Interdisciplinary Sciences: Computational Life Sciences
https://doi.org/10.1007/s12539-024-00611-4

ORIGINAL RESEARCH ARTICLE

SeFilter-DIA: Squeeze-and-Excitation Network for Filtering High-Confidence Peptides of Data-Independent Acquisition Proteomics

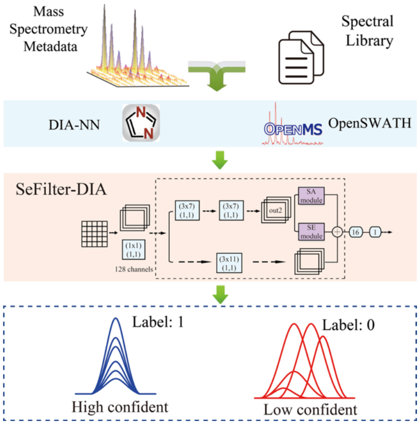
Qingzu He[1,2] · Huan Guo[1] · Yulin Li[1] · Guoqiang He[2] · Xiang Li[1] · Jianwei Shuai[2,3]

Received: 17 July 2023 / Revised: 12 January 2024 / Accepted: 21 January 2024

Abstract

Mass spectrometry is crucial in proteomics analysis, particularly using Data Independent Acquisition (DIA) for reliable and reproducible mass spectrometry data acquisition, enabling broad mass-to-charge ratio coverage and high throughput. DIA-NN, a prominent deep learning software in DIA proteome analysis, generates peptide results but may include low-confidence peptides. Conventionally, biologists have to manually screen peptide fragment ion chromatogram peaks (XIC) for identifying high-confidence peptides, a time-consuming and subjective process prone to variability. In this study, we introduce SeFilter-DIA, a deep learning algorithm, aiming at automating the identification of high-confidence peptides. Leveraging compressed excitation neural network and residual network models, SeFilter-DIA extracts XIC features and effectively discerns between high and low-confidence peptides. Evaluation of the benchmark datasets demonstrates SeFilter-DIA achieving 99.6% AUC on the test set and 97% for other performance indicators. Furthermore, SeFilter-DIA is applicable for screening peptides with phosphorylation modifications. These results demonstrate the potential of SeFilter-DIA to replace manual screening, providing an efficient and objective approach for high-confidence peptide identification while mitigating associated limitations.

Graphical Abstract



Keywords Proteomics analysis · Mass spectrometry · Data-independent acquisition · Manual screening peptides · Squeeze-and-excitation networks · Deep learning

1 Introduction

Data-dependent acquisition (DDA) and data-independent acquisition (DIA) are common strategies in mass spectrometry. In the DDA-based shotgun experiment [1], the mass

Qingzu He and Huan Guo would be designated as co-first authors due to their equal contributions to the research and preparation of the manuscript.

Extended author information available on the last page of the article

Published online: 12 March 2024

spectrometer performs a full scan to acquire the spectra of peptide precursors (MS1), then selects the precursor ions with the top N intensity for fragmentation. While DDA establishes a clear correlation between precursors and fragments, its dependency on precursor ion intensity leads to stochastic experimental outcomes and limited identification capabilities for low-abundance peptides.

The DIA method divides the mass-to-charge ratio (m/z) interval of precursor ions into several independent windows, and breaks all the precursor ions in each window in turn and records the signals of all fragment ions. However, the fragment ions in DIA data are from different precursor ions and mixed within MS2 spectra, which makes the analysis of DIA data extremely difficult. Accurately quantifying DIA mass spectrometry data is fundamental for various biologically significant studies in bioinformatics. For instance, scientists have leveraged SWATH-MS technology [2], a DIA mass spectrometry technique, to achieve quantitative insights into the crosstalk between apoptotic and necroptotic pathways [3]. Additionally, biokinetic modeling based on SWATH-MS mass spectrometry data has enabled valuable investigations [4]. Successfully analyzing DIA mass spectrometry data is crucial for advancing studies, fostering a highly active research field in tool development.

Two primary methods are employed for DIA data analysis. Library-dependent analysis involves creating a library from DDA experimental data and matching experimental DIA spectra with library spectra for quantitative peptide analysis. Common tools include OpenSWATH [5], SWATH-Prophet [6], and Specter [7]. Conversely, library-free analysis, like DIA-Umpire [8], eliminates DDA experiments. It leverages precursor ion fragmentation efficiency to establish a pseudo-database. Tools include Group-DIA [9], PIQED [10], directDIA (a component of Spectronaut [11]), PECAN [12], and MaxDIA [13], providing alternative DIA data analysis without DDA dependency.

Deep learning algorithms, superior to traditional techniques [14–17], have significantly advanced biological image analysis [18–20]. Recent years witnessed the development of deep learning-based library-free methods. For instance, the Prosit algorithm [21] predicted theoretical spectra and retention times using recurrent neural network (RNN) models, enhancing mass spectrometry identification. The DeepNovo-DIA algorithm [22] enabled direct amino acid sequencing of peptides. These advancements have rendered DDA experiments unnecessary, as deep learning methods predict spectra, establish databases, and perform subsequent analysis. Notable tools include DeepMass [23], pDeep [24], and DeepDIA [25].

Dear-DIA^{XMBD} used a deep variational autoencoder to extract ion signal features for identified and quantified peptide and protein analysis [26], and focused on spectrogram-centric DIA mass spectrometry data processing. It employed autoencoder and triplet loss to learn features from fragment ion chromatograms, grouping similar fragments using k-means clustering. Additionally, a deep learning-based identification software, DreamDIA [27] used Long Short-Term Memory (LSTM) and fully connected networks to score input data and improve accuracy by normalizing spectral library results' retention times to predict others.

Among these advancements, DIA-NN stands as an integrated software utilizing deep learning for processing DIA proteomics data, marking a significant milestone in proteomics [28]. DIA-NN enables high-throughput, reliable, and quantitative large-scale experiments, though it might produce a fraction of low-confidence peptides even with algorithmic control of false positives. Hence, researchers manually filter high-confidence peptides by extracting fragment ion chromatographic peak sets (XICs) and filtering based on peak shape similarity across six fragment ions. Available chromatographic peak visualization tools include Skyline [29], TOPPView [30], MSSort-DIA^{XMBD} [31], and DrawAlignR [32].

MSSort-DIA^{XMBD} serves as a final step in the OpenSWATH workflow. This tool specifically focuses on visualizing and categorizing peptide precursor ions using MS/MS data. Leveraging OpenSWATH's output, it reconstructs and visualizes chromatographic curves for each peptide and its matched fragment ion groups. Deep convolutional neural networks extract valuable information from this data, employing a double-threshold segmentation strategy to automatically identify high and low-confidence peptides. Essentially, this method acts as a re-screening of OpenSWATH's output, easing the burden of manual inspection.

DIA-NN generates extensive peptide reports, ranging from thousands to hundreds of thousands. However, manually checking these peptides' ion chromatograms can take an extensive amount of time, potentially up to 83 h, due to the large number of peptides. Manual screening, reliant on subjective criteria, can vary among individuals, affecting error rates.

The manual filtering process involves two steps: extracting chromatographic curves from reported data and visually inspecting these images. To streamline this laborious process, we introduce SeFilter-DIA, a deep learning-based algorithm. SeFilter-DIA automatically reclassifies and filters DIA-NN reported peptides, aiming to identify high-confidence peptides and replace manual screening. We assessed its performance using the getXIC tool to label 86,443 peptides and train the SeFilter-DIA model. Results show SeFilter-DIA effectively discerns high and low-confidence peptides, offering an efficient alternative to manual DIA-NN peptide filtering. This highlights SeFilter-DIA's potential as a valuable tool in enhancing efficiency and standardization in proteomic data analysis.

2 Materials and Methods

2.1 Building of Benchmarked Dataset

High-quality training datasets play a crucial role optimizing deep learning model parameters. To achieve this, we constructed a benchmark dataset specifically designed for training and testing our models. This dataset included diverse raw DIA mass spectrometry data from various species and instruments. Our primary analysis tool was DIA-NN, supplemented by data from the traditional workflow, OpenSWATH-PyProphet-TRIC (OSPT). Integrating these varied datasets expanded our model's adaptability, enhancing its ability to handle diverse data.

In Table 1, we use a naming convention that combines sample names with analysis workflows to create more descriptive dataset names. This helps provide informative details about each dataset.

