Genome Biology

SOFTWARE

Open Access

Full-DIA enables complete single-cell proteomics from diaPASEF using deep learning

Jian Song[1,2*], Amanda Momenzadeh[2], Hebin Liu[3], Chengpin Shen[3*], Jesse G. Meyer[2*] and Xiaohui Wu[1*]

*Correspondence: songjian2022@suda.edu.cn; issac.shen@omicsolution.com; Jesse.Meyer@cshs.org; xhwu@suda.edu.cn

[1] Cancer Institute, Suzhou Medical College, Soochow University, Suzhou 215000, China
[2] Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[3] Shanghai Omicsolution Co., Ltd., Shanghai 200000, China

Abstract

diaPASEF improves ion utilization and sensitivity by synchronizing quadrupole isolation with trapped ion mobility separation, making it suitable for single-cell proteomics. We present Full-DIA, a deep learning–driven software that enhances proteome coverage, quantitative accuracy, and analysis speed over DIA-NN for single-cell diaPASEF data. Notably, Full-DIA generates a missing-value-free protein matrix under stringent global FDR control, enabling downstream analyses without data gaps. Applied to LPS-treated and cell-cycle datasets, this matrix yields pathway enrichment results with fewer off-target and more biologically relevant pathways. Full-DIA highlights the potential of deep learning for four-dimensional diaPASEF analysis and offers a solution to missing values.

Background

Mass spectrometry (MS)-based proteomics has become a cornerstone of systems biology and translational research, owing to its capability to deliver unbiased, quantitative measurements of thousands of proteins across samples [1]. Among the various MS acquisition strategies, data-independent acquisition (DIA) has emerged as a leading approach, valued for its reproducibility, depth of coverage, and suitability for large-scale studies [2, 3]. By systematically fragmenting all precursors within predefined *m/z* windows, DIA avoids the stochastic nature of data-dependent acquisition (DDA) and enables consistent peptide detection. However, conventional DIA methods are limited by their low ion sampling efficiency—typically only 1–5% of the total ion current is captured during a cycle—because each precursor is sampled only once per cycle, while the rest that is not isolated is discarded [4]. Narrow-window DIA improves spectral specificity but does so at the cost of further reducing ion utilization [5, 6]. In contrast, methods like plexDIA [7] increase ion sampling by using broader isolation windows—typically three or four per cycle in single-cell experiments—allowing approximately one-third of all ions to be analyzed, albeit with reduced within-spectrum selectivity [8]. Alternatively, diaPASEF [4] (data-independent acquisition with parallel

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

accumulation–serial fragmentation) achieves substantially higher ion utilization while maintaining spectral specificity by synchronizing quadrupole isolation windows with trapped ion mobility separation. diaPASEF dynamically targets precursors along the *m/z*–mobility diagonal, allowing near-complete sampling of the ion cloud during each cycle. This design enables a dramatic increase in ion utilization—up to 100% in some configurations—resulting in higher signal-to-noise ratios and improved sensitivity. These features make diaPASEF ideally suited for challenging applications such as single-cell proteomics, where limited input material often compromises identification depth and quantification precision.

While diaPASEF substantially enhances ion sampling efficiency, translating these rich four-dimensional data into a high-quality quantitative protein matrix (samples × proteins) for downstream single-cell analysis remains challenging. This holds true even for the widely used tool, DIA-NN [9, 10], which is free for academic use and offers notable advances in speed and sensitivity. Major obstacles include maintaining identification depth under stringent false discovery rate (FDR) control at the global protein level [11], ensuring accurate and precise quantification, minimizing missing values, and accelerating data processing to match the pace of high-throughput acquisition. Overcoming these challenges requires novel software solutions that can fully exploit the ion mobility dimension for denoising and scoring, while producing robust sample-by-protein matrices amenable to integrative analyses and biological interpretation.

Here, we present Full-DIA, a software framework designed to address the above challenges in single-cell diaPASEF data. Full-DIA enhances peptide and protein identification by employing neural network–based scoring on two-dimensional coelution consistency and spectral similarity, improves quantification accuracy through autoencoder-based ion intensity correction, and accelerates analysis via GPU-optimized computation. Furthermore, it produces complete, missing-value-free protein matrices that enable faithful downstream single-cell data integration and biological interpretation. We evaluated Full-DIA on multiple single-cell and bulk datasets, demonstrating substantial gains in identification depth, quantitative precision, and processing efficiency, thereby providing a scalable and generalizable solution for high-throughput single-cell diaPASEF proteomics.

Results

The workflow of Full-DIA

As illustrated in Fig. 1a, Full-DIA performs the peptide-centric [12] analysis per run, assigning a discriminative score and extracting fragment ion intensities for each target peptide. It then calculates the global FDR across runs. For peptides passing the global FDR threshold (e.g., 1%), Full-DIA retrieves fragment intensities across all runs to construct a complete sample-by-fragment-ion matrix. This matrix is denoised using a self-supervised autoencoder, from which peptide and protein matrices are subsequently derived.

Full-DIA automatically optimizes search parameters, including retention time windows, ion mobility, and mass tolerances, and utilizes GPU-accelerated Python to ensure efficient analysis from raw diaPASEF data. Full-DIA uses two deep learning models, DeepProfile (Fig. 1b) and DeepMall (Fig. 1c), to compute 2D (retention time + ion mobility) coelution consistency and spectral similarity, respectively (see

