Automated workflow for optimization of data independent acquisition mass spectrometry settings

Course: BINP 51, 45 credits

Student: Saghar Toresson

Email: saghar.toresson@immun.lth.se

Supervisor: Fredrik Levander

Co-supervisor: Sergio Mosquim Junior

Lund University

August 2024

Abstract

Over recent years, mass spectrometry has become essential for advancing proteomics. A typical approach in proteomics involves digesting proteins into peptides, analyzing these peptides using liquid chromatography and mass spectrometry (LC-MS), and identifying them via tandem mass spectrometry (MS/MS). In Data-Dependent Acquisition (DDA), quantification is often achieved by counting MS/MS spectra or by measuring the intensity of precursor ions at the MS1 level. On the other hand, Data-Independent Acquisition (DIA) typically quantifies peptides using the intensities of fragment ions collected at the MS/MS level.

This study introduces a computational pipeline, implemented using Snakemake, aimed at improving mass spectrometry workflows for proteomic analysis by optimizing Data-Independent Acquisition (DIA) techniques. The pipeline automates the entire process from raw data conversion, through feature extraction, to the generation of detailed analytical reports, supporting both DIA and DDA methods. By comparing different settings for DIA and DDA, we demonstrate that variable window sizes in DIA setups provide more effective protein identification, particularly enhancing detection efficiency and accuracy for low-abundance peptides for a series of test samples. This comparison highlighted that narrower isolation windows are superior in identifying more peptides compared to broader or fixed settings for these samples. These results indicate that the workflow can be used to dynamically adjust DIA settings based on the complexity and other features of samples to enhance the quality and comprehensiveness of proteomic data.

The source code is available at: https://github.com/SagharT/MasterThesis.

Introduction

Proteins, essential for nearly all biological processes, play important roles from structural support in cells to functions in metabolism, biosignaling, and immune responses. Understanding the dynamics of proteins within the proteome is critical, especially as abnormal protein function is often a key factor in disease pathologies.

Proteomics is a branch of 'omics' science, investigates the interactions, functions, structures, and cellular activities of proteins. The term 'proteome' is a combination of the words protein and genome introduced by Marc Wilkins in 1995 (1), includes all proteins that an organism produces or modifies. While proteomics provides a more intricate view of an organism's structure and function compared to genomics, it presents greater challenges. This complexity arises because protein levels vary with environmental conditions and change over time (2).

Traditional methods that focus on a few proteins are being supplemented by mass spectrometry (MS), a technique that allows for an expansive view of the proteome by measuring and identifying proteins with high accuracy. This approach is essential as mRNA abundance often fails to accurately predict protein levels, necessitating direct protein activity measurements to understand how proteomes are altered by diseases (3).

Advances in liquid chromatography (LC), coupled with MS, have transformed proteomics by enhancing the analysis of large, delicate biomolecules. The bottom-up approach, which is the most common in MS-based proteomics, involves digesting proteins into peptides, which are then analyzed to determine the sequences and structural characteristics of the parent proteins. This method relies heavily on MS/MS following LC separation, significantly improving the depth and breadth of proteomic analysis, allowing for a more comprehensive exploration of complex biological systems (4,5).

LC-MS/MS has thus become essential in primary structural characterization, not only determining the molecular weight of compounds but also providing detailed structural insights through fragmentation studies in MS/MS setups. By analyzing peptides from complete protein digests and using a protein database to characterize the peptides' origin, "bottom-up" proteomics enables detailed primary structural characterization of proteins (6,7).

The integration of LC-MS with bioinformatics has significantly expanded the capabilities of proteomics, allowing researchers to decode the complex structure of the proteome. For example, this combination enables the identification and quantification of post-translational modifications, the analysis of protein-protein interactions, and the identification of biomarkers for diseases. This powerful combination is crucial for understanding the intricate interactions of proteins in both healthy and disease states, and it plays a pivotal role in advancing precision medicine and targeted therapies, where detailed protein analysis is essential (8).

High-resolution mass spectrometry, in combination with data-dependent acquisition (DDA) and data-independent acquisition (DIA), have significantly advanced the field. DDA, traditionally used in LC-MS/MS, selects peptides for fragmentation based on their signal intensity, which can lead to favoring more common peptides while overlooking less abundant ones. This often results in incomplete data sets, particularly when analyzing complex biological samples (8,9).

On the other hand, DIA offers comprehensive analysis by systematically fragmenting all peptides within a pre-set mass-to-charge (m/z) range. This method ensures that peptides of all abundances are represented in the data, providing a more complete overview of the proteome. This is especially valuable in fields like oncology, where DIA-MS has been instrumental in molecular classification, pathway analysis, and when quantification is important, for example

in biomarker discovery, improving our understanding of cancer biology and therapeutic responses (8).

By not relying on precursor intensity for ion selection, DIA can detect peptides that might otherwise be missed due to their low abundance. Additionally, DIA's methodical fragmentation within defined m/z windows allows for a more detailed and reproducible analysis, reducing the likelihood of missing important peptides. This has made DIA particularly effective in complex proteomic studies (10,11). The comprehensive nature of DIA ensures that every peptide within the targeted m/z range is accounted for, providing a complete picture of the proteome at any given point (11,12).

A novel approach in data-independent acquisition (DIA) uses overlapping MS/MS isolation windows to capture more detailed fragment ion data, significantly improving the detection and quantification of proteins. This method, which employs computational demultiplexing, nearly doubles the precision in selecting precursor ions without affecting the mass range or scan speed, making it a valuable tool for complex proteomic studies (13,14,15,16). This strategy demonstrated a 64% increase in sensitivity and a 17% increase in detected peptides in a test involving a protein mix, proving its effectiveness (13).

Data-Dependent Acquisition (DDA) methods face limitations in fully sequencing all peptides from a specimen, requiring extended injection times (ITs) to collect sufficient ions and demanding efficient use of mass analyzer time. In contrast, Data-Independent Acquisition (DIA) methods provide thorough proteome coverage but can also encounter issues with spectra convolution and low signal intensity (17). It's important to note that low abundance ions present challenges in DIA as well, although the selection for MS/MS is not affected, which is similar to the issues faced in DDA. The optimal DIA method may vary depending on the sample type and requires careful selection of window sizes and the times needed to collect spectra. These

considerations are the key for enhancing the effectiveness of DIA and serve as a motivation for exploring various methodological optimizations in this study.

This study introduces a computational pipeline designed to optimize the analysis of proteomic data through both Data-Independent Acquisition (DIA) and Data-Dependent Acquisition (DDA) methods. The aim is to maximize the extraction of information from diverse samples, considering their abundance and complexity, which can allow for generation of optimal DIA methods for similar samples.