The Yeast_NN dataset, available at ProteoXChange with identifier PXD031160, includes yeast sample data obtained using the ABSciex TripleTOF 6600 mass spectrometer. This dataset conducted proteomic analysis on FACS-sorted and unsorted helper and prototypic subpopulations within a self-established metabolic cooperative community (SeMeCo), along with a control wild-type yeast community. The SCIEX TripleTOF 6600 instrument performed mass spectrometry analysis employing the DIA method. MS1 scanning ranged from 400 to 1250 m/z, with an accumulation time of 50 ms. MS2 involved 40 variable window scans, each with an accumulation time of 35 ms.

Table 1 Dataset Information

Dataset Name	Instrument	Vender	Analytics software
Yeast_NN	TTOF 6600	ABSciex	DIA-NN
Human_NN	Fusion Lumos	Thermo Fischer	DIA-NN
E.coli_NN	TTOF 6600	ABSciex	DIA-NN
L929_NN	TTOF 5600	ABSciex	DIA-NN
HYE124_NN	TTOF 5600/6600	ABSciex	DIA-NN
HYE110_NN	TTOF 5600/6600	ABSciex	DIA-NN
Yeast_OSPT	TripleTOF 5600	ABSciex	OSPT workflow
SGS_OSPT	TripleTOF 5600	ABScex	OSPT workflow
Hela_OSPT	Q Exactive HF-X	Thermo Fischer	OSPT workflow
E.coli_OSPT	TTOF 6600	ABSciex	OSPT workflow
L929_OSPT	TripleTOF 5600	ABSciex	OSPT workflow
BGS_OSPT	Fusion Lumos	Thermo Fischer	OSPT workflow
HYE124_ OSPT	TTOF 5600/6600	ABSciex	OSPT workflow
HYE110_ OSPT	TTOF 5600/6600	ABSciex	OSPT workflow

The columns display: dataset name, mass spectrometer used (abbreviated as TTOF for TripleTOF and Lumos for Orbitrap Fusion Lumos Tribrid), manufacturer of the mass spectrometer, and the analysis software applied (OSPT workflow stands for OpenSWATH-PyProphet-TRIC)

The Human_NN dataset [33] was generated from brain microvascular endothelial cells (HBMECs) using the Thermo Fischer Scientific Orbitrap Fusion Lumos mass spectrometer. Trypsin digestion with Promega trypsin was applied at a weight ratio of 1:50. The MS1 scan range covered 400–1250 m/z at a resolution of 60,000 (FWHM). MS2 utilized 30 variable window scans with 30% HCD prior to precursor ion fragmentation.

The E.coli_NN dataset [34] was produced from E. coli using the TripleTOF 6600 mass spectrometer. Sequencing grade Trypsin-Gold (Promega) was used for protein digestion at a 1:100 ratio. MS1 scans spanned 400 to 1250 m/z, with 100 variable windows for MS2. The experiment had a total duty cycle of 1.7 s for a 15-min gradient and 2.2 s for a 90-min gradient. Five replicate samples were collected in duplicate.

The L929_NN dataset [26] obtained from mouse samples utilized the TripleTOF 5600 mass spectrometer in SWATH mode. L929 cells treated with Tumor Necrosis Factor (TNF) were purified for TNFR1 complexes and subjected to trypsin digestion. MS1 covered 400–1150 m/z with 100 variable windows for MS2 scans.

The HYE124_NN and HYE110_NN datasets [35] are mixed samples of human, yeast, and E. coli analyzed using the TripleTOF 5600 or TripleTOF 6600 mass spectrometers. Samples were proportionally divided into Sample A (65% human, 30% yeast, 5% E. coli) and Sample B (65% human, 15% yeast, 20% *E. coli*). MS1 scanned 400–1200 m/z, and MS2 employed 32 or 64 windows with the choice of fixed or variable window sizes. The experiment was run over a 2-h gradient.

The combined dataset comprises 42,443 peptides processed by DIA-NN and labeled manually, with 22,306 high-confidence and 20,137 low-confidence peptides.

The Yeast_OSPT dataset, available at ProteoXChange with identifier PXD028735, contains yeast sample data acquired using the TripleTOF 5600 mass spectrometer. The experiment used a primary scan range of 400–1200 m/z and 64 fixed windows for secondary scans over a 2-h gradient.

The SGS_OSPT dataset [5] includes human sample data obtained from the TripleTOF 5600 mass spectrometer. The MS1 scan range for this dataset is 400–1200 m/z, with 32 fixed windows employed during the MS2 acquisition. The gradient length for the experiment is set to 2 h.

The Hela_OSPT dataset consists of HeLa cell sample data acquired using the Q Exactive HF-X mass spectrometer. It employed an MS1 scan range of 350–1650 m/z with 45 windows during MS2 over a 2-h experiment.

The BGS_OSPT dataset [36] contains data from BGS mouse samples obtained using the Orbitrap Fusion Lumos mass spectrometer. This dataset used trypsin from Promega (Madison, WI), digesting samples at a fixed enzyme-to-protein ratio of 1:100. The MS1 scan range was 350–1650 m/z with 40 windows during MS2 using a nonlinear gradient over 2 h.

The settings for the E.coli_OSPT, L929_OSPT, HYE110_OSPT, and HYE124_OSPT datasets correspond to their respective *E.coli*_NN, L929_NN, HYE110_NN, and HYE124_NN datasets.

Following analysis using the OpenSWATH-PyProphet-TRIC workflow, a total of 44,316 peptides were labeled manually, with 22,744 classified as high-confidence and 21,572 as low-confidence peptides.

2.2 Introduction of DIA-NN Workflow

To use DIA-NN effectively, it is necessary to download and install MSConvert [37] (V.3.0.19311) and Thermo MS File Reader (3.0 SP3). These tools help convert DIA raw files from various mass spectrometers into formats that work with DIA-NN. MSConvertconverts.wiff and.raw files to a compatible format for DIA-NN, while Thermo MS File Reader (3.0 SP3) suits Thermo Fisher Scientific mass spectrometry data.

In our study, we employed DIA-NN (version: 1.7.11) with DIA data and a FASTA database from Uniprot. DIA-NN used a specific setup: MS1 m/z range from 400 to 1200 and MS2 from 100 to 1800, maintaining a 1% false discovery rate (FDR). Other parameters were kept at their default values. These settings produced a quantitative report (Fig. 1A) detailing peptide identification, quantification, and statistical analysis.

2.3 Introduction of OpenSWATH-PyProphet-TRIC Workflow

In the OpenSWATH-PyProphet-TRIC workflow [38, 39], we used various tools to handle and validate the mass spectrometry data. First, MSConvert (V.3.0.19311) converted raw data into mzXML format. DIA-Umpire then created pseudo-DDA files in mgf format to replicate DDA experiment data. We referred to the UniprotKB/Swiss-Prot database and searched the pseudo-DDA files using Comet (V2017.01) and X!Tandem (V2013.06.15.1, schema native and k-score) [40, 41]. Results were outputted in pep.xml format. PeptideProphet [42] validated these results, followed by iProphet [43] probability determination using mayu (V1.07). We selected peptides based on a 1% protein-level FDR screening. SpectraST [44] generated a required library file (sptxt), which was then processed to normalize retention times using iRT peptides. Finally, the sptxt file underwent conversion to tsv format and was further transformed into TraML format for quantitative analysis in the OpenSWATH-PyProphet-TRIC workflow.

