Nucleic Acids Research, 2026, **54**, D459–D469
https://doi.org/10.1093/nar/gkaf1146
Advance access publication date: 6 November 2025
Database issue

The ProteomeXchange consortium in 2026: making proteomics data FAIR

Eric W. Deutsch [1,*], Nuno Bandeira [2,3,4], Yasset Perez-Riverol [5], Vagisha Sharma [6], Jeremy J. Carver [2,3,4], Luis Mendoza [1], Deepti J. Kundu [5], Chakradhar Bandla [5], Selvakumar Kamatchinathan [5], Suresh Hewapathirana [5], Zhi Sun [1], Shin Kawano [7,8], Shujiro Okuda [9], Brian Connolly [6], Brendan MacLean [6], Michael J. MacCoss [6], Tao Chen [10], Yunping Zhu [10], Yasushi Ishihama [11], Juan Antonio Vizcaíno [5,*]

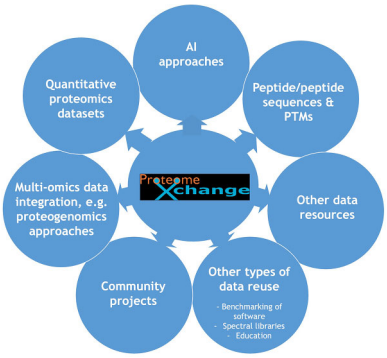
[1]Institute for Systems Biology, Seattle WA 98109, United States
[2]Center for Computational Mass Spectrometry, University of California, San Diego (UCSD), La Jolla, CA 92093, United States
[3]Department of Computer Science and Engineering, University of California, San Diego (UCSD), La Jolla, CA 92093, United States
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego (UCSD), La Jolla, CA 92093, United States
[5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[6]Department of Genome Sciences, University of Washington, Seattle WA 98195, United States
[7]School of Frontier Engineering, Kitasato University, Kanagawa 252-0373, Japan
[8]Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Chiba 277-0871, Japan
[9]Niigata University Graduate School of Medical and Dental Sciences, Niigata 951-8510, Japan
[10]Beijing Proteome Research Center, National Center for Protein Sciences, Beijing Institute of Lifeomics, Beijing 102206, China
[11]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan

*To whom correspondence should be addressed. Email: juan@ebi.ac.uk
Correspondence may also be addressed to Eric Deutsch. Email: eric.deutsch@isbscience.org

Abstract

The ProteomeXchange consortium of proteomics resources (http://www.proteomexchange.org) was established to standardize open data practices in the mass spectrometry (MS)-based proteomics field. Here, we describe the main developments in ProteomeXchange in the last 3 years. The six member databases of ProteomeXchange, spread out in three different continents, are the PRIDE database, PeptideAtlas, MassIVE, jPOST, iProX, and Panorama Public. We provide updated data submission statistics, showcasing that the number of datasets submitted to ProteomeXchange resources has continued to accelerate every year. Through June 2025, 64 330 datasets had been submitted to ProteomeXchange resources, and from those, 30 097 (47%) just in the last 3 years. We also report on the improvements in the support for the standards developed by the Proteomics Standards Initiative, e.g. for Universal Spectrum Identifiers and for SDRF (Sample and Data Relationship Format)-Proteomics. Additionally, we highlight the increase in data reuse activities of public datasets, including targeted reanalyses of datasets of different proteomics data types, and the development of novel machine learning approaches. Finally, we summarize our plans for the near future, covering the development of resources for controlled-access human proteomics data, and for the support of non-MS proteomics approaches.

Graphical abstract



Received: September 17, 2025. **Revised:** October 9, 2025. **Accepted:** October 9, 2025

Introduction

Public data sharing has become the default practice in the proteomics field, which is increasingly prominent in the life sciences as proteomics-only approaches or as part of multi-omics studies. Open data practices have generally been adopted due to the requirements of funding agencies and scientific journals, but also due to the perceived reliability of proteomics public data repositories. In the last decade, there has been a dramatic increase in the amount of proteomics data in the public domain. This has triggered multiple types of data re-use activities that are increasingly contributing to the rapid development of the field, for instance in the context of machine learning (ML) and artificial intelligence (AI) approaches [1], among many others.

Since 2012, the main proteomics data repositories are working together under the umbrella of the ProteomeXchange Consortium (http://www.proteomexchange.org) [2–5]. ProteomeXchange has standardized open data practices internationally via data submission and dissemination of public mass spectrometry (MS)-based proteomics datasets. There are currently six resources that are members of ProteomeXchange: the PRIDE database [6] (European Bioinformatics Institute, EMBL-EBI, Hinxton, UK), MassIVE (University of California San Diego, USA, in 2014) [7], jPOST [8] (Kyoto University and other institutions, Japan), iProX [9] (National Center for Protein Sciences, Beijing, China), and the smaller resources Panorama Public [10] (University of Washington, Seattle, USA), and PASSEL [11] [as part of PeptideAtlas (Institute for Systems Biology, Seattle, USA)]. ProteomeCentral (https://proteomecentral.proteomexchange.org/) is the common web portal to access publicly available datasets across all six ProteomeXchange resources.

ProteomeXchange data resources are committed to implementing the FAIR (Findable, Accessible, Interoperable, Reusable) principles [12] for biological data, supporting reproducible research. For this reason, ProteomeXchange resources are closely aligned with the activities of the Proteomics Standards Initiative (PSI; https://psidev.info/), the organization that develops community-based open data standards in the field [13, 14]. As a result, ProteomeXchange resources support the main MS-related PSI open data formats, and other PSI standards such as Universal Spectrum Identifiers (USIs) [15], the ProForma 2.0 notation [16], and the PSI-MS controlled vocabulary [17]. This support involves the development and maintenance of several open-source parser libraries to support these standards.

In December 2022, the ProteomeXchange resources were included in the initial list of Global Core Biodata Resources (https://globalbiodata.org/what-we-do/global-core-biodata-resources/) created by the Global Biodata Coalition, recognizing ProteomeXchange as an essential biological resource for the scientific community. The PRIDE database is also a core data resource of ELIXIR (http://www.elixir-europe.org) [18], recognizing its key role in the life sciences ecosystem in Europe.

Here we provide an update of the activities of the ProteomeXchange consortium and its individual resources since the previous update paper was published in *Nucleic Acids Research* 3 years ago [5]. We also describe updated submission statistics, demonstrating the year-on-year increase in the number of proteomics datasets in ProteomeXchange resources. As a key point, we also highlight key ongoing data re-use activities, both in the context of the ProteomeXchange resources and led by third parties, and discuss future developments. For more detailed information about the individual data resources in ProteomeXchange, please see the recent manuscripts in the NAR database issue for each [6, 8, 9].

Results

ProteomeXchange current infrastructure

There have not been any substantial changes in the ProteomeXchange data workflow for datasets in the last 3 years. PRIDE, MassIVE, jPOST, and iProX are *universal* archival resources, storing all types of MS-based proteomics experiments, while PASSEL and Panorama Public are *focussed* on targeted proteomics approaches. Table 1 summarizes the main functionality offered by the different ProteomeXchange resources. ProteomeXchange dataset (PXD) identifiers are persistent and unique, and are used as the main dataset identifier for all originally submitted datasets. RPXD identifiers are issued in some cases for reanalysed datasets (i.e. original datasets reanalysed by one of the ProteomeXchange resources). Additionally, MassIVE, jPOST, and iProX use their own identifiers for datasets in addition to the common PXD identifiers for datasets that comply with the ProteomeXchange requirements. In terms of data licenses, most resources assign as default the Creative Commons CC0 license. Panorama Public uses as its default the CC-BY license, which requires attribution, with a CC0 license also available to data submitters.

