CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2025-0031

Vol. **x**, No. x, pp. 1-20
xx 2026

Computational Methods for Mass Spectrometry-Based Single-Cell Proteomics Data

Xiaoran Yu[1] and Zhixiang Lin[1,2,3,*]

[1] *Department of Statistics and Data Science, The Chinese University of Hong Kong, Hong Kong SAR, China.*
[2] *Shenzhen Loop Area Institute, Shenzhen 518000, China.*
[3] *CUHK Shenzhen Research Institute, Shenzhen 518000, China.*

Received 27 November 2025; Accepted 20 February 2026

Abstract. Mass spectrometry-based single-cell proteomics (MS-SCP) enables the quantification of protein abundance in individual cells, offering a molecular perspective on post-transcriptional regulation and heterogeneity that cannot be inferred from transcriptomic data alone. However, MS-SCP data exhibit high rates of missing values, batch effects, and low-input noise, which require tailored computational models. In this review, we examine computational developments across the MS-SCP pipeline, covering protein identification and quantification, public repositories, data enhancement, and downstream analysis. We emphasize algorithms that integrate statistical modeling and deep learning for identification, quantification, and the joint correction of missingness and batch effects. We highlight the unique role of deep learning in modeling non-linear batch-dependent effects and learning robust protein representations from sparse, high-dimensional MS-SCP data. Finally, we outline future directions for method developers, including the incorporation of biological priors, the construction of abundance-level foundation models, the curation of single-cell perturbation datasets, and the integration of proteomic information with spatial and multimodal data.

AMS subject classifications: 92-02, 92C40, 68T07

Key words: Single-cell proteomics, mass spectrometry, computational methods, deep learning, data enhancement.

1 Introduction

Proteins are the primary effectors of cellular function, acting as enzymes, structural components, signaling molecules, and regulators of metabolism and environmental interactions [20]. Unlike nucleic acids, which mainly encode genetic potential, proteins display dynamic abundance, post-translational modifications (PTMs), and interaction states that directly determine cellular phenotypes [35, 47].

*Corresponding author. *Email addresses:* zhixianglin@cuhk.edu.hk (Z. Lin), xryu@link.cuhk.edu.hk (X. Yu)

Despite substantial advances in protein modeling, including protein language models such as ESM-2 [14,29,40] and structure prediction tools such as AlphaFold [2,24,45], these approaches center on individual proteins and do not capture cell-to-cell heterogeneity or context-dependent variation.

Single-cell proteomics (SCP) extends proteome analysis to the level of individual cells, enabling the study of molecular mechanisms within heterogeneous cell states [1]. By directly measuring protein abundances in single cells, SCP can reveal differentiation trajectories, regulatory transitions, and perturbation responses that are averaged out in bulk assays [35]. SCP provides information complementary to scRNA-seq, revealing that correlations between mRNA and protein levels are limited and context dependent, shaped by measurement reproducibility and pervasive post-transcriptional regulation [1, 31]. In biological contexts such as cancer heterogeneity and infectious diseases, SCP provides resolution into rare cell types and signaling pathways, facilitating biomarker discovery and mechanistic modeling [30, 50].

Two technical families currently support SCP. Antibody-based techniques such as CyTOF [3, 4] and CITE-seq [48] enable high-throughput multiplexed detection but are constrained by antibody specificity, cross-reactivity, and the need for predefined targets, typically assaying dozens to low hundreds of proteins per cell [18]. In contrast, mass spectrometry-based single-cell proteomics (MS-SCP) enables quantification of thousands of proteins per cell, achieving deep proteome coverage without prior knowledge of targets [46].

Recent years have seen rapid progress in MS-SCP. Advances in technology now enable quantification of more than 5,000 proteins per cell, with practical throughput reaching over 1,000 cells per day with multiplexed protocols [8]. Nonetheless, intrinsic limitations persist, including sample losses during preparation, challenges in detecting low-abundance proteins, and data sparsity such as high missing values from stochastic ion sampling [47]. This contrast between technological gains and quality issues highlights the central role of computational approaches in enhancing and analyzing MS-SCP data.

In this review, we focus on computational approaches for MS-SCP, tailored for method developers in computational biology. We first summarize current advances in MS-SCP technology (Section 2). We then organize key computational strategies for abundance data generation, data resources, enhancement to improve data quality, and downstream analysis to ensure biologically meaningful insights (Section 3). We discuss evaluation and reproducibility issues (Section 4). Finally, we outline future directions, including integrating biological priors for data enhancement, developing foundation models, and modeling perturbation proteomics and spatial proteomics (Section 5).

2 Current advances for MS-SCP

The evolution of MS-SCP shows rapid gains in proteome depth and throughput (Fig. 1). Early feasibility work (2018–2020) demonstrated that proteomic measurements could, in

