ABSTRACT: Mass spectrometry (MS) is one of the primary techniques used for large-scale analysis of small molecules in metabolomics studies. To date, there has been little data format standardization in this field, as different software packages export results in different formats represented in XML or plain text, making data sharing, database deposition, and reanalysis highly challenging. Working within the consortia of the Metabolomics Standards Initiative, Proteomics Standards Initiative, and the Metabolomics Society, we have created mzTab-M to act as a common output format from analytical approaches using MS on small molecules. The format has been developed over several years, with input from a wide range of stakeholders. mzTab-M is a simple tab-separated text format, but importantly, the structure is highly standardized through the design of a detailed specification document, tightly coupled to validation software, and a mandatory controlled vocabulary of terms to populate it. The format is able to represent final quantification values from analyses, as well as the evidence trail in terms of features measured directly from MS (e.g., LC-MS, GC-MS, DIMS, etc.) and different types of approaches used to identify molecules. mzTab-M allows for ambiguity in the identification of molecules to be communicated clearly to readers of the files (both people and software). There are several implementations of the format available, and we anticipate widespread adoption in the field.

INTRODUCTION

It is now commonplace for high-throughput quantitative technologies to be used for analysis of biological, biomedical,
and environmental samples. Technologies include those for measurements of gene expression using microarrays or RNA sequencing (transcriptomics), proteins by mass spectrometry (MS, proteomics), and MS or nuclear magnetic resonance (NMR) spectroscopy for measuring small molecules/metabolites (metabolomics) and lipids (lipidomics). These methods can provide the source data for systems biology/medicine investigations into the complex network of interactions that reflect both their functional and dysfunctional states, as well as reflect nutritional and environmental impacts. There is now an accepted principle in scientific research that data should be made openly and easily accessible to allow groups other than the initial data generators to verify the findings or search for new interpretations. Such guidelines are now commonly referred to as the “FAIR” principles, data being findable, accessible, interoperable, and reusable. Furthermore, data from omics experiments are typically expensive to generate and often have potential uses beyond their initial purpose, including in meta-analyses, in data integration, or for testing and assisting in the development of new software. In omics research, there is always some heterogeneity in the approaches taken in different laboratories, such as different instrument platforms or analysis software, which usually have their own file formats. To allow data sets to be open for reuse generally requires the formulation of nonproprietary data formats, or more ideally, agreed data standards to which different producers of data must adhere. Without agreed standards (or ubiquitous formats originating from one package), data reuse is highly challenging, since informatics groups would need to write file format converters for every possible source of data, as well as keep these converters updated whenever data-producing software makes a format change. This scenario makes development of analysis software or a specific usage of public databases very challenging.

In a typical MS-based metabolomics/lipidomics pipeline, samples are analyzed by liquid or gas chromatography, coupled to MS (LC-MS/GC-MS), or by direct infusion (DIMS). Measurement of molecular intensity is typically done via software that detects features formed from isotopic patterns (or single peaks) along the time axis. For LC-MS, ionization can be performed in either positive or negative mode to produce protonated or deprotonated ions. It is also common for ion adducts to be formed, including metal adducts (Na+, K+), which have the same time elution profile but different m/z values. Many software packages perform adduct grouping, such that quantification values are reported both for individual features, as well as for the summed abundance across different adduct forms assumed to have come from the same starting molecule. For quantification across different samples, software may perform retention time alignment to ensure that the same features are quantified in each sample. In GC-MS, analysis is performed on volatile molecules and, in some cases, a derivatization step is applied to increase the volatility of compounds of interest.

Molecular identification remains challenging in metabolomics. Typically, some combination of the following steps can assist with identification via searching a pre-existing library or database: accurate neutral mass, the relative abundance of isotopomers, the retention time, masses of fragmentation products (MS/MS and MS² spectra), collisional cross section for platforms with ion mobility, etc. (see the reviews in refs 2–4 for more details). In the case of MS/MS and MS² fragmentation, the spectra can be compared against an in-house spectral library or databases storing reference spectra for molecules including Metlin, The Human Metabolome Database, Global Natural Products Social Molecular Networking (GNPS), MassBank, and others (see ref 9 for a review) or analyzed by in silico identification software.

