Bioinformatics Advances, 2025, **00**, vbaf301
https://doi.org/10.1093/bioadv/vbaf301
Advance Access Publication Date: 17 December 2025
Perspective

OXFORD

Proteomics and metabolomics

Perspectives in computational mass spectrometry: recent developments and key challenges

Timo Sachsenberg[1,2,†], Lindsay K. Pino[3,†], Marie Brunet[4,†], Isabell Bludau[5,†], Oliver Kohlbacher[1,2,6], Juan Antonio Vizcaino[7], Wout Bittremieux[8,*]

[1]Department of Computer Science, University of Tübingen, Tübingen, 72074, Germany
[2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72074, Germany
[3]Talus Bio, Seattle, WA 98122, United States
[4]Pediatrics Department, University of Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada
[5]Department of Neuropathology, Institute of Pathology, University Hospital Heidelberg, Heidelberg, 69120, Germany
[6]Translational Bioinformatics, University Hospital Tübingen, Tübingen, 72074, Germany
[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, United Kingdom
[8]Department of Computer Science, University of Antwerp, Antwerp, 2020, Belgium

*Corresponding author. Department of Computer Science, University of Antwerp, Middelheimlaan 1, Antwerp, 2020, Belgium.
E-mail: wout.bittremieux@uantwerpen.be.

[†]= equal contribution.

Associate Editor: Aida Ouangraoua

Abstract

Summary: Mass spectrometry (MS) is a cornerstone technology in modern molecular biology, powering diverse applications across proteomics, metabolomics, lipidomics, glycomics, and beyond. As the field continues to evolve, rapid advancements in instrumentation, acquisition strategies, machine learning, and scalable computing have reshaped the landscape of computational MS. This perspective reviews recent developments and highlights key challenges, including data harmonization, statistical confidence estimation, repository-scale analysis, multi-omics integration, and privacy in clinical MS. We also discuss the increasing importance of machine learning and the need to build corresponding literacy within the community. Finally, we reflect on the role of the Computational Mass Spectrometry (CompMS) Community of Special Interest of the International Society for Computational Biology in supporting collaboration, innovation, and knowledge exchange. With MS-based technologies now central to both basic and translational research, continued investment in robust and reproducible computational methods will be essential to realize their full potential.

1 Introduction

Mass spectrometry (MS) is a central analytical technology in molecular biology, widely used to characterize the molecular composition of complex biological systems. Its capacity to measure the mass-to-charge ratio of ions with high sensitivity underpins its widespread adoption across diverse omics disciplines—including proteomics, metabolomics, lipidomics, and glycomics—to investigate phenotypes in both health and disease.

Ongoing advances in MS instrumentation and computational methods have significantly expanded the scale and resolution of MS-based experiments. These developments have enabled emerging applications such as single-cell proteomics, spatially resolved molecular profiling, and high-throughput clinical studies. In parallel, computational MS has evolved to address increasingly complex tasks, including real-time data interpretation, machine learning (ML)-driven molecular property prediction, and large-scale integration across datasets.

However, the growing volume and complexity of MS data also introduce new challenges. Scalable computational infrastructure is needed to support high-throughput workflows. Data harmonization across instruments, batches, and modalities remains difficult, especially when integrating multi-omics data. While proteomics has developed advanced methods for statistical confidence estimation, challenges persist for non-standardized applications. In contrast, metabolomics often lacks similar mature statistical approaches, making confidence assessment even more difficult in this domain. Additionally, emerging concerns about data privacy in clinical applications, alongside the need for robust metadata standards and FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson *et al.* 2016) compliance, further underscore the need to address infrastructural and ethical dimensions.

In this perspective, we provide an overview of major recent developments in computational MS, outline key challenges facing the field, and discuss emerging opportunities for continued innovation. We also highlight the role of the Computational Mass Spectrometry (CompMS) Community of Special Interest of the International Society for Computational Biology (ISCB) in fostering collaboration, innovation, and community building. As MS continues to support increasingly diverse and complex omics applications, the

Received: July 29, 2025; **Revised:** October 27, 2025; **Accepted:** November 19, 2025

development of robust, scalable, and transparent computational methods will be essential to ensure accurate data interpretation and meaningful biological insight.

2 Recent developments in computational mass spectrometry

2.1 Emergence of novel mass spectrometry instrumentation

Recent advances in MS instrumentation have substantially improved sensitivity, speed, and analytical depth across a range of applications. Instruments such as Thermo Fisher Scientific's Astral (Heil *et al.* 2023), Sciex's ZenoTOF (Wang *et al.* 2022), and Bruker's timsTOF (Demichev *et al.* 2022) now support high-throughput workflows capable of generating tens of gigabytes of data per instrument hour. While these capabilities enable deeper interrogation of biological systems, they also present considerable challenges in data handling, necessitating scalable solutions for storage, processing, and downstream analysis.

To manage increasing data volumes, the development of computational tools optimized for efficiency has become essential. In proteomics, software such as MSFragger (Kong *et al.* 2017) and Sage (Lazear 2023), and in metabolomics, tools like Clustering+/Networking+ (Mongia *et al.* 2024) and mzMine (Schmid *et al.* 2023), exemplify recent efforts to support high-throughput analysis at scale, reflecting the broader trend toward improved performance and scalability in computational MS. However, meeting the demands of modern instrumentation requires more than faster algorithms; it calls for rethinking entire analysis workflows to ensure that they can accommodate heterogeneous data structures and acquisition modes while maintaining reproducibility and interoperability.

A major enabler of such interoperability has been the vendor-agnostic mzML format (Martens *et al.* 2011), which provides a universal interface between raw instrument data and downstream computational tools. This openness allows researchers to apply shared analysis pipelines regardless of the instrument manufacturer, supporting reproducibility and the development of cross-platform software ecosystems. Yet, rapid technological innovation continually tests the flexibility of this standard. For example, the growing use of ion mobility spectrometry adds an additional analytical dimension that enhances molecular separation but greatly increases data volume and complexity. Likewise, the expanding range of fragmentation techniques, including electron-transfer dissociation and ultraviolet photodissociation in addition to conventional collision-induced and higher-energy collisional dissociation, requires software to adapt to distinct fragmentation behaviors and scoring models. Together, these advances create a moving target for developers, who must continually update algorithms and data models to remain compatible with emerging instrument capabilities.

At the same time, the efficiency of underlying data storage formats has become a growing concern. The mzML format remains the primary open format for MS data exchange, but its XML-based structure can lead to file sizes an order of magnitude larger than the corresponding vendor raw files. Moreover, mzML is not natively compatible with modern data science frameworks, requiring specialized parsers for access and manipulation. To address these limitations, the community is exploring more compact and computation-friendly alternatives, such as the proposed "mzPeak" format (Van Den Bossche *et al.* 2025), which leverages a columnar Parquet-based design to improve compression, query performance, and interoperability with contemporary data science tools.

Beyond computation, the growing size of raw MS data and experiment sample size imposes logistical and financial pressure on infrastructure. Laboratories must find cost-effective strategies for data storage, transfer, and archival, particularly as datasets accumulate over time. Approaches such as spectrum clustering to reduce redundancy (Griss *et al.* 2016) and numerical compression techniques (Teleman *et al.* 2014) have shown promise in limiting storage overhead while preserving analytical utility. Continued investment in such methods is essential to ensure that advances in instrumentation remain accessible and sustainable across diverse research settings.

2.2 Advances in mass spectrometry data acquisition

Recent developments in MS data acquisition strategies are expanding the types of biological questions that can be addressed, while also placing new demands on computational analysis. One major trend is the emergence of techniques for low-input and single-cell proteomics, which enable molecular profiling at the level of individual cells (Bennett *et al.* 2023). These approaches rely on sensitive instrumentation and require computational frameworks capable of addressing the variability and sparsity inherent to single-cell data. As single-cell MS becomes more widely adopted, bioinformatics tools must evolve to support reproducible and statistically robust analyses at this resolution.

