Feature Detection for the Hidden Proteome

Danielle A. Faivre

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Michael J. MacCoss, Chair
Matthew F. Bush
Elizabeth M. Blue

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Feature Detection for the Hidden Proteome

Danielle A. Faivre

Chair of the Supervisory Committee:

Michael J. MacCoss

Department of Genome Sciences

Proteomics primarily focuses on identifying and quantifying peptides using database searches. However, there are instances where certain features of interest, such as unexpected byproducts, abnormal cleavages, unknown modifications, sequence variants, and glycans, remain unidentified. These unidentified peptide-like features can be collectively referred to as the "Hidden Proteome." Unfortunately, these hidden features are often overlooked in quality control techniques, discovery experiments, and targeted analyses. This research aims to address this gap and complement traditional proteomics approaches. To investigate the "Hidden Proteome" and enhance data quality assessment, the Python package *msions* is introduced. This package allows users to assess the extent to which their signal falls into the "Hidden Proteome" and identify issues in the mass spectrometry (MS) data.

Functionality of *msions* was also used to study the mechanism by which high-field asymmetric waveform ion mobility spectrometry (FAIMS) improves MS results. Ion mobility approaches can provide valuable insights into the "Hidden Proteome" since the partially

orthogonal separation aids in sample characterization. Further characterization can be performed using a workflow developed to detect persistent MS1 features and enable users to examine unidentified features alongside features with assigned peptide identities. These "Hidden Proteome" features play a crucial role in Multi-Attribute Method (MAM) for quality control of biopharmaceuticals and other new feature detection approaches, such as the analysis of proteoforms, protein-protein interactions, metabolic responses to stimuli, and impurities. Overall, these projects offer valuable packages, workflows, and tools that can greatly benefit researchers in the field of proteomics. With the integration of *msions* into Limelight[1] and the incorporation of the MAM project into Skyline,[2] these endeavors are expected to continue to grow and reach a wider audience in the future.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I have been lucky to receive an outstanding amount of support during graduate school. First, thank you to my advisor, Mike MacCoss, for providing me the opportunity to improve my mass spectrometry and separation knowledge, learn computational skills, and investigate potential careers. I would also like to thank my supervisory committee, Liz Blue, Judit Villén, Leo Pallanck, and Matt Bush. You asked me what I wanted to learn in graduate school besides research training and forced me to be mindful from the very beginning.

I would like to thank the members of the MacCoss lab for being my research family during my PhD. I enjoyed brainstorming and troubleshooting with all of you. A special thanks to Rich Johnson for training me while I was rotating through the lab. DIADA and DIADDA made us laugh but also convinced me that I was in the right place. I would like to thank Lindsay Pino for being a great mentor and role model. Thank you to Deanna, Lilian, and Kristine for the emotional support and small favors throughout our training. I would also like to thank Genn Merrihew for doing everything for the MacCoss lab.

I have grown a large network of friends while I have been in Seattle. My very first set of friends were from Pokémon Go, and I would like to thank them for calling me "Dr. Dani" for years because it motivated me to live up to that expectation. I would like to thank my cohort for the Zoom happy hours during the pandemic. Thanks to my Oula dance fitness community for keeping me healthy and sane. Thank you to my climbing community for pushing me outside of my comfort zone and showing me that persistence pays off. Thank you to my Notre Dame friends for your unwavering support during undergrad and afterwards.

I would also like to thank the people that helped grow my love for science. Thank you to my previous research advisor, Norm Dovichi. While I was at Notre Dame, you welcomed

me into the lab, supported my growth as a scientist for 3 years, and convinced me that Genome Sciences was where I needed to go next. You were right. I would like to thank Bill Boggess for his support and advice at Notre Dame and since then. I always think fondly of my time in the Mass Spectrometry and Proteomics Facility. Thank you to Nicole Schiavone and Jenn Arceo for training me in the Dovichi lab and providing support when I became more independent. I would also like to thank Greg Côté and Suzanne Unser for giving me my very first research experience at the USDA National Center for Agricultural Utilization Research.

Finally, thank you to my family for supporting me in all areas of my life. Thank you to my parents, Frank and Marla, for shaping me into the person I am today. Thank you to my sister, Francesca, for being my first student. Thank you to my parents and grandparents for the gift of Notre Dame. Lastly, thank you to Louie for being my rock through graduate school and for reminding me to have a good work-life balance. You deserve an honorary doctorate.

DEDICATION

To my family and closest friends

"We do this not because it is easy,
but because we thought it would be easy."
- Garry Shutler

Chapter 1
INTRODUCTION

1.1 Mass spectrometry for proteomics

Mass spectrometry is a useful technique for measuring biological samples. The technique requires ionization of the sample and separates the ions with an electric or magnetic field based on their mass-to-charge ratio (m/z). In 1897, J. J. Thomson was the first person to measure charge-to-mass while he was trying to investigate cathode rays.[3] He later built the first mass spectrometer,[4] which generated ions with gas discharge tubes, passed them through electric and magnetic fields, and detected them on a photographic plate. In modern instruments, the ions are detected with a mass analyzer, which generates a mass spectrum.

Mass spectrometry is an important tool for studying proteomics. The term *proteomics* came from merging *protein* and *genomics* in the 1990s.[5–7] Proteomics is the global analysis of proteins and their functions. Proteins are biomolecules containing one or more chains of amino acids. The genome determines the sequence of amino acids, and the sequence determines the structure and function of the protein. Proteins are vital for many biological processes, such as tissue structure, biochemical reactions, gene expression, and cell signaling. Mass spectrometry allows the identification and quantification of thousands of proteins in a single experiment.[8]

In addition to identifying and quantifying proteins, proteomics also covers cellular localization, protein-protein interactions, post-translational modifications (PTMs), and turnover.[9] By examining these areas, proteomics can be used to characterize the structure and function of proteins and improve our understanding of fundamental biological processes. It can also be applied in medicine and industry and used to diagnose and monitor disease, identify new

drug targets, and develop personalized medicine.

1.1.1 Bottom-up proteomics

Bottom-up proteomics involves identifying and quantifying the proteins in a sample by digesting proteins into peptides and analyzing them by mass spectrometry.[10] When the bottom-up approach is applied to a mixture of proteins, it is called shotgun proteomics.[11–13] A typical workflow includes isolating a protein mixture from a biological sample, quantifying the protein concentration, digesting the mixture, measuring the peptides by mass spectrometry, and identifying the peptides and proteins with a database search.

1.1.2 Electrospray ionization

Before samples can be analyzed in a mass spectrometer, they must be ionized. Electrospray ionization (ESI) is a widely used method for generating ions from a sample.[14,15] The ESI process involves applying a high voltage to a liquid sample to create a charged aerosol of droplets. Gas-phase ions are generated when the solvent evaporates from the droplets and the charge becomes concentrated. The charged ions enter the mass spectrometer, are separated by their m/z, and detected by the mass analyzer.

One advantage of ESI is the ability to ionize large biomolecules without significant fragmentation.[16] This ability allows the detection of intact molecules and the analysis of PTMs and other structural features. In addition to proteomics, ESI is also used for metabolomics and lipidomics, and the method is commonly used in environmental and forensic analysis and in drug discovery and development.

1.1.3 Reverse-phase liquid chromatography

High-performance liquid chromatography (HPLC) is an analytical technique used for separation of compounds in complex mixtures prior to identification and quantification. HPLC uses a high-pressure pump to force a sample through a column packed with a stationary

phase.[17] The stationary phase separates the sample based on the molecules' chemical and physical properties and the properties of the stationary phase. Compounds will interact differently with the stationary phase, which will result in varying retention times and elution profiles. HPLC is commonly used in pharmaceuticals, environmental analysis, food and beverage analysis, and forensics.

Reverse-phase liquid chromatography (RPLC) is a type of high-performance liquid chromatography that is widely used for separating peptides and proteins based on their hydrophobicity.[18] Molecules with higher hydrophobicity will have stronger interactions with the hydrophobic stationary phase and elute later in the gradient than the more hydrophilic molecules. In RPLC, the stationary phase is composed of hydrophobic beads that are normally made of modified silica or polymers. Silica is naturally hydrophilic, so alkyl chains (e.g., C18) are bonded to the beads to create the hydrophobic stationary phase. The sample is added to the packed column and eluted with a mobile phase solution. The mobile phase contains an aqueous buffer and an organic solvent, which is typically acetonitrile or methanol. The organic solvent is gradually increased over a gradient to elute the molecules from the column based on their increasing hydrophobicity.

The main advantage of RPLC is the ability to separate a large range of molecules with high resolution and sensitivity. The technique is useful for separating and analyzing complex mixtures of peptides and proteins that can be found in biological samples. RPLC is applied in many fields, such as proteomics, metabolomics, drug discovery, and environmental and forensic analysis.

1.2 Improving results by reducing data complexity

Sample complexity is a major challenge for protein analysis. The protein abundances in a mammalian cell can range from a few hundred copies to tens of millions of copies.[19] Instrumentation improvements and separation techniques are two of the ways that the dynamic range challenge has been addressed.

1.2.1 Automatic gain control

Automatic gain control (AGC), first developed by the Finnigan Corporation (now Thermo Fisher Scientific),[20,21] is an example of an instrumentation improvement to help dynamic range and sensitivity. AGC adjusts the number of ions accumulated before detection by filling for a calculated amount of time. The AGC is determined by monitoring the ion signal and adjusting the ion trapping parameters to maintain an appropriate signal intensity.

In mass spectrometry, AGC works as a feedback loop. As ions are injected into the trap, the ion current is measured, and the trapping voltage is altered to maintain a consistent signal. This adjustment ensures that the number of ions in the trap is in a target range even if the ion flux varies over time. The injection time is the period of time when the ions can pass into the trap. The MS or MS/MS events are controlled by either the AGC target value or the maximum ion injection time, whichever is reached first.[22]

AGC is important in high-resolution mass spectrometry because the signal-to-noise ratio is crucial for accurate measurements. AGC can improve the dynamic range and increase sensitivity because adjusting the ion accumulation can increase the intensity of lower abundance species and prevent the accumulation from exceeding a threshold that could cause space charge effects.[23] AGC can be configured differently depending on the instrument and application. It could be set to a specific target value or allowed to vary within a certain range. On some instruments, the AGC parameters can be adjusted to optimize performance based on the specific type of sample. High AGC target values can have some mass deviation on precursor ions' mass accuracy due to space charge effects, which results in a decrease in the percentage of identified MS/MS spectra.[24]

1.2.2 Quadrupole gas-phase fractionation

Quadrupole gas-phase fractionation (quadrupole GPF) is used in mass spectrometry-based proteomics to reduce sample complexity and increase the depth of proteome coverage. Quadrupole GPF uses a quadrupole mass filter to isolate and selectively fragment peptides of a specific

m/z range for further analysis.[25,26] The quadrupole mass filter consists of four parallel rods with an oscillating electric field. After the ions are filtered by the electric field, they can be passed to the detector or fragmented using collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD). After fragmentation, the new ions are analyzed by tandem mass spectrometry (MS/MS).

Quadrupole GPF has advantages over other gas analysis techniques because it can analyze complex mixtures quickly and accurately, has high sensitivity, and is able to detect trace amounts of impurities. In proteomics, quadrupole GPF can allow the analysis of less abundant peptides that would otherwise be obscured by more abundant peptides. However, its limitations are that it requires careful calibration and that ions with similar m/z's can cause interference. To further increase the proteome coverage depth, quadrupole GPF can be combined with other fractionation techniques, such as chromatography methods.

1.2.3 High-field asymmetric waveform ion mobility spectrometry

High-field asymmetric waveform ion mobility spectrometry (FAIMS) is another option for reducing sample complexity and increasing the depth of proteome coverage. FAIMS was originally used to measure peptides by Guevremont and coworkers[27–29] and was later used with a biological sample by Venne *et al.*[30] FAIMS separates complex mixtures of ions based on their mobility in a high-frequency electric field. The FAIMS device generates asymmetric waveforms with two electrodes, and the ions are separated based on their difference in mobility as they travel between the electrodes (Figure 1.1). The differential mobility is affected by the collision cross-section of ions and their interaction with the carrier gas.[31] The asymmetry of the electric field and the differential ion mobility cause ions to acquire a net displacement perpendicular to their direction of motion. A DC offset (compensation voltage or CV) can be applied to one electrode to counter this displacement and enable the transmission of ions.

FAIMS can be used with mass spectrometry to enhance the detection of low-abundance compounds and reduce interference from co-eluting compounds. The advantages of FAIMS over other ion mobility techniques include selectively filtering out interfering ions, its high

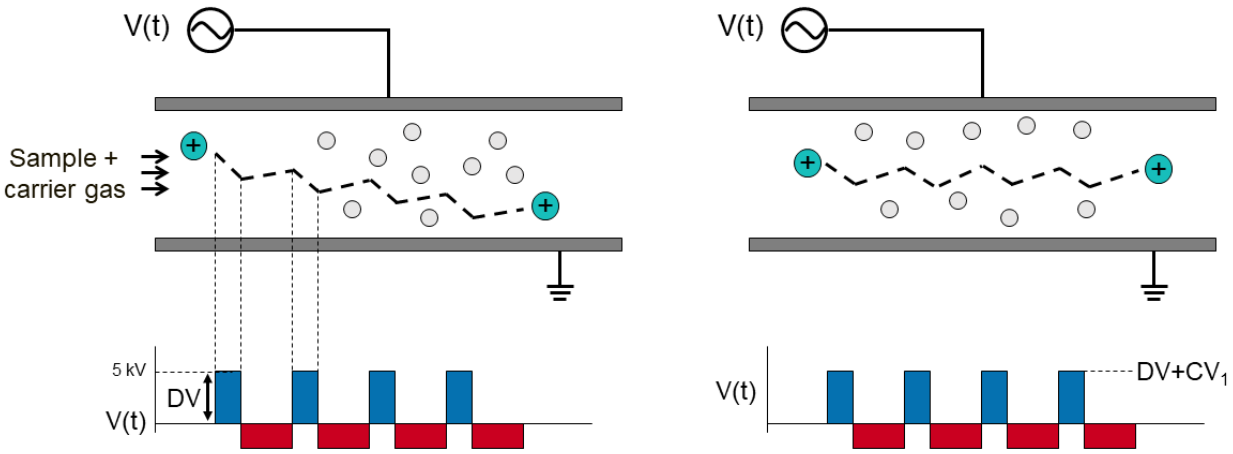


Figure 1.1: Schematic of high-field asymmetric waveform ion mobility spectrometry (FAIMS).

sensitivity, and its compatibility with a range of mass spectrometer instruments. However, FAIMS also has limitations, such as the potential for ion loss and an additional need for calibration and optimization.

1.3 Mass spectrometry acquisition methods

Mass spectrometry acquisition methods are the different techniques used to acquire mass spectrometry data. Choosing a method can depend on the instrumentation available, the sample availability, the expertise of the researchers, and the research question.

1.3.1 Data-dependent acquisition

Data-dependent acquisition (DDA) is a commonly used acquisition method for mass spectrometry.[32,33] The method is typically used for peptide identification in shotgun proteomics. In DDA, the mass spectrometer selects the most intense ions for fragmentation and analysis. DDA works by performing a survey scan of the ions, selecting the 10-20 most intense ions, fragmenting the ions, and analyzing them in another MS scan (Figure 1.2). The analysis by a second scan (MS2) is also known as tandem mass spectrometry (MS/MS).

