Received: 5 May 2022 | Revised: 22 November 2022 | Accepted: 21 December 2022

DOI: 10.1002/pmic.202200179

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Improving SWATH-MS analysis by deep-learning

Bo Sun[1] | **Pawel Smialowski**[2,3] | **Wasim Aftab**[1] | **Andreas Schmidt**[1] | **Ignasi Forne**[1] | **Tobias Straub**[3] | **Axel Imhof**[1]

[1]Faculty of Medicine, Biomedical Center, Protein Analysis Unit, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[2]Institute of Stem Cell Research, Helmholtz Center Munich, German Research Center for Environmental Health, Germany

[3]Faculty of Medicine, Biomedical Center, Computational Biology Unit, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

Correspondence
Biomedical Center, Protein Analysis Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, Großhaderner Strasse 9, 82152 Planegg-Martinsried, Germany.
Email: imhof@lmu.de

Funding information
Chinese Scholarship Council, Grant/Award Number: 201506230154; Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 2133249687, 219249687, 325871075; German Federal Ministry of Education and Research, Grant/Award Number: BMBF FKZ161L0214F

Abstract

Data-independent acquisition (DIA) of tandem mass spectrometry spectra has emerged as a promising technology to improve coverage and quantification of proteins in complex mixtures. The success of DIA experiments is dependent on the quality of spectral libraries used for data base searching. Frequently, these libraries need to be generated by labor and time intensive data dependent acquisition (DDA) experiments. Recently, several algorithms have been published that allow the generation of theoretical libraries by an efficient prediction of retention time and intensity of the fragment ions. Sequential windowed acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) is a DIA method that can be applied at an unprecedented speed, but the fragmentation spectra suffer from a lower quality than data acquired on Orbitrap instruments. To reliably generate theoretical libraries that can be used in SWATH experiments, we developed deep-learning for SWATH analysis (dpSWATH), to improve the sensitivity and specificity of data generated by Q-TOF mass spectrometers. The theoretical library built by dpSWATH allowed us to increase the identification rate of proteins compared to traditional or library-free methods. Based on our analysis we conclude that dpSWATH is a superior prediction framework for SWATH-MS measurements than other algorithms based on Orbitrap data.

KEYWORDS
proteomics, deep learning, spectral Library, data independent acquisition

1 | INTRODUCTION

The analysis of the proteomic composition of biological samples promises to provide a rich source of information, which could greatly improve our molecular understanding of a wide range of biological processes. It has the potential to revolutionize molecular diagnostics and treatment of disease. Despite a substantial improvement of the instruments (mostly mass spectrometers) used to perform proteomic measurements, the field still suffers from a substantial undersampling of peptides in shot gun proteomics studies (also called data dependent acquisition or DDA) and therefore a very low coverage of all possible peptides. To overcome this problem data independent acquisition (DIA) strategies have been developed that result in the fragmentation of all possible ions, which should (at least in theory) substantially improve peptide coverage. To achieve this task, extremely fast tandem mass spectrometers (such a quadrupol time of flight or Q-TOF instruments) need to be used, which results in a decrease of fragment spectrum

Abbreviations: BiLSTM, Bidirectional long-short term memory; CNN, Convolutional neural network; DDA, data-dependent acquisition; DIA, data-independent acquisition; dpMC, deep learning for missed cleavage; dpMS, deep learning for MS fragment ion prediction; dpRT, deep learning for retention time prediction; dpSWATH, deep learning for SWATH analysis; FDR, false discovery rate; LC-MS, Liquid chromatography coupled mass spectrometry; PCC, Pearson correlation coefficient; PSM, Peptide spectral matches; Q-TOF, Quadrupol Time of Flight; RNN, Recurrent neural network; RPKM, reads per kilobase per million mapped reads; SWATH-MS, sequential window acquisition of all theoretical mass spectra.

quality. One of these methods is the so called sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) using a quadrupole-TOF instrument [1]. In SWATH-MS mode, typically a precursor ion (MS1) spectrum is recorded, followed by a series of fragment ion (MS2) spectra recordings with wide precursor isolation windows (for example 25 m/z). A comprehensive data set is recorded through repeated cycling of consecutive precursor isolation windows over a defined mass range, which includes continuous information on all detectable fragment and precursor ions [1]. SWATH-MS has been implemented in many aspects of research, which including quantitative proteomics [2], clinical biomarker research [3], histone post-translational modification (PTM) analysis [4] and the analysis of protein-protein interactomes [5].

In addition to a better peptide coverage SWATH-MS also has advantages in reproducibility [6] and speed of analysis [7] and allows a retrospective targeting [1], which is not possible when using targeted workflows.

A disadvantage of all DIA methods is the requirement of specific fragment ion libraries for identification. Currently most of these libraries are experimentally generated using DDA measurements of a highly fractionated sample pool measured prior to SWATH-MS acquisition on the same instrument [8]. A lot of efforts have been put into building the assay library to improve the coverage and quality of proteomic research [9]. In 2016, J. Wu et al. have compared the SWATH mass spectrometry performance using local seed libraries integrated with external assay libraries and local assay libraries alone [10] and showed that the first one had a better performance with regard to peptide identification and quantification. In addition, software tools like SpectraST [11] have been developed to improve the building of consensus mass spectrum libraries [12].

Nowadays, deep-learning methods have empowered proteomic research. Especially the predictions based on the information inferred from peptide sequence have gained a lot of attention, such as the prediction of retention time [13] and fragment ion intensities [14, 15]. In addition to the prediction of peptide properties, deep-learning is also used for the identification of peptides and proteins. For example, the detection of LC-MS features is performed by deep-learning models [16]. Moreover, the deep-learning approach can also be used for de novo sequencing, such as the work that has been done by DeepNovo [17].

More recently, tools have been developed that allow an extension of the used libraries by applying both experimental [12, 18] and theoretical approaches [14, 15, 19]. However, most of the theoretical approaches used mass spectra recorded in an orbitrap instrument, which are of higher quality than the ones measured in a Q-TOF mass spectrometer. To improve the SWATH-MS analysis, we developed a novel framework and strategy to build high-quality in silico libraries by deep-learning.

2 | MATERIALS AND METHODS

2.1 | Datasets

2.1.1 | Datasets used for training and testing of dpSWATH

We used datasets generated by TripleTOF 5600 and 6600 (ABSciex, Concord, Ontario, Canada) from *Homo sapiens* and *Drosophila melanogaster*, respectively. We used the DDA datasets from the Pan-Human project (PXD000953) [12] as a pre-training datasets for the TripleTOF 5600 measurements. All peptide spectra matches (PSMs) of the Pan-Human project were extracted from file *PHL.pep.xml* and split into training and testing datasets. For the training dataset, we selected 2,000,029 unmodified PSMs containing 94,878 unique unmodified peptides. To test the model, we used 499,999 unmodified PSMs with 23,436 unique peptides. No DIA datasets were used for testing on Pan-Human project.

