Identifying Site-Specific Crosslinks Originating from a Genetically Incorporated, Photoreactive Amino Acid

Lindsey D. Ulmer,[1] Daniele Canzani,[1] Christopher N. Woods,[2] Natalie L. Stone,[2]

Maria K. Janowska,[2] Rachel E. Klevit,[2] Matthew F. Bush[1,*]

Contribution from:

[1] University of Washington, Department of Chemistry, Box 351700, Seattle, WA 98195-1700

[2] University of Washington, Department of Biochemistry, Box 357350, Seattle, WA 98195-7350

*Address correspondence to mattbush@uw.edu

Abstract

In traditional 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. In this study, we employ an alternative approach using benzoylphenylalanine (BPA), a photoreactive amino acid, incorporated into a disordered region of the human protein HSPB5. BPA is 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 can react with all amino acids, which also overcomes limitations of selective reactivity. However, this broad reactivity imposes additional challenges for crosslink identification. We report and characterize the experimental methods and informatics pipeline used to identify and visualize residue-level interactions originating from BPA. We routinely identify 30 to 300 crosslinked peptide spectral matches with this workflow, depending on the site of BPA incorporation. Most identified crosslinks are assigned to a precision of one or two residues, which is supported by a high degree of overlap between technical replicates. Based on these results, we anticipate that this approach will be a powerful, general strategy for characterizing the structures of proteins that have resisted high-resolution characterization, including disordered and heterogeneous proteins.

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 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] 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 from photo-methionine, photo-leucine, and BPA have been identified using database searching with programs such as StavroX,[18–22] MeroX,[23,24] and Crossfinder.[19] 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 interpret their results at the protein level, e.g., using gel-based assays to determine whether crosslinks were formed to a target[25–27] or using quantitative proteomics to characterize the preferential co-isolation of interacting proteins after exposure to

UV light.[28,29] Some studies have localized the site of crosslinking, but often only a single crosslink is reported.[30,31]

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 pipeline (Figure 1) and then characterized the performance of that pipeline for variants of the human small heat shock protein (sHSP) HSPB5 that each contain BPA incorporated at a single site. The informatics pipeline benefits from the use of msconvert,[32] Comet,[33] Kojak,[34,35] PeptideProphet,[36] and other tools from the Trans Proteomic Pipeline (TPP),[37] 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 pipeline with ones reported previously using StavroX[38] and MeroX.[39,40] 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,[41,42] 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.[43–46] Here, we report and characterize the strategy that we applied recently to identify novel, residue-level interactions that were key factors in characterizing NTR-NTR interactions[47,48] and the origin of selectivity[47] 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.

Methods

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.[47,48] Purified BPA-containing variants were UV-treated to form crosslinks, and the product mixtures were subjected to SDS-PAGE on a precast 4–20% acrylamide gradient gel (Bio-Rad, 4561096) as described elsewhere.[47,48] 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.[49] The weight of the peptides in these samples was estimated based on the initial weight of proteins loaded onto the gel and the relative color density of the band excised from the gel. 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 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.

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 reactants. TPP version 6.3.2 was used; some effects of the software version are discussed in the Supporting Information. First, Comet[33] 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,[50] the pertinent HSPB5-BPA variant, 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 (FDR) and a minimum of 2 peptide spectral matches (PSMs), this protein yields a validated protein database for the 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[34,35] and the validated-protein database for the monomeric reactants. 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.

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[51] 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.

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 pipeline, which uses Comet,[33] Kojak,[34,35] PeptideProphet,[36] and other open-source tools from the Trans Proteomic Pipeline (TPP).[37] 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).[43–46] 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 structures of sHSPs, which is evidenced by the significant differences between the structures proposed for 24mers of HSPB5.[52,53] 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 in the disordered NTR to characterize a highly heterogeneous protein that has resisted characterization by conventional structural biology approaches.

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. To gain the broadest understanding of the proteins present in the sample, we include all proteins from the *E. coli* expression system and potential contaminants (e.g., keratins and common proteins associated with molecular biology)[50] 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).

