DIA-NN: Deep neural networks substantially improve the identification performance of Data-independent acquisition (DIA) in proteomics

Vadim Demichev[1,2*], Christoph B. Messner[2], Kathryn S. Lilley[1], Markus Ralser[1,2,3]

1. Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, 80 Tennis Court Rd, Cambridge, CB2 1GA, UK
2. The Molecular Biology of Metabolism laboratory, The Francis Crick Institute, 1 Midland Rd, NW1 1AT London, UK
3. Department of Biochemistry, Charité, Universitaetsmedizin, Berlin, GER

*To whom correspondence should be addressed. vadim.demichev@gmail.com

Abstract

Data-independent acquisition (DIA-MS) strategies, like SWATH-MS, have been developed to increase consistency, quantification precision and proteomic depth in label-free proteomic experiments. They aim to overcome stochasticity in the selection of precursor ions by utilising (mass-) windowed acquisition that is followed by computational reconstruction of the chromatograms. While DIA methods increasingly outperform typical data-dependent methods in identification consistency and precision specifically on large sample series, possibilities remain for further improvements. At present, only a fraction of the information recorded in the complex DIA spectra is extracted by the software analysis pipelines. Here we present a software tool (DIA-NN) that introduces artificial neural nets and a new quantification strategy to enhance signal processing in DIA-data. DIA-NN greatly improves identification of precursor ions and, as a consequence, protein quantification accuracy. The performance of DIA-NN demonstrates that deep learning provides opportunities to boost the analysis of data-independent acquisition workflows in proteomics.

Introduction

Mass spectrometry-based quantitative proteomics is an invaluable tool in a wide range of clinical and research applications. Data-independent acquisition (DIA) is devoid of the inherent stochasticity of data-dependent acquisition (DDA) workflows, which manifests itself as missing measurements between successive runs. Its identification performance is hence not limited by stochastic elements and only indirectly dependent on the sampling velocity of the mass analyzer. In addition, quantification via DIA is more reliable, as it is performed on the fragment (MS/MS or MS2) level and is thus less susceptible to interferences [1]. Introduced already in 1970s, with new instruments and acquisition strategies available (like SWATH-MS [2,3]), DIA is becoming increasingly popular. In an increasing number of instances, DIA workflows outperform label-free DDA in terms of reproducibility [1,4]. Recently, a highly-optimised DIA workflow has been shown to be capable of identifying more precursors than the theoretical maximum number of tandem spectra that can be acquired in the DDA mode [4]. Thus, considerable attention has been dedicated to devising advanced algorithms for processing DIA-proteomics data [5–17].

A main restriction remains in the computational processing of DIA data, as the available algorithms are still underperforming compared to the theoretical information content of DIA data. In particular, the use of wide isolation windows, which are needed to achieve fast scan times, is associated with high levels of interferences in the tandem-MS spectra [18]. The interferences cause false-positive discoveries as well as lower accuracy and precision in proteomic experiments.

Here, we demonstrate the application of artificial neural nets to the extraction of information from the complex SMATH-MS data. Our approach allows to increase the precursor identification performance and, therefore, the number of protein groups that can be accurately quantified. Neural networks have been used previously to classify spectra in DDA-proteomics data [19–21]. However, to the best of our knowledge, DIA-proteomics data processing has so far only employed linear classifiers such as linear discriminant analysis or support vector machines [7–9,13,16,17]. In addition, we introduce an efficient method for the removal of interferences from tandem-MS spectra, thus significantly improving quantification precision. Our software tool, DIA-NN (Data Independent Acquisition by Neural Networks), implements all stages of DIA data processing in a single program, taking a set of raw data files as input and reporting quantitative values for precursor ions and protein groups. DIA-NN is written in C++ and is designed to be fast. Its memory requirements are independent of the number of LC-MS runs in the experiment, as the computationally intensive processing of raw data is performed for each run separately, and the results are saved to the hard drive (the space requirements are several orders of magnitude less than the volume of the respective raw data files). This makes DIA-NN ideal for automated handling of data generated in large-scale experiments.

Results

1. General architecture of DIA-NN

DIA-NN takes as input a collection of centroided mass spectrometry data files (corresponding to individual runs) and a spectral library. Each of the files is processed separately to match precursor ions to elution peaks, the result being saved to the hard drive. Quantification of fragment ions generated from the precursors is also performed at this stage. Interferences are detected and removed for each fragment.

To speed up the data processing, DIA-NN supports optional restriction of the number of elution peaks considered as potentially valid matches for a given precursor. For this, DIA-NN can use a defined set of added or internal reference precursors. These can be provided as a list by the user or automatically generated by DIA-NN using high confidence identifications in previous analyses with the same spectral library. Once elution peaks for these precursor ions are identified, they allow the relationship between the retention times in the data file and in the spectral library to be inferred. DIA-NN then only searches for elution peaks within an automatically generated retention time window.