Method

A computational pipeline was generated with the aims to provide a systematic approach for the analysis of LC-MS/MS data, as illustrated in **Figure 1**, including various stages from raw data processing to report generation.

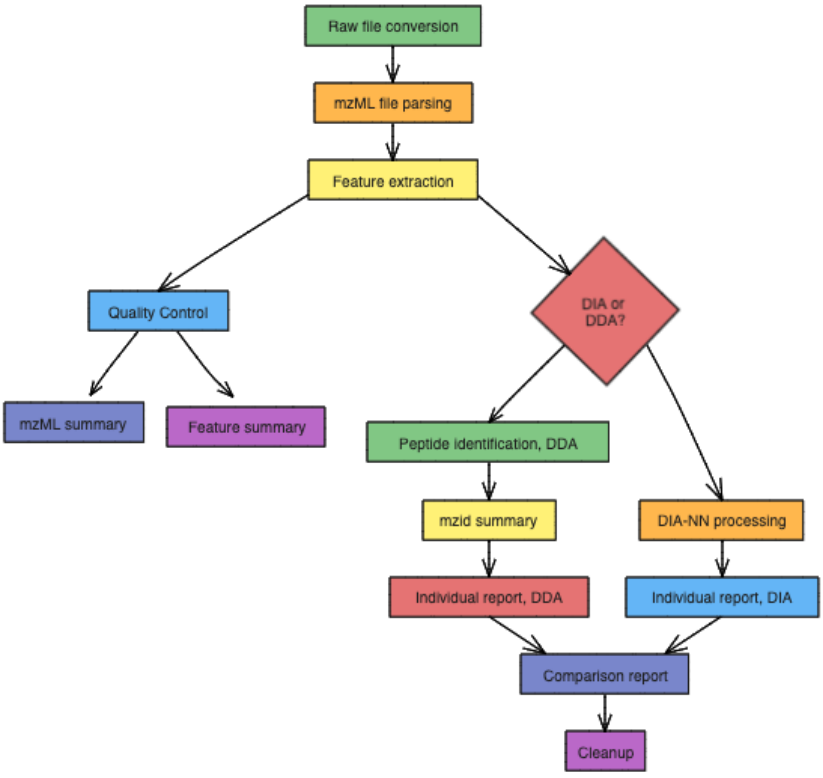

***Figure 1**: **Workflow Diagram of the Computational Pipeline.** This figure outlines the structured workflow used in the study, starting from raw file conversion and proceeding through various stages such as mzML file parsing, feature extraction, quality control, and finally report generation.*

The methodology integrates data extraction, feature extraction, quality control summary, peptide and protein identification, quantification, and the creation of both individual and comparison reports.

Data Processing Workflow

A Snakemake (18) workflow was implemented to automate the entire analysis process for both Data-Independent Acquisition (DIA) and Data-Dependent Acquisition (DDA) files. Snakemake was chosen due to its capability to manage complex workflows efficiently and reproducibly by automatically handling dependencies between various stages of the analysis. The configuration of paths and parameters is managed through a yaml file, allowing flexibility and easy adaptation to different datasets and analysis requirements. This file specifies the directories for input and output files, as well as parameters for different tools used in the pipeline.

Data Acquisition

A combined dataset of LC-MS/MS DIA and DDA data, including samples from HeLa cells (19) and breast cancer tumor biopsies, previously acquired in the laboratory, was collected.

Data acquisition had been performed previously using an Evosep One LC system (EvoSep Biosystems) connected to a Q-Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer with electrospray ionization (20). The Evosep One LC system is known for its high-throughput capabilities, making it ideal for large-scale proteomics experiments. The Q-Exactive HF-X mass spectrometer combines quadrupole and Orbitrap technologies, providing high resolution and sensitivity for peptide analysis.

Samples were separated using a 15 cm long fused silica capillary column equipped with an emitter tip and frit (dimensions: 360 µm outer diameter, 75 µm inner diameter, 50 cm length,

15 µm tip inner diameter[1]. The column was internally packed with ReproSil-Pur 1.9 µm C18 material[2].

The peptides were separated using a gradient of solvents, with 0.1% formic acid in water as Solvent A and 0.1% formic acid in acetonitrile as Solvent B. Formic acid helps in maintaining the ionization efficiency of peptides during mass spectrometry. A column oven was attached and set to 40°C to maintain a constant temperature, which improves the reproducibility and efficiency of peptide separation by reducing the viscosity of the solvents and leading to better separation and peak shapes.

A HeLa Protein Digest Standard was used in varying amounts, ranging from 0.1 ng to 50 ng, serving as a quality control (QC) for liquid chromatography separation and mass spectrometry (MS) method development. The HeLa files had been acquired using different settings over a long period of time. These HeLa samples were acquired using different settings over an extended period. For proteome analysis, the Whisper 40SPD and 20SPD methods were used, corresponding to 31-minute and 58-minute gradients, respectively, in both data-dependent acquisition (DDA) and data-independent acquisition (DIA) modes. The DIA samples included those with overlapping windows, which were subsequently demultiplexed.

Breast cancer tumor samples, specifically pooled flowthroughs after DNA and RNA extraction, underwent PAC digestion and C18 desalting. For full proteome analysis, the Whisper 20SPD method (Evosep Biosystems), corresponding to a 58-minute gradient, was utilized. Data acquisition was performed in data-independent acquisition (DIA) mode, employing methods based on taggered (also referred to as overlapping) windows (21).

[1] . Manufactured by MS Wil B. V., Aarle-Rixtel, The Netherlands
[2] . Dr. Maisch GmbH, Ammerbuch-Entringen, Germany.

For the full proteome analysis, the standard run time was set between 4 and 58 minutes. The MS1 AGC (Automatic Gain Control) target was configured at 3E6 ions with a maximum injection time of 55 ms and a resolution of 60 000, covering a scan range from 395 to 1005 m/z. DIA (Data-Independent Acquisition) MS2 spectra were typically collected with staggered windows in 75 loops, using an 8 m/z isolation window and a normalized collision energy (NCE) of 27, with the maximum injection time set to "auto" and a resolution of 15 000. However, this setup represents the standard approach; the study also explored variations including different numbers of windows and configurations with and without overlapping windows.

Data Processing

Raw data files were automatically detected and converted to mzML format (22) using the msconvert tool from the Proteowizard (23) project. This tool is run inside a docker container[3]. Within the Docker container, the msconvert tool uses a demultiplexing filter to sort out overlapping signals in the data. If demultiplexing isn't possible, the tool defaults to peak picking instead.