For further training and testing we used a DDA dataset of Hela extracts (PXD009273) [20]. To retrain the retention time and mass spectral models to build the in silico library, 12 DIA datasets were used for DIA searching followed by identification and quantification of the proteins.

For TripleTOF 6600, an aliquot corresponding to 500 μg of proteins *of a Drosophila* embryo extract [21] was precipitated with TCA. The protein pellet was dissolved in 6 M urea for subsequent protein cleavage by LysC and trypsin, disulfide reduction and alkylation with DTT and iodoacetamide, respectively. The obtained polypeptide mixture was desalted over C18 stage tips before further high pH-reversed phase separation. Individual fractions were injected onto an Exigent 425 nanoLC system, operated in micro-flow mode at 5 μl/min and separated on a 300 μm x 15 cm column directly coupled to the TripleTOF 6600 mass spectrometer (both ABSciex). For peptide separation a 50 min gradient from 2% to 35% acetonitrile in water was employed followed by 5 min washing at 80% acetonitrile. Peptides eluting from the column were detected in information-dependent detection mode acquiring a survey scan from 350 to 1500 m/z. Maximally 20 precursors with charge state 2+ or higher and a signal intensity of min. 160 counts were selected for MS/MS analysis to obtain high quality data for peptide identification. To further increase the number of detected peptides and proteins, DDA experiments of 72 fractions of a Drosophila embryo extract fractionated by size exclusion chromatography (Superose 6 10/300 GE Healthcare, Chicago, IL). All the PSM information was extracted using ProteinPilot (ABSciex, Concord, Ontario, Canada) or SpectroMine (Biognosys AG, Schlieren, Switzerland) and deposited on the Pride database (PXD038407). To evaluate the performance of dpSWATH, the precursors of 72 fractionated DDA runs including peptide sequences and precursor charges were extracted from

experimental library based on the searching results of Pulsar in Spectronaut (15.2.210819, Biognosys AG, Schlieren, Switzerland). For the 72 fractionated library, 3655 unique peptides were extracted to transfer-train the retention time and mass spectral models, the left 40,000 peptides with corresponding precursor charges were used to build the validation library with prediction by dpSWATH.

For all of the training, testing and validation datasets, the fragment ions were normalized which divided by the highest peak for each mass spectrum. The minimum and maximum length of peptides is 7 and 60 respectively and, the precursor charge ranges from 1 to 6 and a maximum charge of fragment ions of 2+.

2.1.2 | Datasets used for building the theoretical library

Fasta files of *D. melanogaster and H. sapiens* were downloaded from FlyBase (http://flybase.org/) and UniProt (https://www.uniprot.org/) respectively. Then the protein sequences were selected based on the entries recorded in the DDA libraries. For *D. melanogaster*, 5006 protein groups were extracted while 10524 and 4460 protein groups were extracted from the in Pan-Human library or the DDA experiment prepared from HeLa extracts respectively. The peptide sequences were prepared based on the cleavage standard rules of trypsin [22]. Up to two missed cleavages and cleavages followed by proline were predicted by dpMC. The length of the peptides is from 7 to 60 and the charges for peptides range from 2+ to 4+. Data of mRNA expression profiles for different stages of embryos of *D. melanogaster* (gene_rpkm_report_fb_ 2017_05.tsv) were downloaded from Flybase, while mRNA expression data for HeLa cell-lines were used from the ProteomeXchange repository (PXD009273) [20].

2.2 | Preprocessing of datasets for modeling (dpMScore)

All datasets were preprocessed using the newly developed dpMScore and used for both training and testing of the performance of dpSWATH (Figure 1). dpMScore uses hierarchical clustering to choose the most abundant and consistent fragmentation of each peptide. The dpMScore is calculated by the following formula:

$$\text{dpMScore} = -ln(Dist)\sum_{1}^{Nc}\prod_{i}^{Nc} pi \exp^{\prod_{1}^{Nc} pi(1-\prod_{1}^{Nc} pi)}$$

where Dist is the distance among different fragmentations for the same peptide, which ranges from 0.01 to 0.2 based on Pearson Correlation Coefficient (PCC); *Nc* is the number of clusters split at one certain bar; p_i is the proportion of i_{th} cluster calculated by the number of fragmentations in this cluster divided by the total number of fragmentations for this peptide, which ranges from 1 to *Nc*.

The dpMScore was only calculated for peptides that had more than three replicates whereas peptides with less than three replicates were kept in the training or testing datasets for dpSWATH without a score attached to it.

2.3 | Retention time prediction

As the prediction of retention time is crucial to SWATH-MS analysis, we developed dpRT as part of the dpSWATH program for a highly accurate retention time prediction and an increased sensitivity and identification of peptides and proteins (Figure 1, Figure S1). The framework of dpRT takes advantage of both convolutional neural network (CNN) and recurrent neural network (RNN) with self-attention mechanism (Figures 1 and S1). CNN performs very well on image and lingual work which benefits from its powerful feature extraction function. In dpRT, we use one-dimensional CNN as feature extractor to analyze the peptide sequence by setting the kernel size as 3. It is beneficial for next level RNN to use these features to predict the fragment ions' intensities. As for the RNN work, we chose two parallel bidirectional Long-Short Term Memory (BiLSTM) layers. The BiLSTM is very good at dealing with sequence or sentence cases, which has the advantage of processing information in both directions; for each predicted vector, BiLSTM makes the prediction combining the past and future states simultaneously. However, BiLSTM still shows lack of capability of dealing with long sequences, which could be complemented by the advantage of self-attention algorithm which is able to assign different weights to different features and has strong capability to deal with long sequences. Besides the BiLSTM layers, we also adopted self-attention layers to enhance the capability of model on dealing with the distant information along the sequences. Then two dense layers with 256 units and 1 unit respectively were connected to above RNN layers to generate the single predicted value.

2.4 | Fragment ion prediction

For the prediction of fragment ions, we developed dpMS. In dpMS, we also used one-dimensional CNN as feature extractor to analyze the peptide sequence by setting the kernel size as 2, in this way CNN could extract features from each two adjacent amino acids which have a strong and direct effect to the fragment ions that lies between them, which is beneficial for next level RNN to use these features to predict the fragment ions' intensities. As for the RNN work, we keep the similar architecture as dpRT but modify the units of RNN and self-attention layer with width as 49. For the fragment ions used to construct mass spectra, we take b ions and y ions that are generated by one time-distributed dense layer as the output layer of dpMS and the dimension of output is 59*4.

