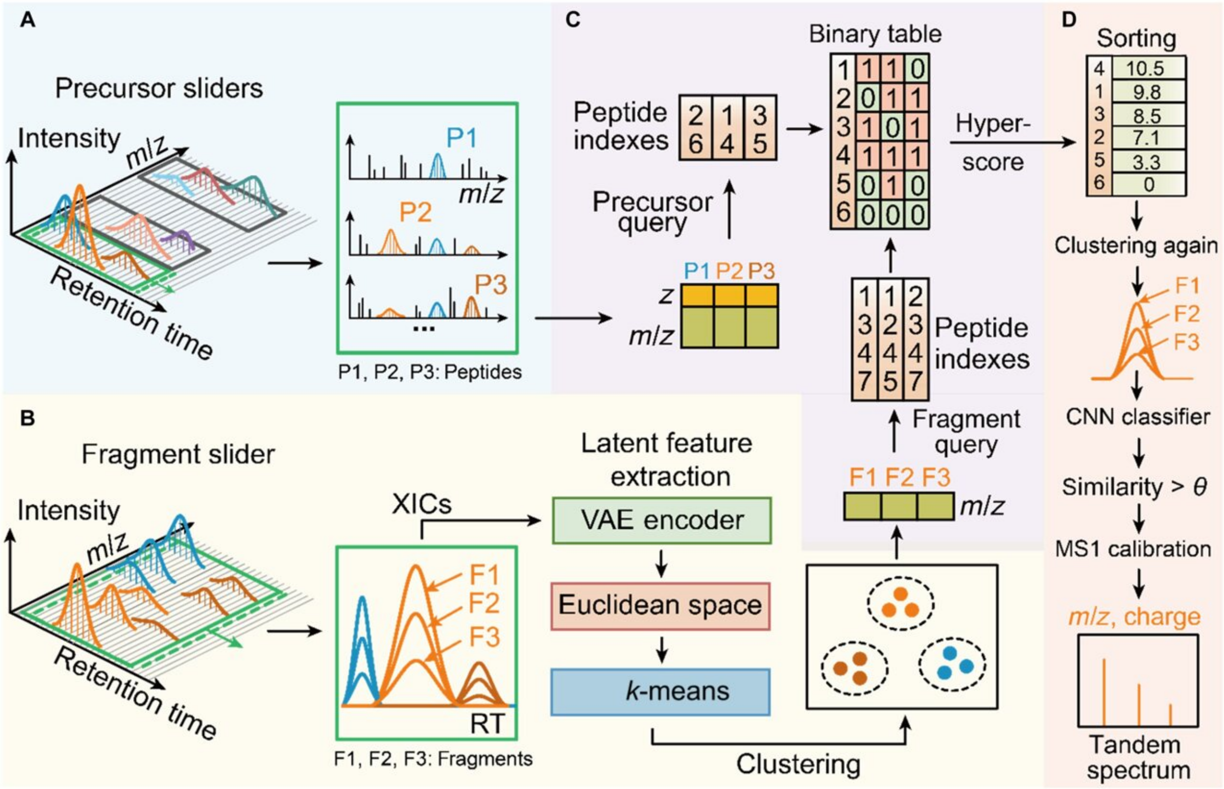
PHYS ORG

A deep variational autoencoder for proteomics mass spectrometry data analysis

November 3 2023



Schematic diagram of Dear-DIA. Credit: *Research*

Jianwei Shuai's team and Jiahuai Han's team at Xiamen University have developed a deep autoencoder-based data-independent acquisition data analysis software for protein mass spectrometry, which realizes the analysis of relevant peptides and proteins from complex protein mass spectrometry data, and demonstrates the superiority and versatility of the method on different instruments and species samples. The study was published in *Research* as "Dear-DIA^{XMBD}: deep autoencoder for data-independent acquisition proteomics".