Once a dataset is submitted, all ProteomeXchange resources provide private password-controlled access for reviewers and journal editors during the manuscript review process (private datasets remain unreleased to the public during that process). Once the manuscript is published, the corresponding dataset(s) are made publicly available in each resource and their metadata are also made available via the ProteomeCentral web portal, so that they may easily be found, accessed, and reused. For all submitted datasets, a set of common experimental metadata at the level of each dataset must be sent to ProteomeCentral (the common data model is encoded in the PX-XML format, http://proteomecentral.proteomexchange.org/schemas/proteomeXchange-1.4.0.xsd), together with the raw MS run files and the processed results (identification and/or quantification data).

In the context of data submissions, the FTP file transfer protocol continues to be supported by most of the resources, while Aspera, HTTPS, and WebDAV are also supported in some cases (Table 1). The jPOST repository employs the in-house developed PRESTO upload protocol [19]. In 2024, PRIDE added support for the Globus file transfer service (https://www.globus.org/data-transfer), which is recommended for very large datasets, especially for institutions where it is not possible to use Aspera due to IT restrictions [6] (https://www.ebi.ac.uk/pride/markdownpage/globus).

Updates in ProteomeCentral

The ProteomeCentral user interface (https://proteomecentral.proteomexchange.org/) has undergone a complete rewrite to support the ever-increasing number of datasets available in ProteomeXchange. The back-end is based on the PROXI web service interface for datasets (https://github.com/HUPO-PSI/proxi-schemas). It provides advanced filtering, free text search, faceted search, summarization, and pagination via an

Table 1. Main functionality offered by the ProteomeXchange resources

Functionality	PRIDE	PASSEL	MassIVE	jPOST	iProX	Panorama public	Peptide atlas
Types of data access							
Web interface	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Application Programming Interface	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Protocol for file transfer (download/upload)	FTP, Aspera, Globus	FTP	FTP	FTP, HTTPS, PRESTO, TripleStore	HTTPS, Aspera	WebDAV, HTTPS	FTP
Reviewer private access	File download	File download	File download, web interface	File download	File download, web interface	File download, web interface	N/A
General functionality/web visualization							
Dataset centric view	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Protein centric view across resource	No	Yes	Yes	No	Yes	Yes	Yes
Annotated mass spectra	Yes	Yes	Yes	Yes	Yes	Yes	Yes
USIs	Yes	Yes	Yes	Yes	Yes	No	Yes
Chromatograms	No	Yes	Yes	No	No	Yes	No
Data license	CC0	CC0	CC0	CC0	CC0	CC-BY (default), CC0 (optional)	CC0

OpenAPI endpoint. The back end is coupled with a custom JavaScript front end that enables a smooth user experience in searching for datasets of relevance, downloading lists of datasets, summarizing subsets of PXDs, and examining individual datasets. The OpenAPI endpoint is publicly accessible and can be used by other applications. A similar system based on the same tools is being developed for spectral libraries.

ProteomeXchange and the PSI: new developments and supported formats

ProteomeXchange resources support the main open PSI data standards for MS: mzML [20] (for MS data), mzIdentML [21], and mzTab [22] (for the representation of peptide/protein identifications and quantification results). SDRF-Proteomics (Sample and Data Relationship Format)-Proteomics is the data standard for encoding sample metadata and experimental design information and their relationship to raw data files [23]. Adoption of SDRF is growing in the community, and therefore, the number of submitted datasets containing this information is also increasing in parallel. Although the most usual way to create SDRF-Proteomics files is by using spreadsheet-based software, the lesSDRF [24] web tool has been developed to facilitate the process. Additionally, some popular analysis tools in the community, such as MaxQuant, are starting to support it as well [25]. Apart from promoting the creation of SDRF-Proteomics files, there are ongoing developments for the extension of the format to support various different proteomics subdomains (e.g. metaproteomics).

In this context, as a unique initiative, the *Journal of Proteome Data and Methods* (JPDM) is a proteomics data journal that accepts data description papers using SDRF spreadsheets as supplementary material. The journal operates in conjunction with jPOST [8]. The jPOST team contacts data submitters, sends them draft SDRF files created based on jPOST data and published papers, and encourages them to review the content and submit back to JPDM. SDRF-Proteomics files are prepared by researchers registering data in the repository, enriching the dataset metadata available. Furthermore, when original data is reused, the data article is cited in addition to the PXD identifier, providing authors with an incentive to justify the extra effort of better describing the dataset metadata.

Recently, the first data descriptor paper using this workflow was published from JPDM [26], featuring supplementary material that includes jeSDRF (JPDM-empowered SDRF) files compliant with SDRF-Proteomics. The jPOST team is extending this approach to datasets from other ProteomeXchange resources, promoting the conversion of metadata for PRIDE datasets created using lesSDRF into the jeSDRF format, and encouraging the original authors to submit it to JPDM.

USI services: visualizing every mass spectrum in ProteomeXchange

USI [15] is a standardized method for expressing a multipart key identifier for every spectrum deposited to a ProteomeXchange resource (https://psidev.info/usi), allowing for greater transparency of mass spectral evidence. All ProteomeXchange resources, apart from PanoramaPublic, support them. ProteomeCentral provides an API service and a spectrum visualization tool (https://proteomecentral.proteomexchange.org/usi/) to retrieve any spectrum from any of the ProteomeXchange partners. The central service (ProteomeCentral USI) uses the USI representation to query all resources that implement the PROXI specification and to determine whether and where the spectrum is available. If found, the spectrum is then displayed through the Lorikeet spectrum viewer (http://uwpr.github.io/Lorikeet/).

Every PX partner is also able to display spectra via USIs within their own resource, and their visualization interfaces are referenced from ProteomeCentral. This is a flexible approach because every resource spectrum viewer provides additional functionalities, for example, linking to other datasets, providing additional metadata, or spectrum prediction for

