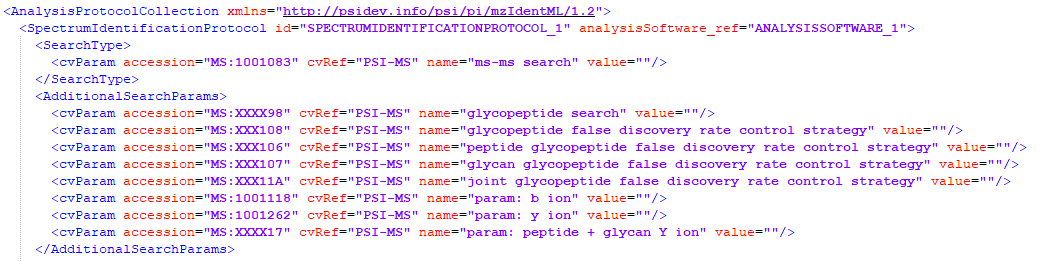
### Encoding results of glycopeptide searches

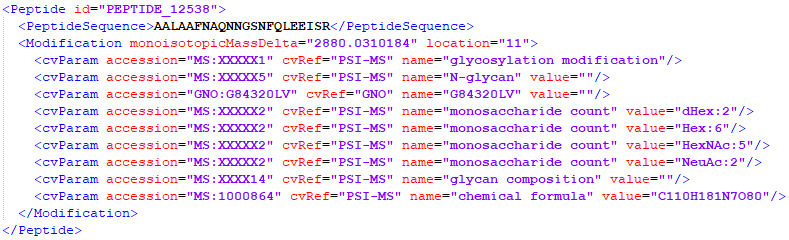
In version 1.2.1, a mechanism for encoding identification of glycopeptides was added. Because glycopeptide identification involves identifying a peptide along with one or more glycans and where they are attached to the peptide, additional information must be encoded to accurately describe what was searched for and what was identified. Additionally, because experimental methods for glycopeptide identification are diverse, what kinds of expectations the identification software was told to make must also be recorded. To address these needs, mzIdentML 1.2.1 adds the following amendments.



**Figure 1.** Additional search parameters added for a multidimensional FDR controlled glycopeptide search experiment.

***Encoding a glycosylation modification***

A <Modification> representing a glycosylation event may be declared by using the CV term “glycosylation modification” (MS:XXXXX1) and by including either a CV term referencing an externally defined glycan, such as those in the GNOme namespace, or provide one or more CV terms describing the glycan from known (MS:XXXXX2) or unknown/user-defined (MS:XXXXX3) saccharide and/or free substituent counts. Additionally, the type of glycosylation MAY be specified using a CV term derived from “glycan class” (MS:XXXXX4), such as “N-glycan” (MS:XXXXX5) or “O-glycan” (MS:XXXXX6) based upon how the search engine treated the glycosylation event, conveying the search engine’s assumptions without making assertions about the actual biological process. If the glycan specified is a structure with a defined topology, the CV term “glycan structure” (MS:XXXX13) should be used, otherwise the term “glycan composition” (MS:XXXX14) should be used. If a single glycan composition is used to represent more than one glycosylation site occupant, “glycan aggregate” (MS:XXXX15) should be used instead. The total mass difference applied to the peptide sequence by the glycan should be provided as the <Modification> element’s *monoisotopicMassDelta* attribute. In the case of a “glycan aggregate”, sites beyond the first SHOULD be marked with a zero mass glycosylation modification and include term “glycan aggregate” (MS:XXXX15), though further description is recommended if possible.



**Figure 2.** A glycan composition modification encoded with a GNOme reference and its monosaccharide count.

If the search engine used an explicit glycan database, that database should be listed in <ModificationParams> for the <SpectrumIdentificationProtocol>, along with other modifications rules, following the same rules as above.

***Encoding glycopeptide search and FDR estimation protocols***

A document containing glycopeptide identifications must include “glycopeptide search” (MS:XXX101), in the <SpectrumIdentificationProtocol> as described in Table 1. Additionally, one or more terms derived from “glycopeptide false discovery rate control strategy” (MS:XXX105) should be specified in <AdditionalSearchParams> to denote how the false discovery rate of the identifications were controlled. Specification of multiple CV terms indicates that the false discovery rate was controlled along multiple dimensions. If a function is used to combine the separate false discovery rate statistics from multiple sources of evidence, the CV term “joint glycopeptide false discovery rate control strategy” (MS:XXX11A) SHOULD be used in addition. If results were filtered along multiple dimensions, these criteria SHOULD be described in the <Threshold> element.

***Encoding a glycopeptide identification***

*Approach #1*

In addition to normally expected elements, a <SpectrumIdentificationItem> referencing a glycopeptide is RECOMMENDED to provide a description of the expectations the search engine made about the dissociation occurring by specifying one or more terms derived from “glycan dissociation mode” (MS:XXX110). These methods should describe the types of fragment ions searched for, distinct from the activation method but are inspired by commonly used dissociation techniques, e.g. “glycan dissociating, peptide preserving” (MS:XXX111) which may describe CID spectra, “glycan preserving, peptide dissociating” (MS:XXX112) for ETD spectra, “glycan eliminated, peptide dissociating” (MS:XXX114) for HCD spectra, or “glycan dissociating, peptide dissociating” (MS:XXX113) for EThcD spectra. Alternatively, if a single set of modes is assumed for all spectra, these CV terms MAY be given in the <AdditionalSearchParams> and deviations from that assumption may be specified explicitly on each <SpectrumIdentificationItem>.

*Approach #2*

If a search engine considered glycosylated product ions during scoring, it is RECOMMENDED that it register this in the <AdditionalSearchParams> with the appropriate terms derived from “glycosylated ion series” (MS:XXXX16). Additional context may be drawn from the spectrum data files’ activation method if provided. If an ion series was not considered to carry more than the reducing end monosaccharide, it may be assumed that this is included in the non-glycosylated ion series-related terms.

If the search engine assigned separate scores to separate sources of structural evidence, reporting each score is RECOMMENDED, with scoring statistics reported with terms derived from “glycopeptide confidence measure” (MS:XXX101). If multiple false discovery rate control strategies were used, reporting each false discovery rate, local FDR, or q-value is RECOMMENDED with statistics reported with terms derived from “glycopeptide statistic” (MS:XXX117), or using one of “peptide-specific statistic” (MS:XXX116) or “glycan-specific statistic” (MS:XXX115) if appropriate. If a joint statistic summarizing statistics from multiple sources of evidence was used, a term derived from “joint glycopeptide statistic” (MS:XXX118) SHOULD be used to report it.

As in the cases of combining multiple fragmentation methods across separate spectra for cross-linked peptides in section 5.2.9, if the spectrum combination is a form of post-processing, a peptide-level score SHOULD be used. Otherwise, the <SpectraData> element SHOULD specify that it is using a “combined spectra” type of input file format (Table 2). The referenced spectrum (via the spectrumID attribute) SHOULD contain a comma-separated list of identifiers e.g. spectrumID=“index=1001, index=1007”.