In parallel, spatial approaches, including MS imaging, are gaining traction in both proteomics (Method of the Year 2024: spatial proteomics 2024), metabolomics (Wadie *et al.* 2024), and lipidomics. These spatial techniques contribute crucial insights into the localization and abundance of molecules within biological tissues, enhancing our understanding of molecular organization within complex biological structures. Recent developments such as deep visual proteomics (Mund *et al.* 2022) have demonstrated the potential of spatially resolved MS for clinical research, enabling molecular characterization of tissue architecture at near-cellular resolution and revealing disease-associated spatial phenotypes. A related and promising direction lies in integrating MS-based measurements with other imaging modalities, for example combining protein–protein interaction data from BioPlex (Huttlin *et al.* 2015, 2017) with fluorescence-based localization data from the Human Protein Atlas (Uhlen *et al.* 2010) to uncover new molecular relationships within subcellular systems (Qin *et al.* 2021). Collectively, these multimodal and spatially informed strategies hold considerable promise for advancing systems-level biology. However, the need to analyze large, multidimensional datasets calls for robust computational solutions that can handle both the volume and specificity of spatial data. As spatial resolution and data complexity continue to increase, scalable and domain-specific bioinformatics tools are needed to manage data volume, integrate data across modalities, and extract biologically meaningful spatial patterns. Developing such frameworks, capable of reconciling heterogeneous data types and scales, remains a central open challenge.

The increasing mass range of MS instruments has enabled the data acquisition of intact proteins without prior digestion

[top-down proteomics (Roberts *et al.* 2024)]. The resulting data exhibit different characteristics compared to classical bottom-up proteomics and offer more comprehensive insights into individual proteoforms (Marx 2024). Fully interpreting these data requires specialized algorithms and software capable of deconvolving complex isotope distributions and charge states in top-down spectra, as well as accurate identification and quantification of proteoforms (Kou *et al.* 2016, Jeong *et al.* 2020). Moreover, representing the resulting proteoform-level information necessitates standardized formats that capture this added molecular complexity, such as the ProForma notation (LeDuc *et al.* 2022).

Another trend in MS data acquisition is the emphasis on high-throughput acquisition. Strategies combining short chromatographic gradients with fast scanning instruments enable the analysis of large sample cohorts, including clinical and pharmacological studies (Messner *et al.* 2020). While this facilitates more comprehensive experimental designs, it also shifts the computational bottleneck downstream, requiring highly automated and efficient data processing pipelines to keep pace with acquisition.

In addition, specialized techniques such as real-time searching (Schweppe *et al.* 2020, Bills *et al.* 2022) and advanced data acquisition methods (Mendes *et al.* 2024) are being adopted to increase experimental robustness and depth. Real-time search engines provide immediate feedback during acquisition, while complex data-independent (Mendes *et al.* 2024) and iterative acquisition (Zuo *et al.* 2021) approaches aim to reduce sampling bias and improve reproducibility.

Overall, the increasing diversity and complexity of MS acquisition strategies call for equally flexible computational tools. Although existing software often provides a starting point, data generated by cutting-edge methods frequently necessitates the development of new algorithms and data models tailored to their specific characteristics. Addressing these needs will be key to ensuring that technological advances in MS acquisition are matched by reliable and scalable computational interpretation.

2.3 Demand for enhanced computational resources

The growing scale and complexity of MS experiments are placing increasing demands on computational infrastructure. Traditionally, many laboratories have relied on local compute workstations or institutional high-performance computing (HPC) clusters for data analysis. However, with modern instruments producing increasingly large and information-rich datasets, these resources are often pushed to their limits. The need for scalable solutions spans not just data processing, but also storage, retrieval, and integration across multiple projects and platforms.

Scalability in computational MS encompasses three interrelated dimensions: algorithmic, systems, and operational scalability. First, algorithmic scalability refers to the efficiency of the computational methods themselves. Tools must be designed to handle rapidly growing data volumes through more efficient algorithms and parallelization strategies. Recent progress in advanced fragment-ion indexing approaches (Kong *et al.* 2017, Lazear 2023, Mongia *et al.* 2024, Li and Fiehn 2023) exemplifies how algorithmic innovation can drastically accelerate spectral matching and reduce computational overhead. Continued development of such methods will be essential to ensure that gains in instrumentation throughput are matched by corresponding improvements in computational performance.

Second, systems scalability involves the ability of computational infrastructure to accommodate expanding workloads and data sizes. Cloud-based computing environments have emerged as a flexible complement to local infrastructure (Neely 2021). These platforms offer access to specialized hardware, such as graphics processing units and high-memory nodes, that are well-suited for intensive tasks like ML, large-scale (re)analysis, and complex statistical modeling. Their scalability allows users to allocate resources dynamically based on project needs, and the pay-as-you-go model can help manage costs. Increasingly, cloud solutions are becoming a key enabler of reproducible, high-throughput MS workflows that can scale seamlessly from individual projects to community-wide data reanalyses.

Third, operational scalability concerns the reproducibility, portability, and maintainability of computational workflows across different environments and laboratories. Achieving this requires transparent, containerized, and workflow-managed solutions, for example built on Nextflow (Di Tommaso *et al.* 2017), that allow analyses to be executed consistently regardless of local infrastructure. This form of scalability ensures that pipelines developed in one laboratory can be deployed in another without modification, facilitating collaboration and cross-institutional benchmarking. Operational scalability thus bridges algorithmic advances and systems capacity, enabling reproducible science at scale.

ML plays an increasingly important role across many stages of MS data analysis. In proteomics, ML models are now widely used to predict peptide properties such as retention time (Gessulat *et al.* 2019, Bouwmeester *et al.* 2021, Zeng *et al.* 2022), ion mobility (Zeng *et al.* 2022, Devreese *et al.* 2025), and fragment ion intensities (Degroeve and Martens 2013, Gessulat *et al.* 2019, Zeng *et al.* 2022), increasing sensitivity by enhancing downstream tasks like spectrum matching and rescoring (Kalhor *et al.* 2024). Tools such as Percolator (Käll *et al.* 2007) integrate these predictions to improve peptide–spectrum match confidence. ML is also advancing *de novo* peptide sequencing (Bittremieux *et al.* 2024), with recent models achieving substantial improvements over traditional algorithms. By automating aspects of data analysis and improving the interpretability of MS results, ML has become a key asset in modern proteomics.

In metabolomics, ML applications are emerging but are not yet as mature. The chemical diversity and limited ground truth annotations in metabolomics create unique challenges. Nonetheless, progress is being made in tasks such as spectrum modeling (Wang *et al.* 2021, Young *et al.* 2024), analog detection (Huber *et al.* 2021), *in silico* database searching (Dührkop *et al.* 2019, Goldman *et al.* 2023), and *de novo* annotation of unknown compounds (Stravs *et al.* 2022). A key limiting factor remains the availability of high-quality benchmark datasets. While proteomics has benefited from large public training sets, equivalent resources in metabolomics have only recently started to emerge. Notably, MassSpecGym provides a benchmark suite to facilitate model evaluation and development (Bushuiev et al. 2024), helping to advance ML applications in this domain.

A promising direction for future work is the development of foundation models trained on large-scale MS data. Inspired by advances in other omics domains, these models are pre-trained on broad datasets and can be fine-tuned for

specific downstream applications (Sanders *et al.* 2025, Bushuiev *et al.* 2025). Initial foundation models for MS have focused on spectral data, but future iterations could incorporate additional context such as metadata, literature, or a broader representation of molecular complexity across different omics layers. These models have the potential to support cross-domain applications in proteomics, metabolomics, and beyond, providing a unified framework for interpreting complex MS data.

3 Challenges and opportunities in computational mass spectrometry

In this section, we outline major bioinformatics challenges and highlight emerging opportunities across the computational MS landscape (Fig. 1).

3.1 Data harmonization across experiments and modalities

Heterogeneity in MS data remains a major obstacle to integrative analysis. Differences in instrumentation, acquisition settings, and experimental protocols—both within and across laboratories—can introduce systematic variation that complicates comparisons. Even repeated measurements on the same instrument can yield different results due to temporal shifts in sensitivity, ionization efficiency, or calibration (Bittremieux *et al.* 2018). This variability becomes particularly problematic when aggregating datasets from multiple sources, where technical artifacts may mask or confound underlying biological signals.