2.5 | DDA library generation

The search engine Pulsar in Spectronaut (15.2.210819, Biognosys AG, Schlieren, Switzerland) was used to build all the above the DDA

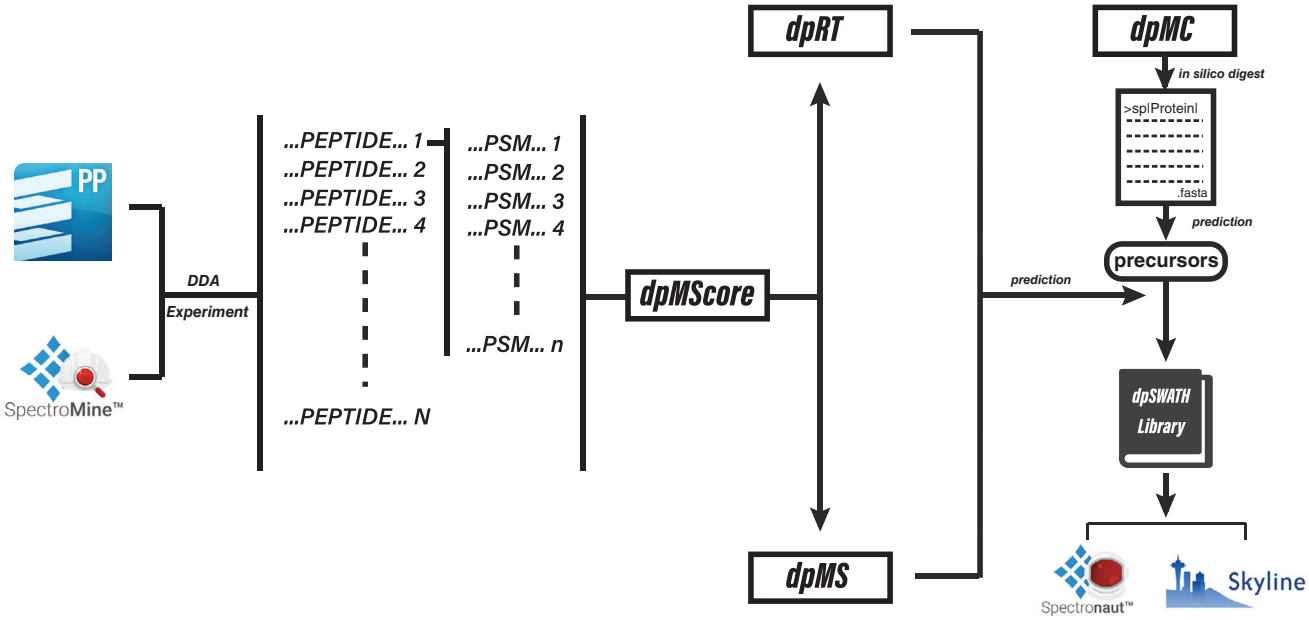

FIGURE 1 The workflow of dpSWATH and strategies applied in this study. Datasets from either ProteinPilot or SpectroMine can be analyzed, and the generated library can be used by Spectronaut or Skyline

libraries. The public Pan-Human library is pre-deposited in the Spectronaut software. To measure the performance of dpSWATH, we built the experimental library of unmodified peptides with length from 7 to 60 amino acids, precursor charges from 1+ to 6+ and, set Cysteine carbamidomethylation as fixed structural modification and no variable modification are selected. The maximum missed cleavage was set as 2.

2.6 | Construction of dpSWATH library

After the prediction of fragment ions' intensities and retention time, we assembled the two parts' results into .txt file which could be read by Spectronaut. The .txt file stores all available mass spectra to build the searching library. We put all of the 10 necessary information (Supplementary Note 1) of mass spectra including the predicted fragment ions and retention time which suggested by Spectronaut into the .txt file (Figure 1). Besides, we also prepared the script for building the library for Skyline.

3 | RESULTS

3.1 | Preprocessing of the datasets

Compared to an orbitrap mass spectrometer, the fragment spectra analyzed within a TripleTOF mass analyzer show a higher variability [11, 23]. The selection of the representative mass spectrum is therefore crucial for efficient identification and quantification of the corresponding peptide. Currently, most spectral libraries were built with the consensus mass spectra from PSMs using clustering algorithms such as SpectraST [11]. The selection of the consensus spectrum is often based on selecting the spectra with a minimal Q-value. However, for TripleTOF datasets, even PSMs with similar Q-values show big differences in the intensities of individual fragment ions as shown in Figure 2. For the training of dpSWATH, we therefore devised dpMScore to preprocess MS datasets and select the most abundant and consistent mass spectra for a given peptide with the same precursor charge. In dpMScore, we take the similarities among mass spectra into consideration and choose the cluster with the smallest distance and the largest number of spectra (see Section 2). In this way, the clusters of mass spectra were not only determined by the intensities, but also by the number of detected fragment ions and their proportions.

3.2 | Benchmarking of dpSWATH

Tandem MS spectra are strongly affected by many different experimental conditions ranging from sample preparation to instrumental set up to ambient environmental conditions such as temperature, humidity, or electrical interference (Gallien et al., 2013). We thereby designed dpSWATH in such a way that it can be trained and tested using data measured on multiple different instruments and under variable conditions and used transfer learning to construct reliable libraries.

To prepare a high-quality predicted library, the algorithm should therefore be able to efficiently extract associated features, which affect the mass spectrum pattern and retention time. To do this, we put the convolutional layer as the first layer to extract the features at a deep level automatically. To address the issue of identifying very long peptides (e.g., longer than 40 amino acids), we also used a self-attention layer to deal with longer sequence peptides (Figure S1).

