mzIdentML 1.3.0 - Essential progress on the support of crosslinking and other identifications based on multiple spectra

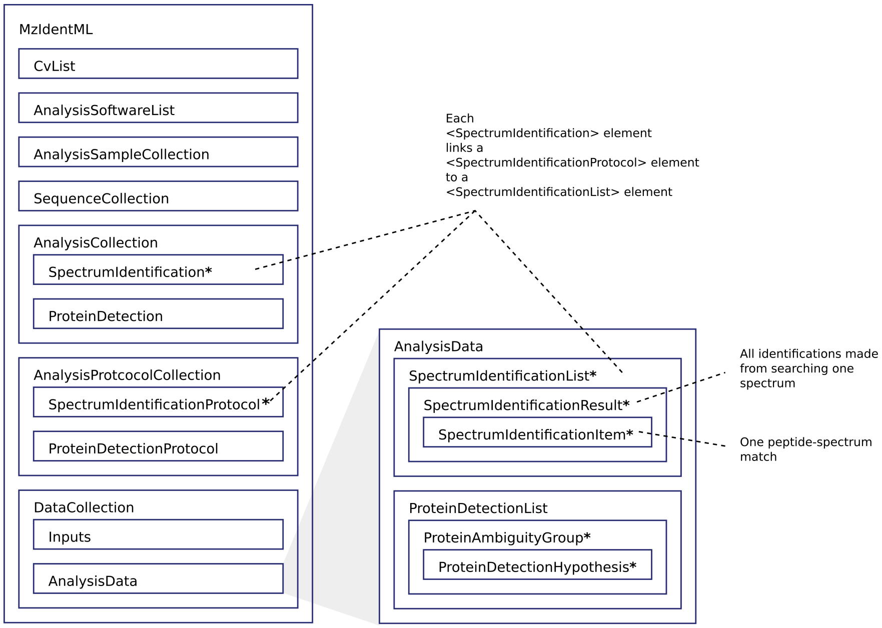
Colin William Combe[1], Lars Kolbowski[2], Lutz Fischer[2], Ville Koskinen[3], Joshua Klein[4], Alexander Leitner[5], Andy Jones[6], Juan Antonio Vizcaino[7], and Juri Rappsilber[1]

[1]University of Edinburgh
[2]Technische Universität Berlin
[3]Matrix Science Ltd
[4]Boston University
[5]ETH Zurich
[6]University of Liverpool
[7]EMBL-EBI