Among the most persistent sources of technical variation are batch effects, which can obscure genuine biological differences and hinder the comparability of datasets across studies and time points (Leek *et al.* 2010). While batch correction methods exist and are routinely used in both proteomics and metabolomics, they are often optimized for relatively homogeneous datasets. In more complex settings, such as multicenter studies or retrospective analyses (Tabb *et al.* 2010), current methods may fail to remove technical variation without also suppressing biological structure. More robust, data-driven methods are needed to support harmonization in increasingly complex MS applications, including multicenter, longitudinal, and multi-omics studies.

In related omics fields such as genomics and transcriptomics, substantial progress has been made in mitigating batch effects through various approaches such as empirical Bayes modeling (Johnson *et al.* 2007) and manifold alignment (Korsunsky *et al.* 2019). These frameworks have proven effective in normalizing large, multi-batch datasets, including single-cell transcriptomic studies, by aligning shared biological signals while reducing technical variance. However, the challenge is even greater for MS-based omics, where data are inherently noisier, feature detection and quantification are less standardized, and measurement variability can be introduced at multiple stages of the workflow. Consequently, many methods that can succeed in genomics domains do not transfer straightforwardly to MS data, where missingness, non-linear signal distortion, and feature overlap complicate direct correction.

Addressing these challenges is central to realizing heterogeneous data integration in MS-based omics. Without effective batch harmonization, datasets from different laboratories, instruments, or acquisition protocols remain only partially comparable, limiting opportunities for repository-level meta-analysis and ML model training. Overcoming this limitation would enable disparate experiments to contribute to a shared, cumulative understanding of biological systems, reducing the need for redundant experiments and accelerating discovery through reanalysis of existing data resources.

Deep learning offers a potential path forward. With the ability to learn complex, non-linear relationships across datasets, neural networks have been proposed as tools for data harmonization and imputation (Webel *et al.* 2024). Initial applications suggest that deep learning can model batch variation while preserving signal, even in the presence of missing data and instrument-specific noise. If trained on diverse and well-annotated datasets, such models could enable reuse of previously incompatible data, enhancing the utility of existing MS repositories. However, these methods must be applied with care, as overly aggressive batch correction can inadvertently suppress biologically meaningful variation or rare phenotypes. Rigorous validation and benchmarking across

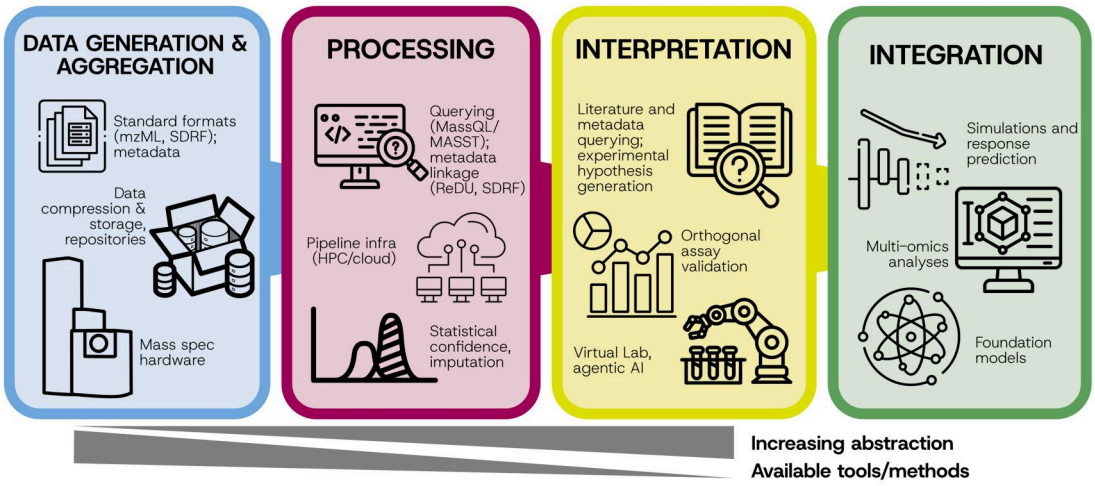

Figure 1. Computational MS spans a continuum from raw data generation to systems-level integration. Key challenges and opportunities arise at each stage of the workflow, from hardware and data formats, to cloud-scale processing, statistical confidence, interpretation frameworks, and multi-omics integration. Progress along this pipeline creates new opportunities for innovation, including virtual labs, agentic AI, and foundation models that can enable more autonomous and integrative data interpretation.

independent datasets are therefore essential to ensure that batch correction improves data comparability without obscuring true biological differences.

Beyond harmonization within a single omics modality, the field is also moving toward multi-modal data integration. In proteomics, combining MS-based measurements with non-MS, affinity-based technologies such as Olink or SomaLogic can increase coverage and sensitivity (Beimers *et al.* 2025). While these approaches are particularly useful in large cohorts (Ferkingstad *et al.* 2021, Sun *et al.* 2023), they cannot provide information on post-translational modifications (PTMs) and protein isoforms, where MS retains a clear advantage. In metabolomics, combining MS data with nuclear magnetic resonance (NMR) spectroscopy can improve structural elucidation and quantification. Despite the complementarity of these technologies, their integration is nontrivial. Differences in resolution, dynamic range, data structure, and software pipelines pose significant barriers. Addressing these will require not only computational innovation but also greater standardization in data formats and metadata reporting across platforms.

3.2 Statistical confidence for robust data interpretation

Reliable statistical assessment is fundamental to interpreting MS data with confidence. Among the most widely used approaches, false discovery rate (FDR) estimation (Storey 2011) plays a central role in distinguishing true positives from random matches. Accurate FDR control is critical for ensuring reproducibility, enabling cross-study comparisons, and building trust in downstream biological conclusions.

In proteomics, FDR estimation methods are well established, particularly through the use of target–decoy competition strategies (Elias and Gygi 2007). These methods have become integral to peptide and protein identification pipelines, allowing researchers to systematically control the rate of false positives in large proteomic datasets. Nevertheless, proper FDR estimation remains a complex and nuanced challenge, and even subtle methodological mistakes can compromise the validity of downstream analyses (Frejno *et al.* 2025, Wen *et al.* 2025). As proteomics applications become more complex (e.g. PTM analysis, non-canonical sequences), the need for continued refinement of confidence estimation remains.

In contrast, small-molecule MS applications such as metabolomics lack comparably mature strategies for statistical validation. The high chemical diversity of metabolites, combined with a heterogeneous fragmentation behavior and limited ground-truth annotations, makes the application of target–decoy-based FDR estimation difficult. As a result, confidence measures in metabolomics often rely on heuristic thresholds or scoring metrics that lack formal statistical calibration. While there are some emerging tools toward establishing FDR strategies for metabolomics (Scheubert *et al.* 2017, Alka *et al.* 2022), these approaches have not been widely adopted yet and require further validation to reach the robustness seen in proteomics.

3.3 Harnessing repository-scale data for biological discovery

Public MS data repositories have become essential infrastructure for the field, supporting open data sharing, reproducibility, and secondary analyses. Initiatives aligned with the FAIR principles (Wilkinson *et al.* 2016) have led to the growth of domain-specific repositories such as PRIDE (Perez-Riverol *et al.* 2022), MassIVE, PeptideAtlas (Desiere *et al.* 2006), iProx (Ma *et al.* 2019), jPOST (Moriya *et al.* 2019), and Panorama Public (Sharma *et al.* 2018) under the ProteomeXchange umbrella (Vizcaíno *et al.* 2014) for proteomics; GNPS/MassIVE (Wang *et al.* 2016), MetaboLights (Haug *et al.* 2013), and Metabolomics WorkBench (Sud *et al.* 2016) for metabolomics, which are now beginning to coordinate under the emerging MetabolomicsHub initiative; and GlycoPOST (Watanabe *et al.* 2021) for glycomics. These resources collectively enable data reanalysis, ML development, cross-study integration, and others at scales that were previously unattainable.

