GDCh Research Article Angewandte International Edition Chemie www.angewandte.org

How to cite: *Angew. Chem. Int. Ed.* **2025**, *64*, e202507348
doi.org/10.1002/anie.202507348

Mass Spectrometry

Protein Secondary Structure Patterns in Short-Range Cross-Link Atlas

Alice Vetrano[+], *Alessio Di Ianni*[+], *Nico Di Fonte, Gianluca Dell'Orletta, Samantha Reale, Isabella Daidone, and Claudio Iacobucci**

Abstract: Cross-linking mass spectrometry (XL-MS) has become a powerful tool in structural biology for investigating protein structure, dynamics, and interactomics. However, short-range cross-links, defined as those connecting residues fewer than 20 positions apart, have traditionally been considered less informative and largely overlooked, leaving significant data unexplored in a systematic manner. Here, we present a system-wide analysis of short-range cross-links, demonstrating their intrinsic correlation with protein secondary structure. We introduce the X-SPAN (Cross-link Structural Pattern Analyzer) software, which integrates publicly available XL-MS datasets from system-wide experiments with AlphaFold-predicted protein structures. Our analysis reveals distinct cross-linking patterns that reflect the spatial constraints imposed by secondary structural elements. Specifically, α-helices exhibit periodic cross-linking patterns consistent with their characteristic helical pitch, whereas coils and β-strands display nearly monotonic distributions. A context-dependent protein grammar reinforces short-range cross-link specificity. Short-range cross-links can enhance the statistical inference of secondary structures within integrative modeling workflows. Additionally, our work establishes a framework for benchmarking AlphaFold's local prediction accuracy and provides novel quality control criteria for XL-MS experiments. We anticipate that X-SPAN and our short-range cross-link database will serve as a valuable resource for exploring local secondary structure rearrangements and their potential roles in protein function and allosteric regulation.

Introduction

Structural biology aims to elucidate the three-dimensional (3D) structure of proteins and protein complexes as an essential step for understanding biological functions. Among the structural proteomics toolkit, cross-linking mass spectrometry (XL-MS) emerged as a key approach to study protein structures and interactions on a system-wide scale.[1–6] XL-MS comprehensively captures the morphology of dynamic biological assemblies in their native environment, providing structural data that enhances cryo-electron microscopy (cryo-EM) methods[7,8] and computational predictions. Recent advances highlight the broader applicability of cross-linking mass spectrometry (XL-MS) beyond single-protein structural characterization. In particular, XL-MS has proven highly effective when integrated with cryo electron microscopy (cryoEM) and computational modeling approaches such as molecular dynamics simulations, enabling detailed exploration of protein-protein interaction networks in complex biological assemblies. For instance, recent studies demonstrated the power of this integrative methodology to unravel ambiguous protein localization in mitochondria and to construct high-resolution spatial interactomes in viral particles, emphasizing its versatility and significance in systems biology[9,10] Given the relatively low structural resolution of XL-MS, typically a few 10 of Å, the most informative insights come from cross-links between residues that are distant in the protein sequence or belong to different proteins. These cross-links provide valuable geometrical restraints, indicating the proximity of distinct regions, domains, and chains, thereby defining the global architecture of proteins. They help reveal long-range structural relationships and contribute to a deeper understanding of overall protein organization. Conversely, cross-links between residues that are close in sequence, e.g. fewer than 20 residues apart, are often considered less informative. Residues close in sequence are generally expected to be physically near each other in the 3D space. Cross-links between such residues would impose trivial constraints that offer limited value in determining the 3D structure. The limited interest in short-range cross-links is reflected in how various XL-MS data analysis and visualization tools handle them. For instance, some XL search engines, such as MeroX,[11] offer the option to exclude all potential short cross-links from the analysis to reduce computational costs when performing proteome-wide studies. The important xiVIEW[12]-based PRIDE[13] Crosslinking Database supports filtering cross-links based on

[*] Dr. A. Vetrano[+], A. Di Ianni, N. Di Fonte, G. Dell'Orletta, Prof. S. Reale, Prof. I. Daidone, Prof. C. Iacobucci
Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, L'Aquila 67100, Italy
E-mail: claudio.iacobucci@univaq.it

A. Di Ianni
Present address: Human Technopole, V.le Rita Levi Montalcini 1, Milan 20157, Italy

[+] Both authors contributed equally to this work.

Additional supporting information can be found online in the Supporting Information section

a minimum residue separation to enhance visualization while preserving essential structural information.

Despite aligning with the common understanding that cross-links between distant residues in the sequence or between different proteins are fundamental in XL-MS, we hypothesize that short-range cross-links may still contain hidden structural information. Numerous XL-MS datasets from system-wide experiments are publicly available. In this study, we analyze short-range cross-links from these datasets by mapping them onto AlphaFold-predicted structures at a proteome-wide scale. Our analysis indicates that the spacing between identical residues depends on the local secondary structure and, to some extent, appears to be encoded at the genomic level. Upon normalization for this context-dependent amino acid distribution, clear secondary structure patterns emerge from experimental short-range cross-links. These patterns are determined by mutual orientation and spatial distances of side chains within distinct structural elements. Accurate interpretation of these observations requires precise geodesic measurements of solvent-accessible surface distances (SASD).[14–19] Therefore, we refined existing computational tools to accurately represent the cross-linker pose. Moreover, the residue composition surrounding cross-linked sites reveals context-dependent characteristics, uncovering a specific protein grammar. We leveraged our short-range cross-link atlas to benchmark AlphaFold protein models across various predicted local distance difference test (pLDDT) scores, a per-residue measure of local confidence.[20,21] Additionally, this atlas provides a novel quality-control framework for XL-MS experiments. Specifically, X-SPAN allows distinguishing genuine structural distributions from potential experimental artifacts.

Results and Discussion

Amino Acids Spacing in Secondary Structure Elements

The spacing between pairs of identical amino acids in secondary structural elements was examined across the proteomes of *Homo sapiens* (UP000005640), *Drosophila melanogaster* (UP000000803), and *Escherichia coli* (UP000000625) (Figures 1 and S1, S2). Secondary structural elements, specifically continuous α-helices, coils, β-strands, and mixed elements, were extracted from AlphaFold-predicted models of entire proteomes[22] (Supporting Information). Subsequently, we computed the spacing between identical residues and analyzed their frequency distributions for all proteinogenic amino acids (Figures 1 and S2). Our analysis revealed that genomes encode distinct preferred spacings for amino acid pairs, providing a glimpse into peptide folding upon translation. For instance, acidic and hydrophobic residues exhibit a characteristic spacing of approximately 3.6 residues within sequences encoding α-helices, reflecting helix amphipathicity. Additionally, distinct sequence motifs within secondary structural elements can be identified from residue pair distributions. Notably, the C2H2 zinc finger motif[23] is characterized by a prominent HxxxH and CxxC peak in helices and coils, respectively (Figure 1). Another example is the ubiquitous K homology domain, marked by the conserved motif VIGxxGxxI[24] within helices.