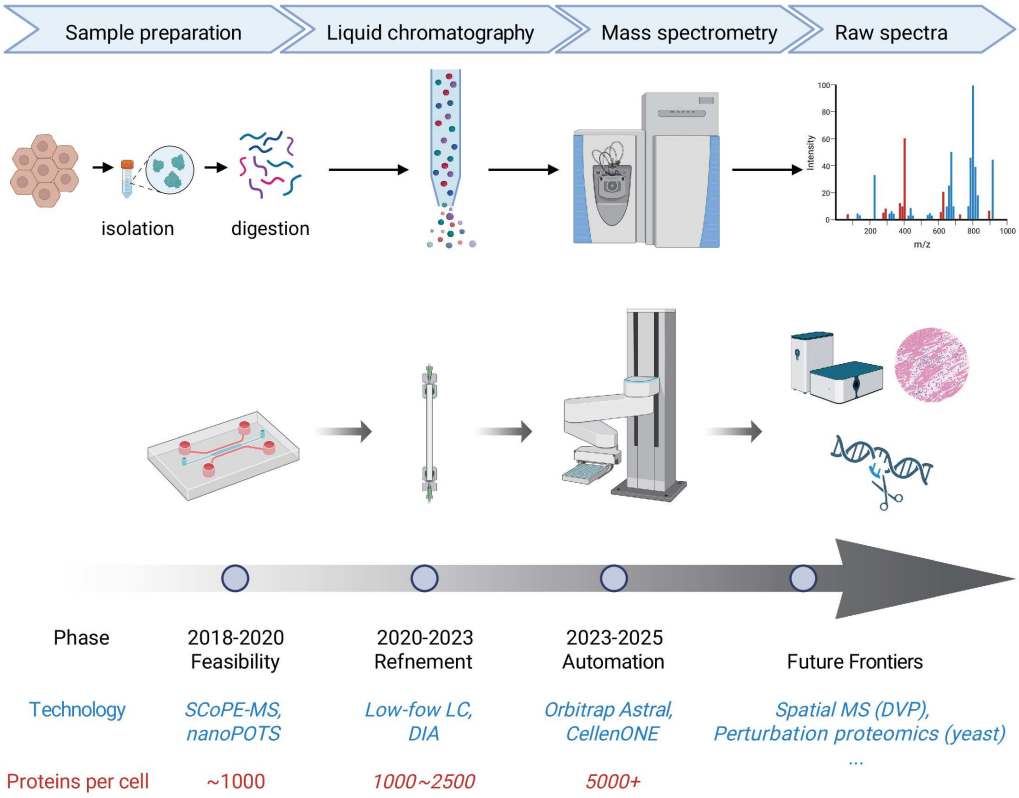

Figure 1: Overview of the MS-SCP experimental workflow and technological evolution. Top panel: The typical MS-SCP experimental pipeline. Single cells are isolated from tissue and lysed to extract proteins. Proteins are digested into peptides (sample preparation), which are then separated via liquid chromatography and ionized for mass spectrometry analysis. The resulting raw spectra serve as the stochastic and noisy input for computational identification and quantification. Bottom panel: The timeline of technological advances in MS-SCP from 2018 to the present. The field has evolved through three distinct phases: Feasibility (2018–2020), demonstrating proof-of-concept with methods like SCoPE-MS and nanoPOTS; Refinement (2020–2023), improving consistency via data-independent acquisition (DIA) and low-flow chromatography; and Automation (2023–2025), characterized by high-throughput instruments (e.g. Orbitrap Astral) and robotic platforms (e.g. CellenONE), pushing proteome coverage from $\sim$1,000 to over 5,000 proteins per single cell. Created in BioRender. X. Yu, 2026. `https://BioRender.com/2jnqph6`.

principle, be performed on individual cells. Pioneering methods such as SCoPE-MS [9] and nanoPOTS [65] established proof of concept by miniaturizing sample preparation to sub-200 nL volumes to reduce surface losses and enable single-cell analyses. These approaches identified $\sim$1,000-2,000 proteins per cell. Nevertheless, throughput remained modest and measurement variability was high. At that time, data-dependent acquisition (DDA), where only the most intense precursor ions from each scan are selected for fragmentation, was the prevailing mode. In practice, nanoPOTS workflows often analyze small pools of 10-100 cells rather than true single-cell runs.

From 2020 to 2023, methodological refinements in separation and ionization efficiency improved peptide recovery. Low-flow nanoLC (less than 100 nL/min) coupled with

electrospray ionization enhanced peptide separation and ion transmission efficiency [56]. A broader adoption of data-independent acquisition (DIA) [16], which fragments all precursors within predefined m/z isolation windows and thereby mitigates the stochastic precursor selection of DDA, enhanced quantitative consistency across runs, resulting in reduced variability and improved data completeness. In practice, these improvements typically delivered 1,000-2,500 proteins per single cell, with higher numbers in best-case studies. Advanced mass analyzers and refined ion optics further increased proteome coverage and quantification accuracy [7], yet missing values and batch effects remained prominent [5].

Since 2023, automation has shifted MS-SCP toward higher throughput. Instruments such as the Orbitrap Astral enabled MS/MS acquisition rates up to ~ 200 Hz [8, 21], and automated platforms such as cellenONE streamlined single-cell isolation and digestion to support scalable workflows [38]. Current label-free experiments report depths of $\sim$5,000 proteins per cell, with upper values around 5,300 in A549 cells, and ultra-low-input benchmarks detect $\sim$7,400 proteins from 250 pg HeLa digests. Throughput commonly reaches $\sim$50-80 single cells per day with short gradients and could be increased with multiplexed strategies at some cost to proteome depth. Multiplexed labeling also reduced inter-batch variability, with reported median CVs around 10-20% under specific setups.

Building on these advances, emerging technologies further expand MS-SCP. Spatial proteomics methods such as deep visual proteomics (DVP) integrate imaging and laser microdissection to map $\sim$1,700 proteins from a cell slice with subcellular resolution [34]. Large-scale perturbational studies, exemplified by proteome-wide analysis of single-gene perturbations in yeast with 3,308 nonessential gene knockouts, reveal global effects of individual perturbations on protein abundance and provide insight into the mechanisms of genetic perturbations [36].

Despite the progress, challenges persist for MS-SCP, including limited sample throughput for large studies, high rates of missing values, noisy spectra from low-input signals, the curse of dimensionality in high-dimensional data, and barriers to accessibility due to cost and specialized instrumentation. Addressing these limitations increasingly depends on computational approaches. To support such progress, there is a need for an ecosystem comprising reliable abundance data, standardized data resources, algorithms for batch reduction and imputation, and integrated pipelines for downstream analysis to foster computational innovation in MS-SCP.

3 Computational approaches for MS-SCP

3.1 Identification and quantification

In mass spectrometry-based single-cell proteomics (MS-SCP), identification and quantification refer to the computational conversion of raw spectra into a cell-by-protein abundance matrix. Each row corresponds to a single cell and each column to a protein, serving

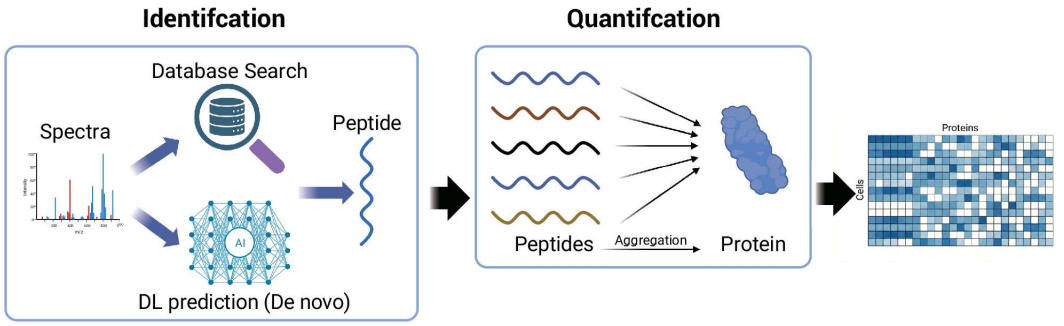

Figure 2: Computational workflow for protein identification and quantification in MS-SCP. Identification: Raw mass spectra are converted into peptide sequences via two primary strategies: database search, which matches spectra against a reference proteome; and de novo prediction, which predicts sequences directly from patterns using neural networks. Quantification: Identified peptide intensities are aggregated (e.g. via sum or linear modeling) to infer protein-level abundance, resulting in the final cell-by-protein quantification matrix used for downstream analysis. Created in BioRender. X. Yu, 2026. `https://BioRender.com/jmldcd0`.

as the input for downstream analyses such as clustering or differential abundance testing. The process involves identifying peptides from spectra and quantifying their abundances to infer protein levels (Fig. 2).

3.1.1 Identification

Peptide identification infers amino acid sequences from tandem-MS fragment spectra, which can be viewed as a pattern-recognition problem where the input is a set of fragment ion peaks and the output is a sequence of amino acids with associated confidence scores. In MS-SCP, spectra from single cells are often sparse and noisy due to low ion counts, making accurate identification analogous to reconstructing sequences from incomplete, high-variance signals.