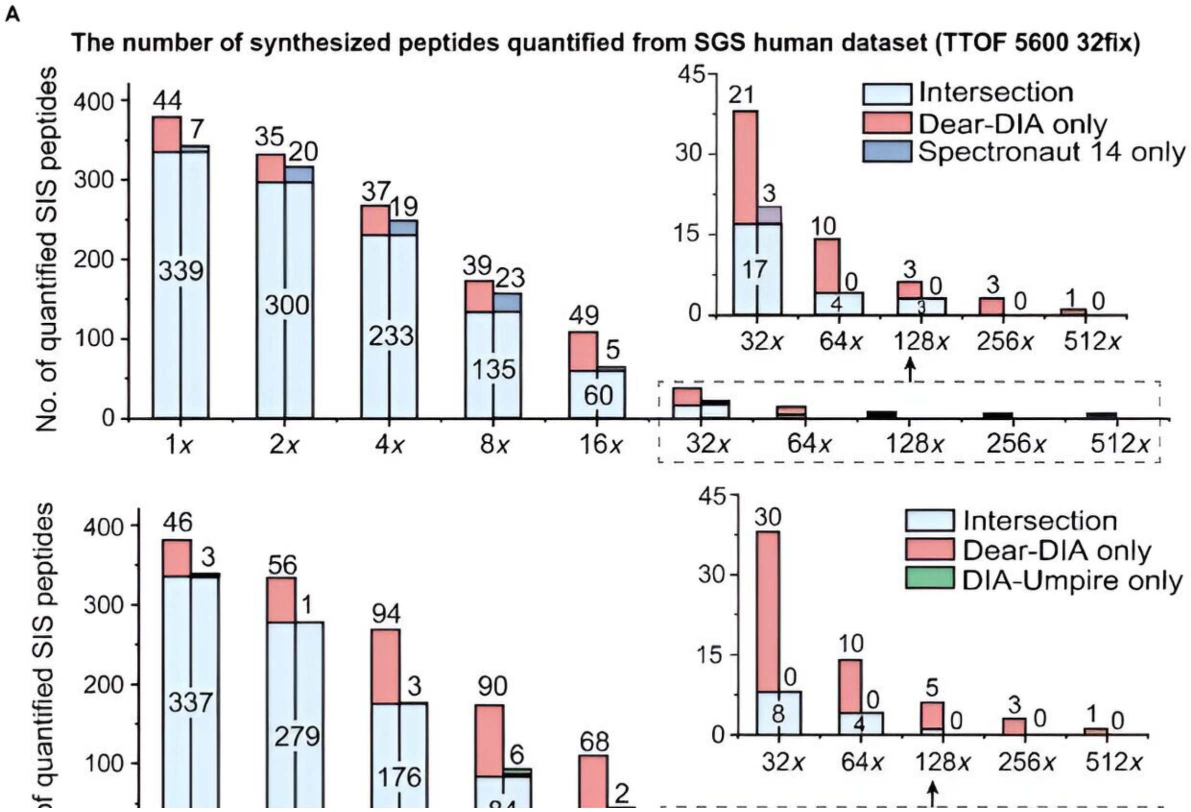
Proteins play a pivotal role as the executors of cellular life activities, driving a myriad of crucial biological processes. Consequently, the field of proteomics has received widespread attention. Proteomics involves the comprehensive study of protein properties, including post-translational modifications, protein expression levels, protein-protein interactions, and more. Its overall goal is to gain a holistic understanding of disease pathogenesis, cellular metabolism, and other vital processes at the protein level.

Among the key analytical techniques in proteomics research, protein mass spectrometry stands out as the most critical. Over time, mass spectrometry technology has evolved to provide researchers with reliable and dynamic tools for proteomics analysis.

Two main approaches to protein mass spectrometry are data-dependent acquisition (DDA) and data-independent acquisition (DIA). In DDA, all peptide precursor ion spectra (MS1) are acquired in full scan mode, followed by selection of the most N-intensive peptide ions for fragmentation to obtain fragment ion spectra (MS2).

Despite its utility, DDA faces challenges related to experimental reproducibility and detection of low-abundance peptides due to the randomness of peptide fragmentation and the preferential selection of high-intensity peptides.