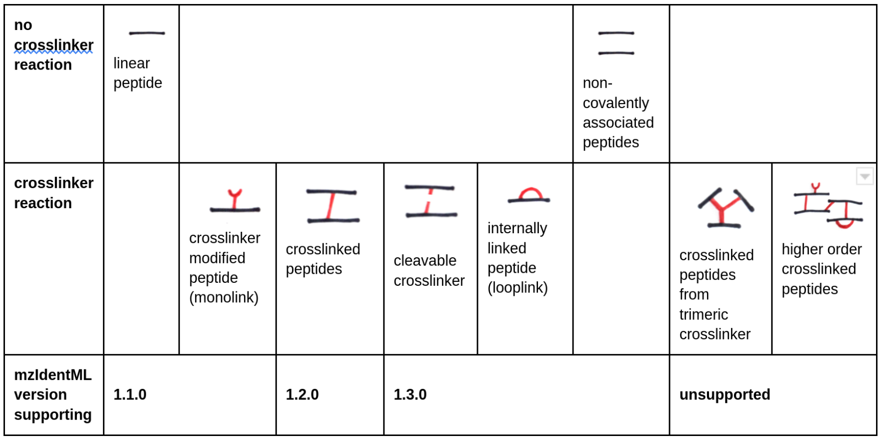
October 9, 2023

Abstract

The mzIdentML file format, originally developed by the Proteomics Standards Initiative in 2011, is the open XML data standard for peptide and protein identification results coming from mass spectrometry. We present mzIdentML version 1.3.0, which introduces new functionality and support for additional use cases. First of all, a new mechanism for encoding identifications based on multiple spectra. Furthermore, the main mzIdentML specification document can now be supplemented by extension documents which provide further guidance for encoding specific use cases for different proteomics subfields. One extension document has been added, covering additional use cases for the encoding of crosslinked peptide identifications. The ability to add extension documents facilitates keeping the mzIdentML standard up to date with advances in the proteomics field, without having to change the main specification document. The crosslinking extension document provides further explanation of the crosslinking use cases already supported in mzIdentML version 1.2.0, and provides support for encoding additional scenarios that are critical to reflect developments in the crosslinking field and facilitate its integration in structural biology. These are: (i) support for cleavable crosslinkers, (ii) support for internally linked peptides, (iii) support for noncovalently associated peptides, and (iv) improved support for encoding scores and the corresponding thresholds.


MzIdentML
CvList
AnalysisSoftwareList
AnalysisSampleCollection
SequenceCollection
AnalysisCollection
SpectrumIdentification*
ProteinDetection
AnalysisProtcocolCollection
SpectrumIdentificationProtocol*
ProteinDetectionProtocol
DataCollection
Inputs
AnalysisData
Each <SpectrumIdentification> element links a <SpectrumIdentificationProtocol> element to a <SpectrumIdentificationList> element
AnalysisData
SpectrumIdentificationList*
SpectrumIdentificationResult*
SpectrumIdentificationItem*
All identifications made from searching one spectrum
One peptide-spectrum match
ProteinDetectionList
ProteinAmbiguityGroup*
ProteinDetectionHypothesis*

* = may be many elements of this type within containing element



no crosslinker reaction	linear peptide					non-covalently associated peptides		
crosslinker reaction		crosslinker modified peptide (monolink)	crosslinked peptides	cleavable crosslinker	internally linked peptide (looplink)		crosslinked peptides from trimeric crosslinker	higher order crosslinked peptides
mzIdentML version supporting	**1.1.0**		**1.2.0**	**1.3.0**			**unsupported**	

Technical Brief

mzIdentML 1.3.0 - Essential progress on the support of crosslinking and other identifications based on multiple spectra

Colin W. Combe[1,2], Lars Kolbowski[2], Lutz Fischer[2], Ville Koskinen[3], Joshua Klein[4], Alexander Leitner[5], Andrew R Jones[6], Juan Antonio Vizcaíno[7], Juri Rappsilber[1,2]

Affiliations

1. University of Edinburgh, Wellcome Centre for Cell Biology, School of Biological Sciences, Edinburgh, EH9 3JR, UK
2. Technische Universität Berlin, Chair of Bioanalytics, 10623 Berlin, Germany
3. Matrix Science Ltd, 64 Baker Street, London W1U 7GB, UK
4. Program for Bioinformatics, Boston University, Boston, Massachusetts, USA,
5. Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, 8093 Zurich, Switzerland
6. Department of Biochemistry & Systems Biology, University of Liverpool, Liverpool, L69 7ZB, UK
7. European Molecular Biology Laboratory, European Bioinformatics
Institute (EMBL-EBI), Wellcome Trust, Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Corresponding Author:

Prof. Juri Rappsilber

Juri.Rappsilber@TU-Berlin.de

TIB 4/4-3

Gebäude 17, Aufgang 1, Raum 476

Gustav-Meyer-Allee 25

13355 Berlin

+49 30 314-72374

Abbreviations

CV	Controlled vocabulary
ETD	Electron transfer dissociation
FAIR	Findable, Accessible, Interoperable and Reusable
HCD	Higher-energy collision-induced dissociation
HUPO-PSI	Human Proteome Organisation - Proteomics Standards Initiative
MS	Mass spectrometry
PDB	Protein Data Bank
PSM	Peptide-spectrum match
XML	Extensible Markup Language

Keywords: crosslinking, mass spectrometry, mzIdentML, proteomics, standards

Word count: 2889

Abstract

The mzIdentML file format, originally developed by the Proteomics Standards Initiative in 2011, is the open XML data standard for peptide and protein identification results coming from mass spectrometry. We present mzIdentML version 1.3.0, which introduces new functionality and support for additional use cases. First of all, a new mechanism for encoding identifications based on multiple spectra. Furthermore, the main mzIdentML specification document can now be supplemented by extension documents which provide further guidance for encoding specific use cases for different proteomics subfields. One extension document has been added, covering additional use cases for the encoding of crosslinked peptide identifications. The ability to add extension documents facilitates keeping the mzIdentML standard up to date with advances in the proteomics field, without having to change the main specification document.