Conventional approaches rely on database search pipelines (e.g. MaxQuant via its Andromeda search engine), which match observed spectra to in-silico-generated peptide candidates from a reference proteome and score the peptide-spectrum matches (PSMs) [10]. These pipelines commonly apply machine-learning-based re-scoring to improve PSM discrimination (for example, semi-supervised re-ranking of PSM features) [25]. A limitation of purely database-constrained searches is their dependence on the completeness of the reference: unexpected sequences (e.g. from proteoforms, alternative splicing, non-canonical translation) or unanticipated modifications can be missed unless explicitly modeled in the search space.

De novo sequencing methods bypass this dependence by predicting sequences directly from spectral patterns. Early deep learning models such as DeepNovo [51] combined convolutional neural networks with an LSTM decoder and beam search, integrating local dynamic programming constraints to ensure mass consistency.

More recently, transformer architectures further improved long-range dependency modeling in spectra. Casanovo [61] formulates de novo sequencing as sequence-to-sequence translation from peak lists to amino acids and achieves consistent peptide re-

covery across species. Extending to data-independent acquisition, Cascadia [42] adapts the transformer framework to handle DIA's multiplexed fragment signals and reports substantially improved performance on DIA datasets. For faster inference, TSARseqNovo [64] introduces a semi-autoregressive decoder with masking refinement, achieving up to $\sim 2\times$ speedups versus Casanovo (and larger gains over older models) while maintaining accuracy. Non-autoregressive designs such as π-PrimeNovo [63] incorporate precise mass-control decoding to boost efficiency and accuracy, with demonstrated advantages in challenging scenarios including phosphopeptide/low-abundance PTMs.

Beyond pure transformers, diffusion and graph formulations add robustness: InstaNovo [15] employs diffusion-based iterative refinement to improve coverage in large-scale de novo applications, and GraphNovo [32] uses a two-stage graph neural network to mitigate missing-fragmentation by constraining decoding with spectrum-graph paths. Finally, foundation-style spectrum encoders pre-trained on large scale MS/MS data show improved transfer to multiple downstream spectrum understanding tasks, suggesting a unifying route for spectrum representation learning [43].

The choice of identification strategy dictates the analytical scope. Database search engines assume the presence of a complete reference proteome, taking spectra and a FASTA file as input to output identified peptides with high specificity but limited discovery potential. Conversely, de novo methods require only raw spectra as input and assume learnable fragmentation patterns, enabling the discovery of novel variants and antibodies at the cost of higher computational demand and potential error propagation in noisy single-cell spectra.

To enable subsequent protein-level quantification, protein inference and grouping aggregate peptide evidence while resolving shared-peptide ambiguity. In practice, peptide-protein grouping rules such as MaxQuant's razor-peptide assignment [10] and probabilistic/Bayesian formulations that estimate protein-level presence from PSM/peptide posteriors [28] are widely used. In single-cell workflows, standardized R/Bioconductor frameworks (e.g. scp [17]) emphasize explicit false-discovery-rate control, peptide-to-protein aggregation, and per-cell quality checks as part of preprocessing.

3.1.2 Quantification

Quantification estimates peptide abundances from mass spectrometric signals and subsequently aggregates peptide intensities to derive protein-level quantities. The intensity of a peptide corresponds to the integrated ion signal detected along its chromatographic elution profile, obtained from precursor (MS1) or fragment (MS2) ion traces. Quantification algorithms integrate these signals over retention time and perform normalization across samples to ensure comparability.

Protein-level estimation aggregates peptide intensities belonging to the same protein. Common aggregation methods include top-N (summing the most intense peptides) and sum-all. These common methods are computationally simple but can propagate measurement noise when peptide detection is incomplete. Recent frameworks incorporate measurement uncertainty or model missingness explicitly. scPROTEIN [27] uses a multi-

task heteroscedastic regression model to estimate peptide-level mean and uncertainty and performs uncertainty-weighted aggregation, while scplainer [54] applies linear modeling to account for missing peptides and technical covariates, yielding interpretable protein estimates.

Together, these approaches infer protein abundances from spectral intensity, providing quantitative inputs for downstream analysis. However, all quantification remains an estimate of relative signal rather than ground-truth measurements, and its accuracy depends on spectrum quality, fragmentation completeness, and preprocessing choices.

3.2 Data resources

After quantification, single-cell proteomic datasets are commonly deposited in public repositories to support reproducibility and method development.

The ProteomeXchange consortium provides coordinated submission and dissemination for proteomics data, including MS-SCP studies[1]. Within this framework, PRIDE (EMBL-EBI) hosts raw MS files, processed results, and metadata; MassIVE (UCSD) supports large-scale datasets; and iProX (China) provides a web-based submission interface. These repositories ensure long-term accessibility but are not always organized for computational benchmarking, motivating the creation of developer-oriented databases.

Specialized databases improve accessibility for method development. The single-cell proteomic database (SPDB) integrates 143 datasets across 12 biotechnologies and 4 species, and provides unified downloads including preprocessed `.rds` files[2]. SingPro aggregates 211 studies with more than 625 million cells and over 16,000 proteins across species, tissues, and diseases, and offers raw and processed files in CSV/TXT formats with batch downloads[3]. The Slavov laboratory site focuses on MS-SCP methods such as SCoPE-MS, SCoPE2, pSCoPE, and plexDIA, and organizes raw and processed datasets by publication with protocols and computational tools[4].

Although not single-cell specific, ProteomicsDB aggregates bulk proteomics and perturbation datasets, including dose–response assays for drug–protein interactions[5]. Such large-scale resources highlight the potential of standardized curation for perturbational and multimodal SCP data. Table 1 summarizes representative features of commonly used repositories.

3.3 Data limitations

Although MS-SCP yields a cell-by-protein abundance matrix that can capture cellular heterogeneity, the data are affected by technical limitations in sample preparation, ion-

[1] `https://proteomecentral.proteomexchange.org/`
[2] (`https:// scproteomicsdb.com/`)
[3] `http://idrblab.org/singpro/`
[4] `https://scp.slavovlab.net/`
[5] `https://www.proteomicsdb.org/`

Table 1: MS-SCP data resources overview.

Feature	ProteomeXchange members	SPDB	SingPro	Slavov Lab	ProteomicsDB
Single-cell	✓	✓	✓	✓	×
Bulk	✓	×	×	×	✓
Raw data	✓	×	✓	✓	×
Abundance	✓	✓	✓	✓	✓
Data formats	Various	RDS	CSV, TXT	RAW, mzML, TXT, CSV	CSV, JSON, XML
Access method	Web portal, FTP, Aspera	Direct web download	Batch web download	Direct web download	Web interface, API

ization and acquisition that introduce inaccuracies and noise. This results in high rates of missing values and pronounced batch effects, which obscure biological signals and hinder downstream analyses.

Missing values stem from the stochastic proteomics workflow, including low peptide abundance below the instrument's limit of detection (LOD), ion suppression, incomplete precursor selection in data-dependent acquisition, and sample preparation losses [41]. These mechanisms give rise to two types of missingness: For low-sensitivity peptides, observations are often missing not at random (MNAR), whereas other losses may approximate missing completely at random (MCAR). Consequently, SCP datasets often exhibit missing rates over 50%, up to 90% in some studies, introducing biases if mishandled [26].