False discovery rate (FDR) is estimated using a modified version of the target-decoy method [22]. Briefly, for a precursor ion, target or decoy, a set of scores is calculated for each potential elution peak. First, one of the scores is used to select the best peak. Second, a classifier is trained to distinguish between the sets of scores corresponding to the best peaks matched to target and decoy precursors. It is then used to generate a "combined score" that allows to refine the selection of peaks. The process is repeated iteratively for a specified number of iterations. In our tests, eight iterations proved to be sufficient for the efficient training of the classifier. After each iteration, the mapping between the actual and library retention times is refined; the normalised retention time of the elution peak as well as the square root of its difference from the library retention time are added to the set of scores. Finally, the ratio of decoy to target precursor numbers with combined scores exceeding a given threshold is used as the FDR estimate.

Once the initial processing of all runs in the experiment is finished, DIA-NN quantifies precursor ions using the previously collected information on the quality of extracted ion chromatograms of their individual fragments. The best three fragments per precursor are selected in a cross-run manner and eventually used for its quantification. DIA-NN also supports automatic cross-run retention time profiling. Briefly, if a precursor is identified in some runs with FDR lower than a specified threshold, the retention time information in the spectral library is corrected based on the run in which this precursor was identified with lowest FDR. All the runs are then reanalysed using corrected retention times.

After precursor ion quantification, optional cross-run normalisation and protein quantification can be performed. All the precursor intensities corresponding to identifications with FDR

estimates above a given threshold are replaced with zeroes. Precursors are then ordered by their coefficients of variation. Top pN precursors are selected, where N is the average number of identifications passing the FDR threshold and p is between 0 and 1. Sums of the intensities of these precursors are calculated and are used for normalisation, i.e. the levels of all precursors are scaled to make these quantities equal in different runs. A "Top N" method is eventually used for protein quantification: protein intensities are obtained as sums of the intensities of top N most abundant precursors identified at FDR lower than a given threshold in a particular run.

1.1. Decoy precursors generation

A decoy precursor is generated for each target precursor. The m/z value of the decoy precursor as well as its reference retention time are set to be equal to those of the target precursor, as specified in the spectral library. The order of amino acid residues (except for the first and the last ones) is reversed. If the resulting sequence of amino acid masses happens to be the same as the one corresponding to the target precursor, then the mass of the central amino acid is increased by an artificial value (i.e. 12 m/z). The fragmentation spectrum of the decoy precursor is then calculated using the same fragmentation pattern as that of the target precursor.

1.2. Detection and scoring of elution peaks

Each precursor is represented in the spectral library by its m/z value as well as the m/z values and reference intensities of its fragments. Six fragments with the highest reference intensities are considered in DIA-NN, and their m/z values are used to extract the respective chromatograms from the data. First, the sequence of spectra is filtered to leave only MS1 spectra and those MS2 spectra that were obtained using precursor ion selection windows containing the m/z value of the precursor. In each of the remaining spectra, the highest peak is chosen within a window centered at the m/z value of interest. The radius of this window is calculated as the product of this m/z value and the specified mass accuracy coefficient.

The chromatograms are scanned using retention time (RT) windows. The window size can be either specified by the user or chosen by DIA-NN automatically. In the latter case, the diameter of the RT window is taken to be $1 + 4.4 \cdot fwhm$, where $fwhm$ is the average peak width at half maximum for the reference precursors. These can be defined by the user or inferred automatically as high-confidence precursors during an initial search with a wide scan window. This procedure is carried out for the first run in the experiment, subsequent runs using the same RT window size. DIA-NN calculates pairwise correlations of elution curves of all fragments of the precursor within the window. The fragment with the largest sum of correlations is then designated as the “best” fragment. It is assumed that this fragment is likely to be the one least affected by interferences; hence its elution curve is expected to be representative of the true elution curve of the precursor. If the level of the best fragment at the center of the window is close to its maximum level within the window, then this window is

considered as an elution peak. The elution curve of the best fragment is smoothed and is designated as the "reference" elution curve. For each elution peak, a set of scores is then calculated.

Signal scores. Total signals of the top five fragments normalised by the total signal of all fragments are used as scores. These scores are ordered by the reference intensities of the respective fragments.

Correlation scores. First, correlations of the fragments' elution curves with the reference curve are calculated. The sum of these correlations is used as a score. Second, each fragment chromatogram is processed by taking minima of each three consecutive values, correlations with the reference curve are calculated, and the sum of these is used as another score. Finally, the correlation between the MS1 elution curve and the reference curve is also used as a score.

Additional scores. The cosine between the set of reference intensities of the fragments (obtained from the spectral library) and their measured intensities is calculated for each time point in the RT window. These values are then weighted by the squared values of the reference elution curve at the respective time points, summed, and the resulting quantity is used as a score. The normalised retention time of the elution peak as well as the square root of its difference from the library retention time are also used as scores.