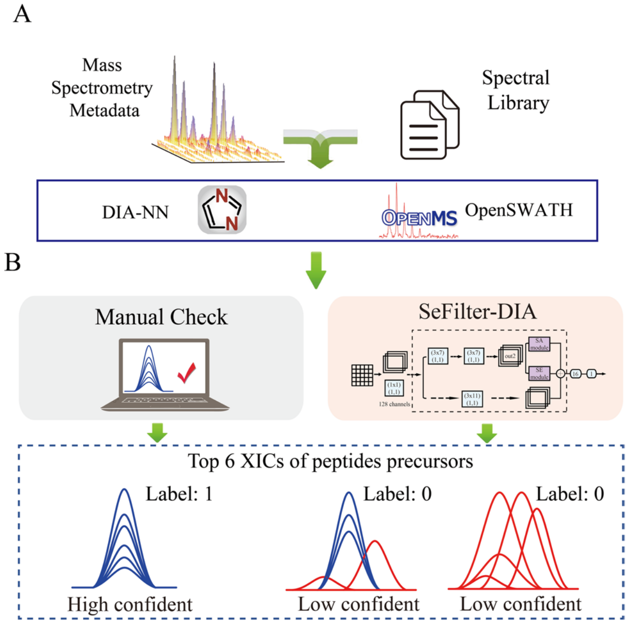


Fig. 1 The workflow of SeFilter-DIA and the workflow of DIA-NN. **A** After analyzing DIA mass spectrometry data with DIA-NN, identified peptides are generated as output. **B** The left side illustrates the manual check stage involving manual screening for high-confidence peptides. The right side demonstrates our workflow: chromatographic peaks are extracted and automatically classified by SeFilter-DIA. This process assigns a label of 1 to high-confidence peptides and 0 to low-confidence ones, aiding in identifying high-confidence peptides

2.4 Dataset Composition and Preprocessing

After obtaining the output report from either DIA-NN or the OpenSWATH-PyProphet-TRIC workflow, we proceeded to extract the fragment chromatographic peak groups from the DIA raw data. This extraction was performed using the script getXIC.py (Fig. 1B). Since the number of fragment ions in the DIA-NN output is not fixed, we selected the six fragment ions with the highest intensity as members of the fragment ion peak group.

We extracted fragment XIC groups from the DIA raw data using getXIC.py (Fig. 1B). We selected the six fragment ions with the highest intensity from the DIA-NN output. After testing various curve lengths (40, 55, 70, 85, 90, 105, and 120), we found that 85 time points yielded the best results. SeFilter-DIA's input data comprises 6 rows and 85 columns, representing six fragment ion extracted ion chromatograms

(XICs). For each XIC, we centered on the time point with the highest intensity and took 42-time points forward and backward. Normalization using sklearn's min–max scaling function ensured intensity values ranged between 0 and 1 (Fig. 2).

The Minmax normalization function equation is:

$$X_{\text{norm}} = \frac{X_i - \min(X_i)}{\max\left(X_i\right) - \min(X_i)}$$

Peptide data were labeled based on peak shape similarity after normalization: similar shapes were labeled as high-confidence, dissimilar as low-confidence (Fig. 1B). The SeFilter-DIA benchmark dataset contains 86,443 peptides with 40,641 high-confidence and 45,802 low-confidence peptides. The dataset was split randomly into training, cross-validation, and test sets in a 6:2:2 ratio.

2.5 The Principle of SeFilter-DIA

To address the classification problem, we developed three deep learning models based on the principles of the convLSTM network [45], the residual network [46], and the compressed excitation model [47] combined with the residual network. Through evaluation and comparison, we selected the deep neural network that demonstrated the best performance as the model for SeFilter-DIA.

To evaluate the performance of the SeFilter-DIA model, we compared MSSort-DIA$^{\text{XMBD}}$, Gradient boosting decision tree (GBDT), AdaBoosting, SVM, RandomForest, Pearson correlation coefficient and Spearman correlation coefficient on the test set. The parameters of the algorithms except MSSort-DIA$^{\text{XMBD}}$ are manually tuned.

The neural network of SeFilter-DIA uses a two-layer residual framework combined with a compression incentive model and self-attention mechanism, forming a three-part connection (Fig. 3). The model's first part includes a 1 × 1 kernel convolutional layer with 256 channels. The second part, the residual structure, has two paths: one with 3 × 7 convolutional kernels, a padding of 1, and a stride of 1, linked to a compressed excitation and self-attention module. The second path has a 3 × 7 convolutional kernel, padding of 1, stride of 1, and Relu activation. Both outputs merge.

The third part has a 16-neuron fully connected layer with Relu activation, followed by a final output layer with 1 neuron activated by Sigmoid (Fig. 3). The convolutional layer in SeFilter-DIA is critical for processing multidimensional data, capturing spatial correlations, and extracting meaningful features. The second route in the residual structure, akin to a skip connection, preserves original data information

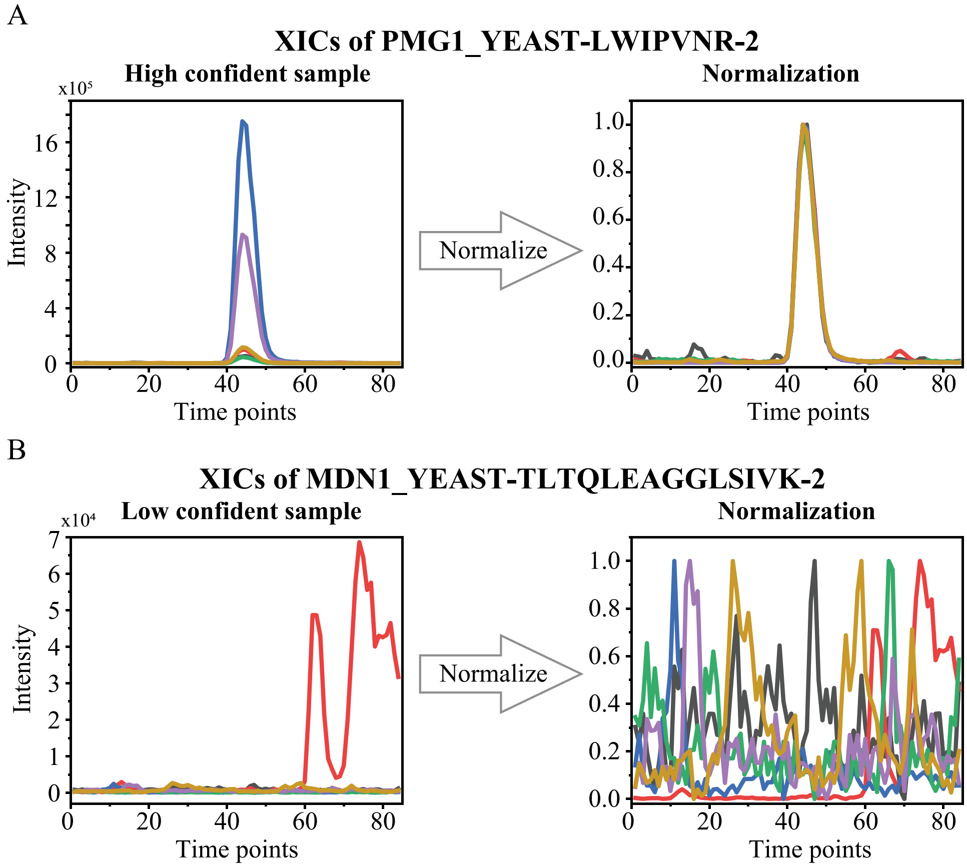


Fig. 2 The XICs Comparison for Peptides. **A** XICs of peptide PMG1_YEAST_LWIPVNR labeled as high-confidence. Intensity versus time (left) and normalized intensity (right) diagrams are shown. **B** XICs of peptide MDN1_YEAST-TLTQLEAGGLSIVK marked as high-confidence. Intensity versus time (left) and intensity-normalized (right) graphs displayed

