted DDA methods with and without dynamic exclusion to determine the effect of dynamic exclusion on the number of crosslink identifications in both targeted and untargeted DDA methods.

4.2 Methods

4.2.1 Sample Preparation

The samples analyzed here use the non-canonical amino acid benzoylphenylalanine (BPA) as a crosslinker. BPA is incorporated into the protein sequence[32] and UV-treated to form a crosslink to any amino acid.[33] As described previously, BPA is incorporated into the human small heat shock protein (sHSP), HSPB5, crosslinked, and analyzed with LC-MS.[34–36] HSPB5 and other sHSPs form oligomers of varying sizes, which leads to the wide variety of crosslinks identified from a single-BPA site.[36] As shown in Figure 1 in this study, we analyzed dimeric products with an in-gel digestion (as described previously) to serve as a sample with enriched crosslinks. The dimeric products are the dimer band that occurs when analyzing a crosslinked reaction mixture with denatured SDS-PAGE. We also analyzed a reaction mixture, an in-solution digestion of the crosslink reaction mixture without any crosslink enrichment or SDS-PAGE analysis. Here, for dimeric products, we analyzed a W9B-variant digested with trypsin, an F61B-variant digested with trypsin, and an F61B-variant digested with a trypsin-GluC dual-enzyme digestion. For reaction mixtures, we analyzed a W9B-variant digested with trypsin, a W9B-variant digested with a trypsin-GluC dual-enzyme digest, and an F61B variant with a trypsin digest. The effect of a trypsin-only digest compared to a trypsin-GluC dual enzyme digest has been described previously.[34]

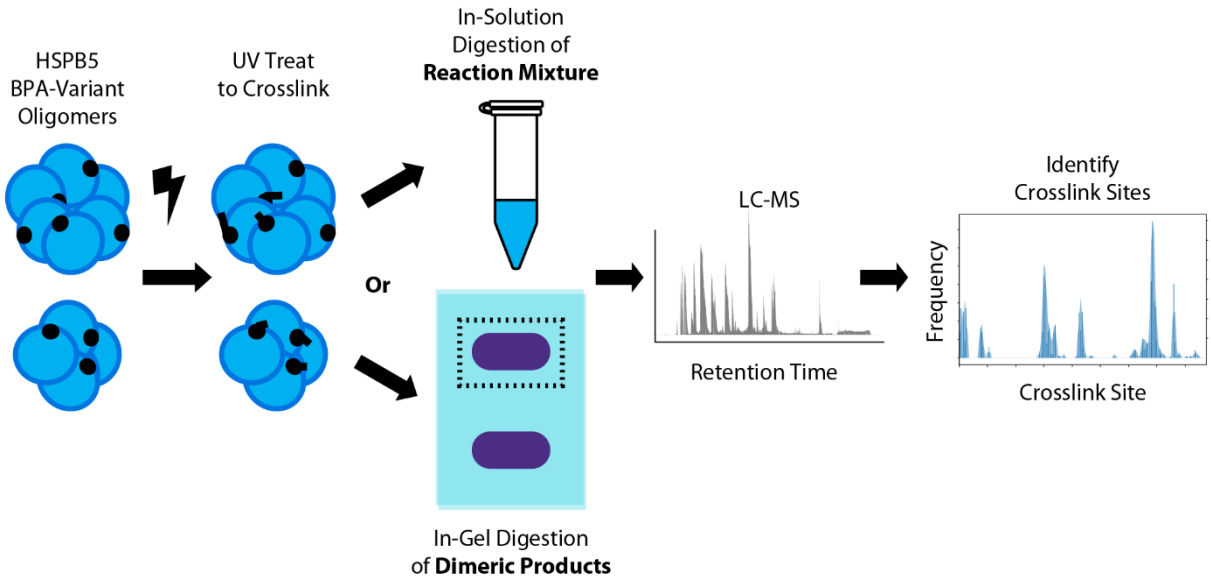

Figure 3. The sample preparation workflow is shown here. BPA variants of the sHSP HSPB5 are UV-treated to form crosslinks. The reaction mixture is analyzed with denatured SDS-PAGE for dimeric product samples, and the dimeric products (dimer band) are in-gel digested. The crosslinked reaction mixture is in-solution digested for reaction mixture samples without SDS-PAGE. Samples are analyzed with LC-MS, and bioinformatic analysis is used to identify the crosslinks present.

4.2.2 Inclusion List Creation

Potential crosslink precursor *m/z* values are calculated based on the masses of unlinked peptides according to the equation below:

$$Crosslink\ Precursor\ {}^{m}/_{z} = \frac{M_B + M_c + (z \times 1.008)}{z}$$

Where M_B is the neutral precursor mass of the BPA-containing peptide, M_c is the neutral precursor mass of the candidate crosslinked peptide, and z is the charge state. Charge states 3-6 were considered. We used two methods to generate a list of candidate peptides: those from non-

crosslinked peptides identified and validated from prior LC-MS data and those from non-crosslinked peptides expected based on the protein sequence and enzyme selection rules. Observed unlinked peptides are defined as peptides with at least one peptide spectral match (PSM) at a 1% PSM-level false-discovery rate (FDR) in a nonenzyme-specific Comet search. Expected peptides must be at least four residues long, have a mass between 500-5000 Da, and allow two missed cleavages for tryptic digests and three missed cleavages for trypsin-GluC dual enzyme digestions.

4.2.3 LC-MS

Untargeted and targeted DDA data were collected using an Easy Nano LC system coupled to a Thermo Orbitrap Fusion Lumos Tribrid. Untargeted DDA methods with 30-second dynamic exclusion were described previously.[34] All methods in this study used a 30-minute LC gradient. Targeted DDA methods used the corresponding precursor list as *m/z* values for analysis, and the option to fragment other species if nothing from the target list was present was enabled. Other Targeted DDA settings matched those previously described and used in the untargeted DDA methods. When not using dynamic exclusion, no dynamic exclusion parameter was defined.

To indicate which method we are using, we use the following notation, ${}^{\text{Exp}}_{(+)\text{DE}}\text{Targeted}$. In this notation, the subscript indicates if dynamic exclusion was used ((+)DE) or if it was not used ((-)DE). ${}^{\text{Exp}}_{(+)\text{DE}}\text{Targeted}$ means the targeted DDA method with the expected peptide inclusion list and dynamic exclusion was used. ${}^{\text{Obs}}_{(+)\text{DE}}\text{Targeted}$ means the PRM method with the observed peptide inclusion list and an inclusive analysis. The abbreviations section defines the notation format for each method.

4.2.3 Data Analysis

Data was analyzed using tools from the Trans-Proteomic-Pipeline,[37] as previously described.[34] Comet[38] was used to identify the proteins present and create validated-protein databases, and Kojak[39] was used for crosslink identification. PeptideProphet was used to validate search results at a 1% PSM level FDR.[40]

4.3 Results and Discussion

To date, BPA crosslinking has been used to identify residue-level HSPB5-HSPB5 crosslinks from multiple BPA sites and digestion conditions using $_{(+)DE}$Untargeted.[34] These residue-level interactions have characterized the selectivity of HSPB4/5[35] and long-range perturbations of HSPB5 accessibility.[36] However, relatively low numbers of PSMs (300 or fewer) are typically identified, which presents some challenges. Figure 2 illustrates crosslinks from $_{(+)DE}$Untargeted on W9B and F61B HSPB5 variants that were digested with Trypsin. The histograms are similar on the y-axis scale (15 and 13 for W9B and F61B, respectively) but differ substantially in the total number of crosslink PSMs represented (195 crosslink PSMs for W9B and 25 crosslink PSMs for F61B). Since there are fewer crosslink PSMs for F61B, it's difficult to determine if certain crosslinks are lower intensity because they are less frequent or because of the few numbers of PSMs detected. When there are fewer PSMs, the ability to detect more crosslink PSMs could significantly change the distribution. The difference in the total number of crosslink PSMs makes it challenging to compare the datasets. In addition, low numbers of PSMs make it especially difficult to generate validation models and interpret FDRs.

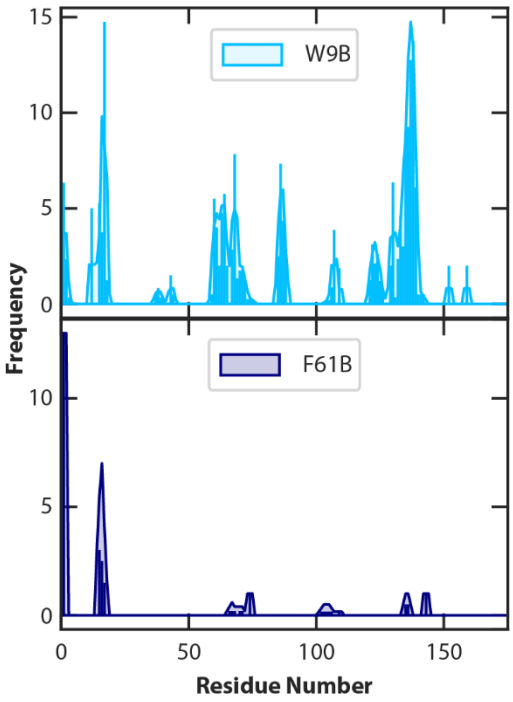

Figure 4. Crosslinks from trypsin-digested, dimeric products of W9B and F61B-HSPB5 using $_{(+)DE}$Untargeted are shown. The x-axis is the residue number of the site crosslinked by BPA, and the frequency axis is the number of crosslink PSMs identifying that crosslink site. W9B has 195 total crosslink PSMs, and F61B has 25 total crosslink PSMs. These datasets are also represented in Table 1.

Our goal was to determine if targeted DDA and altering dynamic exclusion could lead to the identification of more crosslink PSMs. For PRM studies, we used a comprehensive precursor list corresponding to all potential crosslinked products to determine if we could target MS2 acquisition towards the fragmentation of crosslinked products. To do this, we analyzed crosslinked BPA-variants of the sHSP, HSPB5, as previously reported for untargeted DDA methods.[34–36] We analyzed dimeric products (crosslink-enriched samples) and reaction mixtures (no crosslink enrichment) with different BPA positions and digestion enzymes (trypsin only and

trypsin-GluC). Analyzing these two different sample preparations helps us determine how sample complexity, in the form of how intense the crosslinks are, affects which method is most beneficial. All samples analyzed here are single-protein systems.

4.3.1 PRM Methods Used

The two methods for generating inclusion lists differ in what unlinked peptides they consider when calculating potential crosslinks and how the masses of the peptides were found. The observed peptides method uses unlinked peptides identified in previous $_{(+)DE}$Untargeted data, so it considers non-enzymatic peptides and uses experimentally determined precursor masses. The expected peptides method predicts peptides based on the protein sequence and enzyme selection rules, so it does not consider non-enzymatic peptides and uses precursor massesthat were calculated based on the peptide sequence. As shown in Figure 3A, when using a trypsin digestion, the observed peptides $_{(+)DE}$Untargeted method results in four to seven times more precursors than the expected peptides method. When using a trypsin-GluC digestion, the number of precursors in the two methods is much more similar. The W9B-HSPB5 variant has slightly more precursors in the observed peptides PRM method, whereas the F61B-HSPB5 variant has slightly more precursors in the expected peptides method. The number of precursors in the method ranges from 718 up to 7170, depending on the sample. The number of precursors we are considering is much larger than what is typical. Other targeted DDA experiments have used retention time scheduling to reduce the number of precursors considered at a given time.[22,41] Studies did this because the digital signal processor could only handle 500 values at a time.[14] However, advancements in instrument software have removed that limitation as evidences by the much longer inclusion lists that we consider here.

Figures 3B, 3C, 3D, and 3E compare the precursors in the observed and expected inclusion lists for different BPA sites and digestion conditions. Figures 3B and 3C illustrate that when using a tryptic digest on both W9B and F61B, there is a large overlap in the precursors in the inclusion lists even though the sizes of the inclusion lists differ (Figure 3A). Figures 3D and 3E illustrate that when using a trypsin-GluC dual enzyme digest on both W9B and F61B, about half of the precursors in the inclusion lists overlap. The potential effects of the length and identity of the inclusion list are explored in section 4.3.4: Effects of Method Used to Generate Inclusion Lists.

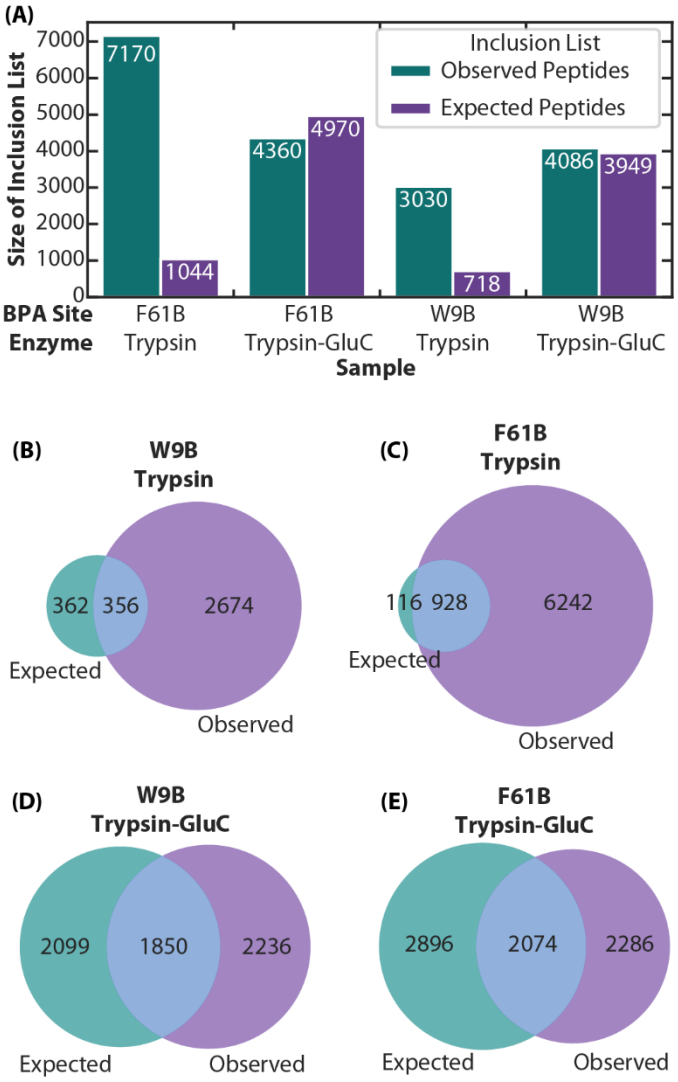

Figure 3. The number of precursors in the inclusion list for the observed and expected peptides targeted DDA methods are shown in Panel A. The number of precursors varies based on the site of BPA incorporation and the digestion enzyme. Panels B, C, D, and E show overlap in the identity of precursor *m/z* values in the expected and observed inclusion lists. Panel B shows results for W9B with a tryptic digestion. Panel C shows results for F61B with a tryptic digestion. Panel D shows results for W9B with a trypsin-GluC dual enzyme

digestion. Panel E shows results for F61B with a trypsin-GluC dual enzyme digestion.

4.3.2 Effects of PRM or DDA

Table 1 illustrates the number of HSPB5-HSPB5 crosslink PSMs identified across methods. When comparing $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, and $^{Obs}_{(+)DE}$Targeted, which method yields the most crosslink PSMs is inconsistent for reaction mixture and dimeric product samples. There is not a consistent difference between the number of crosslink PSMs found in comparing $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, or $^{Obs}_{(+)DE}$Targeted.

When comparing comparing $_{(-)DE}$Untargeted, $^{Exp}_{(-)DE}$Targeted, and $^{Obs}_{(-)DE}$Targeted on dimeric product samples, F61B has the most crosslink PSMs for $_{(-)DE}$Untargeted, and W9B has the most crosslink PSMs for $^{Obs}_{(-)DE}$Targeted. There is not a consistent difference between the number of crosslink PSMs found in $_{(-)DE}$Untargeted, $^{Exp}_{(-)DE}$Targeted, and $^{Obs}_{(-)DE}$Targeted for dimeric product samples.