1.3 Target-decoy classification

DIA-NN implements two different classifiers: a linear classifier and an artificial neural network classifier. Optionally, the set of precursor ions is randomly split in the proportion 3:1 to generate the training and test datasets, respectively. Only the test dataset is then used to calculate the "combined score" level that corresponds to a given FDR threshold.

Linear classifier. DIA-NN calculates $\{\Delta_p\}_p$ - the set of score vector differences between the paired target and decoy precursors. Here p denotes a target-decoy pair. A weight vector w is then obtained as the solution of the equation $Rx = \mu$, where μ is the average of these vector differences and R is their covariance matrix. The "combined score" ws is then calculated for each score vector s.

The *Neural network classifier* is employed for the final two iterations of elution peak scoring. Training of the neural network is only performed during the first of these iterations. In comparison to the linear classifier, some additional scores are used. First, correlations of individual fragments' elution curves with the reference curve are included. Second, the elution curve of each fragment is split into five segments and the total signals in each of these are normalised by their sum and used as scores. In addition, the reference intensities of the fragments (L1-normalised) are also considered. Finally, the normalised total signal of the sixth fragment (skipped in the linear classifier training) is added to the set of scores. A feed-forward neural network is then trained to distinguish between the target and decoy precursors for a specified number of epochs using batch gradient descent with momentum. The network

consists of the input layer, a specified number of hidden fully connected *tanh* layers, and a softmax layer. The size of n-th hidden layer is set to $5(N - n + 1)$, where N is the number of hidden layers. Optionally, standardisation of the input and L2 regularisation of the neural network weights can be performed. Cross-entropy is used as the loss function. Ensemble learning is employed to reduce the effects of overfitting: a specified number of networks having random initial weights are trained independently in parallel, and their predictions are averaged. Optionally, the neural network can be pre-trained in a cross-run manner.

1.4. Interference correction

For quantification, the size of the chromatogram RT window is halved. The elution curve $x(\cdot)$ of each fragment is compared to the reference curve $ref(\cdot)$ - the smoothed elution curve of the best fragment. The "weighted" fragment intensity is calculated as the sum of the fragment elution curve values weighted by the respective squared values of the reference curve. This emphasises the contribution of the data points close to the apex of the reference elution curve, thus making the impact of potential interferences manifesting far from the apex negligible. The ratio r of weighted intensities of the fragment under consideration and the best fragment is calculated. All values of $x(\cdot)$ exceeding $1.5 \cdot ref(\cdot) \cdot r$ are replaced with $1.5 \cdot ref(\cdot) \cdot r$. The area under the resulting curve is then considered to be the intensity of the fragment.

1.5. Cross-run precursor ion quantification

DIA-NN enables cross-run precursor ion quantification. In each run, each fragment is assigned a score which is the correlation score of its elution curve with the respective reference curve, i.e. the smoothed elution curve of the best fragment. For each precursor, three fragments with highest average correlations are selected in a cross-run manner. For this, only runs where the precursor was identified with FDR below a given threshold are considered. The intensities of these fragments are then summed in each run to obtain the precursor ion intensity.

2. Identification performance of DIA-NN

We explored the precursor identification performance of DIA-NN when using the linear classifier and the artificial neural network classifier, which was trained in either run-specific or cross-run manner. We compared the performance characteristics of DIA-NN to the results obtained with Spectronaut Pulsar (version 11.0.15038.17.27438, Biognosys AG) [1], one of the currently most advanced commercial software solutions for the analysis of DIA-data. To perform the comparison, we generated a 40-variable window LC-SWATH-MS dataset using a ten-time injection of yeast (*S. cerevisiae*) full-proteome tryptic digests, and that has been analysed on a microLC system (Waters nanoAcquity) coupled to a Quadrupole Time of Flight (QqTOF) mass spectrometer (TripleTOF 6600, SCIEX). For other benchmarks, we use

a dataset previously generated as part of the LFQbench test, and the spectral library generated therein [23].

As different software tools employ different decoy precursor generation algorithms and use different classifiers, software-reported false discovery rates cannot be compared directly. We therefore followed a strategy suggested previously [4], and 'tested' for false positive hits, by calling *E. coli* peptides in a yeast proteome. Almost every call of an *E.coli*-specific peptide in the pure yeast digest will be a false positive; and the number of such calls is hence an illustrative benchmark that visualizes the FD performance of the different methods (**Fig. 1**). We utilised a compound spectral library, which comprised both yeast and *E.coli* precursors (15153 and 13550, respectively). For this, a human-yeast-*E.coli* spectral library was obtained from the LFQbench test suite (64 var, openswath) [23], and precursors matching to human proteins were removed (calling for human-specific peptides is a less reliable benchmark against calling false positives, as human contamination due to sample handling is difficult to exclude). Both software tools were operated with standard parameters; in Spectronaut, precursor and protein qvalue cutoff thresholds were set to 1.0. The average numbers of total and *E.coli*-specific precursor identifications in the yeast proteome were calculated at different reported FDR levels.

