analytical chemistry

pubs.acs.org/ac

Article

The Proteomics Standards Initiative Standardized Formats for Spectral Libraries and Fragment Ion Peak Annotations: mzSpecLib and mzPAF

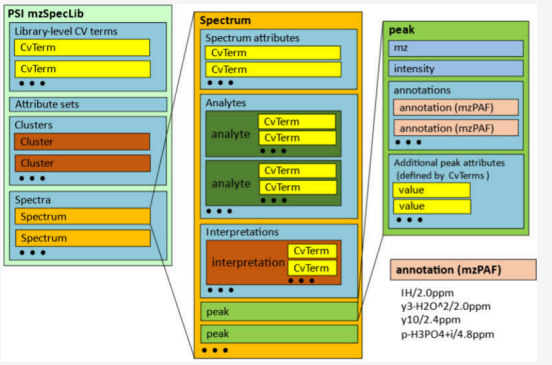
Joshua Klein, Henry Lam, Tytus D. Mak, Wout Bittremieux, Yasset Perez-Riverol, Ralf Gabriels, Jim Shofstahl, Helge Hecht, Pierre-Alain Binz, Shin Kawano, Tim Van Den Bossche, Jeremy Carver, Benjamin A. Neely, Luis Mendoza, Tomi Suomi, Tine Claeys, Thomas Payne, Douwe Schulte, Zhi Sun, Nils Hoffmann, Yunping Zhu, Steffen Neumann, Andrew R. Jones,* Nuno Bandeira, Juan Antonio Vizcaíno, and Eric W. Deutsch*

Cite This: https://doi.org/10.1021/acs.analchem.4c04091

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Mass spectral libraries are collections of reference spectra, usually associated with specific analytes from which the spectra were generated, that are used for further downstream analysis of new spectra. There are many different formats used for encoding spectral libraries, but none have undergone a standardization process to ensure broad applicability to many applications. As part of the Human Proteome Organization Proteomics Standards Initiative (PSI), we have developed a standardized format for encoding spectral libraries, called mzSpecLib (https://psidev.info/mzSpecLib). It is primarily a data model that flexibly encodes metadata about the library entries using the extensible PSI-MS controlled vocabulary and can be encoded in and converted between different serialization formats. We have also developed a standardized data model and serialization for fragment ion peak annotations, called mzPAF (https://psidev.info/mzPAF). It is defined as a separate standard, since it may be used for other applications besides spectral libraries. The mzSpecLib and mzPAF standards are compatible with existing PSI standards such as ProForma 2.0 and the Universal Spectrum Identifier. The mzSpecLib and mzPAF standards have been primarily defined for peptides in proteomics applications with basic small molecule support. They could be extended in the future to other fields that need to encode spectral libraries for nonpeptidic analytes.



INTRODUCTION

Mass spectral libraries (or "spectral libraries" for short hereafter) are collections of fragment ion mass spectra and their originating analytes (if known) that are intended to be used as a reference for future spectral analysis. An MS2 spectrum records the characteristic fragmentation patterns of an analyte, in terms of mass-to-charge ratios and abundances of fragment ions, that reproducibly occur when the analyte is fragmented in the mass spectrometer by one of various methods. Because of this, mass spectral libraries have important applications in analytical chemistry, particularly in the fields of proteomics and metabolomics/lipidomics.[1−4]

In proteomics, spectral libraries are generally compiled from a large amount of data acquired on complex protein mixtures.[5] Often the data were acquired to answer various biological questions as a primary goal rather than to generate spectral libraries. In metabolomics and lipidomics, spectral libraries are more often generated by injecting pure analytes onto the mass spectrometer, thereby conclusively linking the MS2 spectrum and the analyte.[6] This requires the dedicated effort by library builders, most notably the National Institute of Standards and Technology (NIST) of the United States Government, some academic groups (e.g., METLIN spectral libraries)[7] and vendors of mass spectrometers.

Especially in proteomics, applications also exist for libraries that contain unidentified MS2 spectra (exclusively or not). Such spectral libraries are sometimes referred to as spectral archives.[8] For spectral archives to be useful, spectra are often grouped by spectral similarity into clusters by some clustering algorithm. It is often, but not always, assumed that spectra in a cluster originate from the same analyte or from highly related analytes, on

Received: August 3, 2024
Revised: October 16, 2024
Accepted: November 1, 2024

account of their spectral similarity. It should also be noted that in recent years, due to the advances in artificial intelligence, it has become possible to predict the characteristic fragmentation patterns of the analytes of interest, for both peptides and small molecules.[9,10] Such predicted spectra can also be compiled into *in silico* spectral libraries and utilized in a way similar to traditional spectral libraries built from real data.

In proteomics, although sequence database searching is still the method of choice in data-dependent acquisition (DDA) experiments, spectral library searching offers notable advantages.[11] First, they enable faster identification due to a smaller search space of observable peptides. Second, these libraries can increase the specificity and sensitivity of peptide identification tools by using known fragment ion intensities rather than solely relying on the calculated mass-to-charge ratios of the expected fragment ions. On the other hand, in data-independent acquisition (DIA) experiments, where the fragmentation of multiple precursors simultaneously leads to convoluted and hard-to-interpret spectra, spectral libraries play a crucial role in data analysis, in particular in peptide-centric approaches that seek to monitor anticipated fragment ions coeluting over time.[12] Similarly, practitioners of targeted quantitative workflows such as selected reaction monitoring (SRM) depend on spectral libraries to design MS-based "assays" prior to the experiment, so that the mass spectrometer can be instructed to isolate and quantify specific fragment ions.[13]

Peak annotations are useful information for many applications of spectral libraries. Annotation refers to the assignment of observed peaks in a spectrum to the fragment ions responsible for them and can be obtained manually or computationally. The peak annotations (or lack thereof) provide the best indication of whether the observed spectrum is of high quality and can be adequately explained by analyte identification. For example, a spectrum with a large number of unannotated peaks is a telltale sign that the spectrum may be incorrectly identified or is contaminated with another analyte. Some peaks in a spectrum, such as those resulting from cleavages of isobaric labeling reagents for quantifications, have special uses and should be clearly marked. Additionally, spectral library search engines can assign different weights to the matching of fragment ion peaks of different types due to their different information content.[5,14] However, since each spectrum may have hundreds of peaks, peak annotations can account for a large fraction of the storage space and parsing time of a spectral library. Therefore, encoding peak annotations compactly but unambiguously is a highly desirable feature of any spectral library format.

Regardless of the source of the libraries and their intended applications, there is a clear need for various metadata associated with the library and its constituent spectra. There have been for many years several different popular spectral library formats in proteomics, including the MSP format used by NIST, the splib format used by SpectraST,[5,15] the blib format used by Bibliospec,[16] simple yet *ad hoc* tab-separated formats for DIA applications, and others. Similarly, in metabolomics, the MSP format and the Mascot Generic Format (MGF) are commonly used. While these formats are fit for the purpose of storing the spectra themselves, they all lack an adequate and well-documented mechanism for providing rich metadata about the spectra, including their provenance. Several of these formats store limited metadata using *ad hoc* keywords in a comment field. There are common conventions, but these have evolved over the years with no formal documentation or consensus among tools or domains. As spectral libraries become ever more important for mature and emerging applications, a need for a community standard spectral library format with extensive metadata has been identified.[17] It is recognized that a standardized format for spectral libraries will greatly facilitate the systematic compilation, dissemination, and utilization of spectral libraries in the research community.

