Journal of proteome research

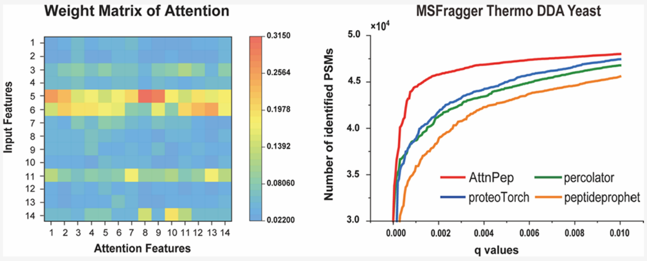
pubs.acs.org/jpr Article

AttnPep: A Self-Attention-Based Deep Learning Method for Peptide Identification in Shotgun Proteomics

Yulin Li, Qingzu He, Huan Guo, Stella C. Shuai, Jinyan Cheng, Liyu Liu, and Jianwei Shuai*

Cite This: *J. Proteome Res.* 2024, 23, 834−843 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In shotgun proteomics, the proteome search engine analyzes mass spectra obtained by experiments, and then a peptide-spectra match (PSM) is reported for each spectrum. However, most of the PSMs identified are incorrect, and therefore various postprocessing software have been developed for reranking the peptide identifications. Yet these methods suffer from issues such as dependency on distribution, reliance on shallow models, and limited effectiveness. In this work, we propose AttnPep, a deep learning model for rescoring PSM scores that utilizes the Self-Attention module. This module helps the neural network focus on features relevant to the classification of PSMs and ignore irrelevant features. This allows AttnPep to analyze the output of different search engines and improve PSM discrimination accuracy. We considered a PSM to be correct if it achieves a *q*-value <0.01 and compared AttnPep with existing mainstream software PeptideProphet, Percolator, and proteoTorch. The results indicated that AttnPep found an average increase in correct PSMs of 9.29% relative to the other methods. Additionally, AttnPep was able to better distinguish between correct and incorrect PSMs and found more synthetic peptides in the complex SWATH data set.

KEYWORDS: *shotgun proteomics, mass spectrometry, peptide identification, deep learning, self-attention*

■ INTRODUCTION

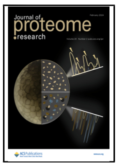
Tandem mass spectrometry (MS/MS) has become the most widely utilized choice for the characterization of proteins in complex biological samples, as it can acquire millions of spectra in a single experiment.[1] Technological advances in mass spectrometry have promoted the development of database search engines such as Sequest,[2] Mascot,[3] Comet,[4,5] MS-GF+,[6] and MSFragger.[7] Database search algorithms assign a score for each peptide-spectrum match (PSM) by measuring the similarity between the theoretically predicted spectrum and the experimentally acquired spectrum.[8] The scored PSMs are ranked, and only the top 1 match for each spectrum will typically be reported by the database search engine. However, the PSM scores resulting from database searches are often uncalibrated,[9,10] which leads to challenges in qualitative comparisons and may result in incorrect interpretations and conclusions.

The target-decoy strategy[11] is developed to estimate the number of false-positive protein identifications in a more systematic way. Data is not only searched against the standard sequence database (target) but also against a reversed protein database (decoy).[12] PSMs obtained from the decoy database can be used to estimate the number of incorrect target PSMs and estimate the PSM-level false discovery rate (FDR). Here, the PSM-level FDR is represented by the *q*-value, which is defined as the minimal FDR at which a PSM is identified as correct.

A variety of approaches have been developed to validate PSMs reported by search engines.[13] In particular, there are two kinds of software widely used in current proteomics analysis pipelines. The first one is PeptideProphet,[14,15] which uses a linear discriminant analysis (LDA) to assess the validity of peptide identifications. The second one, Percolator,[16−19] trains a machine learning model called support vector machine (SVM)[20] to discriminate between target and decoy PSMs. Compared to machine learning models, deep learning models have yielded better results for proteomics data analysis.[21−27] In addition, a deep learning-based algorithm called proteoTorch using deep neural networks (DNNs) further improves the classification of correct target PSMs.[28]

However, while these algorithms have demonstrated state-of-the-art performance in peptide identification, there still remains substantial room for improvement. For example, PeptideProphet incorporates all discrimination properties to model the distributions of correct and incorrect identification scores, but the deviations of real and observed distributions can lead to substantial underestimation or overestimation of computed probabilities;[15] Percolator's adaptive algorithm can postprocess arbitrary sets of MS/MS features,[29−31] but the use of shallow

Received: November 3, 2023
Revised: December 23, 2023
Accepted: January 9, 2024
Published: January 22, 2024



https://doi.org/10.1021/acs.jproteome.3c00729

machine learning models is potentially suboptimal given the cutting-edge performance of deep models in large-scale data analysis; proteoTorch uses the DNNs to improve MS/MS postprocessing accuracy but it has a marginal improvement in efficacy.[28]

In this study, we describe a solution to these problems that uses a Self-Attention-based deep learning model that can be appended to any database search algorithm without adjustment. The algorithm, called AttnPep, is inspired by the Transformer model[32] in the field of Natural Language Processing (NLP).[33] The Self-Attention module[34] in Transformer is used to learn the relationships of sequence elements with one another and how their interaction should be interpreted. AttnPep takes advantage of the Self-Attention module by directly computing the interaction among features. This helps AttnPep enhance the model's understanding of complex relationships in the data, improve representation capability, and handle nonlinear problems. These benefits can improve the accuracy and performance of classification tasks. Meanwhile, we offer an optional algorithm, called AttnProt, which combines a bipartite graph[35] and statistical model to compute probabilities that proteins are presented in a parsimonious protein list.