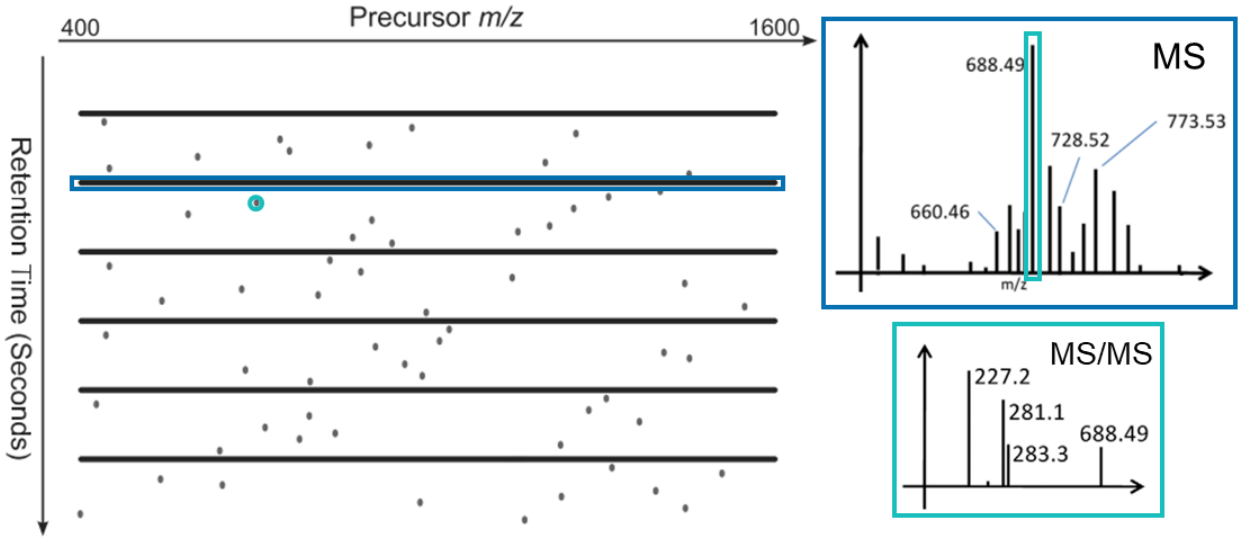


Figure 1.2: Schematic of data-dependent acquisition (DDA).

During the MS/MS analysis, the selected ions are fragmented using collision-induced dissociation (CID)[34] or higher-energy collisional dissociation (HCD).[35] These peptide fragments are known as b- and y-ions, and they are analyzed to determine the amino acid sequence of the peptide. DDA enables the identification of a large number of peptides, but it has limitations. The method is not appropriate for quantification of peptides or proteins because of the stochasticity of the method. The stochasticity can cause the number of MS/MS spectra acquired for each peptide to vary.

1.3.2 Targeted acquisition

Targeted acquisition is used for the quantification of specific peptides or proteins in complex samples. The technique selectively analyzes predefined precursor ions and their corresponding product ions. The two most common targeted acquisition methods are selected reaction monitoring (SRM) and parallel reaction monitoring (PRM).

SRM involves selecting and monitoring specific precursor-product ion pairs (known as transitions) for a particular peptide or protein.[36] The method is typically performed on a

triple quadrupole (QqQ) mass spectrometer. The mass spectrometer isolates precursor ions of interest, fragments them, and monitors the product ions produced. This results in highly accurate and reproducible quantification of target peptides or proteins. Multiple reaction monitoring (MRM) is another name for SRM.

PRM is similar to SRM except it is usually performed on a quadrupole-Orbitrap mass spectrometer (QqOrbi) or quadrupole-time-of-flight mass spectrometer (QqTOF).[37] PRM leverages high resolution and mass accuracy while simultaneously monitoring products of the target peptides. In PRM, the transitions are co-detected and distinguished by the final mass analysis stage. Because targeted acquisition methods offer high specificity, sensitivity, and reproducibility, they are commonly used for biomarker discovery, validation, and clinical applications.

1.3.3 Data-independent acquisition

Data-independent acquisition (DIA) is used for comprehensive analysis and quantification of complex mixtures of peptides and proteins.[38,39] In contrast to DDA, which selects precursors based on their intensity, DIA fragments all ions in a defined mass range. In a DIA experiment, the mass range is divided into predefined, overlapping windows, and the mass spectrometer acquires MS/MS spectra for all the peptides in each window (Figure 1.3). The MS2 spectra are searched against a protein database to identify and quantify the peptides in the sample.

DIA is less biased than other methods because it acquires data for all peptides in a sample and does not rely on precursor ion intensity or predefined targets. However, the complexity of the data is a challenge for data analysis because the spectra contain multiple precursors. Spectra containing multiple precursors are called chimeric spectra.

1.4 Analysis and quantification of mass spectrometry data

Mass spectrometry data analysis involves processing, interpreting, and visualizing data that is generated by a mass spectrometer. In proteomics, the overarching goal is to identify and quantify the peptides in the samples and to extract meaningful insights from the data.

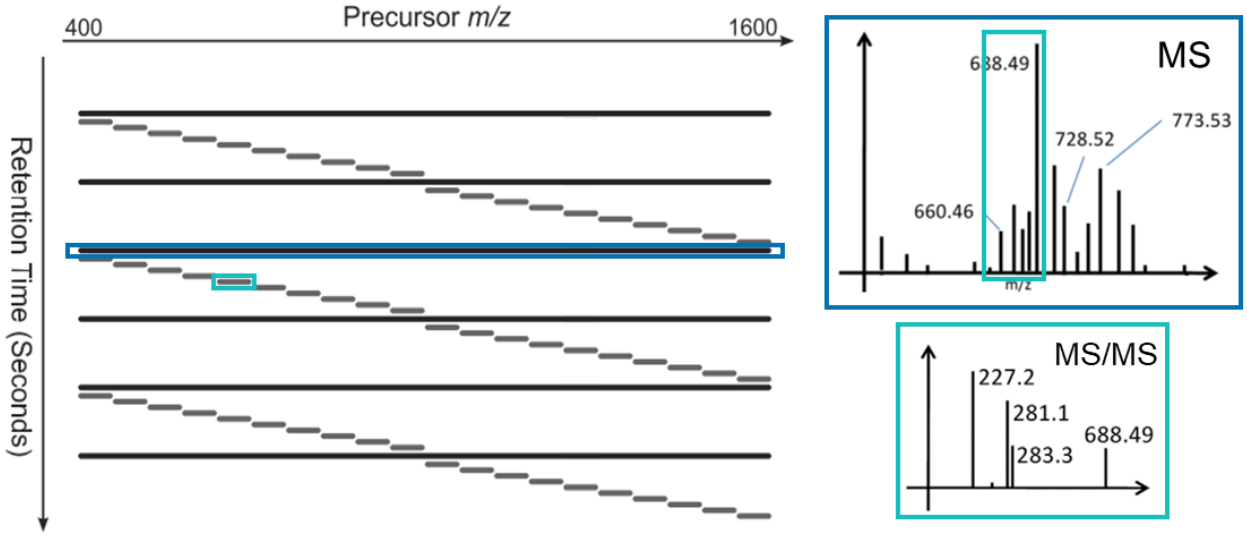


Figure 1.3: Schematic of data-independent acquisition (DIA).

Example workflows can be seen in Figure 1.4. The first step is to convert the raw data into a usable format. The data can also be processed to remove noise and artifacts and to correct for any instrumental biases or drifts. At this point, the data is usually represented as a mass spectrum. Mass spectra are shown with the intensity of ions relative to their m/z. The next step is to identify the components in the sample.

1.4.1 Database searching

Peptide identification is typically done by comparing the fragmentation patterns in mass spectra to a database of known compounds. In proteomics, database searching can be performed with a spectrum-centric approach or a peptide-centric approach. In a spectrum-centric approach, specialized software compares the experimental spectra to a database of theoretical fragment ion spectra and a matched decoy database. This approach is common to DDA. Some of the software used to search DDA data include Comet,[43] SEQUEST,[44–46] MaxQuant,[47] MSFragger,[48] Mascot,[49] and X!Tandem.[50]

The peptide-centric approach is needed for relatively complex mass spectra. For exam-

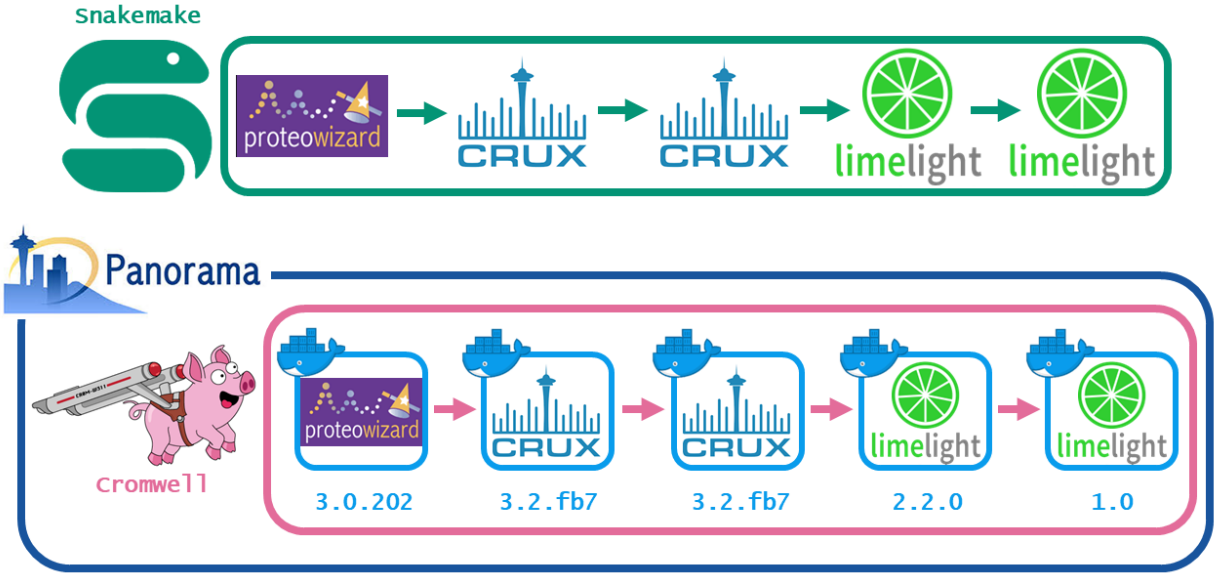


Figure 1.4: Pipeline workflows for data analysis. Workflows allow automated data analysis in parallel. Both example workflows demonstrate the analysis steps for DDA data (i.e., ProteoWizard msconvert,[40] crux Comet,[41] crux Percolator,[42] Limelight convert,[1] and Limelight upload). Snakemake workflows can easily be used locally or with high performance computing. The Cromwell workflows created by the lab use Docker containers and high performance computing to enable reproducible pipelines with tool version tracking. The lab is also planning to have workflow access through Panorama to allow usage by researchers with limited command line and software tool experience.

ple, DIA MS/MS mass spectra are more complex than DDA MS/MS mass spectra because there are more precursors present in the DIA MS2 spectra. In the peptide-centric approach, a reference spectral library is compiled containing the characteristics of peptide precursors of interest, such as retention time and the m/z of precursors and fragments. The spectral library can be generated with a deep learning algorithm, such as Prosit[51] (Appendix A), or experimentally. The approach then tries to find these patterns in the DIA data with statistical algorithms and machine learning. Some examples of software with peptide-centric approaches include EncyclopeDIA[52] (Appendix B), DIA-NN,[53] Spectronaut,,[54] and MaxDIA.[55]

The output of the search is a list of identified peptides or proteins and the statistical

scores indicating the quality of the match. The scoring can be based on many factors, such as the number of matching fragment ions, the mass accuracy of the measurements, and the retention time of the feature. Database searching can be challenging due to the numerous peptide sequences and the complexity of the data, so the accuracy and sensitivity of analyses are affected by the quality of mass spectra, the completeness and accuracy of the database, and the chosen search parameters.

If the features or sample are unknown, software tools can be used to perform *de novo* analysis on the mass spectra. Some examples of *de novo* sequencing tools include Lutefisk,[56,57] SHERENGA,[58] PEAKS,[59] DACSIM,[60] PepNovo,[61] NovoHMM,[62] PILOT,[63] MSNovo,[64] pNovo,[65] UniNovo,[66] and Novor.[67] These tools can assign identifications based on the fragmentation patterns of the ions. If targets are known, one option for quantitative analysis is comparing the ion intensities to a standard curve generated with known concentrations of the target compounds.

1.4.2 Quantification of mass spectrometry data

Methods for quantifying mass spectrometry data are categorized as relative or absolute.[68] Relative quantification compares a sample to a reference sample or a control. This method is commonly used in proteomics and metabolomics to compare the relative abundance of features across different samples. Absolute quantification involves determining the concentration of a specific compound using a known standard or calibration curve. This method is used in clinical diagnostics and pharmaceutical analysis, where measuring biomarkers and drug metabolites with accuracy and precision is critical.

1.5 New feature detection

New feature detection (NFD) or new peak detection (NPD) refers to the process of identifying and quantifying unknown or unannotated peaks in mass spectrometry data.[69] The unannotated features could be due to a variety of reasons, such as novel or previously low-abundance compounds, post-translational modifications, or features that were missing from

the initial database search.

New feature detection requires multiple algorithms and statistical models to process mass spectrometry data. The first step is performing peak detection, which involves identifying the m/z values and intensities of signals in the spectra. The next step is often an alignment, which involves matching peaks across multiple runs and correcting for differences in retention time and mass accuracy. After detection and alignment, new features are detected by identifying peaks that are not in the reference database.

The output of new feature detection is a list of detected features with their m/z values, retention times, intensities, and other parameters. These features can be investigated and annotated to improve the knowledge of the sample. New feature detection can help uncover unexpected aspects of complex biological systems, such as proteoforms, protein-protein interactions, metabolic responses to stimuli, and impurities. For these reasons, it is a powerful tool in the fields of disease diagnosis, environmental monitoring, disease diagnosis, drug discovery, and the development and quality assurance of biopharmaceuticals.

1.5.1 Multi-Attribute Method

Multi-attribute method (MAM) is a quality control approach that is used in the development and manufacturing of biopharmaceuticals.[70] It involves the analysis of multiple attributes to assess the quality of a drug or biologic. The attributes analyzed can include physicochemical properties, identity, potency, and purity. MAM combines multiple analytical techniques to detect and quantify impurities, aggregates, post-translational modifications, and other quality attributes that could affect the safety and efficacy of the product.

MAM is used at several stages of biopharmaceutical development and manufacturing. It can be used to ensure consistent product quality by monitoring batch-to-batch variability, identifying potential sources of variation, and optimizing the production process. MAM can be useful for cell line development, production, formulation, and purification. A key advantage of MAM is the ability to analyze multiple attributes simultaneously. Traditional analytical methods focus on specific attributes, which prevents the ability to see an integrated

view of the product quality. By analyzing multiple attributes in unison, MAM can help identify correlations and potential associations between different parameters. This strength can provide a more accurate and reliable assessment of the product.