First, we split the Pan-Human datasets from TripleTOF 5600 into training, validation, and testing datasets into a ratio of 8:1:1. By

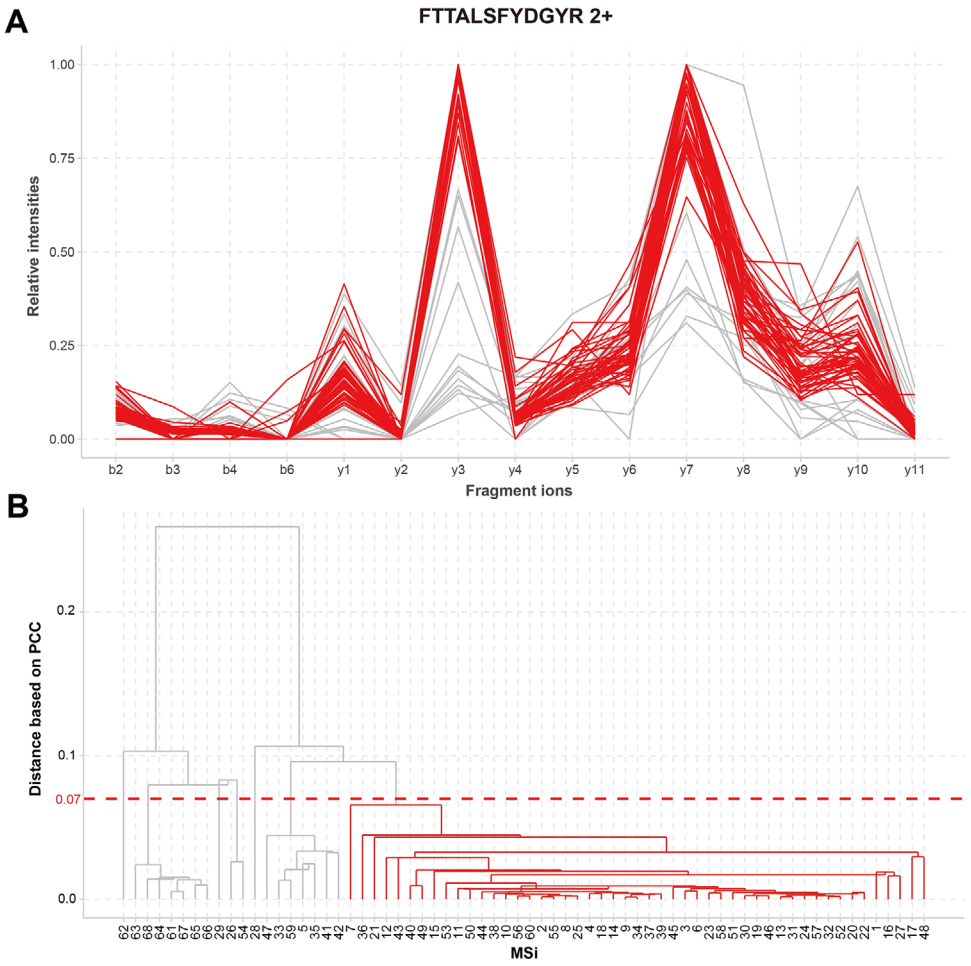

FIGURE 2 Different mass spectra patterns for the same precursor on the same condition. (A) The line plot of mass spectra of peptide "FTTALSFYDGYR" with precursor charge 2, the fragment ions' names are shown in x-axis with relative intensities in y-axis. All mass spectra pattern for this peptide are shown, the red patterns are the cluster chosen based on the dpMScore, the left gray pattern are filtered out by dpMScore for this peptide; (B) the clustering diagram for this peptide, the index of different mass spectra for this peptide are shown in x-axis with the distance among different mass spectra shown in y-axis. The chosen cluster are shown red corresponding to the red mass spectra in (A), which are chosen on the threshold at distance based on PCC 0.07

applying dpMScore as described in the methods, dpMS has achieved a median Pearson Correlation Coefficient (PCC) of 0.968 and median dot-product of 0.973 between the observed and predicted mass spectra (Figure 3A). For all validation and testing datasets, the peptides were not shown in the training datasets. We then applied transfer-learning on human datasets of TripleTOF 6600 with the trained model on TripleTOF 5600 to predict the fragmentation spectra of 57157 peptides from *D. melanogaster*. When doing this, we achieved a median Pearson Correlation Coefficient (PCC) of 0.980 and median dot-product 0.983 between observed and predicted mass spectra (Figure 3A,B). The similarities between observed and predicted mass spectra can directly affect the success of the identification and quantitation of proteins and peptides in the downstream analysis. Compared with DeepDIA on the same datasets, dpMS achieved much higher accuracy, which benefits the following analysis. Besides the higher accuracy given by dpMS, the capability of prediction for the longest sequence has been up to 60 and up to 6 of the highest precursor charges.

Then we applied the same strategy to the prediction of retention time. To eliminate the differences among different experiments and facilitate the prediction, we applied indexed retention time (iRT) throughout this study. The information of retention time can provide a reliable coordinate for mapping corresponding peptides [24, 25] and is usually combined with other analytical coordinates (m/z, intensity) for a reliable identification and quantification [25]. Therefore, we developed dpRT as part of the dpSWATH framework to facilitate the generation of building reliable in silico libraries (Figures 3C and S2).

Based on the high accurate prediction on mass spectra and retention time, we benchmarked the performance of dpSWATH by integrating the results from dpMS and dpRT on the validation datasets (see Section 2), which contains 40,000 peptides in the library. From the results, we got more peptides and proteins compared to the experimental 72 fractionated library and the library built by DeepDIA (Figure 3E,G). Compared to the experimental identified 28,940 peptides and 3301 protein groups, dpSWATH identified 31,012 peptides and 3545 protein groups, which are also more than the results from DeepDIA which