It is common in all approaches for many molecules to remain unannotated or for ambiguity to remain; i.e., software provides a list of possible molecules for each MS feature. Reporting standards and guidelines on these certainties have been developed in several communities. Following quantification (and identification), statistical analysis usually proceeds via univariate approaches, e.g. to find differentially expressed molecules between conditions, or multivariate/machine learning approaches to explore structure within the data and find molecules that can separate sample groups and thus act as potential biomarkers.

There exists a wide range of software, both free and commercial, for processing MS data for metabolomics/lipidomics. Most software produces output data in a unique file format, annotated to different levels of detail, often with the description of preprocessing procedures followed implicit rather than specified, making it highly challenging to compare or integrate the results of different pipelines. For public data sharing, there are several databases that host data sets in support of publications or community data sets, including the European Bioinformatics Institute (EMBL-EBI) MetaboLights database and the NIH Metabolomics Workbench.

In this work, we describe a data standard for MS-based metabolomics analytical pipelines, called mzTab-M, which captures the downstream results of analysis (i.e., excluding raw data), suitable for statistical analysis, result visualization, or submission to a public repository in support of a publication. The standard has been developed in a joint and open process between members of the Metabolomics Standards Initiative (MSI), the Metabolomics Society Data Standards Task Group, and the Proteomics Standards Initiative (PSI), which had originally developed the mzTab format on which it is based. There are several related and complementary efforts, which include efforts to define minimum reporting requirements for different aspects of metabolomics. There is also general agreement among standards groups (MSI, Metabolomics Society) to promote the use of the PSI’s mzML format for raw data storage. mzML is an XML-based standard for MS data, either for profile data as recorded directly from the instrument or for centroided data (peak picked in the m/z domain). The freely available ProteoWizard software embeds software libraries from several vendors of MS instruments, enabling the conversion of vendor raw files into mzML. For NMR metabolomics, the recently released nmrML standard follows a design principle similar to that of mzML, capturing NMR spectra and some metadata within an XML-based standard. For the description of study design, experimental metadata, and sample processing parameters, the ISA framework, while generally applicable to all types of experimental design, has been particularly taken up by the metabolomics field. The PSI previously developed the mzTab format (version 1.0) to act as a simple format for quantified and/or identified peptides and proteins in MS workflows. mzTab version 1.0 also has a section to allow small-molecule data to be captured. However, the data model was rather simple and did not cover some important use cases for metabolomics/lipidomics and, as a result, it has not been extensively used for small molecules or lipids. The development of mzTab-M has thus branched off from the original mzTab format development, and we report it here as a new standard for metabolomics called mzTab-M (“version 2.0” to differentiate it
from mzTab version 1.0). It follows design principles similar to those of mzTab 1.0, but it is not backward compatible.

**METHODS**

The mzTab-M format was designed in a process that was open to any interested parties. All associated materials and code for processing and validating files are fully open source and are hosted on GitHub (https://github.com/HUPO-PSI/mzTab).

mzTab-M started from the design of mzTab version 1.0 format but was further developed to support the specific needs of metabolomics (see the Supporting Information for more details on the relationship). The development took place via face-

