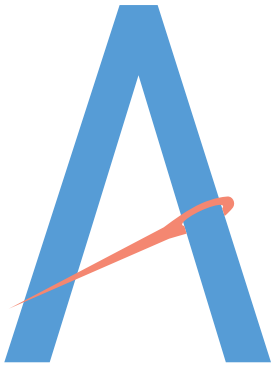


Annotator

A simple tool to help you manually discover the depths of your (complex) spectra, one spectrum at a time. Load your rawfiles, select a spectrum and add your annotation with full control over theoretical fragments. Use the interactive spectrum to discover what your spectrum means and to export gorgeous images.

GitHub: https://github.com/snijderlab/annotator
License: MIT OR Apache-2.0
Version: v1.3.0
DOI: 10.1021/acs.analchem.5c02832[1]

Formatting guide

Buttons are displayed as Button. Keys used in key combinations as `Key`.

> Any side note will be shown like this.

Installing

Using winget

On windows use `winget install --id Snijderlab.Annotator`.

From binary

See GitHub releases for prebuilt binaries for your architecture as well as information on the releases.

From source

To build from source, clone the repository, and build with cargo. Make sure you have installed Rust and Tauri beforehand.

```
git clone https://github.com/snijderlab/annotator annotator
cd annotator
cargo tauri dev
```

Rawfiles

The annotator supports mgf, mzML, IndexedMzML, Bruker .tdf, and Thermo raw files (needs setup see Section 2.3). Rawfiles can be loaded using the button Load raw data file or by dragging in a file from the file explorer. Any number of files can be loaded at the same time. Once opened the file will be added to the list of opened files as 'RX:filename' with two input fields to select spectra and a Close file button to close the file. Spectra can be selected on index, 0 based index of the spectrum in the file, or on native id. Native id is the vendor specific textual identifier for the spectrum, for more detailed info see the documentation in mzdata.

Files in profile mode can be loaded and automatically peak picked in the annotator. For top and middle down data it is recommended to deconvolute the data before loading it in the annotator.

Other formats

Other raw file formats can be converted using available converters. For convenience some common converters are listed below.

Converter	Supported formats
Proteowizard MSConvert	AB/Sciex T2D, Agilent MassHunter, Bruker BAF/Data Exchange/FID/TDF/U2/YEP, Mascot Generic, MS1, MS2, MZ5, mzML, mzXML, Sciex WIFF/WIFF2, Shimadzu LCD, Thermo raw, UIMF, Waters raw
TOPP FileConverter	mzData, mzXML, ANDI/MS
TOPP DTAExtractor	Sequest DTA
Sciex MS Data Converter 2.0	Sciex formats

Selected spectra

Once a spectrum is selected it will be listed below the respective raw file. It will show its index and native id as well as a Unselect button to undo the selection.

Multiple spectra

If multiple spectra are selected at the same time these spectra will be merged before being annotated.

Thermo raw files

The .NET 8.0 runtime is needed to open Thermo raw files, which can be downloaded here. Additionally on windows you can use `winget install Microsoft.DotNet.Runtime.8` for a quick install. Once this is installed Thermo raw files can be loaded as any other file.

Clipboard

Some programs allow copying a spectrum into the clipboard, use the Load Clipboard button to load such a spectrum from the clipboard. Currently spectra from selected Bruker, Stitch, Sciex, and Thermo programs are supported.

> If you find another program that allows this behaviour please open an issue on GitHub for the Annotator, with the name and version of the program in question along with an example of the format.

Universal Spectrum Identifier (USI)

A USI is a text based format that identifies any spectrum in any publicly accessible dataset. It is build using the following parts.

Part	Example	Comment
Prefix	mzspec	required
Collection	PXD043489	required
Dataset	20201103_F1_UM5_Peng0013_-SA_139H2_InS_Elastase.raw	required
Index type	scan	required, can be scan/nativeId/index
Index	11809	required, content depens on Index type
Interpretation	VSLFPPSSEQLTSNASVV	optional, can be any ProForma sequence
Provenance	PR-G47	optional

These parts have to be strung together using colons ':', resulting in the following string:

```
mzspec:PXD043489:20201103_F1_UM5_Peng0013_SA_139H2_InS_Elastase.raw:scan:-
11809:VSLFPPSSEQLTSNASVV
```

Pasting a USI into the Annotator downloads that spectrum from the internet and loads the interpretation (if given) into the peptidoform field.

PSM files

There are quite some programs that export PSMs. The annotator supports a selection of database matching and de novo software file formats. To load these files use the button Load PSM file or drag in a file from the file explorer. Any number of files can be loaded at the same time. See the table below for supported formats. Opened PSM files can be closed with Close file in the PSM details pane.

Open format	Comment
Fasta	No header requirements
mzSpecLib	v1.0
mzTab	v1.0 & Casanovo (3.x, 4.x, & 5.0)
SSL	Spectrum sequence list

Software	Versions/Formats
Basic CSV	A CSV file with the following columns: 'raw_file', 'sequence' (in ProForma format), 'z', 'scan_index', and possibly 'mode' with the fragmentation mode, this ignores any other columns.
DeepNovoFamily	DeepNovo, PointNovo, BiatNovo, PGPointNovo
InstaNovo	1.0.0 & 1.1.4
MaxQuant	msms, msms scans, novo msms scans, & silac
MetaMorpheus	PSM and Peptides
MSFragger	4.2 & 4.3 Fragpipe: 20, 21, & 22, & Philosopher
NovoB	0.0.1
Novor	Denovo and PSM
OPair	common version
Peaks	X, X+, 11, 12, 13 Dia de novo, Ab, DB peptidoforms, DB PSM, & DB protein-peptidoform
PepNet	1.0
PLGS	3.0
pLink	2.3
PowerNovo	1.0.17
Proteoscape	2025B
pUniFind	0.1
Sage	0.14
π-HelixNovo	1.1
π-PrimeNovo	0.1

> If you have data from an unsupported version of a supported program please open an issue on GitHub and give an example file so that the support can be extended.

If you have data from an unsupported program that you think should be supported open an issue on GitHub to discuss.