Written in C++, DIA-NN demonstrated high processing speed, once it converts the input mzML files with the raw data into its own format. For example, when using a neural network with two hidden layers trained in a run-specific manner, DIA-NN was able to process all the ten files in less than 13 minutes on a average processing workstation (2x 6-core Intel Xeon E5645). The fast performance renders DIA-NN suitable for the analysis of very large datasets.

In this comparison, DIA-NN outperformed the Spectronaut-implemented data analysis pipeline both with the linear classifier as well as with the artificial neural network classifier (**Fig. 1**). Application of the latter substantially suppressed false peptide discoveries: the number of falsely discovered *E. coli* precursors (in the yeast dataset) was reduced to about 38-48 at approximately 8000 precursor identifications, which is a 1.6- to 2-fold improvement over the linear classifier. Different settings for the neural network classifier show comparable performance.

3. Quantification performance of DIA-NN

While the identification performance is important, the key application of DIA is accurate, precise and consistent protein quantification in large sample series [24]. We illustrate the quantification performance of DIA-NN by comparing it to Spectronaut Pulsar using the LFQbench test [23]. In this benchmark, human, yeast, and *E.coli* lysates were mixed in different proportions and analysed via SWATH-MS. For each mixture, three injection

replicates were measured. The LFQbench R package takes as input quantitative values for precursor ions and uses these to quantify peptides and proteins. Two characteristics of these are being monitored: median bias ("accuracy") and standard deviation ("precision"). Median coefficients of variation in technical replicates (“technical variance”) are also calculated for human peptides and proteins.

We considered the 64 variable windows acquisition schemes on TripleTOF 6600 (Sciex) datasets, as these were revealed in the LFQbench manuscript to perform best [23]. Both the software tools were operated with standard parameters; in Spectronaut, precursor and protein qvalue cutoff thresholds were set to 1.0; DIA-NN was set to use a run-specific neural network classifier with two hidden layers (**Fig. 2** and **3**). In order to be able to compare the quantification performance of the two methods on the same dataset, we aimed to fix the number of identified peptides, adjusting the FDR threshold supplied to the LFQbench R package accordingly. This strategy was necessary, as the FDR values reported by the software tools can not be compared directly. The number of peptides corresponds to an (FDR) cutoff of 0.01 on DIA-NN, and to 0.002 (HYE124 dataset) and 0.006 (HYE110 dataset) on Spectronaut.

With the FDR thresholds defined to obtain a similar number of peptide IDs from the same raw data, accuracy (bias) of quantification is comparable between DIA-NN and Spectronaut Pulsar. DIA-NN, however, performed better in most individual benchmark values considered (accuracy, precision, and technical variance on all Yeast, *E. coli* or human peptide and protein quantities; better performance values for both methods are highlighted in green). It also demonstrated lower technical variance and better quantification precision on both datasets (**Fig. 2**, **3**). For example, in case of the yeast and *E.coli* proteins in the HYE110 dataset, in which their ratios in different samples were expected to be very high (10:1), DIA-NN shows almost 1.5-times better precision.