We measured AttnPep's ability to identify correct PSMs using 13 high-res MS1/MS2 data sets collected from two different protein databases and compared the number of peptide identifications at q-values less than 0.01 with existing postprocessing software PeptideProphet, Percolator, and proteoTorch. We show that AttnPep significantly improves the classification of correct and incorrect PSMs despite the different acquisition methods, species, or different search engine outputs of these data sets. For example, in the output of MSFragger software for four different sets of species samples and in the output of different search engines for the same species samples, at 1% PSM-level FDR, AttnPep improved results by 9.29% on an average compared to other softwares. In addition, for the SWATH-MS gold standard (SGS) data set, which is a data set consisting of 422 chemically synthesized, stable isotope-labeled standard peptides from samples,[36] AttnPep was able to better distinguish the distribution of correct and incorrect PSMs compared to Percolator and identified more synthetic peptides at a q-value threshold of 0.01. Moreover, we demonstrate that AttnPep amplifies the weights of hyperscore and log10 *e*value through self-attention weights, which are the most relevant features for input during training on MSFragger output features.

■ METHODS

Data Sets and Processing

We used two different high-resolution publicly available data sets. One is a comprehensive DDA/DIA data set that utilizes several of the most commonly used current mass spectrometers to obtain protein mass spectrometry data containing human, yeast, *E. coli*, and mixed samples, which is available through ProteomeXchange (PXD028735).[37] Another is the SGS data set consisting of 422 chemically synthesized, stable isotope-labeled standard peptides from samples that were subjected to the acquisition of protein mass spectrometry data in DIA mode on the AD SCIEX TripleTOF 5600 system. This data set is obtained from the PeptideAtlas raw data repository (PASS00289).[36] The data sets used for testing is shown in Table 1.

Table 1. Dataset Structure

species	instrument	vendor	acquisition method
E. coli	Orbitrap QE HF-X	Thermo Fisher	DDA
	TripleTOF 5600	AB Sciex	DDA SWATH
human	Orbitrap QE HF-X	Thermo Fisher	DDA
	TripleTOF 5600	AB Sciex	DDA SWATH
	TripleTOF 6600+	AB Sciex	SWATH
yeast	Orbitrap QE HF-X	Thermo Fisher	DDA
	TripleTOF 5600	AB Sciex	DDA SWATH
QC (12.5% *E. coli* +22.5% yeast +65% human)	Orbitrap QE HF-X	Thermo Fisher	DDA
	TripleTOF 5600	AB Sciex	DDA SWATH

For the conversion of the raw data, we used msconvert to convert the DDA data to mzML format and qtofpeakpicker to convert the SWATH data to mzXML format.

We used MSFragger to perform database searches for all data sets using the corresponding fasta libraries. The DDA data were analyzed using the default schema and the SWATH data were analyzed using the DIA_DIA-Umpire_SpecLib_Quant schema. All searches were fully tryptic, the mass Modifications of data 5600_SWATH_Human is also K+8.014199 and R+10.008269, and the rest of the data all use the default parameters. We specifically used multiple search engines for 6600_DDA_QC data in addition to MSFragger, Comet, and MS-GF+ for the database search, of which the search parameters are the same as MSFragger. The major database search parameters are listed in Supporting Information Table S2. All search results were output as pepXML and PIN format for subsequent downstream analysis.

AttnPep Model

Model Structure. AttnPep takes the PSM score result output by the search engine as input and ultimately returns a probability representing the correctness of each PSM. AttnPep consists of three main modules: Input Embedding, Self-Attention, and a decoder consisting of resNet and fully connected (FC) strata.

The Input Embedding module transforms each normalized input feature into a 64-dimensional feature vector. This process converts the input features of each set of PSMs into a feature matrix that can be used for self-attention calculation. The Self-Attention module (see Figure 1b) contains four FC layers and a Scaled Dot-Production Attention. The feature matrix output by the Input Embedding module is transformed by the first three FC layers and returns Q, K, and V matrices separately. The Q, K, and V are used to calculate a new feature matrix Z by Scaled Dot-Production Attention (further details of Scaled Dot-Production Attention are described in the next section), and then the matrix Z is transformed into an output feature matrix by the last FC layer. The feature matrix output by Self-Attention passes through the residual layer of dimension 512, the FC layer, and the residual layer. Then, the final sigmoid function normalizes it to the interval [0,1].

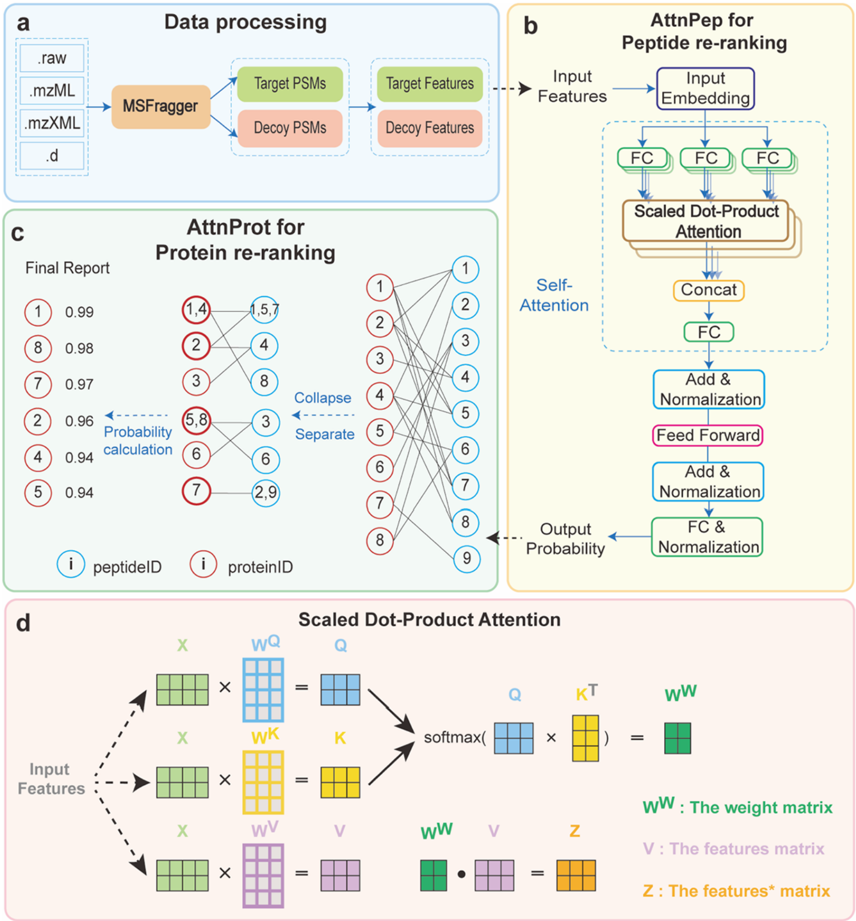


Figure 1. AttnValid workflow. (a) First, raw data are searched for the Target/Decoy database by the search engine, and the PSM results are subjected to feature extraction for AttnPep training. (b) Second, Target/Decoy features are fed into AttnPep to train a classifier for recalculating the confidence level for each PSM reported. (c) Third, the PSMs reported by AttnPep are sent to AttnProt to compute the probability of each protein. (d) Details for Scaled Dot-Product Attention.