MzML Parser

A Python parser was developed to extract essential features from the mzML files. These features included scan ion injection times, retention times, window sizes, number of windows, and scan numbers. The extracted features are saved into CSV files, providing a dataset for further analysis.

[3] . https://hub.docker.com/r/chambm/pwiz-skyline-i-agree-to-the-vendor-licenses

Feature Extraction with Dinosaur

The Dinosaur[4] tool (v1.1.4) (24) was employed with default settings for advanced mass spectrometry MS1 peptide feature detection, suitable for both Data-Independent Acquisition (DIA) and Data-Dependent Acquisition (DDA) analyses. Developed as an improved version of the MaxQuant algorithm, Dinosaur's robust feature detection capabilities include handling overlapping spectra, assessing isotope patterns, and providing detailed quality control metrics. The tool requires input data in mzML format and is written in Java. The output from Dinosaur, which includes retention times, and scan numbers, is used for quality control.

Peptide Identification for DDA Files

For data-dependent acquisition (DDA) files, peptide identification was performed using the MS-GF+[5] tool (v2021.03.22) (25) which scores MS/MS spectra against peptides derived from a protein sequence database. MS-GF+ supports various input file formats, including mzML. This tool is highly precise and versatile, optimized for diverse spectral types, fragmentation methods, and experimental protocols. The specific settings used for MS-GF+ are detailed in Appendix A. This step produced mzid files, which were subsequently used to generate QC summaries.

Quality Control and Summary Generation

Quality control (QC) summaries were generated for both the mzML files and the extracted features file from Dinosaur. The MzMLSummary and FeaturesQCSummary java tools are used to create summary reports, ensuring data quality and consistency. These scripts were developed previously in house and are not yet published.

[4] . github.com/fickludd/dinosaur
[5] . github.com/MSGFPlus/msgfplus

Peptide Identification for DIA Files

For data-independent acquisition (DIA) files, the DIA-NN tool was utilized for peptide identification and quantification. DIA-NN (26) (v1.8.1) leverages deep neural networks and innovative quantification and signal correction strategies to enhance the processing of DIA proteomics experiments. This tool is known for its reliability, robustness, reproducibility, ease of use, and scalability. The DIA-NN tool supports various raw data formats including .mzML file. The specific settings used for DIA-NN are detailed in Appendix B.

DIA-NN can create spectral libraries from any DIA dataset, both in spectral library-based and library-free modes, by selecting the "Generate spectral library" option in the output pane. In this project, the spectral library 50ngHeLas_lib.tsv.speclib was used for running all HeLa samples. For the BC samples, the initial analysis was performed using the same HeLa spectral library. Subsequently, the BC samples were rerun by matching to a predicted in-silico generated library based on the provided fasta file using DIA-NN. The benefit of using a library created from the BC samples themselves is that it provides a more specific reference, potentially improving the accuracy and sensitivity of the protein identifications specific to the BC samples. Additionally, the FASTA file UP000005640_9606_contaminants.fasta, downloaded in August 2022, was used. This is a UniProt fasta file containing human proteome supplemented with list of contaminants after downloading (27)

Report Generation

Individual and comparison reports were generated in PDF format, summarizing key findings from feature extraction and peptide identification. Python scripts were used to compile these reports, which included detailed tables and visualizations. The individual reports provide information on DIA and DDA samples, detailing the number of distinct isolation windows, window sizes, average and median injection times, the number of identified precursors, and

scan numbers for MS1 and MS/MS. Additionally, the reports include plots for retention time vs M/Z. The comparison reports include a comparison of average MS1 and MS/MS spectra and identify precursors for all samples.

Evaluation and Validation

To validate the effectiveness of the pipeline, new data was acquired from various QC samples and biological samples. These included both DIA and DDA samples with different window sizes and, DIA samples with overlapping windows.

Results and Discussion

In this study, a Snakemake workflow was used to automate the processing of proteomic data, as shown in **Figure 1**. The entire process, from initial file conversion to final report generation, was automated. Initially, the workflow was evaluated using HeLa cell samples at varying concentrations, from minimal amounts up to 50 ng. This preliminary evaluation was important to ensure the accuracy and reliability of the workflow before proceeding to more complex analyses.

Our workflow differs from the one in the referenced study (8), which focuses mainly on detailed post-processing of data from specific instruments. Our Snakemake workflow is designed to be flexible and works on various computing platforms, making it useful for labs with different types of equipment. Unlike the referenced system that requires adjustments for each type of mass spectrometry instrument, our workflow is built to work universally, meaning it does not need custom settings for different instruments. This makes it especially useful for labs wanting to use the same process for various studies without changing the setup for each new experiment.

Detailed findings from these initial tests and subsequent experiments are presented in the following sections.

1. HeLa Samples Analysis

1.1 Data generation and documentation

During the workflow development, different HeLa sample concentrations were used (0.1 ng, 1 ng, 2.5ng, 10ng, 30 ng, and 50 ng) using both DIA and DDA methods for investigating the features of the workflow. Each analysis resulted in the generation of a detailed PDF report, named according to the corresponding raw data file. Each file includes information the number of distinct isolation windows, window size, injection times, and the number of identified precursors and scans.

1.2 Example of a Detailed Report Output

For demonstration, **Figure 2** presents a sample from one of the generated PDF reports for a DIA analysis conducted on a HeLa cell sample with a protein concentration of 50 ng. For each DIA sample four plots were generated showing the retention time vs m/z distribution for all features detected at the MS1 level using Dinosaur, features with charge ≥ 2 and identified features and identified features with charge ≥ 2 from Dia-NN analyses.

Features with a charge greater than one typically represent peptides, while singly charged features might also include contaminating compounds. This is why focusing on multiply charged features is beneficial. In Data-Dependent Acquisition (DDA) experiments, the instrument is typically configured to perform MS/MS only on peaks with a charge state of two or higher.

These scatter plots provide a visual representation of the distribution and density of detected features, highlighting the data richness achieved through DIA. The specific settings used was 17 isolation windows each with a window size of 40.5 m/z. The report includes information such as an average injection time, a median injection time, identified precursors, MS1 and MS/MS scans, providing an overview of the experimental setup and its results (**Table 1**).