A central aspect of our analysis focuses on the spacing between lysine residues, the amino acids most frequently targeted in proteome-wide XL-MS experiments. Notably, lysine spacing does not display strongly distinct patterns across different secondary structural elements. Instead, it demonstrates a general decreasing frequency with increased residue spacing, aligning with the length distributions of continuous secondary structural elements within the proteome. This decline is particularly pronounced in β-strands, which typically consist of fewer residues due to their extended structure, and lysine residues are comparatively less abundant within β-strands compared to other elements (Figure S1). In contrast, α-helices, coils, and mixed elements share a more gradual frequency decline. We further investigated whether cross-linked lysine spacing deviates from this general proteomic trend. Such deviations could indicate differential cross-linker reactivity within secondary structural elements, rather than merely reflecting the underlying lysine distribution. Identifying these differences is essential for determining if cross-linking data offer additional structural insights beyond the general proteomic organization.

The sequence composition patterns that arise from secondary structural constraints are ultimately written into the genome and observable across large-scale datasets. Here we provide a systems-level view of this phenomenon, showing how natural selection over evolutionary timescales has shaped protein sequences to reliably fold into functional secondary embedding these amino acid patterns.

Short-Range Cross-Link Analysis

We analyzed 12 publicly available system-wide XL-MS datasets[11,25–32] (Table S1), encompassing a total of 658432 cross-links from four different organisms. These datasets include both amine-reactive and photo-reactive cross-linkers with varying spacer lengths. These 12 datasets were selected because each contained at least 400 unique short-range cross-links, a threshold that ensures sufficient sampling of the distance distribution. Cysteine-reactive cross-linkers were excluded from the analysis, as the presence of disulfide bonds cannot be reliably accounted for at the proteome-wide level, preventing a meaningful comparison between the experimental cross-links and the expected C─C distance distribution.

Our analysis was carried out using X-SPAN, a Python-based software that annotates secondary structural elements for short-range cross-links based on AlphaFold protein models. X-SPAN accepts .mzIdentML 1.3, .zhrm, and .xlsx files as inputs and provides a graphical user interface (GUI) for ease of use. The workflow of the X-SPAN algorithm is outlined in Table 1, and the software is freely accessible at https://github.com/IacobucciLab/X-SPAN. X-SPAN categorizes all short-range cross-links into four groups (continuous α-helices, β-strands, and coils, and mixed secondary structure elements) and plots the distribution of cross-link frequencies as a function of reacted residue spacing. Results from

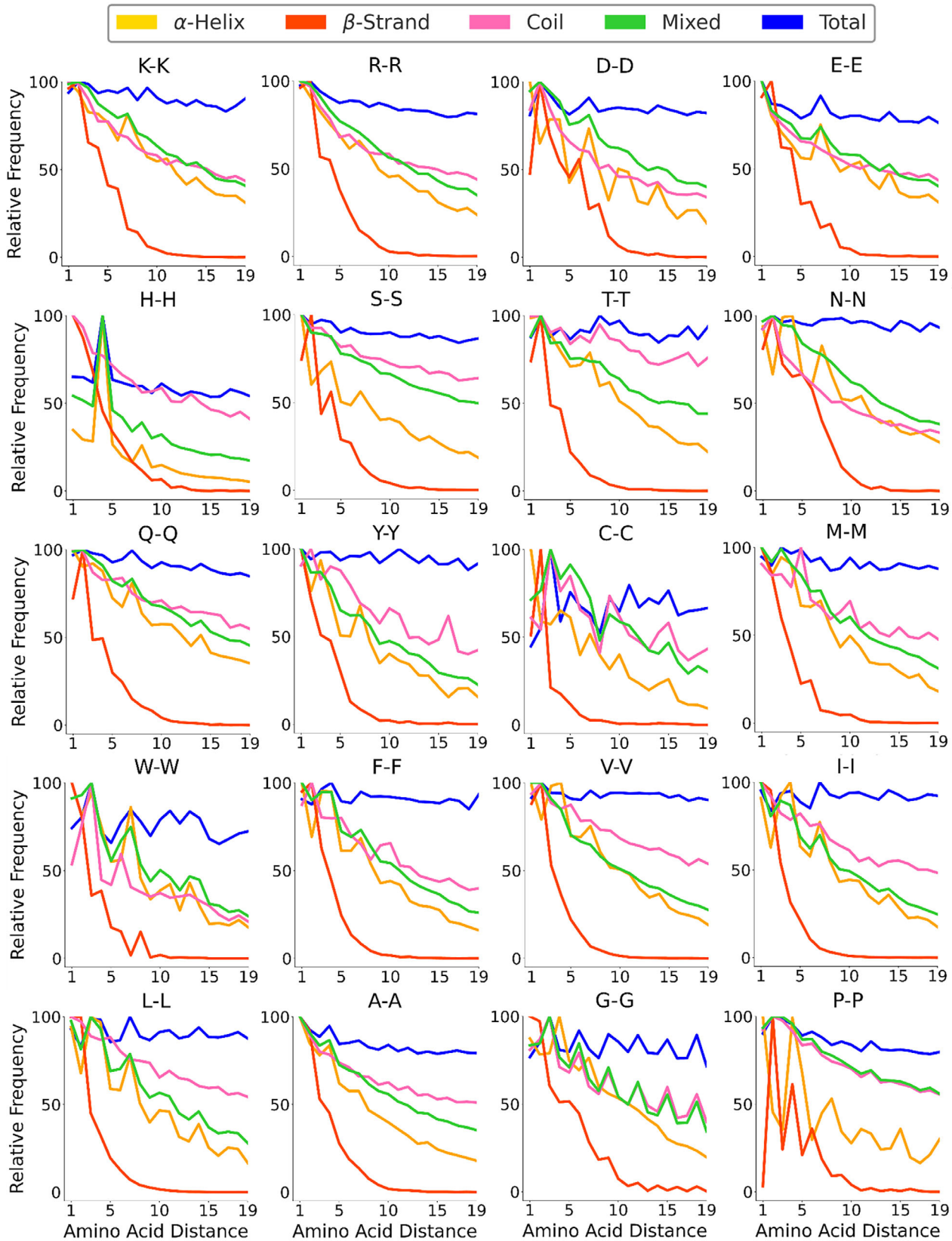

Figure 1. Relative frequency of pairwise amino acid distances across the human proteome. Each plot corresponds to an amino acid pair (e.g., K-K for lysine). The x-axis represents the amino acid spacing, while the y-axis represents the relative frequency normalized within each structural category. Curves are color-coded by secondary structure: α-helix (yellow), β-strand (red), coil (pink), mixed elements (green), and whole proteome (blue). Charged residues are plotted in the first row followed by polar and hydrophobic residues.

Table 1: Concise description of X-SPAN workflow. Each module of the pipeline is described, along with the number of cross-links retained after each processing step.