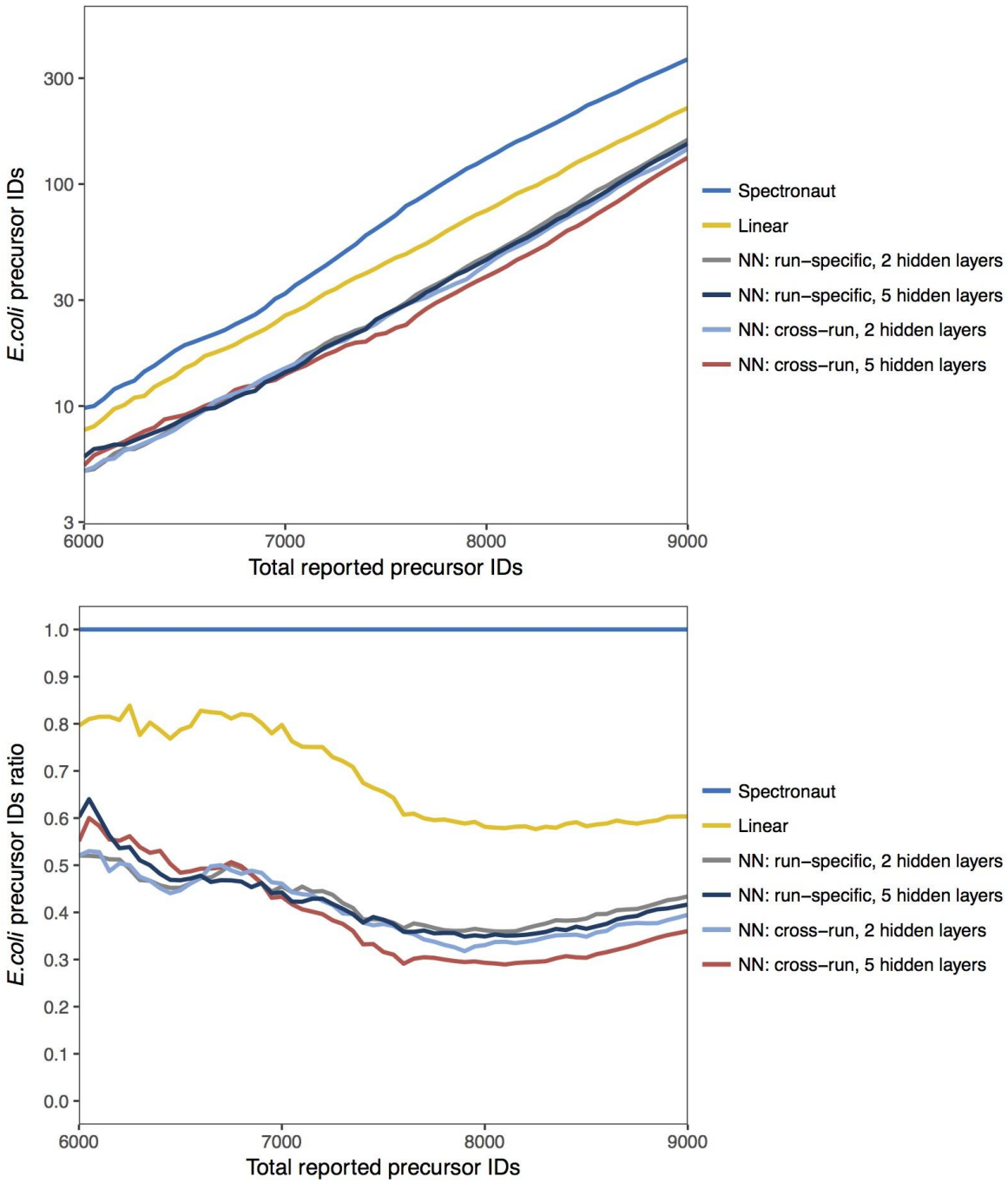

Figure 1. **Identification performance of DIA-NN, in comparison to Spectronaut Pulsar**. Ten yeast full-proteome digest injections were analysed via SWATH-MS. The identification performance of DIA-NN with different settings was compared to that of Spectronaut Pulsar. Data processing was performed with a spectral library which featured both yeast- and *E.coli*-specific precursor ions (15153 and 13550, respectively). Since no *E.coli*-specific precursors are expected to be present in the samples, the numbers of false identifications are proportional to the respective numbers of reported identifications of *E.coli*-specific precursors. For each software configuration tested, these were plotted against the total number of reported identifications (which depends on the reported FDR threshold used) (top panel; lower is better). The numbers of false identifications normalised by the respective numbers of Spectronaut Pulsar are also presented (bottom panel). On this dataset, DIA-NN outperforms the conventional data analysis pipeline even when using a linear classifier; artificial neural networks allow to further reduce the false discovery rate.

	DIA-NN		Spectronaut Pulsar	
FDR cutoff	0.01		0.002	
	peptides	proteins	peptides	proteins
Valid ratios Human	16364	2012	16276	1995
Valid ratios Yeast	11255	1320	11317	1343
Valid Ratios *E.coli*	7780	885	7115	821
Accuracy Human	0.0	0.0	0.0	0.0
Accuracy Yeast	-0.04	-0.02	-0.07	-0.07
Accuracy *E.coli*	0.21	0.14	0.22	0.14
Precision Human	0.24	0.12	0.27	0.13
Precision Yeast	0.30	0.32	0.41	0.36
Precision *E.coli*	0.48	0.40	0.57	0.51
Technical variance	0.045	0.026	0.056	0.030