Batch effects stem from technical variables including instrument calibration, reagent lot differences, operator variation, sample multiplexing (for example TMT labeling), chromatography-gradient changes, ion-source fluctuations and differing solvent or instrument conditions [39]. These cause systematic abundance shifts, inflating variability, masking true differential expression, and leading to false heterogeneity or reduced statistical power in studies [13].

Missing values and batch effects are entangled. Missingness is often batch-dependent, which prevents reliable batch correction. Conversely, if batch effects are not corrected, they introduce bias into the imputation. For example, ComBat [23] batch correction assumes complete data and is unreliable when values are missing non-randomly across batches; on the other hand, kNN [52] imputation can be distorted by batch structure because nearest neighbors may cluster by batch rather than biology. Thus, they should be addressed jointly or with methods that tackle this dependency [22, 54].

Addressing these intertwined challenges requires computational strategies that explicitly account for the characteristics of MS-SCP data, which are introduced in the following section.

3.4 Data enhancement

Data enhancement refers to computational preprocessing and modeling procedures that improve the data quality of the cell-by-protein matrix, so that downstream biological analyses become more reliable. Computational strategies for data enhancement have evolved from adapting transcriptomic tools to developing proteomics-specific models. We categorize these approaches into three distinct frameworks (Fig. 3): Methods adapted from the analysis of transcriptomics, which transfer mature single-cell transcriptomics preprocessing ideas to proteomic abundance matrices, often implemented as a sequential workflow; joint linear and factorization models, which mathematically decouple biological variance from technical noise without explicit imputation; and deep learning frameworks, which utilize neural networks to learn non-linear representations from sparse inputs.

3.4.1 Methods adapted from the analysis of transcriptomics

Methods for analyzing single-cell RNA sequencing (scRNA-seq) data, such as kNN [52] imputation and ComBat [23] batch correction, have been directly applied to SCP data. kNN imputes missing values by averaging abundances from the k nearest cells using metrics like Euclidean distance, while ComBat employs empirical Bayes to model and adjust for batch-induced location and scale shifts.

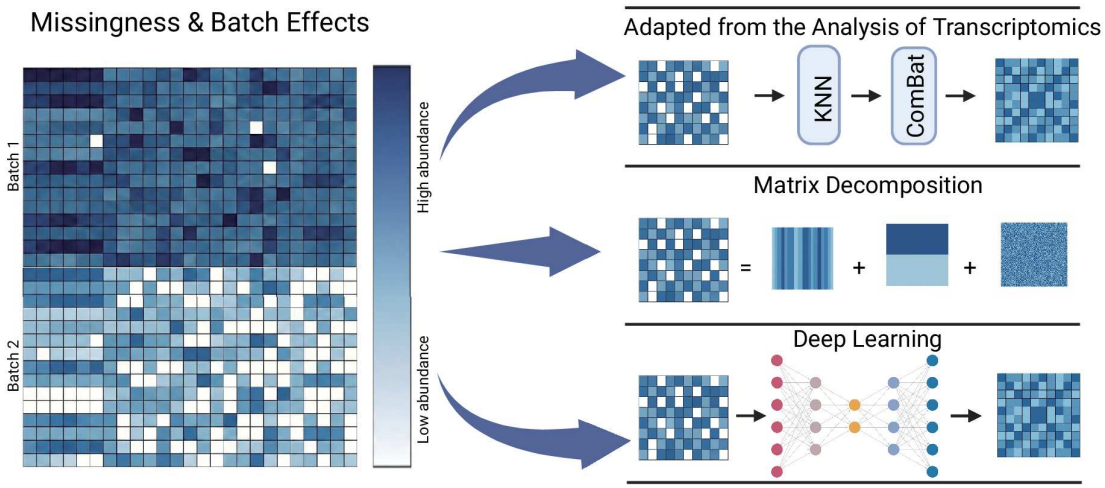

Figure 3: Strategies for MS-SCP data enhancement. Left panel: Illustration of intrinsic data challenges in MS-SCP, highlighting the entanglement of missing values (white blocks) and batch effects. Batch 2 exhibits systematically higher missing rates than Batch 1, indicating batch-dependent missingness consistent with a missing not at random (MNAR) mechanism. Right panel: Three computational paradigms for addressing these challenges, corresponding to the frameworks reviewed in Section 3.4. Top: Methods adapted from the analysis of transcriptomics, which repurpose transcriptomic preprocessing tools to perform imputation (e.g. kNN) and batch correction (e.g. ComBat), typically as a sequential workflow. Middle: Matrix decomposition, which decouples biological signal from technical variation and sparsity through explicit latent-factor modeling. Bottom: Deep Learning, which leverages neural networks to learn non-linear representations and reconstruct abundance profiles from sparse inputs. Created in BioRender. X. Yu, 2026. `https://BioRender.com/9g5vvf4`.

SCPline [19] builds on these two approaches by providing a user-friendly interactive R Shiny platform for SCP preprocessing. It integrates quality screening (e.g. threshold-based cell gating), centered log-ratio normalization, kNN imputation, and ComBat batch correction, with modular design and visualizations enabling user-customized workflows.

BIND [59] is a web platform for analyzing missing-value patterns in MS proteomics. It distinguishes biological versus technical NAs via frequency/pattern analysis, applies weighted kNN to impute values, and leverages NA-aware correlation to enhance protein–protein relationship discovery, providing interactive visualizations for exploratory analysis.

3.4.2 Linear and factorization methods

scplainer [54] leverages linear models as its core framework to perform variance analysis, differential abundance testing, and batch correction. Let $X \in \mathbb{R}^{n\times p}$ encode experimental factors (batches, conditions) and $Y \in \mathbb{R}^{n\times m}$ be log-transformed protein abundances. The model estimates coefficients $\beta \in \mathbb{R}^{m\times p}$ and residuals $E \in \mathbb{R}^{n\times m}$. Fitting directly to observed data partially reduces the impact of batch-dependent missingness without requiring imputation.

omicsGMF [44] unifies dimensionality reduction, batch correction, and imputation within a Gaussian matrix-factorization framework tailored to log-transformed intensities. Let $Y \in \mathbb{R}^{n\times m}$ denote the feature-by-sample log-intensities, and $X \in \mathbb{R}^{n\times p}$ denote the sample-level covariates. omicsGMF decomposes $Y \approx UV^\top + XB$, where U, V captures feature-level and sample-level biological structure respectively, and B parameterizes feature-specific coefficients for the covariates. To estimate the model parameters, omicsGMF minimizes the negative log-likelihood over observed entries with ridge regularization, thereby integrating batch correction into the factorization rather than applying it post hoc.

3.4.3 Deep learning methods

Deep learning methods model non-linear relationships in incomplete or noisy abundance data using neural networks, providing an alternative to parametric statistical models. These methods handle high-dimensional dependencies and data sparsity, although they often require more computation and present challenges for interpretability.