One major benefit of these repositories is their role in enabling ML applications (Mann *et al.* 2021). Public datasets have become valuable resources for training models to predict molecular properties, enhancing identification and scoring in proteomics workflows. ML approaches are also increasingly applied to functional annotation tasks, such as predicting the functional potential of PTMs (Ochoa *et al.* 2020) or inferring protein characteristics from large-scale datasets (Bileschi *et al.* 2022).

Meta-analysis is another powerful use case. In proteomics, community-driven efforts, such as those coordinated by the Human Proteome Organization (HUPO), have reanalyzed thousands of datasets to characterize the human proteome, with evidence now available for approximately 93% of canonical human proteins (Omenn *et al.* 2024). These efforts span a range of objectives, from quantitative reanalysis (Choi *et al.* 2020, Dai *et al.* 2024) and PTM mapping (Ramasamy *et al.* 2020) to proteogenomics approaches for genome annotation (Levitsky *et al.* 2023), including the discovery of non-canonical proteins (Deutsch *et al.* 2024). As data volumes grow and tools mature, similar large-scale efforts are likely to expand into emerging areas such as single-cell proteomics and integrated multi-omics analyses.

In metabolomics, repository-scale reuse is also gaining momentum. ReDU provides a structured system for discovering relevant datasets across repositories based on consistent metadata, facilitating comparative analysis and integration (Jarmusch *et al.* 2020, El Abiead *et al.* 2025). At the spectrum level, tools like MASST (Wang *et al.* 2020) and Flash Entropy (Li and Fiehn 2023) allow researchers to query individual mass spectra against entire repositories to contextualize "unknowns." In addition, the nearest neighbor suspect spectral library (Bittremieux *et al.* 2023) leverages repository-scale data to suggest candidate structures for unannotated molecules, highlighting the potential of large-scale comparisons to accelerate discovery.

Recently, standardized MS query languages have emerged to further improve accessibility and utility of public repositories. The Mass Spec Query Language (MassQL) (Damiani *et al.* 2025) offers a human-readable syntax for querying raw MS data across multiple repositories, enabling searches for spectral features such as specific precursor ions, fragment ions, and neutral losses. By abstracting over software-specific implementations, MassQL empowers researchers, including those without extensive programming skills, to mine large datasets for targeted patterns without needing to download or locally process raw files. This approach has already led to the creation of new domain-specific spectral libraries, such as a large-scale bile acid library (Mohanty *et al.* 2024), and

paves the way for broad hypothesis testing and structured reanalysis across repository-scale public data.

Despite these advances, key challenges remain. Processing repository-scale MS data is computationally intensive, requiring cloud or HPC environments and well-engineered software capable of handling heterogeneous formats and metadata. Unlike smaller-scale projects, these analyses demand maintainable infrastructure and robust codebases that can scale effectively.

As discussed previously, statistical confidence is another concern. While FDR control is well-developed in individual proteomics studies, standard approaches do not readily extend to aggregated datasets, where combining datasets from diverse studies can inflate FDR and lead to an excess of false positives (Serang and Käll 2015). This issue arises because true positives tend to overlap across studies, while false positives are often unique, causing cumulative FDR inflation. Addressing this requires new statistical frameworks that account for dataset heterogeneity while maintaining confidence in discovery rates (Savitski *et al.* 2015).

Metadata quality is a further bottleneck. While experimental metadata is critical for reanalysis and interpretation, it is often inconsistently reported or embedded in unstructured formats. Efforts such as SDRF-Proteomics in proteomics (Dai *et al.* 2021), and ISA-Tab (Rocca-Serra *et al.* 2016) and ReDU (Jarmusch *et al.* 2020) in metabolomics have improved metadata standards, but adoption remains uneven. Better metadata capture at data submission, combined with automated extraction and validation tools (Claeys *et al.* 2023), could greatly improve data reuse and interoperability.

In parallel with metadata, quality control (QC) information is equally important for assessing data reliability and comparability across studies. Providing QC metrics alongside datasets supports transparent evaluation of instrument performance and experimental consistency, which are prerequisites for large-scale data integration and automated reanalysis. Recently, the HUPO Proteomics Standards Initiative (PSI) Quality Control Working Group has developed the mzQC file format, which enables standardized reporting and exchange of QC results (Bittremieux *et al.* 2017, Bielow *et al.* 2024). Incorporating both structured metadata and QC metrics at the point of data submission would substantially enhance the reproducibility, interpretability, and long-term value of public MS data resources.

Finally, the long-term sustainability of repository-scale science depends on incentivizing data sharing. Data generators often bear the burden of curation and formatting, with limited direct benefit. Yet public data are essential not only for reproducibility but also for enabling downstream reuse in various applications such as meta-analysis, benchmarking, and ML. As expectations for standardized data sharing grow, there is increasing recognition of the need to properly acknowledge contributors. Community discussions around improved citation practices, contributor recognition, and dataset-level metrics are ongoing and will be critical to sustaining high-quality data contributions.

Looking ahead, the convergence of large-scale structured data, standardized metadata, and scalable MS data processing opens the door to more autonomous modes of biological discovery using MS. Virtual laboratory concepts, where software agents reason over structured queries, metadata, literature, experimental design, and prompting from human scientists, could enable closed-loop experimentation across public and private datasets. For example, a "virtual lab" could combine spectrum-level search via tools like MASST, coupled to sample metadata via ReDU or SDRF, and AI-assisted summarization to propose new experimental hypotheses or suggest orthogonal validation experiments. These ideas are increasingly plausible as workflows become more digitized and automated; indeed, efforts toward "lights-out" sample preparation robotics (Schär *et al.* 2025) provide the physical infrastructure for seamless, automated experimental execution, while computational infrastructure for repository-scale access and metadata standardization enable contextual reasoning.

3.4 Multi-omics integration for deeper biological insights

Integrating data across multiple omics layers offers the potential for a more holistic understanding of biological systems by linking genomic, transcriptomic, proteomic, metabolomic, and other molecular profiles (Hasin *et al.* 2017, Karczewski and Snyder 2018). While single-omics studies provide valuable insights within their respective domains, multi-omics approaches aim to uncover relationships across molecular layers, such as how genetic variation influences protein expression or how metabolic states reflect transcriptional regulation (Wörheide *et al.* 2021). However, many current applications of multi-omics remain limited in scope, often focusing on simple overlap analyses of top features rather than truly integrative modeling.

Effective multi-omics integration requires methods that can capture dependencies and interactions between omics layers in a statistically sound and biologically meaningful way. Such approaches have the potential to improve biomarker discovery, refine mechanistic hypotheses, and provide context-specific insights into disease or treatment responses. However, realizing these benefits requires moving beyond concatenated datasets and toward analytical frameworks that explicitly model relationships across layers.

Several challenges impede the development of robust multi-omics integration methods (Fig. 2). Omics data types differ in their dynamic ranges, measurement noise, missingness, and resolution, complicating direct alignment and joint modeling. A further practical obstacle is the lack of consistent identifier mapping across omics layers—for example, between genes, proteins, metabolites, and pathways—which hampers accurate data linkage and biological interpretation. Standard pipelines for normalization and integration across omics types remain underdeveloped, and most existing tools are designed for single-modality workflows. Moreover, many statistical models lack mechanisms to account for variation in data quality between layers, leading to biased or overconfident conclusions when integrating heterogeneous datasets.