---

**Figure 1.** Overall structure of an mzTab-M file. (A) Metadata about the experiment, describing experimental design (study variables and assays), links to other files, etc.. (B) The small molecule (SML) table, capturing "final" results table: i.e., overall calculated quantification value (and identity where known) of a metabolite. (C) Quantification value in each (aligned) MS run for MS' features: e.g., mapped to individual adducts or charge states of a molecule. (D) Evidence supporting identification (with ambiguity if needed) for molecules, using CV terms for scores/statistics where available.
The specifications have been submitted to a formal document process for anonymous review, overseen by an editor commissioned jointly between the PSI and the Metabolomics Society. The mzTab-M format is defined by the specification document and example files that demonstrate how to encode certain features (see the GitHub repository). The specification document describes the overall structure of the format: what tables must be present, what columns and rows must be present in those tables, and what terminology is allowed as data values. For various aspects of metadata, the standard enforces (and can be checked by validation software) that controlled vocabulary (CV) terms are used (e.g., for names of software, databases, parameters, statistics, etc.), which can be sourced from the PSI-MS CV26 (https://www.ebi.ac.uk/ols/ontologies/ms), as well as other CVs where appropriate.

**RESULTS AND DISCUSSION**

mzTab-M has been designed to act as a single data format for metabolomics and lipidomics, including an appropriate level of detail about the “final” results: i.e., molecules quantified across samples. The format also contains the ability to represent “intermediate” or supporting data, including the evidence trail for identifications from software (scores or statistics), as well as the quantification values derived directly from MS: i.e., prior to any adduct grouping or summarization steps. The format is represented as tab-separated text, meaning that it can be loaded directly into a spreadsheet editor or into statistical software such as R or SPSS for downstream analysis and visualization, without any need for coding, and can thus replace the use of tables (e.g., in pdf or Excel format) of supplementary data in support of publications. It is also relatively straightforward for informatics groups to develop software to add support for the standard to existing software.

The mzTab-M format consists of four cross-referenced data tables (Figure 1): metadata (MTD), small molecule (SML), small molecule feature (SMF) and the small molecule evidence (SME). The MTD and SML tables are mandatory, and for a file to contain any evidence about how molecules were quantified or identified by software, all four tables must be present. The tables must follow the order MTD, SML, SMF, and SME, with a blank line separating each table. The structure of each table, in terms of the rows and columns that must be present, is tightly specified, as explained in the following sections and formally in the mzTab-M specification document.

**Metadata (MTD) Table.** The metadata table has multiple rows and exactly three columns (Figure 1A). Each row must contain (1) “MTD”, (2) a parameter name, and (3) the parameter value. The types of parameters that must or may be present are described in the specification document, and allowed values from CVs are defined in a mapping file. The MTD and SML tables are mandatory, and for a file to contain any evidence about how molecules were quantified or identified by software, all four tables must be present. The tables must follow the order MTD, SML, SMF, and SME, with a blank line separating each table. The structure of each table, in terms of the rows and columns that must be present, is tightly specified, as explained in the following sections and formally in the mzTab-M specification document.

**RESULTS AND DISCUSSION**

Figure 2. Simple experimental designs in mzTab-M can be represented using a combination of the elements study_variable (SV), assay, ms_run, and sample. Quantitative values can be reported in files for SVs and assays. (A) SV is intended to capture different groups of replicates, which might have resulted from different levels of a given variable: e.g., control versus treated (represented as 2 SVs) and n time points over a treatment course (as n SVs). (B) assay captures a measurement made about a molecule (small molecule/lipid) where multiple assays within the same SV are taken to be replicates of some kind (biological or technical). (C) ms_run captures a single run on an MS instrument. (D) Samples are optional in mzTab, since the quantitative software may often be unaware of the biological samples that have been analyzed. If that information is available, references from assay to the same (technical, upper half) or different (biological, bottom half) samples are used to describe the type of replication performed.
runs (i.e., experimental factors or groups) over which statistical analysis may be done. These values then inform the number of columns present in SML and SMF tables for which (relative or absolute) quantitative values are reported. The following concepts are specified in the MTD table so that they can be referenced and reused elsewhere in the file:

- Assay: the application of a measurement about the sample (in this case through MS), producing values about small molecules or lipids. One assay is typically mapped to one `ms_run` element (see below), although the differentiation between assay and `ms_run` is present to provide a mechanism for grouping multiple MS runs together if the sample has been fractionated and different fractions run on the instrument to increase coverage. The MTD table gives the count of assays with locally unique identifiers, so that they can be referenced by other elements.