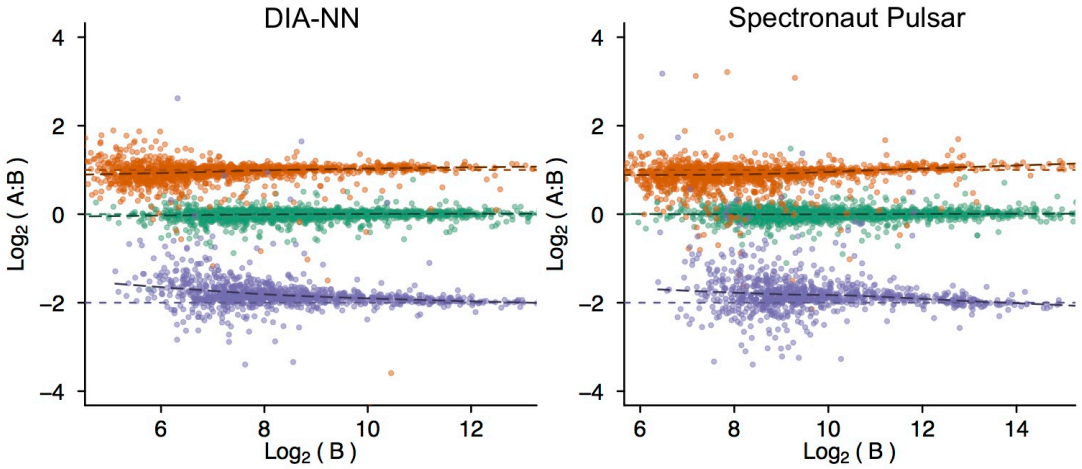

Figure 2. **Quantification performance of DIA-NN on the HYE124 dataset, in comparison to Spectronaut Pulsar**. The LFQbench R package was used to compare the quantitative performance of DIA-NN (neural network classifier with 2 hidden layers, trained in run-specific manner) and Spectronaut Pulsar. In this test suite, human, yeast, and *E.coli* lysates were mixed in different proportions (sample series A and B) and analysed via SWATH-MS. For each data processing tool, median bias ("accuracy") and standard deviation ("precision") are calculated for the detected peptides and proteins from each species (top panel). Median coefficients of variation in technical replicates ("technical variance") are also calculated for human peptides and proteins. Graphic representation of the detected proteins is presented (bottom panel).

	DIA-NN		Spectronaut Pulsar	
FDR cutoff	0.01		0.006	
	peptides	proteins	peptides	proteins
Valid ratios Human	15380	1927	15290	1907
Valid ratios Yeast	2929	422	2918	456
Valid Ratios *E.coli*	3964	498	3975	506
Accuracy Human	0.0	0.0	0.0	0.0
Accuracy Yeast	0.11	0.17	0.04	0.07
Accuracy *E.coli*	0.03	-0.04	0.09	0.09
Precision Human	0.34	0.18	0.37	0.23
Precision Yeast	1.01	0.78	1.43	1.13
Precision *E.coli*	0.74	0.67	1.11	1.02
Technical variance	0.060	0.032	0.070	0.037

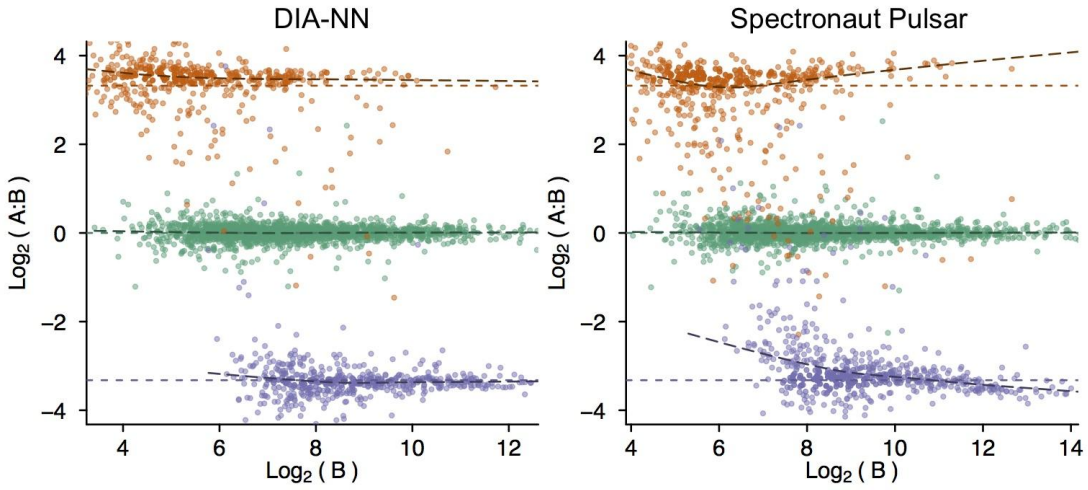

Figure 3. **Quantification performance of DIA-NN on the HYE110 dataset, in comparison to Spectronaut Pulsar**. The LFQbench R package was used to compare the quantitative performance of DIA-NN (neural network classifier with 2 hidden layers, trained in run-specific manner) and Spectronaut Pulsar. In this test suite, human, yeast, and *E.coli* lysates were mixed in different proportions (sample series A and B) and analysed via SWATH-MS. For each data processing tool, median bias ("accuracy") and standard deviation ("precision") are calculated for the detected peptides and proteins from each species (top panel). Median coefficients of variation in technical replicates ("technical variance") are also calculated for human peptides and proteins. Graphic representation of the detected proteins is presented (bottom panel).

Summary and Discussion

In this work, we aimed to maximise the amount of information extracted from DIA-proteomics data. Ultimately, the precursor identification problem relies on the classification of potential elution peaks as false or true positives. Linear classifiers, such as linear discriminant analysis or support vector machines, usually require careful selection of a small set of scores to be calculated for each elution peak. Since DIA-proteomics generates highly information-rich datasets (unlike DDA), conventional approaches to handling DIA data effectively discard a significant portion of information. To address this issue, we developed a set of highly optimised scores, which encodes smoothed elution curves for each fragment of the precursor ion under consideration. We demonstrated that these scores allow for an efficient solution of the classification problem with the use of deep neural networks.