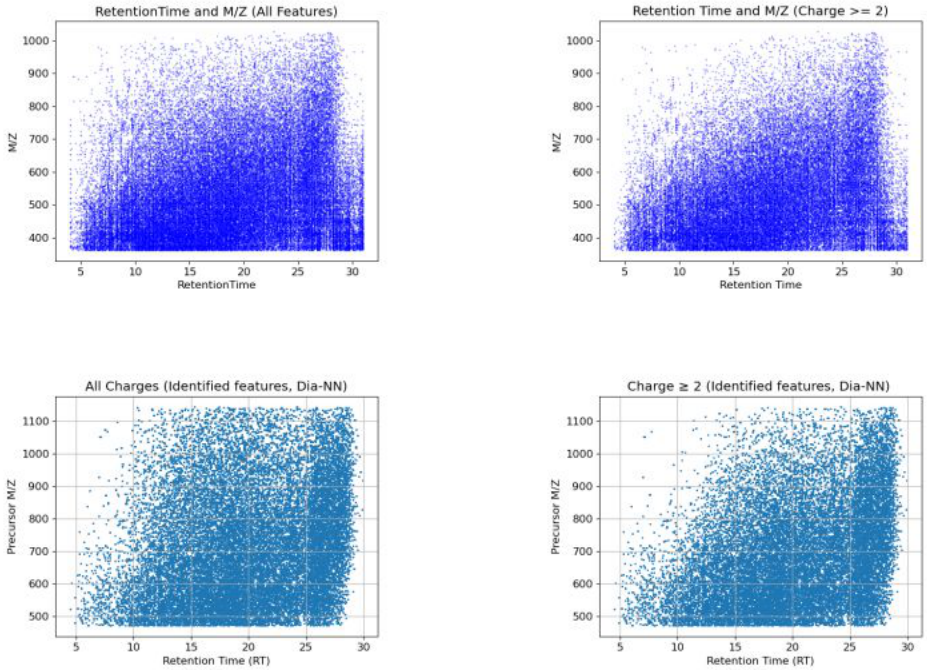

***Figure 2**: **Feature Analysis and Scan Details of DIA Sample.** Top panels: Retention time versus m/z distribution for all features (left) and for features with charge ≥ 2 (right). Bottom panels: Identified features from Dia-NN across all charges (left) and charges ≥ 2 (right).*

DIA SAMPLE	
NUMBER OF DISTINCT ISOLATION WINDOWS USED	17
DETAILED WINDOW SIZE	40.5 for 17 windows
AVERAGE INJECTION TIME	72.88 ms
MEDIAN INJECTION TIME	86.00 ms
PRECURSORS IDENTIFIED	23 930
NUMBER OF MS1 SCANS	760
NUMBER OF MS/MS SCANS	12 910

***Table 1**: Information parsed through the computational workflow for DIA sample.*

Figure 3 presents a sample from one of the generated PDF reports for a DDA analysis conducted on a HeLa cell sample with a protein concentration of 50 ng. For each DDA sample four plots were generated, showing the retention time vs m/z distribution for all features and features with charge ≥ 2 and identified features and features with charge ≥ 2 from MSGFPlus analyses. This LC-MS/MS acquisition utilized 21 078 unique isolation windows, each with a window size of 1.2 m/z, for acquisition of 21 259 MS/MS spectra. The report details key

parameters for each DDA sample including an average injection time, a median injection time, the identification of precursors, and MS1 and MS/MS scans (**Table 2**).

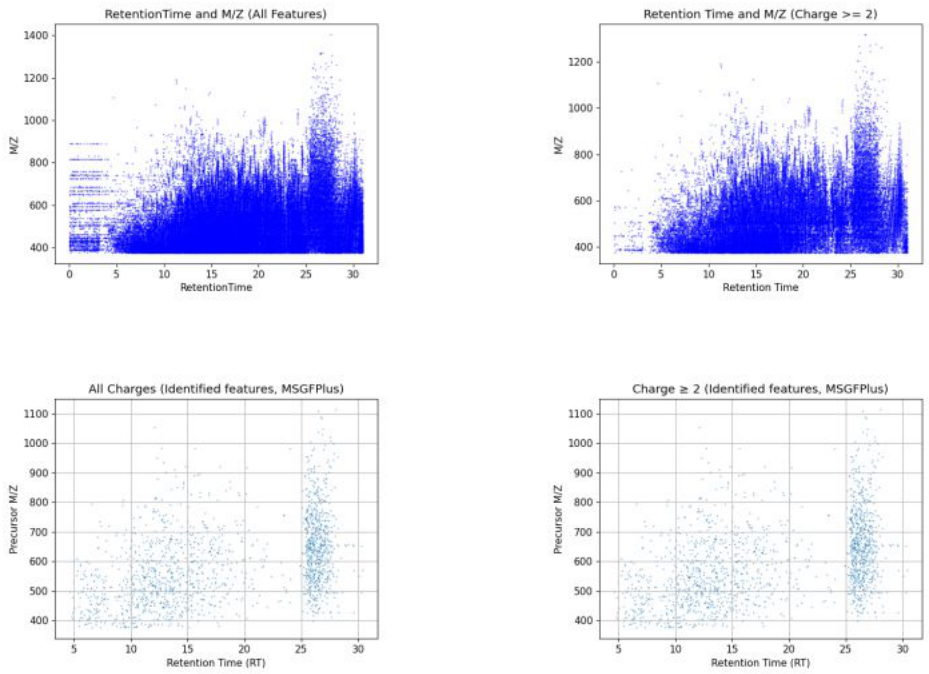

***Figure 3**: **Feature Analysis and Scan Details of DDA Sample.** Top panels: Retention time versus m/z distribution for all features (left) and for features with charge ≥ 2 (right). Bottom panels: Identified features from MSGFPlus across all charges (left) and charges ≥ 2 (right).*

DDA SAMPLE	
NUMBER OF DISTINCT ISOLATION WINDOWS USED	21 078
WINDOW SIZE	1.2
AVERAGE INJECTION TIME	29.87 ms
MEDIAN INJECTION TIME	30.00 ms
PRECURSORS IDENTIFIED	1 642
NUMBER OF MS1 SCANS	3 048
NUMBER OF MS/MS SCANS	21 259

***Table 2**: Information parsed through the computational workflow for DDA sample.*

The analysis of MS1 and MS/MS spectra across different HeLa sample concentrations (0.1ng, 30 ng and 50ng) for both DIA and DDA methods is illustrated in **Figure 4A**. In DDA samples, MS/MS scan counts increase with higher sample concentrations while in DIA samples the total number of scans (both MS1 and MS/MS) remains relatively constant regardless of the sample concentration. **Figure 4B** illustrates the number of identified precursors across various HeLa

sample concentrations for both DIA and DDA methods. The results display a clear trend where the DIA method consistently identifies more precursors than DDA.