When comparing $_{(-)DE}$Untargeted, $^{Exp}_{(-)DE}$Targeted, and $^{Obs}_{(-)DE}$Targeted on reaction mixture samples, all BPA variants, and digestion conditions have more crosslink PSMs for $^{Exp}_{(-)DE}$Targeted and $^{Obs}_{(-)DE}$Targeted than for $_{(-)DE}$Untargeted. The additional crosslink PSMs with $^{Exp}_{(-)DE}$Targeted and $^{Obs}_{(-)DE}$Targeted illustrate that PRM methods with exhaustive inclusion lists can facilitate the identification of additional crosslink PSMs.

4.3.3 Effects of Dynamic Exclusion

For the dimeric product samples with $_{(-)DE}$Untargeted, $^{Exp}_{(-)DE}$Targeted, and $^{Obs}_{(-)DE}$Targeted, we observed between a three- and 15-fold increase in the number of crosslink peptide spectral

matches (PSMs) identified relative to analyses with $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, and $^{Obs}_{(+)DE}$Targeted. For the reaction mixture samples, $_{(-)DE}$Untargeted, $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, and $^{Obs}_{(+)DE}$Targeted yield similar numbers of PSMs. However, for reaction mixture samples using $^{Exp}_{(-)DE}$Targeted and $^{Obs}_{(-)DE}$Targeted, we observed up to a five-fold increase in the number of PSMs identified relative to $_{(-)DE}$Untargeted, $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, and $^{Obs}_{(+)DE}$Targeted. The increase is particularly substantial for the trypsin-digested F61B reaction mixture. This sample yielded only 3 to 7 PSMs for $_{(-)DE}$Untargeted, $_{(+)DE}$Untargeted, $^{Exp}_{(+)DE}$Targeted, and $^{Obs}_{(+)DE}$Targeted and 21 to 25 PSMs for $^{Exp}_{(-)DE}$Targeted and $^{Obs}_{(-)DE}$Targeted. At 3 to 7 PSMs, there are not enough crosslinks to generate a distribution in a histogram, but at 21 to 25 PSMs, there are enough crosslinks to generate a distribution in a histogram. This increase from fewer than 10 to over 20 PSMs results in much more complete, usable data from the same sample.

There is a large increase in the number of crosslink PSMs identified in the dimeric product samples when using $_{(-)DE}$Untargeted. Still, no substantial difference exists in the number of PSMs identified in $_{(-)DE}$Untargeted or $_{(+)DE}$Untargeted in the reaction mixture samples. The only large increase in the number of crosslinks identified for the reaction mixture samples is for $^{Exp}_{(-)DE}$Targeted and $^{Obs}_{(-)DE}$Targeted. This illustrates that targeted DDA can be used to identify more crosslinks and is most useful when crosslinks are lower intensity. The fact that the number of crosslink PSMs identified for targeted DDA methods only increases when not using dynamic exclusion suggest that dynamic exclusion is limiting the number of identifications. Potential reasons for this are explored in future sections.

Sample Type	BPA Site	Enzyme	Untargeted DDA		Targeted DDA			
					Observed		Expected	
			(+)DE	(-)DE	(+)DE	(-)DE	(+)DE	(-)DE
Dimeric Products	F61B	Trypsin	25	397	60	269		
	F61B	Trypsin-GluC	132	482	67	210		
	W9B	Trypsin	195, 298*	618*	182*	696		479
Reaction Mixture	F61B	Trypsin	6, 7	5	4	23, 21	3	25
	W9B	Trypsin-GluC	57, 53	32	39	177, 43	26	139
	W9B	Trypsin	36, 31	40	48	101, 68	49	136

Table 1. The number of HSPB5-HSPB5 PSMs identified at a 1% FDR across methods and samples are indicated here. Two values represented separated by a comma represent replicate data. For the reaction mixture samples, replicate data is from analysis of the same crosslinking reaction sample on different dates. Samples were stored at -80ºC between replicates. For dimeric product samples, replicates come from different crosslink reactions that were prepared and analyzed on different dates. If two replicates are present, the replicate collected at an earlier date is listed first. For the dimeric product sample of Trypsin-digested W9B, the starred values represent samples that were crosslinked at different pH (6.5) than the other samples (7.5). We will repeat these measurements at pH 7.5 for consistency in future work.

4.3.4 Effects of Method Used to Generate Inclusion Lists

As described in Figure 3A, for trypsin-digested samples, observed peptide inclusion lists have many more precursor values than the expected peptide inclusion lists. For the dimeric

product W9B-variant that was trypsin digested, 217 more crosslink PSMs were found with $_{(-)DE}^{Obs}$Targeted (696 crosslink PSMs) than with $_{(-)DE}^{Exp}$Targeted (479 crosslink PSMs). Over 2000 more precursors are in the $_{(-)DE}^{Obs}$Targeted method for trypsin-digested W9B, so this suggests that having fewer precursors in the PRM list does not necessarily mean more crosslinks will be detected.

When looking at the reaction mixture samples, similar numbers of PSMs are found for both the $_{(-)DE}^{Obs}$Targeted and $_{(-)DE}^{Exp}$Targeted methods for the F61B trypsin digested sample (between 21 and 25 PSMs). For the W9B trypsin digested reaction mixture, 35 more PSMs are found with the $_{(-)DE}^{Exp}$Targeted (136 crosslink PSMs) than with the $_{(-)DE}^{Obs}$Targeted method (101 and 68 crosslink PSMs). This is the reverse of the dimeric product sample.

As shown in Figure 3A, for trypsin-GluC digested samples, the number of precursors in the observed and expected peptide inclusion lists is much more similar than with trypsin-digested samples. For the trypsin-GluC digested W9B reaction mixture, the number of crosslink PSMs identified with the $_{(-)DE}^{Exp}$Targeted method (139 crosslink PSMs) lies between the number of crosslink PSMs found with the two replicates of $_{(-)DE}^{Obs}$Targeted (177 and 42 crosslink PSMs). Overall, which method has the most PSMs appears to be sample-dependent.

Notably, there are cases where considering more precursor values (with the trypsin-digested W9B dimeric products) results in the identification of more crosslink PSMs. This illustrates the potential utility of large inclusion lists. Whether additional information can be gained from adding precursors would depend on what precursors are added. In our case of adding non-enzymatic peptides and peptides with experimentally obtained precursor masses in the observed peptides inclusion list, we sometimes identify more crosslink PSMs. Notably, an

enzyme constrained search is used to identify crosslinks, so additional crosslink identifications likely result from the consideration of experimental as opposed to predicted precursor masses. Which inclusion list is optimal ultimately depends on the goals of the study. The observed peptides inclusion list may be optimal if the goal is to identify as many crosslink PSMs as possible. If the goal is to achieve a higher throughput and minimize the number of LC-MS runs, the expected peptides inclusion list would be optimal as it does not require previous LC-MS data.

4.3.5 Reproducibility in Number of PSMs Identified

For the trypsin-GluC digested W9B-variant reaction mixture sample, there is a relatively large difference in the number of crosslinked PSMs identified for the two replicates of $_{(-)DE}^{Obs}$Targeted (43 and 177 crosslink PSMs). The two replicates are from analysis of the same sample on different dates, and the sample was stored at -80ºC between replicates. The two replicates of $_{(+)DE}$Untargeted are also from analysis of the same sample on different dates, but they have a much more similar number of PSMs (57 and 53 crosslink PSMs). The trypsin-digested W9B-variant reaction mixture was also stored at -80ºC between the two replicates of $_{(-)DE}^{Obs}$Targeted and has smaller difference in the number of PSMs (101 and 68 crosslink PSMs), but this difference is still much larger than the difference between the two replicates using $_{(+)DE}$Untargeted (36 and 31 crosslink PSMs). Due to the low number of replicates, we need more data to assess what could be the cause of this difference between the replicates. The reproducibility of dimeric products analyzed with $_{DE}$DDA has been described previously.[34] However, we have yet to assess the reproducibility of reaction mixture samples, so it could be that they do not give as reproducible results as the dimeric product samples (likely because of the lower intensity crosslinks). We need more data on the reproducibility of the identifications from

the targeted methods to determine whether the use of dynamic exclusion, targeted methods, or the sample type affects reproducibility.

4.3.6 Sensitivity and Error

Figure 4 shows the number of crosslinks to different protein types at different PSM-level FDRs. HSPB5-HSPB5 crosslinks represent target matches as HSPB5 is the predominant protein in the samples and contains the BPA residue. Decoy crosslinks contain a peptide from a decoy protein sequence, so they represent known false positives that PeptideProphet uses along with other factors to calculate cutoffs for FDR.[40] Other protein crosslinks contain a peptide to a protein that is not the target, HSPB5, or a decoy. These proteins are most often *E. coli* proteins. Because they increase in line with decoy matches, these other protein crosslinks are likely false positive matches for dimeric products.[34] Figure 4 illustrates that other protein crosslinks increase similarly to decoy crosslinks across methods used on dimeric product samples. Because other protein crosslinks increase similarly to decoys across methods used on dimeric product samples, for each method, other protein crosslinks correspond to false positives, indicating that each method is probing a similar population of crosslinks. When comparing methods with and without dynamic exclusion for dimeric products, using methods without dynamic exclusion shifts curves higher, and curves become slightly more boxlike. The curves shifting higher when not using dynamic exclusion is consistent with methods not using dynamic exclusion identifying more crosslinks. The curves becoming more box-like signify a better ability to validate results in that more true positives (HSPB5-HSPB5 crosslinks) are present while fewer false positives are present.

For the reaction mixture samples shown in Figure 4, the number of other protein crosslinks rises similarly to the number of decoys when dynamic exclusion is used. When not

using dynamic exclusion, the other protein crosslinks still increase across FDRs but are lower in number than the number of decoy crosslinks. The reaction mixture samples do not enrich HSPB5-HSPB5 interactions in the way that the dimeric product samples do, so it is noteworthy that the primary crosslink type detected is still HSPB5-HSPB5. This means that the primary interaction in the samples is HSPB5-HSPB5, whether it is enriched or not, and this corresponds well with the bands we see in the gel (HSPB5-HSPB5 dimer as the main inter-molecular crosslinked product) and little evidence for other proteins. Similarly to the dimeric products, we see more other protein crosslinks as FDR increases for the reaction mixture samples.

For the reaction mixture sample in Figure 4, the curves are about the same height for $_{(-)DE}$Untargeted and $_{(+)DE}$Untargeted, as expected, given that similar numbers of crosslink PSMs were identified for those samples. For the reaction mixture sample in Figure 4, the curves shift higher with $^{Obs}_{(-)DE}$Targeted and $^{Exp}_{(-)DE}$Targeted than with $^{Obs}_{(+)DE}$Targeted and $^{Exp}_{(+)DE}$Targeted. This reflects that more crosslink PSMs are identified using targeted methods without dynamic exclusion. Similarly, without dynamic exclusion for the dimeric product sample across methods, the curves shift higher to reflect the identification of more crosslink PSMs.

For the reaction mixture sample in Figure 4, the curves become slightly more box-like for $^{Obs}_{(-)DE}$Targeted and $^{Exp}_{(-)DE}$Targeted than for$_{(+)DE}^{Obs}$Targeted and $^{Exp}_{(+)DE}$Targeted and much more box-like for $_{(-)DE}$Untargeted than for $_{(+)DE}$Untargeted. More box-like plots indicate a better ability to validate results, so the plots are expected to be more box-like for $^{Obs}_{(-)DE}$Targeted and $^{Exp}_{(-)DE}$Targeted because we identify more crosslink PSMs than for$_{(+)DE}^{Obs}$Targeted and $^{Exp}_{(+)DE}$Targeted. We see a similar effect in the dimeric product samples. However, it is interesting that we see a more box-like plot for $_{(-)DE}$Untargeted for the reaction mixture sample because a

similar number of crosslink PSMs was identified at a 1% FDR. This suggests that validating results from $_{(-)DE}$Untargeted is easier, even though comparable numbers of crosslinks are identified. The intensity of crosslinks section explores potential reasons for this.

Overall, sensitivity error plots indicate that identifying additional crosslink PSMs for $^{Obs}_{(-)DE}$Targeted and $^{Exp}_{(-)DE}$Targeted is driven at least in part by an improved ability to validate crosslink results.

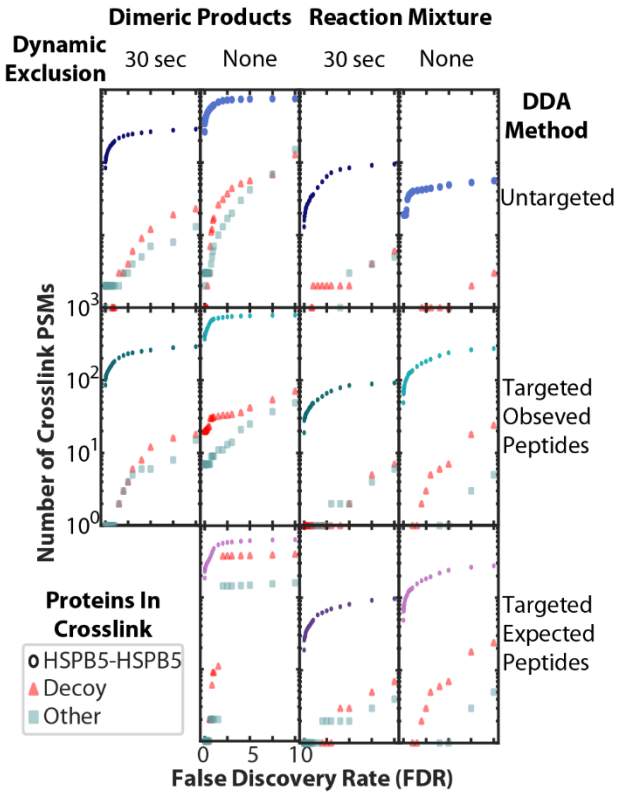

Figure 4. The number of HSPB5-HSPB5, decoy, and other protein crosslink PSMs across different FDRs is shown. The sample represented in these plots is W9B-HSPB5 digested with trypsin. The FDRs are PSM-level and indicated as percents, i.e., 5 is an FDR of 5%. Each plot shows the same x- and y-axis range.

4.3.7 Crosslinks Identified across Methods

Figure 5 shows the identity of the crosslinks found with different methods. For the dimeric product sample, across all methods, the crosslinks identified are similar with high-frequency crosslinks at residues, 137, 65, and 15. Some crosslink sites are only identified in some methods. The crosslinks around residue 40 are not identified in all samples. Residue 40 is in a large tryptic peptide, so crosslinks to that site are expected not to be reliably detected in trypsin-digested samples.[34] Crosslinks in sites 150-160 also vary more between samples, which is unsurprising given that they are relatively low-intensity crosslinks. The $_{(+)DE}$Untargeted for the dimeric products are similar in both the y-axis scale and crosslink sites identified, despite that they were crosslinked at different pH values. The reproducibility of $_{(+)DE}$Untargeted replicates that were crosslinked at the same pH has been established previously.[34]

For the dimeric product samples, when comparing results with and without dynamic exclusion, the crosslinks identified using dynamic exclusion match well with those identified without dynamic exclusion, but the scale of the y-axis is very different. Methods with dynamic exclusion have y-axis scales ranging 10-20 PSMs, but methods without dynamic exclusion have y-axis scales ranging 40-100 PSMs. The increased y-axis scale for methods without dynamic exclusion is helpful because it gives much more confidence in lower PSM hits. For example, a crosslink with 1 or 2 PSMs at a 1% FDR is much lower confidence than a crosslink with 10 PSMs at a 1% FDR. The increase in the y-axis scale helps to identify lower-intensity crosslinks confidently.