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. Figure 2 shows that for samples with BPA incorporated at site 9 (W9B) that were digested with trypsin, the number of crosslink PSMs generally decreases as database size increases and levels off at about 150. Decreasing numbers of crosslink PSMs with increasing database size is consistent with prior reports.[5] Larger databases would allow for the identification of crosslinks with nontarget proteins and decrease the likelihood of misidentifying products. The validated-protein database associated with the analysis in Figure 2 has 12 proteins and is the second smallest database considered. The validated-protein database yielded the most crosslink PSMs (the crosslinks are shown in Figure S2), even more than if the protein database only includes the target protein. The selection of the protein database affects the sensitivity and error of the search. Using a 2 PSM minimum at 1% FDR yields a small database that appears to accurately describe proteins that are present in the sample. A smaller database increases sensitivity because there are fewer candidate identities for each spectrum, which can increase the scores assigned to the remaining candidates. These results suggest that the use of the validated-protein databases minimizes the loss of sensitivity from considering additional proteins, while still providing the benefits of considering signals originating from contaminant proteins that are in the sample.

The protein databases used in Figure 2 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 with the non-UV-treated database and 298 with the UV-treated database. Because similar numbers of crosslink PSMs were identified, using either sample to create the validated protein database appears to be sufficient. In most of the results presented here, we used a monomeric reactant band 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 similar numbers of crosslink PSMs 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).

Performance of Informatic Workflow. Figure 2 also shows the number of crosslink PSMs identified and the corresponding search times for our workflow, StavroX,[38] and MeroX.[39,40] 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[18–22] and

MeroX[23,24] 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 pipeline, 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.

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 (typically 1%), which are 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 proteins, i.e., HSPB5-nontarget crosslink PSMs, are reported for 13 samples containing one of 6 variants in Table S1 and for 9 technical 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 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.

These FDR values are estimated from the validation models, which consider many factors including matches to decoys that are not in the sample.[54] 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. 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. 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 pipelined 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.

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).[43–46] 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 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,[55] 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 figure by indicating the site of BPA incorporation for each crosslink.

Ambiguities in Assigning Residue-Specific Crosslinks. When there are gaps in the coverage of b and y ions in the crosslinked peptide, there can be ambiguity in the crosslink site assignment. As described above, for 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 tworesidues (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).

Reproducibility. A total of nine technical replicates were performed of the preceding analysis; one replicate is shown in Figure 5B and all are shown in Figure 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 edge groove in the full-length protein.[56] 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.[47,48] More generally, these results illustrate the promising potential of this workflow to robustly identify residue-level interactions in highly heterogenous systems with reputations for disorder.

Effects of Digestion. The previous sections only considered samples isolated from the dimeric product 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 6 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.[57–59]

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 HSPB5-HSPB5 dimers, 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.

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.[47] 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 Figure 7. 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 samples. 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.

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 pipeline 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 site 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 technical 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

strategy will be a powerful, general strategy for characterizing the structures of proteins that have resisted high-resolution characterization, including disordered and heterogeneous proteins.

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).

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.

(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) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) 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.

(33) 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.

(34) 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.

(35) 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.

(36) 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.

(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) 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.

(39) 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.

(40) 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.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) 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.

(46) 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.

(47) 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.

(48) 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.

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

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

(51) 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.

(52) 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.

(53) 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.

(54) 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.

(55) 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.

(56) 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.

(57) 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.

(58) 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.

(59) 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.