Module	Description	Cross-links
Data acquisition	Reading the formatted input file of cross-links (.mzIdentML 1.3, .zhrm, or .xlsx)	658432
Sequence Proximity Filtering	Filtering for intra-protein cross-links with a maximum spacing of 19 amino acids in the primary sequence.	78985
AlphaFold Annotation	Downloading AlphaFold-predicted model of all proteins with at least one short-range cross-link. Extracting the secondary structure element[a)] and the pLDDT score of all residues between the two cross-linked sites.	78985
Classification of Cross-Links	Categorizing cross-links into four major categories: continuous[b)] α-helix, β-strand, or coil and mixed[c)] elements.	78985
pLDDT Score Filtering	Filtering cross-links based on the pLDDT score[d)] of all residues between the two cross-linked sites.	48748
Data visualization	Plotting the distribution of cross-link frequencies as a function of sequence span for each structural category.	48748

[a)] Secondary structure assignment is performed using the secondary structure determination method (dss) within PyMOL. For continuous elements, all residues between the cross-linked sites must be assigned to the same structural element. [b)] Cross-links with all residues assigned to the same structural element, either a single α-helix, β-strand, or coil. [c)] Cross-links with residues assigned to more than one structure element. Application of the DSSP algorithm[33] in place of DSS did not result in any substantial differences in the outcomes (Figure S3). [d)] Where not stated differently we retained only cross-links with high confidence (>80) pLDDT score for α-helix and β-strand while any pLDDT score is accepted in case of coils (due to the intrinsic lower pLDDT score typically associated with flexible and disordered regions) and mixed elements.

amino-reactive and photo-activatable cross-linkers are presented in Figures 2 and S4, respectively.

α-Helix

Among short-range cross-links, continuous α-helices represent the most common structural element (27%, Figure S1), independent of cross-linker reactivity or length. The distribution of cross-links by residue spacing strikingly differs from other secondary structures (Figures 2 and S3) and from overall lysine spacing across the proteome (Figures 1 and S2). Certain residue spacings are highly favored for cross-linking, while others appear almost prohibited. Their distribution follows an oscillating pattern modulated by a decreasing exponential trend (Figures 2 and S3). The observed experimental periodicity of approximately 3.6 residues matches the helical pitch characteristic of α-helices. Meanwhile, the exponential curve's slope reflects the length of the cross-linker utilized; shorter cross-linkers produce steeper declines due to their limited ability to reach distant residues.[34] Conversely, longer cross-linkers can bridge residues across multiple helical turns without reducing spatial resolution. Surprisingly, shorter cross-linkers do not capture α-helix structures more effectively than longer cross-linkers, as the spacing between maxima and minima remains consistent across different linker lengths (Figures 2 and S3). This may result from unfavorable entropic contributions associated with longer cross-linkers wrapping around helical structures.[35] It should be noted that our use of the term "short-range" cross-links refers specifically to sequence proximity of residues (less than 20 amino acids apart), differing from previous works,[34] where "short-range" describes the physical length of the cross-linker spacer.

We also examined patterns in non-specific photoactivatable cross-linkers[28] (Figure S4). Such reagents provide high-density structural data[36–38] independent of protein sequence, which could improve the detection of α-helix pattern compared to lysine-selective cross-linkers (Figure 2). Unexpectedly, although photocross-linkers still capture helical periodicity, they exhibit a reduction in the amplitude of the oscillations, evidenced by decreased distances between consecutive maxima and minima. This suggests that, for short-range cross-links, potential benefits of photoreactive cross-linkers might be counterbalanced by their lower precision in cross-linking site identification.[39,40] While this lower accuracy has minimal implications for classical structural interpretations from XL-MS data, it becomes critical for identifying secondary structural elements through short-range cross-links.

Coils

Coils are protein secondary structure elements characterized by irregular and flexible shapes due to the absence of a consistent hydrogen-bonding network. Without a regular hydrogen-bonding pattern, the spacing between cross-linked residues lacks periodicity. Additionally, as the sequence span between cross-linked residues increases, the probability of cross-link formation declines (Figures 2 and S3). Coils account for 33% of all short-range cross-links (Figure S1) and exhibit a gradual exponential decrease in cross-link frequency, reaching a maximum span of approximately 19 residues. Although coils are generally less compact than α-helices, their flexibility occasionally enables longer-range cross-links by bringing distant residues into proximity, albeit at much lower frequencies compared to more structured regions.

β-Strand

Cross-links involving lysine residues in β-strands are notably less frequent compared to other structural elements. This is in agreement with previous reports in which low occurrence of lysine was reported for β-strands.[41,42] Our findings show that short-range cross-links within continuous β-strands are even rarer, representing only 1% of the observed cases