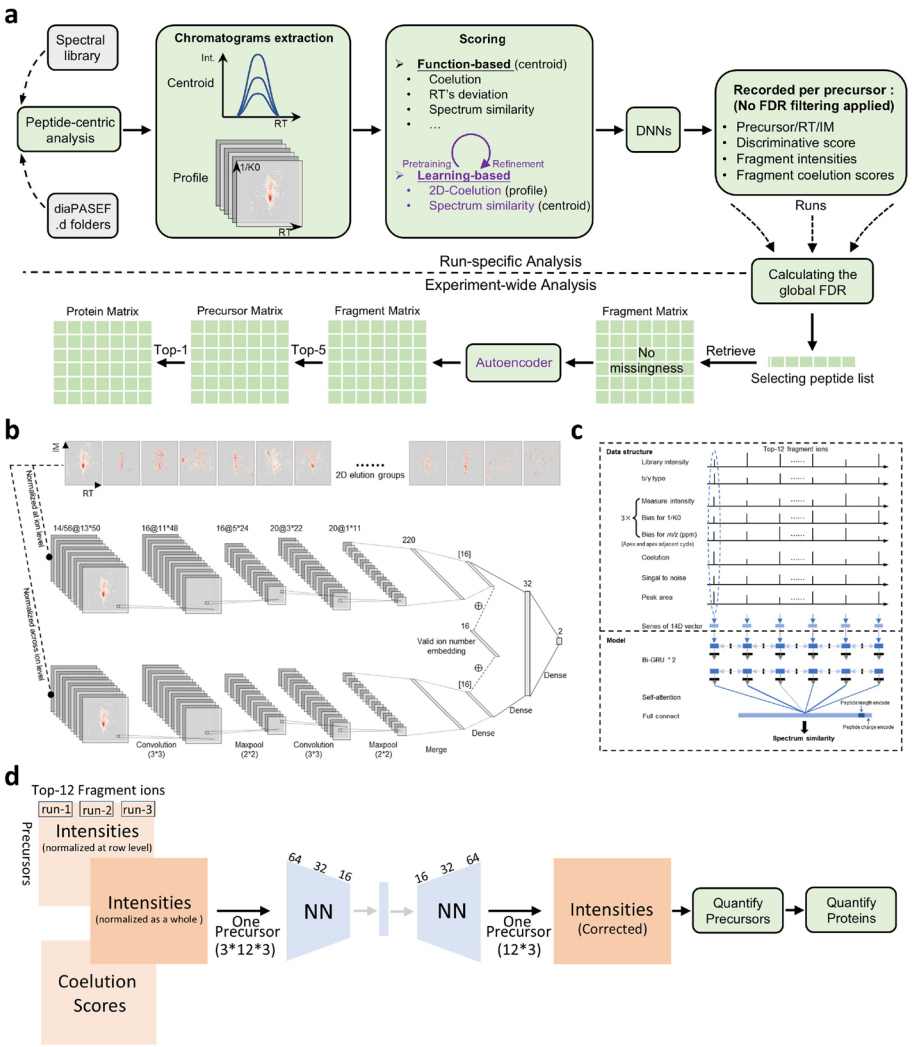

Fig. 1 Overview of Full-DIA and the deep learning models it employs. **a** Full-DIA first performs peptide-centric analysis on centroided and profile chromatograms, integrating functional and learned scores. It retains fragment intensities from each run regardless of q-values, enabling construction of a complete intensity matrix. **b** DeepProfile scores 2D coelution consistency (retention time x ion mobility) using both pre-trained and fine-tuned models. **c** DeepMall evaluates fragment similarity by comparing observed and expected intensities, weighted by ion-specific features like mass accuracy and mobility deviation. **d** DeepQuant autoencoder corrects fragment ion intensities using two normalized intensity matrices and a coelution score matrix; output is a denoised fragment intensity matrix. The figure uses 12 fragment ions and 3 runs as an example

Methods for details). These learned scores are combined with conventional functional features (791 scores in total, see Additional file 1: Table 1) to distinguish target peptides from decoy peptides. DeepProfile operates on 2D elution groups derived from four-dimensional data cubes, preserving chromatogram and mobility patterns. DeepMall compares observed vs. expected fragment intensities, weighted by ion-specific attributes such as mass accuracy and mobility deviation.

Full-DIA determines the most likely elution group for each target peptide and saves all fragment ion intensities, regardless of the run-specific q values. This enables Full-DIA to construct a complete, missingness-free fragment ion matrix as well as the peptide and

protein matrix. Such a protein matrix, with global FDR control based directly on raw MS intensities rather than imputed values, supports robust discovery of differentially expressed proteins.

Full-DIA applies a self-supervised autoencoder (DeepQuant) to denoise the fragment ion matrix (Fig. 1d, see Methods for details). This autoencoder reconstructs the signal using fragment intensities and coelution scores across runs, correcting low-confidence measurements. The top 5 denoised fragment ions are aggregated into peptide intensities, which are then combined into protein intensities.

The performance of Full-DIA

We first benchmarked Full-DIA on plasma and single-cell datasets (the summary table in Fig. 2) against DIA-NN (version 2.2). For Plasma [13] (15 runs) and single-cell SC-293 T [14] dataset (24 runs), both tools used a two-species spectral library (Human + *Arabidopsis Thaliana*), enabling FDR validation via the entrapment approach [15]. For the single-cell SC-LPS [14] (160 runs) and SC-Cycle [16] (229 runs), a predicted human spectral library was used. As shown in Fig. 2, Additional file 2: Fig. S1 and Additional file 2: Fig. S2, Full-DIA consistently achieved higher peptide detections and comparable or higher protein identifications at the global 1% FDR in these datasets, while maintaining an entrapment-based FDR close to 1% despite being marginally less stringent than DIA-NN. The precursor-level gains across the four datasets were 14%, 32%, 55%, and 14%, respectively, while the protein group-level gains were 1%, 7%, 19%, and −4%, respectively. In the SC-LPS dataset, Full-DIA's uniquely identified proteins show greater overlap with in-depth bulk LPS-treated THP-1 proteomics [17] than DIA-NN (Additional file 2: Fig. S3). In the SC-Cycle dataset, where Full-DIA identified more peptides but fewer proteins (Additional file 2: Fig. S4), we speculate that this is likely due to many peptides mapping to proteins that have already been identified, with the newly detected peptides mainly providing additional coverage rather than supporting new protein identifications. Full-DIA also ran faster on Plasma, SC-293 T, and SC-LPS datasets (Fig. 2d).

a

Dataset	Description	Sample	Load amount	Gradient/Throughput	Device	# Runs
Plasma[13]	Plasma by timsTOF HT	Plasma	100-300-600-900-1200 ng	40 min/30 SPD	timsTOF HT	15
SC-293T[14]	Single cells by timsTOF SCP	HEK-293T	Single cell	30 min/40 SPD	timsTOF SCP	24
SC-LPS[14]	Single cells treated with LPS	THP-1	Single cell	30 min/40 SPD	timsTOF Ultra	160
SC-Cycle[16]	Single cells in four cell cycles	HeLa	Single cell	35 min	timsTOF Pro	229

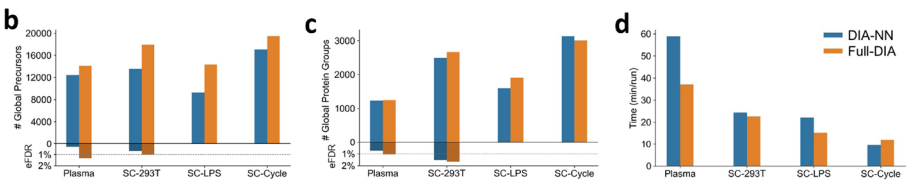

Fig. 2 Benchmarking Full-DIA on test datasets. **a** The summary table of test datasets. **b** Number of precursors at 1% reported global FDR. **c** Number of protein groups at 1% reported global FDR. For Plasma and SC-293 T (analyzed with a two-species library: Human + *Arabidopsis*), entrapment FDR (eFDR) = #*Arabidopsis* hits/#Human hits. **d** Average analysis time per run