In this preprint, we have benchmarked the performance of DIA-NN against only one software tool, Spectronaut Pulsar. Given the widely recognised performance of Spectronaut and its broad use, we think this is the most reasonable choice. This benchmark is hence not yet comprehensive. It is important to mention in this context however, that in the LFQ benchmark study, several of the previously available software tools (DIA-Umpire 2.0, OpenSWATH[1], Skyline 3.1.1.8669, Spectronaut 7.0.8065.1.24792 and SWATH 2.0) demonstrated converging performances. Spectronaut Pulsar is advanced compared to these tools, and hence is expected to perform better.

So far we have benchmarked this first version of DIA-NN only on a small number of datasets; we expect its relative advantage in peptide quantification to be dependent on the dataset. The main advantage of DIA-NN is achieved with the use of an artificial neural network, the performance of which will depend on the number of true and false target precursor identifications used for its training as well as on such parameters as the number of data points per peak. The performance difference between DIA-NN and other tools will hence vary from dataset to dataset (it is not a ‘digital’, universal value) and will also depend on the spectral library used. However, we currently still work on improving DIA-NN; we will conduct a more comprehensive benchmark once a more ‘final’ version of the software is generated and a manuscript prepared for submission. We publish this preprint, to encourage other proteomic labs to try DIA-NN already at this stage; any feedback provided will help to improve it for the use in the community, or for its incorporation in commercial DIA software.

In the identification benchmark conducted with the linear classifier, DIA-NN outperformed the conventional data analysis pipeline, exhibiting lower rate of false positive identifications across the wide range of reported precursor identification numbers considered. The use of artificial neural networks substantially improved this result. Neural networks with two and five hidden layers showed comparable performance; there was also little difference between

[1] https://github.com/OpenMS/OpenMS/commit/4bca6fc

training the network in a run-specific or a cross-run manner. This situation may, however, be different on more variable or complex samples.

We also addressed another crucial issue associated with DIA-proteomics data processing, namely, the problem of interference removal to facilitate accurate and precise quantification of precursor ions and, therefore, proteins. For this, we introduced a novel approach based on selecting the "best" fragment per precursor in each run. Its elution curve is then assumed to be representative of the true elution curve of the precursor, and is used as a template for correcting interferences affecting other fragments.

We validated the high performance of this method using the LFQbench test suite. We observed that while the accuracy (bias) of quantification is comparable between DIA-NN and Spectronaut Pulsar, DIA-NN tends to be superior in quantification precision and exhibits lower technical variance.

In summary, DIA-NN is a fast software tool for processing of DIA proteomics data. Applied in datasets generated to benchmark SWATH platforms, it shows improved precursor identification performance as well as high quantification accuracy and precision. Our study demonstrates the power of using artificial neural networks in the analysis of DIA-proteomics data.

Methods

DIA-NN settings

DIA-NN was run using its default settings. Briefly, the chromatogram scan window size was determined automatically. Mass accuracy was set to 20ppm. For cross-run neural network training, complete libraries were used. For run-specific classifier training, libraries were randomly split into the training and test datasets in a 3:1 ratio. The total number of classifier training iterations was set to eight. The neural network featured either two hidden layers (default) or five hidden layers. Standardisation of the neural network input and regularisation of weights were not used. The number of training epochs was set to 50. Batch size was set to Max(50, N / 100), where N is the number of training observations. Learning rate was set to 10.0, and momentum was set to 0.9. Twelve neural networks were trained in parallel. For cross-run precursor quantification, FDR threshold was set to 0.01. For cross-run normalisation, top 40% of precursors with lowest coefficients of variation and having estimated FDR levels less than 0.001 were used.

DIA-NN implementation

To facilitate the very high processing speed demonstrated by DIA-NN, its code was written in C++. DIA-NN relies on the following third-party libraries:

- Cranium (https://github.com/100/Cranium) provides functionality necessary for the implementation of the artificial neural network classifier;
- MSToolkit (https://github.com/mhoopmann/mstoolkit) provides an interface for files in the mzML format;
- Eigen (http://eigen.tuxfamily.org) is used to solve linear equations.

DIA-NN executable (Windows, x64) as well as the source code are available for download from https://github.com/vdemichev/DiaNN (version 1.0).

Yeast DIA analyses

Saccharomyces cerevisiae (BY4743 rendered prototrophic with a plasmid encoding for *HIS3, LEU2* and *URA3* [25]) were grown to exponential phase in minimal synthetic nutrient media. Proteins were extracted by bead beating for 5min at 1500rpm in 8M urea/0.1M ammonium bicarbonate. Proteins were reduced with 5mM dithiothreitol, alkylated with 10mM iodoacetamide. The sample was diluted to 1.5M urea/0.1M ammonium bicarbonate before the proteins were digested overnight with Trypsin (1:30 Trypsin to total protein ratio). Peptides were cleaned-up with 96-well MacroSpin plates (Nest Group) and iRT peptides (Biognosys AG) were spiked in.