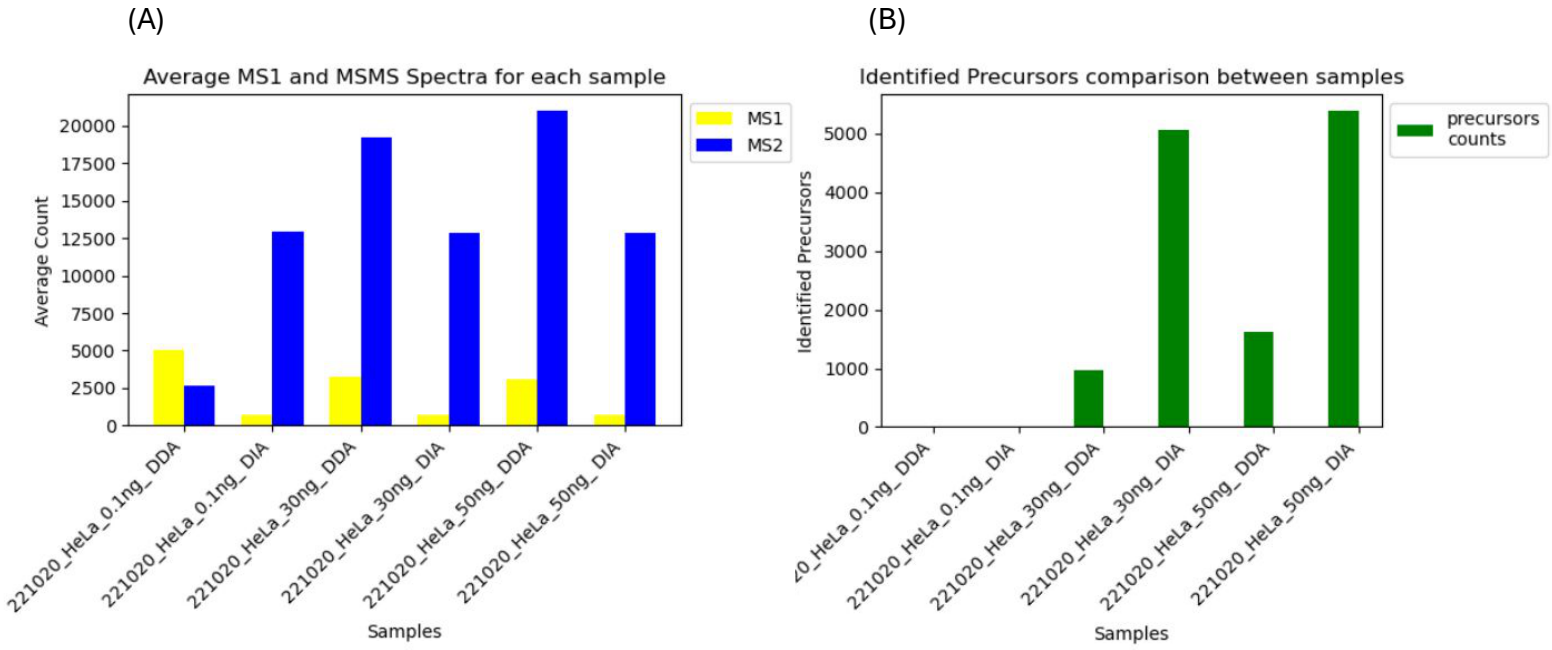

Figure 4: *(A) Average MS1 and MS/MS Spectra for each sample. Yellow bars represent counts for MS1, and blue bars represents counts for MS/MS counts. (B) Identified Precursors comparison between samples.*

2. Comparison of FASTA and Spectral Libraries

The use of an optimized spectral library enhances peptide identification in Data-Independent Acquisition (DIA) analysis. By aligning the spectral library more closely with the actual samples being analyzed, the analysis tools can more effectively match observed spectra to known peptide features. This alignment increases the likelihood of identifying lower-abundance peptides and reduces the background noise in the data. For the test sample '230227_SM_PoolPAC_12mz_44min', the implementation of the newly developed spectral library from breast cancer (BC) samples demonstrated an increase in the number of precursors identified from 21 788 precursors identified using the generic HeLa library to 52 904 with the custom BC library. Figures 5 displays the results from this comparative analysis.

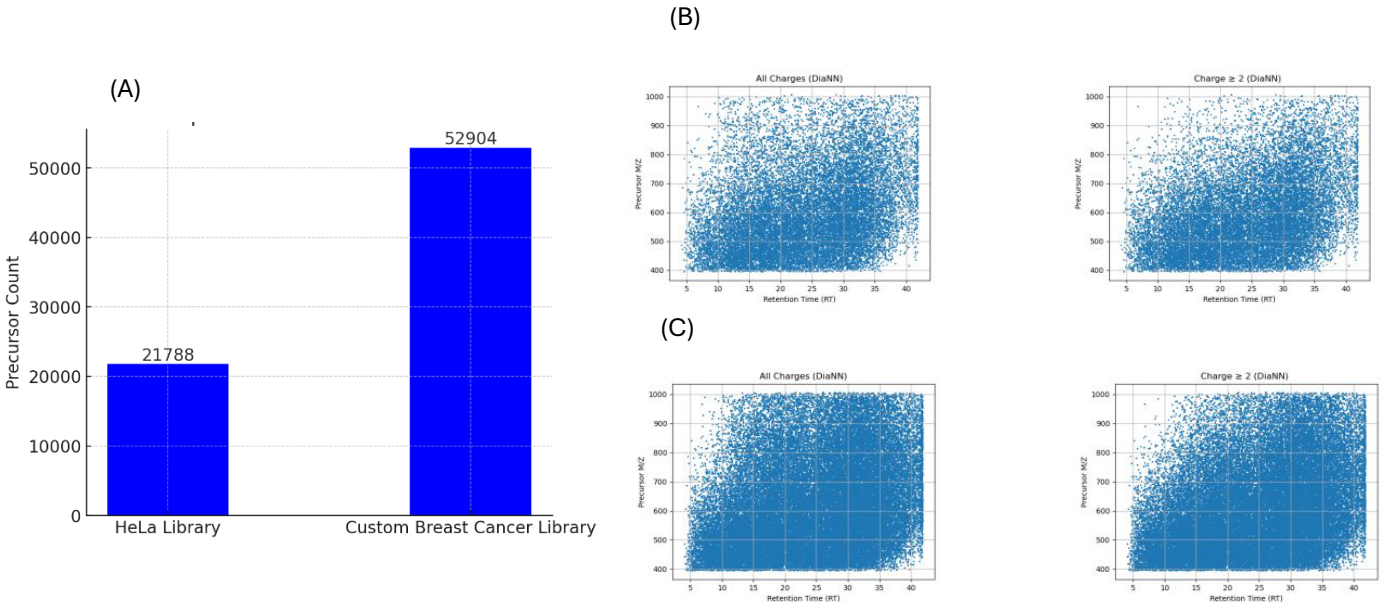

Figure 5: Comparison of precursor counts and retention time versus m/z distribution using different libraries. *(**A**) Bar plot illustrating the number of precursors identified using the HeLa library (27 022) and the custom Breast Cancer library (28 093). (**B**) Scatter plot showing the retention time and m/z distribution for the sample analyzed with the HeLa library. (**C**) Scatter plot showing the retention time and m/z distribution for the sample analyzed with the BC library.*