PSM details

Once a file is opened the PSM details pane opens. Select the right PSM file from the drop down menu and select the right PSM by PSM index. PSMs can be browsed efficiently by selecting the PSM index box and using the arrow up/down keys.

Once a PSM is selected it will show common metadata in a structured format, followed by an overview of the PSM, with the local confidence (if present in the file) depicted by blue squares, and modifications depicted by blue dots. Hover over a modified amino acid to see the full definition. Terminal modifications will be depicted as blue dots on their respective terminal.

Below the structured metadata follows in table format all other metadata from the file format.

Use Load to select the right spectrum (if the corresponding raw file is open) and load the details for annotation. This will load the sequence, charge, and method if these are available.

Peptidoform search

Once at least one PSM file is open all PSM can be searched for sequence patterns. Type the search pattern as a ProForma (see Section 4.1) peptidoform in the search box and hit Search to search. The search is based on mass based alignment[2] so any peptidoform matching the mass pattern of the search will come up. For example, searching for 'WNA' matches 'RWGGAPG'. By default it will show the 25 best matching peptidoforms, but this number can be changed. The search can be restricted with a minimal score for the searched peptidoforms, which also makes the search faster, or with a minimal score for the alignment. Both these minimal score have to be in range 0.0 to 1.0.

Once the search is complete all matching peptidoforms (up to the maximum) will be shown below. The index indicates from which PSM file the peptidoform originated as well as the index in that file. Clicking on the index selects this peptidoform in the peptidoform details pane. The sequence column shows the sequence of the peptidoform, with in blue the section that matched the search term. The match score (normalised mass based alignment score between 0 and 1) as well as the peptidoform score is shown in the last two columns.

Annotate

The annotate sections allow control over the annotation of theoretical spectra on the selected spectrum. The following sections are present:

1. The tolerance for matching theoretical peaks to experimental peaks can controlled and set to ppm or Thompson (mz).
2. The maximum charge for precursors in the theoretical spectrum can be set. If this is not set it takes the charge from the spectrum. If this too is not set it takes +1.
3. The noise filter can be controlled, the noise floor is automatically determined and the noise filter disregards any peak below the factor times the noise floor. Setting this to 0.0 fully removes noise filtering.
4. The match mode indicates the method of determining the mz for theoretical peaks. Set to mono isotopic, average weight, or most abundant isotope.

5. An mz range for the theoretical peaks can be imposed. Setting only one side sets no bounds on the other side.
6. The model can be set to any predefined model. `All` allows most known fragmentation reactions. `None` only allows generation of the precursors. Custom models can also be created see Section 4.3.
7. The peptidoform sequence contains the sequence for the peptidoform in ProForma notation, see Section 4.1.

Hitting Annotate generates the theoretical fragmentation for the given peptidoform with the given settings. The annotated spectrum will be shown below.

ProForma

The Annotator uses the ProForma 2.0 specification to specify the sequence. Here are some examples of valid sequences:

1. Normal amino acids
   `VAEINPSNGGTTFNEKFKGGKATJ`
2. Modifications using UNIMOD, PSI-MOD, RESID, XL-MOD, and GNO
   `EM[L-methionine sulfoxide]EVEES[UNIMOD:21]PEK`
3. Modifications using raw masses
   `TFNEKF[+15.9949]KGGKATJ`
4. Modifications using elemental formula
   `TFNEKF[Formula:O]KGGKATJ`
5. Modifications glycan compositions
   `TFNEKF[Glycan:HexNAc1Hex2]KGGKATJ`
6. Terminal modifications
   `[+16]-TFNEKFKGGKATJ-[Methyl]`
7. Global isotope modifications (all Nitrogen is 15N)
   `<15N>TFNEKFKGGKATJ`
8. Global modifications (all C are carboxamidomethylated)
   `<[S-carboxamidomethyl-L-cysteine]@C>AVYYCSRWGGDGFYAMDYWGQG`
9. Modifications where the location is unknown
   `[UNIMOD:374]?TFNEKFCKGGCKATJ`
10. Modification of unknown position specified on two positions
    `TFNEKFC[UNIMOD:374#g1]KGGC[#g1]KATJ`
11. Modification of unknown position specified on a subset of the peptidoform
    `TFNEKF(CKGGCK)[UNIMOD:374#g1]ATJ`
12. Chimeric spectra, meaning two separate peptidoforms are in your spectrum at the same time
    `VAEINPSNGGTT+FNEKFKGGKATJ`
13. Defined charge, especially good for chimeric cases that have different charges
    `VAEINPSNGGTT/2`
14. A DSSO cross-link between two lysines on two peptidoforms (note the use of `//` versus `+` to indicate cross-linked peptidoforms)
    `VAEINK[X:DSSO#XL1]SNGGTT//WAK[#XL1]INK`
15. A hydrolysed DSSO cross-linker
    `VAEINK[X:DSSO#XL1]SNGGTT`
16. An antibody Fab, encoding the disulfide bridge as `L-cystine (cross link)` from PSI-MOD

```
EVQLVESGGGLVQPGGSLRLSC[L-cystine (cross link)#XL1]AASGFNIKDTYIHWVRQAPGKGL
EWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYC[#XL1]SRWGGDGFYAMDYWG
QGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGC[L-cystine (cross link)#XL2]LVKDYFPEP
VTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYIC[#XL2]NVNHKPSNTKVDKKVEPKS
C[L-cystine (cross link)#XL3]DKT//DIQMTQSPSSLSASVGDRVTITC[L-cystine
(cross link)#XL4]RASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQ
PEDFATYYC[#XL4]QQHYTTPPTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVC[L-cystine
(cross link)#XL5]LLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHK
VYAC[#XL5]EVTHQGLSSPVTKSFNRGEC[#XL3]
```

Custom modifications

Custom modifications can be defined by opening the custom modifications section and hitting Create new. Add the chemical formula, if only the monoisotopic mass is known that can also be defined. The numerical ID is preset but the name must be set. For a custom modification metadata can be set, with description, synonyms, and identifiers for other identification systems (cross IDs).