Beyond traditional omics layers, structural proteomics approaches such as cross-linking mass spectrometry (XL-MS) (O'Reilly and Rappsilber 2018) and hydrogen–deuterium exchange (HDX) (Konermann and Scrosati 2024) offer complementary insights by probing the spatial organization and dynamics of proteins and protein complexes. These techniques are increasingly integrated into multi-omics workflows to add a structural dimension to systems-level analyses. The introduction of high-accuracy structure prediction tools such as AlphaFold (Jumper *et al.* 2021) has dramatically expanded our ability to model protein structures *in silico*. These predicted structures now serve as valuable priors for interpreting

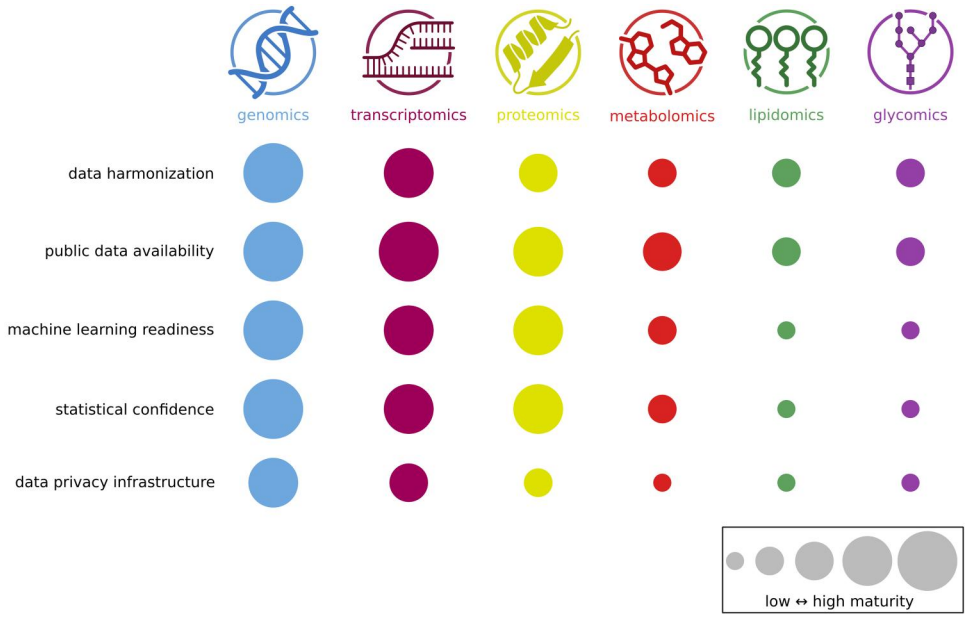

Figure 2. Relative computational maturity of different omics domains. Genomics and transcriptomics lead in maturity, reflecting well-established standards and infrastructure. Proteomics occupies an intermediate position, while metabolomics, lipidomics, and glycomics still face substantial challenges across several dimensions. Uneven maturity across omics fields contributes to the difficulties of multi-omics integration. Circle size represents relative maturity, from low (small) to high (large).

XL-MS and HDX data, facilitating more accurate cross-link mapping and conformational analysis (Drake *et al.* 2022). Conversely, experimental data from XL-MS can validate or constrain *in silico* predictions, and even guide refinement of structural models in regions of low confidence (Stahl *et al.* 2023). Integrating such structural information into multi-omics analyses may reveal mechanistic insights that are inaccessible through sequence-based data alone, such as conformational regulation, allostery, or the physical basis of protein–protein interactions.

Broader visions are also emerging in systems biology, such as the ambition to construct an AI-powered "virtual cell": a multi-scale, multi-modal neural architecture capable of simulating molecular, cellular, and tissue-level behavior across diverse biological states (Bunne *et al.* 2024). MS data, especially from public repositories enriched with high-quality annotations and cohort-level metadata, provide a crucial training set for such efforts (Sun *et al.* 2025). Unlike sequencing-based technologies, MS uniquely captures diverse molecular species, including proteins, lipids, metabolites, and glycans, that are essential for accurately modeling biological complexity at multiple levels.

Ultimately, the integration of multi-omics data represents an essential goal in computational biology, promising to unlock new layers of understanding in complex biological systems. Achieving this goal will require overcoming significant challenges in data compatibility, statistical modeling, and computational resource allocation.

3.5 Building machine learning literacy for mass spectrometry

ML and AI have become increasingly central to computational MS, enabling advances in spectrum interpretation, molecular property prediction, quality control, and data integration, among other topics. However, effectively applying ML to MS data remains challenging. The complexity of MS datasets requires careful methodological choices and a strong understanding of both domain-specific and computational principles.

While ML tools are becoming more accessible—for example, through software libraries like Scikit-Learn (Pedregosa *et al.* 2011) and various deep learning frameworks—this accessibility can mask underlying complexities. Without sufficient training in ML fundamentals, users may select inappropriate models, overlook preprocessing needs, or apply flawed evaluation strategies (Adams and Bittremieux 2025). In MS contexts, where the structure of the data often reflects subtle biological phenomena, such missteps can lead to incorrect conclusions or irreproducible findings. Model selection, hyperparameter tuning, and evaluation metrics must be tailored to the specific characteristics of MS data to ensure meaningful results.

A particularly critical issue in omics-based ML is the risk of data leakage, where information from outside the training set inadvertently informs model training, resulting in overestimated performance. This can occur through improper data splitting (Desaire 2022), reusing test data during model optimization, or embedding confounding factors in input features. Data leakage often goes unnoticed, even in published studies (Quinn 2021), and undermines trust in reported outcomes. Overfitting is another common challenge, particularly given the large number of features relative to sample size in many omics datasets. Regularization, robust cross-validation, and careful feature selection are essential but must be adapted to the nuances of MS data, including frequent missing values and noise (Adams and Bittremieux 2025).

Transparent reporting is key to enabling reproducibility and scientific rigor. ML-based MS studies should clearly describe preprocessing steps, model architecture and parameters, validation strategies, and performance metrics, and where possible, register models in dedicated resources such as the DOME (Data, Optimization, Model, and Evaluation)

(Walsh *et al.* 2021) registry to ensure standardized documentation and accessibility (Palmblad *et al.* 2022). As models grow more complex, particularly with the adoption of deep learning, such documentation becomes even more important for peer evaluation and reuse.

Addressing educational gaps is also essential. Many researchers applying ML to MS come from biology or analytical chemistry backgrounds and may not have formal training in statistical learning. Core concepts such as overfitting, cross-validation, and proper data partitioning are frequently underappreciated. To address this, hands-on training opportunities, such as workshops, tutorials, and summer schools, are increasingly being offered within the MS community. These initiatives aim to equip researchers with practical skills and promote best practices in ML application (Rehfeldt *et al.* 2023).

Finally, interdisciplinary collaboration remains one of the most effective ways to ensure methodological rigor. Involving data scientists and statisticians early in project planning helps align study design with appropriate computational analysis, prevents common pitfalls, and fosters development of more robust and interpretable models. Likewise, involving biologists and chemists throughout an analysis helps ensure that computational frameworks and analyses are grounded and interpretable. As MS increasingly intersects with advanced computational methods, fostering this cross-domain dialogue will be essential for responsible and impactful research.

3.6 Privacy challenges in clinical omics

As MS becomes increasingly integrated into clinical research, the handling of sensitive human data raises important ethical and privacy considerations (Mann *et al.* 2021). While public data sharing is vital for reproducibility and scientific progress, clinical datasets, particularly those containing individual-level health or molecular information, require careful management to ensure data privacy is not compromised.

Privacy risks in proteomics share parallels with those in genomics (Lin *et al.* 2004). Protein sequences, derived from expressed genes, can carry identifying features, particularly when combined with phenotypic or clinical metadata. Although proteomic data lacks certain elements found in genomic sequences, such as non-coding regions, variations in protein-coding regions or characteristic expression patterns may still allow for re-identification under certain conditions (Geyer *et al.* 2021). Metabolite, lipid, and glycan expression patterns potentially carry identifiable information as well (Keane *et al.* 2021). These risks are amplified in the context of multi-omics datasets, where linking across molecular layers can increase the resolution of individual profiles.

Regulatory frameworks such as the General Data Protection Regulation in the European Union and the Health Insurance Portability and Accountability Act in the United States impose strict requirements for managing sensitive personal data. These regulations mandate stringent technical and organizational measures to protect the privacy of the affected persons. However, public MS repositories, originally designed to promote open access and broad reuse, currently lack the infrastructure for secure authentication, access control, or compliance tracking (Bandeira *et al.* 2021). As a result, it becomes difficult to share clinically sensitive data in a privacy-compliant manner, creating obstacles for researchers who wish to adhere to the FAIR data principles while protecting individual privacy. Without privacy-aware repositories or controlled-access solutions, clinical proteomics—and increasingly, clinical metabolomics and lipidomics—risks falling behind in data sharing. This challenge needs to be addressed urgently in order to ensure that the positive role data sharing has been playing in MS-based omics can continue in clinical settings as well. Ideally, the MS community can learn from—and align itself with—the efforts in clinical genomics, where suitable repositories have been developed (D'Altri *et al.* 2025).