To overcome these limitations, the DIA acquisition method has been introduced. This technique divides the mass-to-charge ratio range of parent ion spectra into multiple windows and sequentially fragments all peptides within each window to obtain daughter ion spectra. A common DIA method is Sequential Window Acquisition of all Theoretical fragment ions (SWATH).



Comparison of analysis results on human SGS dataset and mouse L929 cell dataset. Credit: *Research*

While DIA acquisition data retains more comprehensive proteomic information, its large data size, high dimensionality, and complex spectral signals pose challenges to its analysis. As a result, DIA data mining has become a major focus in the proteomics community.

Jianwei Shuai's team and Jiahuai Han's team collaborated to develop

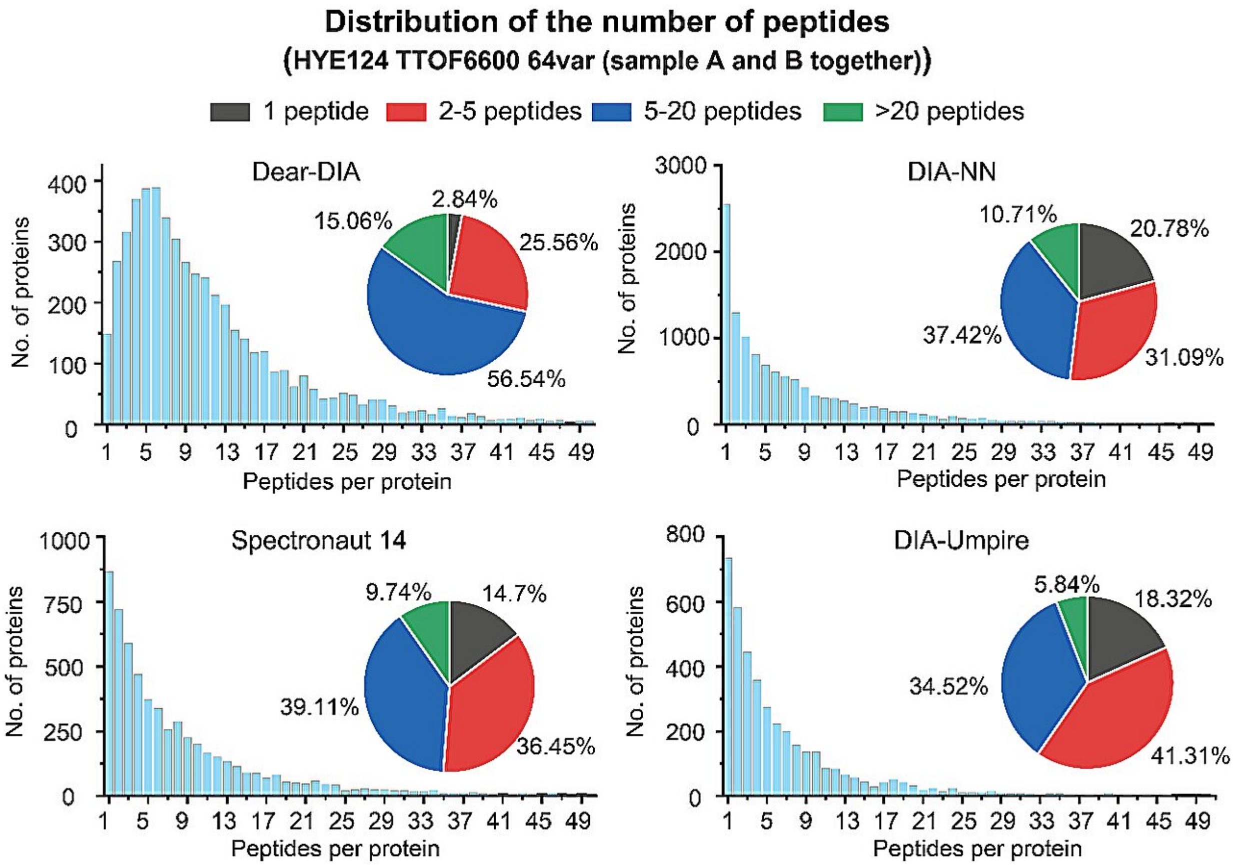
Dear-DIA, a deep learning-based data-independent acquisition data analysis software, which realizes the identification of fragment ions corresponding to different peptides from complex DIA acquisition spectra and demonstrates the generalization to complex samples from different species.