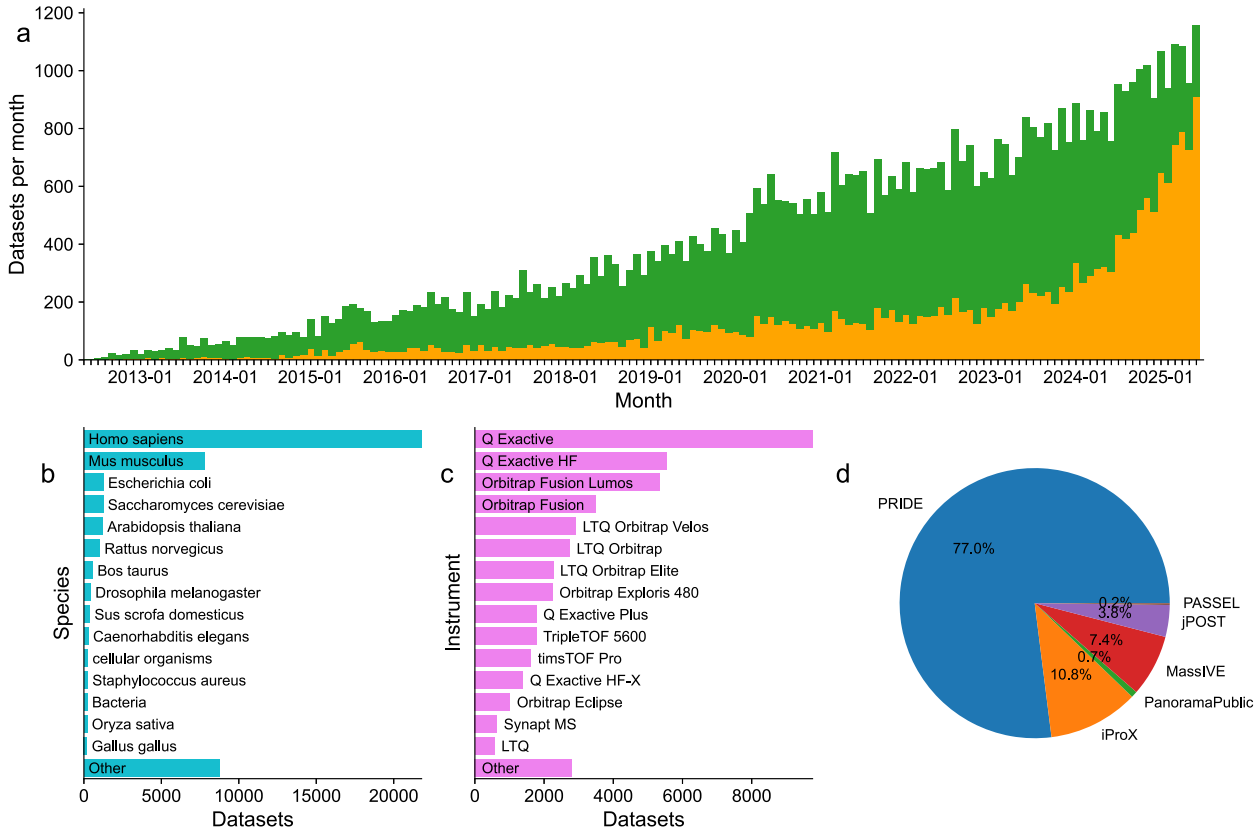


Figure 1. Summary statistics for datasets deposited to ProteomeXchange resources since 2012. (A) Trend in publicly released (green) and not-yet released (orange) datasets from May 2012 through June 2025. A total of 1156 datasets were submitted in June 2025. (B) Summary of the top 15 species for publicly released datasets since 2012. (C) Summary of the top 15 instruments as reported by submitters for publicly released datasets since 2012. (D) Summary of the relative number of all datasets by the receiving repository.

the identified peptide in the USI. There have been multiple recent improvements in the support for USIs in the different resources. First, the USI functionality in PRIDE (PRIDE Archive USI, https://www.ebi.ac.uk/pride/archive/usi) can now extract the specified scan directly from the MS raw files via ThermoRawFileParser [27] for Thermo Scientific raw files, providing access to >80% of the raw files in PRIDE. A new spectral viewer has also been recently developed in PRIDE that (i) enables visualization of multiple interpretations for the same spectrum; (ii) compares USI interpretations with predicted spectra; and (iii) provides additional metadata if SDRF is available for the correspondin dataset.

The PeptideAtlas team developed and deployed a new visualization tool and fragment ion annotation algorithm Quetzal [28], enabling more in-depth annotation of Higher-Energy Collisional Dissociation (HCD) spectra, including internal fragments, immonium ions, isotopic label-associated ions, and known contaminant low-mass ions. In addition to returning the spectrum for a given USI, the MassIVE team has also developed USI query tools to find other repository spectra and/or spectral library for the same peptide precursor (https://massive.ucsd.edu/ProteoSAFe/usi.jsp), and has also co-developed the FASST tool for repository-scale searches of spectra of the same or related (e.g. modified) variants of a given USI spectrum [(https://fasst.gnps2.org/fastsearch/, co-developed with the GNPS (Global Natural Products Social) team].

The ProteomeCentral USI system has been updated to include the ability to visualize predicted spectra from MS2PIP [29] alongside the spectra of specified USIs. Spectra are displayed via Lorikeet, but the peak-by-peak annotation display uses the Quetzal annotation engine. A version of Quetzal is also hosted at ProteomeCentral, allowing users to produce publication-quality PDFs and SVGs of annotated spectra via USIs or pasted-in peak lists.

Data submission statistics

Through the end of June 2025, a total of 64 330 datasets had been submitted to ProteomeXchange resources. Of those, 44 248 datasets (69%) were already publicly available, whereas the rest were still private or unreleased (20 082 datasets, 31%). The number of submitted datasets has increased every year, a trend that has not stopped yet (Fig. 1). In the past 3 years, 30 097 datasets have been submitted to ProteomeXchange resources, meaning that 47% of PXDs were submitted within just the past 36 months through June 2025. This again showcases the continuous significant increase of proteomics datasets in the public domain. During 2024 alone, a record number of 10 686 datasets were submitted to ProteomeXchange resources (890 datasets per month on average). During the first 6 months of 2025, the number of submissions was 6294 datasets (1049 per month on average).

In terms of distribution of datasets submitted across individual ProteomeXchange resources, 49 528 datasets (77%),