The crosslinking extension document provides further explanation of the crosslinking use cases already supported in mzIdentML version 1.2.0, and provides support for encoding additional scenarios that are critical to reflect developments in the crosslinking field and facilitate its integration in structural biology. These are: (i) support for cleavable crosslinkers, (ii) support for internally linked peptides, (iii) support for noncovalently associated peptides, and (iv) improved support for encoding scores and the corresponding thresholds.

Main text

The Human Proteome Organization - Proteomics Standards Initiative (HUPO-PSI) is in charge of developing open data standards in the proteomics field, including mass spectrometry (MS)-based approaches and molecular interaction data [1], [2], [3]. mzIdentML is the PSI XML-based standard for peptide and protein identification results based on MS data. The first stable version of the standard (version 1.1) was formalised in 2012 [4]. In 2017, version 1.2 of the data standard was formalised, adding support for new use cases such as the scoring of the protein modification position, the reporting of peptide level statistics, proteogenomics approaches and initial support for crosslinking data [5]. mzIdentML is implemented widely (https://www.psidev.info/tools-implementing-mzidentml) and is the recommended format for submitting peptide and protein identification results to the public proteomics data repositories which are part of ProteomeXchange [6], including e.g. PRIDE [7] and jPOST [8]. Being the primary open data standard for protein and peptide identifications, mzIdentML needs to be updated regularly in parallel to the developments in the field, as is the case for other data standards.

A driver for updating the mzIdentML standard has been the crosslinking MS field. While initial support of crosslink identifications was added to mzIdentML in version 1.2, important aspects of crosslinking data were not reflected. To follow the recommendations of wwPDB [9] and the crosslinking field [9], [10] and make data more FAIR (Findable, Accessible, Interoperable and Reusable) a transparent linking between the PDB (Protein Data Bank) and ProteomeXchange is required for crosslinking data. This mandates clean reporting of the false discovery rates and score thresholds [11], [12]. Furthermore, crosslinked peptides are frequently investigated with multiple fragmentation modes in parallel and also selecting fragments for further fragmentation [13]. This results in multiple fragmentation spectra being linked to one identification. In addition, peptides can be detected together not only as a result of crosslinking, but also because of non-covalent interactions [13], [14]. Finally, crosslinks may be found not only between peptides but also within [15].

To address these needs and challenges, we present mzIdentML version 1.3.0. Arriving at version 1.3.0 of this data standard followed the procedures established by HUPO-PSI for the revision of specification documents [15], [16]. The initial discussion of the new use cases to be addressed took place at the 2022 HUPO-PSI meeting (May 2022). Then, over the course of a year, a working group composed of key stakeholders met to discuss the details of the changes. Finally, the revised documents have been submitted to the HUPO-PSI document process, a three-level process of review that must be completed before any proposal is declared a ratified standard [3]. All documents in their most recent form are available at the mzIdentML page in the HUPO-PSI website (http://psidev.info/mzIdentML) and at the mzIdentML GitHub page (https://github.com/HUPO-PSI/mzIdentML/). The latest specification document, which we will be referencing, is available at https://github.com/HUPO-PSI/mzIdentML/tree/master/specification_document/specdoc1_3.

[Special note to the reviewers of this manuscript (to be removed before publication): This manuscript has been submitted to this journal for peer review and released via bioRxiv in parallel with a submission of the full technical specification document and related files to the PSI Document Process. In the PSI process, a PSI editor calls out to anonymous reviewers not involved in the development of this format to carefully scrutinise the technical specification (not this manuscript) to ensure its suitability. Until both the PSI review and journal review are complete to the satisfaction of the respective editors, the proposed mzIdentML revisions format will not yet be ratified as version 1.3.0, and your comments will influence the final product. Until ratified, the URL for the documents is https://github.com/HUPO-PSI/mzIdentML/tree/master/specification_document/specdoc1_3-draft]