Modifications can be defined as 'Modification' or 'Cross-linker'. For the first category placement rules can be defined that list the positions where this modification can be placed together with place-specific neutral losses and diagnostic ions.

For cross-linkers the length of the cross-linker can be defined as additional metadata item. Cross-linker placement rules can be defined as a symmetric cross-linker, which binds two identical positions on both sides of the cross-linker, or asymmetrical, where both sides bind different sides. For custom cross-linkers MS cleavage patterns can be defined. These are defined as two molecular formulas separated by `:`. If the theoretical fragmentation model allows MS cleavable cross-linkers these are allowed as breakage patterns. Additionally diagnostic ions can be added.

The custom modifications are stored in a separate JSON file on your computer, the path will be shown in the custom modification section. Updating the annotator will not remove any previously defined modifications. Additionally copying this file to another computer will copy the whole database to that computer. This can be used to move all your definitions to a new computer when upgrading or to aid a colleague with your definitions.

Custom model

Custom models can be defined by opening the custom model section and hitting Duplicate on any existing model. For all main ion series (`abcxyz`) control is given over the section of the peptidoform that produces these fragments.

Satellite ions series ('dvw') are formed at all location where the parent fragment ion are formed when turned on. The set of amino acid that gives rise to satellite ions can be controlled, if this field is left empty all amino acids will give rise to satellite ions. Additionally the maximal distance from the parent ion cleavage can be controlled. This is the number of side chains between the parent cleavage and the side chain that fragments. For example on a given peptidoform `HKSLG` the z3 fragment would contain `SLG`. The normal satellite ion w3 would be the side chain of Serine falling off (with a loss of 16 Da). The non-standard satellite ion 1w3 contains the same amino acids `SLG` but experiences a loss of the Leucine side chain in reference to the z3 fragment, with a mass of 43 Da. 1w3 means 1 side chain between the z3 cleavage and the w fragment site.

Neutral losses or gains can be Selected from a set of common losses/gains and custom losses/gains can be specified. If there are losses that only occur when a certain amino acid is present in the fragment this can also be specified. For example COOH loss is seen from Asparagine in ETD. Additionally, side chain losses can sometimes be seen, primarily in ETD, where any side chain in the fragment can be lost. The maximal number of side chains lost can be controlled, commonly this should be kept at 1 unless there is good evidence for losing multiple side chains from the same fragment. As well as the amino acids that give rise to side chain losses can be controlled, if this is not specified all amino acids are assumed to be able to loose their side chains.

The charge range for any series can be controlled. Both sides of the charge range can be set to an absolute value or a value relative to the precursor. A charge range of 1 to the precursor charge is the default for most ions.

For all peptidoform fragmentation series ('abcdvwxyz') the existence of variant ions can be controlled. Variant ions are fragments that do not follow the common pathway for that ion but end up with some hydrogen difference. For all ions all options between −2 hydrogens and + 2 hydrogens can be turned on. Variant ions are depicted using a single quote `'` for each hydrogen lost or a middle dot `·` for each hydrogen gained. For example `z·` indicates a z ion with one hydrogen gained and `c''` a c ion with two hydrogens lost.

For precursors the neutral losses/gain and charge range are available.

For glycans the neutral losses/gains can be set for all glycan fragment types. The expected losses from diagnostic glycan fragments of one monosaccharide can be set. For glycans that reside on peptidoform fragments the glycan can be fragmented further depending on the fragmentation technique. This can be controlled with custom rules. Each rules applies to a set of amino acids, for example Asparagine for N-glycans, and potentially a set of fragment kinds ('abcdvwxyz'). Each rule defines if contained glycans do not undergo fragmentation ('full') or if they do ('core'), and in the latter case a range can be given. This range is the minimal and maximal depth in the glycan structure for structural glycans, so 0-1 would indicate an absence of the full glycan or the presence of the first monosaccharide. For compositional glycans this range indicates the number of monosaccharides expected, so 0-1 would indicate complete loss of the glycan or the inclusion of one the possible monosaccharides, all options will be generated. The charge range can be set separately for Y and B ions. Additionally, the fragmentation of structures from GNO can be turned on/off as well as the generation of fragmentation from compositional glycans (e.g. Hex1HexNac2) can be controlled.

Some modifications generate specific diagnostic ions, these can be turned on or off. Additionally, some modifications generate specific neutral losses, which can also be turned on or off.

Immonium ions are internal fragments that break on both sides of a single amino acid. These can be turned on/off. For immonium ions there are a lot of common neutral losses and gains, these can be controlled separately for each amino acid.

If there are cross-links in the peptidoform it can be controlled if these are allowed to cleave in theoretical fragmentation. This only work for cross-link modifications that have defined cleavage rules, see Section 4.2.

There are some reference sheets available and the end of this manual to help keep an overview of all fragmentation chemistry. Or see douweschulte/reference-sheets on GitHub.

Interactive spectrum

Legend

Once an annotation is done the annotated spectrum will be shown below the annotation settings. It starts with a legend of the colours. For any of these legend elements if hovered over it highlights all peaks that match. If clicked this highlighting will stay until clicked again. Multiple highlights can be stacked which will show any peaks matching any of the criteria.

Peptidoform legend

Below the legend is an overview of the peptidoform(s) that are annotated. All main ion series are displayed as flags in the corner between the amino acids. Hovering highlights any fragment from that position similar to highlighting in the legend. Clicking makes the highlighting permanent in the same way. If there are multiple peptidoforms they are prefixed by the peptidoform index (peptidoform index followed by a dot followed by the peptidoform index). Hovering/clicking highlights all peptidoform fragments.

In the settings section the peptidoform can be set to use a more compact representation. Additionally, parts of the peptidoform (and any matching fragment) can be highlighted in a different colour.