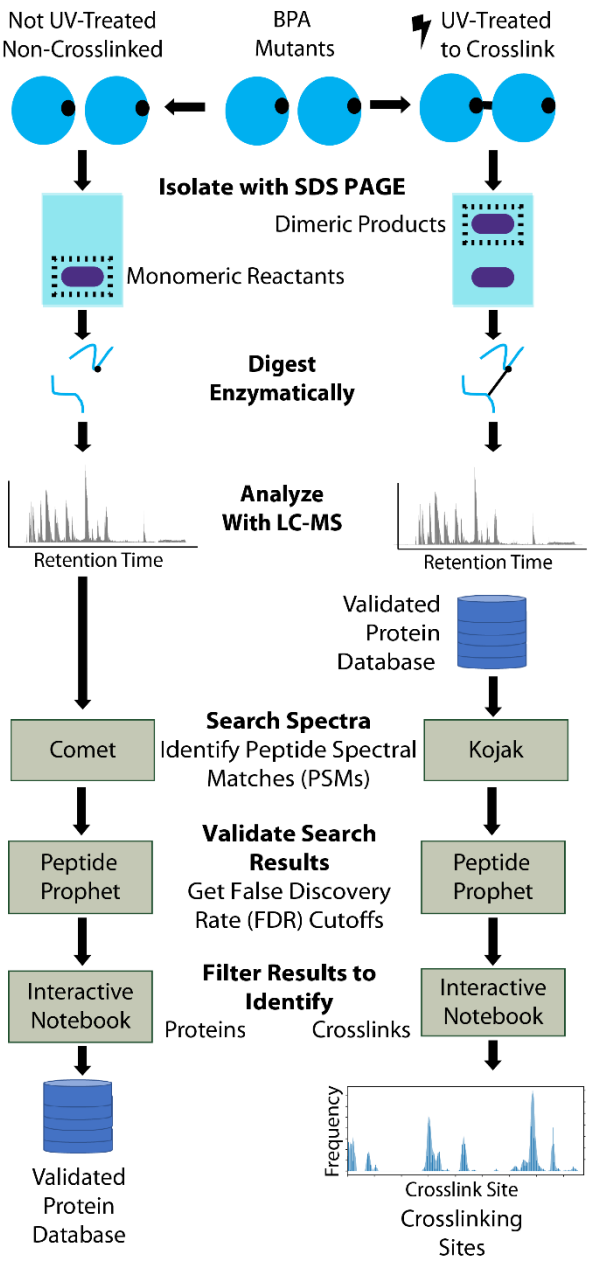

Figure 1. Samples are divided into two fractions prior to the crosslinking reaction to have a non-crosslinked control. The left describes the workflow for the analysis of the monomeric reactants, which was used to create the validated-protein database. The right describes the workflow for the analysis of the dimeric products, which was used to identify site-specific crosslinks.

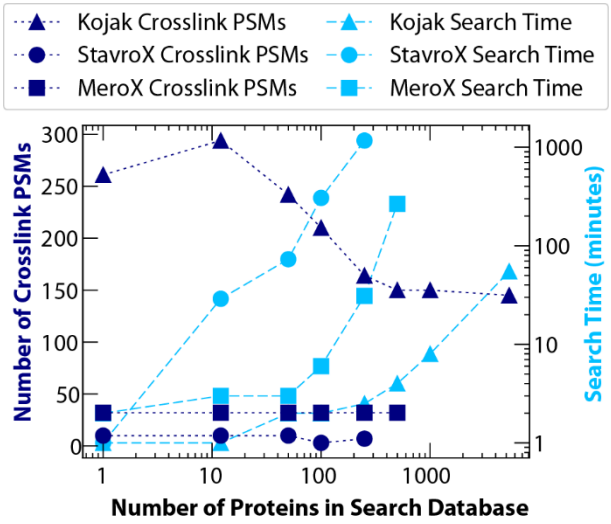

Figure 2. The number of crosslink PSMs found and the corresponding search times for varying protein-database sizes using Kojak, StavroX, or MeroX. These searches were all performed on the same LC-MS data of trypsin-digested, dimeric products of W9B. StavroX identified 10 or fewer crosslink PSMs across 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.

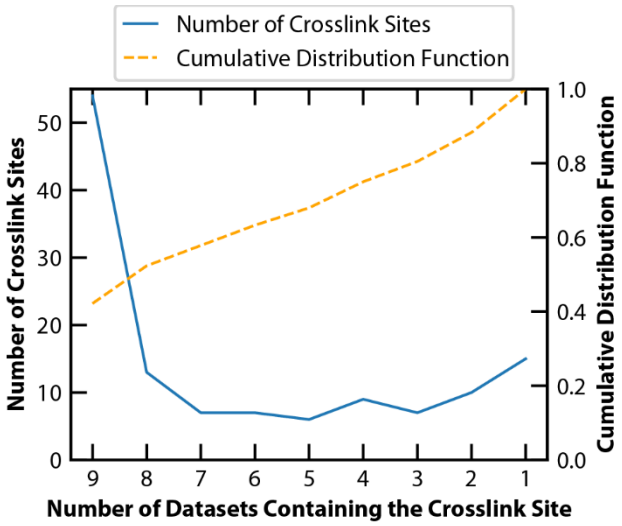

Figure 3. The results of 9 technical replicates of trypsin-digested, dimeric products of W9B at pH 6.5 (Figure S3 and Table S3) are compared here. The number of crosslink sites identified (defined as residue values with a frequency value greater than 0 in Figure S3) across differing numbers of replicates is indicated. A value of 9 datasets indicates that a given crosslink was identified in all 9 datasets, whereas a value 1 indicates that a given crosslink was only identified in a single dataset. Of the 128 total crosslink sites identified 54 are identified in all 9 replicates.