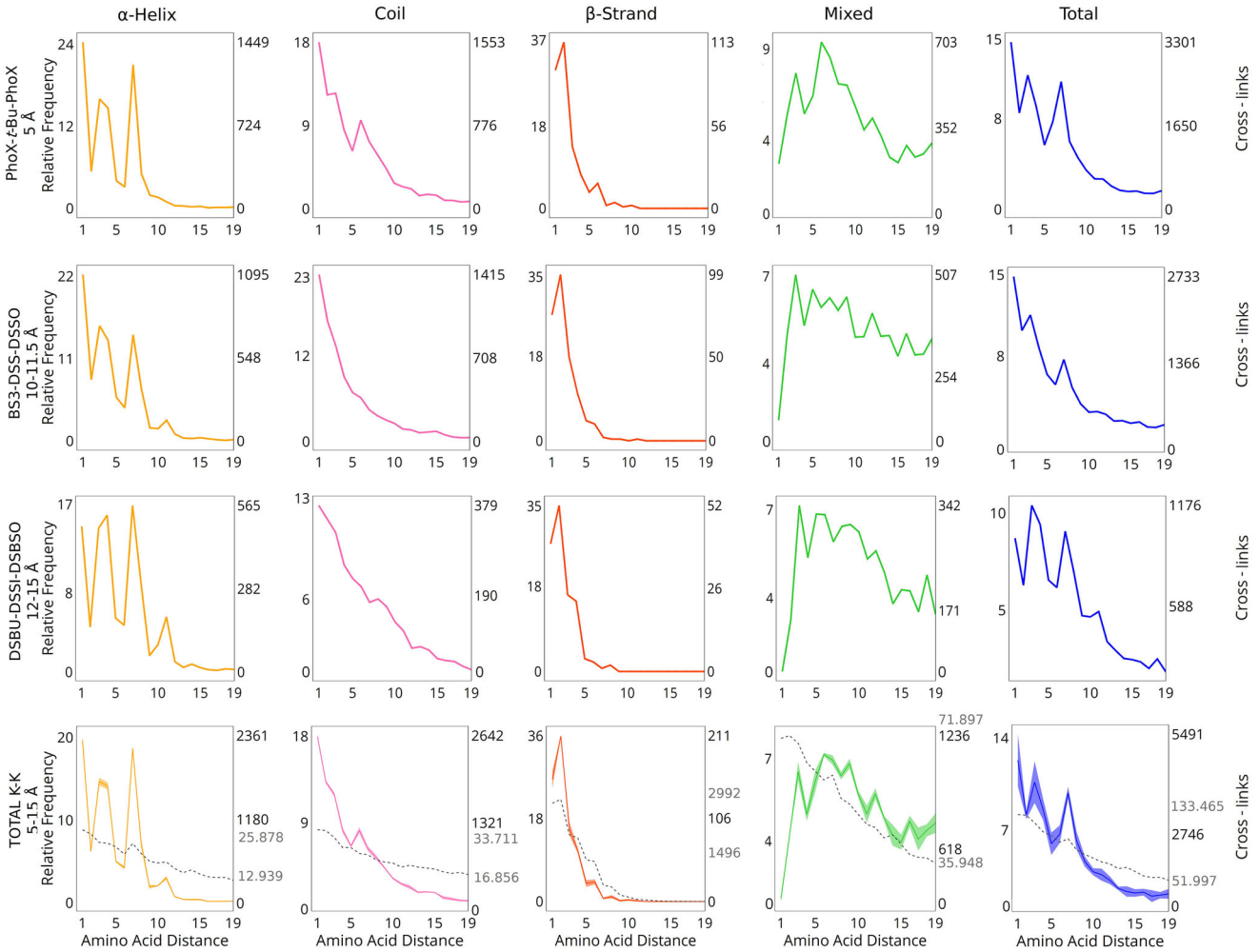

Figure 2. Relative frequency of K—K cross-links at increasing amino acid spacing. The *x*-axis represents the distance between amino acid pairs. The *y*-axis reports both the relative frequency of cross-links (left), normalized within each structural category to a maximum of 100, and the absolute counts (right). The four classes of structural elements are presented in separate columns and are color coded as follows: α-helix–yellow lines, β-strand–red lines, coil–pink lines, mixed elements–green lines, and total–blue lines. Distributions from cross-linkers with similar spacer length are presented in separate lines. Cumulative data are presented in the bottom row. In addition to the 1% false discovery rate (FDR) associated to each dataset, the shaded areas around the cumulative curves in the bottom panels represent the standard error, estimated by dividing the cumulative dataset into three equal random subsets. Additionally, the bottom panels display the relative frequency of pairwise amino acid distances across the human proteome for comparison. Cross-link redundancy has been removed within each dataset.

(Figure S1). The limited presence of lysine residues in β-strands (Figure S1) contributes significantly to their scarcity. Despite their low occurrence, specific preferred lysine distances are identifiable within experimental cross-links. In β-strands, side chains alternate orientation, projecting residues in opposite directions. This alternating pattern typically prohibits cross-link formation between residues with consecutive even-to-odd numbering. Therefore, a periodic distribution similar to that seen in α-helices could be expected, albeit with higher frequency. However, experimentally observed cross-links within β-strands exhibit a largely random distribution, decreasing rapidly with increased residue spacing, closely resembling the proteome-wide distribution (Figures 1, 2, S2, and S3). This distribution pattern can be explained by the elongated and twisted nature of β-strands, where side-chain distances are non-uniform, enabling cross-linking at diverse spacings. Notably, a distinct peak occurs at a residue spacing of two, coinciding with optimal side-chain orientation despite the strand's torsion. Interestingly, this spacing corresponds to a minimum observed in α-helix cross-link distributions.

Mixed Elements

Cross-links within mixed structural elements exhibit a distinctive hump-shaped frequency distribution. Compared to other structural elements, their frequency is relatively low at short residue spacings but increases at intermediate distances before gradually decreasing. Unlike the distributions observed in other structural elements, cross-link frequency in mixed elements does not approach zero at longer spacings but stabilizes at a low plateau. At these extended distances, cross-links in mixed elements become predominant, likely due to compact structural motifs such as hairpins, loops, and

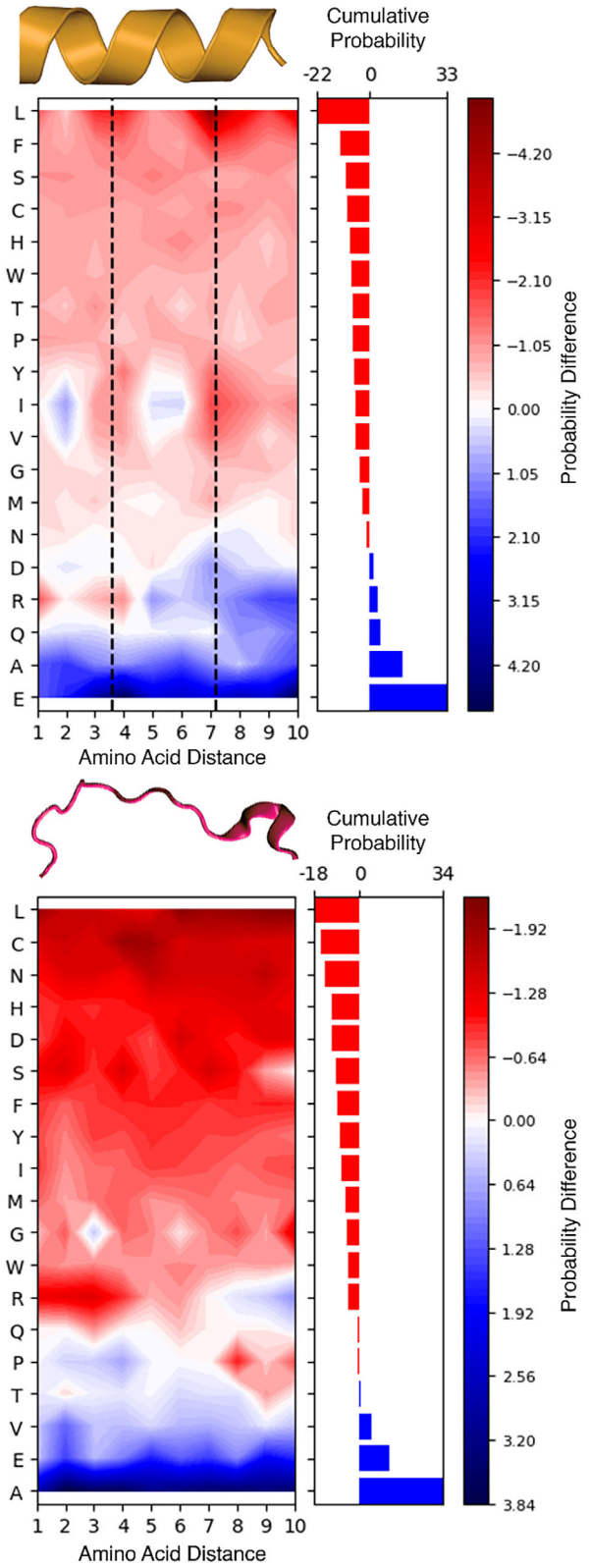

Figure 3. Influence of neighboring amino acids on lysine cross-linking propensity within α-helices (upper panel) and coils (bottom panel). Each heatmap represents the percentage difference in amino acid composition surrounding cross-linked lysine compared to the background proteome. The *y*-axis lists amino acids, while the *x*-axis indicates their sequence distance from a cross-linked lysine. Amino acids are ranked based on their overall effect on lysine cross-linking, with red indicating cross-linking inhibition and blue indicating cross-linking promotion. The right-side histogram illustrates this cumulative effect across all positions. Dashed lines mark the helical pitch, highlighting periodic cross-linking patterns in α-helices.

turns, which bring distant residues into close spatial proximity. The short-range cross-link distribution (Figures 2 and S3) differs significantly from overall lysine spacing within mixed elements at the proteome-wide level (Figures 1 and S2). Notably, the occurrence of cross-links between consecutive residues strongly suggests these residues are part of either an α-helix or a coil.