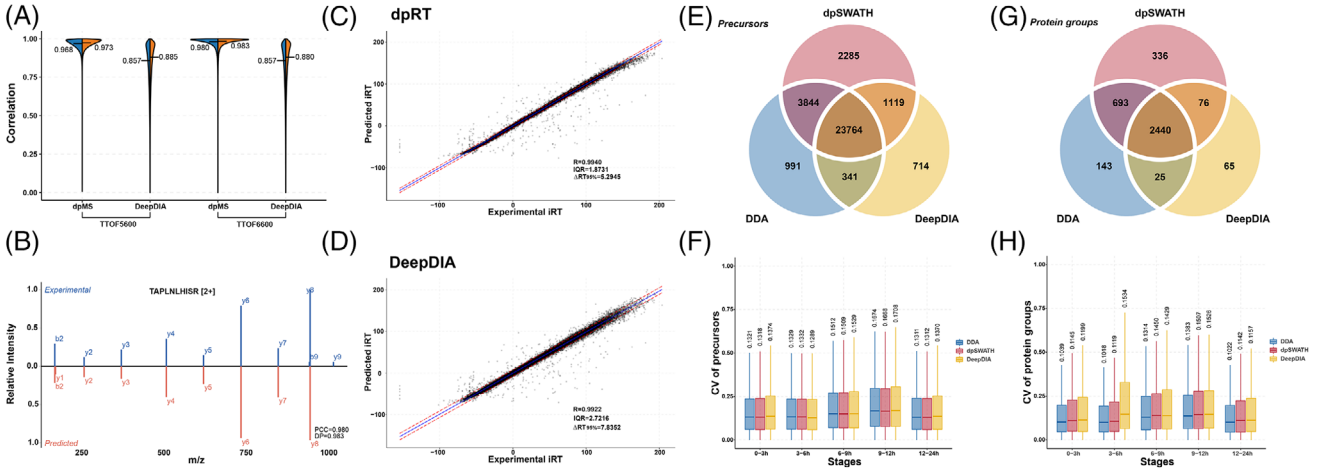

FIGURE 3 Benchmarking of dpSWATH. (A) The performance of dpMS and DeepDIA on datasets from TripleTOF 5600 and TripleTOF 6600, the blue histograms show the distribution of PCC, while the orange histograms show the distribution of dot-product, the median PCC and dot-product are shown. (B) The mirror plot for peptide 'TAPLNLHISR' with precursor charge of 2. The experimental mass spectra is shown in upper blue while the predicted shown in lower red. (C) The prediction of retention time by dpRT on datasets of *D. melanogaster*, the correlation of PCC, interquantile range (IQR) and distance of 95% datapoints are shown; (D) the prediction of retention time by DeepDIA on the same datasets as (C); (E) The overlapping of precursors among libraries of DDA, dpSWATH and DeepDIA. (F) The coefficient of variance (CV) of precursors for each stage of the embryo development in *D. melanogaster*. (G) The overlapping of protein groups among libraries of DDA, dpSWATH and DeepDIA. (H) The coefficient of variance (CV) of protein groups for each stage of the embryo development in *D. melanogaster*

identified 25,938 peptides and 2606 proteins. The libraries built by experimental (DDA) or theoretical approaches (dpSWATH, DeepDIA) are based on very different strategies. The DDA library was built on the identified PSMs of given precursors, which was based on the consensus mass spectra generation algorithm like SpectraST. For the library built by dpSWATH, the training process was based on the filtered PSMs, and then the mass spectra pattern and retention time were predicted by dpMS and dpRT, respectively. The library built by DeepDIA, only the PSMs with minimum Q-values were used for training which leads to relatively higher specificity but lower sensitivity.

To estimate the applicability of theoretical libraries, we also measured the coefficient of variance when quantifying protein groups from two technical replicates of five different developmental stages of *Drosophila* embryos (Figure 3F,H). In each case the CV is very similar between analyses made using the dpSWATH predicted library and derived from a DDA experiment (Figure S3A–S3F). Even when comparing the quantification of individual precursor ions both DDA and dpSWATH libraries performed equally well (Figure S3G). From this comparative analysis we conclude that dpSWATH not only identifies more peptides and protein groups, but also provides a robust and reproducible quantitation similar to the DDA approach but on this higher number of identified peptides.

3.3 | The interpretation of mass spectra on amino acids level

To understand the inner mechanism of our algorithm we investigated the amino acid contributions and therefore analyzed the impact of different amino acids on the prediction of the pattern of mass spectra.

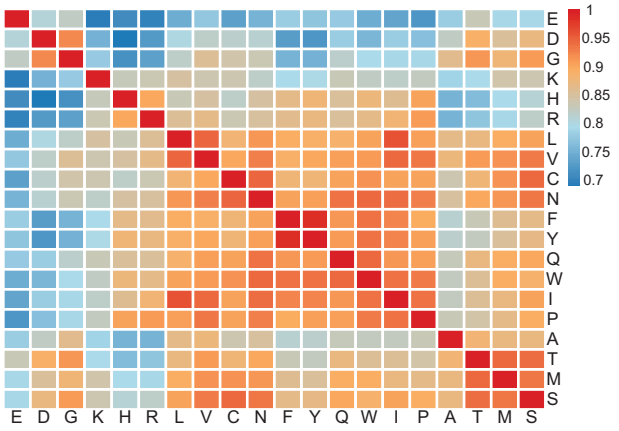

FIGURE 4 The heatmap of the correlation among amino acids based on their features

In the process of predicting mass spectra by dpMS, the properties of amino acids are encoded into each neuron, which is given different weights depending on the peptide sequence. The heatmap illustrates the weights of each amino acid assigned during predictions. We could see that some amino acids such as the aromatic amino acids F and Y cluster together due to their biochemical properties and structures (Figure 4).

3.4 | Missed cleavage prediction by dpMC

In proteomic analysis, trypsin is widely used to digest proteins into peptides. Despite being a robust and efficient protease

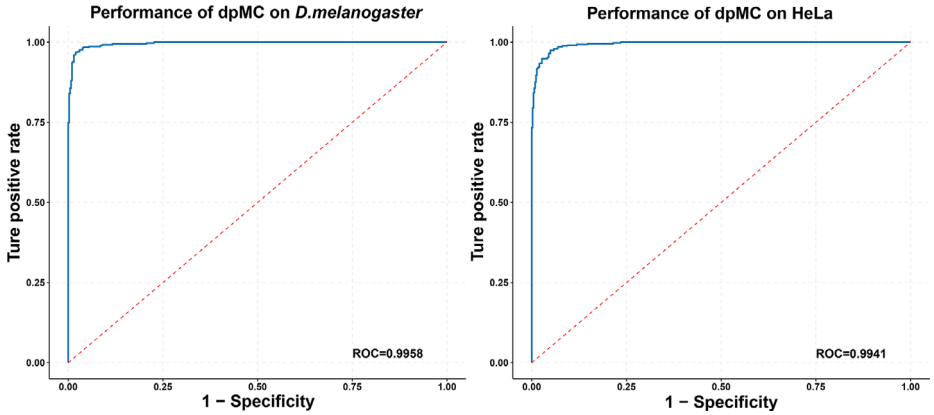

FIGURE 5 The performance of dpMC on datasets of *D. melanogaster* and HeLa

tryptic cleavage rarely reaches a100% efficiency. To predict the sites of inefficient cleavage most search engines use the Keil rules [22]. When it comes to building a large library based on entire proteomes, one problem is how to accurately predict missed cleavages. DeepDIA simply adopts the Keil rules to fully predict missed cleavages. To improve this prediction, we developed dpMC [26] (Figure 5). For the application of dpMC in dpSWATH, we also optimized parameters for the combinations of trypsin and LysC.

Besides, since the training of dpMC is based on the detected peptides in experiments, so the cleaved peptides also have the information of detectability, which is mentioned by AP3 [27]. Thus, the candidate peptides are most detectable for DIA analysis. In this way, we not only reduce the search time while maintaining a high specificity, but also improve the recovery rate and control FDR of theoretical libraries effectively.

