The Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work

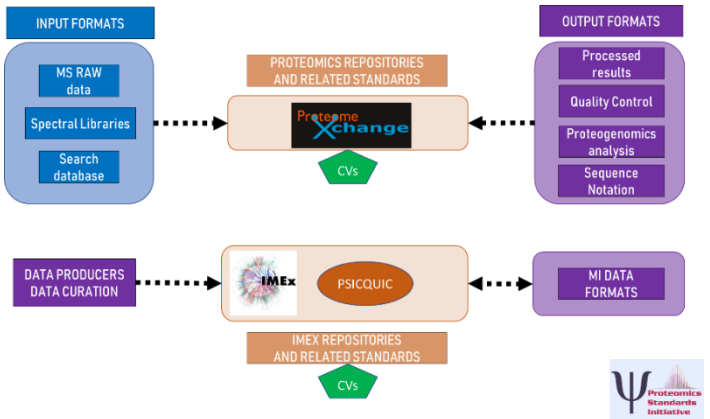
Eric W. Deutsch[1,*], Juan Antonio Vizcaíno[2], Andrew R. Jones[3,*], Pierre-Alain Binz[4], Henry Lam[5,6], Joshua Klein[7], Wout Bittremieux[8], Yasset Perez-Riverol[2], David L. Tabb[9], Mathias Walzer[2], Sylvie Ricard-Blum[10], Henning Hermjakob[2,11], Steffen Neumann[12,13], Tytus D. Mak[14], Shin Kawano[15,16], Luis Mendoza[1], Tim Van Den Bossche[17,18], Ralf Gabriels[17,18], Nuno Bandeira[8,19], Jeremy Carver[19], Benjamin Pullman[19], Zhi Sun[1], Nils Hoffmann[20], Jim Shofstahl[21], Yunping Zhu[22], Luana Licata[23,24], Federica Quaglia[25,26], Silvio C.E. Tosatto[26] and Sandra E. Orchard[2]

[1] Institute for Systems Biology, Seattle, Washington 98109, United States
[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[3] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 3BX, United Kingdom.
[4] Lausanne University Hospital, Lausanne, Switzerland.
[5] Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, P. R. China.
[6] Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, P. R. China.
[7] Program for Bioinformatics, Boston University, Boston, MA 02215, USA
[8] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093, USA
[9] SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[10] Univ. Lyon, Université Lyon 1, ICBMS, UMR 5246, 69622 Villeur-banne, France
[11] State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine; National Center for Protein Sciences;Beijing, Beijing 102206, China.
[12] Bioinformatics and Scientific Data, Leibniz Institute of Plant Biochemistry, Halle, Germany
[13] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[14] Mass Spectrometry Data Center, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[15] Database Center for Life Science, Joint Support Center for Data Science Research, Research Organization of Information and Systems, Chiba, Japan.
[16] Faculty of Contemporary Society, Toyama University of International Studies, Toyama, Japan
[17] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
[18] Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[19] Center for Computational Mass Spectrometry, Department of Computer Science and Engineering, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 92093-0404, USA
[20] Institute for Bio- and Geosciences (IBG-5), Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
[21] Thermo Fisher Scientific, 355 River Oaks Parkway San Jose, CA 95134, USA
[22] National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, #38, Life Science Park, Changping District, Beijing 102206, China
[23] Fondazione Human Technopole, Milan, Italy.
[24] Department of Biology, University of Rome Tor Vergata, Rome, Italy
[25] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy.
[26] Department of Biomedical Sciences, University of Padova, Padova, Italy.

Abstract

The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) has been successfully developing guidelines, data formats, and controlled vocabularies (CVs) for the proteomics community and other fields supported by mass spectrometry since its inception twenty years ago. Here we describe the general operation of the PSI, including its leadership, working groups, yearly workshops, and the document process by which proposals are thoroughly and publicly reviewed in order to be ratified as PSI standards. We briefly describe the current state of the many existing PSI standards, some of which remain the same as when originally developed, some of which have undergone subsequent revisions, and some of which have become obsolete. Then the set of proposals currently being developed are described, with an open call to the community for participation in the forging of the next generation of standards. Finally, we describe some synergies and collaborations with other organizations and look to the future in how the PSI will continue to promote the open sharing of data and thus accelerate the progress of the field of proteomics.

Keywords: Human Proteome Organization, mass spectrometry, proteomics, Proteomics Standards Initiative, molecular interactions, standards.



Introduction

The field of proteomics has seen tremendous advances over the past 20 years, with the emergence of faster and more sensitive instruments, new acquisition workflows that collect more data on more ions per run, and more advanced software capable of better analysis on far greater volumes of data. The identification and quantification of half of the entire proteome in a single run is becoming feasible, facilitating the collection of information about protein abundances, molecular interactions, and protein functions at tremendous scale.[1,2]

Driving these advances is a diverse ecosystem of data analysis software packages, both from academic laboratories as well as from commercial companies. The benefit of such diversity is enhanced if there is also a robust set of standardized data formats that enables the interoperability of the software and also between software and bioinformatics data resources, such as the members of the ProteomeXchange[3–5] and IMEx[6] consortia, for MS proteomics and molecular interaction data, respectively. Although some data types are relatively simple such that *ad hoc* tab-delimited formats are sufficient, most proteomics data types are sufficiently complex that information-rich structured formats are necessary to avoid massive loss of metadata and provenance information. An important effort in the biomedical research field overall is to promote making all data findable, accessible, interoperable, and reusable (FAIR),[7] and officially approved and recognized standards is a major component in the effort to make data FAIR[8].

The Human Proteome Organization[9] (HUPO) Proteomics Standards Initiative[10,11] (PSI) was formed 20 years ago in 2002, under the leadership of Rolf Apweiler, Ruedi Aebersold, and others committed to the formation of an organization that would develop standards for the field of proteomics. At the time there were only vendor formats for raw mass spectrometry data and a few over-simplified plain text formats specific to the software tools of the time. Yet it was recognized that standardized formats would allow not just tool interoperability, but also promote the sharing and reuse of data between laboratories.

The mission of the PSI was, and still is, to bring together tool developers from academia, software vendors, and hardware vendors to create, maintain, and promote data standards that will be used throughout the proteomics and computational mass spectrometry community. These products include standardized data formats, minimum information guidelines, and controlled vocabularies used to drive the formats. Standard formats are only effective if they are widely implemented in software tools, and thus the PSI also includes extensive outreach efforts and works with software developers to implement its standards. As a result, adoption and wide-spread implementation of PSI standards has enabled the development of software, such as Cytoscape, and APIs, such as PSICQUIC[12] and ProXI, which enable easy access to rich data streams by a broad spectrum of the research community.

In this article we first provide an overview of the operation of the PSI, highlighting its most recent workshops. We then describe the controlled vocabularies, guidelines, and data formats that have been developed and ratified by the PSI over the years. Next, we describe the set of standards that are currently in various phases of development, with an open call for participation by anyone willing to contribute to our efforts. Finally, we provide a brief discussion of synergies with other related organizations in the life sciences community and conclude with a vision of future contributions to the field.

Operation of the HUPO-PSI

The PSI is organized as a set of working groups (WGs) and an overall steering group. Each working group consists of a chair, one or two co-chairs, and several named positions such as secretary, editor, guidelines coordinator, CV coordinator, and web content maintainer. In addition to this leadership group, each working group consists of other members contributing to the group, with substantial overlap in membership between the groups. The currently active working groups are the Mass Spectrometry, Molecular Interactions, Proteomics Informatics, Protein Modifications, Quality Control, Intrinsically Disordered Proteins, and Metabolomics Coordination Working Groups. The PSI Steering Group consists of an overall chair, two co-chairs, a secretary, one or more editors, a guidelines coordinator, ontology coordinator, and web site maintainer, plus all of the chairs and co-chairs of the active working groups. The Steering Group generally meets monthly to coordinate working group activities, plan workshops, and develop outreach efforts. Current organizational structure is documented at https://psidev.info/roles.

The general membership of the PSI is open to all who wish to participate. Everyone is encouraged to post to the corresponding mailing list or contact any one of the leadership members of a working group to indicate interest in a specific project, and be included in the ongoing activities. Organizations that have been in operation for so long can seem closed and cliquey, but the PSI actively seeks to dissuade this notion. Individual standards or projects are not necessarily led by working group chairs, but ideally led by those who are most interested in completing a project, regardless of their status. In general, the PSI only develops standards where at least one working group participant is an active champion of the standard and drives its forward progress. Proposed or desired standards without a champion usually are not developed. However, anyone in the community with the desire to see a new or updated PSI standard is very much encouraged to become a member of the PSI and champion that project within the umbrella of the PSI.

The PSI ratifies standards via a mechanism called the Document Process[13] (DocProc). The DocProc is a formal process by which a proposed specification is thoroughly reviewed and refined before becoming an officially ratified standard of the PSI. Once a draft specification has been prepared by a working group, the process begins with the submission of a proposed specification to one of the PSI editors who has not been involved in the development of the specification. If the specification is deemed ready by the editor, it is sent to the Steering Group for a 14-day internal review period to assess initial suitability. Steering Group comments are then addressed by the proposers and the revision is resubmitted. Next, the editor selects at least two external reviewers familiar with the subject matter but not part of the development of the standard. The peer reviewers are generally anonymous, although there is precedent for reviewers requested to be listed in acknowledgements in recognition of their often-substantial time spent reviewing a specification. The specification is then revised based on the comments of the reviewers, after which the revision is subjected to a 4-week open community commenting period during which the proposal is widely advertised as a nearly complete standard and any additional comments from the community at large are sought. Ultimately when all comments received have been addressed to the satisfaction of the handling editor, the specification is declared ratified as an official PSI standard. The DocProc may be revised with the approval of the Steering Group, and has undergone several revisions in the past 20 years. The current version 1.1.2 is available at the PSI web site at https://www.psidev.info/psi-doc-process. It is common that a journal article describing the standard in brief is prepared, submitted, and reviewed independently (ideally in parallel to the PSI review process) by a journal and generally, the standard is not declared ratified until both the DocProc review and journal review are complete.