PIMMS [57] employs self-supervised deep models, including collaborative filtering (CF), denoising autoencoders (DAE), and variational autoencoders (VAE), for imputation in label-free quantitative MS proteomics data. CF exploits cross-sample similarity to impute the missing entries. DAE/VAE reconstruct the signals from noisy and incomplete inputs. The models are applicable at precursor, peptide, and protein-group levels and are provided in the pimms-learn Python package.

scPROTEIN [27] integrates heteroscedastic regression for peptide-level uncertainty estimation with graph contrastive learning for protein-level denoising, batch-aware representation, and cell embedding generation. It predicts per-peptide mean/variance for uncertainty-guided protein aggregation, builds kNN graphs from protein features,

Table 2: Comparison of data enhancement methods in SCP.

Method	SCPline	BIND	scplainer	omicsGMF	PIMMS	scPROTEIN
Model	kNN+ ComBat	Weighted kNN	Linear model	Matrix factorization	CF / DAE / VAE	Graph contrastive learning
Imputation	✓	✓	×	✓	✓	✓
Batch correction	✓	×	✓	✓	×	✓
Cell embedding	×	×	×	✓	×	✓
Protein embedding	×	×	×	✓	×	×
Assumption	Minimal	MNAR, MCAR	Gaussian residuals	Gaussian residuals Minimal	Minimal	Graph structure
Implementation	R / Shiny	Web server	R / Bio-conductor	R / Bio-conductor	Python	Python
Required inputs	Abundance +batch	Abundance	Abundance	Abundance +covariates	Abundance	Abundance
Typical outputs	Corrected matrix	NA-aware association	Adjusted effects	Denoised factors	Imputed matrix	Embedding +matrix
Potential limitations	Sequential bias	Exploratory only	Linear assumption	Likelihood sensitive	Black box	Black box, risk of over-smoothing

and applies contrastive objectives with topology/attribute denoising, mitigating batch effects without explicit batch labels and producing embeddings for clustering and annotation.

Synthesis and trade-offs. Choosing a data enhancement strategy requires balancing interpretability, assumptions, and computational complexity (Table 2). Linear models (e.g. scplainer) offer high interpretability and rigorous statistical inference but may fail to capture complex, non-linear batch effects. In contrast, deep learning frameworks (e.g. scPROTEIN, PIMMS) excel at modeling non-linear technical noise and learning robust cell embeddings, yet they often act as "black boxes" and require larger datasets to avoid overfitting. Factorization methods (e.g. omicsGMF) provide a middle ground, effectively decoupling biological signals from technical covariates through low-rank approximations, though their performance depends heavily on the validity of the chosen likelihood function (e.g. Gaussian vs. Poisson) for the specific data distribution. Notably, despite the growing interest in tailored frameworks, traditional pipelines based on kNN imputation and ComBat-style batch correction sometimes remain preferable in small-scale or low-throughput MS-SCP studies, where data are extremely sparse and model complexity is difficult to justify.

3.5 Downstream analysis

Downstream analysis in MS-SCP interprets processed protein-abundance data to identify cell populations, test for differential protein levels and reconstruct dynamic processes.

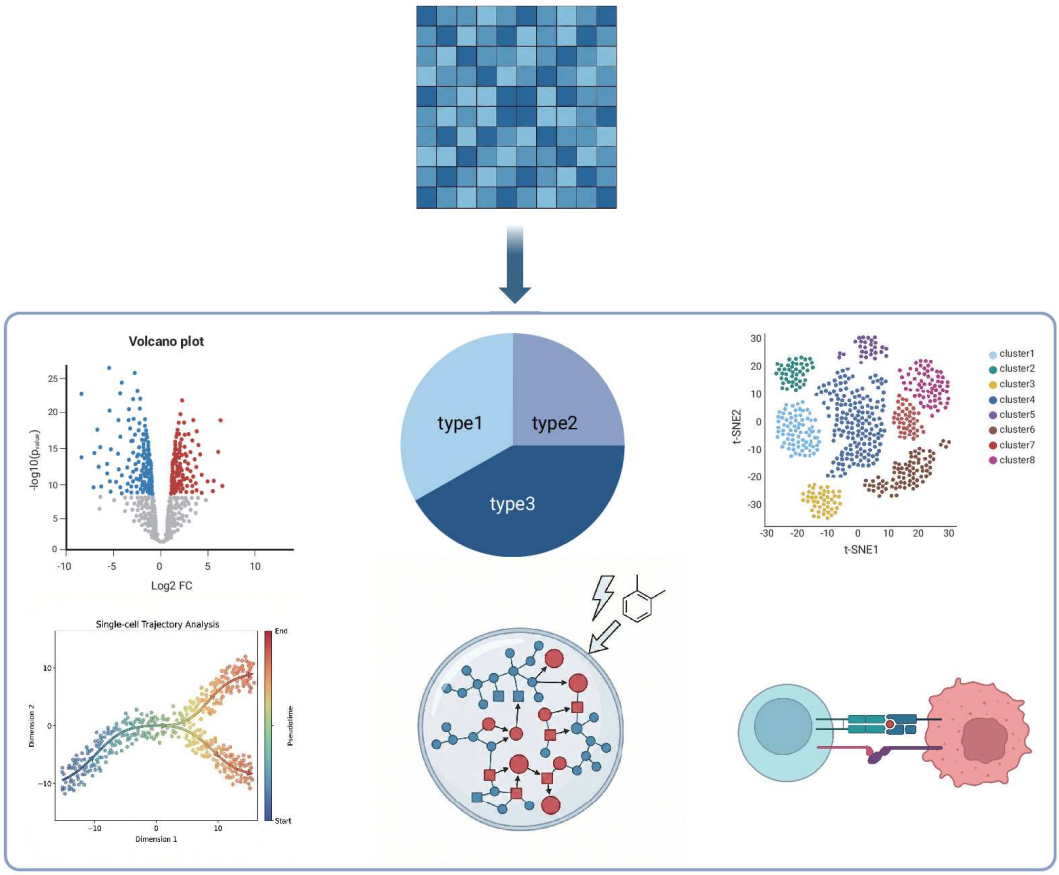

Figure 4: Overview of downstream analysis tasks in MS-SCP. Starting from the processed cell-by-protein abundance matrix (top), computational methods translate high-dimensional data into biological insights. Key analysis modules include: Top row: Differential abundance analysis (left) to identify significant protein changes; Cell composition analysis or deconvolution (center) to determine cell type proportions; and Dimensionality reduction and clustering (right) to define cell states and heterogeneity. Bottom row: Trajectory inference (left) to reconstruct dynamic differentiation processes; Perturbation analysis (center) to model protein network dynamics under drug or genetic interventions; and cell-cell communication inference (right), which reconstructs intercellular signaling based on protein-level receptor and downstream effector activities. Created in BioRender. X. Yu, 2026. `https://BioRender.com/m15bvri`.

These analyses characterize cellular heterogeneity but are limited by technical noise, missing observations, and systematic differences between bulk and single-cell measurements. Building on preprocessed data, this section outlines representative computational approaches (Fig. 4).

3.5.1 Differential abundance analysis

Differential abundance analysis aims to pinpoint proteins with significant changes across cell populations or conditions. In MS-SCP, variance instability and limited sensitivity for low-abundance proteins complicate testing. scplainer [54] addresses these challenges by

fitting linear models that include technical and biological covariates (e.g. batch, condition) and performing inference on model coefficients without explicit imputation, thereby reducing imputation-induced bias and avoiding obscuring biological variation.