Spectrum

In the spectrum hovering over a peak shows the mz value as well as all annotations if there are multiple annotations for that peak. Zooming in can be done by clicking at one point to select one mz bound and move the mouse and release it at another point to set the other mz bound as well as the intensity bound. A selection window will appear to indicate the section shown after zooming in. Additionally zooming can also be done with the scroll wheel. Scrolling zooms in or out on the spot where the mouse is located. Scrolling with `Shift` pans the spectrum to the left or right on the mz axis. Scrolling with `Control` zooms in on the intensity axis. Lastly, using `Control` with `+` and `-` zooms in or out on the middle of the spectrum, and `Control` with `0` zooms to the original zoom level of the spectrum. The distance between peaks can be annotated by clicking on one peak, dragging to the next and releasing on another peak. Such labels can be removed by clicking on them, or all labels can be removed at once in the settings.

Peaks settings

The theoretical spectrum can be shown below the x axis. Displaying the unassigned peaks can be toggled. The colouration can be set to the ion type (default), to the peptidoform, the peptidoform, or grey (none). Also there is an option to remove all distance labels.

Label settings

By default the labels for peaks are shown for any peak within the 90% of intensity, meaning any peak having an intensity of at least $(1.0 - 0.9) * \text{max\_intensity}$. This threshold can be controlled in the label settings section. Showing the mz value can be controlled in the same way for peaks. The 'Manually force show' allows to show either the label or mz value for any peak. Select the mode of interest and any peak clicked will show that information. The 'Hide' mode allows to hide the information (label and mz) for any peak. The button Clear clears all manually forced

labels. The checkboxes 'Show in label' allow the hiding/showing of certain parts of the label to slim down the information shown. Lastly, the labels can be set to be 90 degrees rotated to make overlap less likely. Any of these changes will also show up in the exported image.

Spectrum settings

This section allows zooming to precise numbers. Additionally, the number of tickmarks for the x and y axis can be set. The peaks can be set to use square root intensity instead of linear intensity to see a bigger dynamic range. The y axis can be changed to show percentages instead of raw intensities as well.

Error graph

The graph below the spectrum by default shows the ppm error for all theoretical peaks to their experimental matches. The points are at the same location on the x axis and any zooming in the spectrum will zoom in the error graph as well. In the settings section the points can be set to reflect the peak intensity with the point diameter and the range of y axis can be controlled.

Using the settings the y axis can be changed from relative (ppm) to absolute (Thompson). Additionally, the x axis can be set to reflect errors of unassigned peaks instead of the default assigned peaks. The errors of the unassigned peaks are calculated as being the error to the closest theoretical fragment for all unassigned peaks. In this mode at least one ion series needs to be picked to work as reference series. Commonly the unassigned mode is used together with absolute error mode.

Using the unassigned mode is most helpful to detect a series of peaks that is a constant offset from the reference peaks. This indicates that the assignment is off just before the constant offset series. See Figure 1 for an example where an annotation from the N terminus matched the reference series but has a constant offset from a certain point. This indicates that at that point the actual peaks are 16 Thompson off from the reference series likely indicating an oxidation. This way of using the error graph is most helpful when annotating de novo top down or middle down sequences to determine where the annotation deviates from the experimental evidence.

Figure 1: Example error graph in absolute error mode with the errors for unassigned peaks relative to some N terminal ion series. Visible is that four peaks match perfectly but right after that four additional peaks have a constant offset of 16 Thompson to the reference series.

Statistics overview

The statistic overview gives detailed statistics about the match. Listed are:

1. The precursor mass for each peptidoform. If a peptidoform is part of a peptidoform no mass is listed.
2. The number of fragments found out of all theoretical fragments.
3. The number of peaks annotated out of all peaks.

4. The total intensity annotated out of the sum of the intensity of all peaks.
5. The sequence positions covered by at least one matched fragment. If some positions do not generate theoretical fragments, for example with cross-linking loop links, this statistic will be split into one regarding the full peptidoform and one only taking the possible locations into account.
6. Peaks false match chance a false discovery metric at the peaks level. Estimating the FDR by permutation. The spectrum is shifted by −25 to +25 Da plus π (to have non integer offsets) and the number of matches is counted. The percentage is the number of actual matches divided by the average found for the shifted spectra. The number between brackets denotes the number of standard deviations the actual matches are from the shifted matches. This number is the average chance for a single peak annotation to be based on chance. For some rough guidelines, 10% or less is commonly seen for bottom up data while up to 50% can be seen for top or middle down data.
7. Intensity false match chance. This is based on the same calculation as the peaks false match chance, However this does not count the number of peaks matched but the sum intensity. This number should be similar or lower compared to the peaks false match chance.

The toggle switch above the table allows to open ion series specific statistics. These are exactly the same as above but split per ion series.

Fragment table

This table contains all details on the spectrum in table format. It can display unassigned peaks, annotated peaks, and missing fragments. For each peak/fragment it displays all data. This whole table can be copied to other software for other analysis. Additionally, above the table is a normalised output for the ProForma definition, which removes any implementation specific notation and returns a fully specification compliant ProForma sequence.

Settings

This section below the general statistics allows fine tuning the annotated spectrum and related sections. The graphic section allows fine control over the graphics. These boxes allow sizes set in any CSS unit. The other settings sections are detailed in their related sections above.

Tools

Search modifications

This can be used to find modifications in any of the ontologies (modification databases) or find more information on a single modification. The search box can be used as follows:

1. When a mass is given any modification that fits within the tolerance is returned. For example searching for `16` gives the following results:

Name	Id	monoisotopic mass	Formula
U:Oxidation	UNIMOD:35	15.995 Da / 15.999 Da / 15.995 Da	O1
U:Methyl:2H(2)	UNIMOD:284	16.028 Da / 16.039 Da / 16.028 Da	C12H2
U:Carboxy->Thiocarboxy	UNIMOD:420	15.977 Da / 16.068 Da / 15.977 Da	O-1S1
U:Ala->Ser	UNIMOD:540	15.995 Da / 15.999 Da / 15.995 Da	O1

2. When a formula is given any modification that has the same molecular formula is returned. For example searching for `Formula:O` gives the following results:

Name	Id
U:Oxidation	UNIMOD:35
U:Ala->Ser	UNIMOD:540
U:Phe->Tyr	UNIMOD:569
M:(2S,3R)-3-hydroxyasparagine	MOD:35

3. When a glycan composition is given any glycan with the same composition is returned (all isomeric information is ignored), which makes it possible to find topologies for given glycan compositions. Another way of finding glycan compositions based on composition is using the GlyCosmos GNOme structure browser.
4. When a modification is given the details for that modification are displayed.

