New Bioinformatic Algorithms for Proteome-Wide Cross-Linking Mass Spectrometry

Micha J. Birklbauer
PhD Pre-Defense, 27. February 2025
Johannes Kepler University Linz, Austria

Supervisor
Stephan Winkler
Institute for Symbolic Artificial Intelligence, JKU Linz, Austria

Motivation

Proteins are the core components of life, they are integral to controlling cellular function, ultimately shaping the characteristics of any organism [1]. Proteomics is the study of all proteins aimed at understanding the pathological processes of diseases and other conditions that affect an organism's life [2], [3] which facilitates the subsequent study of discovering treatment options [4]. In order to uncover these processes, it is fundamental to determine the structure of proteins [5], [6] and their interaction partners [7], [8] in their native cellular environment. In this regard crosslinking mass spectrometry (XLMS) has become the method of choice in the past decades [9]. In XLMS the proteins of a sample are crosslinked with a chemical compound called the crosslinker, a molecule that covalently links two amino acids either within the same protein or between different proteins. Subsequently, the proteins are enzymatically cut into smaller parts called peptides and are analyzed by means of liquid chromatography tandem mass spectrometry (LC-MS/MS), creating one or more mass spectra per crosslinked species [10]. Bioinformatic algorithms which the field calls "crosslink search engines" (XL-DBSEs) are needed to identify which crosslinked peptides produced each spectrum. Even though a plethora of XL-DBSEs exits [11]–[14], each search engine comes with its unique strengths and weaknesses, explaining the on-going high demand for specialized software to make sense of XLMS data [15].

State of the art in computational crosslink identification

Several XL-DBSEs have been published in recent years [11]–[14], [16], all with the aim of simplifying the identification of crosslinked peptides. However, many major challenges have yet to be addressed [17], [18]:

Most XL-DBSEs only support a limited number of crosslink acquisition workflows, frequently that is single MS2-acquisition, even though multi-stage MS2-MS3-acquisition is among the most commonly used workflows in experiments [17]. Moreover, the two existing XL-DBSEs capable of analyzing MS2-MS3 data [12], [19] both lack in sensitivity and specificity [17], [20], [21] which are crucial for biological application. **Research Problem 1: This highlights the need for an accurate and robust computational approach to detect crosslinks in MS2-MS3-acquisition workflows.**

Crosslinkers can be categorized into two kind of compounds, cleavable and non-cleavable reagents [22], [23]. Non-cleavable crosslinkers are the most commonly used cross-linking reagents [24] but contrary to cleavable crosslink data, the analysis of mass spectra originating from non-cleavable crosslinks is much more computationally expensive as the search space grows with the square of the protein database, a phenomenon called the n-squared problem [25]. Historically this has limited non-cleavable crosslink experiments to small and low complexity samples. Nevertheless, the computational analysis of complex samples up to proteome-wide searches is of fundamental interest to the research community, for example to identify new pathways and potential drug targets [18]. **Research Problem 2: This underscores the need for an efficient search algorithm for identification of non-cleavable crosslinks.**

One of the most widely discussed topics in XLMS is the estimation of false positives within the reported results which is referred to as false discovery rate (FDR) estimation [26]. FDR estimation has been a large problem since the inception of XLMS and most XL-DBSEs report substantially

more false positives than what they estimate [20], [21]. This is a significant problem as the drawn biological conclusions may be wrong as a result. **Research Problem 3: Therefore, a search engine that can accurately estimate the false discovery rate is needed.**

Approach and results

Micha Birklbauer has been developing algorithms and software in the field of computational crosslinking proteomics for the past three years. He conducted his research as part of a project funded by the Austrian Science Fund (FWF) at the Bioinformatics Research Group at the FHOÖ Campus Hagenberg, and in cooperation with the Institute of Molecular Pathology (IMP), Vienna. In the following, the author describes a selection of results that tackle the highlighted research problems.