Chapter 2

MSIONS: A PYTHON PACKAGE FOR EVALUATING THE QUALITY OF MASS SPECTROMETRY DATA.

2.1 Summary

When most researchers perform quality checks of mass spectrometry data, they only examine the number of peptide and protein identifications. Identifications can be a useful metric to monitor when people are familiar with the sample they are analyzing. However, a high or low number of identifications may not be a good representation of a new sample and could lead to biases in the conclusions that are made. *msions* is a Python package that can be used to check the quality of mass spectrometry data and allow more accurate comparisons with single molecule counting methods. *msions* is freely available for Python 3.9+, has documentation available online, and can be installed with *pip*. The code is open source under the Apache 2.0 license at `https://github.com/dafaivre/msions`.

2.2 Introduction

Mass spectrometry (MS) is a powerful analytical technique for analyzing the chemical and biological properties of samples. However, the accuracy and reliability of MS-based experiments depend on the quality of generated data. MS data quality is affected by multiple factors, including instrument performance, sample preparation, data acquisition, and data processing. This requires paying careful attention to the experimental design and data analysis, such as optimizing instrument parameters, introducing appropriate quality control measures, and using robust data processing algorithms. In recent years, the MS research community has begun focusing on the development of methods and tools to improve data quality and reproducibility.

Rudnick *et al.* created a comprehensive list of performance metrics for QC analysis that have been used by many software tools and pipelines.[71] RawMeat and LogViewer[72] were two of the earliest QC tools. Both tools provided identification-free information about instrument performance, but they were limited to Thermo instruments and not updated for the latest instrumentation. MSQC used search engine results but required multiple format conversions and lacked clear visualizations. QuaMeter[73] solved some challenges of MSQC by using a generic format for the spectral data (mzML).[74] However, the output of the tool still required downstream analysis.

In this chapter, we present the *msions* package for MS data quality evaluation. The *msions* package allows users to easily examine the percentage of their signal that is being identified by visualizing the data and printing statistics. The *msions* package also enables the ability to work with the outputs of the tools used in the analysis and has additional functionality that is described in the documentation `https://msions.readthedocs.io/`. *msions* is freely available under the open-source Apache 2.0 license at `https://github.com/dafaivre/msions`.

2.3 Methods and Results

The *msions* package empowers users to analyze mass spectrometry data and create visualizations. When most people perform quality checks of MS data, they only care about the counts of peptide and protein identifications. Identifications can be a useful metric to monitor when researchers are familiar with the sample they are analyzing. However, a high or low number of identifications may not be a good representation of a new sample and can lead to biases in the results and conclusions that are discussed. *msions* has the ability to plot the total signal and the portion of the signal that has been identified. To generate these plots, the package needs an mzML[74] file and the mzML's corresponding search engine results and Hardklor[75] file of MS1 features. In the examples shown below, EncyclopeDIA[76] was the search engine used to analyze data-independent acquisition (DIA) data and generate the results files.

Use Cases. To demonstrate the functionality of *msions*, we provide the following three

use cases. First, *msions* can be valuable for determining the quality of data and finding mistakes in the data processing. If the wrong FASTA is used to create a spectral library, the percentage of identified signal will be much lower than expected (Figure 2.1). In this case, a human plasma sample was enriched for extracellular vesicles (EVs). When searched against a human database, over 44 thousand peptides were identified, which covered 47.0% of the total signal (Figure 2.1a). If a mouse database was mistakenly used, 17 thousand peptides would be identified, which could seem like an impressive value without knowledge of the human database results (Figure 2.1b). However, only 13.7% of the signal was identified. This should trigger an audit of the data analysis process because around 30% or higher is a reasonable amount of identified signal based on the experiments that have been analyzed. If no mistake can be found in the data processing step, the researchers should back track farther in their experiments. There could be a mistake in the sample type and that could be the reason for the poor database matches.

Second, the package can be useful for inspecting experiments that include an enrichment step. In Figure 2.2, the results are shown for a human plasma sample (a) and the EV enrichment of the plasma sample (b). Using the same database, the plasma sample had 8,100 peptide identifications covering 42.0% of the signal, and the EV enrichment had 44,615 peptide identifications covering 47.0% of the signal. While the large numbers of peptide identifications are impressive, the consistency in the amount of signal being identified can support the argument for the quality of the mass spectrometry data. In (c) and (d), the albumin protein sequence was removed from the database and thus could not be identified. For plasma, the removal of identified signal from albumin dropped the percentage of identified signal by over half–from 42.0% to 20.2%. This agrees with the relative abundance of albumin to total plasma protein.[77,78] On the other hand, the amount of identified signal barely changed for the EV enrichment sample. This is expected because albumin should not be a major component after EV enrichment.

Lastly, in addition to plotting the total and identified TIC, *msions* is also able to calculate the number of ions that the mass spectrometer has been analyzing (Figure 2.3). This can be

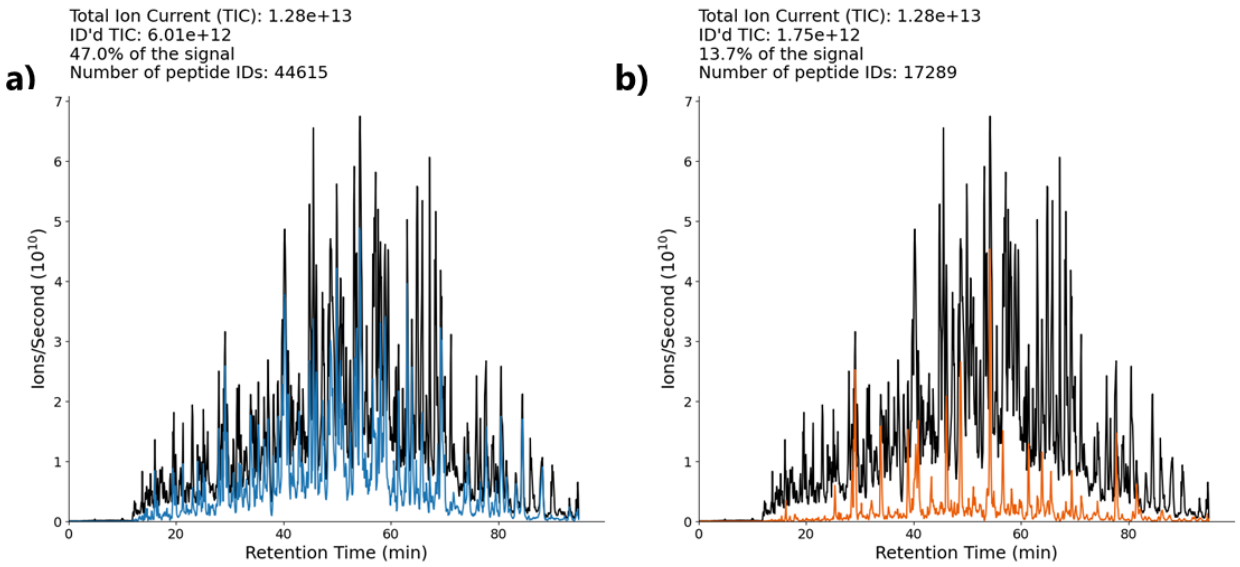


Figure 2.1: *msions* use case for wrong database. (a) The total ion current (TIC) plot for a human plasma sample enriched for extracellular vesicles (EVs). The TIC signal is plotted in black, and the blue represents the fraction of the MS1 signal that has been confidently identified. The y-axis represents an approximation of counts (ions per second). The data were analyzed only for unmodified and fully tryptic peptides from the canonical human protein sequences obtained from Uniprot. (b) The TIC signal of the EV enrichment seen in *a*. However, the data has been searched with a mouse database instead of a human database. The orange represents the fraction of the MS1 signal that has been confidently identified.

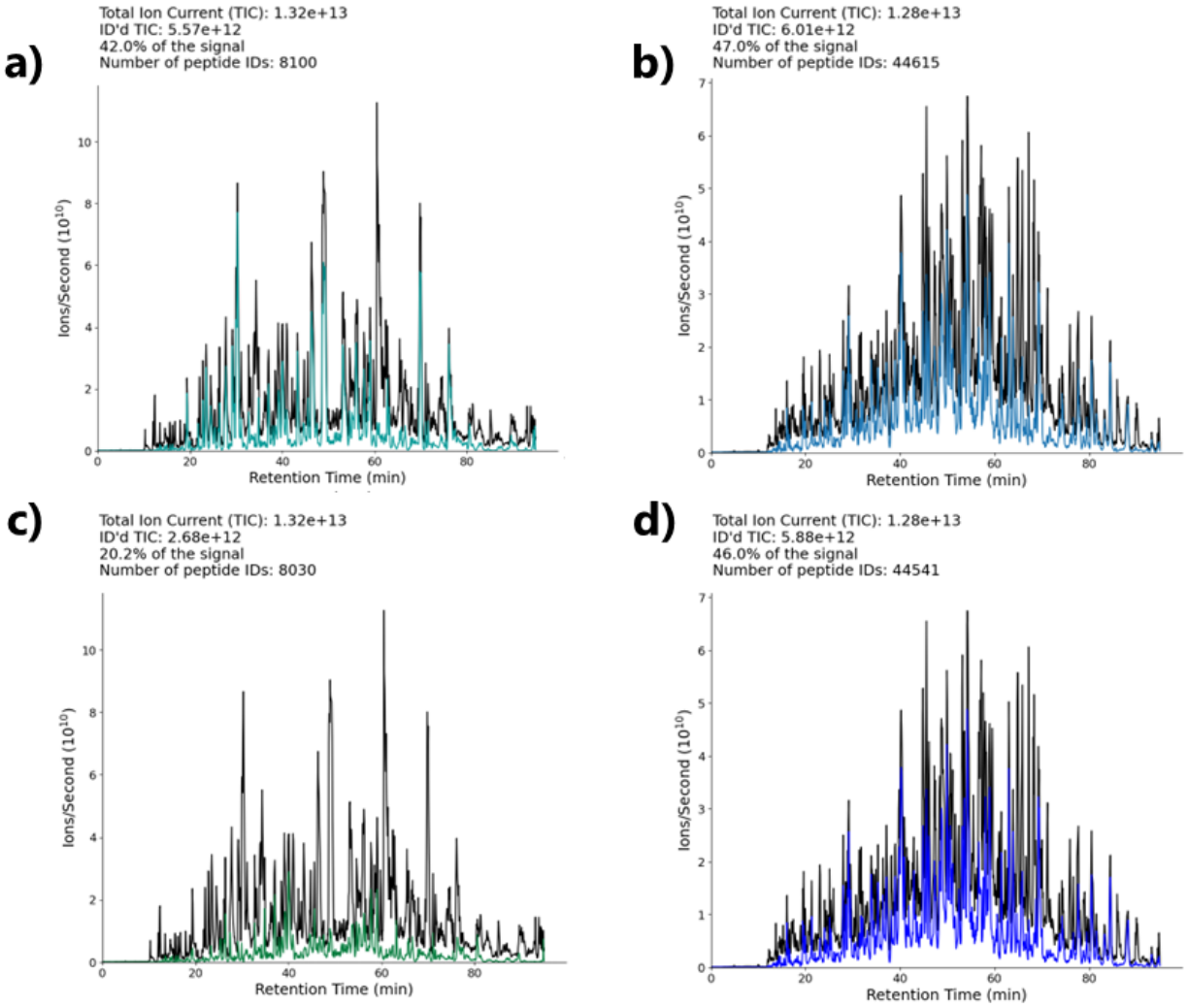


Figure 2.2: *msions* use case for enrichment protocols. (a) The total ion current (TIC) plot for a human plasma sample. The TIC signal is plotted in black, and the teal represents the fraction of the MS1 signal that has been confidently identified. The y-axis represents an approximation of counts (ions per second). (b) The total ion current (TIC) plot for the EV enrichment of the human plasma sample in *a*. The blue represents the fraction of the MS1 signal that has been confidently identified. This is the same plot as seen in Figure 2.1a. (c) The TIC signal of the human plasma sample seen in *a*. However, the protein sequence for albumin was removed from the human database. The green represents the fraction of the MS1 signal that has been confidently identified, not including albumin. (d) The TIC signal of the EV enrichment of the plasma sample in *a*. However, the protein sequence for albumin was removed from the human database. The indigo represents the fraction of the MS1 signal that has been confidently identified, not including albumin.

valuable to users that want a reasonable comparison to single-molecule counting methods. Traditionally, the MS proteomics field reports lists of peptides detected and the proteins they are derived from. As peptides elute off the high-performance liquid chromatography column, the instrument counts large numbers of peptide ions based on their mass-to-charge (m/z) ratio, independently of their sequence identification (Figure 2.4a). The abundance of each analyte is often determined from a background-subtracted peak area. Depending on the method used, the peak area can be obtained from the unfragmented MS1 spectra or from tandem mass spectra (MS/MS or MS2) collected. The peak area is derived from the detector ion current, measured by either the flow of ions to an electron multiplier[79] or the generation of an image current in a Fourier transform mass analyzer.[80] The current is a measure of the number of ions counted, normalized by the amount of time spent sampling the signal. The measured signal is an approximation of counts and proportional to ions per second, and thus it can be converted into a number of counted ions for direct comparison with single-molecule counting methods.[81–83]

LC-MS/MS methods can improve the sensitivity to low-abundance analytes by changing the time spent sampling the signal (also known as dwell time or injection time). In some MS instruments, such as ion traps, the time spent sampling ions changes dynamically depending on the signal at that time.[84] This dynamic adjustment of the injection time, known as automatic gain control (AGC), provides an ideal ion population for the MS measurement (Figure 2.4b). However, an added benefit of AGC is that it enables the instrument to spend less time on abundant molecular species but scale the current into a larger quantity while maintaining quantitative linearity. Likewise, it enables the instrument to spend more time on less abundant peptides to enable the measurement of the weaker signal. This increases the dynamic range and the total number of ions identified (Figure 2.4c). Dividing each spectrum intensity by the time taken to acquire the spectrum gives a normalized signal for each spectrum that is analogous to normalizing the counts obtained between flow cells in a single-molecule counting experiment.[85]

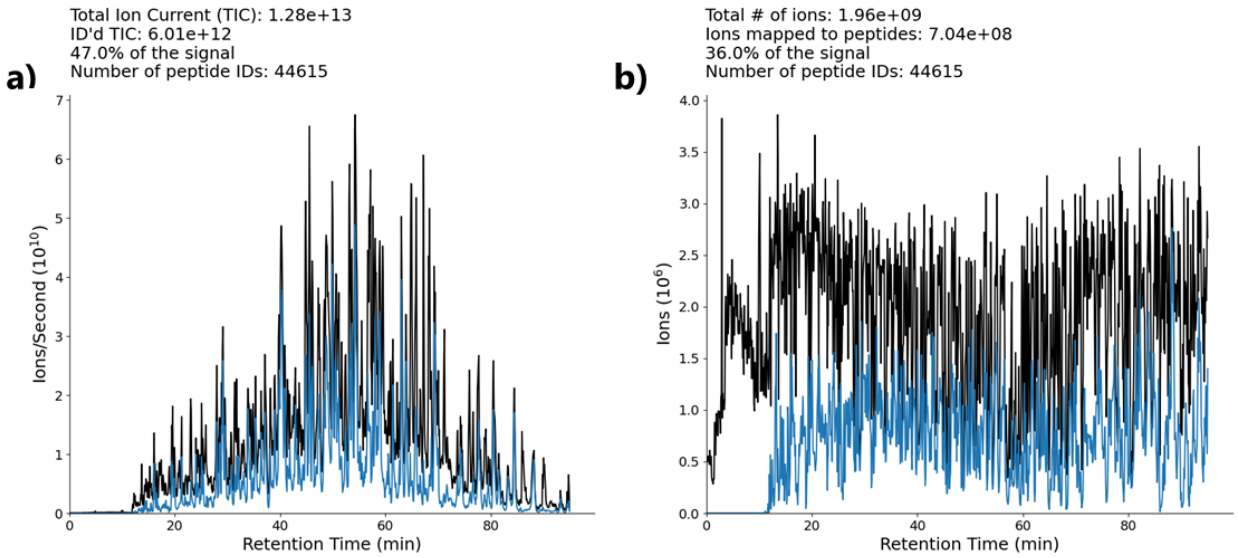


Figure 2.3: Example of TIC and ion plots from *msions* package. (a) The same total ion current (TIC) plot for the EV enrichment of the human plasma sample. The TIC signal is plotted in black, and the blue represents the fraction of the MS1 signal that has been confidently identified. The y-axis represents an approximation of counts (ions per second). The data were analyzed only for unmodified and fully tryptic peptides from the canonical human protein sequences obtained from Uniprot. (b) Representation of the same data plotted in *a* but with the y-axis of each spectrum adjusted to an estimate of ions by multiplying the counts by the Orbitrap injection time. The variable fill times allow peptides with relatively low abundances to be measured with a similar number of ions as the most abundant peptides in the analysis.