Isotopic distribution

Here the isotopic distribution for any molecular formula can be generated. All weights of the formula will be displayed as well as a graph with all isotopic peaks. Hovering over the peaks gives additional details on that specific peak. For the generation of the isotopic distribution an averagine model is used that slightly overestimates the prevalence of higher weight isotopes, especially for elements with multiple isotopes. Any element with a defined monoisotopic weight, so any element that is stable and naturally occurring, can be used in the formula.

Exporting

Vector graphic

To export an annotated spectrum click the export button in the top left corner. This opens the system print dialog. Make sure to enable background graphics otherwise some elements may be missing in the export. Choose print to pdf and save the file. This file can be opened in any vector graphics application for manual touch ups.

Data

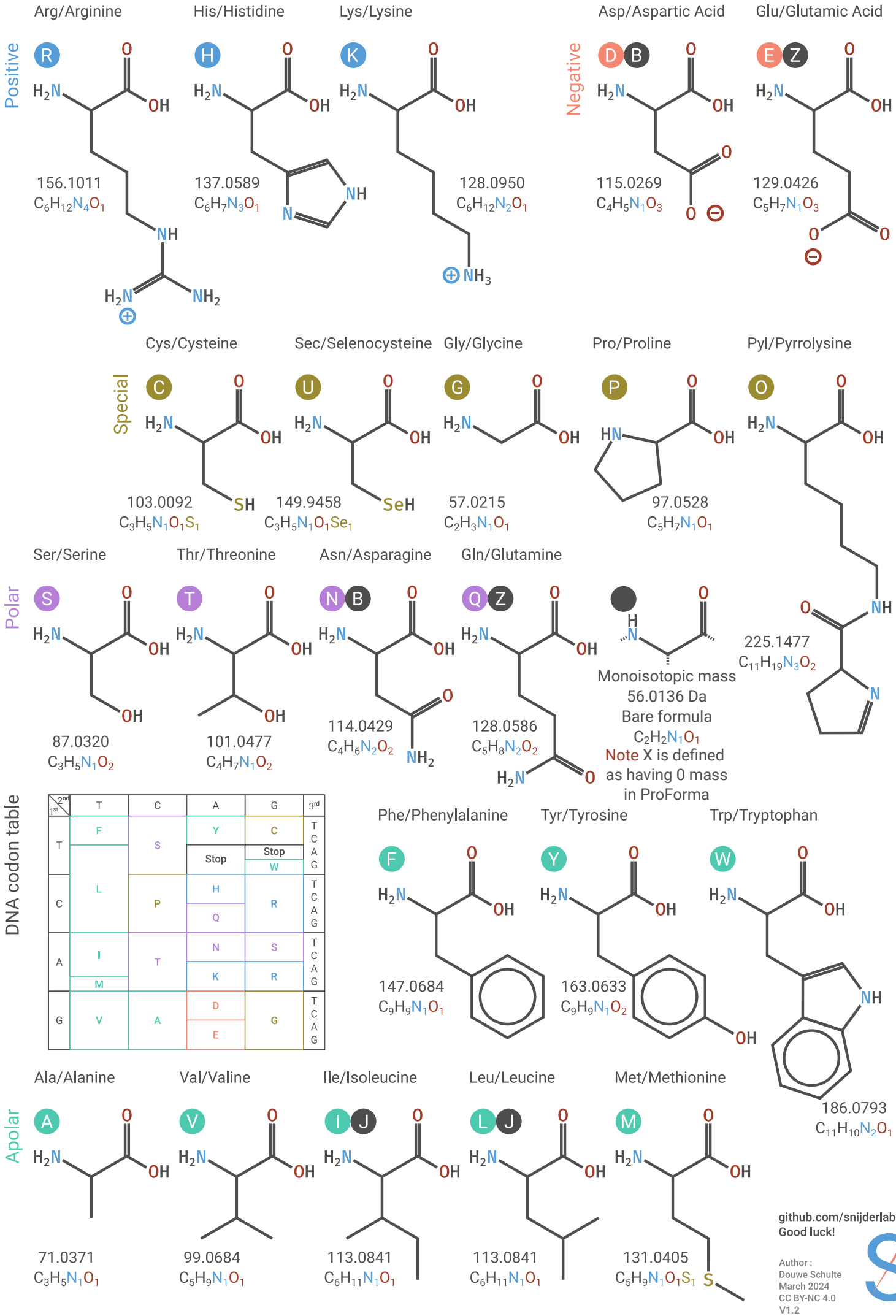
Once a spectrum is annotated the selected spectrum can be saved as a mzSpecLib, MGF, or mzML file. mzSpecLib files contain the spectrum and the peak annotations. MGF and mzML contain only the spectrum.

Ontologies

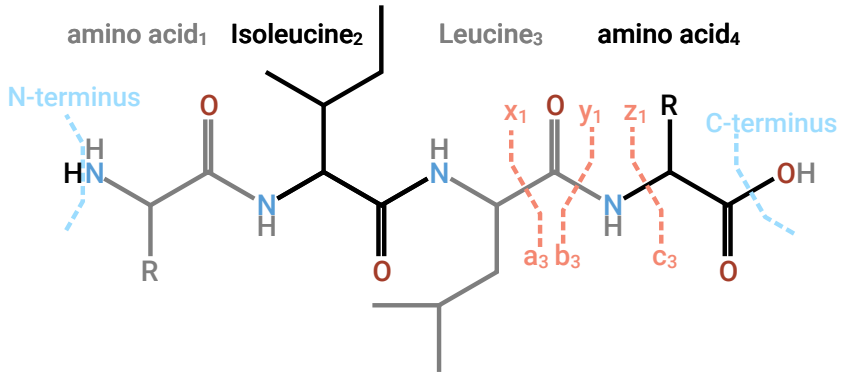
In the last section the current state of the ontologies can be seen. For each ontology the version and last updated date are displayed (if set properly by the ontology). The SHA256 of the loaded file is shown alongside the number of contained modifications. Most ontologies can be updated by clicking on the Internet, this automatically downloads the ontologies from the canonical locations, parses the files and updates the data. Any error will be shown below the table. For more advanced use cases or when the ontologies have moved (or for RESID which cannot be automatically downloaded) a local file can be specified to update the ontologies with. Note that it depends on the ontology how many files and of which file type have to be supplied. Any of these files can also be supplied in gzipped form when supplied with the `.<ext>.gz`. If the update was successful the ontology is directly updated and can immediately be used and will