Given that the Plasma dataset contains multiple loading amounts (100, 300, 600, 900 and 1200 ng), we evaluated the relative quantification performance of DIA-NN and Full-DIA by calculating the Pearson correlation coefficient (PCC) between measured intensities and injection amounts. As shown in Additional file 2: Fig. S1-c, the median PCCs for DIA-NN at the peptide and protein levels were 0.882 and 0.946, respectively, whereas those for Full-DIA were 0.934 and 0.953, indicating a higher correlation between measured and expected values. We further assessed quantification performance using a two-species (Human + *E. coli*) dataset [18] with known mixing ratios. As shown in Fig. 3, Full-DIA demonstrated superior performance in both quantitative precision and accuracy. Specifically, with expected median $\log_2$ fold-changes of 0 for human and 0.586 for *E. coli*, the observed values were − 0.137 and − 0.063 for human peptides (DIA-NN and Full-DIA, respectively), 0.658 and 0.559 for *E. coli* peptides, − 0.112 and − 0.074 for human proteins, and 0.631 and 0.590 for *E. coli* proteins. Collectively, these results indicate that Full-DIA achieves higher quantitative accuracy than DIA-NN across both the Plasma and the two-species dataset.

In response to the question of whether Full-DIA is applicable to bulk proteomics, we further evaluated its performance using a HeLa dilution series dataset (0.2–1–10–50–200 ng; 93-min gradient at 300 nL/min) [10]. As shown in Additional file 2: Fig. S5, Full-DIA exhibits a clear advantage at the lowest input level (0.2 ng), achieving approximately a 16% increase in peptide identifications compared to DIA-NN. At higher input amounts, the numbers of identified peptides and proteins are comparable between the two methods. We speculate that at low sample loads, the diaPASEF signals are characterized by a lower signal-to-noise ratio, under which the learning-based scoring components of Full-DIA may be better suited to extracting informative features, thereby improving identifications for low-input samples. Together, these results support the applicability of Full-DIA to bulk proteomics and highlight its particular strength in low-input samples, as in single-cell proteomics.

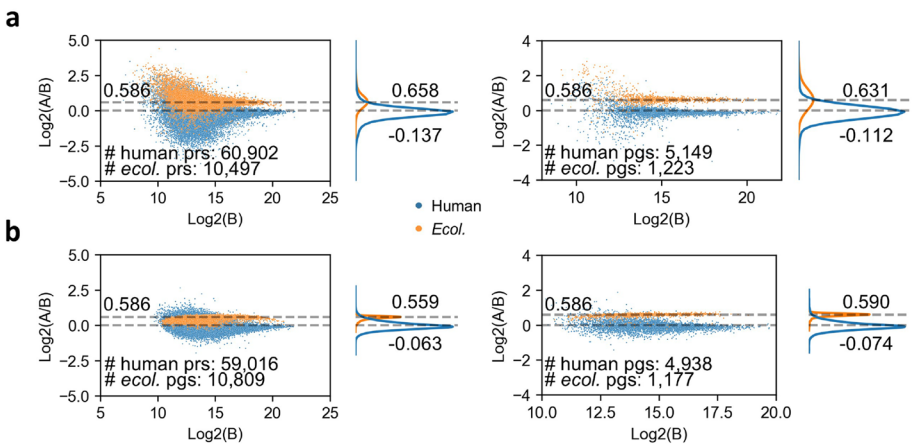

Fig. 3 Quantitative performance comparison between DIA-NN and Full-DIA on the two-species (Human + *E. coli*) mixed dataset with known ratios. The dashed line indicates the theoretical ratio. The values labeled on the density subplot represent the median of the measured ratios. Left: precursor-level results; Right: protein group-level results. **a** by DIA-NN. **b** by Full-DIA with species-level intensity correction

The application of Full-DIA

One innovation of this study is the ability to directly generate a complete, missing-value-free protein matrix at a global 1% FDR. To assess its potential advantage in biological interpretation, we compared it with the sparse matrix produced by DIA-NN under a global 1% FDR combined with a run-specific 1% FDR threshold. In the SC-LPS dataset (86 LPS-treated vs. 74 DMSO-treated THP-1 cells; Fig. 4), Full-DIA recovered a set of pathways closely matching the established TLR4-driven immune program, including TRAF6–IRF7 activation, IRAK1-mediated recruitment of the IKK complex, MAPK signaling cascades, amino acid stress responses via GCN2, and interleukin-mediated inflammatory signaling. These collectively represent the core molecular events expected from LPS stimulation. By contrast, DIA-NN-unique pathways showed a mixed profile: some, such as FasL/CD95L signaling and DEx/H-box helicase-mediated interferon activation, plausibly reflect secondary immune regulation, whereas many others—including viral genome replication, meiotic synapsis, and neurodegenerative processes—are biologically implausible in the context of acute LPS challenge and likely represent noise or off-target enrichments. In the SC-Cycle (91 G1, 41 G1–S, 52 G2, 45 G2–M cells; Fig. 5), Full-DIA achieved clearer PCA-based separation of the four cell-cycle phases compared to DIA-NN, with a higher silhouette score (0.25 vs. 0.20). The G2–M transition is characterized by chromosome condensation, spindle apparatus assembly, activation of the spindle checkpoint, and extensive reprogramming of RNA processing to support mitotic progression. Consistent with these hallmarks, Full-DIA recovered core mitotic pathways, including metaphase–anaphase transition, cytokinesis, cohesin loading onto chromatin, SUMOylation of DNA replication proteins, and multiple mRNA splicing and decay processes—providing a biologically coherent view of G2–M–specific activity. In contrast, DIA-NN enrichment was dominated by general translation-related processes (e.g., elongation, ribosome recycling, selenoamino acid metabolism), which, while relevant to proliferative cells,

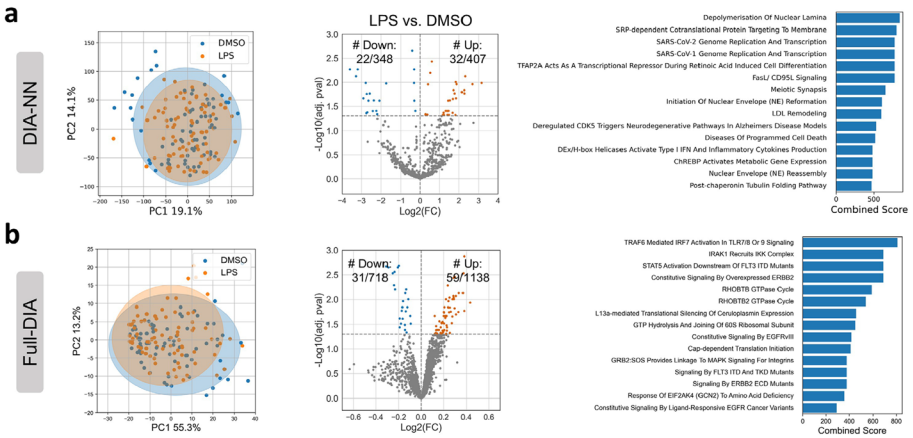

Fig. 4 Biological interpretability assessment of protein matrices derived from DIA-NN (**a**) and Full-DIA (**b**) on the SC-LPS dataset. Shown are principal component analyses (PCAs; ellipses denote 95% confidence intervals), volcano plots (significance threshold: adjusted *p*-value < 0.05 and $|\log_2$fold-change$| > 0.2$), and Reactome pathway enrichment analyses (top 15 pathways with adjusted *p*-value < 0.05, ranked by combined score). Statistical significance was determined using a two-sided *t*-test, and *p*-values were adjusted for multiple testing via the Benjamini–Hochberg procedure