3.5 | SWATH-MS analysis improvement using theoretical libraries generated by dpSWATH

Combined with the predictions of peptide fragment spectra by dpMS, retention time by dpRT and accurate missed cleavage sites by dpMC, we built an in silico library of all proteins detected in DDA experiments. To generate the theoretical library, we prepared the precursor candidates for each protein group detected in DDA library. This resulted in a library based on 5006 protein groups in *D. melanogaster*. We prepared the library using either the same peptide entries as observed in the DDA library, or the library predicted from the protein groups identified up to two missed cleaved sites by Keil rules or up to two missed cleaved sites predicted by dpMC. Then we searched data from the corresponding SWATH runs using these libraries with the same settings in Spectronaut. A comparison showed that we got the most identifications with the library built by dpSWATH with the predicted missed cleavages by dpMC. For the DDA based theoretical library, the protein groups' recovery rate of the library from 66.36% (3560/5006) to 95.65% (4788/5006) of the DDA library. We also performed the searching with directDIA 2.0 developed by Spectronaut, which showed only half of the identifications compared to the library built by dpSWATH (Figure 6).

Next, we built an in silico library referring to the peptide entries in Pan-Human library [12] and a DDA library of HeLa extracts [20] from TripleTOF 5600. We built the libraries with similar strategies except the library with up to two missed cleaved sites by Keil rules. Similar to our *Drosophila* data set, we got the most identifications of protein groups when we searched publicly available SWATH runs of HeLa extracts using a library built by dpSWATH with the predicted missed cleavages by dpMC (Figures 6C, S4C,E). Also in this case, the recovery rate of the Pan-Human library increased substantially from 32.40% (3410/10524) to 69.09% (7271/10524). The recovery rate increases when we use a library based on the entries from a DDA experiment performed on HeLa extracts, which is due to the same source. Even in this case the library built by dpSWATH performs better than the library built from experimental data only (86.32% (3850/4460) to 96.39% (4299/4460)) (Figure S4E). We also performed searches using directDIA 2.0 in Spectronaut, which resulted in far less identifications than by dpSWATH (Figure 6).

The prior DDA analysis to define the proteomic space used for the generation of a theoretical library was essential to keep a low FDR of the DIA search. In fact, when the library is generated from the entire proteome many DIA searches result in a low rate of peptide identifications and quantifications, which is often due to a high FDR. To limit the search space without the need of a prior extensive DDA measurement we built the in silico libraries based on transcriptomic data from the corresponding source. To do this, theoretical fragment spectra were generated from all protein candidates where the corresponding gene had an average number of reads per kilobase per million mapped reads (RPKM) [28] greater than or equal to 1. For the different developmental stages of *D. melanogaster*, this resulted in the inclusion of 17299 proteins. Compared with the DDA library, this strategy resulted in a much higher identification of protein groups (6156/3322), and peptides (64782/31538). The same effect is also observed when using the transcriptomic data from HeLa cells where we predicted the fragment spectra of peptides derived from 8758 proteins (Figure S4).