For the reaction mixture crosslinks in Figure 5, crosslink sites are similar across methods used with major crosslink sites at 15, 60, and 80, and crosslinks of varying intensity at site 137. The site 137 crosslink is of higher intensity in the dimeric product sample than in the reaction

mixture sample. This may be due to differences between samples. The dimeric product samples enrich inter-molecular interactions by excising an in-gel dimer. Because there is one BPA in each HSPB5, the dimer may have inter- and intra-molecular crosslinks, but at least half of the crosslinks will be inter-molecular. In contrast, the reaction mixture samples do not enrich dimeric products, so they may contain a much higher proportion of intra-molecular crosslinks because any crosslinks in a monomer band in a gel are not excluded. Most likely, the crosslink to site 137 is an inter-molecular interaction because of the difference in intensity between the two sample types. This corresponds well with the biology of HSPB5. The site 137 crosslink is an interaction between site nine and a binding pocket in HSPB5.[42]

The $_{(+)DE}$Untargeted replicates of the reaction mixture appear to differ more than the dimeric product replicates. The crosslink distributions of the two replicates do not overlap as clearly and are more jagged. However, the actual crosslink identifications are quite similar with the distributions not overlapping only in the region of the large tryptic peptide (20-40) and residues 150-160. These are the same regions with higher variability in the dimeric product samples. The $_{(+)DE}$Untargeted replicates look visually different than the same method on the dimeric products because the y-axis scale is much lower for the reaction mixture (a maximum of 4 instead of a maximum over 20 PSMs). The low number of PSMs results in more jagged crosslink distributions and illustrates the utility of identifying more crosslink PSMs.

When comparing the $_{(+)DE}$Untargeted and $_{(-)DE}$Untargeted on the reaction mixture sample, the total number of PSMs identified is similar (Table 1), but the crosslinks identified differ. With $_{(-)DE}$Untargeted, crosslinks are identified to sites 60 and 80 with higher frequency, but the crosslink at site 137 is not identified. With $_{(+)DE}$Untargeted, the crosslinks to sites 60 and 80 are identified with lower frequency, and the crosslink at site 137 is identified. This shows

the utility of using dynamic exclusion with untargeted DDA for low-intensity samples in that it facilitates the identification of a more comprehensive array of species. In contrast, $_{(+)DE}$Untargeted and $_{(-)DE}$Untargeted identified very similar crosslinks with different frequency values for the dimeric product sample that enriches crosslinked products.

Similar crosslinks are identified when comparing $_{(-)DE}^{Obs}$Targeted and $_{(-)\ DE}^{Exp}$Targeted with $_{(+)\ DE}^{Obs}$Targeted and $_{(+)\ DE}^{Exp}$Targeted on the reaction mixture sample. Crosslinks to sites 137 and 1 are identified with similar frequency in $_{(-)DE}^{Obs}$Targeted, $_{(-)\ DE}^{Exp}$Targeted, $_{(+)\ DE}^{Obs}$Targeted, and $_{(+)\ DE}^{Exp}$Targeted. Crosslinks at sites 20, 60, and 80 are identified with higher frequency in $_{(-)DE}^{Obs}$Targeted and $_{(-)\ DE}^{Exp}$Targeted. The higher frequency values create smoother distributions of crosslinks and enable comparisons to the dimeric product samples. Comparing samples with many fewer PSMs (reaction mixtures with $_{(+)\ DE}^{Obs}$Targeted and $_{(+)\ DE}^{Exp}$Targeted) to samples with more PSMs (dimeric products) is difficult because with low numbers of crosslink PSMs, there is less confidence in results and more potential for crosslink distributions to change in shape if more crosslinks are identified. The relative increase in crosslink PSMs at specific sites when using $_{(-)DE}^{Obs}$Targeted and $_{(-)\ DE}^{Exp}$Targeted on the reaction sample is an example of how distributions can change as more crosslink PSMs are identified.

Overall, for dimeric product samples, the crosslinks identified are very similar across methods, with the scale of the y-axis changing as more crosslink PSMs are identified. With the reaction mixture sample, the pattern of identified crosslinks changes slightly as more crosslink PSMs are identified, and a distribution can better be created. The increase in crosslink PSMs and the creation of a smoother, more complete distribution for the reaction mixture sample revealed the difference in relative intensity of the site 137 crosslink between the samples, which suggests

that it is an inter-molecular interaction. This is an example of the utility of identifying more crosslink PSMs through targeted DDA methods.

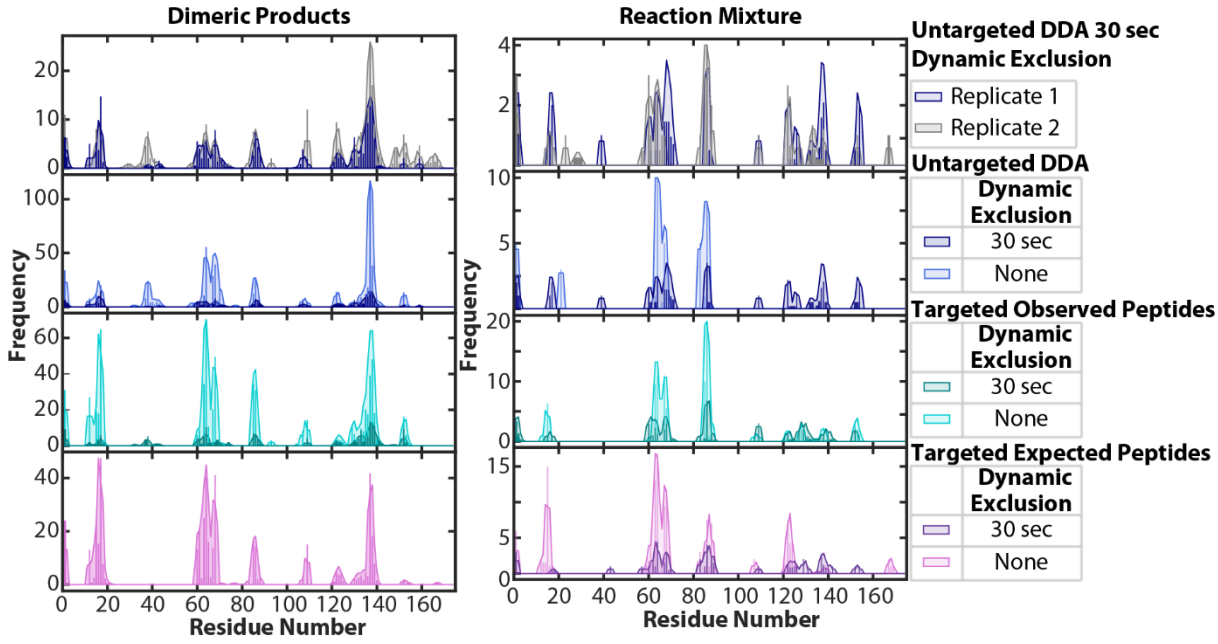

Figure 5. The crosslinks found across methods and samples are illustrated here. All plots here are from W9B-HSPB5 digested with trypsin. The x-axis is the residue within HSPB5 that was crosslinked by the BPA residue at site 9. The frequency axis is the number of crosslink PSMs with a crosslink site at that residue. The shaded windows are a rolling average of three across the x-axis. The dimeric products DDA replicates were crosslinked at different pH values. This affects some other dimeric product data shown here (See Table 1) as well and will be corrected with future work.

4.3.8 Number of Spectra and Precursors

Figure 6 illustrates how many searchable MS2 spectra (MS2 spectra with a candidate match in Kojak) correspond to unlinked and crosslinked peptides across methods for both dimeric products and reaction mixtures of trypsin-digested W9B. There are many more unlinked

than crosslinked MS2 spectra for dimeric products, and without dynamic exclusion, there are more MS2 spectra than if dynamic exclusion was used. In addition, when there are more total MS2 spectra, more MS2 spectra are identified at a 1% FDR. This suggests that the additional MS2 spectra correspond to additional PSMs, which could be due to having more MS2 spectra to create validation models.

For the reaction mixture samples and targeted DDA methods, there are similarly more unlinked and crosslinked total and identified spectra without dynamic exclusion. With the reaction mixture sample, there is an increase in the number of unlinked peptide MS2 spectra with $_{(-)DE}$Untargeted compared to $_{(+)DE}$Untargeted, but not the number of crosslinked peptide MS2 spectra. This is consistent with results from Table 1, showing that the number of crosslink PSMs from $_{(-)DE}$Untargeted and $_{(+)DE}$Untargeted on reaction mixture is similar.

Figure 6 demonstrates that $_{(-)DE}$Untargeted, $_{(-)DE}^{Exp}$Targeted, and $_{(-)DE}^{Obs}$Targeted identify similar numbers of unlinked MS2 spectra for dimeric product and reaction mixture samples. In addition, $_{(+)DE}$Untargeted, $_{(+)DE}^{Exp}$Targeted, and $_{(+)DE}^{Obs}$Targeted, identify similar numbers of unlinked MS2 spectra for dimeric product and reaction mixture samples. Because $_{(-)DE}$Untargeted, $_{(-)DE}^{Exp}$Targeted, and $_{(-)DE}^{Obs}$Targeted have an increase in the unlinked PSMs relative to $_{(+)DE}$Untargeted, $_{(+)DE}^{Exp}$Targeted, and $_{(+)DE}^{Obs}$Targeted, the increase most likely results from the lack of dynamic exclusion.

Interestingly, we detect many unlinked PSMs with the targeted DDA methods. The targeted DDA methods had the option to fragment something else if nothing from the inclusion list was present enabled. In Figure 6, the transparent bars represent the number of MS2 spectra in the sample, and the solid bars represent the number of those MS2 spectra with precursors in the

inclusion list. When using the observed peptides inclusion list with more precursor values (Figure 3A), about half to a third of the total unlinked and total crosslinked MS2 spectra have precursors in the inclusion list. Similar proportions of identified unlinked MS2 spectra are in the inclusion list. When using the expected peptides inclusion list with fewer precursor values (Figure 3A), fewer MS2 spectra have precursors in the inclusion list. That close to half or a third overlap with the larger precursor list and less overlap with the smaller precursor list suggest that it is random overlap of unlinked peptides with the PRM list causing some identifications of the unlinked peptide spectra. The option to fragment something else if no precursor from the inclusion list are present causes any unlinked peptide spectra identifications without precursors in the inclusion list. Nearly all the identified crosslinks are in the inclusion list (only a couple of spectra were not across all samples). This suggests that the inclusion list used are comprehensive of potential crosslinked products.

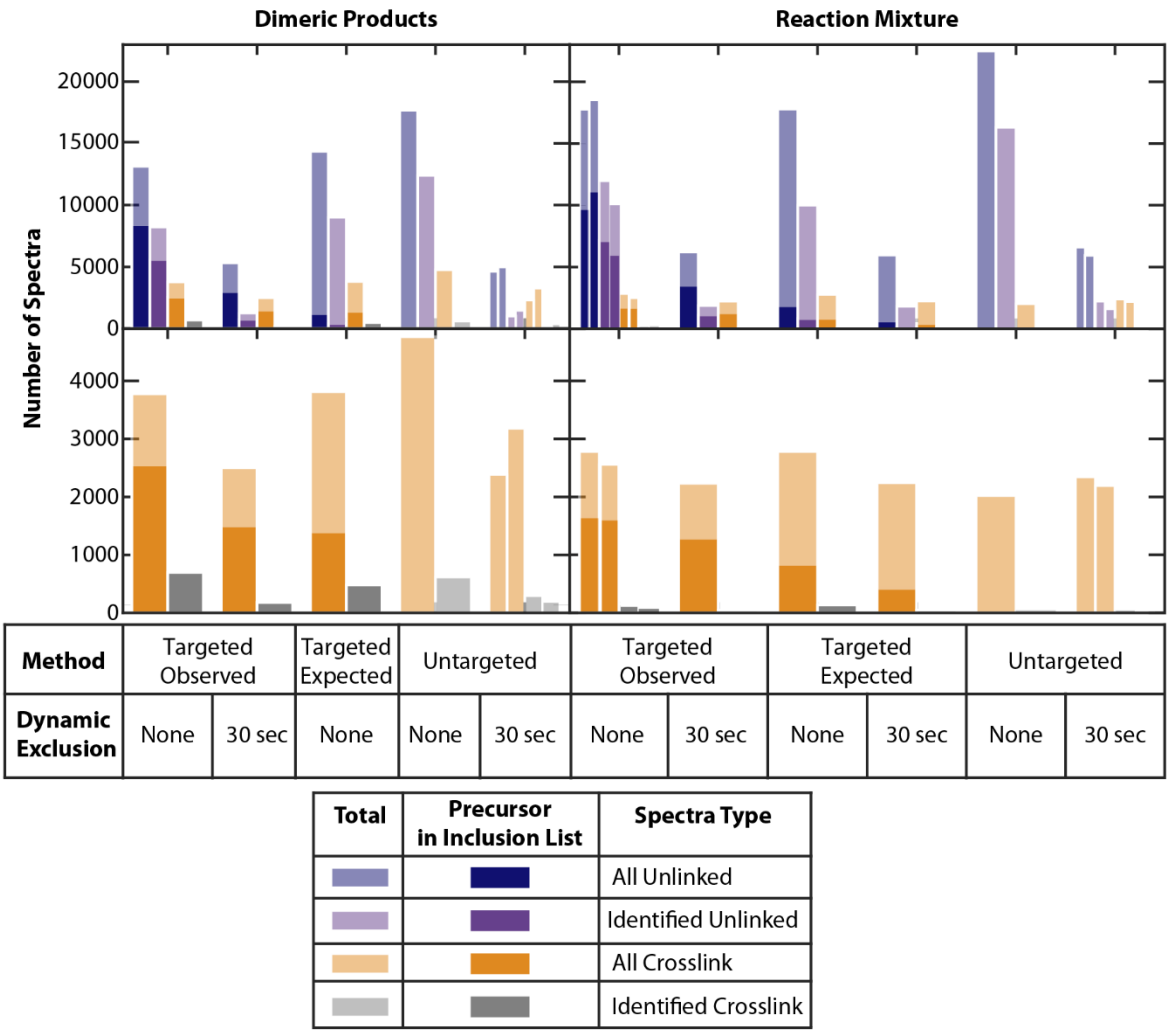

Figure 6. The number of MS2 spectra corresponding to unlinked and crosslinked peptides for trypsin-digested W9B dimeric products and reaction mixture across methods is indicated. The spectra type indicates what peptides correspond to the MS2 spectra. All unlinked peptide spectra correspond to spectra with unlinked peptide candidate matches in Kojak (of any confidence level) and identified unlinked peptide spectra meet the requirements for a 1% FDR. All crosslinked peptide spectra correspond to spectra with a crosslinked peptide candidate match in Kojak (any confidence level) and identified crosslink peptide spectra correspond to crosslinked spectra that meet the requirements for a 1% FDR.

Semi-transparent bars represent the total number in the sample with that method, and the solid bars represent the spectra with precursors in the inclusion list. Untargeted DDA only has transparent bars because there is no inclusion list. For the reaction mixture sample, see Table 1 for the number of identified crosslink PSMs for methods that have too low of a number to be visible in this plot.

Figure 7 establishes how many precursors from the PRM lists are detected. For the observed peptide inclusion list with 3030 precursors, a fraction of the precursors are identified across sample types. The maximum number of precursors identified is 742 of the 3030 (24%) for $_{(+)DE}^{Obs}$Targeted. This finding indicates that more precursors are identified for the dimeric product and reaction mixture sample when using $_{(+)DE}^{Obs}$Targeted compared to $_{(-)DE}^{Obs}$Targeted. For the expected peptide inclusion list, a fraction of the 718 precursors are identified with the maximum being 201 of the 718 precursors (28%) for $_{(-)DE}^{Exp}$Targeted. This suggests that more different precursors were detected with dynamic exclusion for the observed peptide inclusion list, which has more transitions, but for the expected peptide inclusion list, which has fewer precursors, more different precursors were detected without dynamic exclusion.

Across methods and samples as shown in Figure 7, many more precursor values are detected when all MS2 spectra are considered than when MS2 spectra identified at a 1% FDR are considered. For the observed peptide inclusion list (larger inclusion list), unlinked peptide spectra at a 1% FDR corresponds to about two to seven-fold more precursors than crosslinked peptide spectra at a 1% FDR. For the expected peptide inclusion list (smaller inclusion list), unlinked peptide spectra at a 1% FDR correspond to about the same or two-fold more precursors than crosslinked peptide spectra at a 1% FDR. When using the smaller inclusion list, identified

unlinked and crosslinked peptide spectra have much more similar numbers of precursors than when using a larger inclusion list. When using a larger inclusion list (observed peptides inclusion list), identified unlinked peptide spectra have many more unique precursors than identified crosslinked peptide spectra. This matches well with Figure 6, which illustrates that using a smaller precursor list (expected peptide inclusion list) results in fewer unlinked peptide MS2 spectra having precursors in the inclusion list. Using a smaller precursor list could help reduce the random overlap of unlinked peptide precursors with values in the inclusion list. However, as shown in Table 1, reducing the inclusion list would not necessarily result in more crosslink PSM identifications.