Training. AttnPep uses an iterative, semisupervised training procedure to recalibrate input targets and decoy PSMs. Positive training examples are estimated as the set of target PSMs with scores achieving a stringent, user-specified *q*-value, while all decoy PSMs were labeled as negative samples. In each training iteration, all PSMs will be rescored and the positive label assignments will be updated. This overall process repeats either for a user-specified number of iterations or until convergence. We use the Adam optimizer with an initial learning rate of 0.001, 32 samples per batch, a training step of 15, and a loss function of Binary Cross Entropy with Logits Loss. The benefit of the iterative semisupervised learning paradigm is that the classifier is free to exploit a variety of specific features of the data, without overfitting to a particular type of spectrum.

Moreover, in order to mitigate the risk of overfitting and enhance the generalizability of the model, a 3-fold cross-validation technique is employed within the procedure. Subsequently, three separate models are trained using each of these test and train splits. This approach effectively safeguards against overfitting issues that may arise from the over-reliance on learned parameters. Consequently, the resulting three models are utilized to reassess all the PSMs. The final aggregate score is computed by weighting and summing the scores generated by the three models, then continues the *q*-value calculation for PSMs and considers PSMs with *q*-value <0.01 to be the correct PSMs.

Self-Attention Calculation

The feature matrix that is output by the Input Embedding module serves as the input for the Self-Attention module, and we denote the feature matrix as X. When X is input into the Self-Attention module, it will be transferred to three different matrices (Q (Query), K (Key), and V (Value)) by three different FC layers, which defines three different weighted matrices (W^Q, W^K, and W^V). Among them, W^Q, W^K, and W^V are the trainable parameter matrices of X linearly mapped to Q, K, and V.

The Q, K, and V are used to calculate the Scaled Dot-Product Attention by eq 1:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the spatial latitude where the eigenvectors are located. In Scaled Dot-Product Attention, we first dot-product the matrix Q with the matrix K. The purpose of doing so is to calculate the inner product between different feature vectors as a way to reflect the association between feature vectors. In order to prevent the oversized correlation matrix QK^T from affecting the subsequent calculation, we divide it by $\sqrt{d_k}$ to reduce the variance of the correlation matrix. Next, we calculate softmax for each column of the correlation matrix, which makes the correlation of each feature vector with the other feature vectors