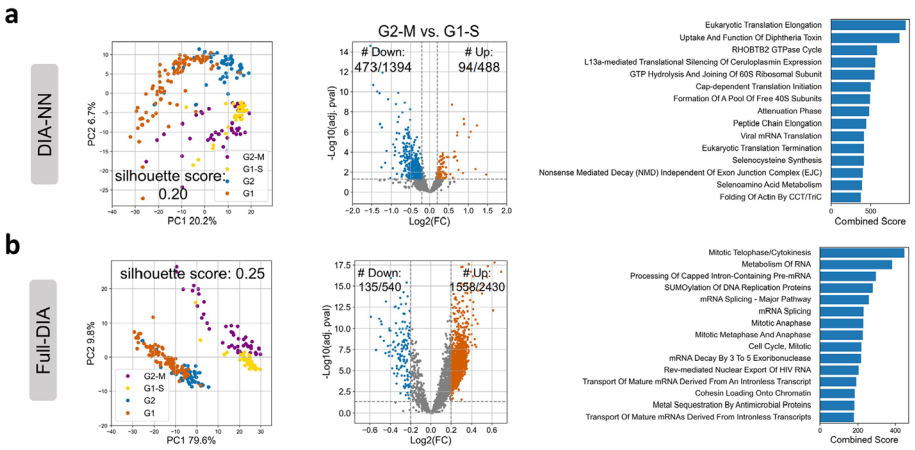

Fig. 5 Biological interpretability assessment of protein matrices derived from DIA-NN (**a**) and Full-DIA (**b**) on the SC-Cycle dataset. Shown are principal component analyses (silhouette scores are used to quantify the degree of separation between four cell phases), volcano plots (significance threshold: adjusted *p*-value < 0.05 and $|\log_2$fold-change$|> 0.2$), and Reactome pathway enrichment analyses (top 15 pathways with adjusted *p*-value < 0.05, ranked by combined score). Statistical significance was determined using a two-sided *t*-test, and *p*-values were adjusted for multiple testing via the Benjamini–Hochberg procedure

lack specificity for G2–M. Moreover, several DIA-NN-unique pathways (e.g., viral mRNA translation, diphtheria toxin uptake) are biologically implausible in the context of cell-cycle regulation.

For the SC-LPS and SC-Cycle datasets, the protein matrices generated by DIA-NN all employed zero-intensity imputation. Considering that different imputation strategies may influence biological interpretation, we additionally applied two alternative imputation methods—MinProb (Probabilistic Minimum Imputation [19]) and BPCA (Bayesian Principal Component Analysis [20])—to the DIA-NN results. As shown in Additional file 2: Fig. S6 and Additional file 2: Fig. S7, neither of these alternative imputations improved DIA-NN's clustering of cells according to distinct cell-cycle stages, nor did they yield more related biologically relevant enriched pathways. These findings further highlight the advantage of the missingness-free matrix generated by Full-DIA for biological interpretation.

Discussion

The extensive integration of deep learning into DIA proteomics represents a paradigm shift in addressing the highly convoluted and intrinsic complexity of DIA data. At the spectral library level, tools such as Prosit [21], DeepDIA [22], and AlphaPeptDeep [23] predict peptide properties to build more accurate libraries, while AlphaDIA [24] and Carafe [25] utilize transfer learning to refine the spectral libraries. For scoring, deep learning–based approaches including Alpha-XIC [26], Alpha-Tri [27], DreamDIA [28], and DIA-BERT [29] assess signal quality with improved precision. In quantification, DIA-BERT further leverages deep learning to estimate peptide abundances, enhancing quantitative accuracy beyond traditional methods.

Full-DIA exemplifies this trend by combining learned scoring of two-dimensional coelution consistency and spectral similarity with an autoencoder-based correction of

ion intensities, effectively bridging identification depth and quantitative precision. The adoption of GPU acceleration further ensures that these computationally intensive models remain practical for large-scale single-cell diaPASEF studies. Importantly, Full-DIA's ability to generate a missing-value-free protein matrix under stringent global FDR control provides an alternative solution to the pervasive issue of missing data in single-cell proteomics, thereby supporting more robust biological interpretation and integrative analyses. This feature is expected to establish a new benchmark for data completeness in high-throughput proteomics.

Overall, Full-DIA represents a significant step forward in harnessing deep learning for comprehensive and accurate single-cell DIA analysis, paving the way for deeper insights into key biological processes at the single-cell level.

Methods

Identification algorithms

Full-DIA takes a spectral library and.d files as inputs. Currently, Full-DIA supports directly reading spectral libraries in the.parquet format as defined by DIA-NN, from which it retrieves target peptide ions and its fragment ion information (including RT, ion mobility, *m/z*, and fragment ion intensities). Full-DIA utilizes the AlphaTims [30] to read.d files as raw profile data, and converts the profile data into centroided data according to DIA-NN's 2D peak-picking algorithm [10].

The identification workflow of Full-DIA can be divided into two phases: run-specific and experiment-wide. The run-specific phase further consists of four stages: the calibration stage, the first round of identification, the second round of identification, and the polishing of results.

In the calibration stage, Full-DIA extracts chromatographic elution groups for each target peptide ion (including the precursor in MS1 scan, the unfragmented precursor and the top 12 fragment ions in MS2 scan, 14 ions in total) from the centroided diaPASEF data with a fixed width in *m/z* (20 ppm) and ion mobility (0.08 Vs cm^{-2}) across the whole RT scale. Full-DIA then calculates the coelution consistency of elution groups using the cosine function within a sliding window, with a window width of 13 cycles and a sliding step of 1 cycle. Full-DIA selects the elution time, ion *m/z*, and ion mobility values corresponding to the apex of the elution group with the highest coelution consistency score as the measured attributes of the target precursors. Based on the measured and the spectral library attributes of target precursors, Full-DIA constructs RT, *m/z*, and mobility error profiles and uses the Calib-RT algorithm [31], a local regression (LOESS) algorithm with noise removal, to calibrate them. It is important to note that FDR control is absent during this calibration stage, which relies on a "rough" identification result that includes false positives. After calibration, Full-DIA randomly selects a subset of target peptides (defaulting to 1% of the spectral library), which undergo strict identification (similar to the first-round identification described below). Subsequently, the retention time, ion mobility, and mass tolerance are optimized based on the identification results.