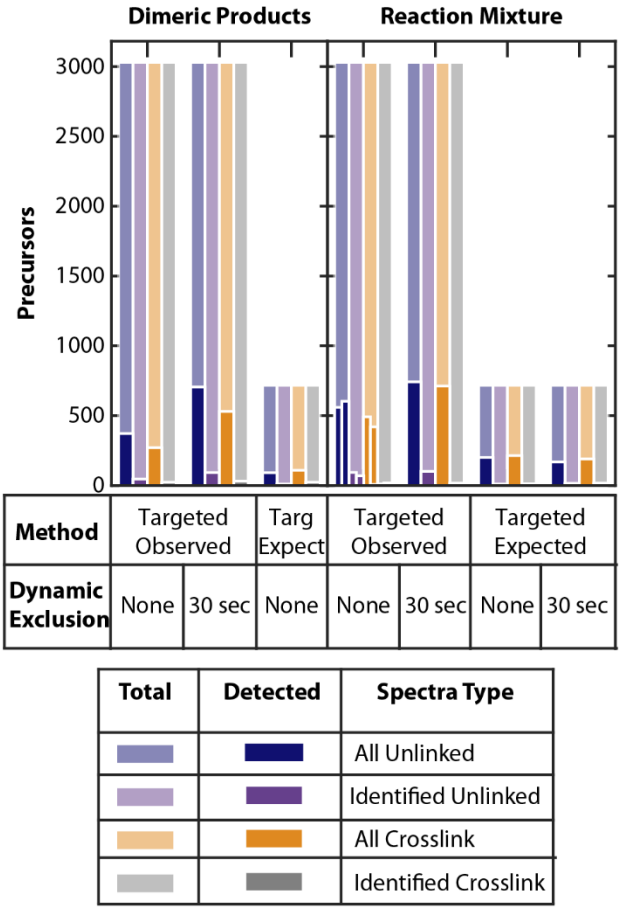

Figure 7. The number of precursors detected in dimeric products and reaction mixtures of trypsin-digested W9B is indicated. The transparent bars represent the total number of precursors, and the solid bars represent how many were detected. The spectra type indicates what peptides correspond to the MS2 spectra. All unlinked spectra correspond to spectra with unlinked peptide candidate matches in Kojak (of any confidence level) and identified unlinked peptide spectra meet the requirements for a 1% FDR. All crosslinked peptide spectra correspond to spectra with a crosslinked peptide candidate match in Kojak (any confidence level) and identified crosslink peptide spectra correspond to crosslinked spectra that meet the requirements for a 1% FDR.

4.3.9 Crosslink Precursor Intensity

Figure 8 shows distributions of the base-10 log of crosslink precursor intensity for dimeric products and reaction mixtures of W9B-HSPB5 digested with trypsin across methods. Crosslink precursor intensities from the same method on different sample types (dimeric products and crosslinked mixture) span similar x-axis ranges and have maxima at similar values. For methods with dynamic exclusion, the maximum of the distribution is at about a log value of 7. For methods without dynamic exclusion, the maximum of the distribution is at about a log value of 8. Disabling dynamic exclusion shifts intensities about a power of 10 higher, even for $_{(-)DE}$Untargeted in the reaction mixture sample when not many more crosslinks were identified.

For the reaction mixture, the shift is most apparent for untargeted data because the maximum of the peak gets narrower for $_{(-)DE}$Untargeted. Interestingly, we see a change in the precursor intensity for the reaction mixture analyzed with $_{(-)DE}$Untargeted, but we don't see much increase in the number of crosslink PSMs identified (Table 1). For targeted methods on the reaction mixture, when using $^{Obs}_{(-)DE}$Targeted or $^{Exp}_{(-)DE}$Targeted, peaks seem to plateau more—i.e., there are more crosslinks around the central intensity, so the most common intensity range increases some. Still, it is not nearly as pronounced as a change as with $_{(-)DE}$Untargeted and $_{(+)DE}$Untargeted on the reaction mixture.

For dimeric products, the count of higher-intensity crosslinks in methods without dynamic exclusion is much higher. The large change in the count values results from the large difference in the total number of crosslinked spectra (Figure 6). The large difference in the number of spectra makes the distributions appear quite different with and without dynamic exclusion. However, they seem to exhibit a similar pattern to the reaction mixture sample where when using $_{(-)DE}$Untargeted, the distribution of the maximum of the peak is narrower than when

using $_{(+)DE}$Untargeted, and when using $^{Obs}_{(-)DE}$Targeted, the distribution plateaus more than with using $^{Obs}_{(+)DE}$Targeted. Overall, crosslink precursor intensity seems to be similar between the sample types and depends heavily on whether dynamic exclusion was used. Since the maximum of the precursor intensity distribution increases when not using dynamic exclusion, we can likely sample higher-intensity precursors when not using dynamic exclusion. The first sampling is likely before the maximum precursor's intensity, so higher-intensity precursors are being ruled out of analysis with dynamic exclusion. This suggests that with high-speed instruments when analyzing low-intensity species, it may be more beneficial to not use dynamic exclusion to try and sample low-intensity species more completely than to try and exclude more intense things. However, we acknowledge that this is the case for a relatively simple sample of predominately a single protein. Other studies working with whole cell lysates have described needing a longer dynamic exclusion window to fragment low-intensity crosslinks.[29]

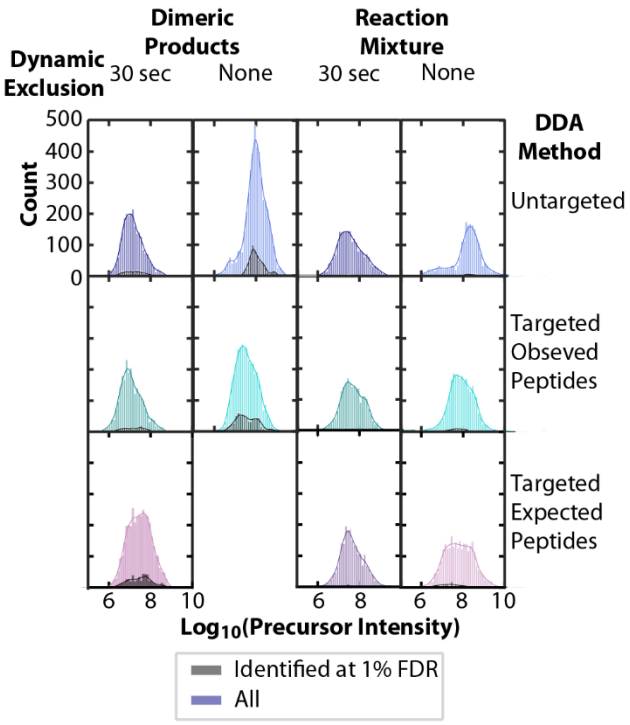

Figure 8. The distribution of crosslink precursor intensity across method types for dimeric products and a reaction mixture of W9B-HSPB5 digested with trypsin is shown. All spectra correspond to those with candidate crosslinked matches from Kojak. The precursors for crosslinks identified at a 1% FDR is shown in gray on top. For some samples, the distribution of identified crosslink precursors is too small in scale to be visible on this plot.

4.4 Conclusion

Here, we applied two targeted DDA methods that calculate all potential crosslinks in a sample based on either previously observed peptides or expected peptides. The two methods yield differing numbers of precursors (Figure 3A), and the precursor values they contain overlap (Figure 3B, 3C, 3D, and 3E). We compared these methods with untargeted DDA on dimeric product (crosslink enriched) and reaction mixture (non-crosslinked enriched) samples (Figure 1).

Not using dynamic exclusion resulted in more crosslink identification for the dimeric product sample for both targeted and untargeted methods. Both $_{(-)DE}^{Obs}$Targeted and $_{(-)DE}^{Exp}$Targeted resulted in more crosslink identifications for the reaction mixture sample relative to untargeted DDA methods (Table 1). The difference in the number of crosslink PSMs is likely partly due to the improved ability to validate crosslink results, as evidenced by more box-like sensitivity error plots (Figure 4). The crosslinks identified across methods are largely similar for dimeric product samples. The crosslink sites identified are mostly similar for reaction mixture samples, but crosslink distributions change some as more crosslink PSMs are identified when using targeted DDA without dynamic exclusion (Figure 5). This change in crosslink distribution reflects additional information gained from the additional crosslink PSMs identified. Figure 6 shows that for dimeric product samples, not using dynamic exclusion results in more unlinked and crosslinked spectra, suggesting that is a driving factor behind the increase in crosslink identifications. For the reaction mixture in Figure 6, there is a similar increase in the total number of spectra for $_{(-)DE}^{Obs}$Targeted and $_{(-)DE}^{Exp}$Targeted compared to $_{(+)DE}^{Obs}$Targeted and $_{(+)DE}^{Exp}$Targeted but not such a change for $_{(-)DE}$Untargeted and $_{(+)DE}$Untargeted. This further supports that targeted DDA, not just $_{(-)DE}$Untargeted, is needed for improved crosslink identification for the reaction mixture samples. Figure 7 shows that we detect a small subset of the precursor in the inclusion lists. Figure 8 illustrates that the maximum of the distribution of crosslinked precursor intensity increases when not using dynamic exclusion, suggesting not using dynamic exclusion results in the detection of higher intensity crosslinked precursors.

Here, we have shown that $_{(-)DE}^{Obs}$Targeted and $_{(-)DE}^{Exp}$Targeted , targeted DDA methods with comprehensive precursor lists of all potential crosslinks, can be used to facilitate the

identification of additional crosslink PSMs, and those additional crosslink PSMs can reveal new information about the sample. In addition, we demonstrated that in samples with enriched crosslinks, $_{(-)DE}$Untargeted results in a similar increase in the number of crosslink PSMs identified, illustrating that using dynamic exclusion is not always beneficial for the identification of crosslinks. Many of the precursors in the PRM lists were not detected (Figure 7), and for some samples, we detected more crosslinks PSMs when using methods with longer precursor lists (Table 1). Therefore, we have no evidence of adverse effects from including extraneous precursor values in the inclusion list, so applying this method to more complicated samples with more potential crosslinks is feasible. The work here was done in a relatively simple, single target-protein sample. Future work will extend this method to more complicated samples with multiple target proteins and higher background noise, such as a whole cell lysate.

4.5 Abbreviations

Crosslinking mass spectrometry, XL-MS

Liquid-chromatography mass spectrometry, LC-MS

Data-dependent acquisition, DDA

Parallel reaction monitoring, PRM

Data-independent acquisition, DIA

Benzoylphenylalanine, BPA

Small heat shock protein, sHSP

Peptide spectral match, PSM

False-discovery rate, FDR

Untargeted DDA with 30-second dynamic exclusion, $_{(+)DE}$Untargeted

Untargeted DDA with an inclusive analysis, $_{(-)DE}$Untargeted

Targeted DDA method with expected peptides inclusion list and 30-second dynamic exclusion, $^{Exp}_{(+)DE}$Targeted

Targeted DDA method with expected peptides inclusion list and no dynamic exclusion, $^{Exp}_{(-)DE}$Targeted

Targeted DDA method with observed peptides inclusion list and 30-second dynamic exclusion, $^{Obs}_{(+)DE}$Targeted

Targeted DDA method with observed peptides inclusion list and no dynamic exclusion, $^{Obs}_{(-)DE}$Targeted

4.6 Acknowledgements

I thank Lucas Narisawa, Christopher N. Woods, Natalie L, Stone, Maria Janowska, Rachel E. Klevit, and Matthew F. Bush for their contributions to this work. This material is based upon work supported by the National Eye Institute through R01EY017370 to R.E.K., the National Institute of General Medical Sciences through T32 GM008268 to C.N.W., the National Institute of Aging through T32 AG066574 to L.D.U., and the University of Washington's Proteomics Resource (UWPR95794).

4.7 References

(1) Klykov, O.; Steigenberger, B.; Pektaş, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990. https://doi.org/10.1038/s41596-018-0074-x.

(2) Yu, C.; Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **2018**, *90* (1), 144–165. https://doi.org/10.1021/acs.analchem.7b04431.

(3) Singh, P.; Panchaud, A.; Goodlett, D. Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Anal. Chem.* **2010**, *82* (7), 2636–2642. https://doi.org/10.1021/ac1000724.

(4) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; Chaignepain, S.; Chavez, J. D.; Claverol, S.; Cox, J.; Davis, T.; Degliesposti, G.; Dong, M.-Q.; Edinger, N.; Emanuelsson, C.; Gay, M.; Götze, M.; Gomes-Neto, F.; Gozzo, F. C.; Gutierrez, C.; Haupt, C.; Heck, A. J. R.; Herzog, F.; Huang, L.; Hoopmann, M. R.; Kalisman, N.; Klykov, O.; Kukačka, Z.; Liu, F.; MacCoss, M. J.; Mechtler, K.; Mesika, R.; Moritz, R. L.; Nagaraj, N.; Nesati, V.; Neves-Ferreira, A. G. C.; Ninnis, R.; Novák, P.; O'Reilly, F. J.; Pelzing, M.; Petrotchenko, E.; Piersimoni, L.; Plasencia, M.; Pukala, T.; Rand, K. D.; Rappsilber, J.; Reichmann, D.; Sailer, C.; Sarnowski, C. P.; Scheltema, R. A.; Schmidt, C.; Schriemer, D. C.; Shi, Y.; Skehel, J. M.; Slavin, M.; Sobott, F.; Solis-Mezarino, V.; Stephanowitz, H.; Stengel, F.; Stieger, C. E.; Trabjerg, E.; Trnka, M.; Vilaseca, M.; Viner, R.; Xiang, Y.; Yilmaz, S.; Zelter, A.; Ziemianowicz, D.; Leitner, A.; Sinz, A. First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961. https://doi.org/10.1021/acs.analchem.9b00658.

(5) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 742. https://doi.org/10.1038/s41467-020-14608-2.

(6) Chen, Z. A.; Rappsilber, J. Protein Dynamics in Solution by Quantitative Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43* (11), 908–920. https://doi.org/10.1016/j.tibs.2018.09.003.

(7) Vidova, V.; Spacil, Z. A Review on Mass Spectrometry-Based Quantitative Proteomics: Targeted and Data Independent Acquisition. *Anal. Chim. Acta* **2017**, *964*, 7–23. https://doi.org/10.1016/j.aca.2017.01.059.

(8) Hu, A.; Noble, W. S.; Wolf-Yadlin, A. Technical Advances in Proteomics: New Developments in Data-Independent Acquisition. *F1000Research* **2016**, *5*, 419. https://doi.org/10.12688/f1000research.7042.1.

(9) Guo, J.; Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **2020**, *92* (12), 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.

(10) Krasny, L.; Huang, P. H. Data-Independent Acquisition Mass Spectrometry (DIA-MS) for Proteomic Applications in Oncology. *Mol. Omics* **2021**, *17* (1), 29–42. https://doi.org/10.1039/D0MO00072H.

(11) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC−MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793. https://doi.org/10.1021/pr101060v.

(12) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol. Cell. Proteomics* **2012**, *11* (11), 1475–1488. https://doi.org/10.1074/mcp.O112.020131.

(13) Picotti, P.; Aebersold, R.; Domon, B. The Implications of Proteolytic Background for Shotgun Proteomics. *Mol. Cell. Proteomics* **2007**, *6* (9), 1589–1598. https://doi.org/10.1074/mcp.M700029-MCP200.

(14) Hoopmann, M. R.; Merrihew, G. E.; Von Haller, P. D.; MacCoss, M. J. Post Analysis Data Acquisition for the Iterative MS/MS Sampling of Proteomics Mixtures. *J. Proteome Res.* **2009**, *8* (4), 1870–1875. https://doi.org/10.1021/pr800828p.