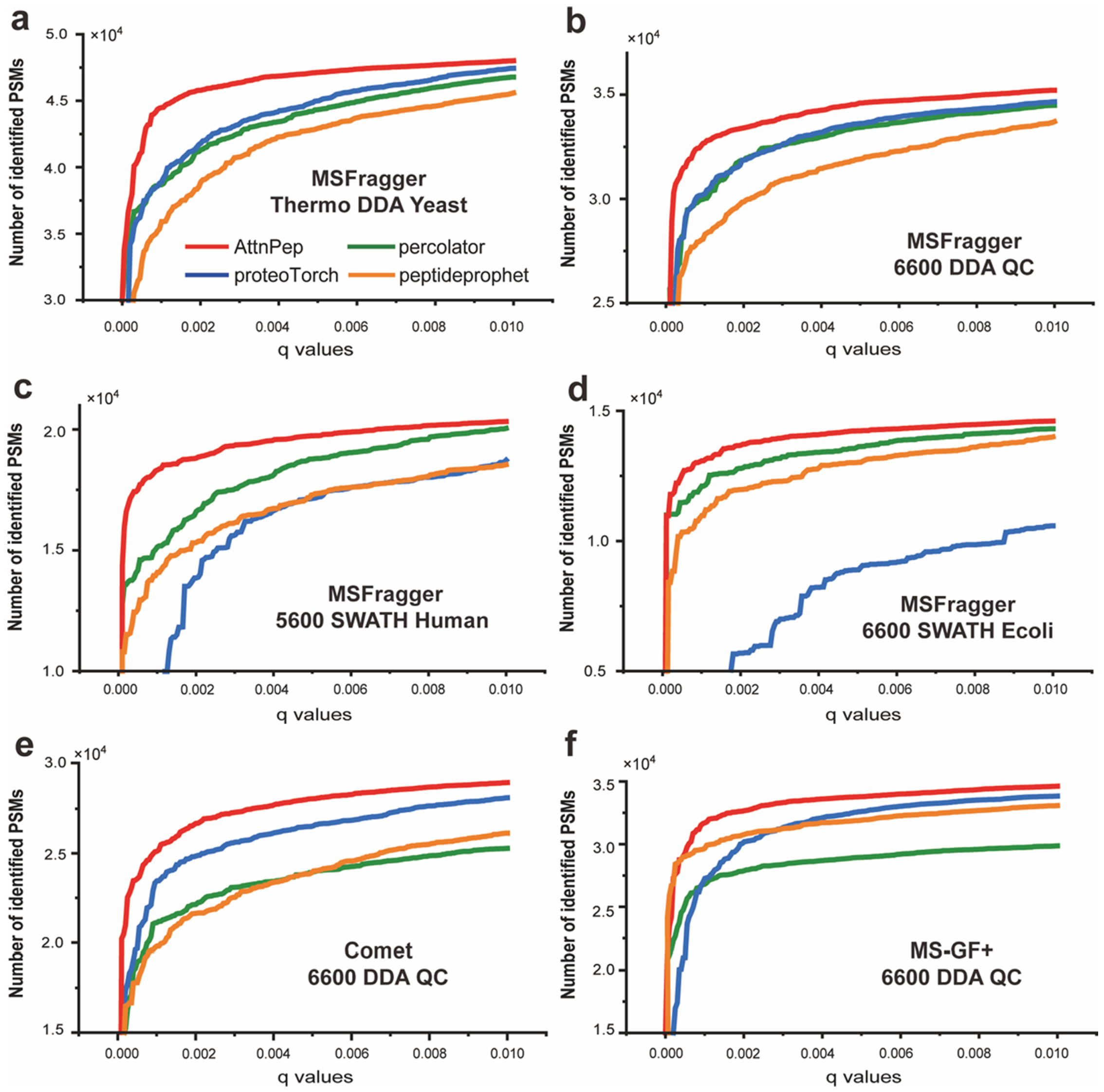


Figure 2. Performance of AttnPep on different data sets. The graph plots the number of identified PSMs as a function of *q*-value. (a) shows the data results of Yeast samples under DDA acquisition mode of Thermo instrument; (b) shows the data results of mixed samples QC under DDA acquisition mode of 6600 instrument; (c) shows the data results of SGS Human samples under 5600 instrument SWATH acquisition mode; (d) shows the data results of *E. coli* samples in 6600 instrument SWATH mode; (e) shows the data results under Comet; and (f) shows the data results under MS-GF+.

sum up to 1. Finally, we multiply the correlation matrix softmax$\left(\frac{QK^T}{\sqrt{d_k}}\right)$ with the matrix V to obtain matrix Z. The output Z of Scaled Dot-Product Attention represents the dependencies between each feature and other features, where the feature vectors of each feature have been taken into account for their correlations and weights with others (see Figure 1d).

■ RESULTS

Workflow of AttnValid

AttnValid is a workflow for validating PSMs' output from database search engines, which first validates PSMs for peptides using the deep learning model AttnPep, followed by the protein reranking model AttnProt using a bipartite graph-based probabilistic (see Figure 1).

The workflow starts with raw data processing (see Figure 1a), where a search engine (MSFragger, Comet, and MS-GF+ were used in this study) performs a search of the theoretical database for raw spectra using the Target/Decoy strategy. We divided the reported PSMs into Target and Decoy PSMs according to whether the database matched by the experimental spectra was Target or Decoy. Then the score information were packaged as the input features for subsequent deep learning models.

To obtain the training data of AttnPep, the scores most relevant to the Target/Decoy Features (e.g., *e*values) are filtered as the initial scores of the data. A user-specified PSM-level FDR filter is then applied to the Target Features using Decoy Features, so that the Target Features with higher confidence are selected as positive samples for training the AttnPep. These positive and negative samples are then input into AttnPep for training (see Figure 1b). These feature matrices are fed to the Self-Attention module (Figure 1d and Methods section) for attention computation, where the resulting feature matrices "notice" the features that are more useful for the classification task and "ignore" the features that are less important. Subsequently, we decode these Feature matrices by using a multilayer residual network and normalization layer, and output an AttnPep score, the size of which can be used to characterize