Efforts to develop specifications continue all year at weekly calls and *ad hoc* meetings. Additionally, the PSI hosts a yearly workshop to bring everyone together to discuss the ongoing work in more depth. The workshops have traditionally been held in the second quarter of each year in different locations throughout the world in an effort to attract new members in different regions. These workshops also have the effect of spurring extra progress in the weeks prior to and after the workshop.

2022 PSI Spring Workshop

The 2022 PSI Spring Workshop was held at the European Bioinformatics Institute (EBI) on the Wellcome Trust Genome Campus in Hinxton, United Kingdom, a fitting location for the 20th anniversary as the inaugural PSI workshop was also hosted at the EBI. This workshop was also the first in-person workshop in three years, since the SARS-CoV-2 pandemic caused the 2020 and 2021 workshops to be held fully online. Previous workshops were held in Cape Town, South Africa in 2019, at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany in 2018, and at the National Center for Protein Sciences, Beijing (Phoenix Center) in Beijing, China in 2017. The 2020 workshop was originally

organized for the University of California San Diego, California, USA, but was forced fully online due to the pandemic.

The 2022 workshop was a hybrid event with 31 in-person participants at the EBI, and 44 members participating via Zoom. The first day began with general overviews of progress and workshop plans by each working group plus an update by Juan Antonio Vizcaíno on the current state of the ProteomeXchange Consortium. After a short break, the Molecular Interactions Working Group split off into a separate track while the Mass Spectrometry, Proteome Informatics, and Quality Control Working Groups continued in a joint session to briefly discuss all of the topics relevant to all three working groups since there is substantial overlap in interests between the members of these groups. The second day was devoted to substantial progress on mzSpecLib and mzQC (both discussed further below) in parallel tracks, whilst the MI group focused on format implementation and data curation. The third day began with more parallel track development and ended with a final plenary session with summaries of progress in each of the parallel tracks. The individual track sessions typically involve minimal presentation and mostly discussion of unresolved items that need group input, general planning, and development and assignment of action items.

Controlled vocabularies and ontologies

An ontology is a collection of concepts with names and definitions and clear relationships between all the terms, typically with "is a" and "part of" relationships (e.g., a wheel is a part of a car, and a car is a vehicle, and a vehicle is a thing, although typically much more complex). A crucial component of the ontology is the relationships. A similar collection of concepts where the focus is on the concept names and definitions and not the relationships is typically thought of as controlled vocabulary (CV). Thus far the PSI collections, widely used as proteomics-related ontologies,[14] are part way between the two extremes and are described further below.

The PSI-MI (Molecular Interactions) CV was first published in 2004[15] and, since then, has been regularly extended to encompass new experimental methodologies and to serve extensions made to the formats to accommodate new data types. The CV is maintained on GitHub (https://github.com/HUPO-PSI/psi-mi-CV/blob/master/psi-mi.obo) and is readily available through the Ontology Lookup Service (www.ebi.ac.uk/ols/ontologies/mi) and BioPortal (https://bioportal.bioontology.org/ontologies/MI). An issue tracker on GitHub enables new term or update requests to be submitted by the user community. Use of the PSI-MI CV is an integral step in describing interaction data using the PSI-MI formats.

The PSI-MS (Mass Spectrometry) CV was initially developed to serve the needs of the PSI mass spectrometry-related data formats, most of which make extensive use of CV terms to precisely describe metadata and provide extensibility as technologies advance[16]. The CV was first developed to support the now-deprecated mzData format and then for its replacement mzML,[17] and has since been adapted for numerous additional formats that will be further described below. It contains many terms for specific instrument models, software packages, and other metadata concepts and continues to grow as more such terms become available. This CV is readily accessible at https://github.com/HUPO-PSI/psi-ms-CV/ and is typically updated at least monthly. The contents of the CV is also browsable via the Ontologies Lookup Service[18,19] (https://www.ebi.ac.uk/ols/ontologies/ms) and the NCBO BioPortal[20] (https://bioportal.bioontology.org/ontologies/MS).

PSI-MOD is an ontology of modified amino acid residues.[21] The terms focus on the end products of modifications rather than modifications themselves. Thus, instead of terms for phosphorylation and oxidation, there are terms such as O-phospho-L-serine and L-methionine sulfoxide. The concepts are organized in several hierarchies by amino acid and by modification type. Although the ontology suffered from a period of inactivity several years ago, volunteers have now stepped forward to maintain it actively. The ontology is maintained on GitHub (https://github.com/HUPO-PSI/psi-mod-CV) or is readily available through the Ontology Lookup Service (https://www.ebi.ac.uk/ols/ontologies/mod). Complementary CVs, maintained by other organizations, which contain information on protein modifications include RESID, PTMList, Unimod, and XLMOD. PSI-MOD was originally based on RESID, which is now mostly deprecated and used by very few tools. UniProtKB uses its own PTMList resource (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/ptmlist), although PTMList is substantially synchronized with PSI-MOD. The popular Unimod community resource of mass modifications (http://unimod.org/) instead focuses on the modifications themselves (with terms like "Phospho" and "Oxidation") and is a flat CV with no relationship structure.

The XLMOD CV[22] (https://github.com/HUPO-PSI/xlmod-CV) is a simple collection of known chemical cross-linkers for use with formats that need to describe which cross-linker was used in the sample handling.

All these ontologies and CVs are maintained in the OBO format. However, each commit is versioned with a new minor release number and an OWL format file is autogenerated using the ROBOT conversion tool (http://robot.obolibrary.org/convert.html) via GitHub actions. Assistance with maintenance of the CVs and ontologies is a good entry point for those with proteomics domain knowledge who wish to begin contributing to the PSI.

Guidelines

In addition to data formats and CVs, the PSI has also developed several sets of guidelines. These guidelines typically describe at minimum which pieces of information should be provided when releasing or sharing data, without specifying the encoding for that information. This minimum information may be provided in supplementary material or in an article or in a PSI standard format. PSI standard formats are often designed with the aid of previously-developed guidelines, ensuring that they are capable of capturing at least all of the minimum information and usually much more.

The first standard published was the Minimum Information about a Molecular Interaction eXperiment (MIMIx), which advises the user on fully describing a molecular interaction experiment, in either a publication or a database, and lists which information is important to capture.[23] The document is designed as a compromise between the necessary depth of information to describe all relevant aspects of the interaction experiment, and the reporting burden placed on the scientist generating the data. The MIMIx standard has remained pertinent for over 15 years and the original Nature Biotechnology publication remains the relevant documentation.

The Minimum Information About a Proteomics Experiment (MIAPE) guidelines[24] were developed following the style of the Minimum Information About a Microarray Experiment (MIAME) guidelines[25] for microarray experiments. The MIAPE guidelines were developed in a modular structure, such that each aspect of a proteomics experiment would correspond to a module, and the modules that pertain to a specific experiment could be used. Modules for the column chromatography[26] (MIAPE-CC), the mass spectrometry[27] (MIAPE-MS), the subsequent informatics analysis[28] (MIAPE-MSI), and finally the quantitative components[29] (MIAPE-Quant) of an experiment were developed. These MIAPE components were implemented in some systems such as the ProteoRed database[30] and used as a guide in other software and during the development of formats. However, it seems that while most researchers would like others to provide rich metadata about their datasets, most researchers balk when asked to provide all that information about their own datasets and avoid systems which require that they do so.

Although not formally ratified by the PSI, the PSI supported the development of the HUPO Human Proteome Project (HPP) Mass Spectrometry Data Interpretation Guidelines. Three such versions have been produced. The 1.0 version of these guidelines described the repository deposition requirement for data contributed to the HPP (https://hupo.org/HPP-Data-Interpretation-Guidelines). Version 2.1 of the

guidelines added additional requirements for false-discovery rate (FDR) thresholding and for evidence supporting claims of detection of human proteins.[31] Version 3.0 of the guidelines further refined these ideas to provide more stringent requirements for exclusion of false positives.[32] Version 3.0 was published in 2019 and no further refinements seem necessary as of 2022.

Existing standard data formats

The PSI has developed numerous standardized data formats in the past 20 years and more are currently under development. An overview of the formats developed by the Molecular Interactions Working Group and their relationships to other components is summarized in Figure 1. The formats of the Mass Spectrometry Working Group and the Proteome Informatics Working Group and their relationships to other components are summarized in Figure 2. In this section we briefly describe each format, assess its current status, and provide links and citations for obtaining more information. Active formats are presented by working group, followed by non-PSI related formats and deprecated formats.

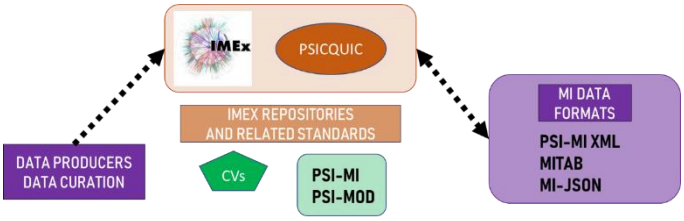


Figure 1. Overview of the formats developed by the Molecular Interactions Working Group and their relationships to other components in the community.

Molecular Interactions Working Group

PSI-MI XML

The MI group released an XML standard (PSI-MI XML) as early as 2004,[15] which allowed the basic information about a protein-protein interaction experiment to be captured and transferred between data resources or visualization/analysis tools. This initial, very simplistic representation was very soon upgraded to enable a full capture of the details of any experiment, including *in silico* and predictive data, which describes an interaction between any number of biomolecules, including experimental details, molecule details (affinity tags, binding sites, amino acid mutations, etc.), and the host organism in which the experiment is undertaken. Additional information, such as kinetic parameters or the method by which a molecule is delivered or engineered into a cell can also be added, if required.[33] The PSI-MI XML2.5 version of the format remains the main workhorse by which the majority of interaction data is exchanged between resources, and enables all of the above use cases. As more specialist use cases have arisen which could not be fully captured in this format, in 2018 a new, backwards compatible version, PSI-MI XML3.0, was developed.[34] This version can be used to describe, for example, the details of a fully curated multiprotein complex with subunit topology, including binding regions, and stoichiometry or to link kinetic parameters to amino acid mutations, sequence deletions or insertions.