(15) Schmidt, R.; Böhme, D.; Singer, D.; Frolov, A. Specific Tandem Mass Spectrometric Detection of AGE-Modified Arginine Residues in Peptides: Tandem Mass Spectrometry for AGE Detection. *J. Mass Spectrom.* **2015**, *50* (3), 613–624. https://doi.org/10.1002/jms.3569.

(16) Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. Automated Approach for Quantitative Analysis of Complex Peptide Mixtures from Tandem Mass Spectra. *Nat. Methods* **2004**, *1* (1), 39–45. https://doi.org/10.1038/nmeth705.

(17) Hao, Y.; Chen, M.; Huang, X.; Xu, H.; Wu, P.; Chen, S. 4D-diaXLMS: Proteome-Wide Four-Dimensional Data-Independent Acquisition Workflow for Cross-Linking Mass Spectrometry. *Anal. Chem.* **2023**, *95* (37), 14077–14085. https://doi.org/10.1021/acs.analchem.3c02824.

(18) Müller, F.; Kolbowski, L.; Bernhardt, O. M.; Reiter, L.; Rappsilber, J. Data-Independent Acquisition Improves Quantitative Cross-Linking Mass Spectrometry. *Mol. Cell. Proteomics* **2019**, *18* (4), 786–795. https://doi.org/10.1074/mcp.TIR118.001276.

(19) Müller, F.; Rappsilber, J. A Protocol for Studying Structural Dynamics of Proteins by Quantitative Crosslinking Mass Spectrometry and Data-Independent Acquisition. *J. Proteomics* **2020**, *218*, 103721. https://doi.org/10.1016/j.jprot.2020.103721.

(20) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184. https://doi.org/10.1038/nmeth.3603.

(21) Leitner, A.; Walzthoeni, T.; Aebersold, R. Lysine-Specific Chemical Cross-Linking of Protein Complexes and Identification of Cross-Linking Sites Using LC-MS/MS and the xQuest/xProphet Software Pipeline. *Nat. Protoc.* **2014**, *9* (1), 120–137. https://doi.org/10.1038/nprot.2013.168.

(22) Zelter, A.; Hoopmann, M. R.; Vernon, R.; Baker, D.; MacCoss, M. J.; Davis, T. N. Isotope Signatures Allow Identification of Chemically Cross-Linked Peptides by Mass Spectrometry: A Novel Method to Determine Interresidue Distances in Protein Structures through Cross-Linking. *J. Proteome Res.* **2010**, *9* (7), 3583–3589. https://doi.org/10.1021/pr1001115.

(23) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Bruce, J. E. In Vivo Conformational Dynamics of Hsp90 and Its Interactors. *Cell Chem. Biol.* **2016**, *23* (6), 716–726. https://doi.org/10.1016/j.chembiol.2016.05.012.

(24) Zhong, X.; Navare, A. T.; Chavez, J. D.; Eng, J. K.; Schweppe, D. K.; Bruce, J. E. Large-Scale and Targeted Quantitative Cross-Linking MS Using Isotope-Labeled Protein Interaction Reporter (PIR) Cross-Linkers. *J. Proteome Res.* **2017**, *16* (2), 720–727. https://doi.org/10.1021/acs.jproteome.6b00752.

(25) Yu, C.; Wang, X.; Huang, L. Developing a Targeted Quantitative Strategy for Sulfoxide-Containing MS-Cleavable Cross-Linked Peptides to Probe Conformational Dynamics of Protein Complexes. *Anal. Chem.* **2022**, *94* (10), 4390–4398. https://doi.org/10.1021/acs.analchem.1c05298.

(26) Picotti, P.; Bodenmiller, B.; Aebersold, R. Proteomics Meets the Scientific Method. *Nat. Methods* **2013**, *10* (1), 24–27. https://doi.org/10.1038/nmeth.2291.

(27) Murray, K. K.; Boyd, R. K.; Eberlin, M. N.; Langley, G. J.; Li, L.; Naito, Y. Definitions of Terms Relating to Mass Spectrometry (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (7), 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06.

(28) Park, J.; Son, A.; Kim, H. A Protein–Protein Interaction Analysis Tool for Targeted Cross-Linking Mass Spectrometry. *Sci. Rep.* **2023**, *13* (1), 22103. https://doi.org/10.1038/s41598-023-49663-4.

(29) Arlt, C.; Götze, M.; Ihling, C. H.; Hage, C.; Schäfer, M.; Sinz, A. Integrated Workflow for Structural Proteomics Studies Based on Cross-Linking/Mass Spectrometry with an MS/MS Cleavable Cross-Linker. *Anal. Chem.* **2016**, *88* (16), 7930–7937. https://doi.org/10.1021/acs.analchem.5b04853.

(30) Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L. Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics. *Anal. Chem.* **2009**, *81* (15), 6317–6326. https://doi.org/10.1021/ac9004887.

(31) Plank, M. J. Modern Data Acquisition Approaches in Proteomics Based on Dynamic Instrument Control. *J. Proteome Res.* **2022**, *21* (5), 1209–1217. https://doi.org/10.1021/acs.jproteome.2c00096.

(32) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proc. Natl. Acad. Sci.* **2002**, *99* (17), 11020–11024. https://doi.org/10.1073/pnas.172226299.

(33) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(34) Ulmer, L.; Canzani, D.; Woods, C.; Stone, N.; Janowska, M.; Klevit, R.; Bush, M. High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid. *J. Proteome Res.* **2024**. https://doi.org/10.1021/acs.jproteome.4c00194.

(35) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(36) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

(37) Deutsch, E. W.; Mendoza, L.; Shteynberg, D. D.; Hoopmann, M. R.; Sun, Z.; Eng, J. K.; Moritz, R. L. Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **2023**, *22* (2), 615–624. https://doi.org/10.1021/acs.jproteome.2c00624.

(38) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(39) Hoopmann, M. R.; Shteynberg, D. D.; Zelter, A.; Riffle, M.; Lyon, A. S.; Agard, D. A.; Luan, Q.; Nolen, B. J.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Improved Analysis of Cross-Linking Mass Spectrometry Data with Kojak 2.0, Advanced by Integration into the Trans-Proteomic Pipeline. *J. Proteome Res.* **2023**, *22* (2), 647–655. https://doi.org/10.1021/acs.jproteome.2c00670.

(40) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392. https://doi.org/10.1021/ac025747h.

(41) Schmidt, A.; Gehlenborg, N.; Bodenmiller, B.; Mueller, L. N.; Campbell, D.; Mueller, M.; Aebersold, R.; Domon, B. An Integrated, Directed Mass Spectrometric Approach for In-Depth Characterization of Complex Peptide Mixtures. *Mol. Cell. Proteomics* **2008**, *7* (11), 2138–2150. https://doi.org/10.1074/mcp.M700498-MCP200.

(42) Klevit, R. E. Peeking from behind the Veil of Enigma: Emerging Insights on Small Heat Shock Protein Structure and Function. *Cell Stress Chaperones* **2020**, *25* (4), 573–580. https://doi.org/10.1007/s12192-020-01092-2.

Chapter 5. Using SILAC to Identify Crosslinked Peptides through Quantifying Depletion of Unlinked Peptides

5.1 Introduction

Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) is an isotopic labeling procedure in which cells are grown in media containing isotopically labeled amino acids. As the cells grow in the heavy media with the isotopically labeled amino acids, they incorporate the isotopically labeled amino acids into their proteins, resulting in proteome-wide heavy labeling.[1] Typically, in SILAC experiments, one group of cells is grown with light media without the isotopically labeled amino acids, and one is grown in heavy media with the isotopically labeled amino acids. The heavy and light cells are combined, lysed, and digested. Therefore, the ratio of heavy to light amino acids in the pooled sample can be used to quantify differences in protein abundance between the heavy and light cells.[2,3] SILAC experiments are especially useful for comparing two groups, such as drug-treated and non-drug-treated.[4,5] SILAC quantification is compatible with many different experiments, including crosslinking mass spectrometry.[6]

Most SILAC experiments use mammalian cell culture. Mammalian cells need amino acids supplied in their media for healthy growth.[7] Most SILAC experiments proteolytically digest proteins using trypsin, which usually cleaves on the C-terminal side of arginine and lysine. Therefore, most SILAC experiments use heavy arginine and lysine because most proteolytic peptides will include at least one heavy amino acid. In mammalian cell culture, arginine can be converted to proline, which can interfere with the specificity of the heavy label incorporation. However, using specific concentrations of arginine and proline can minimize the conversion so that it is not an issue.[8]

Applying the SILAC strategy to bacterial cell culture is much more challenging than for mammalian cell culture because bacteria are better able to biosynthesize and catabolize amino acids.[9] Incorporating heavy lysine in bacteria has been accomplished by growing cells in minimal media and supplying higher concentrations of heavy lysine than other amino acids.[9,10] However, labeling arginine and lysine is more challenging due to arginine-to-proline conversion. This difficulty in arginine incorporation has been overcome by optimizing the media used and by creating new bacteria strains. To optimize media to incorporate heavy arginine in *E. coli,* high concentrations of amino acids are used in the cell culture.[11] To incorporate heavy amino acids by creating new strains, genes that allow *E.coli* to biosynthesize lysine (*LysA*) and arginine (*ArgA*) are knocked out.[12] The resulting strain has been used for protein expression to generate isotopically labeled proteins.[13] Knocking out genes instead of optimizing media allows for the use of four-fold less heavy arginine and eight-fold less heavy lysine, which significantly reduces the cost of expressing the protein.

Figure 1 shows our proposed strategy for quantifying differences in abundance within a single sample. The strategy described in Figure 1 differs from typical SILAC and DIA quantification approaches in several ways. First, it will require high-sequence coverage of the target protein in terms of the number of peptides quantified. Typically, a few well-responding peptides are selected per protein for quantification.[14,15] The strategy presented in Figure 1 will also require quantifying substoichiometric changes in concentration because the amount of the BPA-containing peptide limits the degree of change. Studies using DIA to quantify SILAC data typically measure the degree of change as the degree of 2-fold change.[16] Here, we describe sample preparation, data collection, and data analysis methods that make the approach described

in Figure 1 the most achievable by attempting to gain maximum sequence coverage of quantified peptides while preserving the ability to detect small degree changes.

Here, we use an *E. coli* cell line that was designed for SILAC incorporation[13] to generate heavy-labeled HSPB5 for use as a standard in crosslinking experiments. The crosslinked samples analyzed here use the non-canonical amino acid benzoylphenylalanine (BPA) as a crosslinker. BPA is incorporated into the protein sequence[17] and UV-treated to form a crosslink to any amino acid.[18] As described in Figure 1, the goal was to identify crosslinked peptides through quantifying the depletion of unlinked peptides. Because every crosslink contains the BPA peptide, each crosslink formed depletes one unlinked peptide. We propose to use the heavy-to-light ratio to infer the depletion of unlinked peptides that is concomitant with the participation of that peptide in a crosslink. Peptides with more heavy than light peptides are depleted when crosslinks are formed and represent crosslinked peptides.

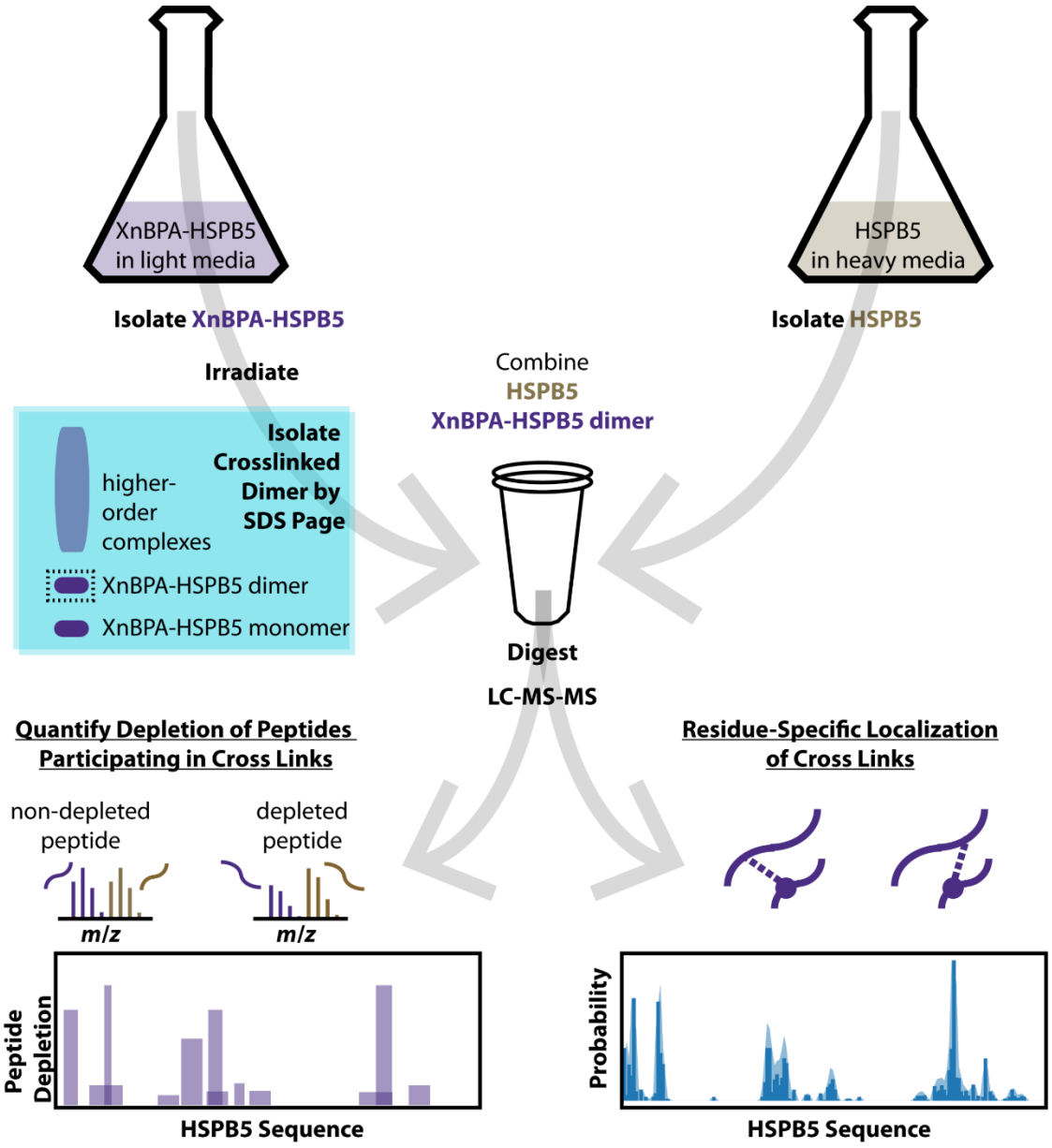

Figure 5. Previous BPA crosslinking experiments used an in-gel digestion of dimeric products as illustrated here. For SILAC experiments, a SILAC labeled WT (non-BPA-containing) variant of the same target protein (HSPB5 in this case) will be combined with the dimeric products before LC-MS analysis. From the analysis of this sample, crosslinks can be identified with residue-level specificity as previously reported,[19–21] or unlinked peptides can be quantified. Unlinked

peptides with more heavy area than light area will indicate crosslinked peptides, and differences in how much heavy area there is will allow for the quantification of crosslinked peptides.

5.2 Methods

5.2.1 Sample Preparation

SILAC HSPB5 was generated using the previously described cell line,[13] $^{13}C_6$ and $^{15}N_4$ arginine (Cambridge Isotopes #CNLM-539-H), and $^{13}C_6$ and $^{15}N_2$ lysine (Cambridge Isotopes #CNLM-291-H). Protein was expressed and purified as previously described.[20,21] BPA crosslinking and preparation of dimeric products (dimers of crosslinked reaction mixtures in denatured SDS-PAGE) was performed as previously described.[19] SILAC protein was added in a near equal molar amount to the BPA-containing protein either in solution or in gel. For SILAC protein added in solution, SILAC protein was added in the solution phase to the gel band prior to digestion. For SILAC protein added in-gel, if the sample was not crosslinked, the SILAC protein was mixed with the BPA prior to running the gel. For SILAC protein added in-gel, if the sample was crosslinked, SILAC protein was run on a separate gel lane than the crosslinked BPA sample and an about equal area of the SILAC monomer band was combined with the crosslinked dimer band when the bands were excised prior to digestion.