The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) has developed multiple open community data standards (e.g., mzML, mzidentML, mzTab), guidelines, and controlled vocabularies/ontologies, among other outputs, since its inception in 2002.[18−20] Here we describe the HUPO PSI mzSpecLib and mzPAF standards designed to encode spectral libraries and annotations of peptide fragment ion peaks, respectively. They have been split into separate standards since fragment ion peaks may be annotated in contexts outside of spectral libraries, such as spectrum display software. However, they are described here together since they will most often be jointly used. We first provide an overview of the mzSpecLib format and then the mzPAF format followed by descriptions of current software implementations that may guide further implementations.

METHODS

The development of mzSpecLib and mzPAF started in 2019. Since then, it has been an open process via conference calls, in addition to discussions at the annual PSI meetings and smaller workshops. Both technical specification documents were submitted to the PSI Document Process[21] for review, during which time external anonymous reviewers not involved in the development process provided their feedback. The document was also made available for comments by the public, enabling broad input on the specifications. The latest versions of the specification documents, up-to-date information on software implementations, and information on future versions of mzSpecLib and mzPAF are available at https://psidev.info/mzSpecLib and https://psidev.info/mzPAF, respectively.

Historically the PSI has developed open and standardized data formats mostly for proteomics applications that have focused on storing mass spectra as generated by an instrument (the mzML format),[22] and formats that have focused on storing the downstream identifications of those generated mass spectra (the mzIdentML[23,24] and mzTab[25] formats). These formats have been kept separate intentionally, with the identification formats designed to refer to the spectrum-containing format. This was done primarily to control file sizes since the serialization of spectra takes substantial space and replicating the spectra in the identification format seemed inefficient.

Spectral libraries are designed with an opposite approach: to unite the spectra and their interpretations into a single format. The significant difference, however, is that spectral libraries are intended to contain single or aggregated spectra that are deemed important, collated across potentially many experiments, to form a reference set of spectra that may be used for downstream processing. Spectral libraries are, however, not intended to capture the complex output of mass spectrum processing pipelines (for which mzIdentML is designed).

During the development process, the working group identified the following design principles for the mzSpecLib format. First, it was decided that the standard should take the form of a common data model with multiple interconvertible serializations that can serve different needs. For example, if human readability is desired, a text-based serialization would be the most suitable, while a JSON file would be best for software-

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

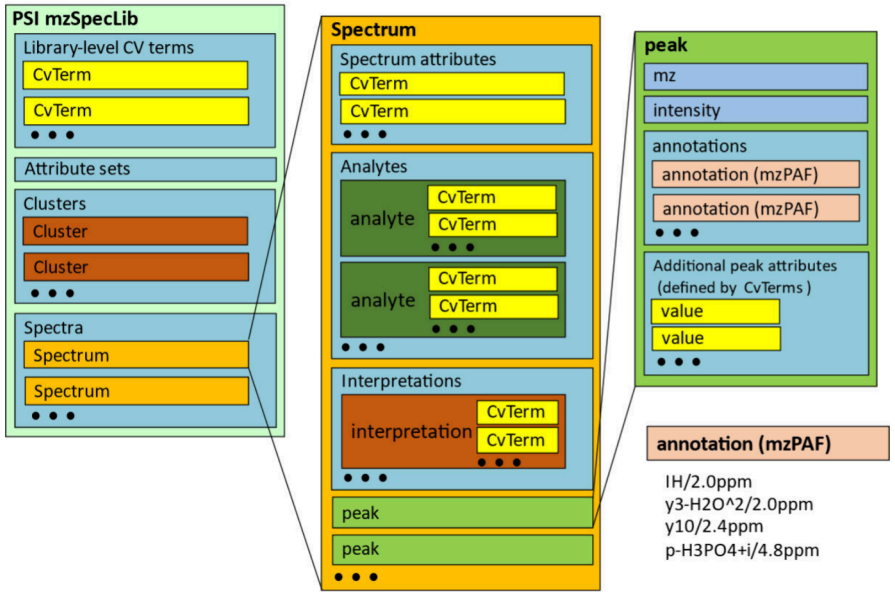


Figure 1. Overview of the mzSpecLib data model. The model consists of six main components: the library metadata, clusters, spectra, analytes, peaks, and the peak annotations in mzPAF.

application data exchange. These two serializations are implemented at the present stage, but more can be added in the future. This design choice is unique among PSI formats but is necessary to reduce the barrier of entry and to address the sometimes-conflicting demands of library builders and users. We expect that applications that demand efficiency will need to develop their own serialization and indexes since it is difficult for all needs to be anticipated. For example, a serialization that optimizes for random access may not be optimal for storing and reading data for machine learning applications. Defining mzSpecLib to be a data model rather than a file format is an important first step in ensuring that different serializations developed by anyone can be interconverted.

Second, all metadata are defined via terms from a controlled vocabulary (CV), primarily the PSI-MS,[26] which is actively maintained by this PSI working group. This forces the producers and consumers of spectral libraries to use mutually agreed terminology with clear definitions to refer to the same concept. For concepts that are outside the scope of the PSI-MS, e.g., sample metadata terms, it is permissible to use compact URIs (CURIEs) (https://www.w3.org/TR/2010/NOTE-curie-20101216/) from other CVs, e.g., EFO:0000434|ecotype = EFO:0005148|Col-0 from the Experimental Factor Ontology (EFO),[27] with prefixes from bioregistry.io.[28] In addition, to ensure flexibility and extensibility, there was no formal restriction in the specification as to what terms may or may not be used in a spectral library as long as they are defined in the CV. However, a table of commonly used terms is provided to guide the adoption of best practices in a separate "living" document maintained online as an addendum to the format. Third, the format should provide a mechanism to minimize redundancy and file size by not encoding repetitive metadata in every spectrum. Meanwhile, any mechanism to do so must also ensure that the data integrity of the library can be maintained as libraries are split or merged and as library entries are added or deleted. Finally, to set a manageable scope for the project, the inaugural version of the format will primarily support proteomics applications with basic small molecule support. Nevertheless, care has been taken so that the overall design can be extended to other MS fields in the future (metabolomics, lipidomics, glycomics, etc.).

For mzPAF, the working group identified the following design principles for the encoding of peak annotations. First, mzPAF should be designed as a compact, human-readable, and unambiguous software-parsable format. Second, the format is also mapped to an object model that may be serialized as a JSON for certain applications. Third, to set a manageable scope for the project, mzPAF is designed for linear peptides with simple modifications, i.e., those routinely identified by typical proteomics pipelines, and for fragmentation methods commonly used in proteomics such as collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), and electron-transfer dissociation (ETD). Finally, as in the case of mzSpecLib, although there are some provisions for annotating small molecules, it is expected that for other major classes of analytes (small molecules, glycans, lipids, glycopeptides, cross-linked peptides, etc.), extended peak annotation formats should be defined in the future, ideally compatible with this format.