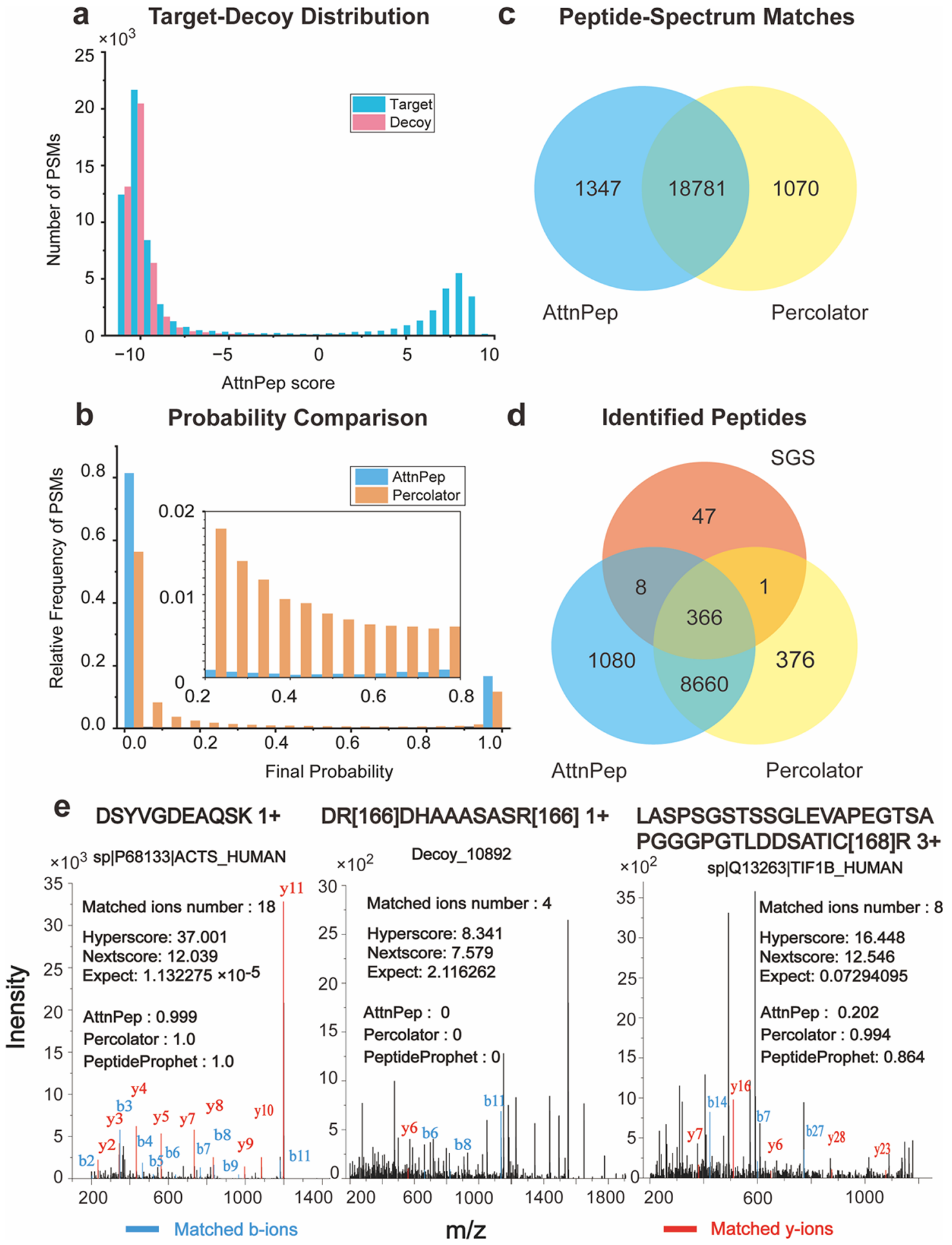


Figure 3. AttnPep shows better performance on the SGS data set. (a) Distribution of Target and Decoy PSMs in AttnPep scores. (b) Confidence probabilities reported at the end of AttnPep versus Percolator. (c) Venn diagram of PSM filtering results with the FDR of 1%. (d) Venn diagram of peptide filtering results for the FDR of 1%. (e) Annotated PSMs output by MSFragger.

the confidence of the input PSMs. Finally, we normalize the AttnPep score to obtain the probability of being correct in characterizing the PSMs. The output of AttnPep is the same as the output of Percolator.

The peptide−protein match output by AttnPep were packaged as input to AttnProt (Figure 1c). AttnProt combines the benefits of existing protein-identifying software ProteinProphet[38] and IDPicker.[39,40] These complex peptide−protein correspondences can be viewed as an undirected bipartite graph model. After initialization, merging, segmentation, filtering, and probability calculation, we get a minimal set of proteins that can cover all peptides reported by the search engine.

AttnPep Improves Peptide Identification

We first analyzed the database search results of MSFragger using AttnPep and three existing software with different strategies and drew the curve of the number of peptides identified with the *q*-value strictly less than 0.01 as a function (Figures 2, S1, and S2) and calculated the average increase rate of the *q*-value from 0 to 0.01. In both data sets where the acquisition mode was the DDA strategy (Figure 2a,b), AttnPep showed a significant improvement compared to the existing software. In the data set MSFragger_Thermo_DDA_Yeast, AttnPep increased the number of identified peptides by 6.01% compared to proteoTorch, by 7.00% compared to percolator, and by 10.68% compared to PeptideProphet. In the mixed data set MSFragger_6600_DDA_QC, AttnPep identified 3.71% more peptides than proteoTorch, 4.08% more than percolator, and

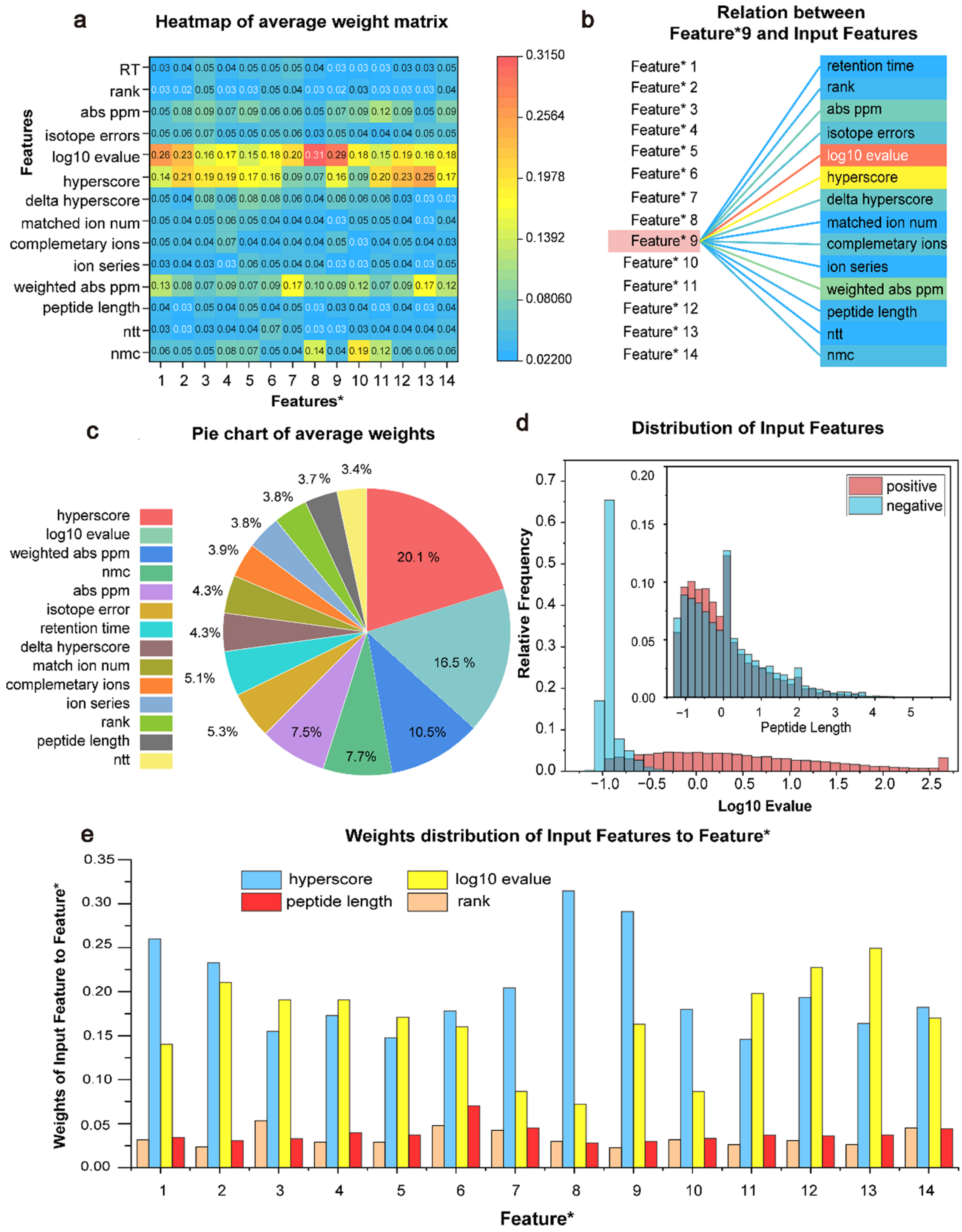


Figure 4. Schematic representation of the weights of Features in the Self-Attention module. (a) Attention is weighted by the original features (line labels) to obtain a new set of features*. (b) Feature*9 is calculated by weighing the original features. (c) The average contribution of input features to the weights of features*. (d) The distribution of positive (log10 *e*value) and negative (Peptide Length) samples of input features. (e) Histogram of the weight distribution of input features (hyperscore, log10 *e*value, peptide length, and rank) on features*.