mzIdentML 1.3.0 introduces the following changes. First, a new mechanism for encoding identifications based on multiple spectra has been introduced. This is applicable across multiple proteomics domains, the examples referred to in the specification document are drawn from crosslinking studies and the encoding of glycopeptides. Second, there is a new, optional, method for encoding a link between <Modification> elements and <SearchModification> elements, see Section 7.12 of the mzIdentML 1.3.0 specification. This allows for more detailed information on modifications to be provided. Third, the main mzIdentML specification document can now be supplemented by “extension documents" that provide further guidance for specific use cases (rather than editing the original specification document as a whole every time that new use cases arising from proteomics subfields need to be added). This provides a much more flexible mechanism. A first extension document has been added, covering the identification of crosslinked peptides. The crosslinking community has recognised the importance of data standards and the work on the mzIdentML 1.3.0 crosslinking extension follows the consensus roadmap put forward by the field [10].

Next, we highlight in detail some of the new use cases supported for crosslinking data and the rationale behind the decisions taken. Some analysis workflows utilise multiple spectra to arrive at a given identification. For more detailed information on encoding such workflows, see Section 7.11 of the mzIdentML 1.3.0 specification. mzIdentML 1.2.0 already provided a way of encoding identifications based on multiple spectra using the “combined spectra” type of input file format (see Section 5.2.9 of the mzIdentML 1.2.0 specification). However, this is inadequate for correctly encoding some workflows due to the restriction that a single <SpectrumIdentificationResult> element can only be associated with a single <SpectrumIdentificationProtocol> (**Figure 1**). The “combined spectra” type of input file format, present in mzIdentML 1.2.0, essentially associates a single <SpectrumIdentificationResult>

element with a comma-separated list of spectrum identifiers. The problem with this can be shown by considering three possible crosslinking search strategies:

(i) spectra from precursors of peptides crosslinked with different isotopic versions of a reagent ("light" and "heavy") are combined together and searched once;

(ii) multiple spectra of the same precursor are acquired, e.g. using different fragmentation techniques like HCD and ETD;

(iii) when using a cleavable crosslinker and both MS3 spectra of the cleaved peptides and the MS2 spectrum of the crosslinked peptide pair are considered in the identification process [17].

When the "combined spectra" type of input file format is used to encode identifications based on multiple spectra, each spectrum identifier in the comma-separated list of identifiers is associated with the same set of acquisition settings and search parameters. This would work for crosslinking search strategy (i) if all the spectra contributing to an identification share the same acquisition settings. However, in use cases (ii) and (iii), the settings usually differ between the spectra contributing to the same identification. Since it is not possible to record multiple sets of acquisition settings in a single <SpectrumIdentificationProtocol>, these search strategies cannot be correctly encoded using the 1.2.0 "combined spectra" type of file format.

This problem has been overcome by introducing a new controlled vocabulary (CV) term [18]. mzIdentML 1.3.0 now advises the use of the new CV term "identification based on multiple spectra" (MS:1003332) instead of the "combined spectra" file format type. This CV term provides an identifier to associate <SpectrumIdentificationItem> elements across multiple <SpectrumIdentificationList> elements. These <SpectrumIdentificationList> elements can then be associated with different <SpectrumIdentificationProtocol> elements. This now allows spectra supporting the same identification to have different acquisition settings encoded. The identifier can be used to distinguish ‘parent’ spectra, which cover the entire identification (in the case of crosslinking, the crosslinked peptide pair), and ‘child’ spectra, which only identify a constituent part of the identification (e.g. an individual peptide in an MS3 scan).

Although the examples above came from crosslinking studies, the underlying mechanism for encoding identifications based on multiple spectra is also applicable to other use cases, such as encoding glycopeptides. For this reason, this change is included in the main mzIdentML 1.3.0 document.

mzIdentML version 1.3.0 introduces two new CV terms to link <SearchModification> elements and <Modification> elements - “search modification id” (MS:1003392) which goes inside <SearchModification> elements, and “search modification id ref” (MS:1003393) which goes inside <Modification> elements. This allows for more detailed information on modifications to be provided, if necessary, without redundant repetition of this information throughout the file. Making this link is optional but recommended where possible. In the case of open modification searches, such a link cannot be made.

As mentioned above, one mzIdentML extension document has been developed, giving further clarification and guidance on the encoding of crosslinking studies, including new use cases not previously supported in mzIdentML 1.2.0.