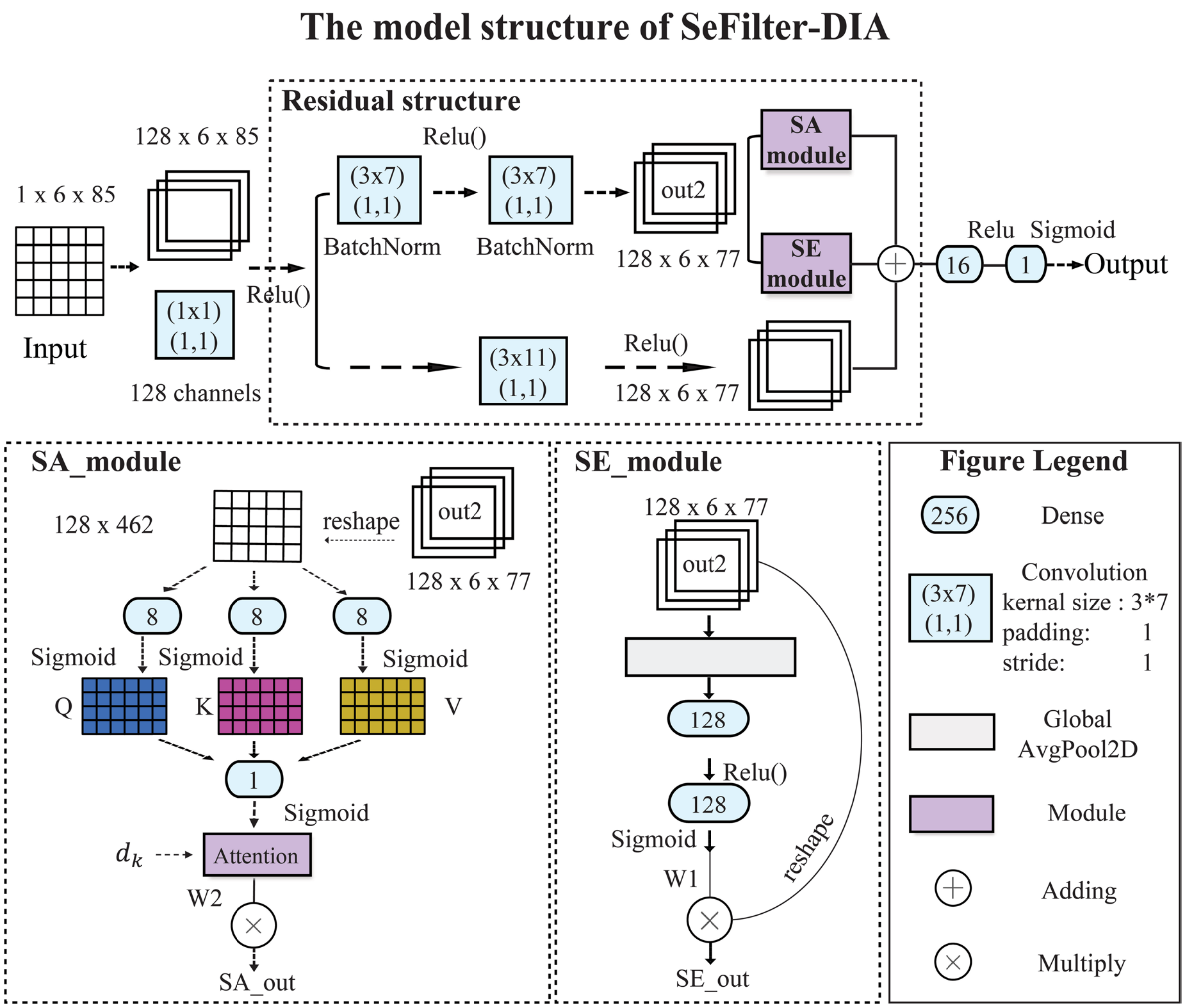


Fig. 3 SeFilter-DIA Model Overview. Residual structure (dashed box), self-attention (SA_module) and squeeze-and-excitation (SE_module) mechanisms highlighted. Legend: blue oval (256) denotes a fully connected layer, blue square (3×7)(1,1) signifies convolutional layer (3 × 7 kernel, padding 1, stride 1), gray square represents global pooling, purple square signifies complex module, plus sign for data matrix summing, and multiplication for data matrix operations

to counter model degradation from information loss. This direct path helps maintain crucial details and prevents performance decline.

Furthermore, the compressive excitation module incorporated in the model enhances its capabilities by explicitly modeling the interdependencies among different channels. This module enables an adaptive recalibration of the channel feature responses, improving the overall representation power and aiding in capturing important patterns and variations within the data. (Fig. 3, where the compressive excitation module is highlighted within the dashed box.)

The self-attention mechanism module incorporated in the SeFilter-DIA model plays a crucial role in capturing internal correlations within the data. By leveraging this mechanism, the model can effectively focus on key information and enhance its performance (Fig. 3). The self-attention mechanism formula is:

$$Attention(X) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$$Q = W^Q X, K = W^K X, V = W^V X$$

The matrices W^K, W^Q, and W^V represent the weights of Q, K, and V, respectively, which are trainable parameters. Here, d_k is a constant, set to the channel value, to prevent the softmax input from becoming too high, which could lead to bias and result in the gradient tending toward 0.

The deep neural network of SeFilter-DIA is trained using the Adam optimizer, with a training batch size of 256. The values of the parameters are as follows: beta1 = 0.9, beta2 = 0.999, epsilon = 1e-5, weight decay = 1e-5, and learning rate = 3e-6. The training is performed for a total of 100 epochs.

2.6 The Principle of the Comparison Methods

The ResNet-based model (Fig. 4A) takes input dimensions of (1 × 6 × 85). The architecture begins with a convolutional layer (3 × 7 kernel, padding of 3, stride of 2) followed by ReLU activation post BatchNorm for channel normalization. Next, a max pooling layer (window size 3, padding 1, stride 2) is applied. The third layer integrates a residual module, featuring parallel convolutional layers (3 × 7 kernel, padding 3, stride 1) using the sigmoid activation function. The outputs are combined into a global pooling layer, leading to a final fully connected output layer with one neuron, employing the sigmoid activation function. Training uses the Adam optimizer with a batch size of 512, learning rate of 6e-6, and 250 epochs. The chosen loss function for the three deep learning networks is cross-entropy loss for binary classification.

The convLSTM-based model, designed with an input dimension of (1 × 6 × 85), starts with the Conv2DLSTMCell from the sklearn library. The output layer consists of 1 neuron, and its activation function is a fully connected layer with a sigmoid activation (Fig. 4B). Training involves the Adam optimizer, a batch size of 128, a learning rate of 0.0001, and 150 training iterations.

The input data for machine learning algorithms is one-dimensional, with each sample having a size of 6 × 85 = 510. Random Forest, part of ensemble learning's bagging method, employs multiple unconnected decision trees for quick training and feature importance assessment (Fig. 4C). We train the Random Forest model using the sklearn library, choosing

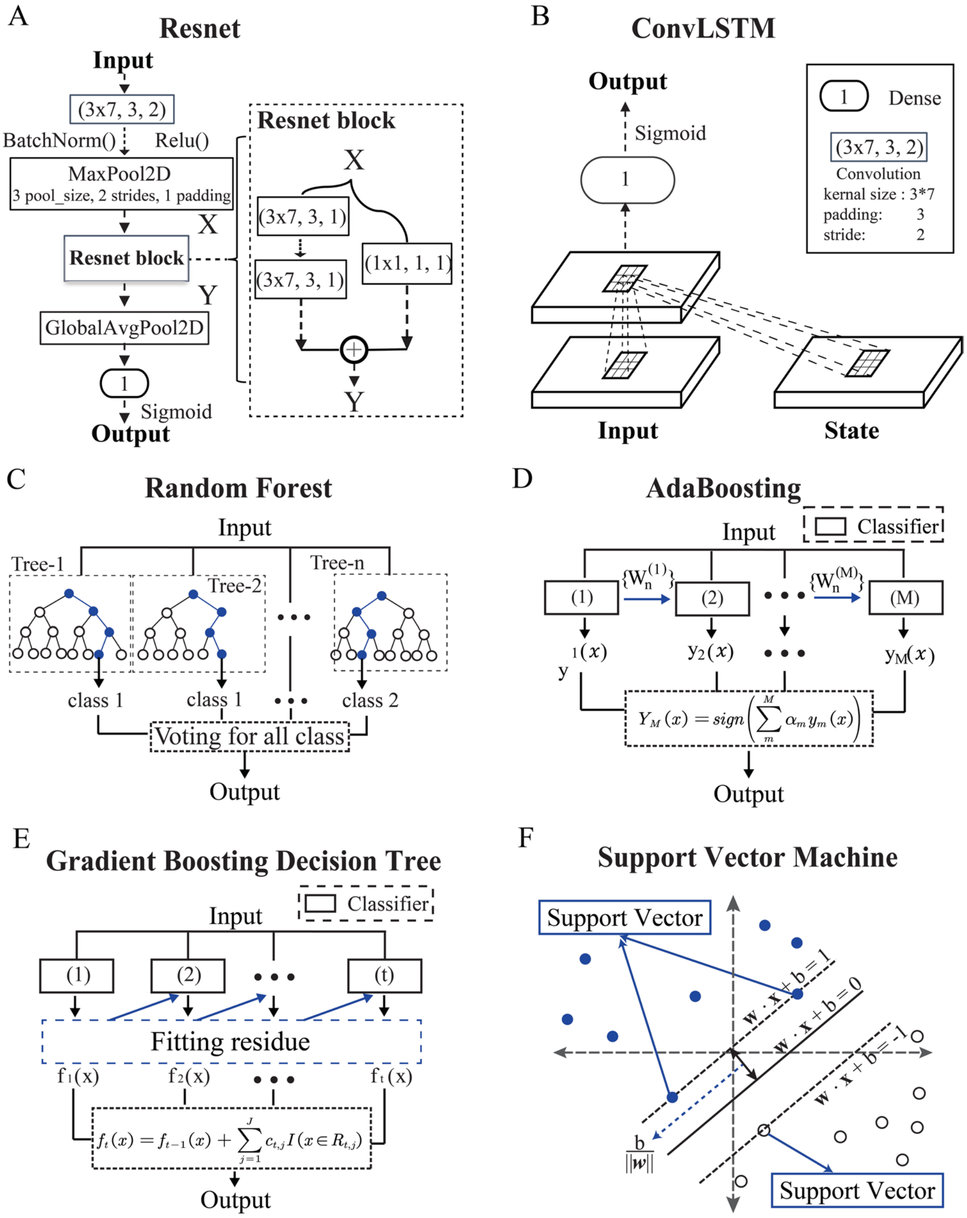


Fig. 4 Model Schematics. **A** Resnet-based deep learning model. **B** ConvLSTM-based deep learning algorithm. **C** Random Forest (RandomForest). **D** AdaBoosting. **E** Gradient boosting decision tree (GBDT). **F** Support Vector Machine (SVM)