Robust crosslink identification from MS2-MS3-acquisition workflows.

Experiments applying multi-stage tandem XLMS are amongst the most common in XLMS, however software for the analysis of the produced data is sparse and limited in sensitivity and specificity [17], [20], [21]. Sensitivity and specificity are crucial to extract biological meaning from crosslinking data, which makes current solutions undesirable in practice. M. Birklbauer wrote an algorithm that can correctly match multi-stage LC-MS/MS information via the isotope distribution of precursor ions and identify crosslinks both at the MS2 level and the MS3 level. Furthermore, the algorithm verifies that the isolated precursor ions were correctly selected in the mass spectrometry instrument during measurement and re-calculates the correct peptide mass upon miss-assignment of the monoisotopic precursor peak. Results are then combined via a novel scoring function that was designed by the author to reflect the relationship across multi-stage tandem information. The approach showed robust performance in practice, outperforming existing XL-DBSEs by up to four times more identifications while reporting less false positives. The algorithm was released as an extension called MS Annika 2.0 to the widely used proteomics software Proteomics Discoverer and published in the *Journal of Proteome Research* (2023) [17].

Efficient crosslink identification of non-cleavable reagents from proteome-wide experiments.

Non-cleavable crosslinkers are the most common crosslink reagents [27], but despite their frequent usage the data analysis of system-wide studies is largely unexplored due to a quadratic search space that make computational interpretation challenging [25]. Proteome-wide experiments are therefore unfeasible. M. Birklbauer introduced a novel approach to this problem by encoding peptides and mass spectra as sparse vectors to calculate an approximate correlation score by matrix multiplication for an optimized selection of crosslink candidates. This implementation allows for fast and efficient search of millions of spectra against large proteome-wide databases. Moreover, this approach is not only faster and more accurate than competing tools, reporting on average 75% less false positives while yielding comparable numbers of crosslink identifications, but also allows processing of large crosslink studies that were previously unfeasible. The algorithm was released as feature update MS Annika 3.0 in the proteomics software Proteome Discoverer and published in *Communications Chemistry* (2024) [18].

Robust FDR estimation for crosslink results.

Validation of results is an integral part of computational proteomics and one of the key metrics to assess the quality of XL-DBSEs [26]. Despite the existence of an abundance of XL-DBSEs, most of them underestimate the true error of the reported results [20], [21]. The author introduced a novel scoring function for MS2-MS3-based acquisition workflows that allows for more accurate estimation of the identification error and showed better performance than competing tools [17]. Moreover, using an optimized crosslink candidate selection the author could show a reduction in false positives for non-cleavable crosslink searches as well, again outperforming competing tools [18].

Future work.

The XL-DBSE developed by M. Birklbauer is mostly feature complete, supporting identification of various types of crosslinkers and acquisition workflows. The author is currently working on a python package to simplify down-stream analysis of crosslink results such as visualization in pyMOL [28] and support for external validation tools like xiFDR [26] which will aid in putting results into biological context.

Research Progress

Software

MS Annika is an XL-DBSE originally implemented by the Bioinformatics Research Group at the FHOÖ Campus Hagenberg. The author took over this project in summer 2021 as the sole maintainer and extended the software with an algorithm for identification of crosslinks from MS2-MS3-acquisition workflows in the feature update ***MS Annika 2.0*** released in 2023 [17]. Another feature update – ***MS Annika 3.0*** – was released in 2024 supporting identification of non-cleavable crosslinks [18]. Aside of the addition of novel algorithms that were also published in renowned proteomics journals, the author optimized the software for newly released mass spectrometry instruments such as Thermo Fisher's Astral and added support for various internally and externally developed down-stream analysis tools.

Crosslink Proteomics Bioinformatics Publications

"MS Annika 2.0 Identifies Cross-Linked Peptides in MS2–MS3-Based Workflows at High Sensitivity and Specificity" journal article published in the *Journal of Proteome Research* (2023) [17]. The author implemented the algorithm, tested the approach, performed data analysis, did the comparisons to other tools, and wrote the manuscript.