8.18% more than PeptideProphet. As for the two data sets where the acquisition mode was SWATH strategy (Figure 2c,d): in the data set MSFragger_5600_SWATH_Human, AttnPep increased the identified peptides by 8.01% over percolator and by 14.44% over PeptideProphet, in the data set MSFragger_6600_SWATH_Ecoli, AttnPep increased the identified peptides by 4.58% over percolator, and 10.61% over PeptideProphet. We also present the results of rescored peptide identifications of AttnPep with the output of Comet and MS-GF+ (Figure 2e,f). In Comet_6600_DDA_QC, AttnPep increased the identified peptides by 6.46% over proteoTorch, by 18.28% over PeptideProphet, and by 18.05% over percolator. In MS-GF+_6600_DDA_QC, AttnPep identified 7.15% more peptides than proteoTorch, 4.87% more peptides than PeptideProphet, and 16.59% more peptides than percolator.

We can see that the growth slope of AttnPep in the interval of q-values less than 0.001 is significantly larger than other methods in all data sets, indicating that AttnPep can learn the difference between Target and Decoy PSMs more accurately and thus achieve better classification performance.

Performance of AttnPep on the SGS Data Set

To validate the accuracy of the PSMs reported by AttnPep, we tested the results using the SGS data set. If the method of reranking the peptide identifications can better distinguish the distribution of Target and Decoy, and can find more synthetic peptides in the peptide output results, the more accurate the method is proven to be.

In Figure 3a, we can see that the distribution of Decoy is in [−10,−3], while the distribution of Target is distinctly split into two. One part of the Target score distribution is in [0,10], and the other part is similar to Decoy's distribution. Figure 3b shows the confidence probability distribution of the AttnPep output and the percolator output. We can see that a clear part of the percolator distribution is distributed in [0.1,0.8], which indicates that AttnPep is able to produce a stricter confidence probability calculation for PSMs.

Figure 3c shows the Venn diagram of the PSMs reported by AttnPep and percolator under the filtering of 1% PSM-level FDR. There are 18,781 PSMs in common between AttnPep and percolator, 1347 independent PSMs in AttnPep, and 1070 independent PSMs in percolator. AttnPep reported 277 more independent PSMs than percolator. Figure 3d shows a Venn diagram of the number of peptides at 1% FDR with the number of synthetic peptides. AttnPep and Percolator together found 366 synthetic peptides on SGS, and AttnPep identified 1080 independent peptides, of which 8 were independent synthetic peptides. Percolator identified 376 independent peptides, of which 1 was an independent synthetic peptide.

Furthermore, we utilized pyteomics to annotate and visualize three randomly selected PSMs reported by MSFragger. As shown in Figure 3e, in addition to annotating the theoretical peptide and protein, we also reported three scores (hyperscore, nextscore, and expect) generated during the matching process, as well as the probabilities obtained from PSM reranking software AttnPep, Percolator, and PeptideProphet.

We conclude that AttnPep has improved the number of reports at the PSM level and especially at the peptides level, where more synthetic peptides could be identified.

Self-Attention Allows AttnPep to Focus More on Useful Features

AttnPep's core is the Self-Attention module, which was first proposed in the field of NLP to compute the degree of association between different words in a sentence. In NLP, the Self-Attention mechanism emerged to compute the connections between complex inputs by adding a Self-Attention layer after the input layer of the network, so that the input vector of each element in turn extracts useful information from each other before being input to the subsequent network for computation.

The score output by the search engines are not completely independent of each other, and some of the scores are not helpful for PSM classification. To address this, Self-Attention was introduced into AttnPep. This helps to simultaneously achieve correlation between extracted features and reduce the effect of invalid scores. Self-Attention can also accommodate scores from different search engines since the composition of scores from each search engine is different.

To explore how the Self-Attention module processes the inputs to the model, the weight matrix was extracted from AttnPep. The weight matrix is the transformation matrix learned by the Self-Attention module when converting the input features to Feature*. Feature* represents the new features learned by the network using Self-Attention. Figure 4a shows the ten-tine average weight matrix heatmap for the data MSFragger_6600_DDA_QC. In Figure 4a, there are several input features that contribute particularly well to the computation of Feature*. In Figure 4b, the computation of Feature*9 is a weighted summation of the input features, with log10 *e*value and hyperscore making the highest contribution.

Figure 4c shows the average contribution of input features to Feature* pairs in the weight matrix, thus reflecting the correlation between input features and Feature*. Among them, hyperscore contributes the highest weight pairs on average, followed by log10_*e*value, weighted abs ppm, nmc, and abs ppm in that order. The total weight contribution of these five input features accounts for a total of 62.3%. This is understandable because these five features are the key features to distinguish positive and negative samples when performing the classification of PSM (Figure 4d, large panel). Meanwhile features like peptide length, rank, etc. (Figure 4d, small panel) only reflect the basic information on the spectra or theoretical peptides and are not helpful for the classification of PSM. Figure 4e plots the distribution of the weights of the two highest and two lowest input features on average contributing to Feature*.

Therefore, we believe that the Self-Attention module can focus attention on the input features that contribute to the classification task and reduce the influence of useless features, which allows us to avoid the need to filter features when using scores from different search software.

AttnProt Combines Bipartite Graphs to Calculate Protein Probabilities

AttnPep is able to recalibrate the PSMs at the peptide level but is unable to score PSMs at the protein level. There are two major challenges in protein-level quality control. The first challenge is that the same peptide will correspond to multiple proteins, and large-scale data sets often contain homologous sequences from multiple species, thus generating many false-positive proteins. The second challenge is that the false-positive rate at the protein level is not equivalent to the false-positive rate at the peptide level, and therefore the proteins cannot be effectively filtered using only the peptide-level PSM scores.