MITAB

Following requests from bench scientists for a simpler version of the format for routine use, the MITAB format was developed to enable most aspects of a molecular interaction experiment to be described. The format shares usage of the PSI-MI CVs and many fields are identical to those in the XML formats, but the data is described in tab-delimited format with increasing number of columns in the different versions (MITAB2.5, 2.6, 2.7, 2.8) allowing increased data capture. The most recent version, MITAB2.8 also referred to as CausalTAB, enables the representation and dissemination of signaling information through the description of the causality of an interaction.

MI-JSON

MI-JSON is the recommended protocol for serving interaction data to web pages and visualization tools. The format is described at https://github.com/MICommunity/psi-jami/blob/master/jami-interactionviewer-json/schema/mi-json-schema.json

Java software library JAMI

JAMI[35] is a single java library and framework which unifies the standard formats such as PSI-MI XML, PSI-MITAB and MI-JSON and also formats not created by the HUPO-PSI. Adopting JAMI avoids conversions between different formats and avoids code/unit test duplication as the code becomes more modular. The JAMI model interfaces are abstracted from each format to hide the complexity/requirements of each and enables the development of software and tools on top of this framework (https://github.com/MICommunity/psi-jami).

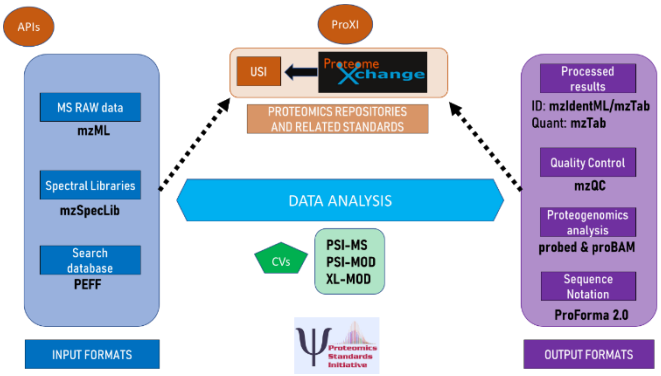


Figure 2. An overview of the formats of the Mass Spectrometry Working Group and the Proteome Informatics Working Group and their relationships to other components in the community.

Mass Spectrometry & Proteomics Informatics Working Groups

mzML

The primary standardized format for encoding the output of mass spectrometer instruments is mzML.[17] All instruments write out their primary data in a binary vendor format. Most vendors provide software libraries to read their data files, but these are often only available on the Microsoft Windows platform. In order to facilitate data analysis on any platform, and ensure that data files could always be read, the PSI developed the open XML-based format mzML. Version 1.0 was released in 2008 and a minor fix version 1.1 was released in 2009; version 1.1 has been stable and widely used ever since.

The focus of mzML has always been universal readability, and thus small file size and fast access speed have not been the primary design drivers. As a result, numerous alternatives to mzML that improve on file size and access speed have been proposed,[36–39] although none seem to have gained substantial usage, perhaps in part because of their dependencies or complexity. A variant of mzML called imzML[40] has become quite common in the imaging MS community, and it is nearly identical to mzML except that the spectra are stored in a more efficient sidecar file instead of the main file. In order to try to consolidate on one particular proposal, the PSI is considering formally approving the HDF5-based mzMLb format[41] as a PSI standard in addition to mzML. The mzMLb format has exactly the same schema and encodes metadata in the same way, and thus interconversion is easy. Adding mzMLb to existing tools is relatively straightforward if the HDF5 dependency can be managed.

Ongoing work: mzML extension for DIA and IMS data

As described above, the mzML format has been stable and widely used since 2009. At the time, data independent acquisition (DIA) and ion mobility spectrometry (IMS) were not widely used and not explicit factors in the design. Now that they have become widely used, there is great interest in extended mzML for use with these technologies. Fortunately the schema definition does not need to be changed, and good support for these technologies can be achieved with some additional controlled vocabulary terms and an explicit best practices document that describes how these types of data should be encoded in mzML. This best-practices document has been drafted and is nearly ready for submission to the DocProc. The current draft is available at the mzML web page (http://psidev.info/mzml/).

mzIdentML

The primary PSI standard format for encoding the peptide/protein identifications that are derived from an MS proteomics experiment is mzIdentML. The version 1.1, considered as the first stable version, was released in 2011[42] and version 1.2[43] in 2017, adding features for workflows such as *de novo* sequencing, crosslinking, proteogenomics approaches and improved encoding of protein inference results. The mzIdentML format is widely implemented in over 35 software tools (https://www.psidev.info/tools-implementing-mzidentml) and an encouraged (but not required) format at the ProteomeXchange data repositories, since it is one of the standard formats that can be used for performing "Complete" submissions (those where the identification data can be parsed and linked to the mass spectra by the receiving repository).

Ongoing work: mzIdentML extension for glycopeptide and cross-linked peptide data

The mzIdentML format still does not fully meet the needs for some special workflows, and efforts are underway to define best practices for improving the encoding of more complex arrangements of cross-linked peptide identifications than in the 1.2 release, and to represent glycopeptide identifications in a consistent manner. This will not require a schema change, but rather, the development of best practice documents and controlled vocabulary terms will mainly define how these data types should be encoded in a consistent manner. These enhancements are expected to be completed in 2023.

Although not supported in the first stable 1.1 version of mzIdentML, initial support for cross-linked peptide identification data was added in mzIdentML version 1.2. However, it has been determined that even mzIdentML 1.2 was underspecified for properly encoding some use cases in cross-linking data, e.g., when working with cleavable reagents. In addition, there is an effort underway as part of the PSI-PI working group to develop more extensive documentation and best-practice guidelines for encoding glycopeptides as well. Participation in this design process is actively sought.

mzTab

Although mzIdentML has been widely adopted, it was felt that for some applications, a simpler tab-delimited format capable of encoding the most important information would be beneficial. The mzTab format[44] was conceived as such a relatively simple format that could encode both identification and quantification, applicable to both bottom-up MS proteomics as well as small-molecule MS metabolomics. It has been implemented in several software packages (mainly for identification data, not for quantification) such as Mascot[45], OpenMS[46], and MaxQuant[47], and ProteomeXchange repositories .

After some years of use, it was concluded that while the attempt to support both proteomics and metabolomics in the same format was a noble idea, the result was a format that did not serve either data type as well as desired. As a result, mzTab-M[48] was redesigned as a format that would fix some perceived design issues and support only metabolomics. Its data model is backed by a JSON-based schema, supporting serialization into a tab-separated main storage format or JSON as a transfer format. Since then, it has been adopted by a number of metabolomics and lipidomics software packages and repositories (https://github.com/HUPO-PSI/mzTab#current-activities-and-software-support). Additionally, similarly to mzIdentML, mzTab can also be used for performing "Complete" submissions to ProteomeXchange data repositories.

Preliminary discussions have started to produce a similarly redesigned mzTab-P 2.0 for proteomics only, based on the experience of the design of mzTab-M, although to date the updated format has not progressed rapidly and hence participation in this design process is actively sought.

proBAM and proBed

The proBAM and proBed formats[49] are relatively straightforward adaptations for proteomics of the tab-delimited BAM/SAM and BED formats widely used in genomics. Existing columns were defined to be applicable to peptide-based MS proteomics, and a few additional columns were added for peptide-specific contexts. Peptide data thus written in proBAM and proBed are compatible with transcript alignment viewers and other similar software already widely used in transcriptomics. No further changes are planned. Unfortunately, the formats have not been widely adopted so far. Most people use their own variation of the original BED format for representing peptide coordinates in a genome context, by using a small number of columns. Additionally, there is not a perceived need yet to use the more verbose proBAM format. In our view, by using these formats, proteogenomics studies would benefit from a greater standardization in data representation. The added advantage is that most genomics viewers developed would be able to visualize these files.

PEFF

The PSI Extended FASTA Format[50] (PEFF) was developed as a format that is broadly compatible with the ubiquitous FASTA format, but defines mechanisms for file-level metadata, multiple sequence collections within a file, collection-level metadata, and how a wide variety of additional information can be encoded on a per-sequence basis. Whereas the description lines in FASTA files are free form and vary widely based on the data provider, the format of description lines for PEFF files are defined in the specification and can be extended with controlled vocabulary terms. This allows PEFF files to encode proteins and their modifications, such as post-translational modifications (PTMs), protein processing such as signal peptides, and sequence variants. Implementations of PEFF support in Comet[51] enables transparent searching for known PTMs and sequence variants.[52] Although it was not originally designed with this use case in mind, PEFF can also be used to represent proteoforms.

ProForma 2.0

The Consortium for Top-Down Proteomics (CTDP) defined the initial 1.0 ProForma notation[53] to encode exact proteoforms with all applicable PTMs on a specific sequence. When the PSI needed a mechanism to encode in a compact way exact peptidoforms (peptide sequences with a specific set of mass modifications), ProForma provided a useful piece of prior work on which to build. In collaboration with the CTDP, the PSI has recently developed the ProForma 2.0 standard,[54] which provides a substantially expanded set of mechanisms to encode a wide variety of modified proteins and peptides (proteoforms and peptidoforms) in a manner that meets the needs of both the CTDP as well as the bottom-up proteomics workflows. The ProForma 2.0 formatting is both easily human readable as well as software parsable. The ProForma 2.0 standard is used in conjunction with other completed and in-development PSI standards as described below.