Taken together, our results indicate that short-range cross-links serve as diagnostic markers for secondary structural elements. By analyzing the distribution of short-range cross-links obtained from all amino-reactive cross-linkers (Figure 2, bottom panel), we estimated the probabilities of peptides adopting specific secondary structures (Figure S5). For instance, at a spacing of two residues, the probability of the peptide forming a coil is 56%, nearly three times greater than the probabilities for α-helices or mixed elements. Conversely, at a residue spacing of seven, the likelihood of forming a continuous α-helix is approximately 50%, nearly double that observed for coils or mixed elements. Therefore, detecting short-range cross-links within an integrative modeling framework could significantly enhance inference of underlying peptide secondary structures.

To interpret these structural patterns (Figure 2), we refined the Biobox[43] Python package to measure geodesic solvent-accessible surface distances (SASD) between reactive side-chain atoms (e.g., ε-nitrogen atom of lysine) while considering the cross-linker thickness (see Supporting Information). Commonly used Euclidean Cα–Cα distances systematically underestimate short-range cross-link distances due to neglecting surface curvature and local structural features (Figures S6 and S7). Our SASD method more realistically represents the spatial pathways accessible to the cross-linker (Figures S6 and S7). This geodesic approach effectively captures the periodic spacing characteristic of α-helices, aligning with experimental observations (Figures 2, S3, S5, and S6).

Context-Dependent Cross-Link Grammar

We also investigated whether the local amino acid environment affects the propensity of lysine to be cross-linked within different secondary structure elements. The composition of neighboring residues might promote or inhibit lysine cross-linking beyond structural constraints imposed by secondary structures. First, we analyzed the natural distribution of amino acids at varying distances around lysine within continuous secondary structures of the human proteome. We then compared these background distributions with those observed in our cross-link atlas (Figure 3). In α-helices, glutamic acid promotes lysine cross-linking when positioned at the first helical turn, at a spacing of three to four residues. At this distance, these residues reach the minimal geodesic separation optimal for hydrogen bonding, thus enhancing lysine

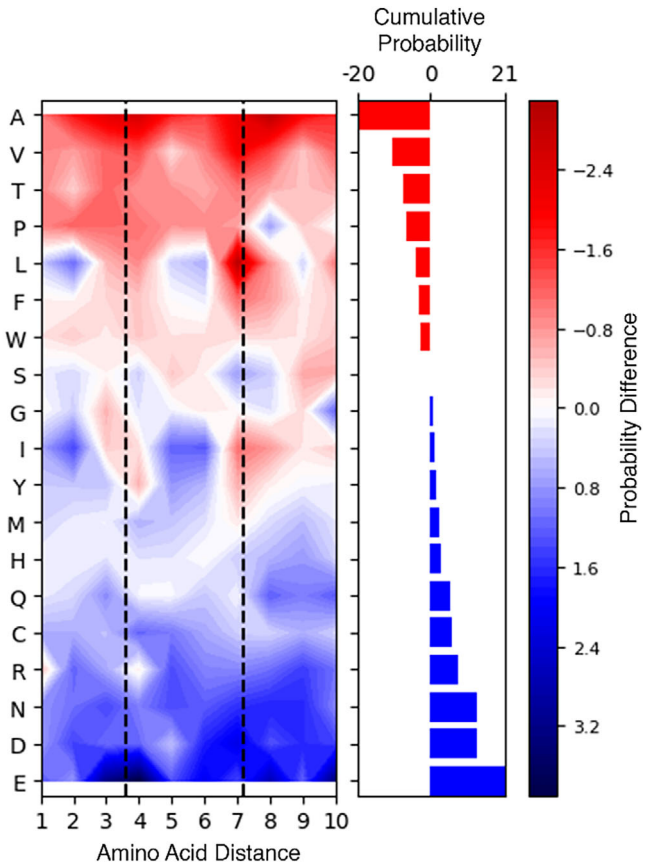

Figure 4. Influence of neighboring amino acids on cross-linked α-helix to coil recognition. The heatmap represents the percentage difference in amino acid composition between α-helices and coils. The *y*-axis lists amino acids, while the *x*-axis indicates their sequence distance from a cross-linked lysine. Amino acids are ranked based on their diagnostic capability, with red indicating residues more abundant in coils and blue indicating residues enriched in α-helices. The right-side histogram illustrates this cumulative effect across all positions. Dashed lines mark the helical pitch, highlighting periodic cross-linking patterns in α-helices.

nucleophilicity. This promoting effect is weaker for glutamine and aspartic acid due to their lack of charge and shorter side chain, respectively. Arginine, possessing a permanent positive charge, initially disfavors cross-linking within the first helical turn but reverses its effect beyond this region. Alanine facilitates lysine cross-linking, particularly when oriented toward the opposite face of the helix. Conversely, bulky hydrophobic residues, notably leucine, positioned within the first two helical turns, inhibit lysine cross-linking. Alanine and glutamic acid are also enriched around cross-linked lysine in coils, though without clear positional preferences. In coils, the inhibitory effects of certain residues, like leucine and cysteine, are weaker and less position-specific compared to α-helices. This aligns with the more extended and less structured nature of coils. Due to the limited experimental data available, the cross-link grammar of β-strands could not be fully characterized (Figure S8).

Finally, we analyzed the cross-link grammar distinguishing α-helices from coils (Figure 4). Residues positioned at the third, fourth, and seventh positions from lysine, corresponding to the first and second helical turns, emerged as diagnostic markers. Specifically, the presence of glutamic acid or aspartic acid at these positions indicates an α-helical structure, whereas hydrophobic residues typically suggest a coil structure.

The limited availability of experimental data prevented us from drawing conclusions from the comparison of the amino acid composition of β-strands versus coils, although this analysis is reported in Figure S9.

Benchmarking AlphaFold-predicted Secondary Structure Elements

Our short-range cross-link atlas provides a novel framework to evaluate the pLDDT score metric on a proteome-wide level. This investigation is based on structural data obtained in the native protein environment and is free from biases associated with the ability to express, purify, and determine high-resolution spectroscopic structures of proteins. We compare the cross-link distributions in secondary structure elements predicted within three pLDDT score ranges (0–60 for low local confidence, 60–80 for medium local confidence, and 80–100 for high local confidence) (Figure 5). The ranges are defined to ensure a sufficient number of cross-links for analysis. For the same reason, we combine all amine-reactive cross-linker datasets and exclude photo-reactive cross-linker datasets from this analysis.

Even at medium pLDDT confidence, the distinct patterns associated with α-helices were largely preserved. This indicates that α-helices with pLDDT scores greater than 60 are reliably predicted. Below this threshold, the α-helical pattern attenuated, and the cross-link distribution resembled that of coils with pLDDT scores above 60. This suggests that a significant portion of predicted α-helices at lower scores may be discontinuous or coil-like. For coils with pLDDT scores below 60, the cross-link distribution flattened, reflecting contributions from disordered regions with higher conformational flexibility, allowing cross-linking between more distant residues. β-strands appeared more sensitive to pLDDT scores, as their characteristic peak at a two-residue spacing disappeared at scores below 80, indicating that these segments might be less structured than predicted. Our analysis provides interpretative depth to the generic confidence metric (pLDDT scores). By correlating experimentally observed short-range cross-links directly to secondary structure elements, we contextualize these scores, to obtain concrete structural insights. Thus, short-range cross-links serve as powerful interpretative tools of AlphaFold's confidence pLDDT score, enhancing its practical utility in structural biology.