IDPicker addresses the first challenge by using the principle of parsimony to filter out the smallest set of proteins that explain the spectra generated by the experiment through an undirected bipartite graph approach that can significantly reduce the number of homologous proteins. For the second challenge, ProteinProphet uses a probabilistic model to assemble candidate peptides with high-scoring values into candidate proteins and gives each candidate protein a scoring value characterizing its confidence level. There are many strategies to deal with those challenges.[41] For example, one straightforward strategy makes use of decoy protein groups (PGs), i.e., PGs consisting entirely of decoy PSMs. A score is constructed for each target and decoy PG, allowing the protein-level FDR to be controlled in a similar way as the PSM FDR is controlled.[42−44] Another less common category of approaches uses the PSM probabilities to calculate a probability for each protein; these probabilistic approaches perhaps more accurately represent the nonbinary evidence for the proteins and do not necessarily require protein grouping.[45−47]

We therefore propose a protein quality control tool, AttnProt, that combines the functions of IDPicker and ProteinProphet to calculate the probability of a protein by combining an undirected bipartite graph and probability statistics. We first use the method of an undirected bipartite graph to cluster proteins by shared peptides and derive the minimum list of proteins. Then we use the statistical method to estimate the probability of protein presence in a sample by the probability of peptide presence. The whole process is divided into 5 steps: generation of bipartite graph, merging of shared peptides/proteins, partitioning of subgraphs, protein filtering, and calculation of protein probabilities. An example of a protein list output by AttnProt is shown in Table S3. Further details of AttnProt are described in Supporting Information.

■ DISCUSSION

In this work, we present AttnPep, a deep learning model for the postprocessing of proteomics mass spectrometry identification results, for improving the accuracy of peptide identification from tandem mass spectrometry. In total, six independent data sets from two data sets, ranging from the output of different database search software to the results of different acquisition methods, were evaluated in this study, which also contains the gold standard data set containing synthetic peptides. In all data sets, AttnPep showed better discrimination performance, was able to more clearly distinguish the distribution of correct and incorrect PSMs in the results of the database search software, and identified more peptides when compared to existing mainstream software.

Furthermore, the study reveals the capability of the Self-Attention module to focus the network's attention on input features that are conducive to the classification effect during the classification process, leading to improved classification outcomes. The Self-Attention module enables AttnPep to process the output of different search software without introducing excessive irrelevant features that might degrade the network's performance.

Despite AttnPep's success in distinguishing correct from incorrect PSMs and identifying more peptides than existing mainstream software, it was unable to fully identify all synthetic peptides from the gold standard identification data. Thus, we postulate that there exist significant features in the experimental data that may further enhance the identification rate of the correct peptides, and we aim to explore these features in future research.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00729.

> Protein probability calculation by bipartite graphs; *q*-value calculation; semilabeled calibration; software and hardware details; features extracted from MSFragger PIN file; database search parameters; outputted protein list; comparison of different methods in Thermo database and 6600 database, separately; performance of AttnPep in database Thermo_DDA_Ecoli, Thermo_DDA_Human, Thermo_DDA_QC, 6600_DDA_Ecoli, 6600_DDA_Yeast, and 6600_DDA_Human; heatmap of average weight matrix for MS-GF+ and Comet; the calibration of PSM-level p values; and performance of AttnPep with different input features (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jianwei Shuai − *Department of Physics, Xiamen University, Xiamen 361005, China; Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325001, China;* orcid.org/0000-0002-8712-0544; Email: shuaijw@wiucas.ac.cn

Authors

Yulin Li − *Department of Physics, Xiamen University, Xiamen 361005, China*

Qingzu He − *Department of Physics, Xiamen University, Xiamen 361005, China; Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325001, China*

Huan Guo − *Department of Physics, Xiamen University, Xiamen 361005, China*

Stella C. Shuai − *Biological Science, Northwestern University, Evanston, Illinois 60208, United States*

Jinyan Cheng − *Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325001, China*

Liyu Liu − *Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325001, China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jproteome.3c00729

Author Contributions

J.S., Y.L., and Q.H. conceived the project. Y.L. developed the algorithm, implemented the software. Y.L. and S.C.S. wrote the manuscript. Y.L., Q.H., J.C., and L.L. analyzed the data and results. Y.L., H.G., J.C., and L.L. analyzed data with software. Y.L., Q.H., H.G., and J.S. discussed the algorithms. Q.H., L.L., and J.S. supervised the project. Y.L. and Q.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The sponsors had no role in the design, execution, interpretation, or writing of the study. This work is supported by the STI2030-Major Projects (2021ZD0201900), the National Natural Science Foundation of China (Grant Nos. 12090052, 12174041, and 11874310).

■ REFERENCES

(1) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743−749.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.

(3) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *ELECTROPHORESIS: An International Journal* **1999**, *20* (18), 3551−3567.

(4) Eng, J. K.; Hoopmann, M. R.; Jahan, T. A.; Egertson, J. D.; Noble, W. S.; MacCoss, M. J. A Deeper Look into Comet—Implementation and Features. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1865−1874.

(5) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22−24.

(6) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5* (1), 5277.

(7) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry−Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513−520.

(8) Aebersold, R.; Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **2016**, *537* (7620), 347−355.

(9) White, F. M. The Potential Cost of High-Throughput Proteomics. *Sci. Signal.* **2011**, *4* (160), pe8.

(10) Nesvizhskii, A.; Proteogenomics, I. Concepts. *Applications and Computational Strategies. Nat. Methods* **2014**, *11* (11), 1114−1125.

(11) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.

(12) Zhou, W.-J.; Yang, H.; Zeng, W.-F.; Zhang, K.; Chi, H.; He, S.-M. PValid: Validation beyond the Target-Decoy Approach for Peptide Identification in Shotgun Proteomics. *J. Proteome Res.* **2019**, *18* (7), 2747−2758.