be in the exact same state when the annotator is closed and opened again. Note that during the updating itself no other calculations can be done with the Annotator and that depending on the internet speed it might take a bit, especially for GNOme as that contains quite a big collection of glycans.

Ontology	File(s)
Unimod	unimod.xml (not unimod_tables.xml)
PSI-MOD	PSI-MOD.obo
GNOme	gno.obo and glycosmos_glycans_list.csv
XLMOD	XLMOD.obo
RESID	RESIDUES.XML

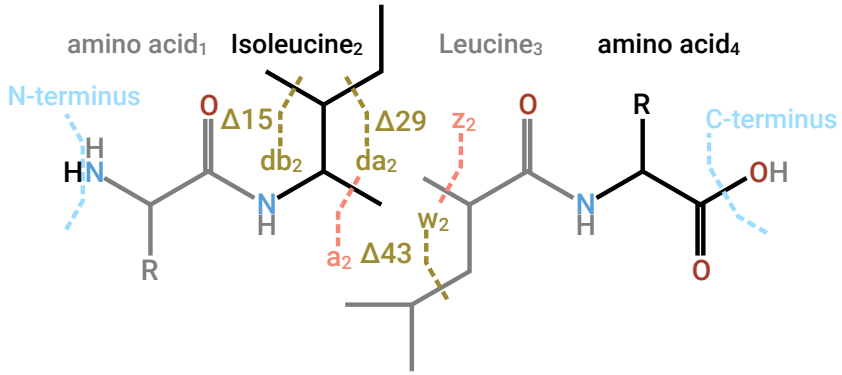

Arg/Arginine
His/Histidine
Lys/Lysine
Asp/Aspartic Acid
Glu/Glutamic Acid
Positive
Negative
156.1011
C6H12N4O1
137.0589
C6H7N3O1
128.0950
C6H12N2O1
115.0269
C4H5N1O3
129.0426
C5H7N1O3
Cys/Cysteine
Sec/Selenocysteine
Gly/Glycine
Pro/Proline
Pyl/Pyrrolysine
Special
103.0092
C3H5N1O1S1
149.9458
C3H5N1O1Se1
57.0215
C2H3N1O1
97.0528
C5H7N1O1
225.1477
C11H19N3O2
Ser/Serine
Thr/Threonine
Asn/Asparagine
Gln/Glutamine
Polar
87.0320
C3H5N1O2
101.0477
C4H7N1O2
114.0429
C4H6N2O2
128.0586
C5H8N2O2
Monoisotopic mass
56.0136 Da
Bare formula
C2H2N1O1
Note X is defined
as having 0 mass
in ProForma
DNA codon table
Phe/Phenylalanine
Tyr/Tyrosine
Trp/Tryptophan
147.0684
C9H9N1O1
163.0633
C9H9N1O2
186.0793
C11H10N2O1
Ala/Alanine
Val/Valine
Ile/Isoleucine
Leu/Leucine
Met/Methionine
Apolar
71.0371
C3H5N1O1
99.0684
C5H9N1O1
113.0841
C6H11N1O1
113.0841
C6H11N1O1
131.0405
C5H9N1O1S1
github.com/snijderlab
Good luck!
Author :
Douwe Schulte
March 2024
CC BY-NC 4.0
V1.2

Backbone fragmentation [1,2]



	Formula (h = H⁺ × z)	Mass (h = 1.008 × z)
a	$^{n}\Sigma$(AA) − CO + h	$^{n}\Sigma$(AA) − 27.995 + h
b	$^{n}\Sigma$(AA) + h	$^{n}\Sigma$(AA) + h
c	$^{n}\Sigma$(AA) + NH_3 + h	$^{n}\Sigma$(AA) + 17.027 + h
d	$^{n-1}\Sigma$(AA) + C_2H_3N + h	$^{n-1}\Sigma$(AA) + 41.027 + h
v	$^{c-1}\Sigma$(AA) + C_2H_2NO + h	$^{c-1}\Sigma$(AA) + 56.014 + h
w	$^{c-1}\Sigma$(AA) + C_3H_3O + h	$^{c-1}\Sigma$(AA) + 55.018 + h
x	$^{c}\Sigma$(AA) + CO − H_2 + h	$^{c}\Sigma$(AA) + 25.979 + h
y	$^{c}\Sigma$(AA) + H_2O + h	$^{c}\Sigma$(AA) + 18.011 + h
z	$^{c}\Sigma$(AA) + O − NH + h	$^{c}\Sigma$(AA) + 0.984 + h

Side chain fragmentation / satellite ions [2,3]



Variant ions [1,2,3]

	"/-2	'/-1	Base	·/+1	··/+2
a			aAU	U	U
b			aAEhU		U
c		e	aAeEU	E	e
d			aAU		
v			aAeU		
w		e	aAeEU	e	
x		U	aAU	U	U
y	U	U	aAeEhU		
z		e	aAeEU	eE	e