Universal Spectrum Identifier (USI)

There are numerous cases when one or more mass spectra should be carefully examined since they provide crucial evidence for a scientific conclusion. Often such spectra are published as figures in journal articles or may appear in supplementary material as static figures that prevent close scrutiny. Additionally, in the context of FAIR data, it is recommended to have unique identifiers for different types of entities (in this case mass spectra) in public data repositories. While not a file format in itself, the PSI has defined the USI[55] as a multipart key that can be copy-pasted into manuscripts or other conversations about important spectra in order to facilitate the identification and retrieval of specific spectra and PSMs. USIs currently identify spectra that have been deposited into one of the

ProteomeXchange public data repositories, although extensions are underway to support USIs for spectra in spectral libraries, both for proteomics as well as metabolomics. See http://proteomecentral.proteomexchange.org/usi for proteomics examples, and https://metabolomics-usi.ucsd.edu/ for metabolomics examples.[56]

Related formats

Although not officially ratified formats of the PSI, there are a few formats that are highly related to the efforts of the PSI and ProteomeXchange, and will be briefly described here.

ProteomeXchange XML (PX XML)

The ProteomeXchange (PX) Consortium[3–5] has been receiving and making public proteomics datasets from the community since 2012, and PX members are highly active in the PSI. The datasets themselves always remain at the receiving repository for a submission, but the most important metadata about each experiment is transmitted from the receiving repository to ProteomeCentral whenever a dataset is made public. The common format for this is PX XML (current version is 1.4, http://proteomecentral.proteomexchange.org/schemas/proteomeXchange-1.4.0.html), which encodes study metadata such as title, description, submitter, publication, location of availability, submitted files, etc. The format is similar to mzML and mzIdentML in basic design and since its development in 2011 has evolved to include an increasing amount of metadata as ProteomeXchange requirements increase.

MAGE-TAB for Proteomics (SDRF-Proteomics and IDF)

Although the PX XML format provides substantial study-level metadata and a list of files submitted with the study, it does not provide for specific sample metadata and a mechanism to link individual files to specific samples of the study. The MAGE-TAB for proteomics format,[57] which is an extension of the original MAGE-TAB format used in transcriptomics,[58] has been recently adapted by ProteomeXchange resources to capture the sample metadata and the experimental design for proteomics experiments. MAGE-TAB for proteomics has two main components: the Investigation Description Format (IDF) and the Sample and Data Relationship Format (SDRF-Proteomics). The MAGE-TAB-Proteomics files are compatible with the original transcriptomics versions, but several adaptations/extensions are included for the proteomics use case.[55] IDF is quite analogous to the PX XML format, although it is less structured and is not as rich in information as PX XML. PX XML files can be easily converted into IDF with some loss of information.

SDRF-Proteomics files provide the previously missing – and much needed – mechanism to provide sample-specific annotations with CV terms, and encode the links from those samples to the submitted data files. Sample attributes are encoded in the columns of the files, where the column definitions are required to be CV terms and the data values are either CV terms or plain scalar values when CV terms are not appropriate. Several ProteomeXchange repositories now accept and process SDRF-Proteomics files upon submission of datasets, and the GitHub repository at https://github.com/bigbio/proteomics-metadata-standard provides a mechanism for anyone in the community to manually curate and generate SDRF-Proteomics files for ProteomeXchange datasets that were previously released. The MAGE-TAB for Proteomics effort was primarily driven by the EuBIC-MS group, in collaboration with the PSI. As mentioned above, submission of IDF files is not required since all the information required to create them is made available at submission time.

Disused formats

Not all formats developed by the PSI have been successful and widely used. The mzData format released in 2005 was deprecated in 2008 with the release of mzML, which incorporated all of mzData's functionality. The sepML and gelML formats[59] were well designed and still potentially usable, but there was no interest in the community to encode the many details of gel-based workflows and for other separations used in proteomics experiments and have been deprecated due to little interest. Should interest revive in capturing such data, they remain good foundational work. The TraML format[60] for encoding SRM transition lists and DDA inclusion lists was used by a few software packages[61] after its release, but the extreme popularity in the SRM field of the Skyline software,[62] which used its own .sky format and never supported TraML, rapidly led to the TraML format becoming largely unused.

mzQuantML was developed as an XML-based format to capture a detailed output of proteomics quantitative workflows.[63] However, its adoption was limited due to a preference from most software developers to export quantitative output data in flat text files. When mzTab was released later, there was the hope that this simpler tab-delimited format could become popular to represent the final results coming from quantitative experiments. As mentioned above, although mzTab has been implemented, it has been mainly used for capturing identification results only so far. Very regrettably, the field has been unable to agree so far on a format for capturing quantitative information. This situation really hinders further progress and the reuse and integration of proteomics data.

A clear lesson is that PSI formats are most successful when there are a variety of tools with developers who want the standardized format and participate in its development, ideally along the entire data life-cycle from data acquisition, preprocessing, identification, quantification to statistical analyses, visualization and eventually data deposition and publication.

New Standards in development

Most of the formats mentioned thus far have completed the standardization process, are complete, and in use. However, the PSI is actively working on several new formats that are in various stages of the standardization process. In all cases, community involvement is always being sought, either for the design phase, early adoption in software or for the review phase. Readers with interest in any of the subsequent formats are encouraged to contribute to their development and ratification.

mzQC

The qcML format[64] was published in 2014 as an XML PSI-like format for encoding mass spectrometry quality control (QC) metrics, but without the PSI ratification process. However, several shortcomings were soon identified and general interest in QC waned, both of which hindered adoption. Renewed interest for a QC format within the PSI has led to a simplified, yet more versatile JSON-based new format, called mzQC, that fixed these shortcomings and includes the participation of members of the community most interested in supporting mzQC in their QC tools.[65] As a result, many QC-relevant concept and metric terms have been proposed for integration into the PSI-MS CV. This JSON-based format is currently under review in the PSI DocProc.

mzSpecLib

There have been several spectral library formats in wide use in the community for some time, including the NIST MSP format, the SpectraST[66,67] speclib format, several versions of the blib format,[68] the hlf format, etc. While any one of these formats does a good job in encoding the spectra in the library, there is general agreement that none of the formats do a good job in encoding the metadata associated with the library in a consistent manner.[69] An example is MSP, wherein nearly all metadata is recorded within the COMMENT field in a highly variable manner across software packages. Various producers and consumers of spectral libraries have come together to create a new standard called mzSpecLib (https://psidev.info/mzSpecLib). It is quite similar to the MSP and speclib formats, with an emphasis on heavy use of CV terms and a standardized data model that can be encoded in several different serialization mechanisms, including an MSP-like text format, a JSON format, and potentially a more performant binary format based on HDF5 or SQLite. mzSpecLib is expected to enter the PSI DocProc in 2023.

Peak Annotation Format

Although originally conceived and developed as part of mzSpecLib, a standardized format for encoding fragment peak annotations has been split into a separately proposed standard (see details at https://psidev.info/mzSpecLib). This was done in order to keep the mzSpecLib specification smaller, because the annotations for other types of molecules, such a lipids and glycans, would be quite different but envisioned as an add-on to mzSpecLib, and finally because a standardized annotation format may be useful in other contexts, such as in spectrum viewers and figures containing annotated spectra. The formatting is based substantially on the NIST MSP peak annotation formatting, which has evolved over the years but was never documented. Several components differ from NIST MSP conventions by general consensus, with participation by NIST as well. This peak annotation format is about to enter the DocProc.

PROXI

The partners of the ProteomeXchange consortium have been developing an API for the communication of proteomics data called ProXI, formally the ProteomeXchange eXpression Interface. The API enables the programmatic access to information about datasets, proteins, peptides, peptidoforms, peptide-spectrum matches (PSMs), and spectra via a consistent interface for all ProteomeXchange partners as well as an aggregator at ProteomeCentral. ProXI remains a work in progress, with several of the endpoints designed and at least partially implemented at several partners. Dedicated funding is likely required to complete it. PROXI is the mechanism that drives the multi-repository USI lookup and display mechanism at ProteomeCentral described above. ProXI uses the OpenAPI 3.0 platform for designing endpoints with a JSON communication format. The dataset endpoint is modeled after the PX XML schema, and the spectrum endpoint is modeled using the mzSpecLib schema described above.

PTM site formats

As part of "PTMeXchange", a project funded by BBSRC/NSF to improve sharing and deposition of high-quality sets of post-translational modifications (PTMs), a set of formats to encode post-translational modification (PTM) results are being developed. These are intended to be as simple as possible while still encoding sufficient information to properly evaluate FDR values and provide transparent validation information such that the results can be filtered to have low false positives and be transferred to knowledge bases such as UniProtKB. The set of formats is envisioned to include a PSM-level format, a peptidoform-level format, and a site-level format. The design is in progress with initial drafts already available, and further participation from the community is welcome.

MIADE

The Minimum Information About Disorder Experiments (MIADE) guidelines aim to provide a standard to improve the reproducibility, interpretation, and dissemination of data generated by the Intrinsically Disordered Proteins (IDP) field. The guidelines provide recommendations for data producers on how to describe the results of their IDP-related experiments, for biocurators on how to annotate experimental data in manually curated community resources, and for database developers on how to disseminate the data. Information about the protein region, the structural state, and the experimental and computational approaches is required to create a MIADE-compliant description of an IDP experiment.

The PSI-IDP working group drafted and submitted for publication an article[70] describing the MIADE guidelines and will submit the full specification to the PSI document process as well. The PSI-IDP working group also implemented these guidelines in DisProt, a manually curated resource of intrinsically disordered proteins and regions from the literature.[71] With the adoption of the MIADE guidelines, DisProt biocurators can now provide more detailed and comprehensive annotation by including information about the sequence construct, the experimental conditions, and experimental components as follows:

- Sequence construct: alterations of the sequence construct with respect to the wild-type sequence. These include - but are not limited to - mutations, PTMs, and tags.

- Experimental conditions: pH, temperature, pressure, and redox potential, along with their respective values and units.