Next is the first round of identification. For each target peptide in the spectral library, Full-DIA generates a decoy peptide via the mutate method (GAVLIFMPWSCTYHKRQEND to LLLVVLLLLTSSSSLLNDQE mutation pattern is used). For each target peptide or decoy peptide, Full-DIA extracts chromatograms of 14 monoisotopic ions

from the centroided diaPASEF data using the optimized tolerances. Full-DIA then seeks potential elution groups and assigns function-based scores to them. Meanwhile, Full-DIA utilizes a pre-trained DeepProfile model (described in detail in a later section) to perform 2D (RT and mobility) co-elution consistency scoring on 2D elution groups retrieved from profile data. Specifically, DeepProfile includes two models: DeepProfile-14 and DeepProfile-56 (unless otherwise specified, DeepProfile refers to both DeepProfile-14 and DeepProfile-56). DeepProfile-14 scores the 2D elution groups of the 14 ions mentioned earlier under -1, 0, $+1$, and $+2$ neutron conditions individually, while DeepProfile-56 scores the 2D elution groups of the 14*4 ions as a whole. After scoring, Full-DIA uses an ensemble neural network to distinguish between target and decoy peptides, resulting in a single run-specific discrimination score for each peptide, by which the local FDR is calculated (#decoy/#target). The ensemble neural network consists of twelve neural network-based binary classifiers. Each classifier is trained with the cross-entropy loss function treating target peptides as positive samples and decoy peptides as negative samples during training. The classifier features a series of hidden layers that are consistent with DIA-NN using [5, 10, 15, 20, 25] neurons. The parameters of each neural network are optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 50. The model is then trained for a single epoch to minimize the risk of overfitting.

For the second round of identification, Full-DIA constructs local training data based on the results of the first round of identification to fine-tune DeepProfile (scoring 2D coelution consistency) and train DeepMall (scoring spectrum similarity) from scratch. Specifically, the best elution groups of target precursors at 1% FDR from the first round are labelled as positive samples, and the second-best elution groups are labelled as negative samples, both of which are used to construct the corresponding 2D elution groups and spectrum similarity data. Based on the sample data, DeepProfile is tuned and DeepMall is trained. The scores from DeepProfile and DeepMall are then integrated with other scores from the first round into an ensemble neural network, and the FDR is updated. Both the refinement of DeepProfile and the training of DeepMall use the cross-entropy loss function and the Adam optimizer (learning rate of 0.0001). The samples are split into a training set and a validation sets in a 4:1 ratio. The batch size is 64. Training is halted when the accuracy on the validation set does not improve for 5 consecutive epochs or after a maximum of 50 training epochs. Notably, the fine tuning of DeepProfile involves only the layers used for calculating class probabilities, while the other feature layers remain frozen to preserve the feature extraction capabilities learned during the pre-training phase. It is essential to note that neither the refinement of DeepProfile nor the training of DeepMall utilizes decoy information, thereby eliminating the possibility of decoy data leakage compromising FDR accuracy. We supplemented ablation experiments comparing learning-based scoring and function-based scoring (see Additional file 2: Fig. S8).

The final stage involves polishing the results from the second round of identification to reduce the redundant use of fragments for multiple precursors from the same spectrum. Full-DIA checks whether the fragment ions of each target precursor could belong to a target precursor with a higher discriminative score. When two target precursors are located within the same fragmentation window, and their RT deviation, ion mobility

deviation, and *m/z* deviation of corresponding fragment ions fall within predefined ranges (3 cycles, 0.03 Vs cm^{-2} and 20 ppm, respectively), the fragment ion of the target precursor with the lower discriminative score is considered suspicious. If a target precursor contains at least three suspicious fragment ions, or if the total coelution score of suspicious fragment ions exceeds two-thirds of the total coelution score of all fragment ions, the target precursor is discarded. Then, the ensemble neural network updates the run-specific discrimination score based on the remaining target and decoy precursors. Full-DIA then saves these scores along with fragment ion intensities and fragment ion elution scores to a temporary file.

After completing the run-specific analysis of each individual run, Full-DIA proceeds to global experiment-wide identification. For each peptide, Full-DIA takes the maximum discrimination score across all runs as its global score and calculates the global peptide FDR based on these scores. Protein identification begins with the IDPicker [32] algorithm, which assigns peptides to protein groups. Since IDPicker determines the assignment of shared peptides based on the number of supporting peptides per protein, Full-DIA evaluates protein inference results under multiple global peptide-level FDR thresholds (1%—5%) and calculates protein-level discrimination scores accordingly (using the formula 1- $\prod(1\text{-}P_i)$, where P_i is the peptide discrimination score), ultimately selecting the configuration that maximizes the number of identified proteins. The global protein-level FDR is also estimated using the #decoy/#target method.

Quantification algorithms

During the run-specific analysis, Full-DIA obtains the best elution group for each peptide ion. Full-DIA then calculates the correlation between each pair of elution profiles and selects the profile with the highest correlation to others as the best profile. Next, Full-DIA iterates over different mass and mobility tolerances to re-extract the elution profiles of ions, selecting the elution profile with the highest correlation to the best profile as the final extracted profile. The mass tolerance space is [4, 8, 12, 16, 20] ppm, and the mobility tolerance space is [0.02, 0.01] Vs cm^{-2}. Full-DIA then stores the fragment ion intensities and their correlation scores in a temporary file. After completing global peptide identification, Full-DIA retrieves these values to construct two matrices: a fragment ion intensity matrix (samples × [#precursors × 12 ions]) and a fragment ion correlation score matrix with the same shape. Since the temporary file contains information for all target peptides, both matrices are complete and contain no missing values. Full-DIA then applies the DeepQuant autoencoder model to correct the fragment ion intensity matrix. Peptide-level intensities are computed as the sum of the top 5 fragment ions with the highest correlation scores across runs, and protein-level intensities are calculated as the sum of the top 1 peptide with the highest global discrimination score.

DeepProfile

DeepProfile is used to score the 2D coelution consistency on 2D elution groups. The pretraining data (Additional file 1: Table S2) of DeepProfile are provided by Shanghai Omicsolution Co., Ltd. These sixteen runs used for training are designed to maximize data diversity by combining different sample types, gradients, and window slicing. They were

identified using DIA-NN, which determined the best elution group for each peptide at 1% FDR. To augment the data, each best elution group was shifted by −1, 0, and 1 cycles, resulting in ~ 2.5 M 2D elution groups as positive samples. We then picked the top 3 suboptimal elution groups based on coelution scoring using cosine similarity across the whole elution time as negative samples. In this way, the number ratio of positive to negative samples keeps 1:1. The 2D elution groups with dimensions of $14 \times 13 \times 50$ (where 14 refers to the aforementioned ions, 13 to the cycles, and 50 to the bins of 0.001 Vs cm^{-2} width) were used to create the training dataset for DeepProfile-14. When accounting for isotopic variations, the 14 ions expand to 56 ions, and the corresponding 2D elution groups with dimensions of $56 \times 13 \times 50$ form the training dataset for DeepProfile-56.