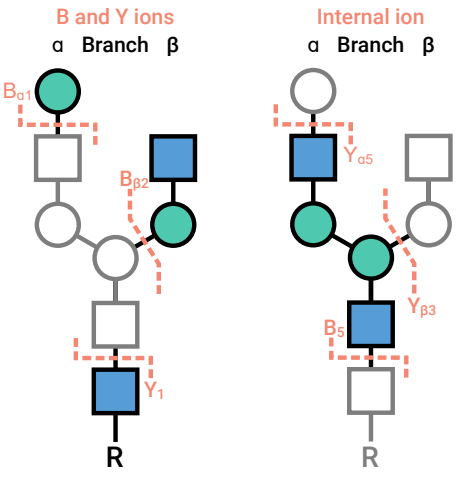
Variant ions are ion series differing by one or two hydrogen atoms in mass from the main ion series. The occurance of these is dependent on the fragmentation technique used.

a=EAD
A=EACID
e=ExD[4-8]
E=EtHCD
h=HCD
U=UVPD[9-12]

Satellite ions are formed as secondary fragmentation, d from an a ion, w from a z ion. For Threonine and Isoleucine there are two possible d and w ions. These can be listed as 'd/wa_n' and 'd/wb_n' where a is the heaviest of the two options. The leaving groups for Threonine are CH_3 for wa and OH for wb, for Isoleucine the leaving groups are CH_3 for wa and C_2H_5 for wb. Glycine, alanine, and proline have no satellite ions. The v ion is full loss of the sidechain from a y ion. For ETD and ECD, side chain loss similar to v ions can be observed from many fragments possibly even with multiple side chains lost[4]. Additionally in ECD, w fragmentation can be found in sidechains away from the backbone cleavage[5,6]. Denoted with ^{d}w, where d indicates the distance from the backbone cleavage, '^{0}w', the standard w ion, is written as 'w'.

Glycan [13]

Collectively B and Internal ions are known as Oxonium ions



Ion	Formula	Mass	Ion	Formula	Mass
H-$C_2H_6O_3$	$C_4H_5O_2$	85.0284	S-H_2O	$C_{11}H_{16}NO_7$	274.0921
H-CH_6O_3	$C_5H_5O_2$	97.0284	S	$C_{11}H_{18}NO_8$	292.1027
H-2H_2O	$C_6H_7O_3$	127.0390	G-H_2O	$C_{11}H_{16}NO_8$	290.0870
H-H_2O	$C_6H_9O_4$	145.0495	G	$C_{11}H_{18}NO_9$	308.0976
H	$C_6H_{11}O_5$	163.0601	H+N	$C_{14}H_{24}NO_{10}$	366.1395
N-$C_4H_8O_4$	C_4H_6NO	84.0444	2N	$C_{16}H_{27}N_2O_{10}$	407.1660
N-$C_2H_6O_3$	$C_6H_8NO_2$	126.0550	H+2N	$C_{22}H_{37}N_2O_{15}$	569.2188
N-CH_6O_3	$C_7H_8NO_2$	138.0550	H+N+S	$C_{25}H_{41}N_2O_{18}$	657.2349
N-$C_2H_4O_2$	$C_6H_{10}NO_3$	144.0655			
N-2H_2O	$C_8H_{10}NO_3$	168.0655			
N-H_2O	$C_8H_{12}NO_4$	186.0761			
N	$C_8H_{14}NO_5$	204.0867			

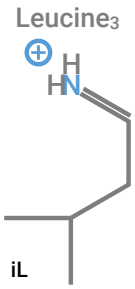
H=Hex
N=HexNAc
S=Neu5Ac
G=Neu5Gc

Formula & Mass are MH+

14

Immonium

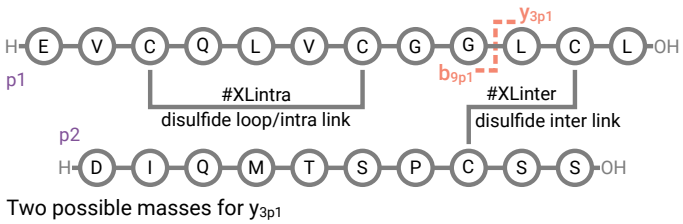
Formed by double backbone cleavage. Commonly fragments further by neutral loss. These fragments indicate existence of certain amino acids, but not all amino acids form immonium ions. The immonium ions have a mass between 0 and 160 Da.
Formula: amino acid - CO + H



Amino Acid	Formula	Mass	Amino Acid	Formula	Mass
A	C_2H_6N	44.050	K	$C_5H_{13}N_2$	101.108
R	$C_5H_{13}N_4$	129.114	M	$C_4H_{10}NS$	104.053
N	$C_3H_7N_2O$	87.056	F	$C_8H_{10}N$	120.081
D	$C_3H_6NO_2$	88.040	P	C_4H_8N	70.066
C	C_2H_6NS	76.022	S	C_2H_6NO	60.045
E	$C_4H_8NO_2$	102.056	T	C_3H_8NO	74.061
Q	$C_4H_9N_2O$	101.071	W	$C_{10}H_{11}N_2$	159.092
G	CH_4N	30.034	Y	$C_8H_{10}NO$	136.076
H	$C_5H_8N_3$	110.072	V	$C_4H_{10}N$	72.081
I/L/J	$C_5H_{12}N$	86.097	Formula & Mass are MH+		

Cross-links

Note that if you want to see breaking cross-linkers you will need to define your own custom linkers in the Annotator as the breaking is not defined in any of the used modification databases.



Two possible masses for y_{3p1}

y_{3p1}└─┘#XLinter LCL + OH - $H_{(broken\ disulfide)}$

y_{3p1}└──┘#XLinter LCL + OH - $2H_{(intact\ disulfide)}$ + p2

One possible mass for b_{9p1}

b_{9p1}└──┘#XLintra EVCQLVCGG + H - $2H_{(disulfide)}$

This disulfide is a loop link, so it could break either in the linker or in any of the contained amino acids without changing the final mass of the fragment.