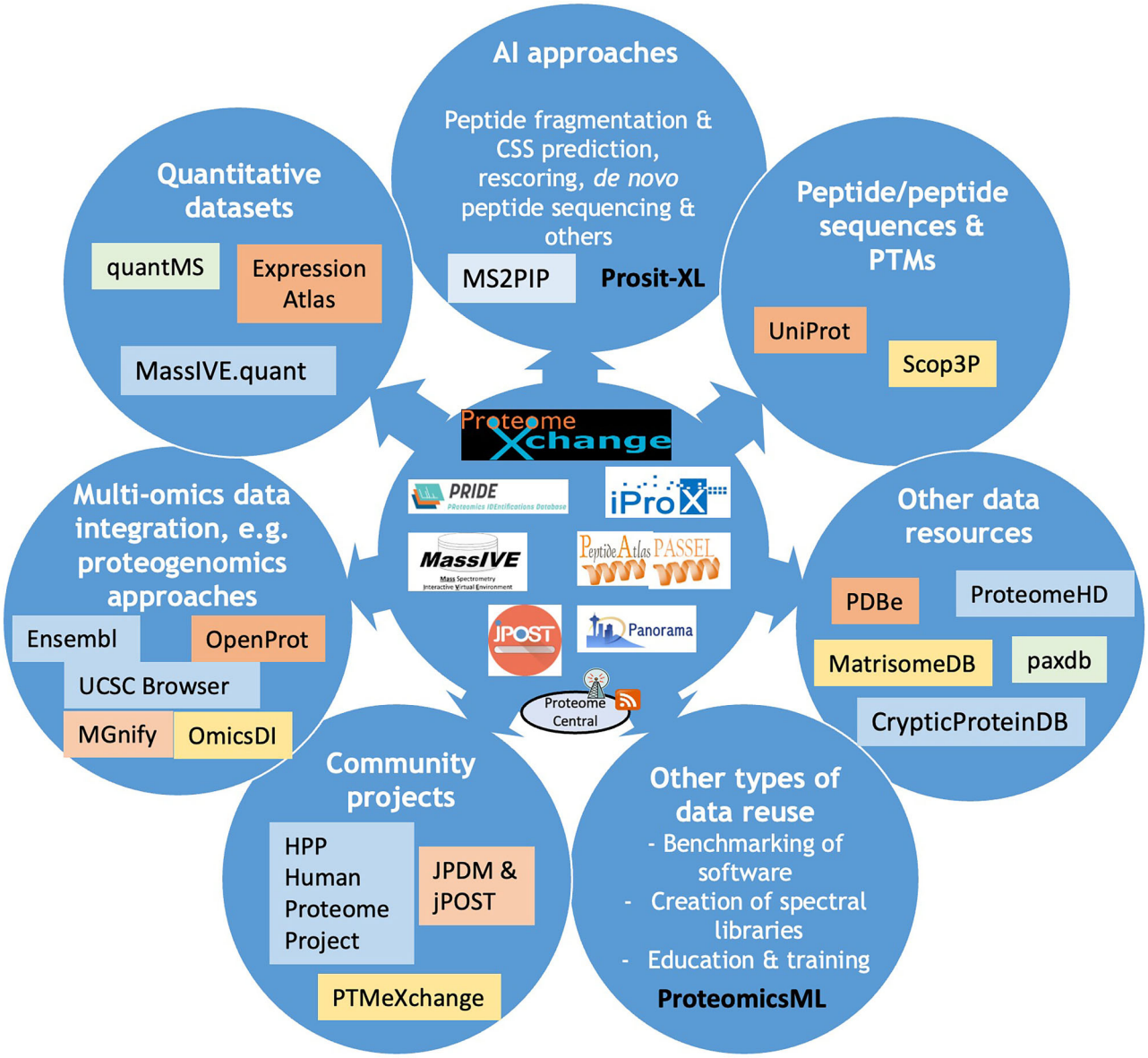


Figure 2. Overview figure including the current ProteomeXchange resources and the main efforts devoted to data reuse of public proteomics datasets. Different types of data reuse are listed and for each of them, the corresponding tools and/or data resources where these data can be accessed are indicated.

had been submitted to PRIDE, followed by iProX (6967 datasets, 11%), MassIVE (4770 datasets, 7.4%), jPOST (2443 datasets, 3.8%), Panorama Public (478 datasets, 0.7%), and PeptideAtlas/PASSEL (144 datasets, 0.2%). During the last 3 years, iProX has become the second resource in terms of submitted PXDs. As of September 2025, datasets came from >80 countries, demonstrating further the global reach of ProteomeXchange. The countries with the largest number of submitted datasets were the USA, Germany, China, UK, and France, in this particular order. See Supplementary File 1 for this detailed information, as available for PRIDE.

Trends in data reuse of public proteomics data

As mentioned above, reuse of public proteomics datasets continues to increase and diversify (Fig. 2). In this section, we highlight data reuse activities by the teams behind the ProteomeXchange data resources, involving different data types.

Peptide/protein sequence data, including post-translational modifications

ProteomeXchange resources regularly provide peptide and protein sequence data to UniProtKB (UniProt Knowledge Base) [30], the world's most used protein knowledgebase. In recent years, these efforts have been driven by the Human Proteome Organization Human Proteome Project, aiming to construct the human proteome blueprint [31]. Led by PeptideAtlas, and with the participation of MassIVE, the team has established the protein-level existence of gene products for 93.6% of the human proteome [32], following established community guidelines [33]. This has been the largest community data re-analysis project so far. In addition to the human proteome, PeptideAtlas has continued to create 'builds' corresponding

to the updated versions of proteomes of various model organisms, reanalysing the relevant public datasets. In recent years, PeptideAtlas has released and analysed the proteomes of *Arabidopsis* [34, 35], maize [36], *Candida albicans* [37], and rice [38].

PTMs are increasingly relevant for explaining protein function and behaviour. However, information about PTMs in databases such as UniProtKB is still limited, and often comes from the literature via manual curation. The PTMeXchange project (https://www.proteomexchange.org/ptmexchange/) was started to improve PTM data availability in UniProtKB, improving its FAIRness. Multiple PTM-enriched datasets from human and several model organisms have been consistently reanalysed, and the resulting high-quality data has been made available in UniProtKB, connecting it to the MS-based proteomics evidence in PeptideAtlas and PRIDE (using USIs). A methodology based on the use of decoy amino acids (mainly alanine) enables the reliable calculation of a false localization rate for phosphorylation [39], a methodology also now extended to other PTMs. The data reanalysis work is organized in 'builds' (related groups of datasets), which correspond to the analysis of one particular PTM in one given species. As of August 2025, the builds already finished and integrated in UniProtKB are phosphorylation for rice [40], *Plasmodium falciparum* [41], and the builds covering phosphorylation for other species such as human, mouse, and *Saccharomyces cerevisiae* are about to be integrated in UniProtKB. Additionally, human 'builds' from other PTMs are now also at different stages of integration into UniProtKB, including ubiquitination, SUMOylation, lysine acetylation and lysine methylation. In the context of PTM data, it is worth noting that Scop3P [42] is an additional resource that provides access to PTM reanalyses of ProteomeXchange PTM-enriched datasets.

Quantitative data reanalysis efforts

The number of reanalysed quantitative proteomics datasets has also increased significantly in recent years. Recognizing the importance of this kind of proteomics data, MassIVE designed and implemented the MassIVE.quant repository infrastructure and data resource [43] for reproducible quantitative MS-based proteomics. Using a branch design, MassIVE.quant stores raw experimental data, metadata of the experimental design, scripts of the quantitative analysis workflow, intermediate input and output files, as well as alternative reanalyses of the same dataset. In this same context, PRIDE has also reanalysed and integrated a wide range of datasets in the resource Expression Atlas [44], where it is possible to access both gene expression and protein abundance data. Most of the integrated datasets come from tissue samples generated in normal/baseline conditions using mainly DDA (Data Dependent Acquisition) but also DIA (Data Independent Acquisition) approaches, including human [45, 46], and model organisms such as mouse, rat [47], and domestic pig [48]. The initial study was performed using mainly cell lines and cancer tissue samples [49]. This reanalysis effort has been followed up by a recent study involving the reanalysis of 12 datasets to detect biomarkers of colorectal cancer [50]. For jPOST, to overcome the differences in highly diverse samples and varying measurement environments and perform quantitative analysis at the repository level, the strategy adopted was to unify data analysis methods for protein identification and to output the protein composition of each individual sample as quantitative values, thereby enabling comparative quantitative analysis between samples. For the former, UniScore was developed as an indicator to integrate and standardize the outputs of multiple search engines in the analysis of DDA data [51]. For the latter, an emPAI algorithm-based spectral count quantification method [52] was adopted, and the reanalysis database is being continually expanded.

MassIVE.quant

MassIVE.quant [43] (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp) is an extension of the MassIVE repository to provide the opportunity for large-scale deposition of data from quantitative MS-based proteomic experiments. MassIVE.quant is compatible with all MS data acquisition types and computational analysis tools. For each dataset, MassIVE.quant systematically stores the raw experimental data, the annotations of the experimental design, the scripts (or descriptions) of every step of the quantitative analysis workflow, and the intermediate input and output files. A branch structure enables MassIVE.quant to store and view alternative reanalyses of the same dataset with various combinations of methods and tools in a way that allows the user to inspect, reproduce or modify any component of the workflow, beginning with well-defined intermediate files. MassIVE.quant also supports infrastructure to fully automate analysis workflows, or to store, and to browse the intermediate results. As of today, MassIVE.quant includes 209 reanalyses of 105 datasets from a variety of species and across all major MS data types.

quantms

The quantms project (https://www.quantms.org) is building one of the most comprehensive resources of quantitative proteomics data by systematically reanalyzing public PXDs using an open-source, large-scale quantms pipeline [53]. To date, quantms has processed >103 large-scale human studies (https://quantms.org/datasets), comprising over 29 000 raw MS files from >13 000 samples, and has quantified upwards of 16 000 proteins across tissues, cell lines, and plasma [53]. These reanalyses not only recover more proteins than many of the original studies but also bring consistency and comparability across experiments that were initially produced with very different analytical methods. The resulting harmonized datasets provide a unified view of protein expression at scale, which is invaluable for identifying proteins expressed in specific tissues and conditions. A recent focus of the project is the generation of baseline expression profiles, both in controlled cell models and in clinically relevant samples such as plasma. By integrating iBAQ quantification through the ibaqpy package [54, 55], quantms has created high-resolution expression maps—for example, a corpus of >11 000 proteins quantified across nearly 5800 HeLa MS runs. In plasma, the project has established robust baseline proteomes that can be used to distinguish healthy variability from disease-specific signatures. The quantms workflow continues to be updated by the community (https://github.com/bigbio/quantms). Recent major releases include the support of DIANN 2.1.0, OpenMS 3.4, and major updates on different libraries like pmultiqc and sdrf-pipelines.