As a practical application, we evaluated the structural predictions for the human ORC4 subunit of the Origin Recognition Complex. This protein was selected because the AlphaFold model complements the available X-ray structure of the complex (PDB ID: 5UJ7[44]) by providing structural information for the unresolved N-terminal region of ORC4 (Figure S10). We applied short-range cross-links from our

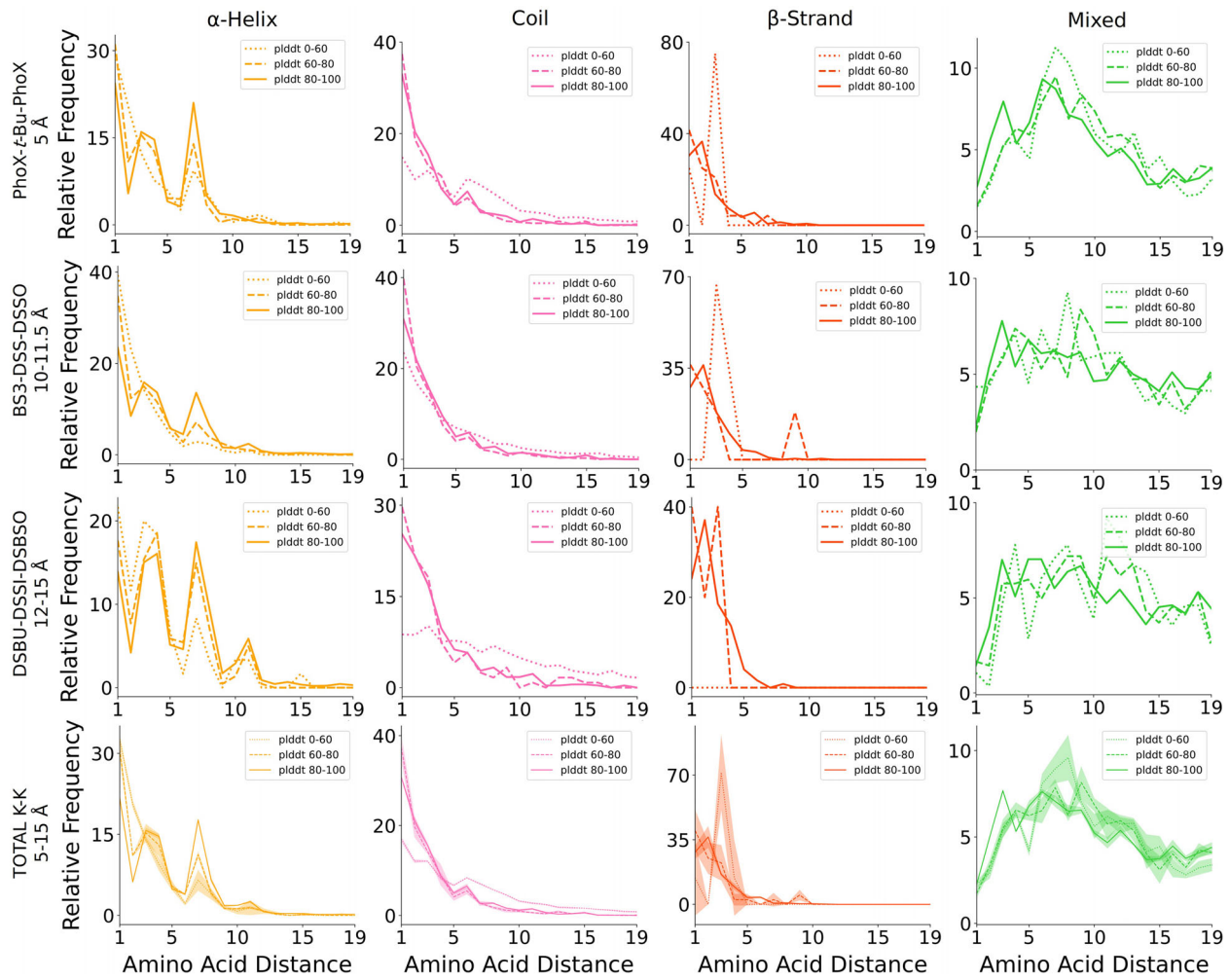

Figure 5. Distribution of K—K cross-links at different pLDDT scores. The x-axis represents the distance between amino acid pairs. The y-axis reports the relative frequency of cross-links, normalized within each structural category to a maximum of 100. The four classes of structural elements are presented in separate columns and are color coded as follows: α-helix–yellow lines, β-strand–red lines, coil–pink lines, and mixed elements–green lines. Distributions from cross-linkers with similar spacer length are presented in separate lines. Cumulative data are presented in the bottom row. In addition to the 1% false discovery rate (FDR) associated to each dataset, the shaded areas around the cumulative curves in the bottom panels represent the standard error, estimated by dividing the cumulative dataset into three equal random subsets. pLDDT score ranges are presented as dotted line (0–60), dashed line (60–80), and solid line (80–100). Cross-link redundancy has been removed within each dataset.

atlas to benchmark the AlphaFold prediction. Two cross-links, K45–K52 and K204–K207, are located within high-confidence α-helices and correspond to three and seven residue spacing. They match the maxima of our α-helix distribution (Figures 2 and S5), thus supporting the assigned secondary structure. In contrast, the K5–K7 cross-link at the N-terminus, found in a predicted low-confidence helix, corresponds to a spacing of two residues, precisely a minimum in the α-helix cross-link distribution. This suggests a coil-like conformation in this region, illustrating how even the shortest cross-links, often considered uninformative, can contribute to the study of protein structure.

Secondary Structure-Based Validation Metric

Validating cross-links in system-wide XL-MS experiments is critical to ensuring the reliability of structural insights. The most commonly used validation approach is structure-based validation, which evaluates cross-links by their agreement with known 3D structures of protein complexes. Although effective for targeted studies of individual proteins and complexes, this method has limitations for proteome-wide XL-MS experiments.[45] Alternatively, a STRING-based scoring approach has been proposed for validating protein–protein interaction (PPI) studies.[11] While meaningful, comparing experimentally detected PPIs against the background distribution in the STRING database does not provide a definitive validation criterion.