- `ms_run`: an MS run is effectively one run on an MS instrument (e.g., by LC-MS, GC-MS, DMS, etc.) and can be referenced from assay elements in different contexts. When an ISA-Tab document from mzTab-M is referenced, `ms_run` should be matched with the ISA “Assay Name” values found in an ISA “Assay Table” file (https://isa-specs.readthedocs.io/en/latest/isatab.html).

- Sample: a biological material that has been analyzed, to which descriptors of species, cell/tissue type, etc. can be attached. Samples are not mandatory, since some software packages that will produce mzTab-M files cannot determine what type of sample was analyzed (e.g., whether biological or technical replication was performed), although it is noted that, without such annotations, downstream statistical analysis of the results will often not be possible.

- `study_variable`: a “study_variable” (SV) element represents a grouping of replicates for which a quantitative value can be reported, for example following averaging of values from individual assays. More accurately, a “study_variable” element usually represents a level of some particular experimental variable, such as the value of time within a time course, dose of a drug, intervention performed on samples, etc. In other contexts, this concept is named differently: e.g., “Factor Value” in ISA format.

Clear definitions of biological and technical replicates are difficult to provide, as the commonly used terminologies are somewhat dependent upon the biological domain. However, we use the following general definitions in mzTab-M: biological replicates represent cases when different samples are analyzed by MS, and technical replicates represent cases where the same samples are analyzed multiple times by MS. As illustrated in Figure 2, a simple form of the experimental design can be captured in mzTab-M using a combination of assay, “study_variable”, and “sample”. In a complex, nested design, linkages between different study variables are not explicitly modeled but captured through the annotated values, as shown in the Supporting Information.

The MTD table also has the (optional) capability to capture additional metadata that can be useful to interpret the study, such as limited details about the sample processing steps performed, the MS instrument, software and parameters, contact details for the study producers, etc. However, it is acknowledged that other formats may capture such details, such as referenced mzML (including instrument information and parameters), other MS data file formats, or ISA-Tab files (containing experimental design and sample processing), which may be more appropriate locations for such information.

**Small Molecule (SML) Table.** The small molecule (SML) table (Figure 1B) is intended to capture the “final” results of the study in terms of molecules that have been quantified (with identification data, where available). If different adduct forms or fragments of a molecule have been observed as different MS features, it is common that feature grouping is performed, and the SML table should contain the final quantitative values after summarization. Thus, SML could be viewed as the equivalent of tabular results presented in a paper for the molecules quantified in different samples. For survey-type data, it is also possible to report quantities as “null”, while still reporting identification evidence, as supported by the SME table.

The header row has “SMH” in the first column, followed by an ordered set of column headers. After the header row, each row reports one molecule with the first cell containing “SML”, followed by the data values for each specified column. The columns include a unique local identifier for the molecule (SML_ID), followed by a cell (SME_ID_REFS) containing references to features in the SML table. The referenced features are the different adduct forms or in-source fragments of the molecular features actually detected by MS. The next set of columns provides different ways to identify the molecule (database_identifier, chemical_formula, smiles, inchi, chemical_name, uri, theoretical_neutral_mass; see section Identification evidence and ambiguity below).

The following columns report quantitative data for the n assays (in n columns, where n is the count of assays reported in MTD) and the m study_variable groups (in m columns) e.g. as an average (e.g., mean) across assay values within that study_variable. A value can also be provided for the variability in the study_variable quantitation value reported e.g. a standard error value. A parameter in MTD specifies how to interpret the quantitative values in these columns in terms of a data type exported from a specific piece of software or where appropriate, absolute values with units.