“n_estimators” as 250 after evaluating various values (50, 100, 150, 200, 240, 250, 260, 300, and 350), while keeping other parameters default.

AdaBoosting, part of the boosting method, iteratively trains weak classifiers, adjusting sample weights based on learning errors to emphasize those with higher rates for subsequent classifiers (Fig. 4D). We train the AdaBoosting model using the sklearn library with a decision tree as the weak classifier (max_depth = 2, min_samples_split = 20, min_samples_leaf = 5). Parameters such as "n_estimators" are set to 200, "learning_rate" to 0.8, and others to their defaults.

Gradient Boosting, also a boosting method, combines weak classifiers to create a strong one using gradients in optimization (Fig. 4E). We use the GradientBoostingClassifier module from sklearn with parameters like “max_depth” set to 8, “min_samples_split" to 500, “min_samples_leaf” to 50, “max_features" to “sqrt,” “subsample" to 0.8, “random_state" to 10, “learning_rate” to 0.1, and “n_estimators" to 300 after evaluating multiple values (150, 200, 250, 300, and 350), while other parameters remain default.

The Support Vector Machine (SVM) algorithm maximizes the margin between training patterns and the decision boundary (Fig. 4F). Using sklearn, we train the sklearn.svm.SVC model with parameters such as "kernel" set to "rbf," "gamma" to 1e-6, "C" to 1e-6, while keeping other parameters default.

We assessed data correlation using Pearson and Spearman correlation coefficients. Specifically, we calculated the mean value of correlations among peaks in six fragment ion chromatograms. The formula for computing the Pearson correlation coefficient between two XICs is:

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{N\sum y_i^2 - \left(\sum y_i\right)^2}}$$

where N is the number of time points, x_i and y_i respectively represent the intensity value of a fragment ion.

The formula for calculating the Spearman correlation coefficient between two fragment ion XICs is as follows:

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2 \sum_i \left(y_i - \overline{y}\right)^2}}$$

Among them, x_i and y_i represent the intensity value of the i-th fragment ion, and $\overline{x}$ and $\overline{y}$ represent the average value of x and y.

2.7 Model Training Process of SeFilter-DIA

The training loss functions for SeFilter-DIA, ConvLSTM, and ResNet deep learning models use binary classification cross-entropy loss combined with the Sigmoid activation function, formulated as:

$$L = -\sum_i label_i * log(pred_i) + (1 - label_i) * log(1 - pred_i)$$

where *pred* is the predicted value output by the algorithm, and *label* refers to the manually labeled label, which can take the values of 0 or 1.

The SeFilter-DIA model reaches a stable cross-validation set loss after 80 iterations, concluding training at 100 iterations (Fig. 5A). The ConvLSTM-based model stabilizes its cross-validation set loss after 140 iterations, concluding training at 150 iterations (Fig. 5B). Similarly, the ResNet-based model's cross-validation set loss stabilizes after 200 iterations, with training concluding at 250 iterations (Fig. 5C). There were no instances of overfitting or underfitting, indicating successful training across these three deep learning models.

2.8 ROC Curves and P/R curves of Training and Testing Process

During the receiver operating characteristic (ROC) curve analysis, the area under the curve (AUC) represents the model's classification performance, with a larger AUC indicating better performance.

In training, the Random Forest model showed the highest AUC of 0.9999, outperforming other models (Fig. 6A, Table 2). On the cross-validation and test sets, the

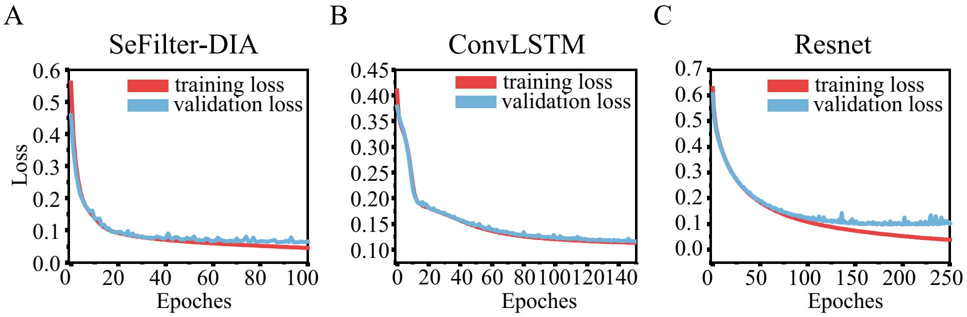


Fig. 5 Training and Cross-validation Loss Curves of Deep Learning Models. (A) SeFilter-DIA model. **B** ConvLSTM-based model. **C** Resnet-based model