RESULTS AND DISCUSSION

mzSpecLib Format. The mzSpecLib specification primarily takes the form of a data model with multiple possible serializations of that data model. The specification describes two serializations: text-based serialization and JavaScript Object Notation (JSON) serialization. Other, potentially more space-efficient serializations for additional use cases are envisioned and are in progress. Lossless interconversion between these serialization methods is provided by a reference implementation Python library (https://github.com/HUPO-PSI/mzspeclib-py), and other implementations are welcome and forthcoming.

The data model consists of six main components (Figure 1). The top level contains basic metadata that describe the library itself. At this level, attribute sets can be defined for entry-level metadata that are shared by many library entries. Optionally, a list of spectrum clusters follows, which consists of references (links) to the library entries belonging to each spectrum cluster. This cluster feature is intended to capture the output of spectrum clustering algorithms,[29−33] which collect sets of similar

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

PSI mzSpecLib Text Format:

```
<Spectrum=1>
MS:1003061|library spectrum name=AAAACALTPGPLADLAAR/2_1(4,C,CAM)_46eV
MS:1003065|spectrum aggregation type=MS:1003066|singleton spectrum
MS:1000044|dissociation method=MS:1000422|beam-type collision-induced dissociation
[1]MS:1000045|collision energy=46
[1]UO:0000000|unit=UO:0000266|electronvolt
MS:1003057|scan number=5538
MS:1003203|constituent spectrum file="CHO-K1_bRPLC_C1.RAW.FT.hcd.ch.MGF"
MS:1003070|number of replicate spectra used=1
MS:1003069|number of replicate spectra available=2
MS:1000002|sample name="jhu_cho_brplc_cam"
MS:1000028|detector resolution=7500
[2]MS:1000828|isolation window lower offset=0.95
[2]UO:0000000|unit=MS:1000040|m/z
[3]MS:1000829|isolation window upper offset=0.95
[3]UO:0000000|unit=MS:1000040|m/z
MS:1003085|previous MS1 scan precursor intensity=8799173.32
MS:1003086|precursor apex intensity=25273307.5
MS:1000512|filter string="FTMS + p NSI d Full ms2 855.96@hcd35.00 [140.00-1725.00]"
MS:1003059|number of peaks=87
MS:1000744|selected ion m/z=855.4538
[4]MS:1003275|other attribute name=Se
[4]MS:1003276|other attribute value=1(^G1:sc=8.13346e-015)
<Analyte=1>
MS:1000224|molecular mass=1710.9076
MS:1000888|stripped peptide sequence=AAAACALTPGPLADLAAR
MS:1000041|charge state=2
MS:1003208|experimental precursor monoisotopic m/z=855.4550
MS:1001117|theoretical mass=1708.89303961159
[1]MS:1001975|delta m/z=1.4
[1]UO:0000000|unit=UO:0000169|parts per million
MS:1003169|proforma peptidoform sequence=AAAAC[Carbamidomethyl]ALTPGPLADLAAR
[2]MS:1003048|number of enzymatic termini=2
[2]MS:1001045|cleavage agent name=MS:1001251|Trypsin
[2]MS:1001112|n-terminal flanking residue=R
[2]MS:1001113|c-terminal flanking residue=L
[2]MS:1000885|protein accession=tr|G3IJB9|G3IJB9_CRIGR UDP-N-acetylhexosamine pyrophosphorylase-like protein 1 OS=Cricetulus griseus GN=I79_023952 PE=4 SV=1
<Interpretation=1>
MS:1003079|total unassigned intensity fraction=0.2848
MS:1003080|top 20 peak unassigned intensity fraction=0.1879
MS:1003289|intensity of highest unassigned peak=0.45
MS:1003290|number of unassigned peaks among top 20 peaks=4
<Peaks>
143.0823 14791.5 b2/5.6ppm
153.2575 5008.6 ?
159.0917 11531.8 ?
162.5977 5804.6 ?
169.0972 12931.0 ?
175.1193 18211.1 y1/2.0ppm,IR/2.0ppm
```

NIST MSP:

```
Name: AAAACALTPGPLADLAAR/2_1(4,C,CAM)_46eV

Comment:

HCD=46eV

Origfile="CHO-K1_bRPLC_C1.RAW.FT.hcd.ch.MGF"

Nreps=1/2

Sample="jhu_cho_brplc_cam"
FTResolution=7500
ms2IsolationWidth=1.90

ms1PrecursorAb=8799173.32
Precursor1MaxAb=25273307.50
Filter="FTMS + p NSI d Full ms2 855.96@hcd35.00 [140.00-1725.00]"

Parent=855.4538
Se=1(^G1:sc=8.13346e-015)

MW: 1710.9076

Charge=2
PrecursorMonoisoMZ=855.4550

Mz_diff=1.4ppm

Single Pep=Tryptic Mods=1(4,C,CAM)
Fullname=R.AAAACALTPGPLADLAAR.L

Scan=5538
Protein="tr|G3IJB9|G3IJB9_CRIGR UDP-N-acetylhexosamine pyrophosphorylase-like protein 1 OS=Cricetulus griseus GN=I79_023952 PE=4 SV=1"

Unassign_all=0.2848
Unassigned=0.1879
max_unassigned_ab=0.45
num_unassigned_peaks=4/20
Num peaks: 87
143.0823 14791.5 "b2/5.6ppm"
153.2575 5008.6 "?"
159.0917 11531.8 "?"
162.5977 5804.6 "?"
169.0972 12931 "?"
175.1193 18211.1 "y1/2.0ppm,IRJ/2.0ppm"
```

Figure 2. A side-by-side comparison of a spectrum from NIST's chinese_hamster_hcd spectral library (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:cho_20180223) in MSP format and the equivalent spectrum in mzSpecLib's text encoding. Note that in the MSP file all of the lines between "Comment" and "Num Peaks" are key-value pairs within a single Comment line but separated out here for better readability. The MSP spectrum comments are color-mapped to their equivalent controlled vocabulary terms. Wide lines that are wrapped for display purposes are marked with a wrapped arrow symbol.

spectra ("clusters") thought to be derived from the same or similar precursor ions, even when the identity of not all precursor ions may be known.