At the right-hand end of the SML table (and SMF and SME tables), it is possible to include user-specified (optional) columns, with a method for annotating that the columns refer to the entire molecule, or the measurement of the molecule in particular assays or study_variables. The user-specified columns thus make mzTab-M extensible to support custom data types not covered in the core model.

**Small Molecule Feature (SMF) Table.** The SMF table contains data on what features were actually measured by the instrument and quantified by software (Figure 1C). The header row of the table has “SFH” in the first cell, followed by a set of columns. Each row of the table is one MS feature recorded across different runs, starting with the code “SMF”. It is assumed that an alignment process has taken place so that the same feature has been seen across different runs, with missing values handled as appropriate (see specification document for guidance on encoding nonaligned workflows). The next column (SME_ID_REFS) is for referencing down to the final table: Small Molecule Evidence (SME) via a set of identifier references, as well as a code telling the file reader how to interpret multiple references (SME_ID_REF_ambiguity_code), explained in Identification Evidence and Ambiguity.

The SMF table next contains information about the type of adduct and charge state observed, the experimental m/z value, the retention time of the feature (in a master or averaged run), and a method for optionally specifying if a given isotopomer has...
been quantified (+1 or +2 peak, 13C peak, etc.) as used in some isotopic labeling/flux studies. The following columns represent the quantitative data within each of the n assays recorded in the MTD section. For SML, a parameter in MTD also describes how to interpret the quantitative values recorded.

Small Molecule Evidence (SME) Table. The SME table represents strands of potentially heterogeneous types of evidence supporting the identification of a molecule (Figure 1D). Each row contains the result of one identification process (library search, pattern match, manual curation, etc.). The header of the table starts with “SEH” followed by a set of columns. The second column is a local identifier for a row of evidence (SME_ID), followed by a local identifier for the input data to the process (evidence_input_id). evidence_input_id is needed for the cases where different rows of evidence are reported for the same input data (MS spectrum, accurate mass + retention time, isotope pattern, etc.). They can be linked by sharing the same value for evidence_input_id. As in the SML table, a set of columns exists to specify the molecular identity from a variety of sources or identifier types (database_identifier, chemical_formula, smiles, inchi, chemical_name, uri). The experimental m/z value of the feature, the charge, and the theoretical m/z value (e.g., from a database) can be recorded, along with scores or confidence measures coming from the software used to support the identification. If a fragmentation spectrum has been used, there is a mechanism for referencing the exact spectrum in the source file (e.g., mzML file) and the MS level of the input data to the identification process.

Identification Evidence and Ambiguity. Small-molecule identification is a well-known challenge in MS metabolomics, and even more so in MS lipidomics, where complete structural elucidation of molecules is often not possible. Different levels of “identification” might be possible, ranging from having the accurate mass only, the chemical formula, a list of possible identifiers to molecules in a database (with the same or different formula), or a complete molecular structure resolved: e.g., if a complementary technique such as NMR has been used. mzTab-M has been designed to accommodate all the different possibilities in a simple yet flexible structure (Figure 3). For further details on how identifications of lipids and other compound classes can be represented see the Supporting Information.

In a row of the final results (SML table), the export software can include one or more identifiers from external databases: e.g., CHEBI:16811 where the prefix is defined in MTD as referencing the ChEBI database10 (with a URL) and the identifier is the ChEBI unique identifier (in this case for methionine). Similarly, the specification allows for the chemical formula in standard notation, simplified molecular-input-line-entry system (SMILES25), or InChi26 to be provided. In all cases, if ambiguity has not been resolved, then a Pipe “|” separated list of identifiers can be provided in the same cell. There are several measures for describing the confidence of identification, including the use of reliability codes such as those developed by the MSI11,12 and the score or confidence measures from identification software where available.