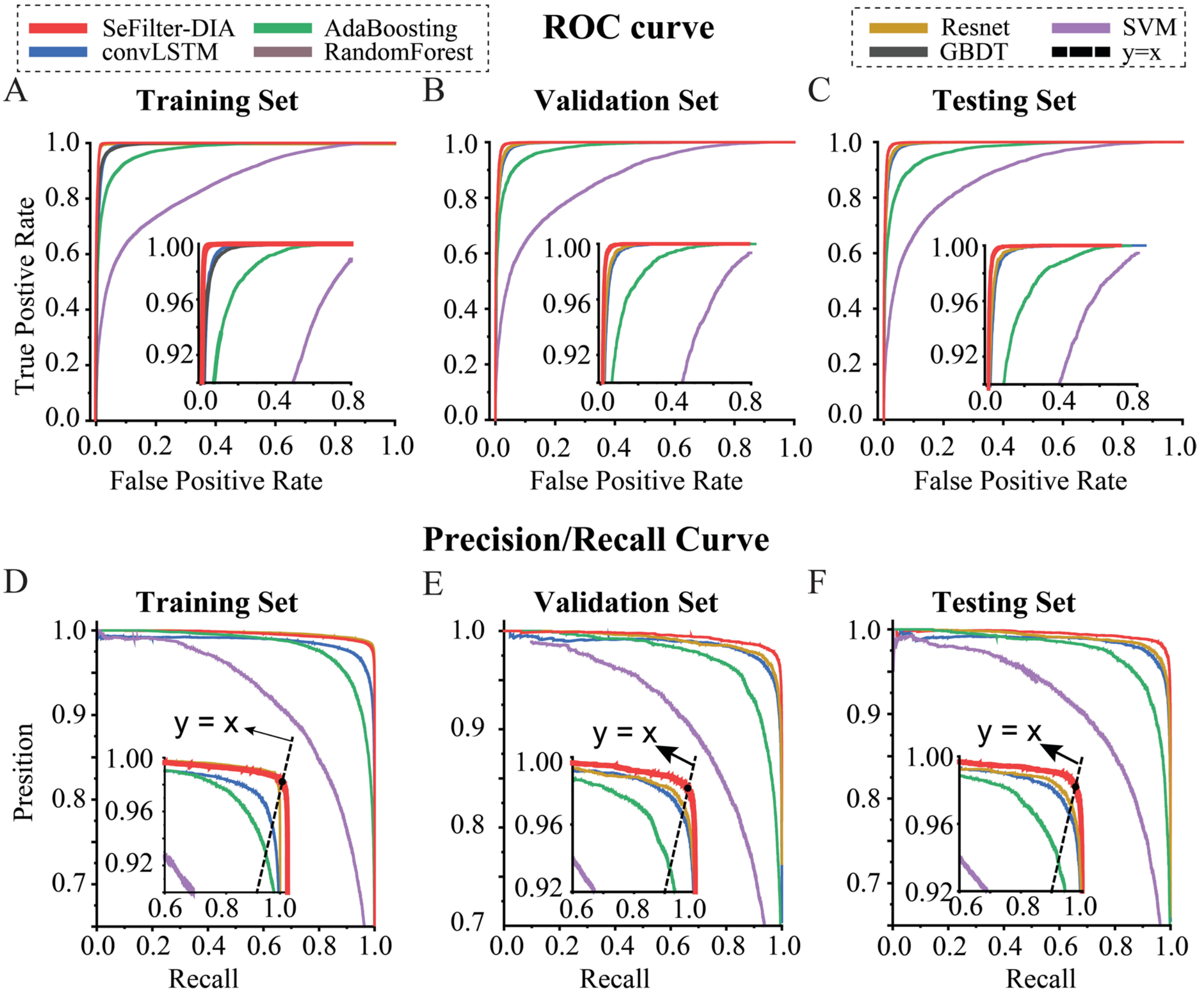


Fig. 6 ROC and Precision/Recall curves of seven models. Different colors represent distinct models: red (SeFilter-DIA), blue (convLSTM), brown (Resnet), green (AdaBoosting), purple (SVM), black (random forest), and black dashed line (GBDT). The black dotted line indicates the y = x line. The balance point, where the SeFilter-DIA curve intersects the y = x line, is marked. **A** ROC curves on the training set. **B** ROC curves on the cross-validation set. **C** ROC curves on the test set. **D** Precision/Recall curves on the training set. **E** Precision/Recall curves on the cross-validation set. **F** Precision/Recall curves on the test set

Table 2 Model Performance Metrics

Metrics\Model	AUC on Testing Set	AUC on Training Set	AUC on Validation Set	Precision	Recall	F1 score	ACC
AdaBoosting	0.9794	0.9878	0.9764	0.944	0.9007	0.9219	0.9282
RandomForest	0.985	**0.9999**	0.9826	0.9544	0.9262	0.9401	0.9447
SVM	0.9161	0.9165	0.9120	0.9636	0.4044	0.5697	0.7128
GBDT	0.9882	0.9923	0.9874	0.9501	0.9467	0.9484	0.9518
MSSort-DIA^{XMBD}	0.9593	–	–	0.9775	0.7326	0.8375	0.8665
convLSTM	0.9914	0.9912	0.9909	0.9581	0.9622	0.9601	0.9624
Resnet	0.992	0.9988	0.992	**0.9779**	0.9143	0.9451	0.9500
SeFilter-DIA	**0.9964**	0.9983	**0.9961**	0.9709	**0.987**	**0.9789**	**0.9799**

"AUC on Testing Set" signifies the AUC value on the test set. "AUC on Training Set" and "AUC on Validation Set" indicate the AUC values on training and cross-validation sets, respectively. Black bold highlights the highest value per metric. Precision is test set accuracy. Recall is the test-set recall rate. F1 score, derived from precision and recall, is the test set harmonic mean. ACC is the test set accuracy rate

SeFilter-DIA model surpassed all others, achieving AUC values of 0.9964 and 0.9961 respectively (Fig. 6B and C, Table 2). This demonstrates the SeFilter-DIA model's better performance on both validation and test sets.

The precision-recall curve evaluates the classifier's performance with a focus on positive samples. It emphasizes the balance point where the curve intersects the y = x line, with a larger area indicating better performance. On all sets (training, cross-validation, and test), SeFilter-DIA's balance point was closest to the upper right corner, showcasing its superior performance (Fig. 6D, E and F). This model excelled in both the ROC and precision-recall curves, especially crucial in the manual screening of protein peptides.

2.9 Evaluation Metrics of Model Performance

Table 2 summarizes the performance metrics of eight models across training, cross-validation, and test sets. While the random forest achieved the highest AUC of 0.9999 on the training set, it showed overfitting, with smaller AUC values around 0.985 on the cross-validation and test sets.

The SVM model exhibited poor performance on the test set with recall, F1 score, and accuracy rates of 0.4044, 0.5697, and 0.7128, respectively. In contrast, SeFilter-DIA, ConvLSTM, and Resnet consistently performed well across all metrics, surpassing 0.9. Their AUC values on all sets exceeded 0.99. MSSort-DIAXMBD showed comparatively lower performance with a test set AUC of 0.9593 and other metrics below 0.9.

Among the deep learning models, SeFilter-DIA stood out with a test set AUC of 0.9964, cross-validation set AUC of 0.9961, and recall, F1 score, and accuracy rates of 0.9870, 0.9789, and 0.9799, respectively. While Resnet had a slightly higher recall rate at 0.9779, SeFilter-DIA demonstrated superior overall classification performance.

2.10 Histogram of Probability Distribution for Classification

In the probability distribution graph for binary classification, higher histograms at both ends and a lower one in the middle indicate better discrimination between sample types.

The test set probability histogram of MSSort-DIA$^{\mathrm{XMBD}}$ shows a prominent distribution around 0–0.1, with a higher middle section compared to SeFilter-DIA (Fig. 7A). The Resnet-based model has a similar pattern to SeFilter-DIA but with slightly higher mid-histogram values (Fig. 7B). ConvLSTM, while mostly spread at both ends, has a more pronounced mid-histogram than SeFilter-DIA (Fig. 7C). Random forest exhibits a relatively uniform distribution (Fig. 7D). AdaBoosting, despite its accuracy, concentrates around 0.4–0.6, indicating limited discrimination (Fig. 7E). Gradient Boosting has higher mid-histogram values than SeFilter-DIA (Fig. 7F). SVM shows a step-like decline, implying poor resolution (Fig. 7G). In comparison, both Pearson and Spearman coefficients exhibit relatively uniform distributions (Fig. 7H and I), indicating lower discrimination power.

Among the models, SeFilter-DIA displays the least distribution in the 0.1–0.9 interval, suggesting superior discrimination between sample types and overall performance.

3 Benchmarking on Phosphorylation Modification Data

We applied the trained SeFilter-DIA model to classify peptides from phosphorylation DIA datasets. The first dataset, obtained from the ProteomeXchange Consortium (PXD014525), was analyzed using a Thermo Fisher Q Exactive HF-X mass spectrometer with an MS1 scan range of 350–1400 m/z and 64 windows in the MS2 analysis [48]. The second dataset, derived from Arabidopsis thaliana and identified by the identifier PXD027512, utilized a Thermo Fisher Orbitrap Fusion Lumos mass spectrometer operating in DIA mode, with an MS1 scan range from 350 to 1500 and segmented into 60 windows [49].