(13) Xiao-Dong, F.; Jie, M.; Cheng, C.; Ming-Ze, B.; Yun-Ping, Z.; Kun-Xian, S. The Application and Progress of Target-Decoy Database Search Strategy in Identification and Quality Control of Tandem Mass Spectrometry Data in Shotgun Proteomics. *Prog. Biochem. Biophys.* **2016**, *43* (7), 661−672.

(14) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.

(15) Choi, H.; Nesvizhskii, A. I. semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **2008**, *7* (01), 254−265.

(16) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923−925.

(17) Halloran, J. T.; Zhang, H.; Kara, K.; Renggli, C.; The, M.; Zhang, C.; Rocke, D. M.; Käll, L.; Noble, W. S. Speeding up Percolator. *J. Proteome Res.* **2019**, *18* (9), 3353−3359.

(18) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. Improvements to the Percolator Algorithm for Peptide Identification from Shotgun Proteomics Data Sets. *J. Proteome Res.* **2009**, *8* (7), 3737−3745.

(19) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1719−1727.

(20) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 1992; pp 144−152.

(21) Gao, M.; Yang, W.; Li, C.; Chang, Y.; Liu, Y.; He, Q.; Zhong, C.-Q.; Shuai, J.; Yu, R.; Han, J. Deep Representation Features from DreamDIAXMBD Improve the Analysis of Data-Independent Acquisition Proteomics. *Commun. Biol.* **2021**, *4* (1), 1190.

(22) Li, X.; Zhong, C.-Q.; Yin, Z.; Qi, H.; Xu, F.; He, Q.; Shuai, J. Data-Driven Modeling Identifies TIRAP-Independent MyD88 Activation Complex and Myddosome Assembly Strategy in LPS/TLR4 Signaling. *Int. J. Mol. Sci.* **2020**, *21* (9), 3061.

(23) Li, X.; Zhong, C.-Q.; Wu, R.; Xu, X.; Yang, Z.-H.; Cai, S.; Wu, X.; Chen, X.; Yin, Z.; He, Q. RIP1-Dependent Linear and Nonlinear Recruitments of Caspase-8 and RIP3 Respectively to Necrosome Specify Distinct Cell Death Outcomes. *Protein Cell* **2021**, *12* (11), 858−876.

(24) Zhong, C.-Q.; Wu, R.; Chen, X.; Wu, S.; Shuai, J.; Han, J. Systematic Assessment of the Effect of Internal Library in Targeted Analysis of SWATH-MS. *J. Proteome Res.* **2020**, *19* (1), 477−492.

(25) He, Q.; Zhong, C.-Q.; Li, X.; Guo, H.; Li, Y.; Gao, M.; Yu, R.; Liu, X.; Zhang, F.; Guo, D. Dear-DIAXMBD: Deep Autoencoder Enables Deconvolution of Data-Independent Acquisition Proteomics. *Research* **2023**, *6*, No. 0179.

(26) Li, Y.; He, Q.; Guo, H.; Zhong, C.-Q.; Li, X.; Li, Y.; Han, J.; Shuai, J. MSSort-DIAXMBD: A Deep Learning Classification Tool of the Peptide Precursors Quantified by OpenSWATH. *J. Proteomics* **2022**, *259*, No. 104542.

(27) Guo, H.; He, Q.; Li, Y.; Shuai, J. A Method for Analyzing DIA-NN Output Peptides Based on Squeeze-and-Excitation Neural Network. *Biophysics* **2023**, *11*, 17.

(28) Halloran, J. T.; Urban, G.; Rocke, D.; Baldi, P. Deep Semi-Supervised Learning Improves Universal Peptide Identification of Shotgun Proteomics Data. *bioRxiv*, 2020; pp 2011−2020.

(29) Granholm, V.; Kim, S.; Navarro, J. C. F.; Sjolund, E.; Smith, R. D.; Kall, L. Fast and Accurate Database Searches with MS-GF+ Percolator. *J. Proteome Res.* **2014**, *13* (2), 890−897.

(30) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176−3181.

(31) Xu, M.; Li, Z.; Li, L. Combining Percolator with X! Tandem for Accurate and Sensitive Peptide Identification. *J. Proteome Res.* **2013**, *12* (6), 3026−3033.

(32) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.*, **2017**, vol *30*.

(33) Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.

(34) Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *arXiv preprint arXiv:1803.02155*, 2018.

(35) Zha, H.; He, X.; Ding, C.; Simon, H.; Gu, M. Bipartite Graph Partitioning and Data Clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, 2001; pp 25−32.

(36) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L. OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nat. Biotechnol.* **2014**, *32* (3), 219−223.

(37) Van Puyvelde, B.; Daled, S.; Willems, S.; Gabriels, R.; Gonzalez de Peredo, A.; Chaoui, K.; Mouton-Barbosa, E.; Bouyssié, D.; Boonen, K.; Hughes, C. J. A Comprehensive LFQ Benchmark Dataset on Modern Day Acquisition Strategies in Proteomics. *Sci. Data* **2022**, *9* (1), 126.

(38) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646−4658.

(39) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency. *J. Proteome Res.* **2007**, *6* (9), 3549−3557.

(40) Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **2009**, *8* (8), 3872−3881.

(41) Shuken, S. R. An Introduction to Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **2023**, *22* (7), 2151−2171.

(42) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets [S]. *Molecular & Cellular Proteomics* **2015**, *14* (9), 2394−2404.

(43) The, M.; Tasnim, A.; Käll, L. How to Talk about Protein-level False Discovery Rates in Shotgun Proteomics. *Proteomics* **2016**, *16* (18), 2461−2469, DOI: 10.1002/pmic.201500431.

(44) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized Ppb-Range Mass Accuracies and

Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.

(45) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data. *J. Proteome Res.* **2010**, *9* (10), 5346−5357.

(46) Serang, O.; Moruz, L.; Hoopmann, M. R.; Kall, L. Recognizing Uncertainty Increases Robustness and Reproducibility of Mass Spectrometry-Based Protein Inferences. *J. Proteome Res.* **2012**, *11* (12), 5586−5591.

(47) Pfeuffer, J.; Sachsenberg, T.; Dijkstra, T. M. H.; Serang, O.; Reinert, K.; Kohlbacher, O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. *J. Proteome Res.* **2020**, *19* (3), 1060−1072.