To trace the evidence source, references via the features (SMF table) and on to the SME table should be provided. In the case that adduct grouping (i.e., multiple SMF rows) has been performed prior to identification, then the different SMF rows will reference the same SME rows. At the SME level, if there are different rows from the same input data (e.g., different database matches), then it is expected that the SMF element(s) references multiple SME elements that share the same value for evidence_input_id. It is also possible to report different evidence streams to support identification, such as searches in different libraries. As such, SMF rows can reference multiple SME elements carrying different values of evidence_input_id. Given that these two cases would both result in multiple SME identifiers referenced from an SMF row, an extra code can be
provided at the SMF level (SME_ID_REF_ambiguity_code) containing values to differentiate whether ambiguity has been resolved or still remains (see the specification document for a full description).

**Using CVs and File Validation.** mzTab-M extensively uses CVs to provide unambiguous terms for annotation. For parameters relating to MS and associated processing, CV terms should generally be sourced from the PSI MS CV. Several other CVs are recommended for describing details about sample types, species taxonomy, sample preparation, etc. (see the specification document). To ensure that valid CV terms are used, we have extended the concept of the PSI semantic validation framework. The framework includes a mapping file that states the groups of CV terms allowed at each position in mzTab-M, enabling the list of terms to expand over time, without changes in the standard or software. New terms can be added straightforwardly by making a request on a mailing list: e.g., for a term describing new software, scores, or statistics. A crucial part of the standard is therefore a validator to ensure that files exported from different packages fulfill the rules defined in the specification, so that they can be read without error by other software. We have developed validation software for mzTab-M, available from jmzTab (project: https://github.com/lifs-tools/jmzTab-m), which checks not only that the structure of the file is correct but also that valid and correct CV terms have been used throughout.

**Implementation in Software and Databases.** The specifications have been verified by both PSI and MSI formal review processes, from which the stable version (mzTab-M 2.0) has been released. It is not expected that there will be changes to the format for several years to allow implementations to be developed. A reference implementation with parser, writer, and validator (in jmzTab-m) has been developed in Java (as for mzTab 1.0). jmzTab-m provides an OpenAPI 2.0 compatible API model that serves as the basis for automatic model generation in a wide number of programming languages (C++, JavaScript, R, Python), reducing the burden of implementation. The library provides parsing, validation, and writing of mzTab-M files and object models. A web-based application (https://apps.lifs.isas.de/mztabvalidator/) provides a user-friendly user interface to perform standard and semantic validation and to display validation results. Additional implementations are under development in software including XCMS, Progenesis QI (Waters), Lipid Data Analyzer, OpenMS, and Metabo-Lights. Over the coming years, we will be promoting the implementation of the standard in a wide variety of both open-source and commercial software to act as a universal standard for metabolomics and lipidomics.

**CONCLUSIONS**

We have developed mzTab-M for metabolomics data representation and sharing. The standard has been developed in an open process with widespread consultation of different approaches taken in the field and involvement of software teams from academic research groups as well as industry. The standard has undergone a rigorous peer review process by both the MSI and PSI to ensure that the resulting standard is of high quality and is stable. The standard is expected to remain stable for several years, except for improvements to documentation and extensions to the CV, allowing research groups and commercial developers to invest time in the implementation. We also encourage other groups interested in standardizing omics data, particularly those using MS (e.g. glycomics), to adopt the mzTab model/design, CV infrastructure, and associated software.

**REFERENCES**

(1) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. 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.; Dunmon, O.; Edmonds, S.; Eulo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; Hutton, R. A. C.; Hooft, R.; Kuhn, T.; Kong, J.; Koller, A.; Lush, 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.; Straw, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Veltrop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Anal. Chem.* 2018, 90, 4249–4257.

(2) Holcapek, M.; Liebisch, G.; Ekkro, K. *Anal. Chem.* 2018, 90, 3308–3310.

(3) Blazencic, I.; Kind, T.; Ji, J.; Fiehn, O. *Metabolites* 2018, 8, 31.
Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. Nat. Methods 2016, 13, 741−748.