Figure 2.4 shows the analysis of an extracellular vesicle (EV) fraction after enriching the EVs from plasma.[86] The fraction was digested using trypsin and analyzed with DIA on an Orbitrap Eclipse instrument. The enriched sample has a lower dynamic range than the whole plasma proteome. The fraction represents about 1-2% of the plasma proteome, is enriched in tissue-derived proteins, and is depleted in abundant plasma proteins. The total ion current from just the MS1 signal was $>10^{12}$ ions per second, of which 46.4% could be assigned to a peptide sequence using the fragment ion data. This current represented >5 billion ions, of which 1.2 billion ions (24.1%) were assigned to peptide sequences.

Implications for Single-Molecule Approaches. To perform similarly, single-molecule methods like Illumina would need to separate a sample into thousands of fractions to collect billions of reads (∼1 million reads per fraction). The signal is normalized between flow cells to achieve counts that can be comparable, with ∼24% of the reads being able to be mapped back to the reference genome. This plasma extracellular vesicle analysis was not sample limited and thus represents the upper end of what can be achieved for the analysis of ions per analysis time.

If we assume that emerging polypeptide counting methods can achieve the current throughput of Illumina NovaSeq for DNA (4 billion reads for $10,000), the cost for analyzing a mammalian proteome would be much higher than the cost by MS analysis. The single-molecule counting approaches would need to be at least 20-fold cheaper than Illumina sequencing to be cost effective when compared with $500 per LC-MS/MS analysis.[87] The typical mammalian cell – e.g., a HeLa cell with a volume of (∼3,000 µm^3) – contains about 300,000 mRNA molecules[88] and about 10 billion protein molecules.[89] Given these estimates of the relative abundance ratio of mRNA to protein molecules, we calculate that about 30,000-fold more counts are required to characterize the protein molecules at an analogous coverage to that which has been achieved with the transcriptome. These implications and Figure 2.4 have been discussed in more detail in a published *Nature Methods* comment.[87]

Code Availability. The *msions* package is available for Python 3.9+ and can be easily installed from the Python Package Index (PyPI) with *pip*. The code for the package is open

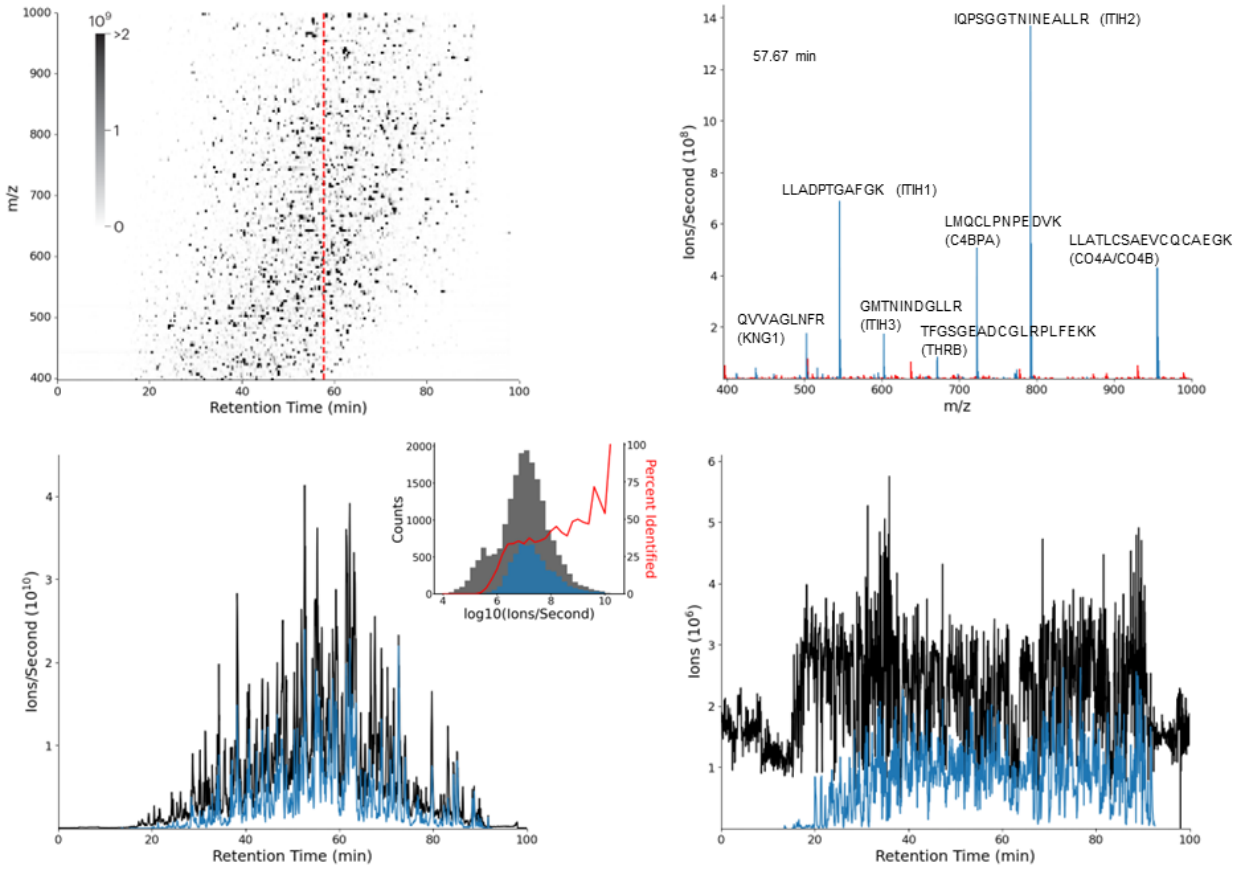


Figure 2.4: *msions* use case for single molecule counting. Signal from the MS1 spectra of an LC-MS run of enriched extracellular vesicles from human plasma using data-independent acquisition on a Thermo Scientific Orbitrap Eclipse. (a) An ion map of the MS1 peptide signal separated in the retention time and m/z dimensions. The red dashed line indicates the location of the spectrum in *b*. (b) Selection of a single MS1 spectrum collected at 57.67 min; blue m/z values have been assigned a peptide sequence and red m/z values are unassigned in the analysis. (c) A total ion current (TIC) plot of the signal intensity from *a* at all time points. The TIC signal is plotted in black, and the blue represents the fraction of the MS1 signal (for example, in *b*) that has been confidently assigned to peptide sequences. The y-axis represents an approximation of counts (ions per second). The insert is a histogram counting distinct molecular entities (features) for different measured intensities. The gray bars of the insert represent all molecular features and the blue represents those assigned a peptide label. The data were analyzed only for unmodified and fully tryptic peptides from the canonical human protein sequences obtained from Uniprot. (d) Representation of the same data plotted in *c* but with the y-axis of each spectrum adjusted to an estimate of ions by multiplying the counts by the Orbitrap fill time. The variable fill times allow peptides with relatively low abundance near 20-30 min to be measured with a similar number of ions as the most abundant peptides in the analysis. The result is billions of peptide ions counted in just 90 min. Data available at `https://panoramaweb.org/Single_Molecule_Counting.url` and under PXD035637. Code available at `https://github.com/uw-maccosslab/single_molecule_counting`.

source under the Apache 2.0 license at `https://github.com/dafaivre/msions`. The documentation for the *msions* package package is available at `https://msions.readthedocs.io/`.

2.4 Conclusions

Here, we have presented *msions* and demonstrated use cases on how *msions* can help investigate the quality of MS data. The *msions* package is a useful tool for detecting issues in data and is a good complement for users that typically only track peptide and protein counts. In its current form, we believe it will be beneficial to users that have limited familiarity with quality control. It can be useful to a greater number of people if it is utilized prior to completion of full mass spectrometry experiments and if it is integrated with other data analysis software, pipelines, and processes.

Chapter 3

COMPARING PEPTIDE IDENTIFICATIONS BY FAIMS VERSUS QUADRUPOLE GAS-PHASE FRACTIONATION

3.1 Summary

Gas-phase fractionation techniques have become popular in the field of mass spectrometry to improve results without using more sample. High-field asymmetric waveform ion mobility spectrometry (FAIMS) coupled to liquid chromatography-mass spectrometry (LC-MS) has been shown to increase peptide and protein detections compared to LC-MS/MS alone. However, FAIMS has not been compared to other methods of gas-phase fractionation, such as quadrupole gas-phase fractionation, which could increase our understanding of the mechanisms of improvement. The goal of this work was to assess whether FAIMS improves peptide identifications because 1) gas-phase fractionation enables the analysis of less abundant signals by excluding more abundant precursors from filling the ion trap, 2) the use of FAIMS reduces co-isolation of peptides during the MS/MS process resulting in a reduction of chimeric spectra, or 3) a combination of both. To investigate these hypotheses, pooled human brain tissue samples were measured in triplicate using FAIMS gas-phase fractionation, quadrupole gas-phase fractionation, or no gas-phase fractionation. To confirm the results, the experiment was reproduced on another instrument. On both instruments, our data confirmed prior observations that FAIMS increased the number of peptides identified. We further demonstrated that the main benefit of FAIMS is due to the reduced co-isolation of persistent peptide precursor ions, which results in a decrease in chimeric spectra.

3.2 Introduction

Ion trap mass spectrometers have been the workhorse instruments in many protein mass spectrometry laboratories for years. What quadrupole ion trap mass spectrometers lack in resolution or mass accuracy, they make up in full scan sensitivity, MS/MS scan speed, and round-the-clock robustness. With the availability of orbitrap mass analyzers, we now have Fourier transform mass accuracy, resolving power, and dynamic range in benchtop capabilities and in combination with a quadrupole ion trap.[90,91]

An important strength of ion trap mass spectrometers is the application of automatic gain control (AGC).[84] AGC accumulates low abundance molecular species to fill the trap and adjusts the accumulation time based on the abundance of analyte ions. In an MS1 survey spectrum covering a wide m/z range, a single abundant peptide can compromise the dynamic range if the majority of the ions within the trap represent the abundant peptide. However, in an MS/MS acquisition, only molecular species within the isolated precursor m/z window are stored within the trap. This facilitates the accumulation of low abundance species in the presence of more abundant ions at different m/z. Thus, it is not uncommon to obtain high quality MS/MS spectra for peptides when there was no detectable precursor during the MS1 survey scan.[92] Because the instrument dynamically adjusts the ion accumulation time to fill the trap, the MS/MS spectra can be of similar quality regardless of the abundance of the analyte in the mixture.

Because of AGC, the use of an ion filter to eliminate abundant peptides is particularly powerful when used in combination with an ion trap mass analyzer. Filtering high abundance molecular species from the analyte stream allows the mass spectrometer to accumulate and fill the trap with low abundance peptides. This idea was the basis of DREAMS reported by the Smith lab where a quadrupole was used to selectively exclude abundant m/z values from the ion cyclotron resonance cell to improve the dynamic range.[25] Additionally, the Goodlett lab made use of gas-phase fractionation to isolate only a narrow m/z range in the acquisition of their MS1 survey scan.[26] More recently, Meier *et al.* implemented BoxCar

where the maximal orbitrap charge capacity could be accumulated over multiple narrow and discontinuous isolation windows[93] — the full mass range covered in subsequent scans. All of these methods enable the exclusion of abundant precursors and facilitate the use of longer injection times to improve the sensitivity of low abundant precursors.

High-field asymmetric waveform ion mobility spectrometry (FAIMS) can be used to decrease the complexity of the mixture before the ions enter the mass spectrometer. FAIMS was first used to measure peptides by Guevremont and coworkers[27–29] and was later applied to a biological sample by Venne *et al.*[30] The FAIMS interface separates ions by applying a high-voltage asymmetric waveform to a set of electrodes and allowing ions, entrained in a carrier gas, to flow through the gap between the electrodes. The ion separation occurs because ions have different mobilities in high and low electric fields. The asymmetry of the applied electric field, coupled with the field-dependent ion mobility, causes ions to acquire a net displacement perpendicular to their direction of motion and collide with one of the electrodes. A DC offset (compensation voltage or CV) can be applied to one electrode, which can counter this displacement and enable the transmission of ions.

The FAIMS device thus acts as a filter, transmitting only a portion of the total ions into the mass spectrometer—reducing the sample complexity, minimizing chemical noise, and increasing the dynamic range.[27–29] Because FAIMS filters ions by differential mobility in a continuous fashion, low abundance ions can be accumulated using AGC in a similar way to using a quadrupole to isolate a subset of the m/z range. Our lab and others have previously demonstrated that the FAIMS CV can be stepped in a synchronized fashion with the acquisition of mass spectra, allowing ions at multiple selected differential ion mobilities to be measured in either separate or the same LC-MS runs.[30,94] Recent improvements in FAIMS hardware have renewed interest in the technology.[95–97] The hardware improvements have made FAIMS into a gas-phase fractionation method that can easily be performed prior to an ion trapping mass spectrometer. While the speed which the FAIMS CV can be stepped is slower than stepping to different m/z regions with a quadrupole filter, an advantage of FAIMS is that the separation is at least partially orthogonal to m/z.[94]

While the improvement of FAIMS on the identification of peptides by data dependent acquisition is clear, the mechanism for how FAIMS improves the identifications is not well understood. Previous FAIMS experiments have benchmarked their experiments against analyses performed simply without FAIMS and never had an alternative gas-phase fractionation method as a control.[96] Thus, it is not clear whether FAIMS acts just like other gas-phase fractionation methods by separating ions from the HPLC into separate bins—enabling the ion trap mass spectrometer's use of AGC to fill longer and improve the sensitivity for lower abundant peptide precursors in the absence of abundant peptides. If FAIMS works simply by improving the dynamic range of the MS1 spectrum, then a gas-phase fractionation method that makes use of the quadrupole to isolate subsets of the m/z range should perform equally as well as, or better than, FAIMS. Thus, the goal of this work was to assess whether FAIMS improves peptide identifications because 1) gas-phase fractionation enables the analysis of less abundant signals by excluding more abundant precursors from filling the ion trap, 2) the use of FAIMS reduces co-isolation of peptides during the MS/MS process resulting in a reduction of chimeric spectra, or 3) a combination of both.