To annote these and more on your own data see: github.com/snijderlab/annotator

Author :
Douwe Schulte
March 2025

References
[1] 10.1002/bms.1200111109
[2] 10.1016/0076-6879(90)93460-3
[3] 10.1016/0168-1176(88)80060-0
[4] 10.1016/S1387-3806(03)00202-1
[5] 10.1021/ja035042c
[6] 10.1021/ac0619332
[7] 10.1002/jms.3919
[8] 10.1021/ja035042c
[9] 10.1002/chem.201103534
[10] 10.1007/s13361-015-1297-5
[11] 10.1021/jacs.6b05147
[12] 10.1021/acs.chemrev.9b00440
[13] 10.1007/BF01049915
[14] 10.1016/j.trac.2018.09.007

Amino acids

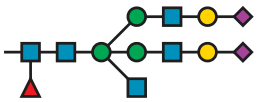
PEPTIDE
PEPTJDE
PE(?PTI)DE

Amino acids are written as the standard IUPAC symbols. With O/U added as pyrolysine, and selenocysteine. J/B/Z are ambiguous amino acids for I/L, N/D, and Q/E respectively. Lastly X is added as any unknown amino acid. X is defined to have a mass of zero. So 'X[+435]' can be used to indicate a mass gap of 435 Da.

Modifications

Modifications are defined between square brackets '[mod]'. Multiple definitions can be combined with a vertical line '|' like so '[mod|mod|mod]'.

Monoisotopic mass	PEPTI[+15.995]DE
Molecular formula	PEPTI[Formula:O]DE
Name (Unimod/PSI-MOD)	PEPTI[Oxidation]DE
Name specific ontology	PEPTI[U:Oxidation]DE
Index specific ontology	PEPTI[UNIMOD:35]DE
Comment or description	PEPTI[INFO:text]DE
Glycan composition	PEPN[Glycan:HexNAc5Hex5dHex1NeuAc2]ITDE
Glycan structure	PEPN[GNO:G75079FY]ITDE → Encodes:
Charged molecular formula	PEPTI[Formula:H-2Zn1:z+2]DE



Modification prefixes

Mass	Name	Index	Description
U	U?	Unimod	Modifications from Unimod
M	M?	PSIMOD	Modifications from PSI MOD
R	R	Resid	Modifications from Resid
X	X	XLMOD	Modifications from XL-MOD
G	G	GNO	Modifications from GNOme
Obs	-	-	To indicate an observed mass
-	INFO	-	To add comments or description
-	Glycan	-	Glycan compositions
-	Formula	-	Molecular formulas

Locations

Localised

PEPT[mod][mod]IDE	(Multiple) modifications on the side chain of an amino acid
[mod][mod]-PEPTIDE	(Multiple) modifications on the N terminus
PEPTIDE-[mod][mod]	(Multiple) modifications on the C terminus
{mod}PEPTIDE	Labile modification, on this peptidoform
<mod@E,N-term>PEPTIDE	Fixed modification

Unlocalised

[mod]^n?PEPTIDE	On the entire peptidoform
PEPT[mod#name]ID[#name]E	On the given locations
PEP(TID)[mod]E	Within the given stretch

The order of all options when used together

<mod@E>[mod]^n?{mod}[mod]-PEP-[mod]

Using unlocalised rules

PEP(TID[mod])[mod|Position:T,D|CoMKP]E

Unlocalised additional rules

Position:E,N-term	Limit the allowed positions, uses the same positions as fixed modifications
Limit:2	Limit the maximal number on one site, only for ^n global unlocalised modifications
CoMKP	Allows colocalising with placed modifications
CoMUP	Allows colocalising with other unlocalised modifications

Combining peptidoforms

PEPTIDE+PEPTIDE	Chimeric peptidoforms (forming one compound peptidoform ion)
PEPT[mod#XLname]IDE//PEPT[#XLname]IDE	Cross-linked peptidoforms (forming one peptidoform ion)
PEPT[mod#BRANCH]IDE//PEP[#BRANCH]TIDE	Branched peptidoforms (forming one peptidoform ion)
PEPT[mod#XLname]IDE[#XLname]	Intra linked cross-linked peptidoform

Charge

PEPTIDE/2	Global charge, assumed to be protons
PEPTIDE/[Na:z+1,H:z+1]	Defined charge carriers, in this case 1 unlocalised sodium and 1 unlocalised proton
PEP[Formula:Na:z+1]TIDE/1	Defined charge carriers, in this case 1 localised sodium and 1 unlocalised proton, this still has a global charge of 2

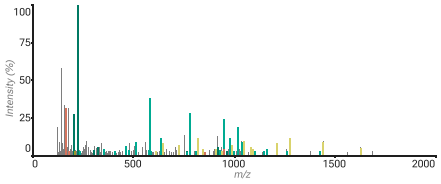
Name

To add a description or metadata to an entire (compound) peptidoform (ion). To do this for single locations or stretches use the INFO tag.

(>Peptidoform name)PEPTIDE
(>>Peptidoform ion name)PEPT[mod#XLname]IDE//PEPT[#XLname]IDE
(>>>Compound peptidoform ion name)PEPTIDE+PEPTIDE

Universal Spectrum Identifier

mzspec:<ID>:<file>:scan:<number>:<peptidoform>
index:<number>
nativeId:<number>(,<number>)*



mzspec:PXD023419:Peng2021_Herceptin_aLP.raw:scan:11368:RWGGDGFYAM[Oxidation]DYWGQGTLVTV/3

USIs are the URL for the mass spectrometry world. A USI points to a Proteomics Exchange Identifier (PXD...), a file, and a spectrum in that file. Additionally, a peptidoform can be added to the end, note that many servers require the peptidoform and a global charge state for the peptidoform.

ProForma v2.1
psidev.info/proforma
USI
psidev.info/usi

Author :
Douwe Schulte
May 2025

Bibliography

(1) Schulte, D.; Leuvenink, R. W.; Jager, S.; Heck, A. J. R.; Snijder, J. A Universal Spectrum Annotator for Complex Peptidoforms in Mass Spectrometry-Based Proteomics. Analytical Chemistry **2025**, acs.analchem.5c02832. https://doi.org/10.1021/acs.analchem.5c02832.

(2) Schulte, D.; Snijder, J. A Handle on Mass Coincidence Errors in De Novo Sequencing of Antibodies by Bottom-up Proteomics. Journal of Proteome Research **2024**, 23 (8), 3552–3559. https://doi.org/10.1021/acs.jproteome.4c00188.