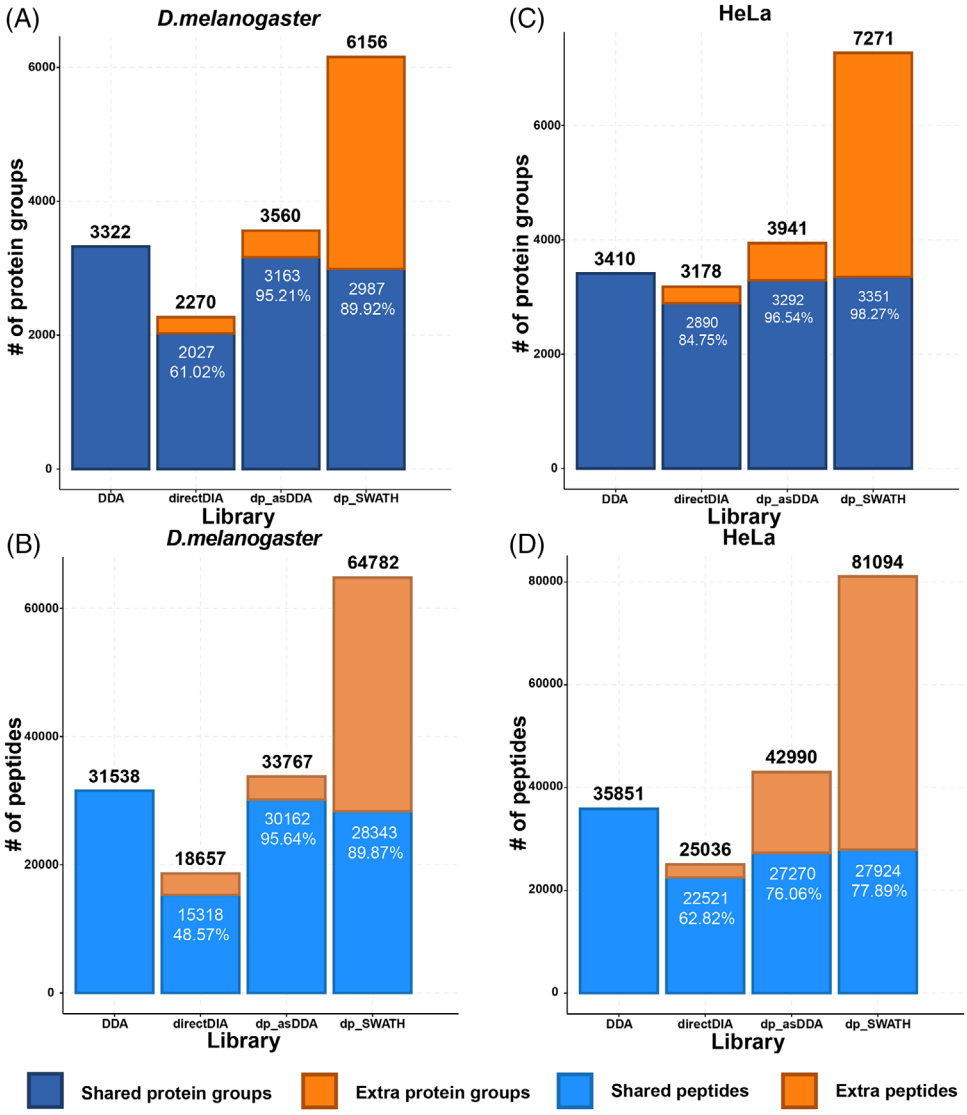

FIGURE 6 The identifications of protein groups and peptides from different libraries. (A) The number of identified protein groups on *D.melanogaster*. "DDA" indicates results from 72 fractionated DDA library; "directDIA" indicates the number of identified protein groups on *D. melanogaster* by directDIA 2.0; "dp_asDDA" indicates results from the in silico library on the same entries as DDA library; "dp_SWATH " indicates results from the in silico library on the digested FASTA sequence of the transcriptome based library by dpMC with up to two missed cleavages combined with no missed cleavages. (B) The number of identified peptides on *D. melanogaster*. (C) The number of identified protein groups on HeLa datasets from TripleTOF 5600 refer to the PanHuman library; "DDA" indicates results from experimental Pan-Human library; "directDIA" indicates the number of identified protein groups on HeLa datasets by directDIA 2.0; "dp_asDDA" indicates results from the in silico library on the same entries as experimental Pan-Human library; "dp_SWATH" indicates results from the in silico library on the digested FASTA sequence of the transcriptome based library by dpMC with up to two missed cleavages combined with no missed cleavages. (D) The number of identified peptides on HeLa datasets from TripleTOF 5600 compared to the PanHuman library. Identifications overlapped with the DDA-based libraries are denoted as "shared." Novel identifications by in silico libraries are denoted as "extra." The numbers and sensitivities of protein groups or peptides are shown

A detailed analysis of the correlation between the predicted mass spectra and the measured ones revealed the strong benefit of using dpMScore, which turned out to be crucial for building high quality libraries on Q-TOF datasets (Figure S5).

For the improved identifications, an estimate of the FDR control is crucial. We estimated the FDR by including predicted spectra from other species. Identifications from these species were counted as false positives. For these libraries from other species, we also digested the protein sequences with dpMC and predict the intensities and retention time by dpMS and dpRT respectively. We prepared the libraries of other species with the same number of proteins as the corresponding libraries built above for *D. melanogaster* and HeLa. We used a *S. cerevisiae* library of 5006 proteins which corresponds to *D. melanogaster* DDA library, the library of *C. elegans* and *D. discoideum* containing 10,524 proteins which corresponds to Pan-Human library, and the library of *S.cerevisiae* including 4460 proteins corresponding to the HeLa DDA library. For the transcriptome wide library, 17,299 proteins and 8758 proteins from *C. elegans* and *D. discoideum*

were prepared for entrapment library of *D.melanogaster* and HeLa, respectively.

We applied the entrapment strategy by pooling the entrapment libraries with their corresponding target libraries together to check the false positives, which revealed the false positives identified by the interferences of each other species. By calculation of the entrapments in the DIA searches based on DDA measurements or the transcriptome, we found the FDR was slightly higher when using larger libraries. For both DDA based libraries of *D. melanogaster* and HeLa, the FDR was around 1%, while it was around 2% for the transcriptome (Figure S6). Such a streamlining of the library is also intrinsically achieved by the use of an accurate prediction algorithm for missed cleavages such as dpMC. Based on the above FDR analysis, we showed the robustness of our method and strategy to buihighly accurate spectral libraries for SWATH-MS analysis.

In agreement with previous findings, the correlation (PCC) between the logarithmically transformed abundance of gene expression (RPKM) and protein intensities is rather moderate with a PCC value of 0.55 and 0.52 for *D. melanogaster* and HeLa respectively (Figure S7). Besides, for different scales of libraries built for *D. melanogaster*, the similarities between replicates for each stage of embryo development were also shown in Figure S8, in which the high correlations between replicates indicate the high quality of in silico libraries built by dpSWATH.

4 | DISCUSSION

The accurate theoretical prediction of peptide fragment spectra holds great promise for an improved quantification of entire proteomes using DIA methods such as SWATH-MS. Recently different models were developed to achieve a higher quality when predicting mass spectra. For example, Prosit [15] uses Collision Energy as an additional feature to train their model. However, for Q-TOF instruments the collision energy only marginally increases the accuracy of prediction [23], suggesting that many other subtle factors that could also affect the behavior of mass spectra. To consider such other, potentially unknown factors, we developed dpMScore to filter out the unreliable fragments spectra, which resulted in a more consistent and high-quality training and testing datasets for dpSWATH, in particular when using lower quality Q-TOF data.

The highly accurate prediction of mass spectra pattern and retention time makes SWATH-MS analysis methods more widely applicable. The reliable and effective workflow of dpSWATH, enables an fast generation and an efficient use of theoretical libraries. Based on the predicted library we built for *D. melanogaster* and *H. sapiens* (HeLa), we identified more proteins and peptides compared to an experimental library. This increase on the proteome coverage will favor a more comprehensive analysis of the biological system of interest.

During the development of the algorithm and its application to a wide range of data sets, we realized that the selection of consensus fragment mass spectra based on the dpMScore clustering algorithm is especially important for lower quality MS/MS spectra as the once recorded with non-trapping Q-TOF instruments. As these fragment spectra are substantially influenced by a various extrinsic factor such as the build of the instrument, humidity, temperature external electric fields et cetera, we suggest building the theoretical library based on the training datasets on the same platform and experimental conditions. Moreover, it turned out that the accuracy of peptide identification can be substantially improved by reducing the search space when building in silico libraries. In our proof-of concept studies we did this by applying a highly accurate prediction of missed tryptic cleavages using dpMC and a restriction to the proteins that are known to be expressed in the samples. The information about the proteins expressed in the studied sample(s) can be relatively easily gathered by RNA-Seq analysis or by a deep proteomic analysis of a pool of all samples. Based on our analysis, the transcriptome-based theoretical library showed the highest identification rate while maintaining FDR as the library based on a DDA measurement.

In summary, dpSWATH allows a robust and reliable prediction of fragment spectra that can be used in SWATH analyses therefore allowing a rapid and efficient quantification of a higher number of proteins and peptides compared to the classical DDA experiments or DIA experiments that rely on experimentally generated libraries.