UniScore-emPAI quantitation

The reanalysis project in jPOST began with developing a methodology to minimize bias in database search engines

for protein identification. This involved re-scoring the results from multiple engines on a common scale, merging them, and then controlling the FDR using the target-decoy approach. By defining a very simple UniScore as the sum of the number of product ions matching the sequence with the number of amino acids flanked by two product ions, it became possible to identify more proteins than either individual search engines or combinations of multiple engines [51]. UniScore can be calculated using only product ion information matching sequences, significantly reducing computational resources compared to various existing re-scoring tools. It has already been applied to reanalyse jPOST global proteome and phosphoproteome data, simultaneously providing not only protein identification results but also quantitative data based on the emPAI algorithm [52].

Proteogenomics and multiomics data reuse

In this context, proteogenomics data reanalysis efforts should be highlighted first, also including immunopeptidomics and metaproteomics approaches. Public datasets can be reanalysed using sequence databases constructed by e.g. using genomics, transcriptomics or Ribo-seq data, among other DNA/RNA sequencing approaches. The original application of these proteogenomics approaches is to improve genome annotation efforts, by providing proteomics experimental evidence for some genomic events.

It has emerged in recent years that thousands of open reading frames (ORFs) beyond the coding DNA sequences of the core proteome (noncanonical ORFs or ncORFs) seem to undergo some degree of translation and can be detected via MS methods [56], although great care must be taken to avoid false positives when searching for these rare events [57]. Community efforts to assemble sequence databases of these ncORFs based on the highest quality Ribo-seq data are led by GENCODE and the TransCODE consortium [58, 59]. Efforts to document the highest quality evidence with USIs for ncORF detections is led by large-scale reprocessing of human datasets from ProteomeXchange by PeptideAtlas and the TransCODE consortium, finding that detections of ncORFs are quite rare and low abundance in ordinary protease digests without enrichment, but far more commonly detected in immunopeptidome enrichments [56].

Also in the context of noncanonical peptides, we have explored the identification of genome population variants (pangenome) in tissue proteomes [60], investigating the potential impact of pangenomes on future proteomics experiments. In the context of proteogenomics approaches involving microbiome data, a pilot study has been performed involving the integration of metaproteomics data available in PRIDE with the corresponding metagenomics and/or metatranscriptomics data coming from the same samples, in the resource MGnify [61].

All PXDs have been integrated in OmicsDI (Omics Discovery Index) (http://www.omicsdi.org) [62]. This portal facilitates the discoverability of multi-omics public datasets submitted to various public data resources, enabling the link, where possible, between proteomics datasets included in multi-omics studies to the corresponding other types of omics public datasets.

Reuse of datasets for AI approaches

Many of the most popular use cases involve the reuse of public datasets for ML or deep learning approaches (e.g. where public datasets are used for training and/or testing purposes), for many different applications, which are revolutionizing the field [63]. The applications include the prediction of peptide fragmentation, prediction of collision cross-section for ion mobility, improvements in algorithms for peptide and protein identification and quantification involving rescoring approaches (e.g. [64–67]), and for the development of *de novo* peptide sequencing algorithms [68], which are being tailored to different data and instrumentation types. For instance, consistently reanalysed public datasets have been used for developing the new peptide fragmentation models of MS2PIP (e.g. for tryptic and nontryptic peptides, and for immunopeptides [29]), Prosit-XL (for crosslinked peptides) [69], and for other peptide types such as glycopeptides [70].

In addition to integrating proteomics data in other resources, one of the main focusses of ProteomeXchange resources will be to continue to provide AI-ready data to the community. In addition to the availability of the mass spectra, this also involves the generation of high-quality reanalysed data (for different data types and approaches), which will need to be well annotated (in terms of experimental metadata), and also provided in suitable data formats.

Recognizing this need, MassIVE has released and continually updates the MassIVE-KB (MassIVE-Knowledge-Base) spectral library [7] constructed from over 1 billion publicly available tandem mass spectra. MassIVE-KB's latest public release contains reference spectra for 5948 126 peptides and has already enabled the development of multiple cutting-edge AI [71] and other data analysis tools [72]. In addition, the role of ProteomeXchange resources in enabling training and education in AI topics is also key. In that context, we have contributed to the development of the ProteomicsML resource (https://proteomicsml.org/), which provides ready-made datasets for ML models together with tutorials on how to work with them [73].

Creation of data resources

Datasets from ProteomeXchange resources are increasingly reused as the basis to create new resources. A few examples are: (i) MatrisomeDB [74], providing an updated view of the human and mouse extracellular matrix; (ii) ProteomeHD [75], a resource providing information about co-expressed proteins; (iii) systeMHC [76], providing a collection of reanalysed immunopeptidomics datasets; (iv) the proteogenomics resource OpenProt [77]; (v) CrypticProteinDB [78], a database of proteome and immunopeptidome derived noncanonical cancer proteins; and (vi) paxDb [79], including protein abundance data from human and several model organisms.

Additional highlights from the ProteomeXchange member databases

PRIDE has started to use large language models to provide extra functionality. A PRIDE chatbot (https://www.ebi.ac.uk/pride/chatbot/) is available, which has been trained on the PRIDE documentation for external users. The main objective is to help PRIDE users navigate PRIDE documentation, therefore decreasing the time required for the team to reply to ever-increasing support queries. Although the initial implementation was done using open source models [80], as of August 2025, the functionality is provided using the Gemini-1.5-pro model.

iProX has developed a new graphical interface tool iProXplorer to support the data submission of the iProX database (https://www.iprox.cn/page/DownloadClient.html). Specifically, iProXplorer presents a data format to describe experimental metadata and files, supports automated data submission, and exchanges experimental summary between different users. iProXplorer provides a user-friendly interface, as well as a command-line tool that can be integrated into analysis processed. iProXplorer, as an automated submission tool for the iProX database, will help facilitate the development of data sharing in the field of proteomics.

Discussion and future plans

We have highlighted the main overall developments in ProteomeXchange resources in the last 3 years. The resources will continue to evolve in parallel to the needs of the field. In the context of data archiving activities, PRIDE and MassIVE are currently extending their functionality for supporting controlled-access datasets, such as proteomics datasets generated from human samples collected under study-specific data access restrictions (i.e. datasets that cannot be made openly available), thus requiring repository-supported controlled access capabilities. The need of controlled access options for proteomics data (analogously to genomics/transcriptomics) is mainly due to three reasons: (i) these data could potentially be used to identify research participants; (ii) requirements related to patient consent; and/or (iii) due to personal data regulations like GDPR (General Data Protection Regulation) in Europe, HIPPA (Health Insurance Portability and Accountability Act) in the USA or any other relevant legislation [81].