Here, we introduce a new metric based on the distribution of short-range cross-links within secondary structure elements as an additional layer of cross-link validation. We demonstrated that true positive short-range cross-links follow predictable patterns dictated by the spatial constraints of α-helices, β-strands, and coils (Figures 2 and S3). In contrast, we expect false positive short-range cross-links to follow

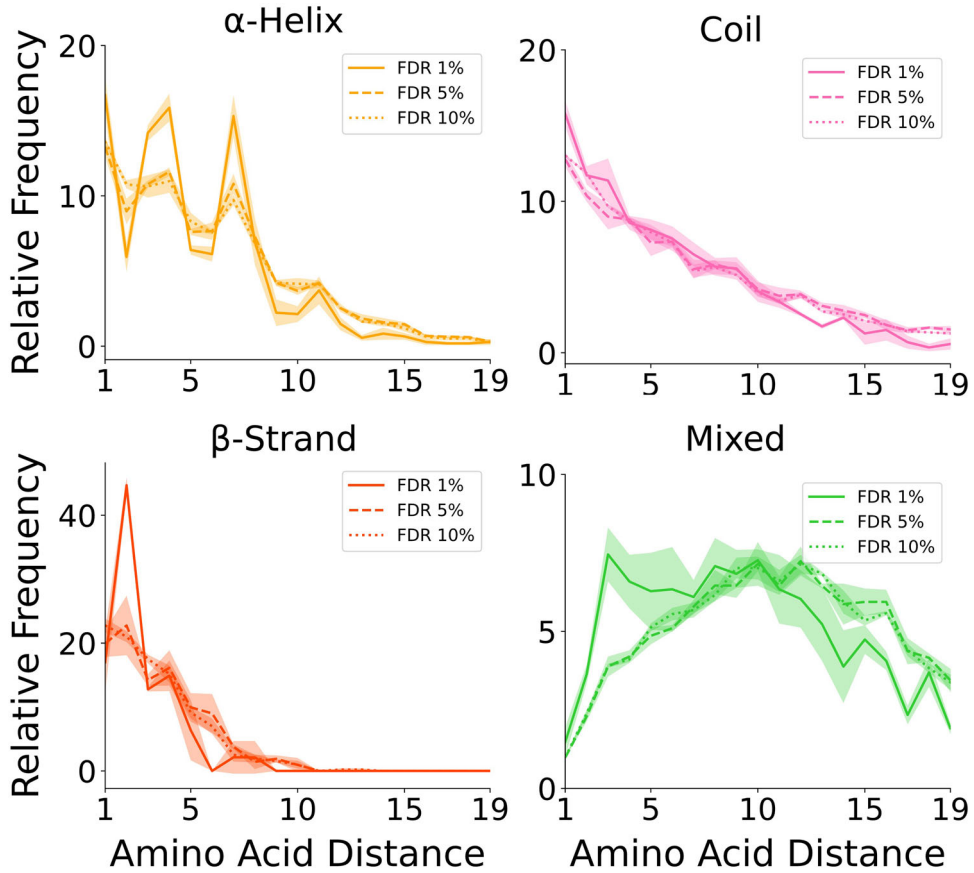

Figure 6. Distribution of DSSI cross-links at different FDR. The *x*-axis represents the distance between amino acid pairs. The *y*-axis reports the relative frequency of cross-links (left), normalized within each structural category to a maximum of 100. The four classes of structural elements are color coded as follows: α-helix–yellow lines, β-strand–red lines, coil–pink lines, and mixed elements–green lines. Different FDR thresholds are represented as dotted (10%), dashed (5%), and solid (1%) lines. The shaded areas around the cumulative curves represent the standard error, estimated by dividing the cumulative dataset into three equal random subsets. Cross-link redundancy has been removed within each dataset.

the natural occurrence of residue pairs within the proteome (Figure 1 and S2). We hypothesized that datasets with a higher proportion of false positives, or those obtained under non-native conditions, would exhibit altered spacing distributions of short-range cross-links.

To test this hypothesis, we selected our recent DSSI[30] cross-link dataset from HEK293T cell lysates (1% FDR at the PSM level) and generated three additional cross-link sets with drastically different quality levels. These sets were obtained by running the MeroX search engine using the original published settings but applying three progressively less stringent FDR criteria (1%, 5%, and 10%; see Supporting Information). We then applied X-SPAN, as defined in Table 1, obtaining 3681 short-range cross-links at 1% FDR, 17 597 at 5% FDR, and 41 422 at 10% FDR across the four categories of secondary structure elements (Figure 6). Notably, the distribution of lower-quality cross-link datasets, which contain a higher proportion of false positives, deviates significantly from that of the original 1% FDR dataset. Since false positives are randomly distributed, they cancel the characteristic α-helix periodic pattern and reduce the steepness of the exponential decay observed in coils and β-strand. This behavior closely resembles what is observed at low-confidence pLDDT scores. These findings demonstrate that secondary structure-based validation of short-range cross-links is an effective approach for distinguishing datasets of varying quality. We propose that X-SPAN can provide an additional metric to define the overall quality of a proteome-wide cross-link dataset, which should align with the distribution patterns observed in our cross-link atlas.

Conclusion

Short-range cross-links, traditionally considered less informative and frequently overlooked in XL-MS studies, have been shown here to carry intrinsic structural information. By systematically analyzing short-range cross-links across multiple proteomes, we demonstrated their strong correlation with secondary structural elements. Introducing X-SPAN, we integrated large, publicly available XL-MS datasets with AlphaFold-predicted proteomes, revealing distinct cross-linking patterns reflective of protein secondary structure organization. α-helices exhibit periodic cross-linking patterns matching their helical pitch, whereas coils and β-strands display nearly monotonic distributions influenced by their

flexibility and extended conformation, respectively. Additionally, we identified a context-dependent protein grammar of cross-links, showing that local amino acid composition influences cross-linking probability. The prevalence of specific residues within secondary structures further supports the potential of short-range cross-links to assist secondary structure inference in integrative studies.

Our results demonstrate that short-range cross-links provide a useful benchmarking tool for evaluating AlphaFold's local accuracy. By comparing experimental cross-link distributions with predicted structural elements across different pLDDT scores, we propose a data-driven method to assess AlphaFold model reliability at a local level. Furthermore, short-range cross-links introduce a novel quality control strategy for XL-MS experiments, distinguishing native structural distributions from potential artifacts.

The short-range cross-link atlas developed here, along with the broader application of our approach, may become a valuable resource for investigating secondary structure transitions and their roles in protein function and allosteric regulation.

Supporting Information

The 12 datasets analyzed in this study were downloaded from the PRIDE database, and detailed information is provided in Table S1. The short-range cross-links extracted from all datasets are available as a .txt file. X-SPAN, along with all scripts used in this work, is freely accessible at https://github.com/IacobucciLab/X-SPAN.

Acknowledgements

C.I. acknowledges financial support by the Italian Ministry of University and Research (MUR) grant PRIN 2022 – Project 20225HNCZK – Master CUP G53D23002450006 – CUP E53D23007110006, by the European Union NextGeneration EU grant PRIN 2022 PNRR – Project P20224WAME – CUP E53D23021440001, and by the European Union NextGeneration EU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 – VITALITY – CUP E13C22001060006.

Open access publishing facilitated by Universita degli Studi dell'Aquila, as part of the Wiley - CRUI-CARE agreement.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in [GitHub] at [10.5281/zenodo.15120243], reference number [15120243]. These data were derived from the following resources available in the public domain: [Resource 1], https://github.com/IacobucciLab/X-SPAN; [Resource 2], https://doi.org/10.5281/zenodo.15120243.