3.3 Methods

Tryptic digestions. The samples were collected at the University of Washington, Kaiser Permanente Washington, and Stanford University. The UW and Stanford Human Subjects Division deemed this project to be non-human subjects research because we used pre-existing de-identified samples. For more information about the samples, see Merrihew *et al.*[98] Briefly, two 25 µm frozen sections of human brain tissue were resuspended in 120 µL of lysis buffer containing 5% SDS, 50mM triethylammonium bicarbonate (TEAB), 2mM MgCl2, 1X HALT phosphatase and protease inhibitors. The suspension was vortexed and briefly sonicated with a Fisher sonic dismembrator model 100 set to setting 3 for 10 s. A microtube was loaded with 30 µL of lysate and capped with a micropestle. The sample was homogenized with a Barocycler 2320EXT (Pressure Biosciences Inc.) for a total of 20 minutes at 35°C with 30 cycles of 20 seconds at 45,000 psi and 10 seconds at atmospheric pressure.

Protein concentration of the homogenate was measured with a BCA assay. Fifty micrograms were added to a process control of 800 ng of yeast enolase protein (Sigma), reduced with 20 mM DTT, and alkylated with 40 mM IAA. The lysate was prepared for S-trap column (Protifi) cleaning by adding 1.2% phosphoric acid and 350 µL of binding buffer (90% Methanol, 100 mM TEAB). The acidified lysate was bound to column incrementally, followed by 3 wash steps with binding buffer to remove SDS, 3 wash steps with 50:50 methanol:chloroform to remove lipids, and a final wash step with binding buffer. Trypsin (1:10) in 50mM TEAB was added to the S-trap column for digestion at 47°C for one hour. Hydrophilic peptides were eluted with 50 mM TEAB and hydrophobic peptides were eluted with a solution of 50% acetonitrile in 0.2% formic acid. Elutions were pooled, speed vacuumed and resuspended in 0.1% formic acid.

NanoLC conditions. On both instruments, 1 µg of sample was loaded into the system. For instrument 1, a Thermo Easy-nLC 1200 was used with a 30 cm fused silica pulled tip column (New Objective, 75 µm inner diameter) and a 4 cm fused silica (150 µm inner diameter) Kasil1 (PQ Corporation) frit trap. The trap and column were loaded with 3 µm Reprosil-Pur C18 (Dr. Maisch) reverse-phase resin. Buffer A was 0.1% formic acid in water and buffer B was 0.1% formic acid in 80% acetonitrile.

The 60-minute LC gradient was 2 to 7% B in 1 minute, 7 to 40% B over 40 minutes, 40 to 60% B over 5 minutes, 60 to 98% B over 5 minutes, a 5 minute wash at 98% B, a return to 2% B in 1 minute, and a 3 minute equilibration at 2% B. The 180-minute LC gradient was 2 to 7% B in 1 minute, 7 to 40% B over 160 minutes, 40 to 60% B over 5 minutes, 60 to 98% B over 5 minutes, a 5 minute wash at 98% B, a return to 2% B in 1 minute, and a 3 minute equilibration at 2% B. Peptides were eluted from the column and electrosprayed into a Thermo Eclipse Tribid Mass Spectrometer with the application of a distal 3 kV spray voltage. Application of the mass spectrometer and LC solvent gradients were controlled by ThermoFisher Xcalibur (version 3.3).

For instrument 2, 1 µg of sample was loaded into the system on an in-house pulled 30

cm C18 (ThermoAccucore, 2.6 Å, 150 µm) column. Buffer A was 5% acetonitrile/0.125% formic acid and buffer B was 0.125% formic acid in 95% acetonitrile. The 60-minute LC gradient was 4 to 35% B over 55 minutes, 35 to 100% B in 5 minutes, and a 5 minute wash at 100% B. The 180-minute LC gradient was 4 to 35% B over 175 minutes, and 35 to 100% B in 5 minutes. Both gradients had a 5 minute wash at 100% B. Peptides were eluted from the column and electrosprayed into a Thermo Eclipse Tribid Mass Spectrometer with the application of a distal 3 kV spray voltage. Application of the mass spectrometer and LC solvent gradients were controlled by ThermoFisher Xcalibur (version 3.5).

Mass Spectrometry. This experiment was designed to replicate a previous experiment,[96] introduce a new control, and test the reproducibility of results on more than one instrument. Eluted peptides were analyzed with two Orbitrap Eclipse Tribrids. The two instruments examined different pooled human brain tissue samples with slightly different LC setups and gradients. The preparation of the pooled human brain tissue samples has been described in detail previously,[98] and the nanoLC conditions can be found in the previous section. Experiments without FAIMS used a 240,000 resolving power MS1 survey scan, Standard AGC Target, and Auto Maximum Injection Time, followed by MS/MS of the most intense precursors for 1 second. The MS/MS analyses were performed by 0.7 m/z isolation with the quadrupole, normalized HCD (higher-energy collisional dissociation) energy of 30%, and analysis of fragment ions in the ion trap using the "Turbo" speed scanning from 200 to 1200 m/z. Dynamic exclusion was set to 10 seconds for the 1-hour analyses and was increased to 30 seconds for the 3-hour analyses. Monoisotopic precursor selection (MIPS) was set to Peptide, maximum injection time was 35 milliseconds, AGC target was 200%, unusual charge states (unknown, +1, or >+5) were excluded, the advanced peak determination was toggled on, and the internal mass calibration was off.

For FAIMS experiments, the settings were identical except the FAIMS device was used between the electrospray source and the mass spectrometer. FAIMS separations were performed with a 100 °C inner electrode temperature, 100 °C outer electrode temperature, 4.7

L/min FAIMS carrier gas flow, and –5000 V dispersion voltage (DV). The FAIMS carrier gas was N_2 sourced from evaporated liquid nitrogen. For external stepping (i.e., single CV or single quadrupole fraction) experiments, the selected CV or quadrupole fraction was applied to all scans throughout the analysis. For internal stepping experiments, each of the 3 selected CVs or quadrupole fractions was applied to sequential survey scans. The MS/MS CV was always paired with the appropriate CV from the corresponding survey scan. For the 3-hour quantitative FAIMS experiments, the survey scan MS resolving power was reduced to 120,000 to permit a cycle time of 0.6 s. The 3 selected CVs (–50, –65, and –85) were chosen based on the results in Hebert et al.[96] The 3 quadrupole fractions were sample-specific and were chosen based on splitting the number of peptide-like features in a normal LC-MS run into thirds.

Data Analysis for Identifications. Raw files were converted into mzMLs using ProteoWizard's msconvert.[40] Peptides and proteins present in the samples were identified using Comet[43] by searching against the human proteome plus common contaminants. The Comet search results were post-processed and a q-value was assigned to each PSM and each peptide using Percolator.[99] The processed results were visualized using Limelight.[1] All reported PSMs and peptides were filtered at a false discovery rate (FDR) of 1%.

mzML Splitting. Raw files from internal stepping experiments cannot be directly processed by Hardklor[75] and Bullseye[100] need features to be present in multiple scans in a row to be considered persistent. To enable this analysis, we generated separate mzML files for each compensation voltage and quadrupole fraction. Available data conversion software such as msConvert[40] can generate compatible mzML files, but it does not currently separate scans by different compensation voltages. Compatible mzML files were created from unseparated mzML files using a Python script developed in-house (`https://github.com/uw-maccosslab/faims_vs_quadgpf`). The script uses functionality provided via the pymzML module (`https://github.com/pymzml/pymzML`).[101]

Data Sharing. All raw data, mzMLs, and other files used for analysis are available on

Panorama Public (`https://panoramaweb.org/faimsvs_quadgpf.url`, ProteomeXchange ID: PXD043458). The database search results can be seen on Limelight (`https://limelight.yeastrc.org/limelight/p/faims_vs_quadgpf`). The figures can be reproduced using the open source code found on GitHub (`https://github.com/uw-maccosslab/faims_vs_quadgpf`).

3.4 Results and Discussion

To assess the mechanism for why FAIMS improves peptide identifications, we analyzed tryptic digests of human brain tissue.[98] Samples were measured in triplicate using data dependent acquisition with FAIMS gas-phase fractionation (3 CV steps: –50, –65, –85 V), with quadrupole gas-phase fractionation (three selected ion monitoring m/z ranges), or without any gas-phase fractionation (Figure 3.1, right side). We also examined what we call external stepping versus internal stepping for the gas-phase fractionation methods (Figure 3.1, left side). For external stepping, a selected CV or limited mass range was applied to all scans throughout the analysis. For internal stepping experiments, each of the selected CVs or limited mass ranges was applied to sequential survey scans and MS/MS cycles. The MS/MS CV was always paired with the same CV from the corresponding MS1 survey scan.

Previously, the internal stepping experiments outperformed the external stepping experiments for peptide and protein detections.[96] This trend was seen in the dataset collected with Instrument 1, but it was not statistically significant in all cases (Figure 3.2a). This result was expected because the instrument can quickly switch between different fractions internally while the external stepping experiments are limited by the dead column loading time at the beginning of each run. For Instrument 2, this trend was not seen (Figure 3.2b). This difference may be the result of the optimized LC gradient reducing dead time at the beginning of each run (Figure 3.3). While internal stepping may not have a significant improvement in identifications, the method is better than external stepping because the multiple runs needed for external stepping require more sample and additional time for gradient preparation, column equilibration, and sample loading.

FAIMS gas-phase fractionation performed significantly better (q-value < 0.05) in peptide

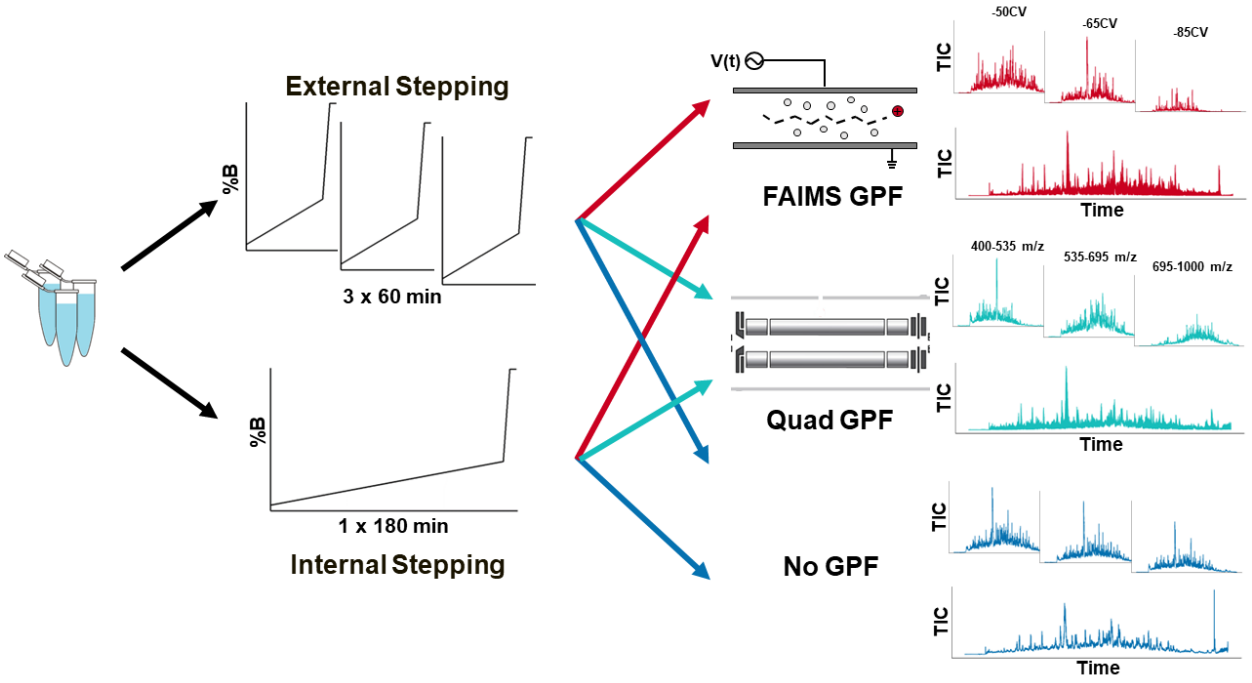


Figure 3.1: Scheme for the comparison of different gas-phase fractionation (GPF) methods to improve data dependent acquisition mass spectrometry. Pooled human brain tissue samples were measured in triplicate using external stepping (three 1-hour runs) and internal stepping (one 3-hour run). The experiments used FAIMS gas-phase fractionation (3 CV steps: –50, –65, –85 V), quadrupole gas-phase fractionation (three selected ion monitoring m/z ranges), or no gas-phase fractionation (right side). Internal stepping applied each of the 3 selected CVs or quadrupole fractions to sequential survey scans in a run while external stepping applied the selected CV or quadrupole fraction to all scans in a run.