The next component describes each individual spectrum entry, both the metadata about each spectrum and a list of peaks. Spectrum metadata describes attributes inherent to the spectrum and is accompanied by metadata that describe the one or more putative analytes that produced it and metadata describing the interpretations that combine one or more of the putative analytes. The spectra themselves may be individual

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

PSI mzSpecLib Text Format:

```
<Spectrum=1>
MS:1003061|library spectrum name=AAAQWVR/2_0
MS:1003065|spectrum aggregation type=MS:1003067|consensus ↪
spectrum
[1]MS:1000045|collision energy=21.53
[1]UO:0000000|unit=UO:0000266|electronvolt
[2]MS:1000138|normalized collision energy=29.82
[2]UO:0000000|unit=UO:0000187|percent
MS:1003070|number of replicate spectra used=38
MS:1003069|number of replicate spectra available=55
MS:1003059|number of peaks=68
[3]MS:1003275|other attribute name=Tissue
[3]MS:1003276|other attribute value={{Skin,55}}
[5]MS:1003275|other attribute name=Quality
[5]MS:1003276|other attribute value=7/7
[6]MS:1003254|peak attribute=MS:1003279|observation
frequency of peak
MS:1000744|selected ion m/z=401.2219
<Analyte=1>
MS:1000888|stripped peptide sequence=AAAQWVR
MS:1000041|charge state=2
[1]MS:1001975|delta m/z=0.4
[1]UO:0000000|unit=UO:0000169|parts per million
MS:1003169|proforma peptidoform sequence=AAAQWVR
MS:1001117|theoretical mass=800.4293214295599
[2]MS:1003048|number of enzymatic termini=2
[2]MS:1001045|cleavage agent name=MS:1001251|Trypsin
[2]MS:1001112|n-terminal flanking residue=K
[2]MS:1001113|c-terminal flanking residue=D
[2]MS:1000885|protein accession=sp|Q8NEX9|DR9C7_HUMAN
[2]MS:1003044|number of missed cleavages=0
<Interpretation=1>
MS:1003079|total unassigned intensity fraction=0.3572
MS:1003080|top 20 peak unassigned intensity fraction=0.264
MS:1003288|number of unassigned peaks=37
MS:1003289|intensity of highest unassigned peak=0.66
MS:1003290|number of unassigned peaks among top 20 peaks=7
MS:1002354|PSM-level q-value=0.0063
<Peaks>
120.0803    48745.9    ?    0.7636
129.0655    43194.2    IQ/-2.7ppm    0.6727
129.1020   297962.0    ?    1
130.0647    85559.3    IW/-3.3ppm    0.8182
130.0859   131952.1    ?    0.9636
130.1053     8821.4    ?    0.4
136.0752   118308.8    ?    1
141.1015     6896.1    ?    0.4
```

NIST MSP:

```
Name: AAAQWVR/2_0
Comment:
Consensus
CE=21.53
NCE=29.82
Nrep=38/55
Tissue={{Skin,55}}
Quality=7/7
Parent=401.2219
Charge=2
Theo_mz_diff=0.4ppm
Mods=0
Fullname=K.AAAQWVR.D
MCtype=Normal
Pep=Tryptic Peptype=<Protein><Peptide><Protein>
Nprot=1
Protein="sp|Q8NEX9|DR9C7_HUMAN(pre=K,post=D)"
MC=0
Unassigned_all_20ppm=0.3572
Unassigned_20ppm=0.2640
num_unassigned_peaks_20ppm=37
max_unassigned_ab_20ppm=0.66
top_20_num_unassigned_peaks_20ppm=7/20
Q-value=0.0063
Num peaks: 68
120.0803    48745.9    "?    42/55"
129.0655    43194.2    "IQD/-2.7ppm    37/55"
129.1020   297962.0    "?    55/55"
130.0647    85559.3    "IWD/-3.3ppm    45/55"
130.0859   131952.1    "?    53/55"
130.1053     8821.4    "?    22/55"
136.0752   118308.8    "?    55/55"
141.1015     6896.1    "?    22/55"
```

Figure 3. A side-by-side comparison of a spectrum from NIST IARPA3_best_tissue_add_info spectral library (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:human_skin_hair) in MSP format and the equivalent spectrum in mzSpecLib's text encoding. See https://github.com/HUPO-PSI/mzSpecLib/blob/master/examples/NIST/IARPA3_best_tissue_add_info.head.mzlib.txt for a more complete example. Wide lines that are wrapped for display purposes are marked with a wrapped arrow symbol.

observed spectra, consensus spectra derived from multiple individual input spectra, or predicted spectra based on one of the many available algorithms. Most libraries comprise collections of consensus spectra that are derived from DDA data, but libraries may also be derived from DIA data, in which fragment ion peaks are selected based on having an ion trace that is compatible in profile (i.e., maximum retention time and shape) with other member fragment ion peaks and/or being a predicted fragment of the putative analyte.

The analytes are precursor ions (typically charged peptidoforms in this initial implementation but potentially extensible to other kinds of charged molecules). The interpretations are the groups of analytes that explain the spectrum. In the simplest and most common case, a single interpretation includes one analyte. However, the format supports additional cases such as the definition of analytes 1, 2, and 3, with two possible interpretations for the spectrum, one of which is a blend of analytes 1 + 2 and the other is a blend of analytes 2 + 3.

The next component describes each peak, including the m/z values, intensities, and aggregation metadata (for when a spectrum is an aggregated spectrum and aggregation metrics are available).

The final component describes one or more proposed annotations for each peak. This information is encoded in the mzPAF peak annotation standard, as described below and found at http://psidev.info/mzPAF.

The text serialization of mzSpecLib is somewhat similar to the popular NIST MSP format but is formally standardized, which MSP never was. Furthermore, it adds important features such as support for spectral clustering output, support for unidentified

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

```
<mzSpecLib>
MS:1003186|library format version=1.0
MS:1003188|library name=IARPA3_best_tissue_add_info
MS:1003191|library
URI=https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:human_skin_hair
MS:1001017|release date=Oct. 01, 2021
MS:1003203|constituent spectrum file=mzspec:USI000000:6donors_sp3/am_24_rg_l1_2021-08-
13_380-2000_120_hcd30_255min_sp3_lysctryp_i_pos.raw
MS:1003203|constituent spectrum file=mzspec:USI000000:6donors_sp3/am_25_rg_l1_2021-08-
13_380-2000_120_hcd30_255min_sp3_lysctryp_i_pos.raw
MS:1003203|constituent spectrum file=mzspec:USI000000:6donors_sp3/am_26_rg_l1_2021-08-
13_380-2000_120_hcd30_255min_sp3_lysctryp_i_pos.raw
<Spectrum=1>
...
```

Figure 4. mzSpecLib library file header snippet taken from the conversion of the NIST IARPA3_best_tissue_add_info spectral library (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:human_skin_hair) to mzSpecLib text format. This header was produced by incorporating the metadata from an external text file (https://chemdata.nist.gov/download/peptide_library/libraries/skin_hair/IARPA3_all.out.zip) because the MSP format had no suitable representation for the relevant details. See https://github.com/HUPO-PSI/mzSpecLib/blob/master/examples/NIST/IARPA3_best_tissue_add_info.head.mzlib.txt for a more complete example.

(but commonly observed) spectra, support for multiple analytes per spectrum, and alternative hypotheses for the interpretation of each spectrum. It also explicitly makes use of the PSI-MS controlled vocabulary, rather than using somewhat *ad hoc* keywords in the comment field that were never properly defined. Thus, while the formal mzSpecLib specification is expected to remain fixed for a long time, its use can be sufficiently flexible via the externally maintained controlled vocabulary. This principle is analogous to that followed in all PSI data standards.

Each of the aspects mentioned above are documented in extensive detail—often with a dedicated subsection to each aspect—in the formal mzSpecLib specification available at https://psidev.info/mzspeclib. Implementers and interested readers are referred to the full specification document for further details.