3. Optimization and Comparison of DIA and DDA Settings

The analysis was conducted using 50 ng Hela samples with various settings for both DIA and DDA, with the aim to optimize parameters for maximizing the identification and quantification of precursors. We compared the performance of various configurations by focusing on the number of distinct isolation windows, window sizes, and the number of identified precursors and scans.

DIA with 12 m/z Windows

This DIA sample used 50 overlapping isolation windows with a window size of 12.0 m/z, followed by computational demultiplexing. This configuration identified 22 347 precursors, with an average injection time of 23.89 ms and a median injection time of 23.00 ms. The number of MS1 scans was 773, and the number of MS/MS scans was 77 256.

DIA with Variable Windows (12 m/z and 24 m/z)

This DIA sample used 38 distinct isolation windows, with 25 windows at 12.0 m/z and 13 windows at 24.0 m/z. This setup identified 19 439 precursors, with an average injection time of 23.38 ms and a median injection time of 23.00 ms. The number of MS1 scans was 1 029, and the number of MS/MS scans was 39 095.

DIA with 24 m/z Windows

This DIA sample used 25 overlapping isolation windows with a window size of 24.0 m/z, followed by computational demultiplexing. This configuration identified 18 617 precursors, with an average injection time of 23.49 ms and a median injection time of 23.00 ms. The number of MS1 scans was 1 483, and the number of MS/MS scans was 74 146.

DDA Sample Analysis

The DDA sample utilized 22 467 distinct isolation windows with a fixed window size of 1.20. This setup identified 7 942 precursors, with an average injection time of 28.93 ms and a median injection time of 30.00 ms. The number of MS1 scans was 2 899, and the number of MS/MS scans was 22 627.

The bar charts in the **figure 6** show the comparison of MS1 and MS/MS scan counts across various DIA window configurations and a DDA setting for analyzing a 50-ng sample of breast cancer cells. The data compares DDA, DIA with fixed 12 m/z and 24 m/z windows, and DIA with variable m/z windows to show how different isolation window sizes influence the number of scans achievable.

DDA results in a higher count of MS1 scans due to its selective nature of precursor ion scanning. In DDA, each MS1 scan is used to select precursors for subsequent MS/MS

fragmentation based on their intensity, leading to a more frequent occurrence of MS1 scans as it cycles through repeatedly scanning the precursor ions. This results in relatively fewer MS/MS scans since only selected precursors are fragmented. DIA configurations with fewer, broader isolation windows resulted in a higher count of MS/MS scans while reducing the frequency of MS1 scans.

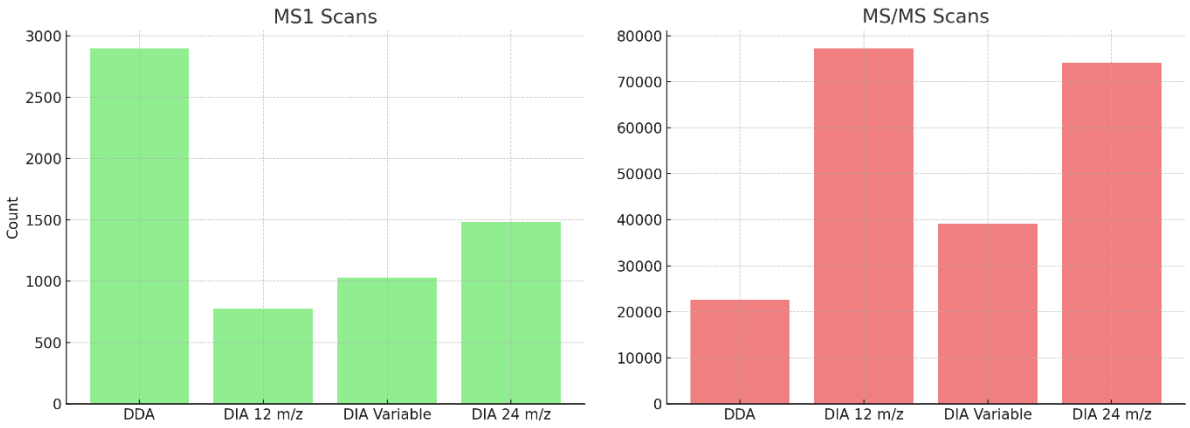

Figure 6: Comparative Analysis of MS1 and MS/MS Scans Across DIA and DDA Configurations. *The effect of DIA window configurations and a DDA approach on the number of MS1 and MS/MS scans performed in the analysis of breast cancer samples. The green bars represent the MS1 scans which provide the initial mass spectrometry data, while the red bars depict the MS/MS scans necessary for detailed peptide identification.*

Figure 7 presents the analysis of precursor identification and retention time versus m/z ratios across different DIA and DDA configurations. The bar chart on the left illustrates the number of precursors identified under different window settings. Notably, the DIA method with 12 m/z windows records the highest number of identified precursors, followed by variable windows and 24 m/z windows, indicating an increase in precursor detection with broader isolation windows.

The scatter plots on the right detail the distribution of identified features across retention time and m/z ratios for the different settings. The top right plot for the 24 m/z window setting shows a dense distribution of features.