- Experimental components: any interacting partners that are part of the experimental setup, e.g. the interacting lipid, interacting small molecule, or interacting nucleic acid.

Future work and synergies with other organizations

The PSI develops and ratifies standards as part of the community and as such attempts to foster synergistic activities with other organizations within the community in order to maximize its impact. As one of the long-standing HUPO initiatives, the PSI undertakes development efforts that further the mission of HUPO, its other initiatives, and major projects such as the Human Proteome Project (HPP). An important example is the participation in the development of the HPP MS Data Interpretation Guidelines (versions 1, 2.1, and 3.0) as described above. These guidelines are not formally a PSI product, but were developed with the extensive experience of PSI members. As also highlighted above, the PSI cooperates extensively with the ProteomeXchange and the IMEx Consortia, which both actively use and promote the PSI standards.

Although primarily focused on proteomics, the PSI does reach out to, and has members from the metabolomics and lipidomics communities to extend the use of PSI standards for metabolomics applications, with substantial success for mzML and mzTab-M, and good future perspectives for USI and mzSpeclib. In this context, the PSI also fosters collaboration with computational-focused groups, such as the CompMS group (https://compms.org), which promotes the development of computational mass spectrometry algorithms and training in their use. The CompMS group includes participation from both proteomics and metabolomics researchers who use MS.

The PSI also collaborates with the European Bioinformatics Community for Mass Spectrometry[72] (EuBIC-MS) (https://eubic-ms.org/), which promotes development of MS bioinformatics tools and provides training on how to apply them. As a concrete collaboration, the EuBIC-MS group was leading the development of the MAGE-TAB for Proteomics format.

The Global Alliance for Genomics and Health[73] (GA4GH) is a policy-framing and technical standards-setting organization, enabling the responsible sharing of clinical and sensitive genomic data through both harmonized data aggregation and federated approaches. The discussion about whether human sensitive proteomics data should be subjected to the same access restrictions as sensitive DNA/RNA sequencing data has just started.[74,75] The PSI will then follow closely the developments of the GA4GH since clearly, some existing standards could be adapted or extended to support proteomics data, if needed. In our view, definitely, it would not make sense to "reinvent the wheel".

Conclusion

We have presented here an overview of the most important aspects of the PSI after 20 years of efforts, including its current operation, the current status of existing standards, ongoing work for standards in progress, and synergies between the PSI and community groups. These activities demonstrate the commitment of the PSI in accelerating the pace of biomedical research by facilitating the dissemination and reuse of data and interoperability of software. There are many ongoing standards in development, and it is perhaps worth reiterating that greater community participation is always welcome, since such greater participation leads to better standards. As the PSI completes these standards, new needs will emerge as the field advances. Already there is growing interest in standardization around the rapidly emerging high-throughput affinity-based protein quantification platforms, and some initial discussions have taken place at PSI workshops, although a clear plan has not yet emerged.

The PSI will continue its efforts to maintain and enhance existing standards to meet emerging needs. While many of the current PSI standards focus primarily on data-dependent acquisition workflows, more efforts are needed to apply standardization to data-independent acquisition workflows.[76] As software development leads to ever greater automation and artificial intelligence, the PSI should continue to foster the development and implementation of APIs and interchange systems that allow intelligent agents and software to interoperate in ways so that end-users need not concern themselves with which standards are being used, but rather rely on an ecosystem of formats, APIs, and software that allows them to focus only on their science.

Supporting Information

None

Author Information

*Address correspondence to:

- Andrew R. Jones: Email: jonesar@liverpool.ac.uk, Phone: , Fax:
- Eric W. Deutsch: Email: edeutsch@systemsbiology.org, Phone: 206-732-1200, Fax: 206-732-1299.

Notes

The authors declare no competing financial interest.

Acknowledgements

EWD acknowledges funding from the National Institutes of Health (NIH) grants R01GM087221, R24GM127667, U19AG023122, and from the National Science Foundation grants DBI-1933311, and IOS-1922871. JAV wants to acknowledge the funding received from BBSRC [BB/S01781X/1, BB/T019670/1,BB/N022440/1, BB/K01997X/1, BB/L024225/1, BB/V018779/1], Wellcome [208391/Z/17/Z, 223745/Z/21/Z], NIH [R24 GM127667-01] and EMBL core funding. ARJ wishes to acknowledge funding from BBSRC [BB/T019557/1, BB/S01781X/1, BB/R02216X/1, BB/L024128/1, BB/K01997X/1]. SK acknowledges funding from the JST NBDC grant [18063028] and JSPS KAKENHI [20H03245]. RG received funding from the Research Foundation Flanders (FWO) [1S50918N]. SO was supported by was supported by the National Human Genome Research Institute (NHGRI), Office of Director (OD/DPCPSI/ODSS), National Institute of Allergy and Infectious Diseases (NIAID), National Institute on Aging (NIA), National Institute of General Medical Sciences (NIGMS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Eye Institute (NEI), National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under Award Number [U24HG007822] (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health) and EMBL core funding.NH and SN acknowledge funding by the Bundesministerium für Bildung und Forschung (de.NBI/BMBF 031L0108A and de.NBI/BMBF 031L0107, respectively). SCET received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 778247 as well as ELIXIR implementation studies. NB acknowledges funding from the National Institutes of Health (1R01LM013115) and National Science Foundation (ABI 1759980).

References

(1) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol Cell Proteomics* **2014**, *13* (1), 339–347. https://doi.org/10.1074/mcp.M113.034769.

(2) Huttlin, E. L.; Ting, L.; Bruckner, R. J.; Gebreab, F.; Gygi, M. P.; Szpyt, J.; Tam, S.; Zarraga, G.; Colby, G.; Baltier, K.; Dong, R.; Guarani, V.; Vaites, L. P.; Ordureau, A.; Rad, R.; Erickson, B. K.; Wühr, M.; Chick, J.; Zhai, B.; Kolippakkam, D.; Mintseris, J.; Obar, R. A.; Harris, T.; Artavanis-Tsakonas, S.; Sowa, M. E.; De Camilli, P.; Paulo, J. A.; Harper, J. W.; Gygi, S. P. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **2015**, *162* (2), 425–440. https://doi.org/10.1016/j.cell.2015.06.043.

(3) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P.-A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H.-J.; Albar, J. P.; Martinez-Bartolomé, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226. https://doi.org/10.1038/nbt.2839.

(4) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. *Nucleic Acids Res.* **2017**, *45* (D1), D1100–D1106. https://doi.org/10.1093/nar/gkw936.

(5) Deutsch, E. W.; Bandeira, N.; Sharma, V.; Perez-Riverol, Y.; Carver, J. J.; Kundu, D. J.; García-Seisdedos, D.; Jarnuczak, A. F.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; Hermjakob, H.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaíno, J. A. The ProteomeXchange Consortium in 2020: Enabling "big Data" Approaches in Proteomics. *Nucleic Acids Res.* **2020**, *48* (D1), D1145–D1152. https://doi.org/10.1093/nar/gkz984.

(6) Porras, P.; Barrera, E.; Bridge, A.; Del-Toro, N.; Cesareni, G.; Duesbury, M.; Hermjakob, H.; Iannuccelli, M.; Jurisica, I.; Kotlyar, M.; Licata, L.; Lovering, R. C.; Lynn, D. J.; Meldal, B.; Nanduri, B.; Paneerselvam, K.; Panni, S.; Pastrello, C.; Pellegrini, M.; Perfetto, L.; Rahimzadeh, N.; Ratan, P.; Ricard-Blum, S.; Salwinski, L.; Shirodkar, G.; Shrivastava, A.; Orchard, S. Towards a Unified Open Access Dataset of Molecular Interactions. *Nat Commun* **2020**, *11* (1), 6144. https://doi.org/10.1038/s41467-020-19942-z.

(7) 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. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* **2016**, *3*, 160018. https://doi.org/10.1038/sdata.2016.18.

(8) Wood-Charlson, E. M.; Crockett, Z.; Erdmann, C.; Arkin, A. P.; Robinson, C. B. Ten Simple Rules for Getting and Giving Credit for Data. *PLoS Comput Biol* **2022**, *18* (9), e1010476. https://doi.org/10.1371/journal.pcbi.1010476.

(9) Hanash, S.; Celis, J. E. The Human Proteome Organization: A Mission to Advance Proteome Knowledge. *Mol. Cell Proteomics* **2002**, *1* (6), 413–414.

(10) Orchard, S.; Hermjakob, H.; Apweiler, R. The Proteomics Standards Initiative. *Proteomics* **2003**, *3* (7), 1374–1376. https://doi.org/10.1002/pmic.200300496.

(11) 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. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J. Proteome Res.* **2017**, *16* (12), 4288–4298. https://doi.org/10.1021/acs.jproteome.7b00370.

(12) del-Toro, N.; Dumousseau, M.; Orchard, S.; Jimenez, R. C.; Galeota, E.; Launay, G.; Goll, J.; Breuer, K.; Ono, K.; Salwinski, L.; Hermjakob, H. A New Reference Implementation of the PSICQUIC Web Service. *Nucleic Acids Res* **2013**, *41* (Web Server issue), W601-606. https://doi.org/10.1093/nar/gkt392.

(13) Vizcaíno, J. A.; Martens, L.; Hermjakob, H.; Julian, R. K.; Paton, N. W. The PSI Formal Document Process and Its Implementation on the PSI Website. *Proteomics* **2007**, *7* (14), 2355–2357. https://doi.org/10.1002/pmic.200700064.

(14) Mayer, G.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Orchard, S.; Montecchi-Palazzi, L.; Vizcaíno, J. A.; Hermjakob, H.; Oveillero, D.; Julian, R.; Stephan, C.; Meyer, H. E.; Eisenacher, M. Controlled Vocabularies and Ontologies in Proteomics: Overview, Principles and Practice. *Biochim. Biophys. Acta* **2014**, *1844* (1 Pt A), 98–107. https://doi.org/10.1016/j.bbapap.2013.02.017.