There is an increasing number of sensitive human proteomics datasets that cannot be made available to the scientific community through a public-access ProteomeXchange resource. We recommend to users at present that if there are any potential legal issues of this type, they should submit their data to an alternative repository outside ProteomeXchange until ProteomeXchange resources can provide controlled access capabilities. However, existing controlled access resources developed for DNA/RNA sequencing data, e.g. the European Genome-Phenome Archive (EGA), the Japanese Genome-Phenome Archive (JGA) and dbGAP, even if they do already store a small number of proteomics datasets at present, are not ideal for proteomics datasets, because their data model cannot appropriately represent such datasets.

There is a growing popularity of non-MS-based proteomics technologies such as the use of affinity reagents (e.g. SomaLogic® and Olink® assays, among others), especially for human plasma datasets. PRIDE is currently starting a new section called PRIDE 'Affinity Proteomics' (https://www.ebi.ac.uk/pride/archive/affinity-proteomics) supporting these technologies. If there is a demand for it, the ProteomeXchange framework could be extended in the future to support non-MS data as well. However, it is important to highlight that a very large proportion of the studies coming from non-MS approaches are generated from human samples (and often from cohorts), and then the data may be considered sensitive so that controlled access mechanisms may apply. This is the case, for instance, of the non-MS proteomics datasets generated by UK Biobank, which are available via their dedicated data platform.

In addition, we plan to keep working extensively in data reuse/reanalysis projects, disseminating high-quality proteomics data into other bioinformatics resources, and making AI-ready data to the community. ProteomeXchange remains open to accept new members, provided that they adhere to the consortium requirements set out in the ProteomeXchange Membership Agreement, which was updated in 2024 (https://www.proteomexchange.org/pxcollaborativeagreement_2024.pdf).

Acknowledgements

The ProteomeXchange partners would like to thank all data submitters and collaborators for their contributions. In these last 3 years, PRIDE activities have been funded by EMBL core funding, Wellcome [grant number 223745/Z/21/Z], BBSRC [grant numbers BB/V018779/1, BB/S01781X/1, BB/X001911/1, BB/T019670/1, BB/Y513829/1), EPSRC [grant number EP/Y035984/1], European Commission H2020 program [grant number 823839], Open Targets (https://www.opentargets.org/) [grant number OTAR3091], the Luxembourg National Research Fund [grant number C19/BM/13684739], and several ELIXIR Implementation Studies. PeptideAtlas acknowledges support from the National Institutes of Health [grant numbers R01 GM087221, R24 GM127667], and the National Science Foundation [grant numbers DBI-1933311, IOS-1922871]. MassIVE activities are partially funded by grants from the National Institutes of Health [grant numbers R24GM148372, U24DK133658]. jPOST is supported by the Database Integration Coordination Program, operated by the National Bioscience Database Center (JST, Japan Science and Technology Agency) [grant numbers 15650519 (2015–2018), 18063028 (2018–2023) and JPMJND2304 (2023–2028)]. iProX has been supported by the Chinese National Infrastructure for Protein Science (Beijing), and National Key Research and Development Program [grant numbers 2021YFA1301603, 2024YFE0202700]. Panorama Public is funded by the National Institutes of Health [grant number R24 GM141156], the Panorama Partners Program (https://panoramaweb.org/partners.url), and by the University of Washington's Proteomics Resource [grant number UWPR95794].

Author contributions: Eric W.Deutsch (Conceptualization, Project administration, Writing - Original draft), Nuno Bandeira (Project administration, Writing), Yasset Perez-Riverol (Writing), Michael J. MacCoss (Project administration, Writing), Yunping Zhu (Project administration, Writing), Yasushi Ishihama (Project administration, Writing), Juan Antonio Vizcaino (Conceptualization [equal], Project administration [equal], Writing – original draft [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Wellcome Trust (223745/Z/21/Z); National Key Research and Development Program of China (2021YFA1301603 and 2024YFE0202700); Biotechnology and Biological Sci-

ences Research Council (BB/S01781X/1, BB/T019670/1, BB/V018779/1, BB/X001911/1, and BB/Y513829/1); Japan Science and Technology Agency (15650519, 18063028, and JPMJND2304); Fonds National de la Recherche Luxembourg (C19/BM/13684739); ELIXIR; University of Washington's Proteomics Resource (UWPR95794); Chinese National Infrastructure for Protein Science; Open Targets (OTAR3091); Panorama Partners Program; National Institutes of Health (R01 GM087221, R24 GM127667, R24 GM141156, R24GM148372, and U24DK133658); Engineering and Physical Sciences Research Council (EP/Y035984/1). Funding to pay the Open Access publication charges for this article was provided by Wellcome.

Data availability

The ProteomeXchange consortium of proteomics resources is freely available to all at http://www.proteomexchange.org.