Although DeepProfile-14 and DeepProfile-56 have different input dimensions, their model designs are identical (Fig. 1b). The raw 2D elution groups are first normalized at both the ion level and the overall level. The normalized data then pass through two rounds of [convolution, max pooling] and a fully connected layer, producing a 16-dimensional feature vector. Two 16-dimensional feature vectors are concatenated with an embedding representing the actual number of fragment ions, resulting in a 48-dimensional vector. This vector undergoes two fully connected transformations to produce the final class probability values. During the refinement process for DeepProfile, the two 16-dimensional feature vectors are frozen and remain unchanged.

DeepProfile's training is implemented by PyTorch, employing the Adam optimizer (with an initial learning rate of 0.001 and 1024 samples per batch) and a cross-entropy loss function. The training data is split into training and validation sets in a 4:1 ratio. Training stops when the validation accuracy does not improve for 5 consecutive epochs or after a maximum of 50 training epochs. Additional file 2: Fig. S9 illustrates the impact of training set size on model performance, with the validation set size kept constant. It shows that when the training set reaches 40% of its maximum size, adding more training data does not result in higher validation accuracy, suggesting that the current data size is sufficient for practical model training. Additional file 2: Fig. S10 illustrates the effect of varying the number of training epochs on identification results. It can be observed that additional training epochs (e.g., 50 epochs) do not lead to the identification of more peptides, indicating that the early stopping method based on a patience of 5 epochs is reasonable.

DeepMall

Compared to coelution consistency, spectrum similarity is a secondary feature in the identification process. Therefore, DeepMall (Fig. 1c), designed to score spectral similarity, is not pre-trained; instead, it is trained on the fly. After determining the best elution group for each peptide at 1% FDR, Full-DIA constructs the input data for DeepMall based on the top-12 fragment ions, including: a) the relative intensities of the fragment ions provided by the spectrum library; b) fragment ions type (b ions coded as 1, y ions coded as 2); c) normalized measurements intensities of the fragment ions from the three cycles around the elution apex, as well as the corresponding mass errors (normalized by dividing by 20 ppm) and mobility errors (normalized by dividing by 0.05 Vs cm^{-2}); d) cosine similarity scores of elution profiles; e) signal-to-noise ratios of elution profiles; f) profile areas. DeepMall treats these 14 features for each fragment ion as a whole and

feeds them into a two-layer bidirectional GRU, which ultimately outputs a probability value indicating spectral similarity. Since the spectrum similarity calculation incorporates not only the spectrum library intensities and measured intensities but also as much additional information about the fragment ions as possible to provide a confidence weight for spectrum similarity, the model is named DeepMall, reflecting its comprehensive amalgamation of ion information. The description of DeepMall's training can be found in the " Identification algorithms" section.

Pretraining DeepProfile with experiment-specific datasets

We downloaded a urine cohort experimental dataset [33], which contains 37.d files, each identified using DIA-NN. Except for the two files with the minimum and maximum identifications by DIA-NN, 35 files were processed to generate experiment-specific pretraining data using the method above, resulting in ~ 3 M 2D elution groups as positive samples. We applied the same pretraining method to complete the pretraining of the DeepProfile model based on experiment-specific data. Then, Full-DIA analyzed the two excluded files using DeepProfile, which was pretrained on either the heterogeneous data or the experiment-specific data, with the results shown in Additional file 2: Fig. S11. As can be seen, the DeepProfile model pretrained on heterogeneous data slightly outperforms the one trained on experiment-specific data. Given that Full-DIA includes a refinement process, it can be considered that, to some extent, Full-DIA eliminates the need for users to pretrain DeepProfile using experiment-specific datasets.

DeepQuant

DeepQuant is designed to calibrate fragment ion intensities through self-supervised reconstruction, aiming to correct low-intensity signals or low-quality measurements (Fig. 1d). For a given peptide, its 12 fragment ion intensities across multiple runs are first globally normalized (by dividing with the maximum intensity across all ions and runs) and locally normalized (by dividing with the maximum intensity among the 12 ions across all runs). These normalized intensities, along with the corresponding fragment ion correlation scores, are concatenated into a vector of length $3 \times 12 \times \#runs$ and fed into DeepQuant. DeepQuant uses a series of linear layers with ReLU activation functions to compress the input vector into 64-, 32-, and 16-dimensional representations, forming the encoder. The decoder then reconstructs the intensity input through successive linear layers (also with ReLU activations) to output a vector of size $12 \times \#runs$. The initial loss is computed as the mean squared error between this output and the globally normalized intensities. This loss is then averaged across fragment ions using their correlation scores as weights to yield a peptide-level reconstruction loss, which is further averaged across peptides using global discrimination scores as weights to obtain the final reconstruction loss. DeepQuant is implemented in PyTorch and optimized using the Adam optimizer (initial learning rate = 0.001) with a batch size of 64. The training data is split into training and validation sets in a 4:1 ratio. Training stops either when validation loss fails to improve for 10 consecutive epochs or after a maximum of 100 epochs.

We disabled DeepQuant and directly used the raw intensity values to demonstrate the effectiveness of DeepQuant (Additional file 2: Fig. S12-a). We also observed that,

for artificial mixed-species data, the relative quantification trends no longer followed a normal distribution. In such cases, DeepQuant's predictions for low-abundance species were biased toward those of high-abundance species (Additional file 2: Fig. S12-b, where DeepQuant was trained without distinguishing species), resulting in quantification bias. Therefore, for synthetic multi-species mixtures, DeepQuant adopts a species-specific training strategy to achieve optimal quantification performance (Fig. 3b). This issue does not arise in other test datasets, as they are all naturally single-species samples and their overall relative quantification trends are more likely to follow a normally distributed.

Spectral library generation

For all analyses, we used DIA-NN (version 2.2.0) to theoretically digest protein sequences and generate predicted spectral libraries. *Homo sapiens* (taxon identifier: 9606, 20,420 sequences), *Mus musculus* (taxon identifier: 10,090, 17,212 sequences), *E. coli* (taxon identifier: 562, 4,401 sequences), and *A. thaliana* (taxon identifier: 3702, 16,310 sequences) from UniProt (reviewed sequences only; downloaded on June, 2024) were used. All UniProt protein sequence files have been compiled in Additional File 3.The enzymatic digestion and prediction parameters by DIA-NN were as follows which are self-explanatory: '–min-fr-mz 200 –max-fr-mz 1800 –met-excision –min-pep-len 7 –max-pep-len 30 –min-pr-mz 300 –max-pr-mz 1800 –min-pr-charge 1 –max-pr-charge 4 –cut K*,R* –missed-cleavages 1 –unimod4 –var-mods 1 –var-mod UniMod:35,15.994915,M'. As a result, the predicted spectral library generated for Plasma, Dilution and SC-293 T datasets contained 9,557,086 precursors (*Homo sapiens* + *A. thaliana*), for SC-Cycle and SC-LPS contained 5,628,093 precursors, for the two-proteome mixture dataset (K562 + *E. coli*) contained 6,322,592 precursors.