In parallel to repository infrastructure development, privacy-preserving computational techniques offer complementary solutions. Federated learning, where ML models are trained across decentralized datasets without transferring the raw data, or approaches based on homomorphic encryption are particularly well suited for clinical MS applications (Burankova *et al.* 2025, Cai *et al.* 2025). These approaches require sophisticated infrastructure, but they allow multi-institutional analyses while maintaining local control over sensitive data.

3.7 Role of the CompMS community of special interest

ISCB's CompMS Community of Special Interest brings together researchers from computer science, bioinformatics, statistics, and molecular biology to advance the analysis and interpretation of MS data. Uniting disciplines under a shared interest in MS-based omics, the CompMS community supports proteomics, metabolomics, lipidomics, glycomics, and related fields where MS serves as a core analytical platform. Through its activities, the community fosters collaboration and knowledge exchange aimed at transforming complex datasets into meaningful biological insights.

CompMS emphasizes community building and open exchange through accessible communication channels. A central hub for this interaction is the CompMS Slack workspace (https://compms.slack.com), where researchers at all career stages seek advice, share software updates, and discuss emerging challenges and opportunities in computational MS (Fig. 3). This environment promotes open dialogue and collaborative problem-solving, encouraging both technical innovation and mentorship.

The community's primary annual event is the CompMS track at the Intelligent Systems for Molecular Biology (ISMB) conference, which showcases the latest advances in computational MS, including method development, bioinformatics, and ML applications. These sessions draw a multidisciplinary audience and provide a venue for presenting novel tools and addressing key challenges across MS-based domains. By facilitating cross-disciplinary interaction, the CompMS track serves as a vital forum for the exchange of ideas and development of shared standards. Beyond ISMB, CompMS actively engages with the broader MS community by contributing to Bioinformatics Hub sessions at the annual conferences of the American Society for Mass Spectrometry (ASMS) and HUPO, and by seeking synergies with initiatives led by organizations such as HUPO, the HUPO-PSI, the Metabolomics Society, the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS), and ELIXIR. Such cross-community involvement helps to bridge the gap between computational and experimental MS communities, connecting developers of algorithms and data standards with researchers driving new MS technologies and applications.

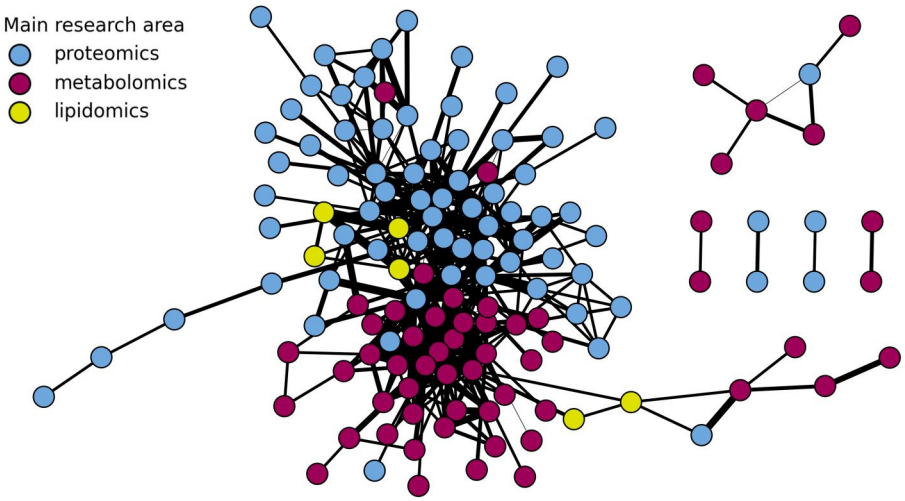

Figure 3. Co-publication network of users from the CompMS Slack workspace. Nodes represent individual users (based on best-effort name matching to PubMed), colored by their primary research area as inferred by PubMed keyword matching. Edges indicate co-authorship of at least one scientific publication, with edge width reflecting the number of shared publications. The network highlights the collaborative and interdisciplinary nature of the CompMS community, bridging proteomics, metabolomics, lipidomics, and related fields.

By facilitating collaboration across these diverse forums, CompMS provides a platform for advancing computational MS globally.

A distinctive feature of the CompMS community is its strong support for early-career researchers. The ISMB track highlights junior contributions through talks, posters, and awards, while dedicated travel fellowships broaden participation by enabling students and postdocs to attend. These opportunities help integrate emerging researchers into the field and support the development of a vibrant, inclusive community committed to advancing computational MS.

4 Conclusions

MS has become a foundational technology in the life sciences, supporting deep molecular analyses across proteomics, metabolomics, lipidomics, glycomics, and other omics disciplines. Its ability to detect and characterize a broad range of molecules makes it indispensable for understanding biological systems in both health and disease. As research increasingly relies on high-throughput, high-resolution, and multidimensional data, the role of MS in capturing phenotypic complexity continues to grow.

This perspective outlined recent advances in computational MS, including developments in acquisition strategies, scalable data processing, and ML/AI. We also highlighted ongoing challenges, such as batch effects and data harmonization, the need for robust statistical confidence across domains, limitations in multi-omics integration, and emerging concerns around privacy in clinical applications. Addressing these issues is critical to ensure that MS data remains reliable, interpretable, and broadly reusable.

Looking ahead, realizing the full potential of computational MS will depend on developing systems that are not only scalable and interoperable but also "AI-ready." In this context, AI-readiness refers to the extent to which MS (and derived) data, metadata, and analytical tools are structured and harmonized to enable ML and automated reasoning. Concretely, this includes the use of FAIR principles, standardized metadata schemas, transparent and reproducible preprocessing pipelines, and programmatic data access through machine-actionable interfaces. Establishing such foundations will be essential for enabling autonomous data analysis, knowledge discovery, and virtual laboratory environments in the future.

To further advance MS data toward AI-readiness, a new AI Working Group has recently been established under the HUPO-PSI (https://www.psidev.info/ai-readiness) (Deutsch *et al.* 2023). The group's goal is to make public proteomics data suitable for AI applications and to support and standardize the incorporation of AI within workflows that generate new data. This is a timely, community-driven effort, open to all interested scientists, and represents an important step toward enabling reproducible and AI-augmented discovery in MS-based related omics disciplines.

These emerging standardization and interoperability efforts complement the broader activities of the CompMS Community of Special Interest. By fostering collaboration across disciplines, advancing methodological development, and supporting early-career researchers, the CompMS community is helping to shape a robust and forward-looking computational MS ecosystem that will continue to empower discovery across diverse biological domains.

Acknowledgements

Figure 2 was created using BioRender.

Author contributions

Wout Bittremieux (Conceptualization [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Timo Sachsenberg (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Lindsay Pino (Conceptualization [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Marie A. Brunet (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Isabell Bludau (Conceptualization [equal], Writing—original draft [equal],

Writing—review & editing [equal]), Oliver Kohlbacher (Writing—original draft [equal], Writing—review & editing [equal]), and Juan Antonio Vizcaino (Writing—original draft [equal], Writing—review & editing [equal])

Conflict of interest

L.P. is a co-founder and employee of Talus Bio, a biotechnology company that uses mass spectrometry in its research and development efforts. T.S. and O.K. are co-founders of OpenMS Inc., a nonprofit organization propagating open-source software in computational mass spectrometry.

Funding

This work was supported by the University of Antwerp's Research Fund and the Research Foundation—Flanders (G087625N and G0AHY25N to W.B.); the Federal Ministry of Research, Technology and Space in the frame of de.NBI/ELIXIR-DE (W-DE.NBI-022 to T.S. and O.K.); the Ministry of Science, Research and Arts Baden-Württemberg (LIBIS to T.S. and O.K.); EMBL core funding (to J.A.V.); Wellcome (223745/Z/21/Z to J.A.V.); BBSRC (BB/W000156/1, BB/Y513829/1, BB/X001911/1, and BB/V018779/1 to J.A.V.); EPSRC (EP/Y035984/1 to J.A.V.); and the Chan Zuckerberg Initiative (2024–350548 to J.A.V.).