In Figure 2 and Figure 3, we show spectra from different NIST spectral libraries and their equivalent mzSpecLib text renderings, highlighting the mapping between the MSP key-value pairs and the controlled vocabulary terms. By cataloging many MSP files from NIST and other sources, we identified synonymous or conditionally equivalent keys. For example, at least eight distinct term keys were found that corresponded to the analyte precursor m/z, six terms for scan polarity, and eight terms for collision energy (CE). For example, in Figure 2 the HCD key is equivalent to the CE key in Figure 3.

In some cases, a single MSP comment implies multiple attributes, as in the case of "Nreps" mapping to "MS:1003070|number of replicate spectra used" (accession of the CV term in the PSI-MS controlled vocabulary and its name) and "MS:1003069|number of replicate spectra available", and implying that peak aggregation of type "MS:1003254|peak attribute = MS:1003279|observation frequency of peak" should be present but may not be available.

mzSpecLib is more precise with respect to concepts that may seem interchangeable in other contexts. For example, the m/z of the selected ion (MS:1000744|selected ion m/z), the theoretical m/z of the annotated analyte (MS:1003053|theoretical monoisotopic m/z), and the experimentally determined monoisotopic peak m/z (MS:1003208|experimental precursor monoisotopic m/z) for the analyte or selected ion might all be called the precursor m/z, but they mean different things. By having explicit contexts, mzSpecLib makes it clear which piece is the subject of each attribute, especially when there are multiple interpretations or multiple analytes per spectrum. The specification also provides recommendations for using standardized formats for defining analytes, using ProForma 2.0[34] for peptides, and either SMILES or InChI for small molecules. The ProForma 2.0 specification allows the description of quite complex post-translational modifications and other artifactual mass modifications, including modification localization ambiguities.

The mzSpecLib text format uses sections to explicitly organize information into different contexts, using "<", the section opening expression, followed by ">" to denote a new section. The format begins with a header section, "<mzSpecLib>", containing library-level metadata, often missing in plain-text formats, describing attributes like where the library came from or was made, its identifiers, and license if that information is available (Figure 4). This header may include information about which raw mass spectrometry data files were included in the library using the Universal Spectrum Identifier (USI)[35] notation for data files. It is not currently standard practice in the field to provide the false discovery rate (FDR) of a spectrum library, although every library probably has at least some false positives. It is encouraged that producers of mzSpecLib libraries provide a global FDR of a library in the header via a term such as MS:1002350|PSM-level global FDR = 0.nnn when it is known. Furthermore, the specification allows the writer to describe the procedure and software tools used to generate FDR (or q-value) estimates, e.g., by including terms such as MS:1001456|analysis software. The library header may also declare sets of shared attributes to reduce repetition in sections below it, in the style of the cascading configuration.

Next, an mzSpecLib library can list any number of "<Cluster = ${key}>" entries, describing related groups of spectra inside or outside the library, followed by any number of "<Spectrum = ${key}>" entries, describing library spectra including any acquisition or generation attributes, interpretations as sets of zero or more analytes, and a peak list that may include peak annotation or aggregation statistics. All of these entities are defined either in terms of controlled vocabulary attributes or via an externally defined grammar-data model pair permitted by the specification.

mzPAF Format. In the mzSpecLib format, it is possible to annotate individual peaks, as is already done in spectral libraries from NIST (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:start), SpectraST,[5] and PeptideAtlas.[36] However, there have been several different styles of annotation in the

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

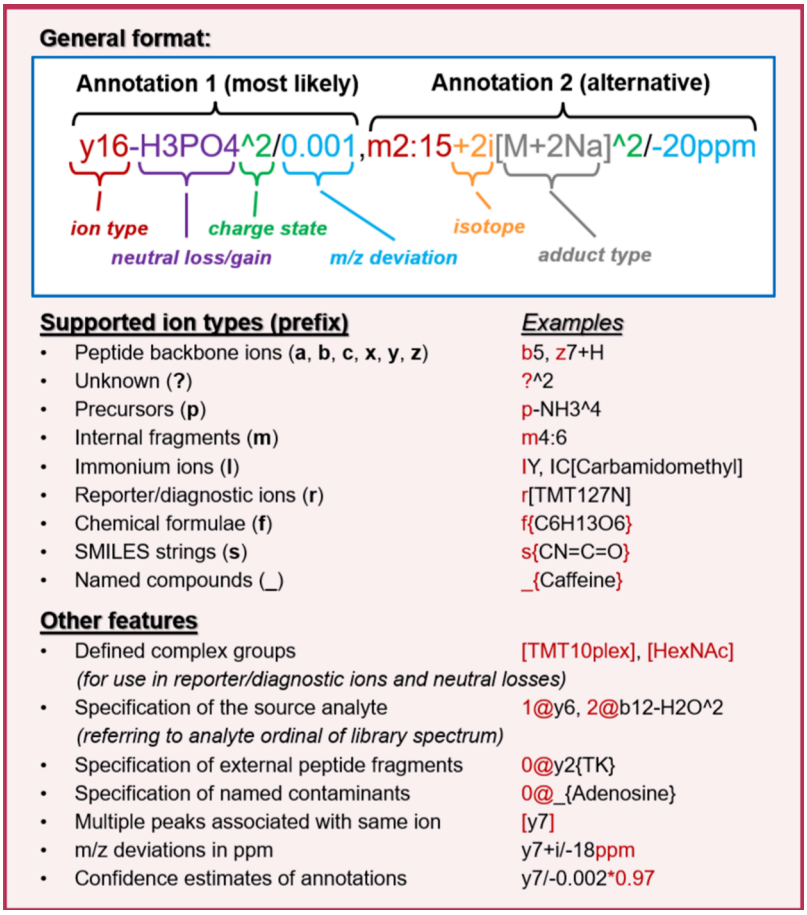


Figure 5. Main features of the mzPAF format. An example peak annotation with common components is shown along with a list of supported ion types (with prefix in parentheses) and other features of the format. Detailed descriptions can be found in the documentation.

past (even from a single provider). To address this, mzPAF describes a single common peak annotation format standard for peptides that is recommended for all peptide libraries and related applications for which peak annotations are desirable.

The mzPAF specification is heavily based on the formatting used in the NIST MSP format and the SpectraST sptxt format. These precursor formats were quite similar but not exactly the same and were never fully documented. NIST MSP annotations have undergone small changes over the years. Those annotation formats are based on the original nomenclature proposals published by Roepstorff and Fohlman,[37] which was further refined by Biemann.[38] Participation from NIST and SpectraST developers and many others has led to the development of this unified standard.

The mzPAF format, as described in the Methods, is designed for linear peptides with simple modifications, i.e., those routinely identified by typical proteomics pipelines (with or without enrichment methods), and for fragmentation methods commonly used in proteomics. However, there are some provisions for annotating small molecules (e.g., contaminants in a predominantly peptide spectrum) as well as unusual fragments.