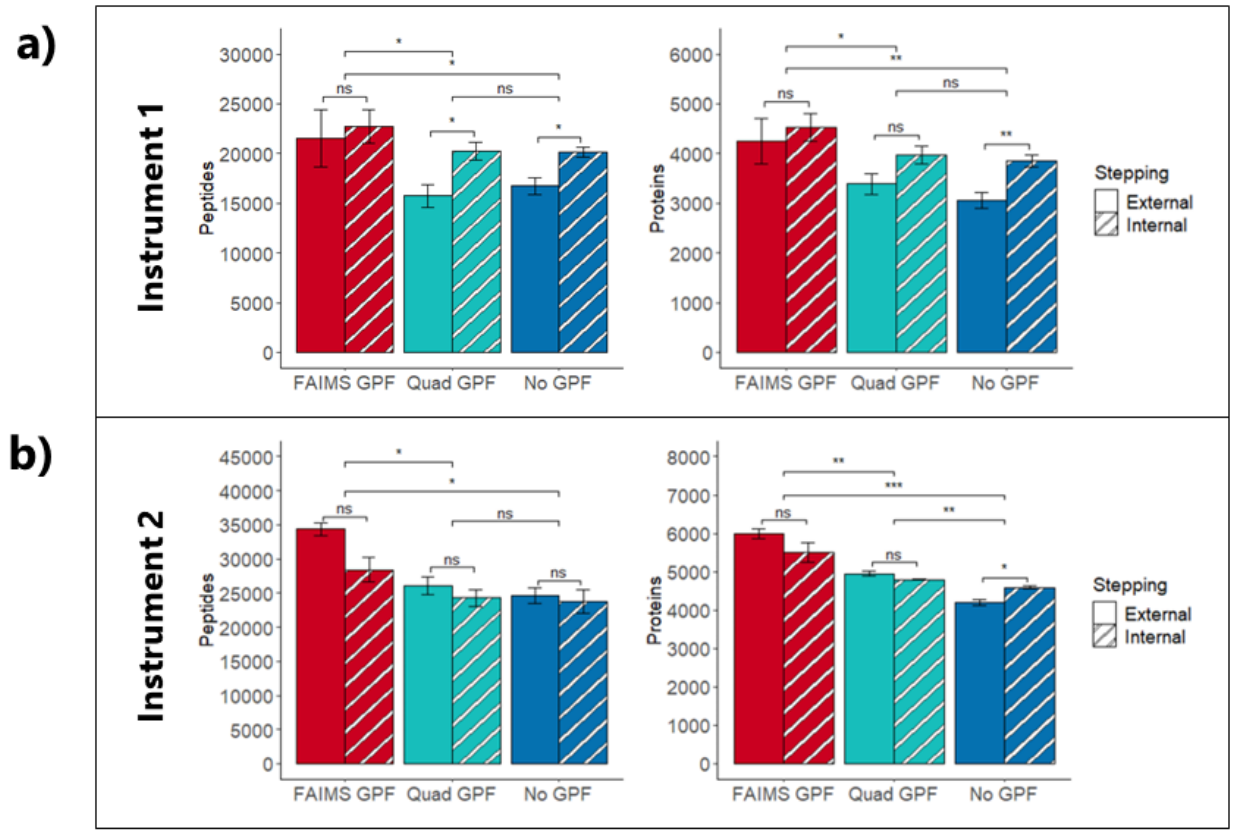


Figure 3.2: Peptide and protein counts for internal and external stepping gas-phase fractionation experiments. On both instruments, FAIMS gas-phase fractionation (red) performed significantly better (q-value < 0.05) in peptide (left) and protein (right) detections than quadrupole gas-phase fractionation (teal) or the absence of gas-phase fractionation (blue). The internal (striped) and external (solid) stepping results were not consistent between both instruments.

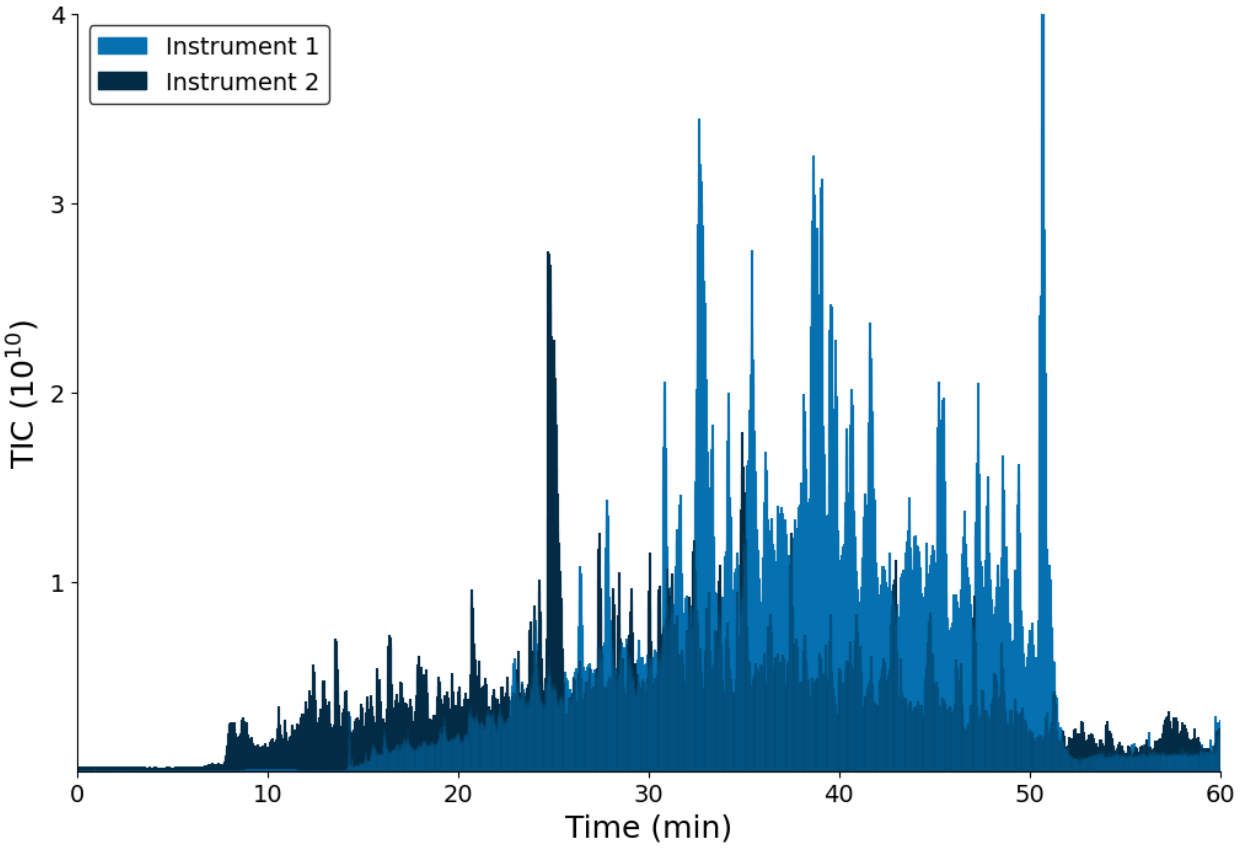


Figure 3.3: Example TICs for gradients used with Instrument 1 and Instrument 2.

and protein detections than quadrupole gas phase fractionation or the absence of gas-phase fractionation on both instruments (Figure 3.2). While the number of detections was similar within the triplicate of a method, there was poor overlap of identifications within the triplicate of a method and between the three methods due to the stochasticity of data dependent acquisition (Figures 3.4-3.5).

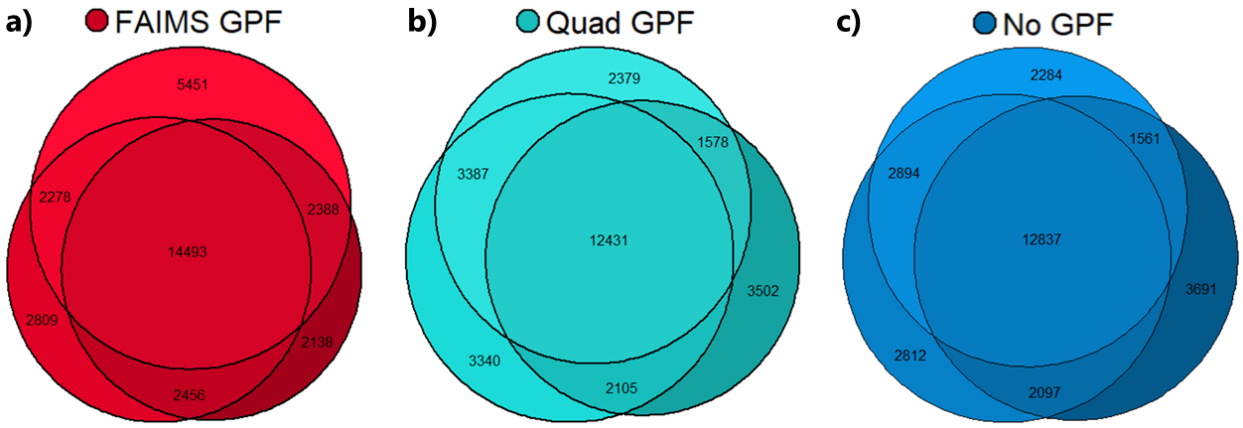


Figure 3.4: Overlap of identified peptides between the triplicate data dependent acquisition runs with FAIMS gas-phase fractionation (internal stepping), quadrupole gas-phase fractionation (internal stepping), and no gas-phase fractionation.

We questioned if the lower number of identifications by quadrupole gas phase fractionation was due to an underfilling of the ion trap, so we examined the ions measured in each method's MS2 spectra. We observed that there was a larger difference in ion filling between FAIMS and non-FAIMS LC-MS (Figure 3.6a) than observed previously by Hebert *et al.*[96] (Figure 3.7). To confirm that the results from our initial experiment on an Orbitrap Eclipse was not an artifact of a single instrument, we repeated the experiments on a second Orbitrap Eclipse. The data from Instrument 2 (Figure 3.6b) minimized these concerns because the data between the two tribrids were similar.

To assess if gas-phase fractionation enables the analysis of less abundant signals by excluding more abundant precursors from filling the ion trap, we examined the number of persistent peptide-like features found with each method. An advantage of quadrupole gas-phase

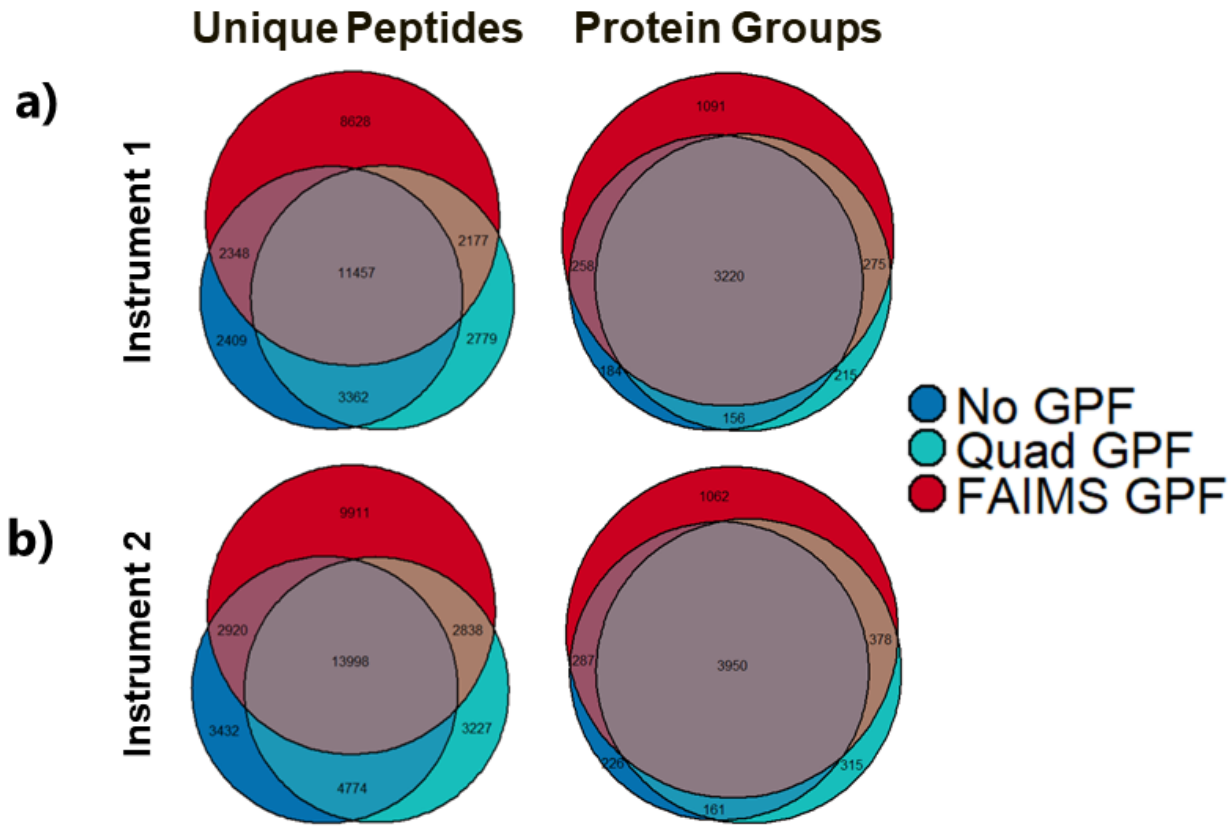


Figure 3.5: Overlap of identified peptides (left) and proteins (right) between FAIMS gas-phase fractionation, quadrupole gas-phase fractionation, or no gas-phase fractionation.

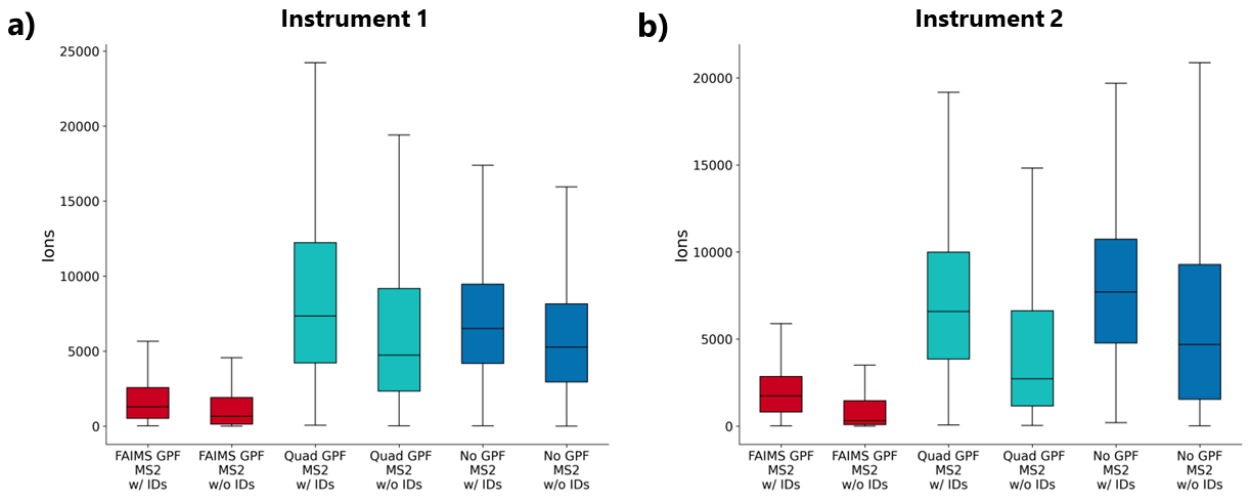


Figure 3.6: The number of ions measured in MS2 spectra for the respective LC-MS/MS runs. The number of ions were estimated by taking the total signal in each spectrum and multiplying by the ion injection time using the *msions* Python package. Boxplots demonstrate the median number of ions (middle line), extend from the first quartile to the third quartile (Q3 - Q1 = Interquartile Range or IQR), and have whiskers that extend to the minimum/maximum or 1.5 x IQR, whichever is less. The data show that FAIMS (red) underfills the ion trap more often than quadrupole gas-phase fractionation (teal) or normal LC-MS (blue). On average, the MS2 spectra with a larger number of ions are assigned a peptide identification with a low q-value at a higher frequency than spectra with fewer ions.