Dear-DIA first divides the spectra into a fixed-width slider with a fixed width along the retention time (RT) direction, and each slider contains a set of precursor spectra MS1 and fragment spectra MS2 as the minimum processing unit. Then, a peak-finding algorithm was used to remove the low signal-to-noise background ions and retain the candidate precursor ions and candidate fragment ions.

Next, Dear-DIA uses a variational autoencoder to extract the peak features of fragment ions and maps the features into Euclidean space, and then clusters the features, with different classes of fragments corresponding to different peptides, thus realizing the spectrogram deconvolution process.

Dear-DIA includes an indexing algorithm called PIndex, which matches the precursors to the fragments clustering results and selects the best pairing results by scoring. Dear-DIA uses a convolutional neural network to recalculate the peak shape similarity of fragments in the same class to eliminate interfering ions and clustering results with low similarity.

The authors first tested the performance of Dear-DIA on a SGS Human dataset containing 422 synthetic peptides of stable isotope-labeled standards divided into 10 dilution gradients (from 1-fold to 512-fold dilution), and DIA data were obtained on an AB SCIEX TTOF5600 mass spectrometer using the SWATH technique to obtain DIA data.



Comparison of peptide number distributions resulting from analysis of the HYE124 TTOF6600 64var dataset. Credit: *Research*

The analysis results showed that Dear-DIA found more synthetic peptides in all diluted solutions compared to the two commonly used analytical methods, Spectronaut 14 and DIA-Umpire. The authors also compared the number of peptides and proteins found by the different analytical methods for the SGS Human and L929 Mouse data sets. The results showed that Dear-DIA was able to find more peptides and proteins compared to Spectronaut 14 and DIA-Umpire, covering more than 85% of their results.

The confidence of proteomics analysis results can also be demonstrated by the number of peptides identified for each protein. Proteins with 2 or more identified peptides are generally considered to be more credible identifications. The authors compared the number of proteins versus peptides reported by Dear-DIA with existing software on a mixed species dataset (HYE124 TTOF6600 64var dataset).

The dataset contains proteins from three species, human, yeast and E. coli, and the data were acquired on an AB SCIEX TTOF6600 mass spectrometer using the SWATH method, with parent ion spectra containing 64 variable windows. The analysis results showed that 97.16% of the proteins found by Dear-DIA could correspond to 2 and more peptides, which is much higher than DIA-NN, Spectronaut 14 and DIA-Umpire.

Data-independent acquisition techniques for proteomics have been widely adopted, and related analysis algorithms have become a research hotspot. Protein discovery from massive mass spectrometry data is an interesting and challenging task. In this paper, the team developed Dear-DIA, an analysis software based on deep learning, which is used to process a variety of highly complex DIA acquisition data, and can discover more peptides and proteins, in addition to reproducing most of the results of Spectronaut and DIA-Umpire.

In addition, although the training dataset is from E. coli, the excellent performance of Dear-DIA on the mixed species dataset demonstrates its strong generalization ability to analyze complex proteomics data. Deep learning, as a widely used tool for big data analysis, has demonstrated excellent data mining capabilities to discover deep intrinsic associations in big data.

The use of deep learning to analyze proteomics mass spectrometry data has great potential and will further promote the study of fundamental issues such as protein signaling networks.

More information: Qingzu He et al, Dear-DIA^{XMBD} : Deep Autoencoder Enables Deconvolution of Data-Independent Acquisition Proteomics, *Research* (2023). DOI: 10.34133/research.0179

Provided by Research

Citation: A deep variational autoencoder for proteomics mass spectrometry data analysis (2023, November 3) retrieved 27 April 2024 from https://phys.org/news/2023-11-deep-variational-autoencoder-proteomics-mass.html