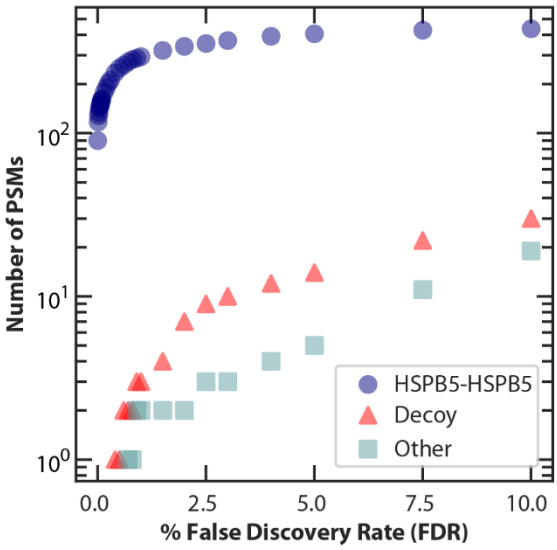

Figure 4. The number of HSPB5-HSPB5 crosslinks, decoy crosslinks, and other crosslinks identified as a function of FDR. The search represented in this data uses the validated-protein database (12 proteins) on the same dataset of trypsin-digested, dimeric products of W9B represented in Figure 2.

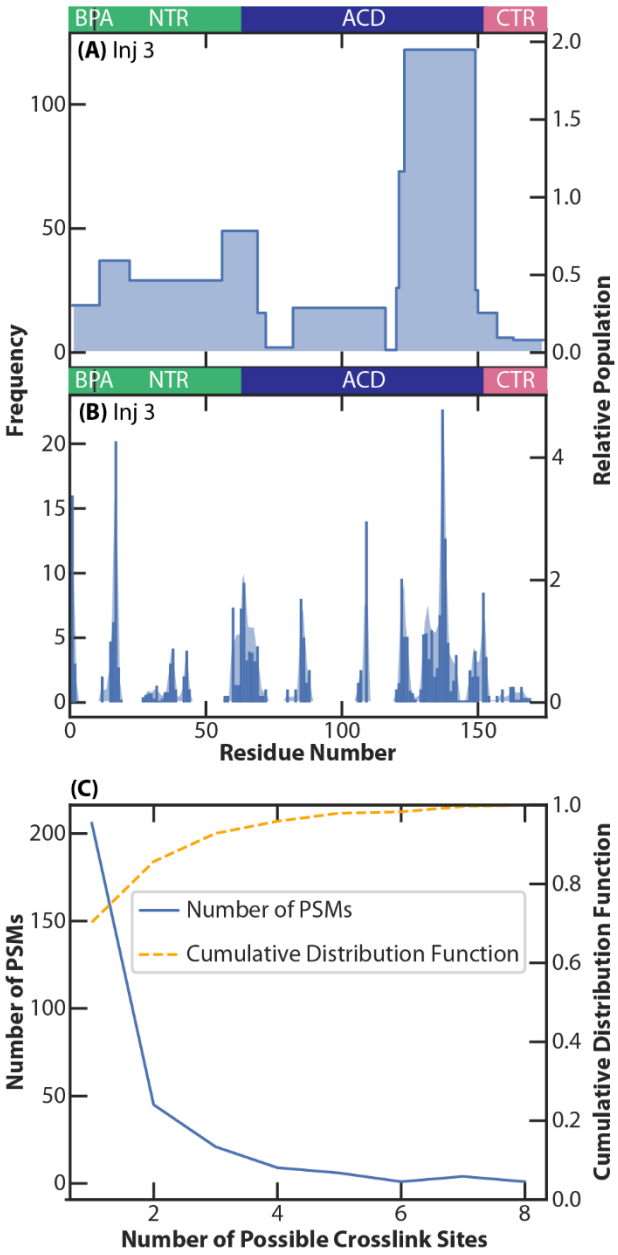

Figure 5. Above panels A and B, the BPA position and domain boundaries are shown for the structural regions of HSPB5: a disordered N-terminal region (NTR), an α-crystallin domain (ACD) that folds into two anti-parallel β-sheets, and a disordered C-terminal region (CTR). All panels represent the same technical replicate of trypsin-digested W9B at pH 6.5, which is injection 3 in Figure S3 and Table S3. Panel A illustrates a peptide level representation of the results shown in panel B. In panel A, every position in a crosslinked peptide received a PSM value of one. In panel B, 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.