Figure 5 highlights the main aspects of mzPAF with an example and a list of major features. At the top are two example annotations, separated by a comma, indicating that one or both are potential interpretations of a peak, with the first deemed most likely. The first annotation describes a doubly charged y16 ion (a chain of 16 residues beginning at the C terminus of a peptide) with an H_3PO_4 neutral loss (a common neutral loss for a phosphorylated peptide) somewhere on those 16 residues. The m/z delta between measured m/z and predicted m/z is 0.001. The second annotation shows a second isotope of a doubly charged internal fragmentation ion ("m" for middle) of residues 2 through 15 (counting from the N terminus) that acquired its charge via two sodium ions. The delta of measured m/z minus computed m/z for the annotated fragment is −20 ppm.

The central panel of Figure 5 lists the mzPAF-supported ion types, including the customary abcxyz designations, "m" for internal fragmentation ions, "?" for unidentified ions, "p" for precursor-related ions, "I" for immonium ions, "r" for reference ions (such as isobaric labeling related ions), "f" for arbitrary chemical formulas, "s" for SMILES strings, and "_" for named compounds. Examples of each of these ion types are also provided.

The bottom panel of Figure 5 lists some notable features of the standard. Certain CV terms from a list that accompanies the specification or the Unimod CV[39] for protein modifications may be specified in square brackets. An analyte number may be specified before each annotation (e.g., "1@" or "2@") for cases where more than one analyte may be present in the spectrum ("0@" indicates an unspecified contaminant ion). Such contaminant ions may take a few different forms, such as "0@y2{TK}" to indicate a y2 type ion from some unspecified ion (ending with TK on the C terminus), or "0@_{Adenine}" to indicate a fragment ion corresponding to a nucleotide fragment perhaps from an RNA molecule that entered the selection

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

window (common in phospho-enriched data sets). If multiple peaks gather the same annotation since they are closer together than the annotation m/z tolerance, the less likely duplicates may be enclosed in square brackets. The m/z delta values are always expressed as measured minus computed, with default units as m/z and deltas in parts per million listed with a suffix of "ppm". Finally, the confidence of one or more annotations may be specified with a "*" followed by a number between 0 and 1, inclusive (e.g., *0.97). It may be used to indicate the relative confidence between alternative explanations or to indicate a confidence that a proposed explanation is truly the source of measured peak (perhaps based on a model of the m/z deltas for the spectrum and likelihood that such a fragment ion would be present).

All of these features are described in substantially greater detail—often in a dedicated subsection—in the formal specification document available at https://psidev.info/mzPAF/. Although the compact text notation described in Figure 5 and the text above is the most common serialization for spectral libraries and annotated-spectrum depictions, a formal object model that may be serialized in JSON is also described in the specification and may be useful in software implementations and transfer of information between software implementations.

Not only is the compact mzPAF text format human readable (with some expertise), it is also unambiguously software-parsable, via either a regular expression or a state machine parser. Such parsers are already implemented in the Python mzPAF package (https://github.com/HUPO-PSI/mzpaf), and also for several other languages. It is hoped that the vast majority of software tool writers can use existing parsing code without needing to implement their own parser (which is a nontrivial task). As explained in the Methods, it is expected that for other major classes of analytes extended peak annotation formats should be defined in the future.

Resources and Implementations. Multiple resources and implementations are already available for mzSpecLib and mzPAF, and it is expected that additional resources and implementations will follow in parallel with the adoption of the format. The reference implementations that have been developed in parallel with the specification, and used to test the specification, are the Python packages mzspeclib and mzpaf, which are available on PyPI and in their GitHub repositories at https://github.com/HUPO-PSI/mzspeclib-py and https://github.com/HUPO-PSI/mzpaf, respectively. These packages provide a complete set of classes and methods needed to read and write both mzSpecLib and mzPAF annotations. Furthermore, the mzspeclib package provides a spectral library data conversion tool that is able to convert various other formats (such as the NIST MSP format) into mzSpecLib files and the reverse as well for a subset of formats. This allows most existing libraries to be converted to mzSpecLib, and mzSpecLib libraries to be converted to another format, for example, in order to be used with a software package that does not yet support mzSpecLib natively.

In addition to the specification documents and related reference files, there are example files in the GitHub repositories listed above. The examples at the mzPAF site provide a series of example spectra that have been annotated with text serialization. They include examples of varying complexity including complex phosphopeptide spectra and a chimeric spectrum. The examples also include an annotated small molecule spectrum from MassBank,[40] using features that are thus far available in the mzSpecLib and mzPAF basic support available for small molecules.

CONCLUSION

We have presented here an overview of mzSpecLib, the new PSI spectral library format, and mzPAF, the new PSI peak annotation format. These data standards focus on the capability of encoding extensive metadata in a standardized mechanism. The standards attempt to be highly similar to what is already common practice (as seen in the commonly used MSP format) but with carefully documented and standardized specifications that have been debated and refined in an open community-based process.

These new standards also support features that were not available in previous formats, thereby providing standardization as well as additional functionality. We expect that these new standards will further accelerate the recent growth of spectral libraries in support of DIA, DDA, and SRM workflows by making high-quality reference libraries and reusable software components more broadly available. In the context of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles,[41] it is essential to have open formats for spectral libraries with the ability to trace library elements back to the original MS source files (the mass spectra), including published sources with USIs,[35] analytes, InChI, SMILES, or protein accessions, enabling full transparency and traceability. The FAIR principles are further promoted when the spectral libraries provide the links to original public data sets with PXD identifiers and SDRF-Proteomics[42] files that fully describe the samples and how the samples and data files are associated. Libraries should themselves be deposited and made publicly accessible in ProteomeXchange repositories with centrally managed identifiers for the libraries. Widespread adoption of the mzSpecLib and mzPAF standards in proteomics software would further accelerate the interoperability among those software tools.

There are many subfields of mass spectrometry, including cross-linking MS and glycoproteomics, that have special needs and potentially huge complexity both in the description of analytes as well as the fragment ion peaks. While we have produced a flexible set of formats that provide a basic foundation for supporting these types of data, practitioners in such subfields should come together if they are sufficiently motivated to produce best practice documents and examples that demonstrate how the many complexities in their spectra can be encoded via mzSpecLib and mzPAF in a common way.

PSI standards are developed via an open process in which all interested individuals and groups are encouraged to participate. Although data standards that are cooperatively developed inevitably take longer to complete, the resulting standards are more broadly applicable to many more use cases than those from independent initiatives. Broad participation is therefore essential for the successful generation of future standards for the proteomics community. See https://psidev.info/ for information about how to contribute to the PSI activities.