"Proteome-wide non-cleavable crosslink identification with MS Annika 3.0 reveals the structure of the C. elegans Box C/D complex" journal article published in *Communications Chemistry* (2024) [18]. The author implemented the algorithm, tested the approach, performed data analysis, did the comparisons to other tools, and wrote the manuscript.

"A journey towards developing a new cleavable crosslinker reagent for in-cell crosslinking" journal article currently in revision at *Communications Chemistry* [24]. The author adapted the software for the unique fragmentation behavior of the proposed crosslinker, performed data analysis and wrote the corresponding section in the manuscript.

"Unified down-stream analysis of crosslink results with pyXLMS" is a planned journal article to be submitted to *Bioinformatics*. The author implemented the python package, did data analysis and will write the manuscript.

"Optimized crosslink quantification using data-independent acquisition" is a planned journal publication to be submitted to *Nature Communications*. The author implemented the underlying software solution.

Proteomics Bioinformatics Publications

"Making sense of internal ions" is a planned journal article to be submitted to the *Journal of Proteome Research*. The author implemented large parts of the described software and contributed to the manuscript.

Conference Talks and Posters

21/03/2022 Poster: *"Extending MS Annika for MS2-MS3-based Cross-linking Workflows"* at the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS) Winter School 2022.

13/09/2022 Talk: *"Identifying Crosslinks in MS2-MS3-based Workflows with MS Annika"* at the Austrian Proteomics and Metabolomics Research Symposium (APMRS) 2022.

15/01/2023 Talk/Poster: *"MS Annika identifies cross-linked peptides in MS2-MS3-based workflows at high sensitivity and specificity"* at the EuBIC-MS Developers Meeting 2023.

27/09/2023 Poster: *"MS Annika 2.0: Identification of cross-linked peptides in MS2 and MS3 spectra"* at APMRS 2023.

15/01/2024 Poster: *"Cleavable crosslink identification from MS2 and MS3 spectra with MS Annika 2.0"* at the EuBIC-MS Winter School 2024.

25/09/2024 Talk: *"Proteome-wide Non-Cleavable Crosslink Identification Using Sparse Matrix Multiplication with MS Annika 3.0"* at APMRS 2024.

Awards

- Best Presentation Award at APMRS 2024

Other Research Activities

- Reviewer for *Analytical Chemistry*.
- Co-organizer of the EuBIC-MS Winter School 2024
- Member of the Junior Board of the Austrian Proteomics and Metabolomics Association (APMA)

Personal References

[17] M. J. Birklbauer, M. Matzinger, F. Müller, K. Mechtler, and V. Dorfer, "Ms annika 2.0 identifies cross-linked peptides in ms2–ms3-based workflows at high sensitivity and specificity," *Journal of Proteome Research*, vol. 22, no. 9, pp. 3009–3021, Aug. 2023.

[18] M. J. Birklbauer, F. Müller, S. S. Geetha, M. Matzinger, K. Mechtler, and V. Dorfer, "Proteome-wide non-cleavable crosslink identification with ms annika 3.0 reveals the structure of the c. elegans box c/d complex," *Communications Chemistry*, vol. 7, no. 1, Dec. 2024.

[24] F. Müller, B. R. Brutiu, I. Saridakis, T. Leischner, M. Birklbauer, M. Matzinger, M. Madalinski, T. Lendl, S. Shaaban, V. Dorfer, N. Maulide, and K. Mechtler, "A journey towards developing a new cleavable crosslinker reagent for in-cell crosslinking," *bioRxiv*, Nov. 2024.

External References

[1] D. Perrett, "From 'protein' to the beginnings of clinical proteomics," *PROTEOMICS – Clinical Applications*, vol. 1, no. 8, pp. 720–738, Aug. 2007.