ACKNOWLEDGEMENTS

B.S. was funded by the Chinese Scholarship Council (201506230154). Work in the lab of A.I. and T.S. were funded by grants of the Deutsche Forschungsgemeinschaft (CRC1064 (project number: 2133249687 (A.I.) and 219249687 (T.S.)) and 1309 (project number 325871075 (A.I.)) and the German Federal Ministry of Education and Research (BMBF FKZ161L0214F, ClinspectM).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The method and tools are open source on https://github.com/dpSWATH-sun/dpSWATH

ORCID

Axel Imhof https://orcid.org/0000-0003-2993-8249

REFERENCES

1. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, *14*(8), e8126. https://doi.org/10.15252/msb.20178126
2. Krasny, L., Bland, P., Kogata, N., Wai, P., Howard, B. A., Natrajan, R. C., & Huang, P. H. (2017). SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. *Scientific Reports*, *7*, 45913. https://doi.org/10.1038/srep45913
3. Liu, Y., Hüttenhain, R., Collins, B., & Aebersold, R. (2013). Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert Review of Molecular Diagnostics*, *13*(8), 811–825. https://doi.org/10.1586/14737159.2013.845089
4. Sidoli, S., Lin, S., Xiong, L., Bhanu, N. V., Karch, K. R., Johansen, E., Hunter, C., Mollah, S., & Garcia, B. A. (2015). Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) analysis for

characterization and quantification of histone post-translational modifications. *Molecular & Cellular Proteomics*, *14*(9), 2420–2428. https://doi.org/10.1074/mcp.O114.046102

5. Lambert, J.-P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z.-Y., Zhong, Q., Lindquist, S., Vidal, M., Aebersold, R., Pawson, T., Bonner, R., Tate, S., & Gingras, A.-C. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nature Methods*, *10*(12), 1239–1245. https://doi.org/10.1038/nmeth.2702
6. Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R L., Ohtsuki, S., … Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, *8*(1), 291. https://doi.org/10.1038/s41467-017-00249-5
7. Messner, C B., Demichev, V., Bloomfield, N., Yu, J. S. L., White, M., Kreidl, M., Egger, A.-S., Freiwald, A., Ivosev, G., Wasim, F., Zelezniak, A., Jürgens, L., Suttorp, N., Sander, L. E., Kurth, F., Lilley, K S., Mülleder, M., Tate, S., & Ralser, M. (2021). Ultra-fast proteomics with Scanning SWATH. *Nature Biotechnology*, *39*(7), 846–854. https://doi.org/10.1038/s41587-021-00860-4 From NLM
8. Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., & Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, *32*(3), 219–223. https://doi.org/10.1038/nbt.2841
9. Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., Maclean, B., & Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, *10*(3), 426–441. https://doi.org/10.1038/nprot.2015.015
10. Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C., & Molloy, M. P. (2016). SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Molecular & Cellular Proteomics*, *15*(7), 2501–2514. https://doi.org/10.1074/mcp.M115.055558
11. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., & Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, *5*(10), 873–875. https://doi.org/10.1038/nmeth.1254
12. Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., Faini, M., Schubert, O T., Faridi, P., Ebhardt, H. A, Matondo, M., Lam, H., Bader, S L., Campbell, D S., Deutsch, E. W., … Aebersold, R. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS, *Science Data*, *1*, 140031. https://doi.org/10.1038/sdata.2014.31
13. Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., & Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, *11*(1), 146. https://doi.org/10.1038/s41467-019-13866-z
14. Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.-M, & Zhang, Z. (2017). pDeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, *89*(23), 12690–12697. https://doi.org/10.1021/acs.analchem.7b02566
15. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., & Wilhelm, M. (2019). Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, *16*(6), 509–518. https://doi.org/10.1038/s41592-019-0426-7
16. Kantz, E. D., Tiwari, S., Watrous, J. D., Cheng, S., & Jain, M. (2019). Deep neural networks for classification of LC-MS spectral peaks. *Analytical Chemistry*, *91*(19), 12407–12413. https://doi.org/10.1021/acs.analchem.9b02983 From NLM
17. Tran, N. H., Zhang, X., Xin, L., Shan, B., & Li, M. (2017). De novo peptide sequencing by deep learning. *PNAS*, *114*(31), 8247–8252. https://doi.org/10.1073/pnas.1705691114 From NLM
18. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., & Bandeira, N. (2018). Assembling the community-scale discoverable human proteome. *Cell Systems*, *7*(4), 412–421.e5 e415. https://doi.org/10.1016/j.cels.2018.08.004
19. Guan, S., Moran, M. F., & Ma, B. (2019). Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular & Cellular Proteomics*, *18*(10), 2099–2107. https://doi.org/10.1074/mcp.TIR119.001412
20. Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., Mehnert, M., Seifert, M., Emmenlauer, M., Sorg, I., Bezrukov, F., Bena, F. S., Zhou, H., Dehio, C., Testa, G., … Aebersold, R. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology*, *37*(3), 314–322. https://doi.org/10.1038/s41587-019-0037-y
21. Völker-Albert, M. C., Pusch, M. C., Fedisch, A., Schilcher, P., Schmidt, A., & Imhof, A. (2016). A quantitative proteomic analysis of in vitro assembled chromatin. *Molecular & Cellular Proteomics*, *15*(3), 945–959. https://doi.org/10.1074/mcp.M115.053553
22. Keil, B. (1992). *Specificity of proteolysis* (p. 335). Springer-Verlag Berlin-Heidelberg.
23. Ammar, C., Berchtold, E., Csaba, G., Schmidt, A., Imhof, A., & Zimmer, R. (2019). Multi-reference spectral library yields almost complete coverage of heterogeneous LC-MS/MS data sets. *Journal of Proteome Research*, *18*(4), 1553–1566. https://doi.org/10.1021/acs.jproteome.8b00819
24. Searle, B. C., Swearingen, K. E., Barnes, C. A., Schmidt, T., Gessulat, S., Küster, B., & Wilhelm, M. (2020). Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature Communications*, *11*(1), 1548. https://doi.org/10.1038/s41467-020-15346-1 From NLM
25. Van Puyvelde, B., Willems, S., Gabriels, R., Daled, S., De Clerck, L., Vande Casteele, S., Staes, A., Impens, F., Deforce, D., Martens, L., Degroeve, S., & Dhaenens, M. (2020). Removing the hidden data dependency of DIA with predicted spectral libraries. *Proteomics*, *20*(3-4), 1900306. https://doi.org/10.1002/pmic.201900306 From NLM
26. Sun, B., Smialowski, P., Straub, T., & Imhof, A. (2021). Investigation and highly accurate prediction of missed tryptic cleavages by deep learning. *Journal of Proteome Research*, *20*(7), 3749–3757. https://doi.org/10.1021/acs.jproteome.1c00346 From NLM
27. Gao, Z., Chang, C., Yang, J., Zhu, Y., & Fu, Y. (2019). AP3: An advanced proteotypic peptide predictor for targeted proteomics by incorporating peptide digestibility. *Analytical Chemistry*, *91*(13), 8705–8711. https://doi.org/10.1021/acs.analchem.9b02520
28. Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226 From NLM

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202200179 in the Supporting Information section at the end of the article.

How to cite this article: Sun, B., Smialowski, P., Aftab, W., Schmidt, A., Forne, I., Straub, T., & Imhof, A. (2023). Improving SWATH-MS analysis by deep-learning. *Proteomics*, e2200179. https://doi.org/10.1002/pmic.202200179