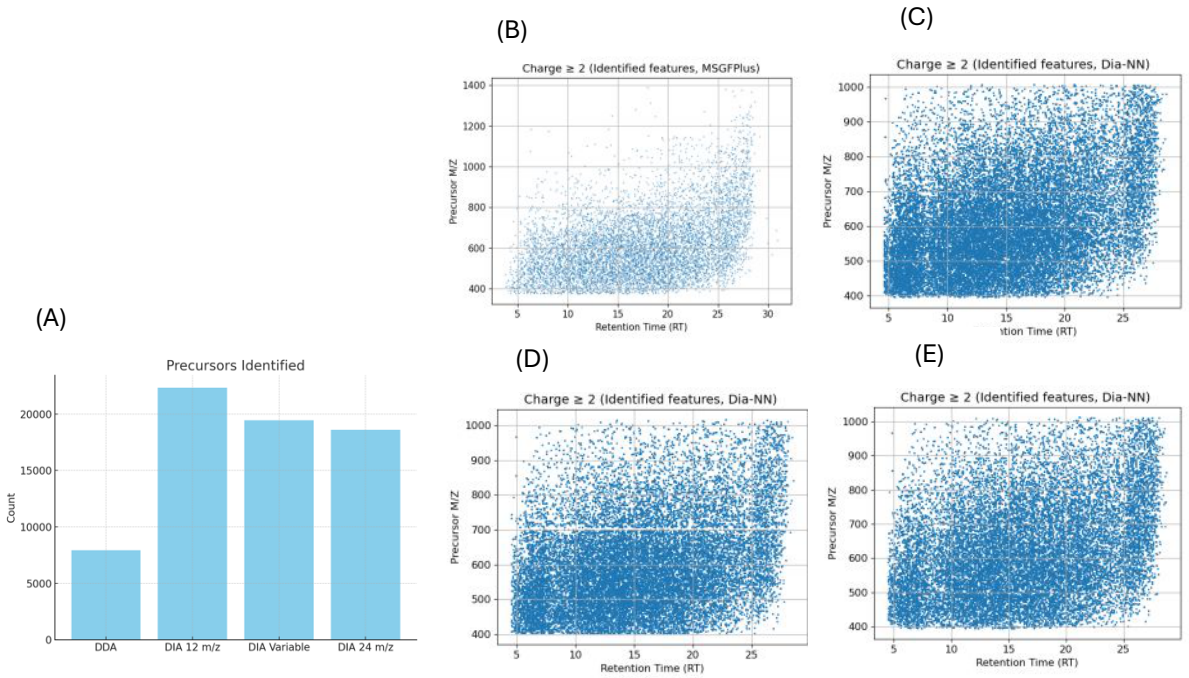

Figure 7: Analysis of precursor identification and feature distribution in DIA and DDA settings. *(**A**) Number of precursors identified across different isolation window configurations, DIA settings with variable, 12 m/z, and 24 m/z windows are compared alongside a standard DDA setup.* ***(B-E)*** *Scatter plots showing the distribution of detected features across retention time and m/z ratios for different window configurations. (**B**) With DDA setting, (**C**) DIA with 12 m/z windows, and (**D**) DIA with variable windows, and (**E**) DIA with 24 m/z windows. The plots emphasize how window settings influence the comprehensiveness of proteomic sampling and the resolution of detected features.*

In contrast, the variable window setting (top left) displays a more dispersed pattern, indicating a varied range of feature detection, likely due to the combined use of narrow and broad windows. The bottom right plot for the 12 m/z setting shows tighter clustering and the bottom left plot represents the DDA setting, highlighting its selective nature in feature detection based on precursor abundance.

The use of variable isolation window sizes in data-independent acquisition (DIA) has proven to enhance the efficiency and sensitivity of proteomic analyses, as demonstrated by our findings and supported by the study by Smith et al. (28). Smith and colleagues show that narrower windows, such as those as small as 5 m/z, improve the detection of peptides by reducing co-elution and background noise, which is essential in complex samples like tumor biopsies.

Our experiments show that a DIA setup with 12 m/z windows clearly outperforms broader or fixed window configurations in precursor identification for the analysed sample. This underscores the advantages of dynamically adjusting window sizes based on the peptide complexity observed in real-time spectral data. By customizing window sizes to the specific needs of each mass spectrum region, the mass spectrometer can use its resources more efficiently, enhancing throughput and the detection of low-abundance peptides.

Conclusions

The integration of variable window sizes, particularly through adaptive methods that respond to live data, could greatly simplify proteomic workflows. Predictive models that refine window settings after initial broad-window scans promise to reduce method development time and improve data quality, aligning with the needs of high-throughput proteomic studies. Moving forwar, by analyzing results from this workflow, we can predict which settings will work best for specific data sets. This predictive capability allows for the experimentation with different window sizes and settings to determine the most effective configuration for given data. A future development could involve automating this process, where the workflow not only suggests but also implements the optimal settings for maximal data quality and throughput.

Acknowledgments

I want to thank my supervisor, Fredrik Levander, and my co-supervisor, Sergio Mosquim Junior, for their invaluable support, guidance and time throughout my thesis. I also thank Parisa Esmaeili and Måns Zamore for providing the data needed for my study, and everyone at the Department of Immunotechnology at Lund University for making it a great place to work.

References

1. Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., & Williams, K. L. (1996). Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews, 13*(1). https://doi.org/10.1080/02648725.1996.10647923

2. Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., & Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, 12(5), 57-69. https://doi.org/10.4331/wjbc.v12.i5.57

3. Chen, C., Hou, J., Tanner, J. J., & Cheng, J. (2020). Bioinformatics methods for mass spectrometry-based proteomics data analysis. *International Journal of Molecular Sciences, 21*(8), 2873. https://doi.org/10.3390/ijms21082873

4. Chen, G., Pramanik, B. N., Liu, Y.-H., & Mirza, U. A. (2007). Applications of LC/MS in structure identifications of small molecules and proteins in drug discovery. *Journal of Mass Spectrometry*, 22(3), 442. https://doi.org/10.1002/jms.1184

5. Cho, W. C. S. (2007). Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics*, 5(2), 77–85. https://doi.org/10.1016/S1672-0229(07)60018-7

6. Karr, U., Simonian, M., & Whitelegge, J. P. (2017). Integral membrane proteins: Bottom-up, top-down and structural proteomics. *Expert Review of Proteomics*, 14(8), 715-723. https://doi.org/10.1080/14789450.2017.1359545

7. Dupree, E. J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B. A., & Darie, C. C. (2020). A critical review of bottom-up proteomics: The good, the bad, and the future of this field. *Proteomes*, 8(3), 14. https://doi.org/10.3390/proteomes8030014

8. Wallmann, G., Leduc, A. and Slavov, N. (2023). Data-Driven Optimization of DIA Mass Spectrometry by DO-MS. *Journal of Proteome Research*, 22(8), pp. 3149-3158. https://doi.org/10.1101/2023.02.02.526809

9. Gillet, L. C., Leitner, A., & Aebersold, R. (2016). Mass spectrometry applied to bottom-up proteomics: Entering the high-throughput era for hypothesis testing. *Annual Review of Analytical Chemistry*, 9, 449-472. https://doi.org/10.1146/annurev-anchem-071015-041535