[2] N. L. Anderson and N. G. Anderson, "Proteome and proteomics: New technologies, new concepts, and new words," *ELECTROPHORESIS*, vol. 19, no. 11, pp. 1853–1861, Aug. 1998.

[3] W. P. Blackstock and M. P. Weir, "Proteomics: Quantitative and physical mapping of cellular proteins," *Trends in Biotechnology*, vol. 17, no. 3, pp. 121–127, Mar. 1999.

[4] A. G. Birhanu, "Mass spectrometry-based proteomics as an emerging tool in clinical laboratories," *Clinical Proteomics*, vol. 20, no. 1, Aug. 2023.

[5] E. V. Petrotchenko and C. H. Borchers, "Crosslinking combined with mass spectrometry for structural proteomics," *Mass Spectrometry Reviews*, vol. 29, no. 6, pp. 862–876, Aug. 2010.

[6] Y. Shi, J. Fernandez-Martinez, E. Tjioe, R. Pellarin, S. J. Kim, R. Williams, D. Schneidman-Duhovny, A. Sali, M. P. Rout, and B. T. Chait, "Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex," *Molecular & Cellular Proteomics*, vol. 13, no. 11, pp. 2927–2943, Nov. 2014.

[7] C. R. Weisbrod, J. D. Chavez, J. K. Eng, L. Yang, C. Zheng, and J. E. Bruce, "In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy," *Journal of Proteome Research*, vol. 12, no. 4, pp. 1569–1579, Feb. 2013.

[8] Z. Ser, P. Cifani, and A. Kentsis, "Optimized cross-linking mass spectrometry for in situ interaction proteomics," *Journal of Proteome Research*, vol. 18, no. 6, pp. 2545–2558, May 2019.

[9] F. J. O'Reilly and J. Rappsilber, "Cross-linking mass spectrometry: Methods and applications in structural, molecular and systems biology," *Nature Structural & Molecular Biology*, vol. 25, no. 11, pp. 1000–1008, Oct. 2018.

[10] A. Leitner, M. Faini, F. Stengel, and R. Aebersold, "Crosslinking and mass spectrometry: An integrated technology to understand the structure and function of molecular machines," *Trends in Biochemical Sciences*, vol. 41, pp. 20–32, 1 Jan. 2016.

[11] Ş. Yılmaz, F. Busch, N. Nagaraj, and J. Cox, "Accurate and automated high-coverage identification of chemically cross-linked peptides with maxlynx," *Analytical Chemistry*, vol. 94, pp. 1608–1617, 3 Jan. 2022.

[12] F. Liu, P. Lössl, R. Scheltema, R. Viner, and A. J. R. Heck, "Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification," *Nature Communications*, vol. 8, no. 1, May 2017.

[13] Z.-L. Chen, J.-M. Meng, Y. Cao, J.-L. Yin, R.-Q. Fang, S.-B. Fan, C. Liu, W.-F. Zeng, Y.-H. Ding, D. Tan, L. Wu, W.-J. Zhou, H. Chi, R.-X. Sun, M.-Q. Dong, and S.-M. He, "A high-speed search engine plink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides," *Nature Communications*, vol. 10, no. 1, Jul. 2019.

[14] G. J. Pirklbauer, C. E. Stieger, M. Matzinger, S. Winkler, K. Mechtler, and V. Dorfer, "Ms annika: A new cross-linking search engine," *Journal of Proteome Research*, vol. 20, no. 5, pp. 2560–2569, Apr. 2021.

[15] M. Matzinger and K. Mechtler, "Cleavable cross-linkers and mass spectrometry for the ultimate task of profiling protein–protein interaction networks in vivo," *Journal of Proteome Research*, vol. 20, no. 1, pp. 78–93, Nov. 2020.