3.5.2 Deconvolution methods

Deconvolution estimates the proportions of cell types in bulk proteomes using single-cell data as a reference. The scpDeconv [55] framework combines an autoencoder (to learn latent representations from bulk and pseudo-bulk data) with a domain adversarial network (to align distributions between bulk and single-cell), followed by a predictor for cell-type proportions. This strategy helps bridge bulk-single-cell distributional gaps and has been demonstrated across multiple species and proteomic technologies.

3.5.3 Dimensionality reduction and clustering

Dimensionality reduction and clustering reveal cell types/states by learning low-dimensional representations while mitigating batch effects and sparsity. scPROTEIN [27] constructs a cell-similarity graph from protein profiles and applies graph-contrastive training to derive cell embeddings following uncertainty-weighted peptide-to-protein aggregation. omicsGMF [44] factorizes the abundance matrix into low-rank biological structure plus covariate effects, providing loadings that support downstream clustering (e.g. k-means) and integration across batches.

3.5.4 Trajectory inference

Trajectory inference orders cells along pseudotime to reconstruct dynamic processes (e.g. differentiation). Many studies reuse established scRNA-seq pipelines for SCP data. In practice, preprocessing interfaces such as SCPline [19] perform quality control and dimensionality reduction (PCA/UMAP) before trajectory inference. A recent perspective emphasizes that sample throughput remains a primary limitation for broad biological use cases, which motivates methods that balance measurement depth with sample throughput and emphasize scalable inference under noisy conditions [35].

3.5.5 Perturbation analysis

Perturbation analysis quantifies causal changes in protein expression and PTMs under drugs or genetic alterations, informing mechanisms and network models. Dose-response proteomics such as decryptE fit proteome-wide dose-response curves for the discovery of mechanism of action (MoA) [62]. Time/dose-resolved PTM profiling such as decryptM has revealed mechanisms such as rituximab killing $CD20^+$ B cells by overactivating BCR signaling [62]. Proteome-scale drug atlases further support network inference (e.g. correlational maps linking JP1302 to FACT inhibition and histone H1 degradation) [33]. ProteinTalks is a neural ODE–based perturbation proteomics foundation model trained on more than 38M perturbed protein measurements from breast cancer cell lines to learn transferable, cell-contextual protein network dynamics [49]. ProteinTalks improves drug

response and synergy prediction and highlights proteins associated with resistance. It also provides a model for estimating how protein networks change under drug perturbation.

3.5.6 Cell-cell communication and signaling inference

While transcriptomic approaches typically infer cell-cell communication (CCC) by matching ligand and receptor expression, MS-SCP provides complementary information by directly measuring downstream signaling proteins that execute these signals. However, due to the high sparsity of MS-SCP data, low-abundance membrane receptors are often difficult to detect reliably. To address this limitation, recent computational strategies shift from direct ligand-receptor matching toward network-based pathway inference. These methods use biological prior knowledge, such as protein-protein interaction networks (e.g. OmniPath [53]), to propagate signals from observed downstream proteins and infer upstream receptor or pathway activity. This idea is conceptually similar to NicheNet [6], but operates at the protein level, enabling the reconstruction of intercellular signaling even when receptor measurements are incomplete.

4 Critical assessment and limitations

MS-SCP has moved from feasibility studies to routine workflows that quantify thousands of proteins per cell and reveal heterogeneity that cannot be inferred from RNA measurements alone. Advances in spectrum interpretation and peptide-to-protein aggregation have improved identification accuracy and stabilized protein quantification. Methods for data enhancement based on linear models, matrix factorization, and self-supervised learning help reduce missing values and batch effects while preserving biological information needed for downstream analyses.

However, several critical bottlenecks limit the reliability and broader adoption of computational methods. This section synthesizes the key methodological constraints and evaluation gaps that define the current state of the field.

4.1 Methodological constraints in identification

A primary limitation in protein identification is the trade-off between sensitivity and false discovery rates (FDR) in low-input spectra. De novo sequencing, while promising for identifying variable sequences, often suffers from error propagation in single-cell data where fragment ion series are incomplete. Current transformers (e.g. Casanovo) trained on bulk data may hallucinate amino acids when applied to noisy single-cell spectra without specifically adapted confidence scoring.

4.2 The entanglement of missingness and batch effects

A pervasive challenge in MS-SCP is the entanglement of missing values and batch effects. Missingness is frequently Missing Not At Random (MNAR) and batch-dependent,

as peptides are more likely to be absent in batches with lower instrument sensitivity. This dependency complicates data enhancement across all major methodological paradigms. Transcriptomics-adapted methods typically apply imputation and batch correction sequentially, relying on assumptions (e.g., batch-free neighbors or complete observations) that are often violated in MS-SCP. Joint linear and factorization models alleviate this issue through unified modeling of biological and technical effects, but still depend on simplified distributional assumptions. Deep learning frameworks provide greater modeling flexibility, yet without explicit constraints they may over-smooth biological variation or learn batch-specific patterns, particularly in sparse or unbalanced datasets.

4.3 Evaluation gaps

Finally, the field lacks standardized benchmarks for evaluating data enhancement methods in MS-SCP. Imputation and denoising methods are commonly assessed by artificially masking observed entries and measuring reconstruction accuracy. However, such synthetic masking schemes may not reflect the structured, batch-dependent MNAR missingness in real MS-SCP, making the resulting performance estimates unreliable and hard to compare across studies. Establishing benchmark datasets with known composition (e.g. spike-in or mixed-species designs) and adopting evaluation criteria tailored to MS-SCP missingness mechanisms are therefore essential for rigorous and reproducible assessment.

5 Perspective and outlook

Looking ahead, we outline four areas for coordinated development: (i) biology-informed priors; (ii) foundation models for MS-SCP abundance data; (iii) standardized single-cell perturbation corpora; and (iv) spatial and multimodal grounding. Advances in these areas could improve data quality, strengthen downstream analyses, and increase model interpretability by linking computational predictions to known biological pathways and protein complexes.

Biology-informed priors. We recommend embedding biological knowledge, including protein–protein interactions, pathway topology, complex stoichiometry, and subcellular organization, as explicit priors in aggregation, denoising, and representation-learning models. In practice, this can be realized via graph learning, structured regularization within linear or factorization models, or topology-aware inductive biases in neural networks. Incorporating such information helps share signals across related proteins and imposes pathway-level constraints that stabilize representations when data are missing.

Foundation models for abundance data. A potential direction is to pretrain models directly on MS-SCP abundance data, and to fine-tune these models for specific objectives such as missing-value prediction, contrastive batch alignment, and perturbation-

response modeling. This pretraining-finetuning paradigm has already succeeded in single-cell transcriptomics (e.g. scGPT, Geneformer, scBERT) [11, 12, 60]. Recent efforts relevant to MS–SCP include building a foundation model for de novo sequencing [43] and a neural ODE-based foundation model for bulk-MS perturbation modeling (e.g. ProteinTalks [49]). Establishing foundation models could provide transferable representations that generalize to new datasets and support downstream tasks including clustering, differential abundance analysis, and trajectory inference. However, to date, no foundation model has been developed specifically for MS-SCP abundance data, largely due to limitations in data scale. While transcriptomic foundation models (e.g. scGPT, Geneformer) leverage training corpora exceeding 30 million cells, the aggregate volume of public MS-SCP data currently resides in the range of hundreds of thousands (10^5), with typical individual experiments profiling 10^3-10^4 cells. Building effective foundation models for proteomics will require substantially expanded and standardized MS-SCP datasets, and such data resources are currently insufficient.