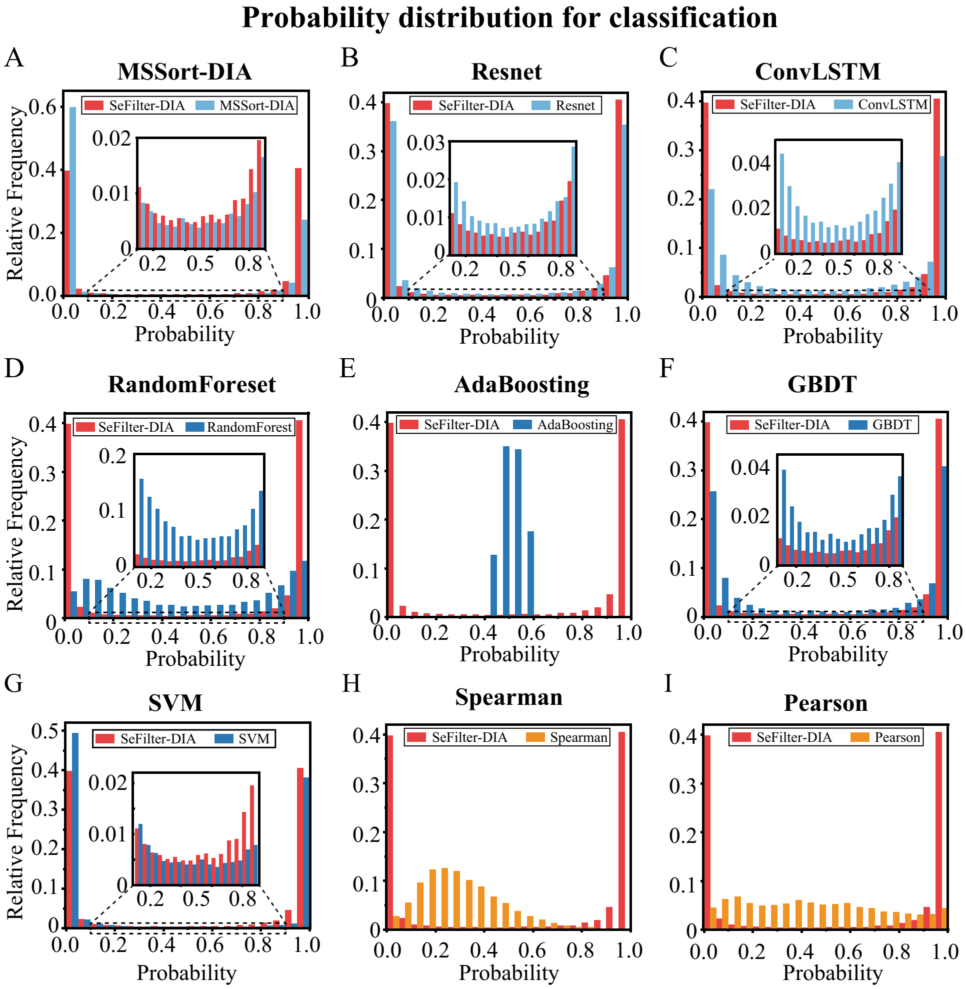
We performed analyses on these benchmark datasets using DIA-NN in library-free mode, considering phosphorylation modifications as variable factors. Subsequently, SeFilter-DIA was employed for the binary classification of peptides identified by DIA-NN. The classification performance was assessed through probability histograms, revealing a predominance of low-confidence peptides in the phosphorylation data. This underscores SeFilter-DIA's effectiveness in replacing manual screening for rapid identification (Fig. 8A and D). Additionally, XIC similarity analysis highlighted SeFilter-DIA's proficiency in distinguishing between high- and low-confidence peptides (Fig. 8).

4 Discussion

In this study, three deep-learning models were proposed for high-confidence peptide identification within DIA-NN or OpenSWATH-PyProphet-TRIC outputs. Among them, the SeFilter-DIA model, integrated a compressed excitation module with a residual network, displaying the most promising performance. It employed a process involving XIC extraction, normalization, and deep learning classification to discern high-confidence peptides.

Comparatively, SeFilter-DIA outperformed traditional machine learning algorithms like GBDT, AdaBoosting, SVM, Random Forest, MSSort-DIA$^{\mathrm{XMBD}}$, and scoring

Fig. 7 Probability Distribution Histograms for Classification. Horizontal axis: Model's Predicted Probability. Vertical axis: Relative Frequency. Red bars represent SeFilter-DIA probability distribution. **A** MSSort-DIAXMBD. **B** Resnet-based Deep Learning Model. **C** ConvLSTM-based Deep Learning Model. **D** Random Forest. **E** AdaBoosting. **F** Gradient Boosting Decision Tree (GBDT). **G** Support Vector Machine (SVM). **H** Pearson Correlation Score Distribution. **I** Spearman Correlation Score Distribution



methods based on Pearson and Spearman coefficients in cross-validation and test datasets. Despite its effectiveness, there's room for model enhancement. Addressing the complexity of DIA data could involve tailored preprocessing for different mass spectrometers, integrating data from diverse manufacturers, and exploring varied network architectures. These improvements will boost our model's capabilities. Furthermore, combining proteomics analysis with network modeling can significantly contribute to understanding regulatory mechanisms and identifying potential disease-related therapeutic targets alongside image analysis [50, 51]. The presence of low-confidence results in proteomics analysis can significantly impact subsequent applications. Se-Filter serves as an effective alternative to manual screening, enhancing the elimination of low-confidence peptides and proteins. This automated refinement not only mitigates false positive rates but also enhances the overall efficacy of proteomics in downstream applications.

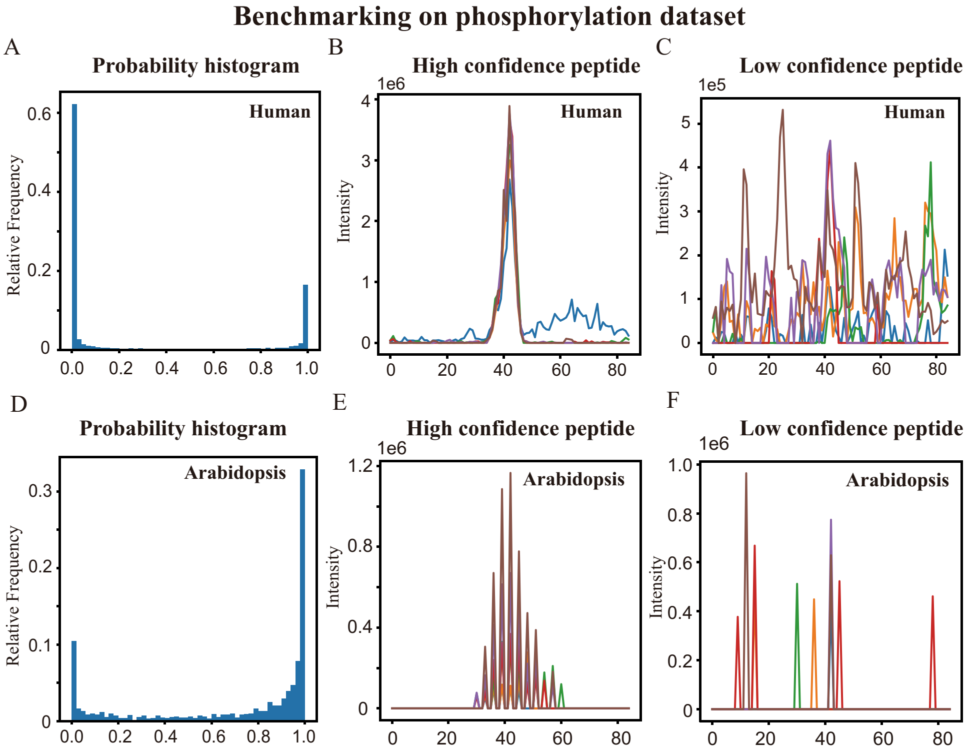


Fig. 8 Probability histograms and fragment XICs of phosphorylation modification datasets. The horizontal axis represents the predicted probability of the model, while the vertical axis displays the relative frequency. XICs show quantified intensity. **A** Probability histogram of human phosphorylation data. **B** High-confidence peptide from human phosphorylation data. **C** Low-confidence peptide from human phosphorylation data. **D** Probability histogram of Arabidopsis thaliana phosphorylation data. **E** High-confidence peptide from Arabidopsis thaliana phosphorylation data. **F** Low-confidence peptide from Arabidopsis thaliana phosphorylation data

Funding This work is supported by the Ministry of Science and Technology of the People's Republic of China (STI2030-Major Projects2021ZD0201900), the National Natural Science Foundation of China under Grant 12090052 and 11874310, the Natural Science Foundation of Fujian Province of China (Grant No. 2023J05002), and the Fundamental Research Funds for the Central Universities (Grant No. 20720230017).

Declarations

Conflict of Interest The authors declare that they have no conflicts of interest.