AUTHOR INFORMATION

Corresponding Authors

Andrew R. Jones − *Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 3BX, United Kingdom;* orcid.org/0000-0001-6118-9327; Email: Andrew.Jones@liverpool.ac.uk

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

Eric W. Deutsch − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0001-8732-0928; Email: edeutsch@systemsbiology.org

Authors

Joshua Klein − *Program for Bioinformatics, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0000-0003-1279-6838

Henry Lam − *Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, 999077 Hong Kong, P. R. China;* orcid.org/0000-0001-7928-0364

Tytus D. Mak − *Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0001-9501-5640

Wout Bittremieux − *Department of Computer Science, University of Antwerp, 2020 Antwerpen, Belgium;* orcid.org/0000-0002-3105-1359

Yasset Perez-Riverol − *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom;* orcid.org/0000-0001-6579-6941

Ralf Gabriels − *VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9052 Ghent, Belgium;* orcid.org/0000-0002-1679-1711

Jim Shofstahl − *Thermo Fisher Scientific, San Jose, California 95134, United States;* orcid.org/0000-0001-5968-1742

Helge Hecht − *RECETOX, Faculty of Science, Masaryk University, 60200 Brno, Czech Republic;* orcid.org/0000-0001-6744-996X

Pierre-Alain Binz − *Lausanne University Hospital, CH-1011 Lausanne, Switzerland;* orcid.org/0000-0002-0045-7698

Shin Kawano − *Database Center for Life Science, Joint Support Center for Data Science Research, Research Organization of Information and Systems, Chiba 277-0871, Japan; School of Frontier Engineering, Kitasato University, Sagamihara 252-0373, Japan;* orcid.org/0000-0002-7969-2972

Tim Van Den Bossche − *VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9052 Ghent, Belgium;* orcid.org/0000-0002-5916-2587

Jeremy Carver − *Center for Computational Mass Spectrometry, Department of Computer Science and Engineering, University of California, San Diego, California 92093-0404, United States;* orcid.org/0000-0003-0384-8130

Benjamin A. Neely − *National Institute of Standards and Technology (NIST) Charleston, Charleston, South Carolina 29412, United States;* orcid.org/0000-0001-6120-7695

Luis Mendoza − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0003-0128-8643

Tomi Suomi − *Turku Bioscience Centre, University of Turku and Åbo Akademi University, FI-20520 Turku, Finland;* orcid.org/0000-0003-3639-979X

Tine Claeys − *VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9052 Ghent, Belgium;* orcid.org/0000-0001-9408-488X

Thomas Payne − *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom*

Douwe Schulte − *Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, 3584, CH Utrecht, The Netherlands;* orcid.org/0000-0003-0594-0993

Zhi Sun − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0003-3324-6851

Nils Hoffmann − *Institute for Bio- and Geosciences (IBG-5), Forschungszentrum Jülich GmbH, 52428 Jülich, Germany;* orcid.org/0000-0002-6540-6875

Yunping Zhu − *National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China;* orcid.org/0000-0002-7320-7411

Steffen Neumann − *Computational Plant Biochemistry, Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany; German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig 04103 Leipzig, Germany;* orcid.org/0000-0002-7899-7192

Nuno Bandeira − *Center for Computational Mass Spectrometry, Department of Computer Science and Engineering, University of California, San Diego, California 92093-0404, United States; Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0001-8385-3655

Juan Antonio Vizcaíno − *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom;* orcid.org/0000-0002-3905-4335

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c04091

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge all other individuals who contributed in various ways to the mzSpecLib and mzPAF specifications. This work was funded in part by the National Institutes of Health grants R01 GM087221 (EWD), R24 GM148372 (NB,EWD), U19 AG023122 (RLM), U24 DK133658 (NB), and by the National Science Foundation grants DBI-2324882 (EWD) (DIAeXchange), DBI-1933311 (EWD) (PTMeXchange), and IOS-1922871 (EWD) (*Arabidopsis*). JAV and YPR would like to acknowledge BBSRC grant BB/X001911/1 (DIAeXchange), APP9749, and the EPSRC grant EP/Y035984/1. HH thanks the RECETOX Research Infrastructure (LM2023069) financed by the Ministry of Education, Youth and Sports, and the Operational Programme Research, Development and Education (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17 043/0009632) for supportive background. This work was supported from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857560. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