(15) Hermjakob, H.; Montecchi-Palazzi, L.; Bader, G.; Wojcik, J.; Salwinski, L.; Ceol, A.; Moore, S.; Orchard, S.; Sarkans, U.; von Mering, C.; Roechert, B.; Poux, S.; Jung, E.; Mersch, H.; Kersey, P.; Lappe, M.; Li, Y.; Zeng, R.; Rana, D.; Nikolski, M.; Husi, H.; Brun, C.; Shanker, K.; Grant, S. G. N.; Sander, C.; Bork, P.; Zhu, W.; Pandey, A.; Brazma, A.; Jacq, B.; Vidal, M.; Sherman, D.; Legrain, P.; Cesareni, G.; Xenarios, I.; Eisenberg, D.; Steipe, B.; Hogue, C.; Apweiler, R. The HUPO PSI's Molecular Interaction Format--a Community Standard for the Representation of Protein Interaction Data. *Nat Biotechnol* **2004**, *22* (2), 177–183. https://doi.org/10.1038/nbt926.

(16) 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.; HUPO-PSI Group. The HUPO Proteomics Standards Initiative- Mass Spectrometry Controlled Vocabulary. *Database (Oxford)* **2013**, *2013*, bat009. https://doi.org/10.1093/database/bat009.

(17) 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. MzML--a Community Standard for Mass Spectrometry Data. *Mol. Cell Proteomics* **2011**, *10* (1), R110.000133. https://doi.org/10.1074/mcp.R110.000133.

(18) Côté, R. G.; Jones, P.; Apweiler, R.; Hermjakob, H. The Ontology Lookup Service, a Lightweight Cross-Platform Tool for Controlled Vocabulary Queries. *BMC Bioinformatics* **2006**, *7*, 97. https://doi.org/10.1186/1471-2105-7-97.

(19) Perez-Riverol, Y.; Ternent, T.; Koch, M.; Barsnes, H.; Vrousgou, O.; Jupp, S.; Vizcaíno, J. A. OLS Client and OLS Dialog: Open Source Tools to Annotate Public Omics Datasets. *Proteomics* **2017**, *17* (19). https://doi.org/10.1002/pmic.201700244.

(20) Whetzel, P. L.; Noy, N. F.; Shah, N. H.; Alexander, P. R.; Nyulas, C.; Tudorache, T.; Musen, M. A. BioPortal: Enhanced Functionality via New Web Services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications. *Nucleic Acids Res* **2011**, *39* (Web Server issue), W541-545. https://doi.org/10.1093/nar/gkr469.

(21) Montecchi-Palazzi, L.; Beavis, R.; Binz, P.-A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S. The PSI-MOD Community Standard for Representation of Protein Modification Data. *Nat. Biotechnol.* **2008**, *26* (8), 864–866. https://doi.org/10.1038/nbt0808-864.

(22) Mayer, G. XLMOD: Cross-Linking and Chromatography Derivatization Reagents Ontology. **2020**. https://doi.org/10.48550/ARXIV.2003.00329.

(23) Orchard, S.; Salwinski, L.; Kerrien, S.; Montecchi-Palazzi, L.; Oesterheld, M.; Stümpflen, V.; Ceol, A.; Chatr-aryamontri, A.; Armstrong, J.; Woollard, P.; Salama, J. J.; Moore, S.; Wojcik, J.; Bader, G. D.; Vidal, M.; Cusick, M. E.; Gerstein, M.; Gavin, A.-C.; Superti-Furga, G.; Greenblatt, J.; Bader, J.; Uetz, P.; Tyers, M.; Legrain, P.; Fields, S.; Mulder, N.; Gilson, M.; Niepmann, M.; Burgoon, L.; De Las Rivas, J.; Prieto, C.; Perreau, V. M.; Hogue, C.; Mewes, H.-W.; Apweiler, R.; Xenarios, I.; Eisenberg, D.; Cesareni, G.; Hermjakob, H. The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx). *Nat Biotechnol* **2007**, *25* (8), 894–898. https://doi.org/10.1038/nbt1324.

(24) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P.-A.; Julian, R. K.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; Dunn, M. J.; Heck, A. J. R.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T. A.; Patterson, S. D.; Ping, P.; Seymour, S. L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T. M.; Whitelegge, J. P.; Wilkins, M. R.; Xenarios, I.; Yates, J. R.; Hermjakob, H. The Minimum Information about a Proteomics Experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25* (8), 887–893. https://doi.org/10.1038/nbt1329.

(25) Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. Minimum Information about a Microarray Experiment (MIAME)-toward Standards for Microarray Data. *Nat Genet* **2001**, *29* (4), 365–371. https://doi.org/10.1038/ng1201-365.

(26) Jones, A. R.; Carroll, K.; Knight, D.; Maclellan, K.; Domann, P. J.; Legido-Quigley, C.; Huang, L.; Smallshaw, L.; Mirzaei, H.; Shofstahl, J.; Paton, N. W.; Minimum Information About a Proteomics Experiment (MIAPE). Guidelines for Reporting the Use of Column Chromatography in Proteomics. *Nat. Biotechnol.* **2010**, *28* (7), 654. https://doi.org/10.1038/nbt0710-654a.

(27) Taylor, C. F.; Binz, P.-A.; Aebersold, R.; Affolter, M.; Barkovich, R.; Deutsch, E. W.; Horn, D. M.; Hühmer, A.; Kussmann, M.; Lilley, K.; Macht, M.; Mann, M.; Müller, D.; Neubert, T. A.; Nickson, J.; Patterson, S. D.; Raso, R.; Resing, K.; Seymour, S. L.; Tsugita, A.; Xenarios, I.; Zeng, R.; Julian, R. K. Guidelines for Reporting the Use of Mass Spectrometry in Proteomics. *Nat. Biotechnol.* **2008**, *26* (8), 860–861. https://doi.org/10.1038/nbt0808-860.

(28) Binz, P.-A.; Barkovich, R.; Beavis, R. C.; Creasy, D.; Horn, D. M.; Julian, R. K.; Seymour, S. L.; Taylor, C. F.; Vandenbrouck, Y. Guidelines for Reporting the Use of Mass Spectrometry Informatics in Proteomics. *Nat. Biotechnol.* **2008**, *26* (8), 862. https://doi.org/10.1038/nbt0808-862.

(29) Martínez-Bartolomé, S.; Deutsch, E. W.; Binz, P.-A.; Jones, A. R.; Eisenacher, M.; Mayer, G.; Campos, A.; Canals, F.; Bech-Serra, J.-J.; Carrascal, M.; Gay, M.; Paradela, A.; Navajas, R.; Marcilla, M.; Hernáez, M. L.; Gutiérrez-Blázquez, M. D.; Velarde, L. F. C.; Aloria, K.; Beaskoetxea, J.; Medina-Aunon, J. A.; Albar, J. P. Guidelines for Reporting Quantitative Mass Spectrometry Based Experiments in Proteomics. *J Proteomics* **2013**, *95*, 84–88. https://doi.org/10.1016/j.jprot.2013.02.026.

(30) Medina-Aunon, J. A.; Martínez-Bartolomé, S.; López-García, M. A.; Salazar, E.; Navajas, R.; Jones, A. R.; Paradela, A.; Albar, J. P. The ProteoRed MIAPE Web Toolkit: A User-Friendly Framework to Connect and Share Proteomics Standards. *Mol. Cell Proteomics* **2011**, *10* (10), M111.008334. https://doi.org/10.1074/mcp.M111.008334.

(31) Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y.-K.; Weintraub, S. T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; Hancock, W. S.; Hermjakob, H.; Aebersold, R.; Moritz, R. L.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016**, *15* (11), 3961–3970. https://doi.org/10.1021/acs.jproteome.6b00392.

(32) Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y.-K.; Weintraub, S. T.; Vandenbrouck, Y.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J. Proteome Res.* **2019**, *18* (12), 4108–4116. https://doi.org/10.1021/acs.jproteome.9b00542.

(33) Kerrien, S.; Orchard, S.; Montecchi-Palazzi, L.; Aranda, B.; Quinn, A. F.; Vinod, N.; Bader, G. D.; Xenarios, I.; Wojcik, J.; Sherman, D.; Tyers, M.; Salama, J. J.; Moore, S.; Ceol, A.; Chatr-Aryamontri, A.; Oesterheld, M.; Stümpflen, V.; Salwinski, L.; Nerothin, J.; Cerami, E.; Cusick, M. E.; Vidal, M.; Gilson, M.; Armstrong, J.; Woollard, P.; Hogue, C.; Eisenberg, D.; Cesareni, G.; Apweiler, R.; Hermjakob, H. Broadening the Horizon--Level 2.5 of the HUPO-PSI Format for Molecular Interactions. *BMC Biol.* **2007**, *5*, 44. https://doi.org/10.1186/1741-7007-5-44.

(34) Sivade Dumousseau, M.; Alonso-López, D.; Ammari, M.; Bradley, G.; Campbell, N. H.; Ceol, A.; Cesareni, G.; Combe, C.; De Las Rivas, J.; Del-Toro, N.; Heimbach, J.; Hermjakob, H.; Jurisica, I.; Koch, M.; Licata, L.; Lovering, R. C.; Lynn, D. J.; Meldal, B. H. M.; Micklem, G.; Panni, S.; Porras, P.; Ricard-Blum, S.; Roechert, B.; Salwinski, L.; Shrivastava, A.; Sullivan, J.; Thierry-Mieg, N.; Yehudi, Y.; Van Roey, K.; Orchard, S. Encompassing New Use Cases - Level 3.0 of the HUPO-PSI Format for Molecular Interactions. *BMC Bioinformatics* **2018**, *19* (1), 134. https://doi.org/10.1186/s12859-018-2118-1.