In the crosslinking extension document, additions are made to provide: (i) support for cleavable crosslinkers, (ii) support for internally linked peptides, (iii) support for noncovalently associated peptides, and (iv) improved support for encoding scores and the corresponding thresholds.

(i) Particular attention has been paid to supporting workflows that use cleavable crosslinkers. MS-cleavable crosslinkers can cleave upon activation in the mass spectrometer, releasing the individual peptides (modified with a crosslinker “stub”) and thus enabling their individual analysis [17]. Supporting this has been approached as two independent tasks. Improving support for identifications based on multiple spectra (see above and Section 7.11 of the mzIdentML 1.3.0 specification), and improving the encoding of derivatives of cleavable crosslinkers (see Section 3.2.2 of the crosslinking extension document).

(ii) Internally linked peptides, commonly known as “looplinks”, are cases where both ends of the crosslinker are within a single peptide, not between two copies of the same peptide (see Section 3.3 and Section 3.4.3 of the crosslinking extension document). This type of crosslinking product is therefore necessarily intramolecular.

(iii) Some spectra show the fragmentation of two peptides which were not crosslinked but stayed associated due to noncovalent interactions [14] (see Section 3.4.2 of the crosslinking extension document). Both peptides appear together as a single precursor species in the instrument, as opposed to ‘chimeric’ spectra where a single peptide is selected as precursor and additional peptide(s) fall into the same selection window. Identifying these noncovalently associated peptides improves the accuracy of crosslinking analyses, as it can prevent them from being misidentified as crosslinked peptides [14].

(iv) Finally, the encoding of scores applicable to crosslinking MS results, and their corresponding thresholds, has been clarified and improved (see Section 4 of the crosslinking extension document). There are different points in the analysis at which thresholds may be applied [12], [11]. These correspond to different levels of consolidation at which error control procedures may be performed. Scores and thresholds are encoded differently in mzIdentML depending on the level of consolidation at which they were applied. For crosslinking studies encoded in mzIdentML, the possible levels are: crosslink containing PSM (peptide-spectrum match); unique peptide-pair; unique residue-pair; protein-protein interaction.

An overview of the different crosslinking product types and their support in mzIdentML is given in **Figure 2**. For a discussion of the product types that are not supported in this version of mzIdentML (trimeric crosslinkers and higher order crosslinked peptides) see Section 5 of the crosslinking extension document.

In terms of software implementations, work is underway to implement the new revisions in *xi-mzidentml-convertor* (https://github.com/Rappsilber-Laboratory/xi-mzidentml-converter), a library used by the tools xiSPEC [19] and xiVIEW (https://xiview.org). Integrating the new version of this tool into PRIDE will bring support for the new use cases to this ProteomeXchange repository.

In conclusion, it is now possible to encode in mzIdentML different spectra contributing to the same identification and for these to be associated with different acquisition settings. This adds new functionality that is applicable to multiple types of proteomics studies. Work to implement the new revisions is underway, including within the key proteomics repository PRIDE. Keeping the mzIdentML standard up to date with advances in the proteomics field is now easier through the use of extension documents. The crosslinking extension document provides an example of this. Additional work is currently in progress on a further extension document covering the encoding of glycopeptides, this will also use the new mechanism for identifications based on multiple spectra.

PSI standards are developed via an open process in which all interested individuals and groups are encouraged to participate. Broad participation is, therefore, essential for successful generation of future standards for the proteomics community. See the PSI web site (https://www.psidev.info/) or PSI mailing lists for information about how to contribute to the PSI.

Acknowledgements

The authors want to acknowledge funding from the BBSRC BBR '3D Proteomics' grant [BB/V018779/1, BB/V018817/1]. JAV would also like to acknowledge funding from BB/S01781X/1 and EMBL core funding. ARJ would like to acknowledge funding from BBSRC [BB/S017054/1].

The authors have declared no conflict of interest.

References

[1] Deutsch, E. W., Albar, J. P., Binz, P.-A., Eisenacher, M., Jones, A. R., Mayer, G., Omenn, G. S., … Hermjakob, H. (2015). Development of data representation standards by the human proteome organization proteomics standards initiative. Journal of the American Medical Informatics Association: JAMIA, 22, 495–506.