Government. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

REFERENCES

(1) Griss, J. *Proteomics* **2016**, *16* (5), 729−740.
(2) Shao, W.; Lam, H. *Mass Spectrom. Rev.* **2017**, *36* (5), 634−648.
(3) Stein, S. *Anal. Chem.* **2012**, *84* (17), 7274−7282.
(4) Bittremieux, W.; Wang, M.; Dorrestein, P. C. *Metabolomics Off. J. Metabolomic Soc.* **2022**, *18* (12), 94.
(5) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. *Nat. Methods* **2008**, *5* (10), 873−875.
(6) Brungs, C.; Schmid, R.; Heuckeroth, S.; Mazumdar, A.; Drexler, M.; Šácha, P.; Dorrestein, P. C.; Petras, D.; Nothias, L.-F.; Nencka, R.; Kameník, Z.; Pluskal, T. Efficient Generation of Open Multi-Stage Fragmentation Mass Spectral Libraries. *ChemRxiv*, May 10, 2024. DOI: 10.26434/chemrxiv-2024-l1tqh.
(7) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. *Nat. Commun.* **2019**, *10* (1), 5811.
(8) Frank, A. M.; Monroe, M. E.; Shah, A. R.; Carver, J. J.; Bandeira, N.; Moore, R. J.; Anderson, G. A.; Smith, R. D.; Pevzner, P. A. *Nat. Methods* **2011**, *8* (7), 587−591.
(9) Gabriel, W.; The, M.; Zolg, D. P.; Bayer, F. P.; Shouman, O.; Lautenbacher, L.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Huhmer, A.; Wenschuh, H.; Reimer, U.; Médard, G.; Kuster, B.; Wilhelm, M. *Anal. Chem.* **2022**, *94* (20), 7181−7190.
(10) Declercq, A.; Bouwmeester, R.; Chiva, C.; Sabidó, E.; Hirschler, A.; Carapito, C.; Martens, L.; Degroeve, S.; Gabriels, R. *Nucleic Acids Res.* **2023**, *51* (W1), W338−W342.
(11) Zhang, X.; Li, Y.; Shao, W.; Lam, H. *Proteomics* **2011**, *11* (6), 1075−1085.
(12) Jones, A. R.; Deutsch, E. W.; Vizcaíno, J. A. *Proteomics* **2023**, *23* (7−8), No. e2200014.
(13) Kusebauch, U.; Campbell, D. S.; Deutsch, E. W.; Chu, C. S.; Spicer, D. A.; Brusniak, M.-Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; Shteynberg, D.; Hoopmann, M. R.; Blattmann, P.; Ratushny, A. V.; Rinner, O.; Picotti, P.; Carapito, C.; Huang, C.-Y.; Kapousouz, M.; Lam, H.; Tran, T.; Demir, E.; Aitchison, J. D.; Sander, C.; Hood, L.; Aebersold, R.; Moritz, R. L. *Cell* **2016**, *166* (3), 766−778.
(14) Bittremieux, W.; Meysman, P.; Noble, W. S.; Laukens, K. *J. Proteome Res.* **2018**, *17* (10), 3463−3474.
(15) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007**, *7* (5), 655−667.
(16) Frewen, B.; MacCoss, M. J. *Curr. Protoc. Bioinform.* **2007**, DOI: 10.1002/0471250953.bi1307s20.
(17) Deutsch, E. W.; Perez-Riverol, Y.; Chalkley, R. J.; Wilhelm, M.; Tate, S.; Sachsenberg, T.; Walzer, M.; Käll, L.; Delanghe, B.; Böcker, S.; Schymanski, E. L.; Wilmes, P.; Dorfer, V.; Kuster, B.; Volders, P.-J.; Jehmlich, N.; Vissers, J. P. C.; Wolan, D. W.; Wang, A. Y.; Mendoza, L.; Shofstahl, J.; Dowsey, A. W.; Griss, J.; Salek, R. M.; Neumann, S.; Binz, P.-A.; Lam, H.; Vizcaíno, J. A.; Bandeira, N.; Röst, H. *J. Proteome Res.* **2018**, *17* (12), 4051−4060.
(18) Orchard, S.; Hermjakob, H.; Apweiler, R. *Proteomics* **2003**, *3* (7), 1374−1376.
(19) Deutsch, E. W.; Orchard, S.; Binz, P.-A.; Bittremieux, W.; Eisenacher, M.; Hermjakob, H.; Kawano, S.; Lam, H.; Mayer, G.; Menschaert, G.; Perez-Riverol, Y.; Salek, R. M.; Tabb, D. L.; Tenzer, S.; Vizcaíno, J. A.; Walzer, M.; Jones, A. R. *J. Proteome Res.* **2017**, *16* (12), 4288−4298.
(20) Deutsch, E. W.; Vizcaíno, J. A.; Jones, A. R.; Binz, P.-A.; Lam, H.; Klein, J.; Bittremieux, W.; Perez-Riverol, Y.; Tabb, D. L.; Walzer, M.; Ricard-Blum, S.; Hermjakob, H.; Neumann, S.; Mak, T. D.; Kawano, S.; Mendoza, L.; Van Den Bossche, T.; Gabriels, R.; Bandeira, N.; Carver, J.; Pullman, B.; Sun, Z.; Hoffmann, N.; Shofstahl, J.; Zhu, Y.; Licata, L.; Quaglia, F.; Tosatto, S. C. E.; Orchard, S. E. *J. Proteome Res.* **2023**, *22* (2), 287−301.
(21) Vizcaíno, J. A.; Martens, L.; Hermjakob, H.; Julian, R. K.; Paton, N. W. *Proteomics* **2007**, *7* (14), 2355−2357.
(22) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.
(23) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P.-A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D. *Mol. Cell. Proteomics* **2012**, *11* (7), M111.014381.
(24) Vizcaíno, J. A.; Mayer, G.; Perkins, S.; Barsnes, H.; Vaudel, M.; Perez-Riverol, Y.; Ternent, T.; Uszkoreit, J.; Eisenacher, M.; Fischer, L.; Rappsilber, J.; Netz, E.; Walzer, M.; Kohlbacher, O.; Leitner, A.; Chalkley, R. J.; Ghali, F.; Martínez-Bartolomé, S.; Deutsch, E. W.; Jones, A. R. *Mol. Cell. Proteomics MCP* **2017**, *16* (7), 1275−1285.
(25) Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q.-W.; Del Toro, N.; Pérez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaíno, J. A.; Hermjakob, H. *Mol. Cell. Proteomics MCP* **2014**, *13* (10), 2765−2775.
(26) Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; Orchard, S.; Vizcaíno, J. A.; Hermjakob, H.; Stephan, C.; Meyer, H. E.; Eisenacher, M. *Database* **2013**, *2013*, bat009.
(27) Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; Parkinson, H. *Bioinforma. Oxf. Engl.* **2010**, *26* (8), 1112−1118.
(28) Hoyt, C. T.; Balk, M.; Callahan, T. J.; Domingo-Fernández, D.; Haendel, M. A.; Hegde, H. B.; Himmelstein, D. S.; Karis, K.; Kunze, J.; Lubiana, T.; Matentzoglu, N.; McMurry, J.; Moxon, S.; Mungall, C. J.; Rutz, A.; Unni, D. R.; Willighagen, E.; Winston, D.; Gyori, B. M. *Sci. Data* **2022**, *9* (1), 714.
(29) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2008**, *7* (1), 113−122.
(30) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaíno, J. A. *Nat. Methods* **2016**, *13* (8), 651−656.
(31) The, M.; Käll, L. *J. Proteome Res.* **2016**, *15* (3), 713−720.
(32) Bittremieux, W.; Laukens, K.; Noble, W. S.; Dorrestein, P. C. *Rapid Commun. Mass Spectrom. RCM* **2021**, No. e9153.
(33) To, P. K. P.; Wu, L.; Chan, C. M.; Hoque, A.; Lam, H. *J. Proteome Res.* **2021**, *20* (12), 5359−5367.
(34) LeDuc, R. D.; Deutsch, E. W.; Binz, P.-A.; Fellers, R. T.; Cesnik, A. J.; Klein, J. A.; Van Den Bossche, T.; Gabriels, R.; Yalavarthi, A.; Perez-Riverol, Y.; Carver, J.; Bittremieux, W.; Kawano, S.; Pullman, B.; Bandeira, N.; Kelleher, N. L.; Thomas, P. M.; Vizcaíno, J. A. *J. Proteome Res.* **2022**, *21* (4), 1189−1195.
(35) Deutsch, E. W.; Perez-Riverol, Y.; Carver, J.; Kawano, S.; Mendoza, L.; Van Den Bossche, T.; Gabriels, R.; Binz, P.-A.; Pullman, B.; Sun, Z.; Shofstahl, J.; Bittremieux, W.; Mak, T. D.; Klein, J.; Zhu, Y.; Lam, H.; Vizcaíno, J. A.; Bandeira, N. *Nat. Methods* **2021**, *18* (7), 768−770.
(36) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. *Nucleic Acids Res.* **2006**, *34*, D655−D658.
(37) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11* (11), 601.
(38) Biemann, K. *Methods Enzymol.* **1990**, *193*, 886−887.
(39) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2004**, *4* (6), 1534−1536.
(40) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka,

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX

K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom. JMS* **2010**, *45* (7), 703−714.
(41) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016**, *3*, 160018.
(42) Dai, C.; Füllgrabe, A.; Pfeuffer, J.; Solovyeva, E. M.; Deng, J.; Moreno, P.; Kamatchinathan, S.; Kundu, D. J.; George, N.; Fexova, S.; Grüning, B.; Föll, M. C.; Griss, J.; Vaudel, M.; Audain, E.; Locard-Paulet, M.; Turewicz, M.; Eisenacher, M.; Uszkoreit, J.; Van Den Bossche, T.; Schwämmle, V.; Webel, H.; Schulze, S.; Bouyssié, D.; Jayaram, S.; Duggineni, V. K.; Samaras, P.; Wilhelm, M.; Choi, M.; Wang, M.; Kohlbacher, O.; Brazma, A.; Papatheodorou, I.; Bandeira, N.; Deutsch, E. W.; Vizcaíno, J. A.; Bai, M.; Sachsenberg, T.; Levitsky, L. I.; Perez-Riverol, Y. *Nat. Commun.* **2021**, *12* (1), 5854.

https://doi.org/10.1021/acs.analchem.4c04091
Anal. Chem. XXXX, XXX, XXX−XXX