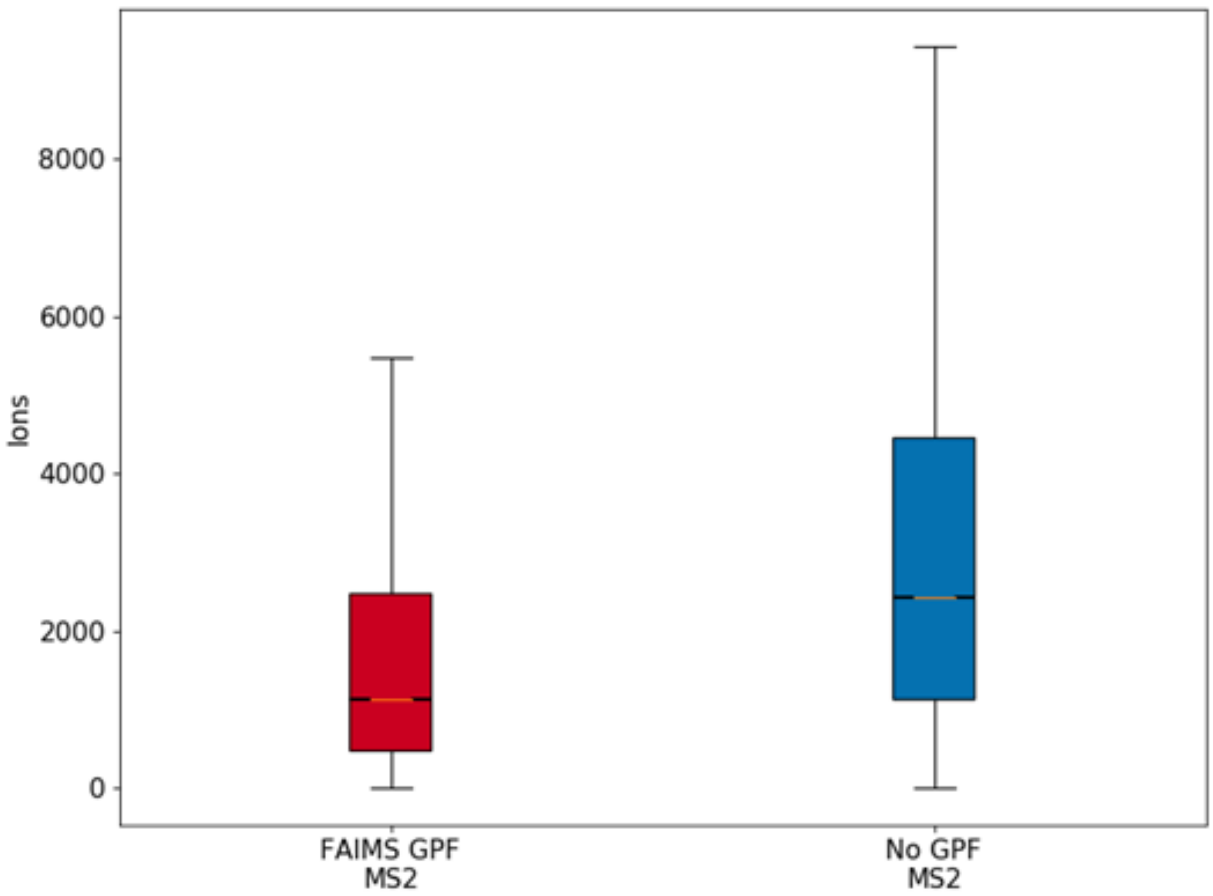


Figure 3.7: Ions injected for each MS2 spectra in Hebert *et al.*

fractionation is that, unlike FAIMS gas-phase fractionation, the MS1 features are distinct between fractions (Figure 3.8). This limited redundancy should increase the likelihood of selecting a low-intensity precursor for fragmentation.

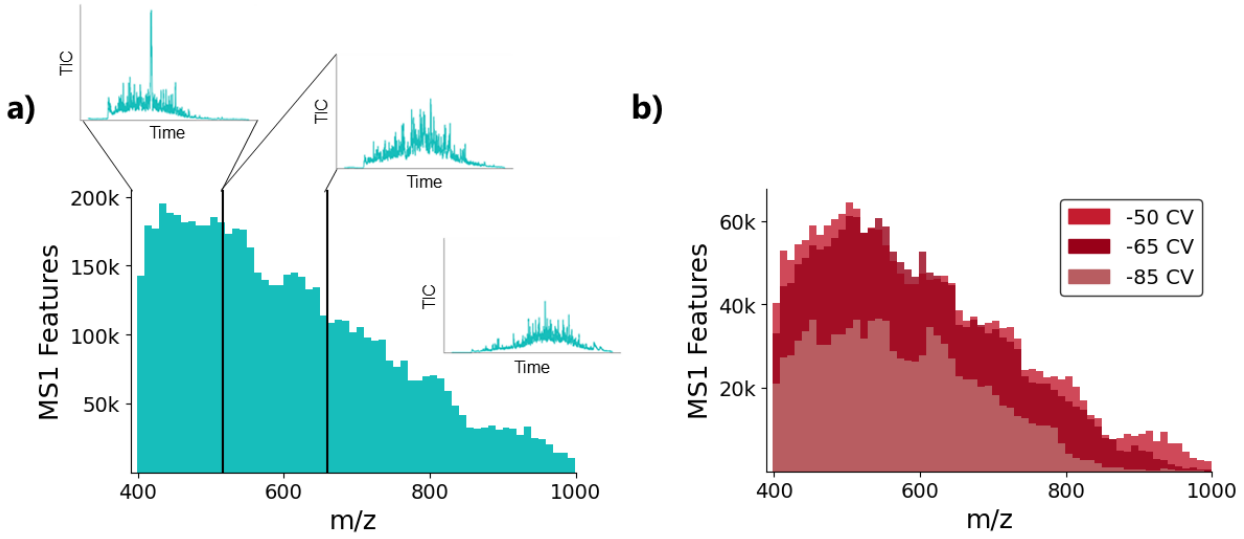


Figure 3.8: Overlap in m/z space for quadrupole (a) and FAIMS (b) gas phase fractionation.

The data collected using FAIMS had a larger percentage of low intensity MS1 precursors identified by MS/MS (Figure 3.9a), but the intensities of the features were also significantly lower (Figure 3.9b). This agreed with the underfilling of the ion trap shown in Figure 3.6. To determine how individual features were affected by the 3 methods, we examined the features that were identified by all 3 methods in one instrument batch. Compared to no gas-phase fractionation, FAIMS gas-phase fractionation lowered the intensity of over 90% of the shared identified features (Figure 3.9c) while quadrupole gas-phase fractionation increased the intensity of 90% of them (Figure 3.9d). These results support hypothesis 1 for quadrupole gas-phase fractionation, but not for FAIMS gas-phase fractionation.

To explore if the use of FAIMS reduces co-isolation of peptides during MS/MS and results in a reduction of chimeric spectra, we examined the number of persistent precursors present within quadrupole isolation windows. An isolation window is the narrow m/z range

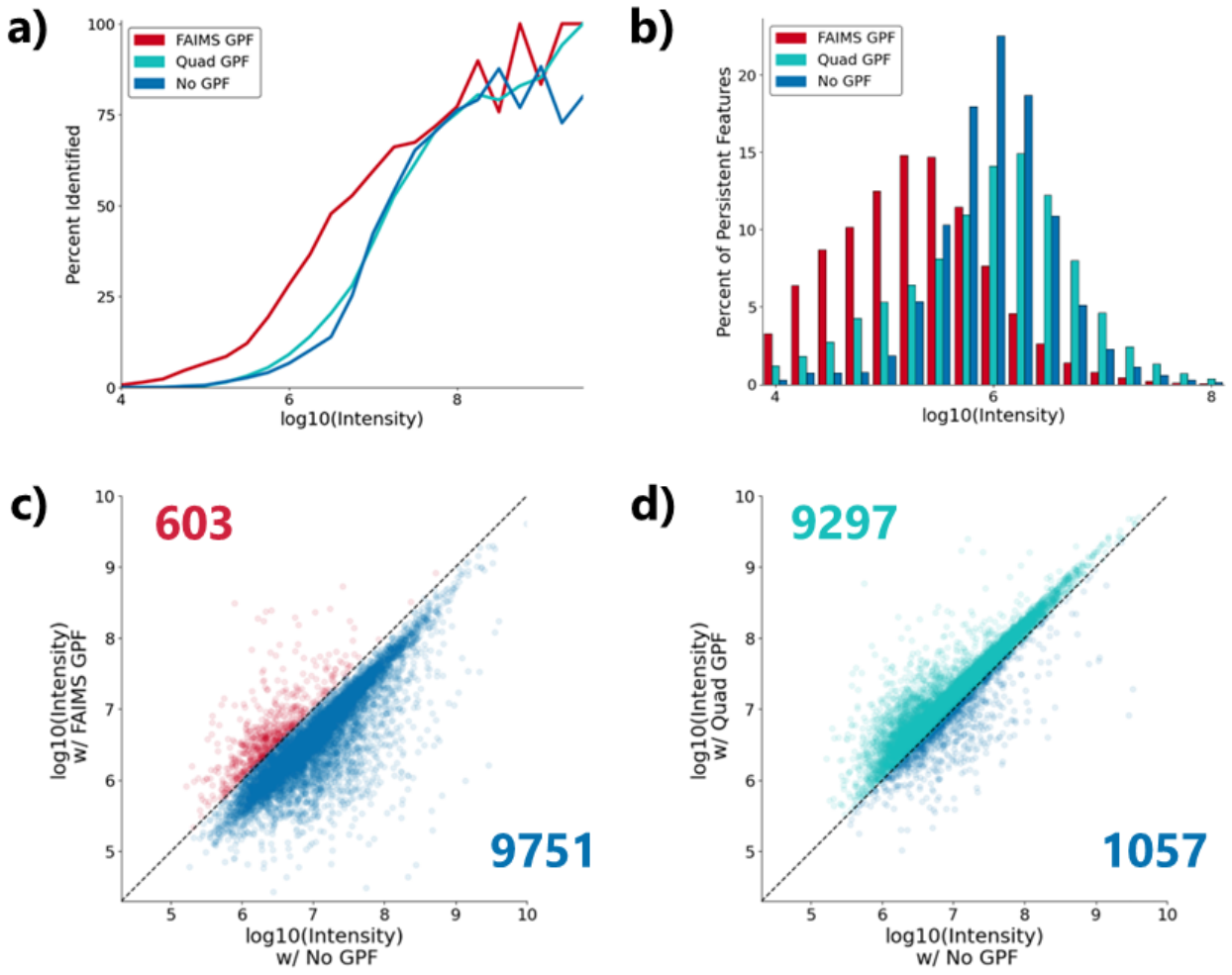


Figure 3.9: Persistent peptide-like features found with different gas-phase fractionation methods. Persistent peptide-like MS1 features were determined using Hardklor and Kronik. (a) The percent of features identified based on their apex intensities (log10). (b) The density of features in each intensity bin (log10). (c) Comparison of the apex intensities of the same identified peptide features in a FAIMS gas-phase fractionated experiment (red) and experiment without gas-phase fractionation (blue). (d) The apex intensities of the same identified features in a quadrupole gas fractionated experiment (teal) and experiment without gas-phase fractionation (blue).

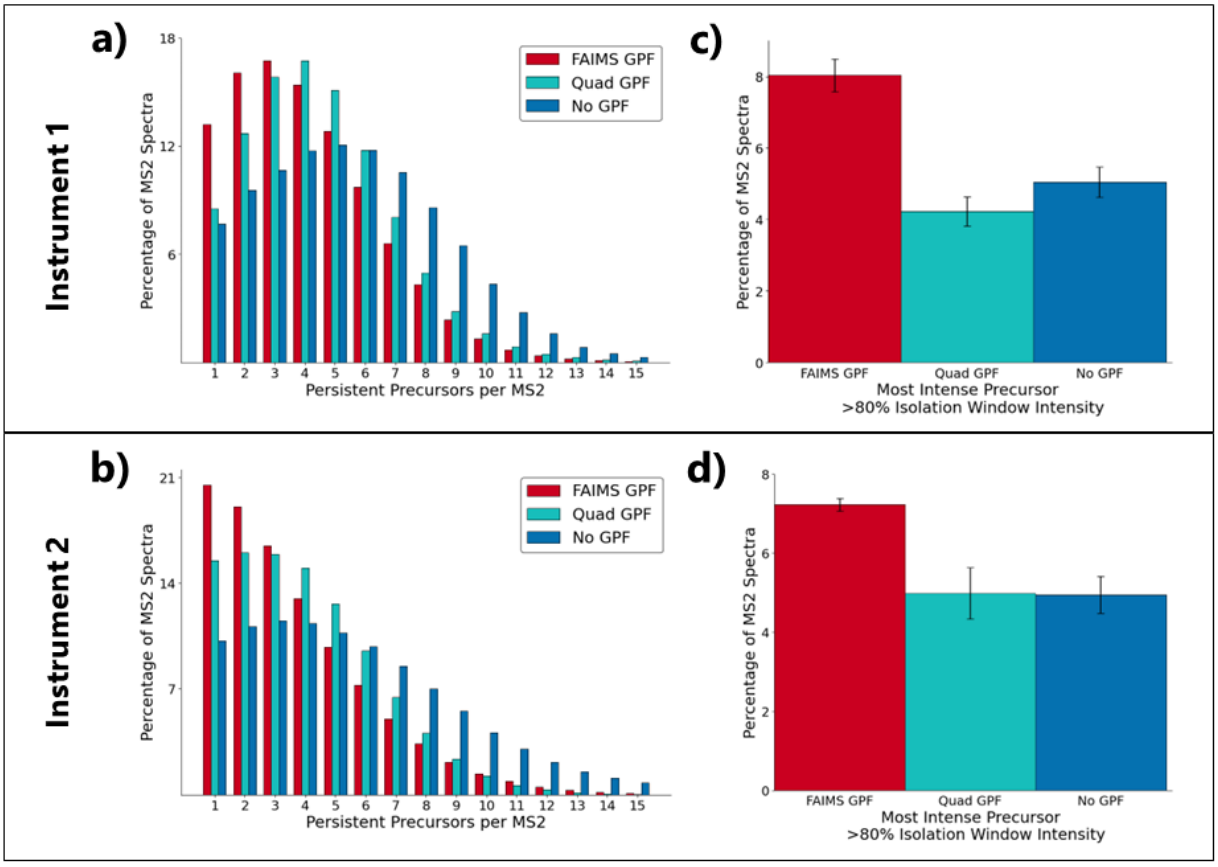


Figure 3.10: a. Percentage of MS2 spectra with multiple precursors for Instrument 1. b. Percentage of MS2 spectra with multiple precursors for Instrument 2. c. Percentage of MS2 spectra where the most intense precursor had >80% of the summed isolation window intensity for Instrument 1. d. Percentage of MS2 spectra where the most intense precursor had >80% of the summed isolation window intensity for Instrument 2. FAIMS gas-phase fractionation reduced the number of chimeric spectra compared to quadrupole gas-phase fractionation and no gas-phase fractionation and improved the relative intensity of the most intense precursor in spectra that were chimeric.

that is analyzed by the quadrupole during an MS2 spectrum acquisition. Non-chimeric spectra would only have one precursor in each MS2 spectrum and have a better chance of being identified the identification rate of chimeric spectra can be 2-fold lower than non-chimeric spectra.[102] Based on this criterion, FAIMS gas-phase fractionation successfully reduced chimeric spectra compared to quadrupole gas-phase fractionation and no gas-phase fractionation (Figure 3.10a & b). Quadrupole gas-phase fractionation also had fewer chimeric spectra than no gas-phase fractionation, but the improvement was smaller than with FAIMS. To investigate how FAIMS affects the chimeric spectra, we checked the relative intensities of the precursors in the isolation windows. On both instruments, the most intense precursor within an isolation window has a higher relative intensity for a higher percentage of FAIMS MS2 spectra (Figure 3.10c & d). This supports FAIMS improving the chances of identifying spectra that are chimeric in addition to reducing the number of chimeric spectra. We believe the different distributions seen between Instrument 1 and 2 could be explained by the sample and LC gradients being different. The data collected with Instrument 2 appeared to separate the peptides better across the elution time (Figure 3.3).