[2] Deutsch, E. W., Orchard, S., Binz, P.-A., Bittremieux, W., Eisenacher, M., Hermjakob, H., Kawano, S., … Jones, A. R. (2017). Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. Journal of Proteome Research, 16, 4288–4298.

[3] Deutsch, E. W., Vizcaíno, J. A., Jones, A. R., Binz, P.-A., Lam, H., Klein, J., Bittremieux, W., … Orchard, S. E. (2023). Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work. Journal of Proteome Research, 22, 287–301.

[4] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., Selley, … Creasy, D. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. Molecular & Cellular Proteomics: MCP, 11, M111.014381.

[5] Vizcaíno, J. A., Mayer, G., Perkins, S., Barsnes, H., Vaudel, M., Perez-Riverol, Y., Ternent, T., … Jones, A. R. (2017). The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. Molecular & Cellular Proteomics: MCP, 16, 1275–1285.

[6] Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L., Kundu, D. J., … Vizcaíno, J. A. (2022). The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Research, 51, D1539–D1548.

[7] Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D. J., … Vizcaíno, J. A. (2021). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research, 50, D543–D552.

[8] Watanabe, Y., Yoshizawa, A. C., Ishihama, Y., & Okuda, S. (2021). The jPOST Repository

as a Public Data Repository for Shotgun Proteomics. Methods in Molecular Biology , 2259, 309–322.

[9] Berman, H. M., Adams, P. D., Bonvin, A. A., Burley, S. K., Carragher, B., Chiu, W., DiMaio, F., … Sali, A. (2019). Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures. Structure , 27, 1745–1759.

[10] Leitner, A., Bonvin, A.M.J.J., Borchers, C. H., Chalkley, R. J., Chamot-Rooke, J., Combe, C. W., Cox, J., … Rappsilber, J. (2020). Toward Increased Reliability, Transparency, and Accessibility in Cross-linking Mass Spectrometry. Structure , 28, 1259-1268. https://doi.org/10.1016/j.str.2020.09.011

[11] Fischer, L., & Rappsilber, J. (2017). Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. Analytical Chemistry, 89, 3829-3833.

[12] Lenz, S., Sinn, L. R., O'Reilly, F. J., Fischer, L., Wegner, F., & Rappsilber, J. (2021). Reliable identification of protein-protein interactions by crosslinking mass spectrometry. Nature Communications, 12, 3564.

[13] O'Reilly, F. J., & Rappsilber, J. (2018). Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. Nature Structural & Molecular Biology, 25, 1000–1008.

[14] Giese, S. H., Belsom, A., Sinn, L., Fischer, L., & Rappsilber, J. (2019). Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses. Analytical Chemistry, 91, 2678–2685.

[15] Sinz, A. (2006). Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. Mass Spectrometry Reviews, 25, 663–682.

[16] Vizcaíno, J. A., Martens, L., Hermjakob, H., Julian, R. K., & Paton, N. W. (2007). The PSI formal document process and its implementation on the PSI website. Proteomics, 7, 2355–2357.

[17] Liu, F., Lössl, P., Scheltema, R., Viner, R., & Heck, A. J. R. (2017). Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. Nature Communications, 8, 15473.

[18] Mayer, G., Jones, A. R., Binz, P.-A., Deutsch, E. W., Orchard, S., Montecchi-Palazzi, L., Vizcaíno, J. A., … Eisenacher, M. (2014). Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochimica et Biophysica Acta, 1844, 98–107.

[19] Kolbowski, L., Combe, C., & Rappsilber, J. (2018). xiSPEC: web-based visualization, analysis and sharing of proteomics data. Nucleic Acids Research, 46, W473–W478.

Figure Legends

Figure 1. Overview of the structure of mzIdentML.

Each <SpectrumIdentifcationResult> can only be associated with a single <SpectrumIdentificationProtocol>. This association is made by <SpectrumIdentification> elements, which link the <SpectrumIdentificationProtocol> to the containing <SpectrumIdentificationList>. <SpectrumIdentificationProtocol> elements encode the parameters and settings of a spectrum identification analysis.

Figure 2. Summary of mzIdentML support for crosslinking product types.