Keywords: Chemical proteomics · Cross-linking · Mass spectrometry · Structural biology · Structural proteomics

[1] F. J. O'Reilly, J. Rappsilber, *Nat. Struct. Mol. Biol.* **2018**, *25*, 1000–1008.
[2] A. Graziadei, J. Rappsilber, *Structure* **2022**, *30*, 37–54.
[3] L. Botticelli, A. A. Bakhtina, N. K. Kaiser, A. Keller, S. McNutt, J. E. Bruce, F. Chu, *Curr. Opin. Struct. Biol.* **2024**, *87*, 102872.
[4] C. Yu, L. Huang, *Curr. Opin. Chem. Biol.* **2023**, *76*, 102357.
[5] M. K. O. Klykov, B. Carragher, A. J. R. Heck, A. J. Noble, R. A. Scheltema, *Mol. Cell* **2022**, *82*, 285–303.
[6] L. Piersimoni, P. L. Kastritis, C. Arlt, A. Sinz, *Chem. Rev.* **2022**, *122*, 7500–7531.
[7] H. Stark, *Methods Enzymol.* **2010**, *481*, 109–126.
[8] S. Neyer, M. Kunz, C. Geiss, M. Hantsche, V.-V. Hodirnau, A. Seybert, C. Engel, M. Scheffer, P. Cramer, A. S. Frangakis, *Nature* **2016**, *540*, 607–610.
[9] F. Schuhmann, K. C. Akkaya, D. Puchkov, S. Hohensee, M. Lehmann, F. Liu, W. Pezeshkian, *Angew. Chem. Int. Ed. Engl.* **2025**, *64*, e202417804.
[10] B. Bogdanow, I. Gruska, L. Mühlberg, J. Protze, S. Hohensee, B. Vetter, J. B. Bosse, M. Lehmann, M. Sadeghi, L. Wiebusch, F. Liu, *Nat. Microbiol.* **2023**, *8*, 1732–1747.
[11] M. Götze, C. Iacobucci, C. H. Ihling, A. Sinz, *Anal. Chem.* **2019**, *91*, 10236–10244.
[12] C. W. Combe, M. Graham, L. Kolbowski, L. Fischer, J. Rappsilber, *J. Mol. Biol.* **2024** *436*, 168656.
[13] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, J. A. Vizcaíno, *Nucleic Acids Res.* **2022**, *50*, D543–D552.
[14] A. Kahraman, F. Herzog, A. Leitner, G. Rosenberger, R. Aebersold, L. Malmström, *PLoS One* **2013**, *8*, e73411.
[15] J. M. A. Bullock, J. Schwab, K. Thalassinos, M. Topf, *Mol. Cell. Proteomics* **2016**, *15*, 2491–2500.
[16] M. T. Degiacomi, C. Schmidt, A. J. Baldwin, J. L. P. Benesch, *Structure* **2017**, *25*, 1751–1757.
[17] Z. Gong, Z. Liu, X. Dong, Y. H. Ding, M. Q. Dong, C. Tang, *Biophys. Rep.* **2017**, *3*, 100–108.
[18] A. Kahraman, L. Malmström, R. Aebersold, *Bioinformatics* **2011**, *27*, 2163–2164.
[19] J. M. A. Bullock, K. Thalassinos, M. Topf, *Bioinformatics* **2018**, *34*, 3584–3585.
[20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, et al., *Nature* **2021**, *596*, 583–589.
[21] K. Tunyasuvunakool, J. Adler, Ž. Bodrikov, T. Green, M. Figurnov, O. Ronneberger, R. Bates, S. Bridgland, A. Cowie, N. A. Paschalidis, A. Žídek, R. Evans, A. W. Senior, D. Hassabis, J. Jumper, *Nature* **2021**, *596*, 590–596.
[22] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, D. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, S. Green, M. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, *50*, D439–D444.

[23] S. A. Wolfe, L. Nekludova, C. O. Pabo, *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 183–212.
[24] N. V. Grishin, *Nucleic Acids Res.* **2001**, *29*, 638–643.
[25] M. A. Gonzalez-Lozano, F. Koopmans, P. F. Sullivan, J. Protze, G. Krause, M. Verhage, F. L. K. W. Li, A. B. Smit, *Sci. Adv.* **2020**, *6*, eaax5783.
[26] S. H. Giese, L. R. Sinn, F. Wegner, J. Rappsilber, *Nat. Commun.* **2021**, *12*, 3237.
[27] P.-L. Jiang, C. Wang, A. Diehl, R. Viner, C. Etienne, P. Nandhikonda, L. Foster, R. D. Bomgarden, F. Liu, *Angew Chem. Int. Ed. Engl.* **2022**, *61*, e202113937.
[28] A. M. Faustino, P. Sharma, E. Manriquez-Sandoval, D. Yadav, S. D. Fried, *Anal. Chem.* **2023**, *95*, 10670–10685.
[29] F. Jiao, C. Yu, A. Wheat, L. Chen, T.-S. M. Lih, H. Zhang, L. Huang, *J. Proteome Res.* **2024** *23*, 3269–3279.
[30] A. Di Ianni, C. H. Ihling, T. Vranka, V. Matoušek, A. Sinz, C. Iacobucci, *JACS Au* **2024**, *4*, 2936–2943.
[31] A. R. M. Michael, B. C. Amaral, K. L. Ball, K. H. Eiriksson, D. C. Schriemer, *Nat. Commun.* **2024**, *15*, 8537.
[32] R. D. H. Ramazanova, T. I. Roumeliotis, J. C. Wright, J. S. Choudhary, *J. Proteome Res.* **2024** *23*, 5209–5220.
[33] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577–2637.
[34] N. I. Brodie, K. I. Popov, E. V. Petrotchenko, N. V. Dokholyan, C. H. Borchers, *Sci. Adv.* **2017** *3*, e1700479.
[35] Z. Gong, S.-X. Ye, C. Tang, *Structure* **2020** *28*, 1160–1167.e3.
[36] A. Belsom, M. Schneider, L. Fischer, O. Brock, J. Rappsilber, *Mol. Cell. Proteomics* **2016** *15*, 1105–1116.
[37] A. Belsom, M. Schneider, O. Brock, J. Rappsilber, *Trends Biochem. Sci.* **2016**, *41*, 564–567.
[38] M. Schneider, A. Belsom, J. Rappsilber, *Trends Biochem. Sci.* **2018** *43*, 157–169.
[39] C. Iacobucci, M. Götze, C. Piotrowski, C. Arlt, A. Rehkamp, C. Ihling, C. Hage, A. Sinz, *Anal. Chem.* **2018** *90*, 2805–2809.
[40] A. V. West, G. Muncipinto, H.-Y. Wu, A. C. Huang, M. T. Labenski, L. H. Jones, C. M. Woo, *J. Am. Chem. Soc.* **2021**, *143*, 6691–6700.
[41] K. Fujiwara, H. Toda, M. Ikeguchi, *BMC Struct. Biol.* **2012**, *12*, 18.
[42] M. Tsutsumi, J. M. Otaki, *J. Chem. Inf. Model.* **2011** *51*, 1457–1464.
[43] L. S. P. Rudden, S. C. Musson, J. L. P. Benesch, M. T. Degiacomi, *Bioinformatics* **2022** *38*, 1149–1151.
[44] A. Tocilj, K. F. On, Z. Yuan, J. Sun, E. Elkayam, H. Li, B. Stillman, L. Joshua-Tor, *Elife* **2017**, *6*, e20818.
[45] K. Yugandhar, T.-Y. Wang, S. D. Wierbowski, E. E. Shayhidin, H. Yu, *Nat. Methods* **2020**, *17*, 985–988.

Manuscript received: April 01, 2025
Revised manuscript received: May 26, 2025
Accepted manuscript online: May 27, 2025
Version of record online: June 02, 2025