(35) Sivade Dumousseau, M.; Koch, M.; Shrivastava, A.; Alonso-López, D.; De Las Rivas, J.; Del-Toro, N.; Combe, C. W.; Meldal, B. H. M.; Heimbach, J.; Rappsilber, J.; Sullivan, J.; Yehudi, Y.; Orchard, S. JAMI: A Java Library for Molecular Interactions and Data Interoperability. *BMC Bioinformatics* **2018**, *19* (1), 133. https://doi.org/10.1186/s12859-018-2119-0.

(36) Shah, A. R.; Davidson, J.; Monroe, M. E.; Mayampurath, A. M.; Danielson, W. F.; Shi, Y.; Robinson, A. C.; Clowers, B. H.; Belov, M. E.; Anderson, G. A.; Smith, R. D. An Efficient Data Format for Mass Spectrometry-Based Proteomics. *J Am Soc Mass Spectrom* **2010**, *21* (10), 1784–1788. https://doi.org/10.1016/j.jasms.2010.06.014.

(37) Wilhelm, M.; Kirchner, M.; Steen, J. A. J.; Steen, H. Mz5: Space- and Time-Efficient Storage of Mass Spectrometry Data Sets. *Mol. Cell Proteomics* **2012**, *11* (1), O111.011379. https://doi.org/10.1074/mcp.O111.011379.

(38) Bouyssié, D.; Dubois, M.; Nasso, S.; Gonzalez de Peredo, A.; Burlet-Schiltz, O.; Aebersold, R.; Monsarrat, B. MzDB: A File Format Using Multiple Indexing Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets. *Mol. Cell Proteomics* **2015**, *14* (3), 771–781. https://doi.org/10.1074/mcp.O114.039115.

(39) Wang, J.; Lu, M.; Wang, R.; An, S.; Xie, C.; Yu, C. StackZDPD: A Novel Encoding Scheme for Mass Spectrometry Data Optimized for Speed and Compression Ratio. *Sci Rep* **2022**, *12* (1), 5384. https://doi.org/10.1038/s41598-022-09432-1.

(40) Schramm, T.; Hester, Z.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M.-F.; Stoeckli, M.; Spengler, B.; Römpp, A. ImzML--a Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data. *J Proteomics* **2012**, *75* (16), 5106–5110. https://doi.org/10.1016/j.jprot.2012.07.026.

(41) Bhamber, R. S.; Jankevics, A.; Deutsch, E. W.; Jones, A. R.; Dowsey, A. W. MzMLb: A Future-Proof Raw Mass Spectrometry Data Format Based on Standards-Compliant MzML and Optimized for Speed and Storage Requirements. *J Proteome Res* **2021**, *20* (1), 172–183. https://doi.org/10.1021/acs.jproteome.0c00192.

(42) 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. The MzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Mol. Cell Proteomics* **2012**, *11* (7), M111.014381. https://doi.org/10.1074/mcp.M111.014381.

(43) 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. The MzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. *Mol. Cell Proteomics* **2017**, *16* (7), 1275–1285. https://doi.org/10.1074/mcp.M117.068429.

(44) 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. The MzTab Data Exchange Format: Communicating Mass-Spectrometry-Based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell Proteomics* **2014**, *13* (10), 2765–2775. https://doi.org/10.1074/mcp.O113.036681.

(45) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20* (18), 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

(46) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* **2016**, *13* (9), 741–748. https://doi.org/10.1038/nmeth.3959.

(47) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat Protoc* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.

(48) Hoffmann, N.; Rein, J.; Sachsenberg, T.; Hartler, J.; Haug, K.; Mayer, G.; Alka, O.; Dayalan, S.; Pearce, J. T. M.; Rocca-Serra, P.; Qi, D.; Eisenacher, M.; Perez-Riverol, Y.; Vizcaíno, J. A.; Salek, R. M.; Neumann, S.; Jones, A. R. MzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Anal Chem* **2019**, *91* (5), 3302–3310. https://doi.org/10.1021/acs.analchem.8b04310.

(49) Menschaert, G.; Wang, X.; Jones, A. R.; Ghali, F.; Fenyö, D.; Olexiouk, V.; Zhang, B.; Deutsch, E. W.; Ternent, T.; Vizcaíno, J. A. The ProBAM and ProBed Standard Formats: Enabling a Seamless Integration of Genomics and Proteomics Data. *Genome Biol.* **2018**, *19* (1), 12. https://doi.org/10.1186/s13059-017-1377-x.

(50) Binz, P.-A.; Shofstahl, J.; Vizcaíno, J. A.; Barsnes, H.; Chalkley, R. J.; Menschaert, G.; Alpi, E.; Clauser, K.; Eng, J. K.; Lane, L.; Seymour, S. L.; Sánchez, L. F. H.; Mayer, G.; Eisenacher, M.; Perez-Riverol, Y.; Kapp, E. A.; Mendoza, L.; Baker, P. R.; Collins, A.; Van Den Bossche, T.; Deutsch, E. W. Proteomics Standards Initiative Extended FASTA Format. *J. Proteome Res.* **2019**, *18* (6), 2686–2692. https://doi.org/10.1021/acs.jproteome.9b00064.

(51) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(52) Eng, J. K.; Deutsch, E. W. Extending Comet for Global Amino Acid Variant and Post-Translational Modification Analysis Using the PSI Extended FASTA Format. *Proteomics* **2020**, *20* (21–22), e1900362. https://doi.org/10.1002/pmic.201900362.

(53) LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P.; Kelleher, N. L.; Smith, L. M.; Ge, Y.; Agar, J. N.; Chamot-Rooke, J.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17* (3), 1321–1325. https://doi.org/10.1021/acs.jproteome.7b00851.

(54) 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. Proteomics Standards Initiative's ProForma 2.0: Unifying the Encoding of Proteoforms and Peptidoforms. *J Proteome Res* **2022**, *21* (4), 1189–1195. https://doi.org/10.1021/acs.jproteome.1c00771.

(55) 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. Universal Spectrum Identifier for Mass Spectra. *Nat Methods* **2021**, *18* (7), 768–770. https://doi.org/10.1038/s41592-021-01184-6.

(56) Bittremieux, W.; Chen, C.; Dorrestein, P. C.; Schymanski, E. L.; Schulze, T.; Neumann, S.; Meier, R.; Rogers, S.; Wang, M. *Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service*; preprint; Bioinformatics, 2020. https://doi.org/10.1101/2020.05.09.086066.

(57) 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. A Proteomics Sample Metadata Representation for Multiomics Integration and Big Data Analysis. *Nat Commun* **2021**, *12* (1), 5854. https://doi.org/10.1038/s41467-021-26111-3.

(58) Rayner, T. F.; Rocca-Serra, P.; Spellman, P. T.; Causton, H. C.; Farne, A.; Holloway, E.; Irizarry, R. A.; Liu, J.; Maier, D. S.; Miller, M.; Petersen, K.; Quackenbush, J.; Sherlock, G.; Stoeckert, C. J.; White, J.; Whetzel, P. L.; Wymore, F.; Parkinson, H.; Sarkans, U.; Ball, C. A.; Brazma, A. A Simple Spreadsheet-Based, MIAME-Supportive Format for Microarray Data: MAGE-TAB. *BMC Bioinformatics* **2006**, *7*, 489. https://doi.org/10.1186/1471-2105-7-489.

(59) Gibson, F.; Hoogland, C.; Martinez-Bartolomé, S.; Medina-Aunon, J. A.; Albar, J. P.; Babnigg, G.; Wipat, A.; Hermjakob, H.; Almeida, J. S.; Stanislaus, R.; Paton, N. W.; Jones, A. R. The Gel Electrophoresis Markup Language (GelML) from the Proteomics Standards Initiative. *Proteomics* **2010**, *10* (17), 3073–3081. https://doi.org/10.1002/pmic.201000120.

(60) Deutsch, E. W.; Chambers, M.; Neumann, S.; Levander, F.; Binz, P.-A.; Shofstahl, J.; Campbell, D. S.; Mendoza, L.; Ovelleiro, D.; Helsens, K.; Martens, L.; Aebersold, R.; Moritz, R. L.; Brusniak, M.-Y. TraML--a Standard Format for Exchange of Selected Reaction Monitoring Transition Lists. *Mol. Cell Proteomics* **2012**, *11* (4), R111.015040. https://doi.org/10.1074/mcp.R111.015040.

(61) Helsens, K.; Brusniak, M.-Y.; Deutsch, E.; Moritz, R. L.; Martens, L. JTraML: An Open Source Java API for TraML, the PSI Standard for Sharing SRM Transitions. *J. Proteome Res.* **2011**, *10* (11), 5260–5263. https://doi.org/10.1021/pr200664h.

(62) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(63) Walzer, M.; Qi, D.; Mayer, G.; Uszkoreit, J.; Eisenacher, M.; Sachsenberg, T.; Gonzalez-Galarza, F. F.; Fan, J.; Bessant, C.; Deutsch, E. W.; Reisinger, F.; Vizcaíno, J. A.; Medina-Aunon, J. A.; Albar, J. P.; Kohlbacher, O.; Jones, A. R. The MzQuantML Data Standard for Mass Spectrometry-Based Quantitative Studies in Proteomics. *Mol. Cell Proteomics* **2013**, *12* (8), 2332–2340. https://doi.org/10.1074/mcp.O113.028506.

(64) Walzer, M.; Pernas, L. E.; Nasso, S.; Bittremieux, W.; Nahnsen, S.; Kelchtermans, P.; Pichler, P.; van den Toorn, H. W. P.; Staes, A.; Vandenbussche, J.; Mazanek, M.; Taus, T.; Scheltema, R. A.; Kelstrup, C. D.; Gatto, L.; van Breukelen, B.; Aiche, S.; Valkenborg, D.; Laukens, K.; Lilley, K. S.; Olsen, J. V.; Heck, A. J. R.; Mechtler, K.; Aebersold, R.; Gevaert, K.; Vizcaíno, J. A.; Hermjakob, H.; Kohlbacher, O.; Martens, L. QcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments. *Mol Cell Proteomics* **2014**, *13* (8), 1905–1913. https://doi.org/10.1074/mcp.M113.035907.