References

1. Mann M, Kumar C, Zeng WF *et al.* Artificial intelligence for proteomics and biomarker discovery. *Cell Systems* 2021;12:759–70. https://doi.org/10.1016/j.cels.2021.06.006
2. Vizcaino JA, Deutsch EW, Wang R *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014;32:223–6.
3. Deutsch EW, Csordas A, Sun Z *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017;45:D1100–6. https://doi.org/10.1093/nar/gkw936
4. Deutsch EW, Bandeira N, Sharma V *et al.* The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res* 2020;48:D1145–52.
5. Deutsch EW, Bandeira N, Perez-Riverol Y *et al.* The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res* 2023;51:D1539–48. https://doi.org/10.1093/nar/gkac1040
6. Perez-Riverol Y, Bandla C, Kundu DJ *et al.* The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res* 2025;53:D543–53. https://doi.org/10.1093/nar/gkae1011
7. Wang M, Wang J, Carver J *et al.* Assembling the community-scale discoverable human proteome. *Cell Syst* 2018;7:412–21.
8. Okuda S, Yoshizawa AC, Kobayashi D *et al.* jPOST environment accelerates the reuse and reanalysis of public proteome mass spectrometry data. *Nucleic Acids Res* 2025;53:D462–7. https://doi.org/10.1093/nar/gkae1032
9. Chen T, Ma J, Liu Y *et al.* iProX in 2021: connecting proteomics data sharing with big data. *Nucleic Acids Res* 2022;50:D1522–7. https://doi.org/10.1093/nar/gkab1081
10. Sharma V, Eckels J, Schilling B *et al.* Panorama public: a public repository for quantitative data sets processed in skyline. *Mol Cell Proteom* 2018;17:1239–44. https://doi.org/10.1074/mcp.RA117.000543
11. Farrah T, Deutsch EW, Kreisberg R *et al.* PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 2012;12:1170–5. https://doi.org/10.1002/pmic.201100515
12. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
13. Deutsch EW, Vizcaino JA, Jones AR *et al.* Proteomics standards initiative at twenty years: current activities and future work. *J Proteome Res* 2023;22:287–301. https://doi.org/10.1021/acs.jproteome.2c00637
14. Deutsch EW, Orchard S, Binz PA *et al.* Proteomics standards initiative: fifteen years of progress and future work. *J Proteome Res* 2017;16:4288–98. https://doi.org/10.1021/acs.jproteome.7b00370
15. Deutsch EW, Perez-Riverol Y, Carver J *et al.* Universal spectrum identifier for mass spectra. *Nat Methods* 2021;18:768–70. https://doi.org/10.1038/s41592-021-01184-6
16. LeDuc RD, Deutsch EW, Binz PA *et al.* Proteomics standards initiative's ProForma 2.0: unifying the encoding of proteoforms and peptidoforms. *J Proteome Res* 2022;21:1189–95. https://doi.org/10.1021/acs.jproteome.1c00771
17. Mayer G, Montecchi-Palazzi L, Ovelleiro D *et al.* The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database* 2013;2013:bat009. https://doi.org/10.1093/database/bat009
18. Drysdale R, Cook CE, Petryszak R *et al.* The ELIXIR core data resources: fundamental infrastructure for the life sciences. *Bioinformatics* 2020;36:2636–42. https://doi.org/10.1093/bioinformatics/btz959
19. Watanabe Y, Aoki-Kinoshita KF, Ishihama Y *et al.* GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. *Nucleic Acids Res* 2021;49:D1523–8. https://doi.org/10.1093/nar/gkaa1012
20. Martens L, Chambers M, Sturm M *et al.* mzML–a community standard for mass spectrometry data. *Mol Cell Proteom* 2011;10:R110.000133. https://doi.org/10.1074/mcp.R110.000133
21. Vizcaino JA, Mayer G, Perkins S *et al.* The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Mol Cell Proteom* 2017;16:1275–85. https://doi.org/10.1074/mcp.M117.068429
22. Griss J, Jones AR, Sachsenberg T *et al.* The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteom* 2014;13:2765–75. https://doi.org/10.1074/mcp.O113.036681
23. Dai C, Fullgrabe A, Pfeuffer J *et al.* A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat Commun* 2021;12:5854. https://doi.org/10.1038/s41467-021-26111-3
24. Claeys T, Van Den Bossche T, Perez-Riverol Y *et al.* lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation. *Nat Commun* 2023;14:6743. https://doi.org/10.1038/s41467-023-42543-5
25. Viegener W, Urazbakhtin S, Ferretti D *et al.* Facilitating analysis and dissemination of proteomics data through metadata integration in MaxQuant. *Nat Commun* 2025;16:8421. https://doi.org/10.1038/s41467-025-64089-4
26. Wu P-S, Lin MH, Hsiao J-C *et al.* Dataset for EGFR-mutant interactome in tyrosine kinase inhibitor sensitive/resistant NSCLC cells. *J Proteome Data Methods* 2025;7. https://doi.org/10.14889/jpdm.2025.0009
27. Hulstaert N, Shofstahl J, Sachsenberg T *et al.* ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. *J Proteome Res* 2020;19:537–42. https://doi.org/10.1021/acs.jproteome.9b00328
28. Deutsch EW, Mendoza L, Moritz RL. Quetzal: comprehensive peptide fragmentation annotation and visualization. *J Proteome Res* 2025;24:2196–204. https://doi.org/10.1021/acs.jproteome.5c00092
29. Declercq A, Bouwmeester R, Chiva C *et al.* Updated MS(2)PIP web server supports cutting-edge proteomics applications. *Nucleic Acids Res* 2023;51:W338–42.
30. UniProt, C. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31. https://doi.org/10.1093/nar/gkac1052
31. Adhikari S, Nice EC, Deutsch EW *et al.* A high-stringency blueprint of the human proteome. *Nat Commun* 2020;11:5301. https://doi.org/10.1038/s41467-020-19045-9
32. Omenn GS, Orchard S, Lane L *et al.* The 2024 report on the human proteome from the HUPO human proteome project. *J*