5.2.2 LC-MS

DDA LC-MS data was collected using an Thermo Orbitrap Fusion Lumos Tribrid and a 30 min LC gradient as described previously.[19] All DIA methods used the same LC gradient and instrument as DDA methods, and a few different DIA methods were used. Different DIA methods were used as described in the relevant figure captions.

5.2.3 Data Analysis

DDA data was analyzed using tools from the Trans-Proteomic Pipeline.[22] Comet[23,24] was used to search DDA data against a database containing the BL21 *E. coli* database from UniProt (UP000431028), the cRAP database from the Global Proteome Machine with all 5 levels of proteins,[25] the pertinent BPA-containing variant of HSPB5, WT-HSPB5, peptides used for quality control (AngioNeuro), and reverse-sequence decoys. Comet results were validated using PeptideProphet,[26] and Skyline was used to create a spectral library from the PeptideProphet-validated results.[27] The resulting spectral library was used to analyze DIA data in Skyline. All spectral libraries were selected for use when analyzing DIA data, but the spectral library from the DDA data collected from the same sample was given top priority. Light and heavy peak areas were exported from Skyline, and an in-house-developed interactive Python notebook processed these files to create the plots shown.

5.3 Results and Discussion

5.3.1 Heavy Amino Acid Incorporation

To test the incorporation of the heavy amino acids, we used DDA data with both a 30- and 85-minute gradient to analyze a sample that contained only SILAC-labeled HSPB5. Figure 2A shows the MS1 spectrum corresponding to a spectral match for the HSPB5 peptide HFSPEELK. Figure 2A and Figure 2B show that the heavy precursors are about 80 times more intense than the light precursors. Figure 2C shows the chromatogram traces for the heavy and light versions of the HSPB5 peptide HFSPEELK. Panel D shows the chromatographic peak area for the heavy and light versions of the same peptide when using both a 30 and 85-minute gradient. The peak area for the heavy peptide is much higher than that for the light peptide. This analysis workflow of detecting peaks corresponding to the heavy and light versions of a peptide,

mapping the chromatogram traces for heavy and light versions of the peptide, and calculating the heavy and light peak areas was repeated for the analysis of DIA data throughout the work presented here. When extending the analysis illustrated in Figure 2 across detected HSPB5 peptides and summing the light and heavy areas, HSPB5 is over 99% heavy labeled. The incorporation efficiency of over 99% is consistent with prior reports using this cell line to generate SILAC-labeled protein.[13]

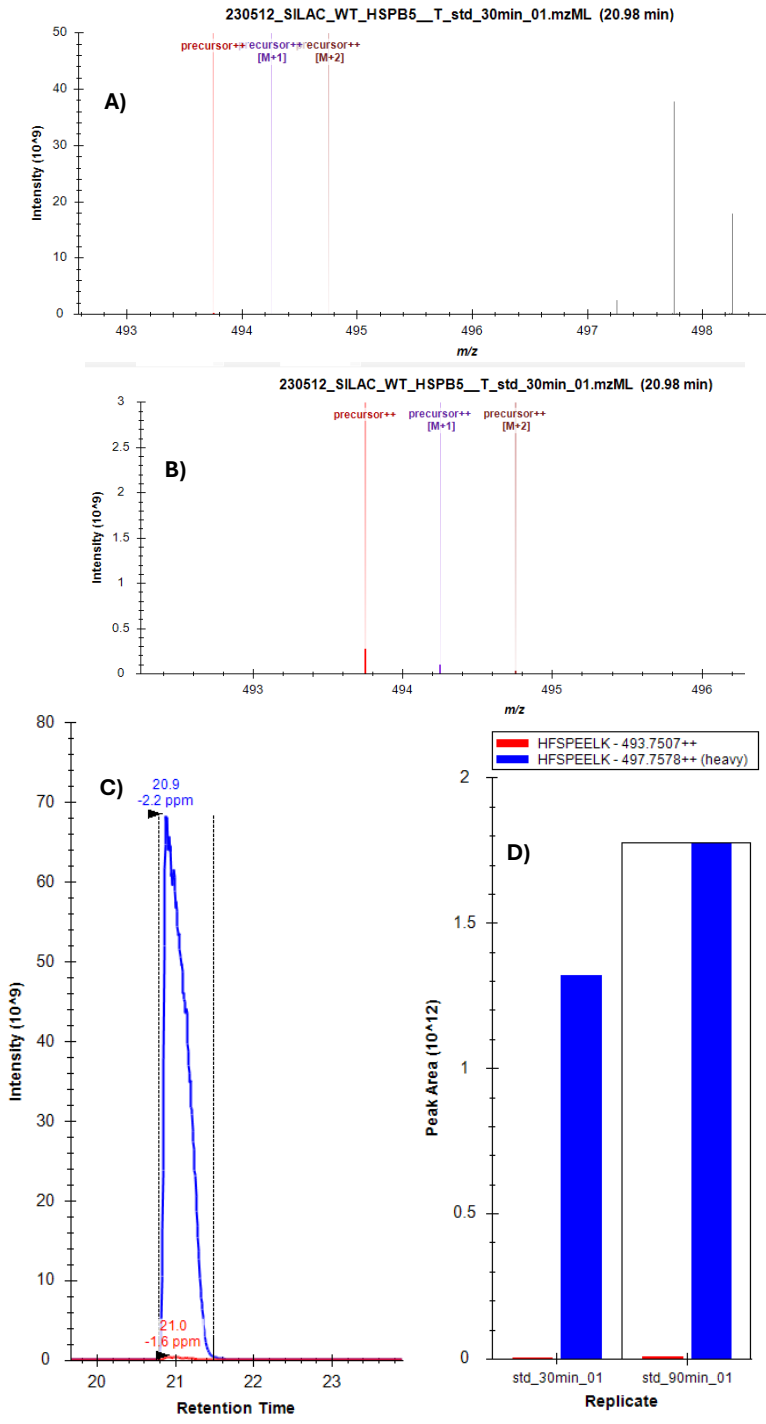

Figure 6. Examples from the analysis of a sample containing only WT-HSPB5 SILAC are shown here. DDA data is represented here. Panel A shows the precursor mass spectrum. The colored bars highlight where light precursors are. The peaks around 498 *m/z* are the heavy precursors. The intensity for the heavy precursors reaches about $40x10^9$. Panel B is a zoomed-in version of Panel A to show the light precursors better. The Light precursor intensity reaches about $0.5x10^9$. Panel C shows the chromatogram trace for both the heavy (blue) and light (red) versions of the HSPB5 peptide HFSPEELK. Panel D shows the chromatographic peak area for the light (red) and heavy (blue) versions of the same peptide from DDA data with a 30-

minute gradient (left) and an 85-minute gradient (right). The plots shown here were generated using Skyline.

5.3.2 Effect of Peptides Considered for Quantification

Figure 3 shows results from a pooled sample containing SILAC WT-HSPB5 and W9B-HSPB5, which was not exposed to UV light and should not contain crosslinks. The heavy and light peak areas for each transition (each ion corresponding to a given peptide) were exported from Skyline, and the $\log_{10}$ of the heavy area and $\log_{10}$ of the light area for each transition were plotted as scatter plots in Panels A, D, and G. Lines of best fit and R^2 values are included as well. A linear response in this plot with few outliers would represent peptides with similar heavy/light ratios with little to no detectable change in abundance. A linear response would be expected for samples without crosslinks, such as those in Figure 3. Peptides containing the position of the BPA residue are excluded from these plots because there is no SILAC-labeled counterpart to the BPA peptides. In Panels B, E, and H, the residuals of the log of the light area are indicated. These residuals were calculated by subtracting the value predicted by the line of best fit from the experimental value of the log of the light area. Panels C, F, and I show distributions of the residuals in plots B, E, and F.

All peptides detected include all HSPB5 peptides that meet the 1% FDR requirement. This included 208 total peptides (excluding BPA-containing peptides) and 1240 transitions. When considering all peptides, the sequence coverage of HSPB5 was 99%. When considering all peptides detected for analysis (Figure 3 Panels A, B, and C), the log of the heavy area covers a vast range from below –30 to above 10, and the R^2 of 0.2 is quite low, illustrating that the linear model does not represent the data well. The low heavy values suggest we are considering low-

intensity peptides that are more difficult to detect reliably, which could result in a lower R^2. When excluding the low-area transitions and retaining the higher-area transitions around a log of heavy area of 0-10, there is still a visible spread in the data, especially at lower light areas. The residuals in Figure 2B have a similar shape to the plot in Figure 2A, further demonstrating how poorly the linear model represents this data. Figure 2C shows the distribution of the residuals in panel 2B, demonstrating that they range from -5 to 5. Overall, when considering all peptides for analysis, there are low-area outliers, and a linear model does not represent the data well. Since this sample does not contain crosslinks, the outliers and lack of linearity indicate variance in the measurements that interferes with the quantification, likely due to the analysis of non-quantifiable peptides.

The peptides considered for analysis were then limited to a list called “good peptides”. The good peptides were selected based on the chromatogram traces. Peptides that did not have chromatogram traces with similar shapes for heavy and light traces or that had interference (overlapping peaks from other peptides or transitions) were removed. When creating this list, the sequence coverage was considered, so some peptides with worse chromatogram fittings were included to maintain higher sequence coverage. The good peptides list contains 72 peptides and 218 transitions. When considering the good peptides, the sequence coverage of HSPB5 was 98%.

Figure 3D shows the scatter plot and line of best fit for the log of the heavy peak area vs the log of the light peak area when considering the good peptides for analysis. The minimum of the log of the heavy is much higher than when considering all peptides (a minimum of 4 instead of below –30), illustrating that the low area outliers from Figure 3A are no longer analyzed in Figure 3D. In Figure 3D, most points lie near the line of best fit with some outliers both above and below the line of best fit in the middle of the x-axis range. In Figure 3D, the R^2 is 0.9, which

is much higher than the value of 0.2 obtained when considering all peptides for analysis. The higher R^2 indicates that the linear model better represents the data. The linear plot represents a near-constant heavy/light area ratio as is expected for samples without crosslinks. Figure 3E shows the residuals of the log of the light area in Figure 3D. Most residuals are centered around 0, with some higher and lower across the middle portion of the x-axis range. This starkly contrasts Figure 3B, which had two clusters of residual values. Figure 3F shows the distribution of residual values in Figure 3E. The residual values range from just below –2 to +2. This is a much narrower range of residuals than in Figure 3C, where all peptides are considered for analysis, which had a residual range of –5 to +5.

Overall, restricting the peptides considered for analysis from all peptides to good peptides based on the chromatogram fittings removes low-area outliers from the analysis, improves the fit of the linear model, and decreases the magnitude of the residuals. This suggests that the response of heavy to the light area is more linear, and the heavy-to-light area ratio is more constant, as would be expected from samples without crosslinks. However, there are outliers in Figure 3D that lie outside the line of best fit. Because we aim to detect sub-molar changes in peptide abundance, our approach must be sensitive to small changes in areas for the heavy peptides. The outliers, such as those in Figure 3D, may interfere with detecting changes in abundance due to crosslink formation because the degree of change may be similar to that with crosslink formation. Therefore, we further limited the peptides considered in our analysis. We used data with varying amounts of added SILAC protein to further limit the peptides analyzed. These samples were prepared by adding SILAC protein in-solution to monomeric reactant BPA (non-crosslinked sample) with increasing concentrations of added SILAC protein. Only peptides with an increasing proportion of heavy area when more SILAC protein was added, were included in

this list of "responding peptides". Sequence coverage was not taken into consideration when selecting these peptides. If the proportion of heavy area of a peptide did not respond to an increasing amount of SILAC protein added, it was not included in the responding peptides list. The responding peptide list contained 47 peptides and 121 transitions. When considering the responding peptides, the sequence coverage of HSPB5 was 82%. This is much lower than the sequence coverages of 98 and 99% obtained when using good peptides and all peptides for analysis. The most significant gap in sequence coverage when using the responding peptides list is from residues 22 to 40, which overlaps with the large tryptic HSPB5 peptide from residues 23 to 56.

Figure 3G shows the scatter plot and line of best fit for the log of the heavy peak area vs the log of the light peak area when considering the responding peptides for analysis. The minimum of the log of the heavy area is 6, which is higher than the analysis with the good peptides list (Figure 3D). In Figure 3G, most points appear to lie around the line of best fit, with a couple of points that appear outside at lower area values. The R^2 value is 0.98 when using the responding peptides for analysis, which is higher than the R^2 when using either the good peptides or all peptides for analysis. The lack of visual outliers and higher R^2 suggests that the linear model best fits the data when using the responding peptides for analysis. Figure 3H shows the residuals for Figure 3G. Most of the residuals are centered around zero, with the residuals appearing to have more spread at heavy area values. The residuals of –1.25 correspond to the visual outliers in Figure 3G. The difference in residual spread at lower values suggests that the model is not a perfect fit. Figure 3I shows the distribution of the residual values in Figure 3H. Most residuals are from –0.5 to +0.5, with a couple values at –1.25. Even considering the values

at –1.25, the residuals span a much narrower range than when considering all peptides or the good peptides for analysis.

In summary, Figure 3 demonstrates the effect of the peptides used for analysis. Limiting the analysis to responding peptides results in the most linear plot of light vs heavy area with the lowest magnitude residuals. Reducing the residuals is likely necessary to be sensitive to the small magnitude changes expected from crosslink formation, even though we decrease our sequence coverage from 98 to 82% when limiting the analysis to the responding peptides.

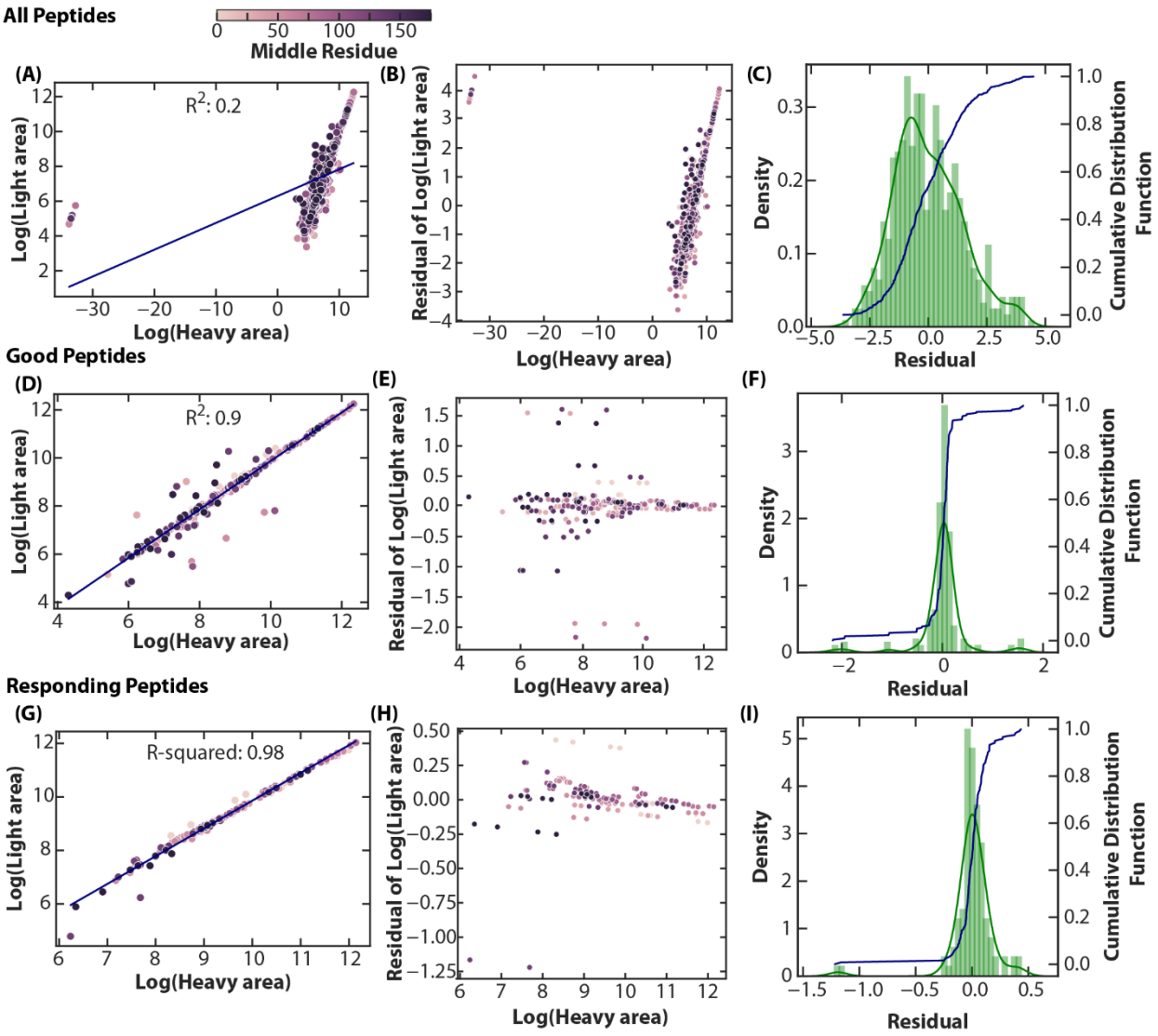

Figure 7. A single dataset from a mixture of non-crosslinked W9B-HSPB5 and SILAC WT-HSPB5 is analyzed with different sets of peptides. All analyzed peptides are from HSPB5. The data analyzed here is from untargeted DIA with a precursor range of 400–1100 *m/z* that was measured in 10 *m/z* intervals. In the sample analyzed here, SILAC and BPA protein were combined in gel. Panels A-C use all peptides for analysis. Panels D-F use selected good peptides for Skyline analysis. Panels G-I use responding peptides for analysis. Panels A, D, and G are plots of the $\log_{10}$ of the heavy area on the x-axis and the $\log_{10}$ of the light area on the y-axis. There is one point for every product ion corresponding to the analyzed peptides. Each of these