Software versions and parameters

For each dataset, DIA-NN (version 2.2.0) was run using the following command:

'diann-2.2.0/diann-linux --dir/dataset/dir --lib/corresponding library.parquet --threads 24 --verbose 1 --out report_diann.parquet --qvalue 0.01 --matrices --out-lib/report-lib.parquet --gen-spec-lib --fasta/corresponding fasta.fasta --met-excision --cut K*,R* --missed-cleavages 1 --unimod4 --var-mods 1 --var-mod UniMod:35,15.994915,M --reanalyse --rt-profiling --no-peptidoforms --no-norm'

We disable intensity normalization in DIA-NN for the plasma dataset due to varying loading amounts, and because other datasets require specific normalization strategies for analysis. Full-DIA (version 1.0) analyzed each dataset using default parameters:

'full_dia -lib/corresponding library.parquet -ws/dataset'.

Data analysis

Full-DIA and DIA-NN each generated a report table in.parquet format for every diaPASEF dataset, with consistent column names across both. This table is used to generate the peptide or protein matrix. For DIA-NN, a global FDR of 1% and a run-specific FDR of 1% were applied. For Full-DIA, only a global FDR of 1% was considered. The analysis

of each dataset was performed as consistently as possible, following the description in the corresponding original publication, as detailed below.

For the Plasma dataset, the peptide and protein matrix by DIA-NN were imputed with zeros. The entrapment FDR was calculated as the ratio of *Arabidopsis Thaliana* hits to human hits. The PCC for each peptide or protein was calculated based on its measured intensities across different runs and the corresponding sample loading amounts for those runs.

For the Human-*E. coli* mixed-species dataset, the peptide and protein matrices by DIA-NN were imputed with zeros. The repeated samples were merged by averaging intensities.

For the HeLa dilution series dataset, performance comparisons were conducted without match-between-runs (MBR); therefore, the DIA-NN identifications were obtained from the 'report-first-pass.parquet' file. Since each input amount included three replicates, the relative increase in identifications was calculated based on the number of peptides (or proteins) detected in at least two replicates.

For the SC-293 T dataset, the entrapment FDR was calculated as the ratio of the number of *Arabidopsis Thaliana* hits to the number of human hits.

For the SC-LPS dataset, proteins with more than 70% missing values were removed from the DIA-NN protein matrix. Missing values in the remaining proteins were then imputed using zero filling, MinProb (implemented in the R package *MSnbase*, version 2.28.1), or BPCA (implemented in the R package *pcaMethods*, version 1.94.0). Protein matrices were then normalized using the R package *SCnorm* (version 1.24.0). After $\log_2$-transformation, batch correction and differential protein expression analysis, including calculation of p-values and adjusted p-values (Benjamini–Hochberg), were performed using the R package *limma* (version 3.58.1). An adjusted p-value < 0.05 was used as the cutoff for statistical significance. Pathway enrichment analysis was conducted using the Python package *gseapy* (version 1.1.8) on *Reactome_2022* gene sets. All pathways with adjusted p-values < 0.05 were reported.

For the SC-Cycle dataset, DIA-NN's protein matrix discarded the proteins with more than 70% missing values, and the remaining were imputed using zero filling, MinProb (implemented in the R package *MSnbase*, version 2.28.1), or BPCA (implemented in the R package *pcaMethods*, version 1.94.0). Protein matrices were then normalized to a total intensity of 10,000 per sample and subsequently $\log_2$-transformed to complete preprocessing. Principal component analysis and silhouette score were computed using scikit-learn version 1.5.2. Differential protein expression was assessed using moderated two-sided, two-sample t-tests. Proteins with an adjusted p-value < 0.05 (Benjamini–Hochberg correction) and an absolute $|\log_2(\text{fold change})| > 0.2$ were considered statistically significant. Pathway enrichment analysis was performed using the Python package *gseapy* (version 1.1.8) based on *Reactome_2022* gene sets. The top 15 pathways with adjusted p-values < 0.05 and the highest combined scores reported by *gseapy* were selected for visualization.

Runtime comparison

The runtimes for Full-DIA and DIA-NN were measured on a Linux desktop equipped with an Intel Core i9-14900KF CPU (32 logical cores), 256 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU. Full-DIA was set to use 12 CPU cores, while DIA-NN was allocated 24 CPU cores.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04087-x.

Additional file 1: Table S1. Detailed descriptions of all scoring items. Table S2. Table of detailed descriptions of training datasets for DeepProfile.

Additional file 2: Fig. S1. Performance comparison between DIA-NN and Full-DIA on the Plasma dataset. Fig. S2. Performance comparison between DIA-NN and Full-DIA on SC-293T dataset. Fig. S3. Venn diagrams showing the overlap of protein groups identified at 1% global FDR by DIA-NN, Full-DIA, and the bulk result from Mulvey et al. on the SC-LPS dataset. Fig. S4. Venn diagrams showing the overlap of protein groups identified at 1% global FDR by DIA-NN and Full-DIA on the SC-Cycle dataset. Fig. S5. Identification results for the HeLa dilution series dataset. Fig. S6. Biological interpretability assessment of protein matrices derived from DIA-NN with two imputation methods on the SC-LPS dataset. Fig. S7. Biological interpretability assessment of protein matrices derived from DIA-NN with two imputation methods on the SC-Cycle dataset. Fig. S8. Results of ablation experiments. Fig. S9. Pretraining results of the DeepProfile under different training data scales and epochs. Fig. S10. The impact of different training epochs for DeepProfile on identification results. Fig. S11. The impact of pretraining DeepProfile using heterogeneous (~2.5M 2D elution groups) or experiment-specific datasets (~3M 2D elution groups) on identification results. Fig. S12. Quantitative performance of Full-DIA's two comparative quantification strategies on the two-species (Human + *E. coli*) mixed dataset.

Additional file 3. The UniProt sequences used in this study.

Acknowledgements
We thank Liu Zhu and Catherine C. L. Wong for their early support of this project.