[16] M. A. Clasen, M. Ruwolt, C. Wang, J. Ruta, B. Bogdanow, L. U. Kurt, Z. Zhang, S. Wang, F. C. Gozzo, T. Chen, P. C. Carvalho, D. B. Lima, and F. Liu, "Proteome-scale recombinant standards and a robust high-speed search engine to advance cross-linking ms-based interactomics," *Nature Methods*, Oct. 2024.

[19] K. Yugandhar, T.-Y. Wang, A. K.-Y. Leung, M. C. Lanz, I. Motorykin, J. Liang, E. E. Shayhidin, M. B. Smolka, S. Zhang, and H. Yu, "Maxlinker: Proteome-wide cross-link identifications with high specificity and sensitivity," *Molecular & Cellular Proteomics*, vol. 19, pp. 554–568, 3 Mar. 2020.

[20] R. Beveridge, J. Stadlmann, J. M. Penninger, and K. Mechtler, "A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes," *Nature Communications*, vol. 11, no. 1, Feb. 2020.

[21] M. Matzinger, A. Vasiu, M. Madalinski, F. Müller, F. Stanek, and K. Mechtler, "Mimicked synthetic ribosomal protein complex for benchmarking crosslinking mass spectrometry workflows," *Nature Communications*, vol. 13, no. 1, Jul. 2022.

[22] P. F. Pilch and M. P. Czech, "Interaction of cross-linking agents with the insulin effector system of isolated fat cells. covalent linkage of 125i-insulin to a plasma membrane receptor protein of 140, 000 daltons.," *Journal of Biological Chemistry*, vol. 254, no. 9, pp. 3375–3381, May 1979.

[23] A. Kao, C.-l. Chiu, D. Vellucci, Y. Yang, V. R. Patel, S. Guan, A. Randall, P. Baldi, S. D. Rychnovsky, and L. Huang, "Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes," *Molecular & Cellular Proteomics*, vol. 10, no. 1, p. M110.002170, Jan. 2011.

[25] J. Rappsilber, "The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes," *Journal of Structural Biology*, vol. 173, no. 3, pp. 530–540, Mar. 2011.

[26] L. Fischer and J. Rappsilber, "Quirks of error estimation in cross-linking/mass spectrometry," *Analytical Chemistry*, vol. 89, no. 7, pp. 3829–3833, Mar. 2017.

[27] B. Steigenberger, P. Albanese, A. J. R. Heck, and R. A. Scheltema, "To cleave or not to cleave in xl-ms?" *Journal of the American Society for Mass Spectrometry*, vol. 31, no. 2, pp. 196–206, Dec. 2019.

[28] B. Schiffrin, S. E. Radford, D. J. Brockwell, and A. N. Calabrese, "Pyxlinkviewer: A flexible tool for visualization of protein chemical crosslinking data within the pymol molecular graphics system," *Protein Science*, vol. 29, pp. 1851–1857, 8 Aug. 2020.

Publications

- M. J. Birklbauer, M. Matzinger, F. Müller, K. Mechtler, and V. Dorfer, "Ms annika 2.0 identifies cross-linked peptides in ms2–ms3-based workflows at high sensitivity and specificity," *Journal of Proteome Research*, vol. 22, no. 9, pp. 3009–3021, Aug. 2023
- M. J. Birklbauer, F. Müller, S. S. Geetha, M. Matzinger, K. Mechtler, and V. Dorfer, "Proteome-wide non-cleavable crosslink identification with ms annika 3.0 reveals the structure of the c. elegans box c/d complex," *Communications Chemistry*, vol. 7, no. 1, Dec. 2024
- F. Müller, B. R. Brutiu, I. Saridakis, T. Leischner, M. Birklbauer, M. Matzinger, M. Madalinski, T. Lendl, S. Shaaban, V. Dorfer, N. Maulide, and K. Mechtler, "A journey towards developing a new cleavable crosslinker reagent for in-cell crosslinking," *bioRxiv*, Nov. 2024