product ions are referred to as transitions. The points in panels A, D, and G are color-coded based on the middle residue of the peptide. The line of best fit is shown in navy with the R^2 displayed. Panels B, E, and H use the same color scheme as Panels A, D, and G and show the residuals of the log of the light area in the preceding panel. Panels C, F, and I show the distributions of the residuals in the preceding panels (Panels B, E, and H). Data here was collected using an untargeted DIA method with a precursor range of 400–1100 *m*/*z* that was measured in 10 *m*/*z* intervals.

5.3.3 Effect of Digestion Phase

Figure 4 describes the effect of the digestion phase on the analysis. In both samples in Figure 4, the BPA protein was in gel. In Figures 4A, 4B, and 4C, the SILAC protein was added in gel, and in Figures 4D, 4E, and 4F, the SILAC protein was added in solution. SILAC protein added in gel means that both BPA and SILAC protein are in the same phase during digestion. SILAC added in solution means that SILAC protein is in solution during digestion and BPA-containing protein is in gel.

Figure 4A shows the scatter plot of the log of the heavy area and the log of the light area as well as the line of best fit for SILAC protein being added in gel. The plot appears very linear and has an R^2 of 0.998. Figure 4D shows the same analysis for when SILAC protein was added in solution. More points are visible above and below the line of best fit than when SILAC protein is added in gel, and the R^2 is 0.99. The plot of the log of the light vs the log of the heavy area has a higher R^2 and appears to have fewer outliers when SILAC protein is added in gel.

Figure 4B shows the residuals of Figure 4A when SILAC protein is added in gel. The residuals are highly concentrated around zero, with a few values at –0.1, 0.2, and 0.3 as extremes. Figure 4E shows the residuals of Figure 4D when SILAC protein is added in solution.

The residuals are highly concentrated around zero, similar to Figure 4B, but have greater magnitude. Figures 4C and 4F are the distributions of the residuals in Figure 4B and Figure 4E, respectively. In Figure 4C, when the SILAC protein is added in gel, residuals range from –0.2 to 0.4. In Figure 4F, when SILAC protein is added in solution, residuals range from just below –0.6 to 0.6, which is a broader range than when SILAC protein was added in gel. The increase in residuals when SILAC is added in a different phase than the BPA protein suggests that having the proteins in different phases adds additional variability to the experiment, affecting the heavy and light peak areas. Therefore, it is advantageous to have the SILAC and BPA proteins in the same digestion phase when preparing samples. When preparing samples without crosslinks, the BPA and SILAC protein could be mixed before running the gel so that both SILAC and BPA protein are in the same monomer, or lanes with SILAC and BPA proteins can be run separately, and monomer bands can be combined before digestion. When preparing crosslinked samples, lanes of crosslinked reaction mixtures and SILAC protein were run separately. Dimers (dimeric products) were excised from the lane with the crosslinked reaction mixture, and monomers were excised from the lane with the SILAC protein. The dimeric products and SILAC monomers were combined to form crosslinked samples with SILAC and BPA crosslinks in the same phase.

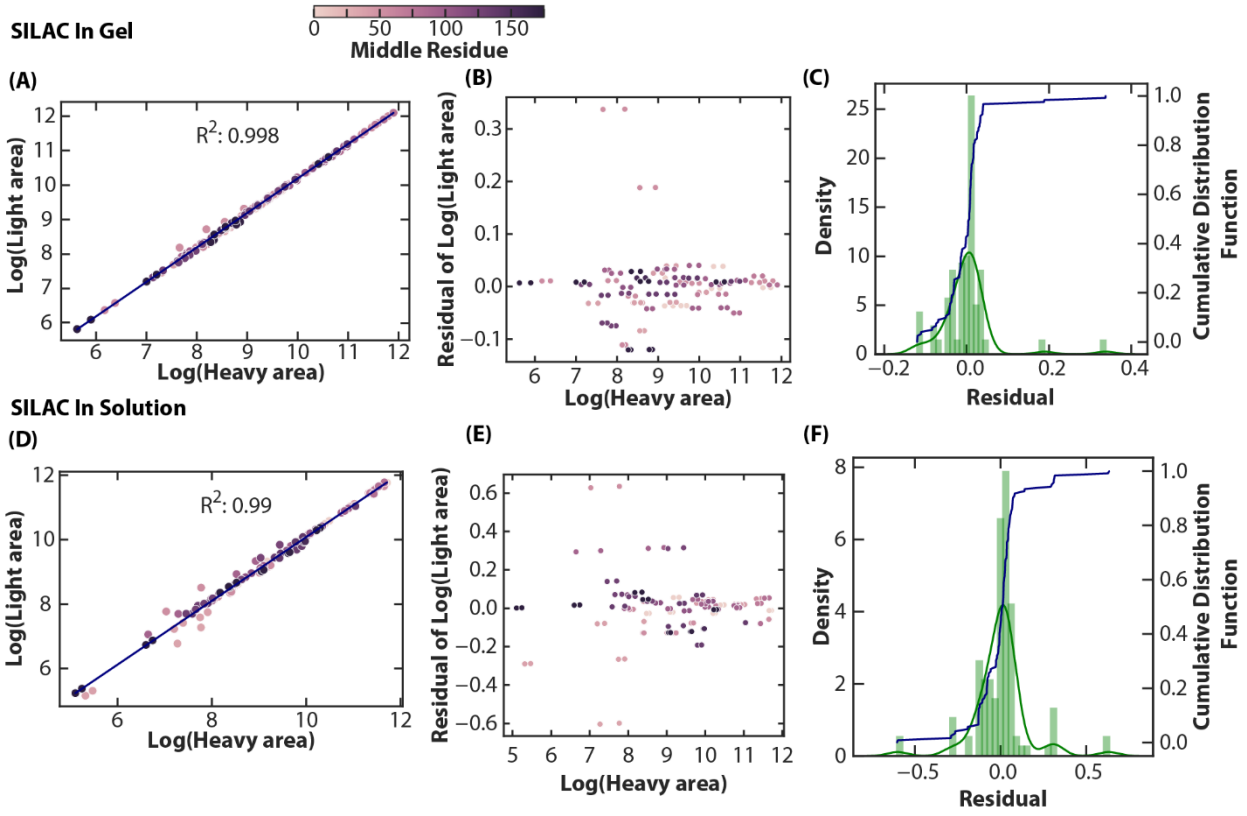

Figure 8. Datasets where SILAC protein was added in solution and where the SILAC protein was added in gel to BPA protein are compared here. Panels A, B, and C are from a sample where the SILAC protein was added in gel. Panels D, E, and F are from a sample where the SILAC protein was added in solution. The data analyzed here is from untargeted DIA with a precursor range of 400–900 *m/z* that was measured in 10 *m/z* intervals, and analysis used the responding peptides list. Panels A and D are plots of the $\log_{10}$ of the heavy area on the x-axis and the $\log_{10}$ of the light area on the y-axis. There is one point for every product ion corresponding to the analyzed peptides. Each of these product ions are referred to as transitions. The points in panels A and D are color-coded based on the middle residue of the peptide. The line of best fit is shown in navy with the R^2 displayed. Panels B and E use the same color scheme as Panels A and D, and show the residuals of the log of the light area in the preceding panel. Panels C and F show the distributions of the residuals in the preceding panels (Panels B and E). The data represented here

was collected using a DIA method with a precursor range of 400-900 *m/z* that was measured in 10 *m/z* intervals. The precursor range was shortened from a maximum of 1100 *m/z* to a maximum of 900 *m/z* compared to Figure 3 because only a few peptides were identified in that region.

5.3.4 Results for Crosslinked Samples

To compare results from crosslinked and non-crosslinked samples, we considered two different BPA sites: W9B and F61B. The crosslinks from these two positions are shown in Figure 5. W9B-HSPB5 has many crosslinks across the protein sequence, and F61B-HSPB5 has crosslinks primarily at the beginning of the protein sequence. Because many different peptides are crosslinked in W9B-HSPB5, many different non-crosslinked peptides would be depleted to form crosslinks. The depletion of many peptides with crosslink formation in W9B-HSPB5 may make it more challenging to identify which peptides are crosslinked because multiple peptides could have similar changes in the proportion of heavy area. In contrast, for F61B-HSPB5, which crosslinks primarily to the beginning of the protein sequence, many fewer peptides would be crosslinked, so the difference in heavy area for crosslinked peptides may be more apparent.

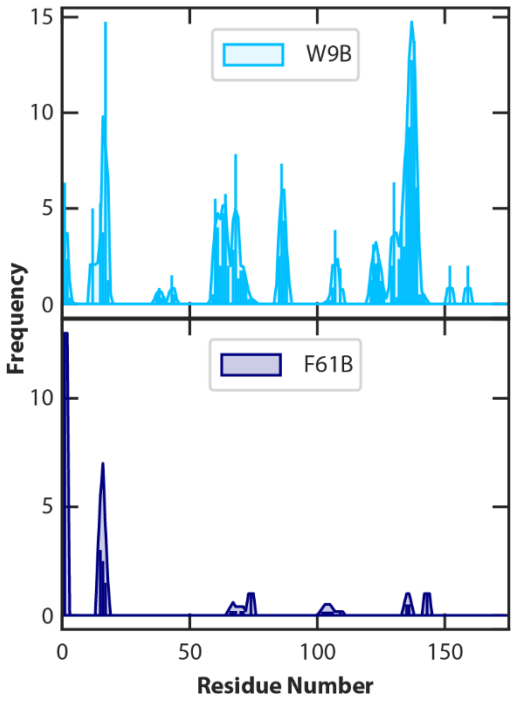

Figure 5. This Figure is reproduced from Chapter 4. Crosslinks from trypsin-digested, dimeric products of W9B and F61B-HSPB5 are shown. The x-axis is the residue number of the site crosslinked by BPA, and the frequency axis is the number of crosslink PSMs identifying that crosslink site. W9B has 195 total crosslink PSMs, and F61B has 25 total crosslink PSMs.

Figure 6 shows results from crosslinked samples of crosslinked W9B-HSPB5 (6A, 6B, 6C) and F61B-HSPB5 (6D, 6E, 6F) and non-crosslinked F38B-HSPB5 (G, H, I) that were all mixed with SILAC protein in gel. In crosslinked samples, we expected a depletion of the light area as more of a given peptide is crosslinked. Because we expect depletion of the light peptides, we expect any crosslinked peptides to lie below the line of best in the plots of log of light area vs log of heavy area. Figure 6A shows the plot of the log of light area vs the log of heavy area for a crosslinked W9B-HSPB5 sample. Most points appear visually to lie well within the line of best

fit, and the R^2 of 0.99 further supports that the linear model represents the data well. Figure 6D, which has crosslinked F61B-HSPB5, is similar, with an R^2 of 0.99. Figure 6G, which has no crosslinks, also has points that appear to lie closely along the line of best fit and a high R^2 of 0.996. Overall, plots of heavy vs light area with and without crosslinks appear similarly linear, and there are no significant outliers that can be attributed to the depletion of peptides that participate in crosslinks.

When looking at the residuals for the sample with crosslinked W9B-HSPB5 in Figure 6B, they appear to be most concentrated from +0.1 to –0.1, and there are a couple points at higher residual values near 0.25. This is very similar to Figure 4B, which represents a non-crosslinked sample, so a few higher residual points do not necessarily indicate crosslinked peptides. When looking at the residuals for the sample with crosslinked F61B-HSPB5 in Figure 6E, they appear to be most concentrated from 0.1 to –0.1, and there are a couple of points at higher and lower values residual values of –0.4 and +0.4. This shape with a few points at higher magnitude residual values is quite similar to those shown in Figure 4E, which had no crosslinks. The residuals for the sample with non-crosslinked F38B-HSPB5 in Figure 6H are scattered evenly throughout the residual range. This does differ from Figures 6B and 6E, which had crosslinks, but other samples without crosslinks do not showcase this pattern (Figures 3H and 4B).

The distributions of residuals for crosslinked W9B-HSPB5 in Figure 6C range from –0.2 to +0.4, with just a couple of outlier points at 0.3 and the main distribution from –0.2 to +0.2. The distributions of residuals for crosslinked F61B-HSPB5 in Figure 6F range from –0.5 to +0.5, and the distribution is relatively symmetric. The distributions of residuals for non-crosslinked F38B-HSPB5 in Figure 6I range from –0.2 to +0.2, and the distribution is relatively symmetric. The residuals for the crosslinked samples are slightly higher in magnitude than for the non-

crosslinked samples. Still, there are no lower residuals that would suggest peptide depletion corresponding to crosslinked formation.

The very similar R^2 and residual values for samples with and without crosslinks suggest that no significant difference could be attributed to the depletion of peptides upon crosslink formation. This is most likely because the substoichiometric changes in peptide abundance from crosslink formation lie within the noise of our system.