Peer review information
Yasset Perez-Riverol and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
JS and JGM contributed to the conceptualization of the study. JS was responsible for data curation (together with HL), formal analysis (together with AM), methodology design (together with JGM), software development and validation (together with HL and AM). Funding acquisition and supervision were provided by XW and JGM. Resources were provided by CS, XW and JGM. JS wrote the original draft of the manuscript. All authors participated in the review and editing of the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by National Natural Science Foundation of China [T2222007 to X.W.] and Outstanding Postdoctoral Program of Jiangsu Province, China. This work was partially funded by NIGMS R35GM142502.

Data availability
Full-DIA is available on GitHub (https://github.com/xomicsdatascience/full_dia) [34] and archived on Zenodo (https://doi.org/10.5281/zenodo.19448789) [35] under the Apache-2.0 License. Full-DIA can be installed via PyPI. Python/R scripts for data analysis can be found at the same repository. DIA-NN and Full-DIA reports on test datasets can be download from Figshare (https://doi.org/10.6084/m9.figshare.29527934) [36].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 September 2025 Accepted: 16 April 2026
Published online: 21 April 2026

References

1. Guo T, Steen JA, Mann M. Mass-spectrometry-based proteomics: from single cells to clinical applications. Nature. 2025;638(8052):901–11.
2. Gillet LC, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Molecular & Cellular Proteomics. 2012. https://doi.org/10.1074/mcp.O111.016717.
3. Ludwig C, Gillet L, Rosenberger G, et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol. 2018;14(8):e8126.

4. Meier F, Brunner AD, Frank M, et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nat Methods. 2020;17(12):1229–36.
5. Guzman UH, Martinez-Val A, Ye Z, et al. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition[J]. Nat Biotechnol. 2024;42(12):1855–66.
6. Kuster B, Tüshaus J, Bayer FP. A new mass analyzer shakes up the proteomics field. Nat Biotechnol. 2024;42(12):1796–7.
7. Derks J, Leduc A, Wallmann G, et al. Increasing the throughput of sensitive proteomics by plexDIA. Nat Biotechnol. 2023;41(1):50–9.
8. MacCoss MJ, Alfaro JA, Faivre DA, et al. Sampling the proteome by emerging single-molecule and mass spectrometry methods. Nat Methods. 2023;20(3):339–46.
9. Demichev V, Messner CB, Vernardis SI, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 2020;17(1):41–4.
10. 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(1):3944.
11. Rosenberger G, Bludau I, Schmitt U, et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods. 2017;14(9):921–7.
12. Ting YS, Egertson JD, Payne SH, et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data[J]. Mol Cell Proteomics. 2015;14(9):2301–7.
13. Vitko D, Chou WF, Nouri Golmaei S, et al. timsTOF HT improves protein identification and quantitative reproducibility for deep unbiased plasma protein biomarker discovery[J]. J Proteome Res. 2024;23(3):929–38.
14. Ctortecka C, Clark NM, Boyle BW, et al. Automated single-cell proteomics providing sufficient proteome depth to study complex biology beyond cell type classifications[J]. Nat Commun. 2024;15(1):5707.
15. Bruderer R, Bernhardt OM, Gandhi T, et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results[J]. Mol Cell Proteomics. 2017;16(12):2296–309.
16. Brunner AD, Thielert M, Vasilopoulou C, et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation[J]. Mol Syst Biol. 2022;18(3):e10798.
17. Mulvey CM, Breckels LM, Crook OM, et al. Spatiotemporal proteomic profiling of the pro-inflammatory response to lipopolysaccharide in the THP-1 human leukaemia cell line[J]. Nat Commun. 2021;12(1):5773.
18. Szyrwiel L, Gille C, Mülleder M, et al. Fast proteomics with dia-PASEF and analytical flow-rate chromatography[J]. Proteomics. 2024;24(1–2):2300100.
19. Lazar C, Gatto L, Ferro M, et al. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J Proteome Res. 2016;15(4):1116–25.
20. Stacklies W, Redestig H, Scholz M, et al. pcaMethods—a bioconductor package providing PCA methods for incomplete data[J]. Bioinformatics. 2007;23(9):1164–7.
21. Gessulat S, Schmidt T, Zolg DP, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 2019;16(6):509–18.
22. Yang Y, Liu X, Shen C, et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun. 2020;11(1):146.
23. Zeng WF, Zhou XX, Willems S, et al. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. Nat Commun. 2022;13(1):7238.
24. Wallmann G, Skowronek P, Brennsteiner V, et al. AlphaDIA enables DIA transfer learning for feature-free proteomics[J]. Nat Biotechnol. 2025:1–10.
25. Wen B, Hsu C, Shteynberg D, et al. Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics[J]. Nat Commun. 2025;16(1):9815.
26. Song J, Yu C. Alpha-XIC: a deep neural network for scoring the coelution of peak groups improves peptide identification by data-independent acquisition mass spectrometry. Bioinformatics. 2022;38(1):38–43.
27. Song J, Yu C. Alpha-Tri: a deep neural network for scoring the similarity between predicted and measured spectra improves peptide identification of DIA data. Bioinformatics. 2022;38(6):1525–31.
28. Gao M, Yang W, Li C, et al. Deep representation features from DreamDIAXMBD improve the analysis of data-independent acquisition proteomics[J]. Communications Biology. 2021;4(1):1190.
29. Liu Z, Liu P, Sun Y, et al. DIA-BERT: pre-trained end-to-end transformer models for enhanced DIA proteomics data analysis[J]. Nat Commun. 2025;16(1):3530.
30. Willems S, Voytik E, Skowronek P, et al. AlphaTims: indexing trapped ion mobility spectrometry–TOF data for fast and easy accession and visualization[J]. Mol Cell Proteomics. 2021;20:100149.
31. Zhang Y, Hu C, Wu X, et al. Calib-RT: an open source python package for peptide retention time calibration in DIA mass spectrometry data. Bioinformatics. 2024;40(7):btae417.
32. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency[J]. J Proteome Res. 2007;6(9):3549–57.
33. Tian W, Zhang N, Jin R, et al. Immune suppression in the early stage of COVID-19 disease[J]. Nat Commun. 2020;11(1):5859.
34. Song Jian, Momenzadeh Amanda, Liu Hebin, Shen Chengpin, Meyer Jesse, Wu Xiaohui. Full-DIA enables complete single-cell proteomics from diaPASEF using deep learning. Github. 2026 https://github.com/xomicsdatascience/full_dia
35. Song Jian, Momenzadeh Amanda, Liu Hebin, Shen Chengpin, Meyer Jesse, Wu Xiaohui. Full-DIA enables complete single-cell proteomics from diaPASEF using deep learning. Zenodo. 2026 https://doi.org/10.5281/zenodo.19448789
36. Song Jian, Momenzadeh Amanda, Liu Hebin, Shen Chengpin, Meyer Jesse, Wu Xiaohui. Full-DIA enables complete single-cell proteomics from diaPASEF using deep learning. Figshare. 2026 https://doi.org/10.6084/m9.figshare.29527934

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.