Single-cell perturbation proteomics. Perturbation experiments and virtual-cell modeling support mechanistic analysis, causal inference, and predictive simulations of drug and genetic interventions. At present, MS-SCP perturbation datasets remain limited in size and number. This contrasts with single-cell transcriptomics, where large and well-curated resources such as scPerturb [37] and PerturBase [58] have enabled systematic benchmarking. We advocate coordinated efforts to generate MS-based single-cell perturbation datasets, particularly those involving genetic perturbations in human cells, as these perturbations are widely used to probe causal mechanisms but remain underrepresented in current MS-SCP resources.

Spatial proteomics and multimodal integration. MS-SCP can be jointly analyzed with spatial proteomics and other single-cell modalities to provide information that MS-SCP alone cannot capture. Spatial proteomics retains microanatomical context and reveals how protein abundance varies across tissue structures, complementing the cell-level measurements from MS-SCP. Additional modalities such as scRNA-seq, spatial transcriptomics, and scATAC-seq measure transcriptional programs, spatial expression patterns, and chromatin accessibility. Integrating these modalities enables the mapping of protein-level variation to transcriptional and chromatin states, uncovering cell states, microenvironmental niches and regulatory interactions that are not detectable from any single assay in isolation.

Acknowledgments

The work of Z. Lin is supported in part by the Chinese University of Hong Kong Science Faculty's Collaborative Research Impact Matching Scheme (Grant CRIMS 4620033), by the 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund (Grant 2025A0505000057) and by the Research Grants Council, University Grants Committee (Grants GRF 14301120, GRF 14300923).

References

[1] R. Ahmad and B. Budnik, *A review of the current state of single-cell proteomics and future perspective*, Anal. Bioanal. Chem., 415:6889–6899, 2023.

[2] M. Baek et al., *Accurate prediction of protein structures and interactions using a three-track neural network*, Science, 373:871–876, 2021.

[3] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, *Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry*, Anal. Chem., 81:6813–6822, 2009.

[4] S. C. Bendall et al., *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum*, Science, 332:687–696, 2011.

[5] H. Boekweg and S. H. Payne, *Challenges and opportunities for single-cell computational proteomics*, Mol. Cell. Proteom., 22(4):100518, 2023.

[6] R. Browaeys, W. Saelens, and Y. Saeys, *Nichenet: Modeling intercellular communication by linking ligands to target genes*, Nat. Methods, 17:159–162, 2020.

[7] A.-D. Brunner et al., *Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation*, Mol. Syst. Biol., 18:e10798, 2022.

[8] J. A. Bubis et al., *Challenging the astral mass analyzer to quantify up to 5,300 proteins per single cell at unseen accuracy to uncover cellular heterogeneity*, Nat. Methods, 22:510–519, 2025.

[9] B. Budnik, E. Levy, G. Harmange, and N. Slavov, *SCoPE-MS: Mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation*, Genome Biol., 19(1):161, 2018.

[10] J. Cox and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*, Nat. Biotechnol., 26:1367–1372, 2008.

[11] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, *scGPT: Toward building a foundation model for single-cell multi-omics using generative AI*, Nat. Methods, 21:1470–1480, 2024.

[12] Z. Cui, T. Xu, J. Wang, Y. Liao, and Y. Wang, *GeneFormer: Learned gene compression using transformer-based context modeling*, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 8035–8039, 2024.

[13] J. Čuklina et al., *Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial*, Mol. Syst. Biol., 17:e10240, 2021.

[14] A. Elnaggar et al., *ProtTrans: Toward understanding the language of life through self-supervised learning*, IEEE Trans. Pattern Anal. Mach. Intell., 44:7112–7127, 2021.

[15] K. Eloff et al., *InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments*, Nat. Mach. Intell., 7:565–579, 2025.

[16] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis*, Mol. Cell. Proteomics, 11(6):O111.016717, 2012.

[17] S. Grégoire, C. Vanderaa, S. P. dit Ruys, C. Kune, G. Mazzucchelli, D. Vertommen, and L. Gatto, *Standardized workflow for mass-spectrometry-based single-cell proteomics data processing and analysis using the scp package*, in: Mass Spectrometry Based Single Cell Proteomics. Methods in Molecular Biology, Vol. 2817, Humana, 177–220, 2024.

[18] S. M. Guldberg, T. L. H. Okholm, E. E. McCarthy, and M. H. Spitzer, *Computational methods for single-cell proteomics*, Annu. Rev. Biomed. Data Sci., 6:47–71, 2023.

[19] S. Guo, S. Zhou, G. Wang, and F. Wang, *SCPline: An interactive framework for the single-cell proteomics data preprocessing*, Brief. Bioinform., 26(3):bbaf256, 2025.
[20] T. Guo, J. A. Steen, and M. Mann, *Mass-spectrometry-based proteomics: From single cells to clinical applications*, Nature, 638:901–911, 2025.
[21] U. H. Guzman et al., *Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition*, Nat. Biotechnol., 42:1855–1866, 2024.
[22] L. Heumos et al., *Best practices for single-cell analysis across modalities*, Nat. Rev. Genet., 24:550–572, 2023.
[23] W. E. Johnson, C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical bayes methods*, Biostatistics, 8:118–127, 2007.
[24] J. Jumper et al., *Highly accurate protein structure prediction with alphafold*, Nature, 596:583–589, 2021.
[25] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, *Semi-supervised learning for peptide identification from shotgun proteomics datasets*, Nat. Methods, 4:923–925, 2007.
[26] W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh, *Dealing with missing values in proteomics data*, Proteomics, 22:2200092, 2022.
[27] W. Li, F. Yang, F. Wang, Y. Rong, L. Liu, B. Wu, H. Zhang, and J. Yao, *scPROTEIN: A versatile deep graph contrastive learning framework for single-cell proteomics embedding*, Nat. Methods, 21:623–634, 2024.
[28] Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng, and H. Tang, *A Bayesian approach to protein inference problem in shotgun proteomics*, J. Comput. Biol., 16:1183–1193, 2009.
[29] Z. Lin et al., *Evolutionary-scale prediction of atomic-level protein structure with a Language model*, Science, 379:1123–1130, 2023.
[30] J. Liu, C. Bao, J. Zhang, Z. Han, H. Fang, and H. Lu, *Artificial intelligence with mass spectrometry-based multimodal molecular profiling methods for advancing therapeutic discovery of infectious diseases*, Pharmacol. Ther., 263:108712, 2024.
[31] M. S. Mansuri, K. Williams, and A. C. Nairn, *Uncovering biology by single-cell proteomics*, Commun. Biol., 6:381, 2023.
[32] Z. Mao, R. Zhang, L. Xin, and M. Li, *Mitigating the missing-fragmentation problem in de novo peptide with a two-stage graph-based deep learning model*, Nat. Mach. Intell., 5:1250–1260, 2023.
[33] D. C. Mitchell, M. Kuljanin, J. Li, J. G. Van Vranken, N. Bulloch, D. K. Schweppe, E. L. Huttlin, and S. P. Gygi, *A proteome-wide atlas of drug mechanism of action*, Nat. Biotechnol., 41:845–857, 2023.
[34] A. Mund et al., *Deep visual proteomics defines single-cell identity and heterogeneity*, Nat. Biotechnol., 40:1231–1240, 2022.
[35] A. A. Nitz, J. H. Giraldez Chavez, Z. G. Eliason, and S. H. Payne, *Are we there yet? Assessing the readiness of single-cell proteomics to answer biological hypotheses*, J. Proteome Res., 24:1482–1492, 2024.
[36] M. Öztürk, A. Freiwald, J. Cartano, R. Schmitt, M. Dejung, K. Luck, B. Al-Sady, S. Braun, M. Levin, and F. Butter, *Proteome effects of genome-wide single gene perturbations*, Nat. Commun., 13:6153, 2022.
[37] S. Peidli et al., *scPerturb: Harmonized single-cell perturbation data*, Nat. Methods, 21:531–540, 2024.
[38] V. Petrosius et al., *Quantitative label-free single-cell proteomics on the orbitrap astral MS*, Mol. Cell. Proteomics, 24(6):100982, 2025.
[39] S.-X. Phua, K.-P. Lim, and W. W.-B. Goh, *Perspectives for better batch effect correction in mass-spectrometry-based proteomics*, Comput. Struct. Biotechnol. J., 20:4369–4375, 2022.