10. Creative Proteomics. (n.d.). About us. Retrieved 13 june, 2024, from https://www.creative-proteomics.com/ngpro/resource-dia-vs-dda-mass-spectrometry-a-comprehensive-comparison.html

11. Doerr, A. (2015). DIA mass spectrometry. *Nature Methods*, 12, 35. https://www.nature.com/articles/nmeth.3234

12. Xin, L., Qiao, R., Chen, X. *et al.* A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nat Commun* **13**, 3108 (2022). https://doi.org/10.1038/s41467-022-30867-7

13. Amodei, D., Egertson, J., MacLean, B. X., Johnson, R., Merrihew, G. E., Keller, A., Marsh, D., Vitek, O., Mallick, P., & MacCoss, M. J. (2019). Improving precursor selectivity in data-independent acquisition using overlapping windows. *Journal of the American Society for Mass Spectrometry*, 30(4), 669-684. https://doi.org/10.1007/s13361-018-2122-8

14. Moseley, M. A., Hughes, C. J., Juvvadi, P. R., Soderblom, E. J., Lennon, S., Perkins, S. R., Thompson, J. W., Steinbach, W. J., Geromanos, S. J., Wildgoose, J., Langridge, J. I., Richardson, K., & Vissers, J. P. C. (2017). Scanning quadrupole data-independent acquisition, Part A: Qualitative and quantitative characterization. *Journal of Proteome Research*, 17(2), 770-779. https://doi.org/10.1021/acs.jproteome.7b00464

15. Williams, B. J., Ciavarini, S. J., Devlin, C., Cohn, S. M., Xie, R., Vissers, J. P. C., Martin, L. B., Caswell, A., Langridge, J. I., & Geromanos, S. J. (2016). Multi-mode acquisition (MMA): An MS/MS acquisition strategy for maximizing selectivity, specificity, and sensitivity of DIA product ion spectra. *Proteomics*, 16, https://doi.org/10.1002/pmic.201500492

16. Heaven, M. R., Cobbs, A. L., Nei, Y.-W., Gutierrez, D. B., Herren, A. W., Gunawardena, H. P., Caprioli, R. M., & Norris, J. L. (2018). Micro-Data-Independent Acquisition for High-Throughput Proteomics and Sensitive Peptide Mass Spectrum Identification. *Analytical Chemistry*, 90(15), 8905-8911. https://doi.org/10.1021/acs.analchem.8b01026

17. Petrosius, V. et al. (2022). Enhancing single-cell proteomics through tailored DataIndependent Acquisition and micropillar array-based chromatography. 2022.11.29.518366 Preprint at https://doi.org/10.1101/2022.11.29.518366

18. Köster, J., & Rahmann, S. (2012). Snakemake - A scalable bioinformatics workflow engine. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/bts480

19. Thermo Fisher Scientific. (n.d.). Pierce™ HeLa Protein Digest Standard (Catalog number: 88328). Retrieved June 15, 2024, from https://www.thermofisher.com/order/catalog/product/88328

20. Thermo Fisher Scientific. (n.d.). Evolution™ 350 UV-Vis Spectrophotometer. Retrieved May 06, 2024, from https://www.thermofisher.com/order/catalog/product/912A0959

21. Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villén, J., & MacCoss, M. J. (2018). Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, 9, Article 5128. https://doi.org/10.1038/s41467-018-07454-w

22. Deutsch, E. W. (2010). Mass spectrometer output file format mzML. *Methods in Molecular Biology*, 604, 319-331. https://doi.org/10.1007/978-1-60761-444-9_22

23. Chambers, M. C., MacLean, B., & Burke, R. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30, 918-920. https://doi.org/10.1038/nbt.2377

24. Teleman, J., Chawade, A., Sandin, M., Levander, F., & Malmström, J. (2016). Dinosaur: A refined open-source peptide MS feature detector. *Journal of Proteome Research*, 15(7), 2143-2151. https://doi.org/10.1021/acs.jproteome.6b00016

25. Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5, Article 5277. https://doi.org/10.1038/ncomms6277

26. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2019). DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17, 41–44. https://doi.org/10.1038/s41592-019-0638-x

27. Frankenfield, A. M., Ni, J., Ahmed, M., & Hao, L. (2022). Protein contaminants matter: building universal protein contaminant libraries for DDA and DIA proteomics. *Journal of Proteome Research*, 21(9). https://doi.org/10.1021/acs.jproteome.2c00145

28. Li, W., Chi, H., Salovska, B., Wu, C., Sun, L., Rosenberger, G., & Liu, Y. (2019). Assessing the relationship between mass window width and retention time scheduling on protein coverage for data-independent acquisition. *Journal of The American Society for Mass Spectrometry, 30*, 1396–1405. https://doi.org/10.1007/s13361-019-02243-1

Appendix A. MS-GF+ Settings

The following settings were used for MS-GF+:

PrecursorMassTolerance: 10 ppm: Specifies the mass tolerance for precursor ions.

Target Decoy Strategy: 1: Indicates the use of a target-decoy search strategy, with 1 meaning the search includes both forward and reverse proteins.

InstrumentID: 3: Specifies the instrument type as Q-Exactive (ID 3).

Number of concurrent threads to be executed: 10: Sets the number of concurrent threads to 10 for parallel processing.

Maximum number of dynamic (variable) modifications per peptide: 3: Limits the number of variable modifications per peptide to 3.

NumMods = 3

01, M, opt, any, Oxidation

Appendix B. DIA-NN Settings

The following settings were used for DIA-NN, as referenced from the DIA-NN GitHub command line documentation:

--qvalue: 0.01

--met-excision: Enables protein N-terminal methionine excision as a variable modification for the in-silico digest.

--cut K*, R*: Specifies cleavage specificity for the in-silico digest at lysine (K) and arginine (R) residues.

--relaxed-prot-inf: Instructs DIA-NN to use a heuristic protein inference algorithm, ensuring no protein is present simultaneously in multiple protein groups. This mode is recommended for method optimization, benchmarks, gene set enrichment analysis, and related downstream processing. Note that the alternative protein inference strategy of DIA-NN is more reliable for differential expression analyses.

--smart-profiling: Enables an intelligent algorithm that determines how to extract spectra when creating a spectral library from DIA data. This is highly recommended and should almost always be enabled.

--peak-center: Instructs DIA-NN to integrate chromatographic peaks only in the vicinity of the apex, equivalent to the "Robust LC" quantification mode.

--no-ifs-removal: Turns off interference subtraction from fragment ion chromatograms, equivalent to the "high precision" quantification mode.

--gen-spec-lib: Instructs DIA-NN to generate a spectral library.