(65) Bittremieux, W.; Walzer, M.; Tenzer, S.; Zhu, W.; Salek, R. M.; Eisenacher, M.; Tabb, D. L. The Human Proteome Organization-Proteomics Standards Initiative Quality Control Working Group: Making Quality Control More Accessible for Biological Mass Spectrometry. *Anal. Chem.* **2017**, *89* (8), 4474–4479. https://doi.org/10.1021/acs.analchem.6b04310.

(66) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667. https://doi.org/10.1002/pmic.200600625.

(67) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building Consensus Spectral Libraries for Peptide Identification in Proteomics. *Nat. Methods* **2008**, *5* (10), 873–875. https://doi.org/10.1038/nmeth.1254.

(68) Frewen, B.; MacCoss, M. J. Using BiblioSpec for Creating and Searching Tandem MS Peptide Libraries. *Curr Protoc Bioinformatics* **2007**, *Chapter 13*, Unit 13.7. https://doi.org/10.1002/0471250953.bi1307s20.

(69) 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. Expanding the Use of Spectral Libraries in Proteomics. *J. Proteome Res.* **2018**. https://doi.org/10.1021/acs.jproteome.8b00485.

(70) Mészáros, B.; Hatos, A.; Palopoli, N.; Quaglia, F.; Salladini, E.; Van Roey, K.; Arthanari, H.; Dosztányi, Z.; Felli, I. C.; Fischer, P. D.; Hoch, J. C.; Jeffries, C. M.; Longhi, S.; Maiani, E.; Orchard, S.; Pancsa, R.; Papaleo, E.; Pierattelli, R.; Piovesan, D.; Pritisanac, I.; Viennet, T.; Tompa, P.; Vranken, W.; Tosatto, S. C.; Davey, N. E. *MIADE Metadata Guidelines: Minimum Information About a Disorder Experiment*; preprint; Scientific Communication and Education, 2022. https://doi.org/10.1101/2022.07.12.495092.

(71) Quaglia, F.; Mészáros, B.; Salladini, E.; Hatos, A.; Pancsa, R.; Chemes, L. B.; Pajkos, M.; Lazar, T.; Peña-Díaz, S.; Santos, J.; Ács, V.; Farahi, N.; Fichó, E.; Aspromonte, M. C.; Bassot, C.; Chasapi, A.; Davey, N. E.; Davidović, R.; Dobson, L.; Elofsson, A.; Erdős, G.; Gaudet, P.; Giglio, M.; Glavina, J.; Iserte, J.; Iglesias, V.; Kálmán, Z.; Lambrughi, M.; Leonardi, E.; Longhi, S.; Macedo-Ribeiro, S.; Maiani, E.; Marchetti, J.; Marino-Buslje, C.; Mészáros, A.; Monzon, A. M.; Minervini, G.; Nadendla, S.; Nilsson, J. F.; Novotný, M.; Ouzounis, C. A.; Palopoli, N.; Papaleo, E.; Pereira, P. J. B.; Pozzati, G.; Promponas, V. J.; Pujols, J.; Rocha, A. C. S.; Salas, M.; Sawicki, L. R.; Schad, E.; Shenoy, A.; Szaniszló, T.; Tsirigos, K. D.; Veljkovic, N.; Parisi, G.; Ventura, S.; Dosztányi, Z.; Tompa, P.; Tosatto, S. C. E.; Piovesan, D. DisProt in 2022: Improved Quality and Accessibility of Protein Intrinsic Disorder Annotation. *Nucleic Acids Res* **2022**, *50* (D1), D480–D487. https://doi.org/10.1093/nar/gkab1082.

(72) Bittremieux, W.; Bouyssié, D.; Dorfer, V.; Locard-Paulet, M.; Perez-Riverol, Y.; Schwämmle, V.; Uszkoreit, J.; Van Den Bossche, T. The European Bioinformatics Community for Mass Spectrometry (EuBIC-MS): An Open Community for Bioinformatics Training and Research. *Rapid Commun Mass Spectrom* **2021**, e9087. https://doi.org/10.1002/rcm.9087.

(73) Rehm, H. L.; Page, A. J. H.; Smith, L.; Adams, J. B.; Alterovitz, G.; Babb, L. J.; Barkley, M. P.; Baudis, M.; Beauvais, M. J. S.; Beck, T.; Beckmann, J. S.; Beltran, S.; Bernick, D.; Bernier, A.; Bonfield, J. K.; Boughtwood, T. F.; Bourque, G.; Bowers, S. R.; Brookes, A. J.; Brudno, M.; Brush, M. H.; Bujold, D.; Burdett, T.; Buske, O. J.; Cabili, M. N.; Cameron, D. L.;

Carroll, R. J.; Casas-Silva, E.; Chakravarty, D.; Chaudhari, B. P.; Chen, S. H.; Cherry, J. M.; Chung, J.; Cline, M.; Clissold, H. L.; Cook-Deegan, R. M.; Courtot, M.; Cunningham, F.; Cupak, M.; Davies, R. M.; Denisko, D.; Doerr, M. J.; Dolman, L. I.; Dove, E. S.; Dursi, L. J.; Dyke, S. O. M.; Eddy, J. A.; Eilbeck, K.; Ellrott, K. P.; Fairley, S.; Fakhro, K. A.; Firth, H. V.; Fitzsimons, M. S.; Fiume, M.; Flicek, P.; Fore, I. M.; Freeberg, M. A.; Freimuth, R. R.; Fromont, L. A.; Fuerth, J.; Gaff, C. L.; Gan, W.; Ghanaim, E. M.; Glazer, D.; Green, R. C.; Griffith, M.; Griffith, O. L.; Grossman, R. L.; Groza, T.; Auvil, J. M. G.; Guigó, R.; Gupta, D.; Haendel, M. A.; Hamosh, A.; Hansen, D. P.; Hart, R. K.; Hartley, D. M.; Haussler, D.; Hendricks-Sturrup, R. M.; Ho, C. W. L.; Hobb, A. E.; Hoffman, M. M.; Hofmann, O. M.; Holub, P.; Hsu, J. S.; Hubaux, J.-P.; Hunt, S. E.; Husami, A.; Jacobsen, J. O.; Jamuar, S. S.; Janes, E. L.; Jeanson, F.; Jené, A.; Johns, A. L.; Joly, Y.; Jones, S. J. M.; Kanitz, A.; Kato, K.; Keane, T. M.; Kekesi-Lafrance, K.; Kelleher, J.; Kerry, G.; Khor, S.-S.; Knoppers, B. M.; Konopko, M. A.; Kosaki, K.; Kuba, M.; Lawson, J.; Leinonen, R.; Li, S.; Lin, M. F.; Linden, M.; Liu, X.; Udara Liyanage, I.; Lopez, J.; Lucassen, A. M.; Lukowski, M.; Mann, A. L.; Marshall, J.; Mattioni, M.; Metke-Jimenez, A.; Middleton, A.; Milne, R. J.; Molnár-Gábor, F.; Mulder, N.; Munoz-Torres, M. C.; Nag, R.; Nakagawa, H.; Nasir, J.; Navarro, A.; Nelson, T. H.; Niewielska, A.; Nisselle, A.; Niu, J.; Nyrönen, T. H.; O'Connor, B. D.; Oesterle, S.; Ogishima, S.; Wang, V. O.; Paglione, L. A. D.; Palumbo, E.; Parkinson, H. E.; Philippakis, A. A.; Pizarro, A. D.; Prlic, A.; Rambla, J.; Rendon, A.; Rider, R. A.; Robinson, P. N.; Rodarmer, K. W.; Rodriguez, L. L.; Rubin, A. F.; Rueda, M.; Rushton, G. A.; Ryan, R. S.; Saunders, G. I.; Schuilenburg, H.; Schwede, T.; Scollen, S.; Senf, A.; Sheffield, N. C.; Skantharajah, N.; Smith, A. V.; Sofia, H. J.; Spalding, D.; Spurdle, A. B.; Stark, Z.; Stein, L. D.; Suematsu, M.; Tan, P.; Tedds, J. A.; Thomson, A. A.; Thorogood, A.; Tickle, T. L.; Tokunaga, K.; Törnroos, J.; Torrents, D.; Upchurch, S.; Valencia, A.; Guimera, R. V.; Vamathevan, J.; Varma, S.; Vears, D. F.; Viner, C.; Voisin, C.; Wagner, A. H.; Wallace, S. E.; Walsh, B. P.; Williams, M. S.; Winkler, E. C.; Wold, B. J.; Wood, G. M.; Woolley, J. P.; Yamasaki, C.; Yates, A. D.; Yung, C. K.; Zass, L. J.; Zaytseva, K.; Zhang, J.; Goodhand, P.; North, K.; Birney, E. GA4GH: International Policies and Standards for Data Sharing across Genomic Research and Healthcare. *Cell Genom* **2021**, *1* (2), 100029. https://doi.org/10.1016/j.xgen.2021.100029.

(74) Keane, T. M.; O'Donovan, C.; Vizcaíno, J. A. The Growing Need for Controlled Data Access Models in Clinical Proteomics and Metabolomics. *Nat Commun* **2021**, *12* (1), 5787. https://doi.org/10.1038/s41467-021-26110-4.

(75) Bandeira, N.; Deutsch, E. W.; Kohlbacher, O.; Martens, L.; Vizcaíno, J. A. Data Management of Sensitive Human Proteomics Data: Current Practices, Recommendations, and Perspectives for the Future. *Mol Cell Proteomics* **2021**, *20*, 100071. https://doi.org/10.1016/j.mcpro.2021.100071.

(76) Jones, A. R.; Deutsch, E. W.; Vizcaíno, J. A. Is DIA Proteomics Data FAIR? Current Data Sharing Practices, Available Bioinformatics Infrastructure and Recommendations for the Future. *Proteomics* **2022**, e2200014. https://doi.org/10.1002/pmic.202200014.