References

1. Zhang Y, Fonslow BR, Shan B et al (2013) Protein analysis by shotgun/bottom-up proteomics. Chem Rev 113:2343–2394. https://doi.org/10.1021/cr3003533
2. Gillet LC, Navarro P, Tate S et al (2012) Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteom 11(O111):016717. https://doi.org/10.1074/mcp.O111.016717
3. Li X, Zhong C, Wu R et al (2021) RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes. Protein Cell 12:858–876. https://doi.org/10.1007/s13238-020-00810-x
4. Li X, Zhong C, Yin Z et al (2020) Data-driven modeling identifies TIRAP-independent MyD88 activation complex and myddosome assembly strategy in LPS/TLR4 signaling. Int J Mol Sci 21:3061. https://doi.org/10.3390/ijms21093061
5. Röst HL, Rosenberger G, Navarro P et al (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol 32:219–223. https://doi.org/10.1038/nbt.2841
6. Keller A, Bader SL, Shteynberg D et al (2015) Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition Mass Spectrometry (MS) using SWATHProphet. Mol Cell Proteom 14:1411–1418. https://doi.org/10.1074/mcp.O114.044917
7. Peckner R, Myers SA, Jacome ASV et al (2018) Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat Methods 15:371–378. https://doi.org/10.1038/nmeth.4643
8. Tsou C, Avtonomov D, Larsen B et al (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods 12:258–264. https://doi.org/10.1038/nmeth.3255
9. Li Y, Zhong C, Xu X et al (2015) Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Methods 12:1105–1106. https://doi.org/10.1038/nmeth.3593
10. Meyer JG, Mukkamalla S, Steen H et al (2017) PIQED: automated identification and quantification of protein modifications from DIA-MS data. Nat Methods 14:646–647. https://doi.org/10.1038/nmeth.4334
11. Bruderer R, Bernhardt OM, Gandhi T et al (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteom 14:1400–1410. https://doi.org/10.1074/mcp.M114.044305
12. Ting YS, Egertson JD, Bollinger JG et al (2017) PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods 14:903–908. https://doi.org/10.1038/nmeth.4390

13. Sinitcyn P, Hamzeiy H, Salinas Soto F et al (2021) MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat Biotechnol 39:1563–1573. https://doi.org/10.1038/s41587-021-00968-7
14. Qian X, Qiu Y, He Q et al (2021) A review of methods for sleep arousal detection using polysomnographic signals. Brain Sci 11:1274. https://doi.org/10.3390/brainsci11101274
15. Hu H, Feng Z, Lin H et al (2023) Modeling and analyzing single-cell multimodal data with deep parametric inference. Brief Bioinform 24:bbad005. https://doi.org/10.1093/bib/bbad005
16. Wang W, Zhang L, Sun J et al (2022) Predicting the potential human lncRNA–miRNA interactions based on graph convolution network with conditional random field. Brief Bioinform 23:bbac463. https://doi.org/10.1093/bib/bbac463
17. Zhao J, Sun J, Shuai SC et al (2023) Predicting potential interactions between lncRNAs and proteins via combined graph autoencoder methods. Brief Bioinform 24:bbac527. https://doi.org/10.1093/bib/bbac527
18. Zhong J, Song Z, Zhang L et al (2022) Assembly of guanine crystals as a low-polarizing broadband multilayer reflector in a spider, phoroncidia rubroargentea. ACS Appl Mater Interfaces 14:32982–32993. https://doi.org/10.1021/acsami.2c09546
19. Chen X, Zhu R, Zhong J et al (2022) Mosaic composition of RIP1–RIP3 signalling hub and its role in regulating cell death. Nat Cell Biol 24:471–482. https://doi.org/10.1038/s41556-022-00854-7
20. Wang J, Chen F, Ma Y et al (2023) XBound-former: toward cross-scale boundary modeling in transformers. IEEE Trans Med Imaging 42:1735–1745. https://doi.org/10.1109/tmi.2023.3236037
21. Gessulat S, Schmidt T, Zolg DP et al (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 16:509–518. https://doi.org/10.1038/s41592-019-0426-7
22. Tran NH, Qiao R, Xin L et al (2019) Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nat Methods 16:63–66. https://doi.org/10.1038/s41592-018-0260-3
23. Tiwary S, Levy R, Gutenbrunner P et al (2019) High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods 16:519–525. https://doi.org/10.1038/s41592-019-0427-6
24. Zhou X, Zeng W, Chi H et al (2017) pDeep: predicting MS/MS spectra of peptides with deep learning. Anal Chem 89:12690–12697. https://doi.org/10.1021/acs.analchem.7b02566
25. Yang Y, Liu X, Shen C et al (2020) In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun 11:146. https://doi.org/10.1038/s41467-019-13866-z
26. He Q, Zhong C, Li X et al (2023) Dear-DIAXMBD: deep autoencoder enables deconvolution of data-independent acquisition proteomics. Research 6:0179. https://doi.org/10.34133/research.0179
27. Gao M, Yang W, Li C et al (2021) Deep representation features from DreamDIAXMBD improve the analysis of data-independent acquisition proteomics. Commun Biol 4:1190. https://doi.org/10.1038/s42003-021-02726-6
28. Demichev V, Messner CB, Vernardis SI et al (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods 17:41–44. https://doi.org/10.1038/s41592-019-0638-x
29. MacLean B, Tomazela DM, Shulman N et al (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–968. https://doi.org/10.1093/bioinformatics/btq054
30. Sturm M, Kohlbacher O (2009) TOPPView: an open-source viewer for mass spectrometry data. J Proteome Res 8:3760–3763. https://doi.org/10.1021/pr900171m
31. Li Y, He Q, Guo H et al (2022) MSSort-DIAXMBD: A deep learning classification tool of the peptide precursors quantified by OpenSWATH. J Proteomics 259:104542. https://doi.org/10.1016/j.jprot.2022.104542
32. Gupta S, Sing J, Mahmoodi A et al (2020) DrawAlignR: an interactive tool for across run chromatogram alignment visualization. Proteomics 20:1900353. https://doi.org/10.1002/pmic.201900353
33. Tatjana V, Domitille S, Jean-Charles S (2021) Paraquat-induced cholesterol biosynthesis proteins dysregulation in human brain microvascular endothelial cells. Sci Rep 11:18137. https://doi.org/10.1038/s41598-021-97175-w
34. Midha MK, Kusebauch U, Shteynberg D et al (2020) A comprehensive spectral assay library to quantify the Escherichia coli proteome by DIA/SWATH-MS. Sci Data 7:389. https://doi.org/10.1038/s41597-020-00724-7
35. Navarro P, Kuharev J, Gillet LC et al (2016) A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol 34:1130–1136. https://doi.org/10.1038/nbt.3685
36. Muntel J, Kirkpatrick J, Bruderer R et al (2019) Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. J Proteome Res 18:1340–1351. https://doi.org/10.1021/acs.jproteome.8b00898
37. Chambers MC, Maclean B, Burke R et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920. https://doi.org/10.1038/nbt.2377
38. Reiter L, Rinner O, Picotti P et al (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods 8:430–435. https://doi.org/10.1038/nmeth.1584
39. Röst HL, Liu Y, D'Agostino G et al (2016) TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat Methods 13:777–783. https://doi.org/10.1038/nmeth.3954
40. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: An open-source MS/MS sequence database search tool. Proteomics 13:22–24. https://doi.org/10.1002/pmic.201200439
41. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467. https://doi.org/10.1093/bioinformatics/bth092
42. Keller A, Nesvizhskii AI, Kolker E et al (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392. https://doi.org/10.1021/ac025747h
43. Shteynberg D, Deutsch EW, Lam H et al (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteom. https://doi.org/10.1074/mcp.M111.007690
44. Lam H, Deutsch EW, Eddes JS et al (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteom 7:655–667. https://doi.org/10.1002/pmic.200600625
45. Shi X, Chen Z, Wang H et al (2015) Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. Proceed Int Conf Neural Inform Process Syst 1:802–810. https://doi.org/10.5555/2969239.2969329
46. He K, Zhang X, Ren S et al (2016) Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016:770–778. https://doi.org/10.1109/CVPR.2016.90
47. Hu J, Shen L, Albanie S et al (2020) Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 42:2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
48. Bekker-Jensen DB, Bernhardt OM, Hogrebe A et al (2020) Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat Commun 11:787. https://doi.org/10.1038/s41467-020-14609-1

49. Zhou Q, Meng Q, Tan X et al (2021) Protein phosphorylation changes during systemic acquired resistance in *Arabidopsis thaliana*. Front Plant Sci. https://doi.org/10.3389/fpls.2021.748287
50. Li X, Zhang P, Yin Z et al (2022) Caspase-1 and gasdermin d afford the optimal targets with distinct switching strategies in NLRP1b inflammasome-induced cell death. Research 2022:9838341. https://doi.org/10.34133/2022/9838341
51. Xu F, Miao D, Li W et al (2023) Specificity and competition of mRNAs dominate droplet pattern in protein phase separation. Phys Rev Res 5:023159. https://doi.org/10.1103/PhysRevResearch.5.023159

Authors and Affiliations

Qingzu He[1,2] · Huan Guo[1] · Yulin Li[1] · Guoqiang He[2] · Xiang Li[1] · Jianwei Shuai[2,3]

✉ Xiang Li
xianglibp@xmu.edu.cn

✉ Jianwei Shuai
jianweishuai@xmu.edu.cn

[1] Department of Physics, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China

[2] Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325001, China

[3] Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou 325001, China