The digested peptides were analysed on a nanoAcquity (Waters) coupled to a TripleTOF 6600 (Sciex). Peptides were separated with a 23 minute non-linear gradient (4% Acetonitrile/0.1 % formic acid to 36% Acetonitrile/0.1% formic acid) on a Waters HSS T3 column (150mm x 300µm, 1.8µm Particles) with a 5µl/min flow rate. The DIA method consisted of an MS1 scan from m/z 400 to m/z 1250 (50ms accumulation time) and 40 MS2 scans (35ms accumulation time) with variable precursor isolation width covering the mass range from m/z 400 to m/z 1250.

The raw mass spectrometry data files have been deposited to Zenodo repository (http://dx.doi.org/10.5281/zenodo.1187150).

Raw mass spectrometry data files conversion

SCIEX wiff files were converted to the mzML format using MSConvert with the following settings: binary encoding precision was set to 32-bit, MS1 and MS2 vendor peak picking was used; all other options were turned off.

Acknowledgements

This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001134), the UK Medical Research Council (FC001134), and the Wellcome Trust (FC001134), and received specific funding from the BBSRC (BB/N015215/1 and BB/N015282/1).

References

1. Bruderer R, Bernhardt OM, Gandhi T, Miladinović SM, Cheng L-Y, Messner S, et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. Mol Cell Proteomics. 05/2015;14: 1400–1410.

2. Venable JD, Dong M-Q, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004;1: 39–45.

3. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Mol Cell Proteomics. 06/2012;11: O111.016717.

4. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, et al. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. Mol Cell Proteomics. 12/2017;16: 2296–2309.

5. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem. 2010;82: 833–841.

6. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26: 966–968.

7. Reiter L, Rinner O, Picotti P, Hüttenhain R, Beck M, Brusniak M-Y, et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods. 5/2011;8: 430–435.

8. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014;32: 219–223.

9. Keller A, Bader SL, Shteynberg D, Hood L, Moritz RL. Automated Validation of Results and Removal of Fragment Ion Interferences in Targeted Analysis of Data-independent Acquisition Mass Spectrometry (MS) using SWATHProphet. Mol Cell Proteomics. 05/2015;14: 1411–1418.

10. Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras A-C, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods. 2015;12: 258–264.

11. Li Y, Zhong C-Q, Xu X, Cai S, Wu X, Zhang Y, et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Methods. 2015;12: 1105–1106.

12. Wang J, Tucholska M, Knight JDR, Lambert J-P, Tate S, Larsen B, et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat Methods. 2015;12: 1106–1108.

13. Teleman J, Röst HL, Rosenberger G, Schmitt U, Malmström L, Malmström J, et al. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. Bioinformatics. 2015;31: 555–562.

14. Bilbao A, Zhang Y, Varesio E, Luban J, Strambio-De-Castillia C, Lisacek F, et al. Ranking Fragment Ions Based on Outlier Detection for Improved Label-Free Quantification in Data-Independent Acquisition LC–MS/MS. J Proteome Res. 2015;14: 4581–4593.

15. Röst HL, Liu Y, D'Agostino G, Zanella M, Navarro P, Rosenberger G, et al. TRIC: an

automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat Methods. 2016;13: 777–783.

16. Ting YS, Egertson JD, Bollinger JG, Searle BC, Payne SH, Noble WS, et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods. 2017;14: 903–908.

17. Peckner R, Myers SA, Egertson JD, Johnson RS, Abelin JG, Carr SA, et al. Specter: linear deconvolution as a new paradigm for targeted analysis of data-independent acquisition mass spectrometry proteomics. bioRxiv. 2017; 152744.

18. Zhang Y, Bilbao A, Bruderer T, Luban J, Strambio-De-Castillia C, Lisacek F, et al. The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition. J Proteome Res. 2015;14: 4359–4371.

19. Spivak M, Weston J, Bottou L, Käll L, Noble WS. Improvements to the Percolator Algorithm for Peptide Identification from Shotgun Proteomics Data Sets. J Proteome Res. 2009;8: 3737–3745.

20. Spivak M, Weston J, Tomazela D, MacCoss MJ, Noble WS. Direct Maximization of Protein Identifications from Tandem Mass Spectra. Mol Cell Proteomics. 02/2012;11: M111.012161.

21. Raczynski L, Rubel T, Zaremba K. Neural Network-Based Method for Peptide Identification in Proteomics. Information Technologies in Biomedicine. Springer Berlin Heidelberg; 2012. pp. 437–444.

22. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 3/2007;4: 207–214.

23. Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 2016;34: 1130–1136.

24. Vowinckel J, Zelezniak A, Bruderer R, Mülleder M, Reiter L, Ralser M. Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. Sci Rep. 2018;8: 4346.

25. Mülleder M, Campbell K, Matsarskaia O, Eckerstorfer F, Ralser M. Saccharomyces cerevisiae single-copy plasmids for auxotrophy compensation, multiple marker selection, and for designing metabolically cooperating communities. F1000Res. 2016;5: 2351.