References

Adams C, Bittremieux W. A 2025 perspective on the role of machine learning in clinical proteomics. *Expert Rev Proteomics* 2025; **22**:363-74.

Alka O, Shanthamoorthy P, Witting M *et al.* DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. *Nat Commun* 2022; **13**:1347.

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 Proteomics* 2021; **20**:100071.

Beimers WF, Overmyer KA, Sinitcyn P *et al.* Technical evaluation of plasma proteomics technologies. *J Proteome Res* 2025;**24**:3074–87.

Bennett HM, Stephenson W, Rose CM *et al.* Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat Methods* 2023;**20**:363–74.

Bielow C, Hoffmann N, Jimenez-Morales D *et al.* Communicating mass spectrometry quality information in mzQC with Python, R, and Java. *J Am Soc Mass Spectrom* 2024;**35**:1875–82.

Bileschi ML, Belanger D, Bryant DH *et al.* Using deep learning to annotate the protein universe. *Nat Biotechnol* 2022;**40**:932–7.

Bills B, Barshop WD, Sharma S *et al.* Novel real-time library search driven data acquisition strategy for identification and characterization of metabolites. *Anal Chem* 2022;**94**:3749–55.

Bittremieux W, Ananth V, Fondrie WE *et al.* Deep learning methods for de novo peptide sequencing. *Mass Spectrom Rev* 2024.

Bittremieux W, Avalon NE, Thomas SP *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nat Commun* 2023;**14**:8488.

Bittremieux W, Tabb DL, Impens F *et al.* Quality control in mass spectrometry-based proteomics. *Mass Spectrom Rev* 2018; **37**:697–711.

Bittremieux W, Walzer M, Tenzer S *et al.* The human proteome organization–proteomics standards initiative quality control working group: making quality control more accessible for biological mass spectrometry. *Anal Chem* 2017;**89**:4474–9.

Bouwmeester R, Gabriels R, Hulstaert N *et al.* DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat Methods* 2021;**18**:1363–9.

Bunne C, Roohani Y, Rosen Y *et al.* How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024; **187**:7045–63.

Burankova Y, Abele M, Bakhtiari M *et al.* Privacy-preserving multicenter differential protein abundance analysis with FedProt. *Nat Comput Sci* 2025;**5**:675–88.

Bushuiev R, Bushuiev A, de Jonge NF *et al.* MassSpecGym: a benchmark for the discovery and identification of molecules. In: *Proceedings of the 38th Conference on Neural Information Process System NeurIPS 2024 Datasets Benchmarks Track*, Vol. 37. Vancouver, BC, Canada: Curran Associates, Inc., 2024.

Bushuiev R, Bushuiev A, Samusevich R *et al.* Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat Biotechnol* 2025.

Cai Z, Boys EL, Noor Z *et al.* Federated deep learning enables cancer subtyping by proteomics. *Cancer Discov* 2025;**15**:1803–18.

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.

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.

Dai C, Füllgrabe A, Pfeuffer J *et al.* A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat Commun* 2021;**12**:5854.

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.

D'Altri T, Freeberg MA, Curwin AJ *et al.* The federated European genome–phenome archive as a global network for sharing human genomics data. *Nat Genet* 2025;**57**:481–5.

Damiani T, Jarmusch AK, Aron AT *et al.* A universal language for finding mass spectrometry data patterns. *Nat Methods* 2025; **22**:1247–54.

Degroeve S, Martens L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 2013;**29**:3199–203.

Demichev V, Szyrwiel L, Yu F *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat Commun* 2022;**13**:3944.

Desaire H. How (not) to generate a highly predictive biomarker panel using machine learning. *J Proteome Res* 2022;**21**:2071–4.

Desiere F, Deutsch EW, King NL *et al.* The PeptideAtlas project. *Nucleic Acids Res* 2006;**34**:D655–8.

Deutsch EW, Kok LLW, Mudge JM *et al.* High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. bioRxiv, https://doi.org/2024.09.09.612016, 2024, preprint: not peer reviewed.

Deutsch EW, Vizcaíno JA, Jones AR *et al.* Proteomics standards initiative at twenty years: current activities and future work. *J Proteome Res* 2023;**22**:287–301.

Devreese R, Nameni A, Declercq A *et al.* Collisional cross-section prediction for multiconformational peptide ions with IM2Deep. *Anal Chem* 2025;**97**:15113–21.

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9.

Drake ZC, Seffernick JT, Lindert S *et al.* Protein complex prediction using rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat Commun* 2022;**13**:7846–9.

Dührkop K, Fleischauer M, Ludwig M *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302.

El Abiead Y, Strobel M, Payne T *et al.* Enabling pan-repository reanalysis for big data science of public metabolomics data. *Nat Commun* 2025;**16**:4838.

Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;**4**:207–14.

Ferkingstad E, Sulem P, Atlason BA *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021; **53**:1712–21.

Frejno M, Berger MT, Tüshaus J *et al.* Unifying the analysis of bottom-up proteomics data with CHIMERYS. *Nat Methods* 2025; **22**:1017–27.

Gessulat S, Schmidt T, Zolg DP *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* 2019;**16**:509–18.

Geyer PE, Mann SP, Treit PV *et al.* Plasma proteomes can be reidentifiable and potentially contain personally sensitive and incidental findings. *Mol Cell Proteomics* 2021;**20**:100035.

Goldman S, Wohlwend J, Stražar M *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Mach Intell* 2023;**5**:965–79.

Griss J, Perez-Riverol Y, Lewis S *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* 2016;**13**:651–6.

Hasin Y, Seldin M, Lusis A *et al.* Multi-omics approaches to disease. *Genome Biol* 2017;**18**:83–15.

Haug K, Salek RM, Conesa P *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 2013;**41**:D781–6.

Heil LR, Damoc E, Arrey TN *et al.* Evaluating the performance of the astral mass analyzer for quantitative proteomics using data-independent acquisition. *J Proteome Res* 2023;**22**:3290–300.

Huber F, van der Burg S, van der Hooft JJJ *et al.* MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminform* 2021;**13**:84.

Huttlin EL, Bruckner RJ, Paulo JA *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* 2017;**545**:505–9.

Huttlin EL, Ting L, Bruckner RJ *et al.* The BioPlex network: a systematic exploration of the human interactome. *Cell* 2015;**162**:425–40.

Jarmusch AK, Wang M, Aceves CM *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat Methods* 2020; **17**:901–4.

Jeong K, Kim J, Gaikwad M *et al.* FLASHDeconv: ultrafast, high-quality feature deconvolution for top-down proteomics. *Cell Syst* 2020;**10**:213–8.e6.

Johnson WE, Li C, Rabinovic A *et al.* Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 2007;**8**:118–27.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kalhor M, Lapin J, Picciani M *et al.* Rescoring peptide spectrum matches: boosting proteomics performance by integrating peptide property predictors into peptide identification. *Mol Cell Proteomics* 2024;**23**:100798.

Käll L, Canterbury JD, Weston J *et al.* Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 2007;**4**:923–5.

Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;**19**:299–310.

Keane TM, O'Donovan C, Vizcaíno JA *et al.* The growing need for controlled data access models in clinical proteomics and metabolomics. *Nat Commun* 2021;**12**:5787.

Konermann L, Scrosati PM. Hydrogen/deuterium exchange mass spectrometry: fundamentals, limitations, and opportunities. *Mol Cell Proteomics* 2024;**23**:100853.

Kong AT, Leprevost FV, Avtonomov DM *et al.* MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat Methods* 2017;**14**:513–20.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019; **16**:1289–96.

Kou Q, Xun L, Liu X *et al.* TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 2016;**32**:3495–7.

Lazear MR. Sage: an open-source tool for fast proteomics searching and quantification at scale. *J Proteome Res* 2023;**22**:3652–9.

LeDuc RD, Deutsch EW, Binz P-A *et al.* Proteomics standards initiative's ProForma 2.0: unifying the encoding of proteoforms and peptidoforms. *J Proteome Res* 2022;**21**:1189–95.