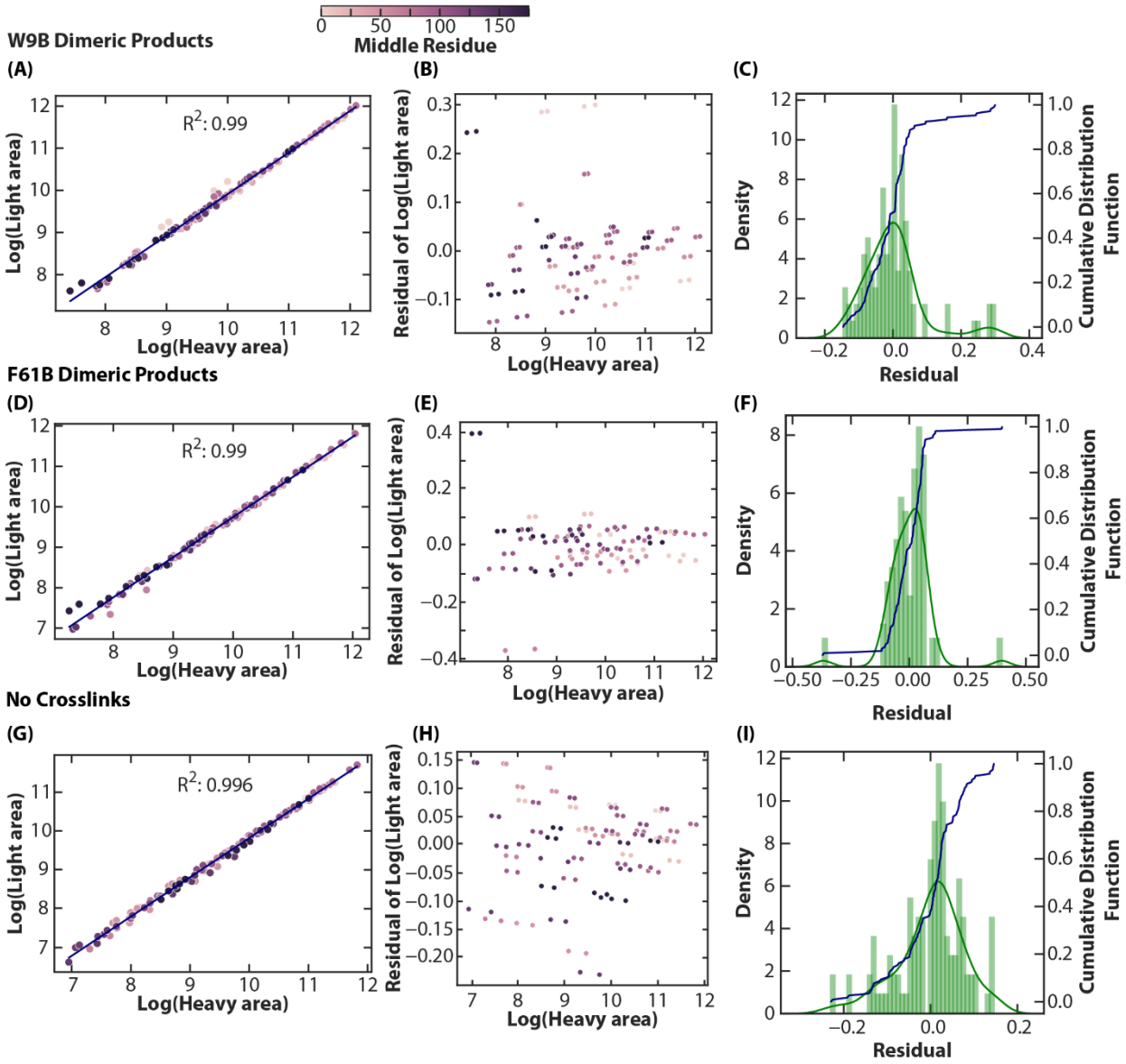

Figure 6. Datasets with and without crosslinks are compared here. The data analyzed here is from untargeted DIA with a precursor range of 400–900 *m*/*z* that was measured in 10 *m*/*z* intervals. Analysis used the responding peptides list; all samples had SILAC and BPA protein added in gel. Panels A, B, and C are from a sample W9B-HSPB5 crosslinks. Panels D, E, and F show results from a sample with F61B crosslinks. Panels G, H, and I are from a non-crosslinked sample with F38B-HSPB5. Panels A, D, and G are plots of the $\log_{10}$ of the heavy area on the x-

axis and the $\log_{10}$ of the light area on the y-axis. There is one point for every product ion corresponding to the analyzed peptides. The points in Panels A, D, and G are color-coded based on the middle residue of the peptide. The line of best fit is shown in navy with the R^2 displayed. Panels B, E, and H use the same color scheme as Panels A, D, and G, and show the residuals of the log of the light area in the preceding panel. Panels C, F, and I show the distributions of the residuals in the preceding panels (Panels B, E, and H). The data represented here was collected using a DIA method with a precursor range of 400-900 *m/z* that was measured in 10 *m/z* intervals.

5.4 Challenges and Potential Next Steps

In this approach, we aim to quantify the small-magnitude, sub-stoichiometric changes in abundance from crosslink formation to identify crosslinked peptides. This strategy aims to quantify differences in abundance within a single sample as opposed to comparing two different samples. Although this method of crosslink identification will not be site-specific, it should be easier to identify crosslinked peptides because unlinked peptides are much higher in concentration than crosslinked peptides. Hence, they are much further above the limit of detection and better able to generate validation models. With this SILAC approach, we can still identify residue-level crosslinks because the crosslinked data is not isotopically labeled,[19] and we can also identify crosslinks based on quantification. Assuming the BPA peptide is in a protein with 20 well-defined enzymatic peptides, if it crosslinked equally to each of those 20 peptides, there would be a 5% change in abundance due to the crosslinking reaction. The assumption that estimates a 5% change in abundance assumes the formation of more crosslinked peptides than we observe and a much cleaner, more complete digestion than we observe. 208 different peptides are observed when considering non-enzymatic peptides in a non-crosslinked sample. Crosslinked

samples would even complicate this even further, as the presence of crosslinks is known to interfere with digestion.[28] The difference in digestion with crosslinks may also limit this approach. For discussion, we will use the 5% change estimation as a benchmark for the small magnitude of change that needs to be detectable to identify crosslinked peptides through depletion.

` To detect a 5% change in abundance, the precision of the data needs to be high enough that it is clear that a 5% change is significant and not due to the variance of the measurement. The coefficient of variation (CV, ratio of the standard deviation to the mean) is commonly reported as a measure of the precision of quantification studies. These CVs are typically reported per peptide or per protein quantified or as the median value for the peptides/proteins analyzed and have been described as being less than or greater than 10%.[29] To detect changes of 5% in magnitude, we would need a CV of less than 5% (ideally lower) to determine if changes of 5% are outside the variability of the system. A 5% CV would be extremely low compared to commonly reported values.[5,30] However, there have been reports that simplifying samples (in the form of depleting plasma samples) can reduce CVs when analyzing plasma samples with Multiple Reaction Monitoring (MRM) on a triple quadruple.[31] Half of the MRM methods used in that study had CVs of less than 5% when analyzing depleted samples. Because our study focuses primarily on samples containing a single protein, obtaining the lower CVs necessary to quantify small-magnitude changes is likely more feasible. However, the similarity of plots representing data with and without crosslinks in Figure 6 indicates that our current system is not sensitive to sub-stoichiometric changes in magnitude. Potential changes to help make the measurements more precise and more sensitive to small magnitude changes would be to use targeted DIA methods or a quadrupole time-of-flight or triple quadruple instrument instead of an orbitrap.

5.5 Conclusion

Figure 1 outlines the overall goals of this study: to identify crosslinks through quantifying the depletion of non-crosslinked peptides. Figure 2 shows an example of how heavy and light peak area were quantified on a sample of only SILAC protein and illustrates that our incorporation of SILAC residues is over 99%. Figure 3 demonstrates that minimizing the list of peptides used for quantification improves the linearity of results at the cost of reducing sequence coverage. Figure 4 illustrates that combining SILAC and BPA proteins in the same phase prior to digestion increases linearity and reduces noise in the results. Figure 5 shows example crosslink results for W9B-HSPB5 and F61B-HSPB5 to demonstrate that we expect fewer different peptides to be crosslinked with F61B, and that could result in higher magnitude depletions upon crosslink formation. Figure 6 compares results from samples containing W9B-HSPB5 crosslinks, samples containing F61B-HSPB5 crosslinks, and a sample without crosslinks and demonstrates that results are fairly similar with no significant differences between samples with and without crosslinks.

Overall, here we described how aspects of the sample preparation and data analysis could interfere with the linearity of the results and determined that limiting the quantification of the peptides and combining SILAC and crosslinked protein in the same phase yield the most linear results. To help increase the sensitivity towards small magnitude changes by increasing the precision, future work could use targeted DIA methods or a different instrument such as a quadrupole time-of-flight or a triple quadruple.

5.6 Acknowledgements

I thank Natalie L, Stone, Lucas Narisawa, Christopher N. Woods, Maria Janowska, Rachel E. Klevit, and Matthew F. Bush for their contributions to this work. This material is

based upon work supported by the National Eye Institute through R01EY017370 to R.E.K., the National Institute of General Medical Sciences through T32 GM008268 to C.N.W., the National Institute of Aging through T32 AG066574 to L.D.U., and the University of Washington's Proteomics Resource (UWPR95794).

5.7 References

(1) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376–386. https://doi.org/10.1074/mcp.M200025-MCP200.

(2) Ong, S. Mass Spectrometric-Based Approaches in Quantitative Proteomics. *Methods* **2003**, *29* (2), 124–130. https://doi.org/10.1016/S1046-2023(02)00303-1.

(3) Mann, M. Functional and Quantitative Proteomics Using SILAC. *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (12), 952–958. https://doi.org/10.1038/nrm2067.

(4) Kurokawa, N.; Kishimoto, T.; Tanaka, K.; Kondo, J.; Takahashi, N.; Miura, Y. New Approach to Evaluating the Effects of a Drug on Protein Complexes with Quantitative Proteomics, Using the SILAC Method and Bioinformatic Approach. *Biosci. Biotechnol. Biochem.* **2019**, *83* (11), 2034–2048. https://doi.org/10.1080/09168451.2019.1637244.

(5) Pino, L. K.; Baeza, J.; Lauman, R.; Schilling, B.; Garcia, B. A. Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation. *J. Proteome Res.* **2021**, *20* (4), 1918–1927. https://doi.org/10.1021/acs.jproteome.0c00938.

(6) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Zheng, C.; Taipale, A.; Zhang, Y.; Takara, K.; Bruce, J. E. Quantitative Interactome Analysis Reveals a Chemoresistant Edgotype. *Nat. Commun.* **2015**, *6* (1), 7928. https://doi.org/10.1038/ncomms8928.

(7) Salazar, A.; Keusgen, M.; Von Hagen, J. Amino Acids in the Cultivation of Mammalian Cells. *Amino Acids* **2016**, *48* (5), 1161–1171. https://doi.org/10.1007/s00726-016-2181-8.

(8) Chen, X.; Wei, S.; Ji, Y.; Guo, X.; Yang, F. Quantitative Proteomics Using SILAC: Principles, Applications, and Developments. *PROTEOMICS* **2015**, *15* (18), 3175–3192. https://doi.org/10.1002/pmic.201500108.

(9) Soufi, B.; Macek, B. Stable Isotope Labeling by Amino Acids Applied to Bacterial Cell Culture. In *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)*; Warscheid, B., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2014; Vol. 1188, pp 9–22. https://doi.org/10.1007/978-1-4939-1142-4_2.

(10) Ping, L.; Zhang, H.; Zhai, L.; Dammer, E. B.; Duong, D. M.; Li, N.; Yan, Z.; Wu, J.; Xu, P. Quantitative Proteomics Reveals Significant Changes in Cell Shape and an Energy Shift after IPTG Induction via an Optimized SILAC Approach for *Escherichia Coli*. *J. Proteome Res.* **2013**, *12* (12), 5978–5988. https://doi.org/10.1021/pr400775w.

(11) Han, J.; Yi, S.; Zhao, X.; Zheng, Y.; Yang, D.; Du, G.; Yang, X.-Y.; He, Q.-Y.; Sun, X. Improved SILAC Method for Double Labeling of Bacterial Proteome. *J. Proteomics* **2019**, *194*, 89–98. https://doi.org/10.1016/j.jprot.2018.12.011.

(12) Waugh, DavidS. Genetic Tools for Selective Labeling of Proteins with Alpha-15N-Amino Acids. *J. Biomol. NMR* **1996**, *8* (2). https://doi.org/10.1007/BF00211164.

(13) Matic, I.; Jaffray, E. G.; Oxenham, S. K.; Groves, M. J.; Barratt, C. L. R.; Tauro, S.; Stanley-Wall, N. R.; Hay, R. T. Absolute SILAC-Compatible Expression Strain Allows

Sumo-2 Copy Number Determination in Clinical Samples. *J. Proteome Res.* **2011**, *10* (10), 4869–4875. https://doi.org/10.1021/pr2004715.

(14) Parker, S. J.; Venkatraman, V.; Van Eyk, J. E. Effect of Peptide Assay Library Size and Composition in Targeted Data-independent acquisition-MS Analyses. *PROTEOMICS* **2016**, *16* (15–16), 2221–2237. https://doi.org/10.1002/pmic.201600007.

(15) Cappadona, S.; Baker, P. R.; Cutillas, P. R.; Heck, A. J. R.; Van Breukelen, B. Current Challenges in Software Solutions for Mass Spectrometry-Based Quantitative Proteomics. *Amino Acids* **2012**, *43* (3), 1087–1108. https://doi.org/10.1007/s00726-012-1289-8.

(16) Casavant, E. P.; Liang, J.; Sankhe, S.; Mathews, W. R.; Anania, V. G. Using SILAC to Develop Quantitative Data-Independent Acquisition (DIA) Proteomic Methods. In *SILAC*; Luque-Garcia, J. L., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2023; Vol. 2603, pp 245–257. https://doi.org/10.1007/978-1-0716-2863-8_20.

(17) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proc. Natl. Acad. Sci.* **2002**, *99* (17), 11020–11024. https://doi.org/10.1073/pnas.172226299.

(18) Dorman, G.; Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **1994**, *33* (19), 5661–5673. https://doi.org/10.1021/bi00185a001.

(19) Ulmer, L.; Canzani, D.; Woods, C.; Stone, N.; Janowska, M.; Klevit, R.; Bush, M. High-Performance Workflow for Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid. *J. Proteome Res.* **2024**. https://doi.org/10.1021/acs.jproteome.4c00194.

(20) Woods, C. N.; Ulmer, L. D.; Guttman, M.; Bush, M. F.; Klevit, R. E. Disordered Region Encodes α-Crystallin Chaperone Activity toward Lens Client γD-Crystallin. *Proc. Natl. Acad. Sci.* **2023**, *120* (6), e2213765120. https://doi.org/10.1073/pnas.2213765120.

(21) Woods, C. N.; Ulmer, L. D.; Janowska, M. K.; Stone, N. L.; James, E. I.; Guttman, M.; Bush, M. F.; Klevit, R. E. HSPB5 Disease-Associated Mutations Have Long-Range Effects on Structure and Dynamics through Networks of Quasi-Ordered Interactions. *bioRxiv*. https://doi.org/10.1101/2022.05.30.493970.

(22) Deutsch, E. W.; Mendoza, L.; Shteynberg, D. D.; Hoopmann, M. R.; Sun, Z.; Eng, J. K.; Moritz, R. L. Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **2023**, *22* (2), 615–624. https://doi.org/10.1021/acs.jproteome.2c00624.

(23) Eng, J. K.; Hoopmann, M. R.; Jahan, T. A.; Egertson, J. D.; Noble, W. S.; MacCoss, M. J. A Deeper Look into Comet—Implementation and Features. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1865–1874. https://doi.org/10.1007/s13361-015-1179-x.

(24) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(25) *cRAP protein sequences*. https://www.thegpm.org/crap/ (accessed 2022-04-19).

(26) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392. https://doi.org/10.1021/ac025747h.

(27) Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; MacLean, B.; MacCoss, M. J. The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. *Mass Spectrom. Rev.* **2020**, *39* (3), 229–244. https://doi.org/10.1002/mas.21540.

(28) Dau, T.; Gupta, K.; Berger, I.; Rappsilber, J. Sequential Digestion with Trypsin and Elastase in Cross-Linking Mass Spectrometry. *Anal. Chem.* **2019**, *91* (7), 4472–4478. https://doi.org/10.1021/acs.analchem.8b05222.

(29) Steigerwald, S.; Sinha, A.; Fort, K. L.; Zeng, W.-F.; Niu, L.; Wichmann, C.; Kreutzmann, A.; Mourad, D.; Aizikov, K.; Grinfeld, D.; Makarov, A.; Mann, M.; Meier, F. Full Mass Range ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis. *Mol. Cell. Proteomics* **2024**, 100713. https://doi.org/10.1016/j.mcpro.2024.100713.

(30) Hanke, S.; Besir, H.; Oesterhelt, D.; Mann, M. Absolute SILAC for Accurate Quantitation of Proteins in Complex Mixtures Down to the Attomole Level. *J. Proteome Res.* **2008**, *7* (3), 1118–1130. https://doi.org/10.1021/pr7007175.

(31) Anderson, L.; Hunter, C. L. Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Mol. Cell. Proteomics* **2006**, *5* (4), 573–588. https://doi.org/10.1074/mcp.M500331-MCP200.