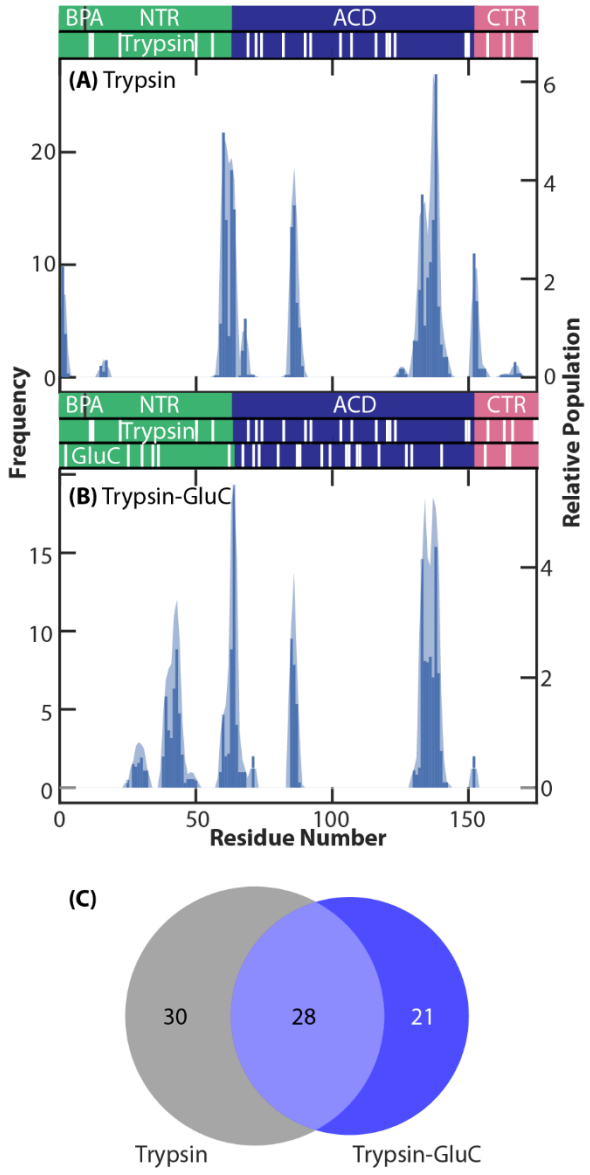

Figure 6. Above panels A and B, the top row shows the BPA position and domain boundaries. The following rows show the expected cleavage sites for trypsin or GluC. Panels A and B depict W9B samples with varying digestion enzymes. Panel A has 277 crosslink PSMs from trypsin-digested W9B at pH 6.5. Panel B has 195 crosslink PSMs from the trypsin-GluC-digested W9B at pH 6.5. The large tryptic peptide is from sites 23-56, and we do see crosslinks to that region when using a trypsin-GluC, parallel digestion. The raw file represented in panel B has been reported previously.[47] 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).

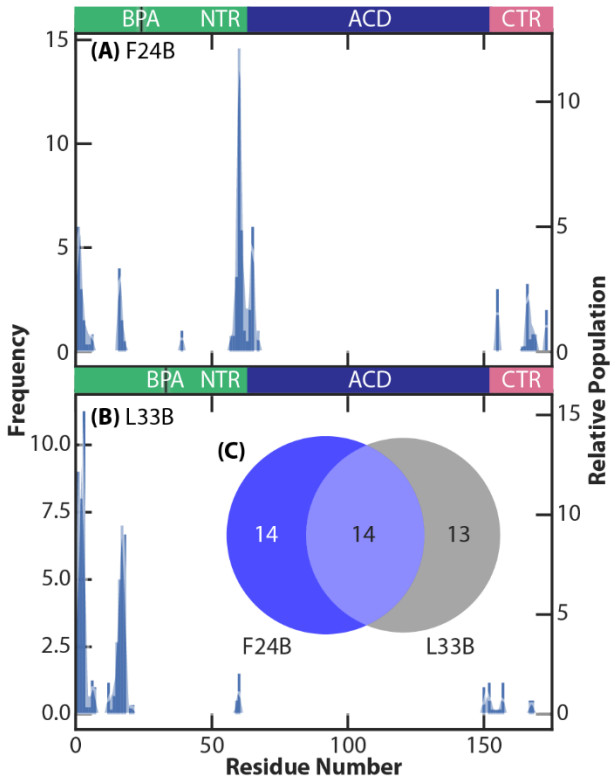

Figure 7. Panels A and B depict trypsin-GluC-digested samples with differing sites of BPA incorporation. Panel A has 66 crosslink PSMs from F24B. Panel B has 62 crosslink PSMs from L33B. The Venn diagram illustrates the overlap in crosslink sites identified in F24B (panel A) and L33B (panel B). About half of the crosslink sites identified are found in both samples. Both raw files represented in this figure have been reported previously.[48]

For Table of Contents Only