Proteome Res 2024;23:5296–311. https://doi.org/10.1021/acs.jproteome.4c00776
33. Deutsch EW, Lane L, Overall CM *et al.* Human proteome project mass spectrometry data interpretation guidelines 3.0. *J Proteome Res* 2019;18:4108–16. https://doi.org/10.1021/acs.jproteome.9b00542
34. van Wijk KJ, Leppert T, Sun Q *et al.* The arabidopsis PeptideAtlas: harnessing worldwide proteomics data to create a comprehensive community proteomics resource. *Plant Cell* 2021;33:3421–53. https://doi.org/10.1093/plcell/koab211
35. van Wijk KJ, Leppert T, Sun Z *et al.* Detection of the arabidopsis proteome and its post-translational modifications and the nature of the unobserved (Dark) proteome in PeptideAtlas. *J Proteome Res* 2024:23:185–214. https://doi.org/10.1021/acs.jproteome.3c00536
36. van Wijk KJ, Leppert T, Sun Z *et al.* The Zea mays PeptideAtlas: a new maize community resource. *J Proteome Res* 2024;23:3984–4004. https://doi.org/10.1021/acs.jproteome.4c00320
37. Gomez-Artiguez L, de la Camara-Fuentes S, Sun Z *et al.* Candida albicans: a comprehensive view of the proteome. *J Proteome Res* 2025;24:1636–48. https://doi.org/10.1021/acs.jproteome.4c01020
38. Contreras-Moreira B, Sharma E, Saraf S *et al.* A pan-gene catalogue of Asian cultivated rice. bioRxiv, https://doi.org/10.1101/2025.02.17.638606, 23 February 2025, preprint: not peer reviewed.
39. Ramsbottom KA, Prakash A, Riverol YP *et al.* Method for independent estimation of the false localization rate for phosphoproteomics. *J Proteome Res* 2022;21:1603–15. https://doi.org/10.1021/acs.jproteome.1c00827
40. Ramsbottom KA, Prakash A, Perez-Riverol Y *et al.* Meta-analysis of rice phosphoproteomics data to understand variation in cell signaling across the rice pan-genome. *J Proteome Res* 2024;23:2518–31. https://doi.org/10.1021/acs.jproteome.4c00187
41. Camacho OJM, Ramsbottom KA, Prakash A *et al.* Phosphorylation in the *Plasmodium falciparum* proteome: a meta-analysis of publicly available data sets. *J Proteome Res* 2024;23:5326–41. https://doi.org/10.1021/acs.jproteome.4c00418
42. Ramasamy P, Turan D, Tichshenko N *et al.* Scop3P: a comprehensive resource of human phosphosites within their full context. *J Proteome Res* 2020;19:3478–86. https://doi.org/10.1021/acs.jproteome.0c00306
43. Choi M, Carver J, Chiva C *et al.* MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat Methods* 2020;17:981–4. https://doi.org/10.1038/s41592-020-0955-0
44. George N, Fexova S, Fuentes AM *et al.* Expression Atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Res* 2024;52:D107–14. https://doi.org/10.1093/nar/gkad1021
45. Prakash A, Garcia-Seisdedos D, Wang S *et al.* Integrated view of baseline protein expression in human tissues. *J Proteome Res* 2023;22:729–42. https://doi.org/10.1021/acs.jproteome.2c00406
46. Prakash A, Collins A, Vilmovsky L *et al.* Integrated view of baseline protein expression in human tissues using public data independent acquisition data sets. *J Proteome Res* 2025;24:685–95. https://doi.org/10.1021/acs.jproteome.4c00788
47. Wang S, Garcia-Seisdedos D, Prakash A *et al.* Integrated view and comparative analysis of baseline protein expression in mouse and rat tissues. *PLoS Comput Biol* 2022;18:e1010174. https://doi.org/10.1371/journal.pcbi.1010174
48. Wang S, Collins A, Prakash A *et al.* Integrated proteomics analysis of baseline protein expression in Pig tissues. *J Proteome Res* 2024;23:1948–59.
49. Jarnuczak AF, Najgebauer H, Barzine M *et al.* An integrated landscape of protein expression in human cancer. *Sci Data* 2021;8:115. https://doi.org/10.1038/s41597-021-00890-2
50. Robles J, Prakash A, Vizcaino JA *et al.* Integrated meta-analysis of colorectal cancer public proteomic datasets for biomarker discovery and validation. *PLoS Comput Biol* 2024;20:e1011828. https://doi.org/10.1371/journal.pcbi.1011828
51. Tabata T, Yoshizawa AC, Ogata K *et al.* UniScore, a unified and universal measure for peptide identification by multiple search engines. *Mol Cell Proteom* 2025;24:101010. https://doi.org/10.1016/j.mcpro.2025.101010
52. Ishihama Y, Oda Y, Tabata T *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteom* 2005;4:1265–72. https://doi.org/10.1074/mcp.M500061-MCP200
53. Dai C, Pfeuffer J, Wang H *et al.* quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nat Methods* 2024;21,1603–7. https://doi.org/10.1038/s41592-024-02343-1
54. Zheng P, Audain E, Webel H *et al.* Ibaqpy: a scalable Python package for baseline quantification in proteomics leveraging SDRF metadata. *J Proteomics* 2025;317:105440. https://doi.org/10.1016/j.jprot.2025.105440
55. Wang H, Dai C, Pfeuffer J *et al.* Tissue-based absolute quantification using large-scale TMT and LFQ experiments. *Proteomics* 2023;23:e2300188. https://doi.org/10.1002/pmic.202300188
56. Deutsch EW, Kok LW, Mudge JM *et al.* High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. bioRxiv, https://doi.org/10.1101/2024.09.09.612016, 24 July 2025, preprint: not peer reviewed.
57. Wacholder A, Deutsch EW, Kok LW *et al.* Detection of human unannotated microproteins by mass spectrometry-based proteomics: a community assessment. bioRxiv, https://doi.org/10.1101/2025.02.19.639069, 24 July 2025, preprint: not peer reviewed.
58. Mudge JM, Ruiz-Orera J, Prensner JR *et al.* Standardized annotation of translated open reading frames. *Nat Biotechnol* 2022;40:994–9. https://doi.org/10.1038/s41587-022-01369-0
59. Chothani S, Ruiz-Orera J, Tierney JAS *et al.* An expanded reference catalog of translated open reading frames for biomedical research. bioRxiv, https://doi.org/10.1101/2025.07.03.662928, 7 July 2025, preprint: not peer reviewed.
60. Wang D, Bouwmeester R, Zheng P *et al.* Proteogenomics analysis of human tissues using pangenomes. bioRxiv, https://doi.org/10.1101/2024.05.24.595489, 28 May 2024, preprint: not peer reviewed.
61. Richardson L, Allen B, Baldi G *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;51:D753–9. https://doi.org/10.1093/nar/gkac1080
62. Perez-Riverol Y, Bai M, da Veiga Leprevost F *et al.* Discovering and linking public omics data sets using the omics discovery index. *Nat Biotechnol* 2017;35:406–9.
63. Neely BA, Dorfer V, Martens L *et al.* Toward an integrated machine learning model of a proteomics experiment. *J Proteome Res* 2023;22:681–96. https://doi.org/10.1021/acs.jproteome.2c00711
64. Chang CH, Yeung D, Spicer V *et al.* Sequence-specific model for predicting peptide collision cross section values in proteomic ion mobility spectrometry. *J Proteome Res* 2021;20:3600–10. https://doi.org/10.1021/acs.jproteome.1c00185
65. Nakai-Kasai A, Ogata K, Ishihama Y *et al.* Leveraging pretrained deep protein language model to predict peptide collision cross section. *Commun Chem* 2025;8:137. https://doi.org/10.1038/s42004-025-01540-z
66. Willems P, Thery F, Van Moortel L *et al.* Maximizing immunopeptidomics-based bacterial epitope discovery by multiple search engines and rescoring. *J Proteome Res* 2025;24:2141–51. https://doi.org/10.1021/acs.jproteome.4c00864
67. Siraj A, Bouwmeester R, Declercq A *et al.* Intensity and retention time prediction improves the rescoring of protein-nucleic acid

cross-links. *Proteomics* 2024;24:e2300144. https://doi.org/10.1002/pmic.202300144
68. Bittremieux W, Ananth V, Fondrie WE *et al.* Deep learning methods for *de novo* peptide sequencing. *Mass Spectrometry Reviews* 2024. https://doi.org/10.1002/mas.21919
69. Kalhor M, Saylan CC, Picciani M *et al.* Prosit-XL: enhanced cross-linked peptide identification by fragment intensity prediction to study protein interactions and structures. *Nat Commun* 2025;16:5429. https://doi.org/10.1038/s41467-025-61203-4
70. Yang Y, Fang Q. Prediction of glycopeptide fragment mass spectra by deep learning. *Nat Commun* 2024;15:2448. https://doi.org/10.1038/s41467-024-46771-1
71. Yilmaz M, Fondrie WE, Bittremieux W *et al.* Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat Commun* 2024;15:6427. https://doi.org/10.1038/s41467-024-49731-x
72. Kang J, Xu W, Bittremieux W *et al.* Accelerating open modification spectral library searching on tensor core in high-dimensional space. *Bioinformatics* 2023;39:btad404. https://doi.org/10.1093/bioinformatics/btad404
73. Rehfeldt TG, Gabriels R, Bouwmeester R *et al.* ProteomicsML: an online platform for community-curated data sets and tutorials for machine learning in proteomics. *J Proteome Res* 2023;22:632–6. https://doi.org/10.1021/acs.jproteome.2c00629
74. Shao X, Gomez CD, Kapoor N *et al.* MatrisomeDB 2.0: 2023 updates to the ECM-protein knowledge database. *Nucleic Acids Res* 2023;51:D1519–30.
75. Fischer SN, Claussen ER, Kourtis S *et al.* hu.MAP3.0: atlas of human protein complexes by integration of >25,000 proteomic experiments. *Mol Syst Biol* 2025;21:911–43.
76. Huang X, Gan Z, Cui H *et al.* The SysteMHC Atlas v2.0, an updated resource for mass spectrometry-based immunopeptidomics. *Nucleic Acids Res* 2024;52:D1062–71.
77. Leblanc S, Yala F, Provencher N *et al.* OpenProt 2.0 builds a path to the functional characterization of alternative proteins. *Nucleic Acids Res* 2024;52:D522–8. https://doi.org/10.1093/nar/gkad1050
78. Othoum G, Maher CA. CrypticProteinDB: an integrated database of proteome and immunopeptidome derived non-canonical cancer proteins. *NAR Cancer* 2023;5:zcad024. https://doi.org/10.1093/narcan/zcad024
79. Huang Q, Szklarczyk D, Wang M *et al.* PaxDb 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts. *Mol Cell Proteom* 2023;22:100640. https://doi.org/10.1016/j.mcpro.2023.100640
80. Bai J, Kamatchinathan S, Kundu DJ *et al.* Open-source large language models in action: a bioinformatics chatbot for PRIDE database. *Proteomics* 2024;24:e2400005. https://doi.org/10.1002/pmic.202400005
81. Bandeira N, Deutsch EW, Kohlbacher O *et al.* Data management of sensitive human proteomics data: current practices, recommendations and perspectives for the future. *Mol Cell Proteom* 2021;20: 100071. https://doi.org/10.1016/j.mcpro.2021.100071

Received: September 17, 2025. **Revised:** October 9, 2025. **Accepted:** October 9, 2025