Leek JT, Scharpf RB, Bravo HC *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.

Levitsky LI, Ivanov MV, Goncharov AO *et al.* Massive proteogenomic reanalysis of publicly available proteomic datasets of human tissues in search for protein recoding via adenosine-to-inosine RNA editing. *J Proteome Res* 2023;**22**:1695–711.

Li Y, Fiehn O. Flash entropy search to query all mass spectral libraries in real time. *Nat Methods* 2023;**20**:1475–8.

Lin Z, Owen AB, Altman RB *et al.* Genomic research and human subject privacy. *Science* 2004;**305**:183.

Ma J, Chen T, Wu S *et al.* iProX: an integrated proteome resource. *Nucleic Acids Res* 2019;**47**:D1211–7.

Mann M, Kumar C, Zeng W-F *et al.* Artificial intelligence for proteomics and biomarker discovery. *Cell Syst* 2021;**12**:759–70.

Mann SP, Treit PV, Geyer PE *et al.* Ethical principles, constraints, and opportunities in clinical proteomics. *Mol Cell Proteomics* 2021; **20**:100046.

Martens L, Chambers M, Sturm M *et al.* mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011; **10**:R110.000133.

Marx V. Inside the chase after those elusive proteoforms. *Nat Methods* 2024;**21**:158–63.

Mendes ML, Borrmann KF, Dittmar G *et al.* Eleven shades of PASEF. *Expert Rev Proteomics* 2024;**21**:367–76.

Messner CB, Demichev V, Wendisch D *et al.* Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Syst* 2020;**11**:11–24.e4.

Method of the Year 2024: spatial proteomics. *Nat Methods* 2024; **21**:2195–6.

Mohanty I, Mannochio-Russo H, Schweer JV *et al.* The underappreciated diversity of bile acid modifications. *Cell* 2024;**187**: 1801–18.e20.

Mongia M, Yasaka TM, Liu Y *et al.* Fast mass spectrometry search and clustering of untargeted metabolomics data. *Nat Biotechnol* 2024; **42**:1672–7.

Moriya Y, Kawano S, Okuda S *et al.* The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res* 2019;**47**:D1218–24.

Mund A, Coscia F, Kriston A *et al.* Deep visual proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* 2022;**40**:1231–40.

Neely BA. Cloudy with a chance of peptides: accessibility, scalability, and reproducibility with Cloud-Hosted environments. *J Proteome Res* 2021;**20**:2076–82.

Ochoa D, Jarnuczak AF, Viéitez C *et al.* The functional landscape of the human phosphoproteome. *Nat Biotechnol* 2020;**38**:365–73.

Omenn GS, Lane L, Overall CM *et al.* The 2023 report on the proteome from the HUPO human proteome project. *J Proteome Res* 2024; **23**:532–49.

O'Reilly FJ, Rappsilber J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat Struct Mol Biol* 2018;**25**:1000–8.

Palmblad M, Böcker S, Degroeve S *et al.* Interpretation of the DOME recommendations for machine learning in proteomics and metabolomics. *J Proteome Res* 2022;**21**:1204–7.

Pedregosa F *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.

Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;**50**:D543–52.

Qin Y, Huttlin EL, Winsnes CF *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* 2021;**600**:536–42.

Quinn TP. *Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning*. 2021. https://arxiv.org/abs/2107.03611

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.
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.
Roberts DS, Loo JA, Tsybin YO *et al.* Top-down proteomics. *Nat Rev Methods Primer* 2024;**4**:38.
Rocca-Serra P, Salek RM, Arita M *et al.* Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* 2016;**12**:14.
Sanders J, Yilmaz M, Russell JH *et al.* Foundation model for mass spectrometry proteomics. arXiv, https://doi.org/10.48550, 2025, preprint: not peer reviewed.
Savitski MM, Wilhelm M, Hahne H *et al.* A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol Cell Proteomics* 2015;**14**:2394–404.
Schär S, Räss L, Malinovska L *et al.* A flexible end-to-end automated sample preparation workflow enables reproducible large-scale bottom-up proteomics. *Anal Chem* 2025;**97**:22116–31.
Scheubert K, Hufsky F, Petras D *et al.* Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun* 2017;**8**:1494.
Schmid R, Heuckeroth S, Korf A *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* 2023;**41**:447–9.
Schweppe DK, Eng JK, Yu Q *et al.* Full-Featured, Real-Time database searching platform enables fast and accurate multiplexed quantitative proteomics. *J Proteome Res* 2020;**19**:2026–34.
Serang O, Käll L. Solution to statistical challenges in proteomics is more statistics, not less. *J Proteome Res* 2015;**14**:4099–103.
Sharma V, Eckels J, Schilling B *et al.* Panorama public: a public repository for quantitative data sets processed in skyline. *Mol Cell Proteomics* 2018;**17**:1239–44.
Stahl K, Graziadei A, Dau T *et al.* Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat Biotechnol* 2023;**41**:1810–9.
Storey JD. False discovery rate. In: *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer, 2011, 504–8.
Stravs MA, Dührkop K, Böcker S *et al.* MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;**19**:865–70.
Sud M, Fahy E, Cotter D *et al.* Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* 2016;**44**:D463–70.
Sun BB, Chiou J, Traylor M *et al.*; Regeneron Genetics Center. Plasma proteomic associations with genetics and health in the UK biobank. *Nature* 2023;**622**:329–38.
Sun Y, Jun A, Zhiwei L *et al.* Strategic priorities for transformative progress in advancing biology with proteomics and artificial intelligence. *Nat Methods* 2025.
Tabb DL, Vega-Montoto L, Rudnick PA *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* 2010;**9**:761–76.
Teleman J, Dowsey AW, Gonzalez-Galarza FF *et al.* Numerical compression schemes for proteomics mass spectrometry data. *Mol Cell Proteomics* 2014;**13**:1537–42.
Uhlen M, Oksvold P, Fagerberg L *et al.* Towards a knowledge-based human protein atlas. *Nat Biotechnol* 2010;**28**:1248–50.
Van Den Bossche T, Alexandrov T, Bilbao A *et al.* mzPeak: designing a scalable, interoperable, and future-ready mass spectrometry data format. *J Proteome Res* 2025;**24**:5329–35.
Vizcaíno JA, Deutsch EW, Wang R *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014;**32**:223–6.
Wadie B, Stuart L, Rath CM *et al.* METASPACE-ML: context-specific metabolite annotation for imaging mass spectrometry using machine learning. *Nat Commun* 2024;**15**:9110.
Walsh I, Fishman D, Garcia-Gasulla D *et al.*; ELIXIR Machine Learning Focus Group. DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* 2021;**18**:1122–7.
Wang F, Liigand J, Tian S *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;**93**:11692–700.
Wang M, Jarmusch AK, Vargas F *et al.* Mass spectrometry searches using MASST. *Nat Biotechnol* 2020;**38**:23–6.
Wang M, Carver JJ, Phelan VV *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 2016;**34**:828–37.
Wang Z, Mülleder M, Batruch I *et al.* High-throughput proteomics of nanogram-scale samples with zeno SWATH MS. *Elife* 2022;**11**:e83947.
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.
Webel H, Niu L, Nielsen AB *et al.* Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. *Nat Commun* 2024;**15**:5405.
Wen B, Freestone J, Riffle M *et al.* Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nat Methods* 2025;**22**:1454–63.
Wilkinson MD, Dumontier M, Aalbersberg IJJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
Wörheide MA, Krumsiek J, Kastenmüller G *et al.* Multi-omics integration in biomedical research—a metabolomics-centric review. *Anal Chim Acta* 2021;**1141**:144–62.
Young A, Röst H, Wang B *et al.* Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell* 2024;**6**:404–16.
Zeng W-F, Zhou X-X, Willems S *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat Commun* 2022;**13**:7238.
Zuo Z, Cao L, Nothia L-F *et al.* MS2Planner: improved fragmentation spectra coverage in untargeted mass spectrometry by iterative optimized data acquisition. *Bioinformatics* 2021;**37**:i231–i236.

Bioinformatics Advances, 2025, 00, 1–12
https://doi.org/10.1093/bioadv/vbaf301
Perspective