[40] A. Rives et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*, Proc. Natl. Acad. Sci. USA, 118:e2016239118, 2021.
[41] X. Sanchez-Avila, R. M. de Oliveira, S. Huang, C. Wang, and R. T. Kelly, *Trends in mass spectrometry-based single-cell proteomics*, Anal. Chem., 97:5893–5907, 2025.
[42] J. Sanders, B. Wen, P. A. Rudnick, R. S. Johnson, C. C. Wu, M. Riffle, S. Oh, M. J. MacCoss, and W. S. Noble, *A transformer model for de novo sequencing of data-independent acquisition mass spectrometry data*, Nat. Methods, 22(7):1447–1453, 2025.
[43] J. Sanders, M. Yilmaz, J. H. Russell, W. Bittremieux, W. E. Fondrie, N. M. Riley, S. Oh, and W. S. Noble, *Foundation model for mass spectrometry proteomics*, arXiv:2505.10848, 2025.
[44] A. Segers, C. Castiglione, C. Vanderaa, L. Martens, D. Risso, and L. Clement, *omicsGMF: A multi-tool for dimensionality reduction, batch correction and imputation applied to bulk-and single cell proteomics data*, bioRxiv, 2025, doi: `https://doi.org/10.1101/2025.03.24.644996`
[45] A. W. Senior et al., *Improved protein structure prediction using potentials from deep learning*, Nature, 577:706–710, 2020.
[46] L. R. Sinn and V. Demichev, *Entering the era of deep single-cell proteomics*, Nat. Methods, 22:459–460, 2025.
[47] N. Slavov, *Single-cell proteomic technologies: Tools in the quest for principles*, arXiv:2506.18198, 2025.
[48] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert, *Simultaneous epitope and transcriptome measurement in single cells*, Nat. Methods, 14:865–868, 2017.
[49] R. Sun et al., *A perturbation proteomics-based foundation model for virtual cell construction*, bioRxiv, 2025, doi: `https://doi.org/10.1101/2025.02.07.637070`
[50] Y. Sun et al., *Strategic priorities for transformative progress in advancing biology with proteomics and artificial intelligence*, arXiv:2502.15867, 2025.
[51] N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li, *De novo peptide sequencing by deep learning*, Proc. Natl. Acad. Sci. USA, 114:8247–8252, 2017.
[52] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, *Missing value estimation methods for dna microarrays*, Bioinformatics, 17:520–525, 2001.
[53] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, *Omnipath: Guidelines and gateway for literature-curated signaling pathway resources*, Nat. Methods, 13:966–967, 2016.
[54] C. Vanderaa and L. Gatto, *scplainer: Using linear models to understand mass spectrometry-based single-cell proteomics data*, Genome Biol., 26:237, 2025.
[55] F. Wang, F. Yang, L. Huang, W. Li, J. Song, R. B. Gasser, R. Aebersold, G. Wang, and J. Yao, *Deep domain adversarial neural network for the deconvolution of cell type mixtures in tissue proteome profiling*, Nat. Mach. Intell., 5:1236–1249, 2023.
[56] K. G. Webber, S. Huang, H.-J. L. Lin, T. L. Hunter, J. Tsang, D. Jayatunge, J. L. Andersen, and R. T. Kelly, *Gradient-elution nanoflow liquid chromatography without a binary pump: Smoothed step gradients enable reproducible, sensitive, and low-cost separations for single-cell proteomics*, Mol. Cell. Proteomics, 23(12):100880, 2024.
[57] H. Webel, L. Niu, A. B. Nielsen, M. Locard-Paulet, M. Mann, L. J. Jensen, and S. Rasmussen, *Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning*, Nat. Commun., 15:5405, 2024.
[58] Z. Wei, D. Si, B. Duan, Y. Gao, Q. Yu, Z. Zhang, L. Guo, and Q. Liu, *Perturbase: A comprehensive database for single-cell perturbation data analysis and visualization*, Nucleic Acids Res., 53:D1099–D1111, 2025.

[59] G. Weiheng, J. Wenyi, Z. Jieyi, P. Yilin, W. Rui, Z. Jian, F. Xikang, C. Lingxi, and Z. Liang, *Biologically informative NA deconvolution (BIND) excavates hidden features of the proteome from missing values in large-scale datasets*, bioRxiv, 2025, doi: `https://doi.org/10.1101/2025.06.19.660508`

[60] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao, *scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data*, Nat. Mach. Intell., 4:852–866, 2022.

[61] M. Yilmaz, W. E. Fondrie, W. Bittremieux, C. F. Melendez, R. Nelson, V. Ananth, S. Oh, and W. S. Noble, *Sequence-to-sequence translation from mass spectra to peptides with a transformer model*, Nat. Commun., 15:6427, 2024.

[62] J. Zecha et al., *Decrypting drug actions and protein modifications by dose- and time-resolved proteomics*, Science, 380:93–101, 2023.

[63] X. Zhang et al., *π-PrimeNovo: An accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing*, Nat. Commun., 16:267, 2025.

[64] Y. Zhao, S. Wang, J. Huang, B. Meng, D. An, X. Fang, Y. Wei, and X. Dai, *A transformer-based semi-autoregressive framework for high-speed and accurate de novo peptide sequencing*, Commun. Biol., 8:234, 2025.

[65] Y. Zhu et al., *Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells*, Nat Commun., 9(1):882, 2018.