3.5 Conclusions

Overall, the results supported the use of FAIMS gas-phase fractionation over either quadrupole gas-phase fractionation or no gas-phase fractionation for this data acquisition setup and sample type. FAIMS reduced the co-isolation of persistent precursors during the MS/MS process, which resulted in a reduction of chimeric spectra, even though FAIMS gas-phase fractionation was less efficient at transmitting ions than quadrupole gas-phase fractionation. The lower ion transmission led to FAIMS lowering the intensity of over 90% of the shared identified features. We confirmed these results on two Thermo Eclipse Tribrid Mass Spectrometers while using two separate FAIMS interfaces, slightly different chromatography gradients, similar samples, and internal and external stepping. While internal stepping may not have a statistically significant improvement in identifications, the method should be chosen because the multiple runs needed for external stepping require more sample and additional time for

each gradient preparation, column equilibration, and sample loading.

Chapter 4

COMBINING IDENTIFIED PEPTIDES WITH PERSISTENT UNIDENTIFIED FEATURES IN A SKYLINE WORKFLOW

4.1 Summary

Proteomics generally focuses on the presence and abundance of peptides that have been identified by a database search. While that information is useful, there are cases where the features of interest are not easily identified (e.g., unexpected byproducts, abnormal cleavages, unknown modifications, sequence variants, glycans, etc.). To address this issue, our lab has developed a workflow that detects persistent MS1 features and enables the user to examine unidentified persistent features alongside features that have been assigned a peptide identity. Our pipeline can start from a RAW or converted file (e.g., mzML), detect and summarize the features, match features between runs, create aligned files, and output the transition list for import into Skyline. This workflow is a promising tool for multi-attribute method (MAM) applications.

4.2 Introduction

Proteomics frequently revolves around the presence and intensity of peptides that have been identified by a database search. While that information is useful, there are cases where the features of interest are not easily identified (e.g., unexpected byproducts, abnormal cleavages, unknown modifications, sequence variants, glycans, etc.). Those features would be lost in a discovery experiment, and potential features of interest are also missed while performing targeted experiments. New feature detection (NFD) or new peak detection (NPD) is the process of identifying and quantifying unknown or unannotated peaks.[69]

NFD workflows can be particularly useful as a multi-attribute method (MAM) applica-

tion. Multi-attribute method (MAM) is a quality control approach used while developing and manufacturing biopharmaceuticals.[70,103] The method uses mass spectrometry to assess the quality of a drug or biologic by analyzing multiple attributes, such as physicochemical properties, identity, and purity. One MAM function is to monitor a test sample by quantitatively measuring product quality attributes at the peptide level. A second function is to identify new and changing features by performing a differential analysis of a test sample versus a reference. Traditional analytical methods, including hydrophilic interaction liquid chromatography (HILIC) for glycan profiling, cation exchange chromatography (CEX) for charge variant analysis, and reduced capillary electrophoresis-sodium dodecyl sulfate (rCE-SDS) for clipped variant analysis, focus on single attributes, which prevents the ability to see an integrated view of the product quality. Many purity methods also use spectroscopic detection (e.g., UV-visible absorption), which may mask the presence of an impurty because the method can be easily confounded by co-eluting species. By analyzing multiple attributes simultaneously, MAM can help identify correlations and potential associations while monitoring batch-to-batch variability, identifying potential sources of variation, and optimizing the production process.

Mass spectrometry's use in MAM is well positioned for the biopharmaceutical industry because nearly 20% of pharmaceuticals are protein-based[104] and over 95% of biologics license applications (BLAs) already use mass spectrometry for protein or impurity characterization.[105] MS has previously been used for QC testing of molecular mass measurements for less complex products, such as small molecule and peptide drug products. However, protein therapeutic BLAs have typically not used MS for more detailed purity assessments or for quality control (QC) testing in Current Good Manufacturing Practice (cGMP) laboratories.

Recent advances in high resolution mass spectrometers and software for data analysis are promising for MAM because we can now distinguish between similar features and perform quantitative measurements. Our lab has developed a workflow that detects persistent MS1 features and enables the user to examine these features that have not been assigned a peptide identity. Hardklor,[75] Kronik,[100] BullseyeSharp, and Skyline[2] can be used to investigate these

unidentified features. In addition to its use for MAM, the workflow is also useful for general data quality control because it can demonstrate the amount of signal that is being identified (Figure 4.1). In the future, we intend to implement this pipeline directly within Skyline and use it to support MAM for the QC of therapeutic proteins.

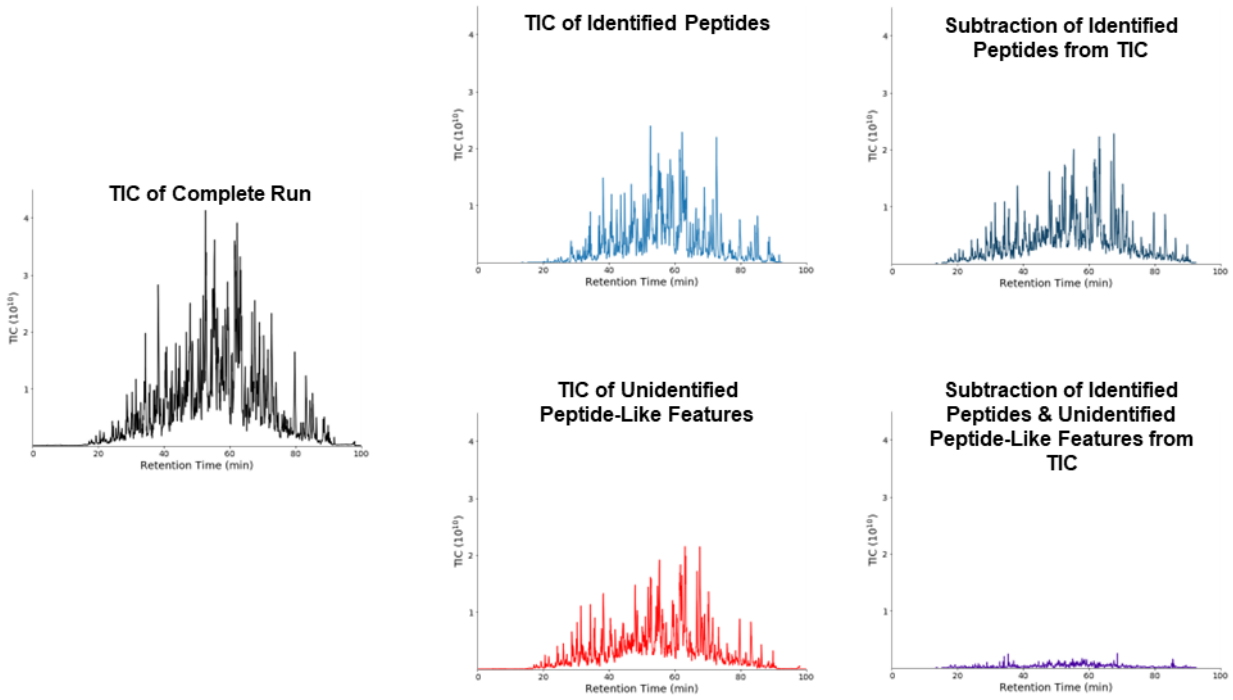


Figure 4.1: Total ion current (TIC) chromatograms for identified and unidentified peptide-like features. TIC chromatograms have TIC on the y-axis and retention time on the x-axis. Left - TIC for the complete run. Middle - TIC for identified peptides and unidentified peptide-like features. Right - Middle plots subtracted from left plot.

4.3 Methods

New feature detection requires multiple algorithms and statistical models to process mass spectrometry data. The first step is performing peak detection, which involves identifying the m/z values and intensities of signals in the spectra. Our workflow starts with Hardklor finding features on the MS1 level by analyzing peptide isotope distributions (Figure 4.2). The

workflow's feature detection algorithm has improved speed and can handle overlapping and in-phase isotope distributions. After analyzing each scan, the workflow outputs the mass, charge, intensity, and retention time of each feature in each scan. We have summarized these features to determine which features are persistent in multiple scans.

The next step is an optional alignment, which involves matching peaks across multiple runs and correcting for differences in retention time and mass accuracy. We use the N most intense persistent features from each run (e.g., 10,000 most intense persistent features) in an algorithm to match features between runs. Based on the retention time difference of features that were matched between runs, we train a support vector regression (SVR) to align the runs. The trained SVR is used to update feature retention times and create aligned mzMLs that can be imported into Skyline.

The output of the workflow is a list of detected features with their m/z values, retention times, intensities, and other parameters. The union of persistent features with updated retention times are exported into a transition list for import into Skyline. We have successfully imported our feature list and the aligned mzMLs into Skyline using our "Molecule interface" and our "Mixed interface." The updated retention times improve Skyline's peak picking accuracy for these unidentified features. For a case where users want to use the "Mixed interface" and have peptide transitions as well as unidentified features, we remove the features that match to an identified peptide transition. This gives a list of unique features, some of which are identified, and the rest are sorted by decreasing intensity and designated by their specific mass, charge, and retention time. We have created a pipeline that can start from a RAW or converted file (e.g., mzML), detect and summarize the features, match features between runs, create aligned files, and output the transition list for import into Skyline.

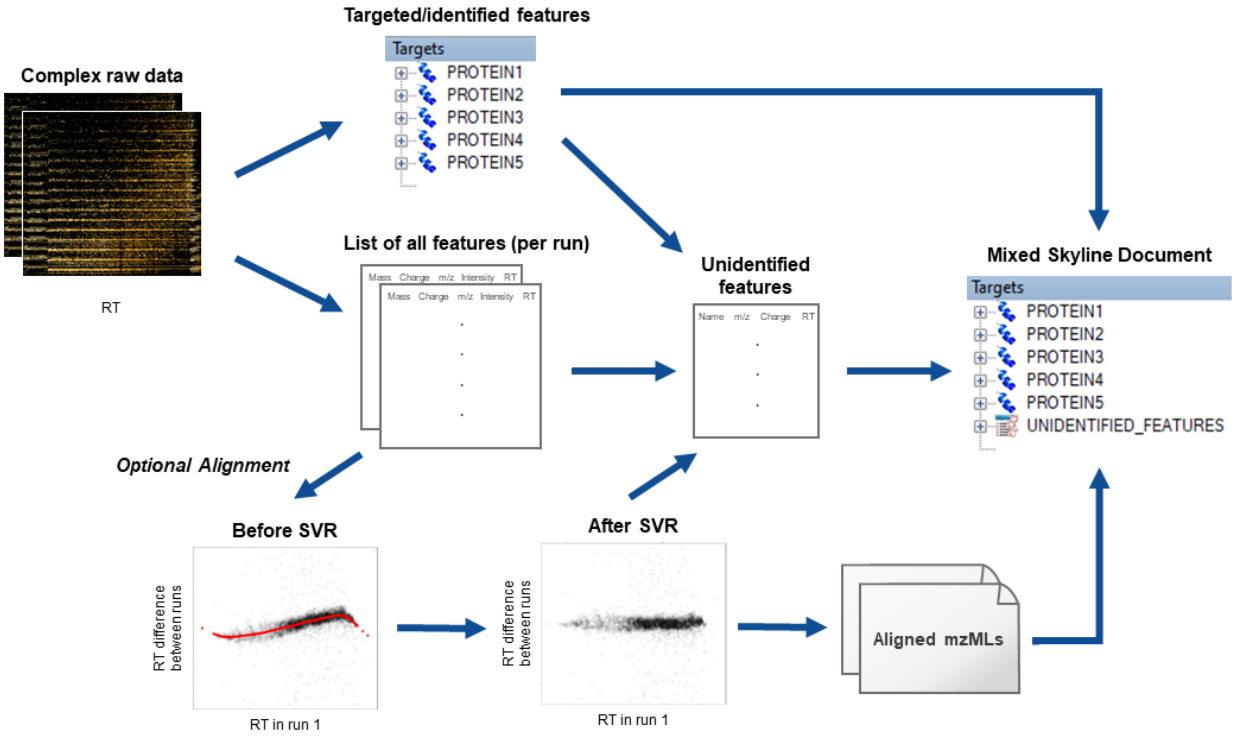


Figure 4.2: Mixed Skyline workflow diagram. Hardklor detects features on the MS1 level by analyzing peptide isotope distributions. Each feature is defined with a mass, charge, m/z, intensity, and retention time. These features are summarized with Kronik to determine which features are persistent across time. Optionally, we can align the runs by support vector regression. The trained support vector regression is used to update feature retention times and create aligned mzMLs that can be imported into Skyline. The features can be filtered by the N most intense persistent features, by the percentage of the total ion current (TIC) in a scan, or by an intensity threshold.

4.4 Results and Discussion

As a proof of concept, we ran replicates (n = 5) of a plasma sample with two concentrations of Pierce Peptide Retention Time Calibration (PRTC) Mixture and Bovine Serum Albumin (BSA) spiked in. The samples were provided as a mixture of peptides because the plasma proteins and BSA had been digested after PRTC and BSA were spiked in. PRTC is a known mixture of peptides and was not affected by the digestion. At the time of analysis, we were blind to the relative concentrations. Because the LC-MS data were acquired within a short time frame, we did not perform an alignment for this proof of concept experiment. The workflow was completed with a database containing the PRTC and BSA peptides (Figure 4.3a) and a second database that was missing 1 BSA peptide (Figure 4.3b). In *b*, the missing peptide (LVNELTEFAK) was found to be the most intense unidentified feature by the workflow. We also examined the relative concentrations of the targeted peptides by annotating each file with its group and creating a "Group Comparison" in Skyline (Figure 4.3c). Spiked-in peptides clustered around a log2 fold change of 2, which would suggest the concentration of spike-in 1 (PRTC1) was four times the concentration of spike-in 2 (PRTC2). This result agreed with the relative concentrations when the blind was removed, which demonstrates that changes in abundance of unidentified features can also be flagged using the workflow.

After demonstrating the proof of concept, we applied the workflow to a set of files we received from the National Institute of Standards and Technology (NIST). The data had been collected by the MAM Consortium in an interlaboratory study to evaluate new peak detection performance metrics.[69] The MAM Consortium was formed to bring together government agencies, biopharmaceutical companies, manufacturing organizations, and mass spectrometer, software, and reagent vendors to share MAM knowledge and experience. The Consortium wants to utilize MAM as a QC approach and potential replacement for multiple single-attribute assays. For the interlaboratory study, the MAM Consortium, in collaboration with NIST, the Institute for Bioscience and Biotechnology Research, and Just-Evotec Biologics, initiated an NPD round robin (NPDRR) to survey the current performance metrics

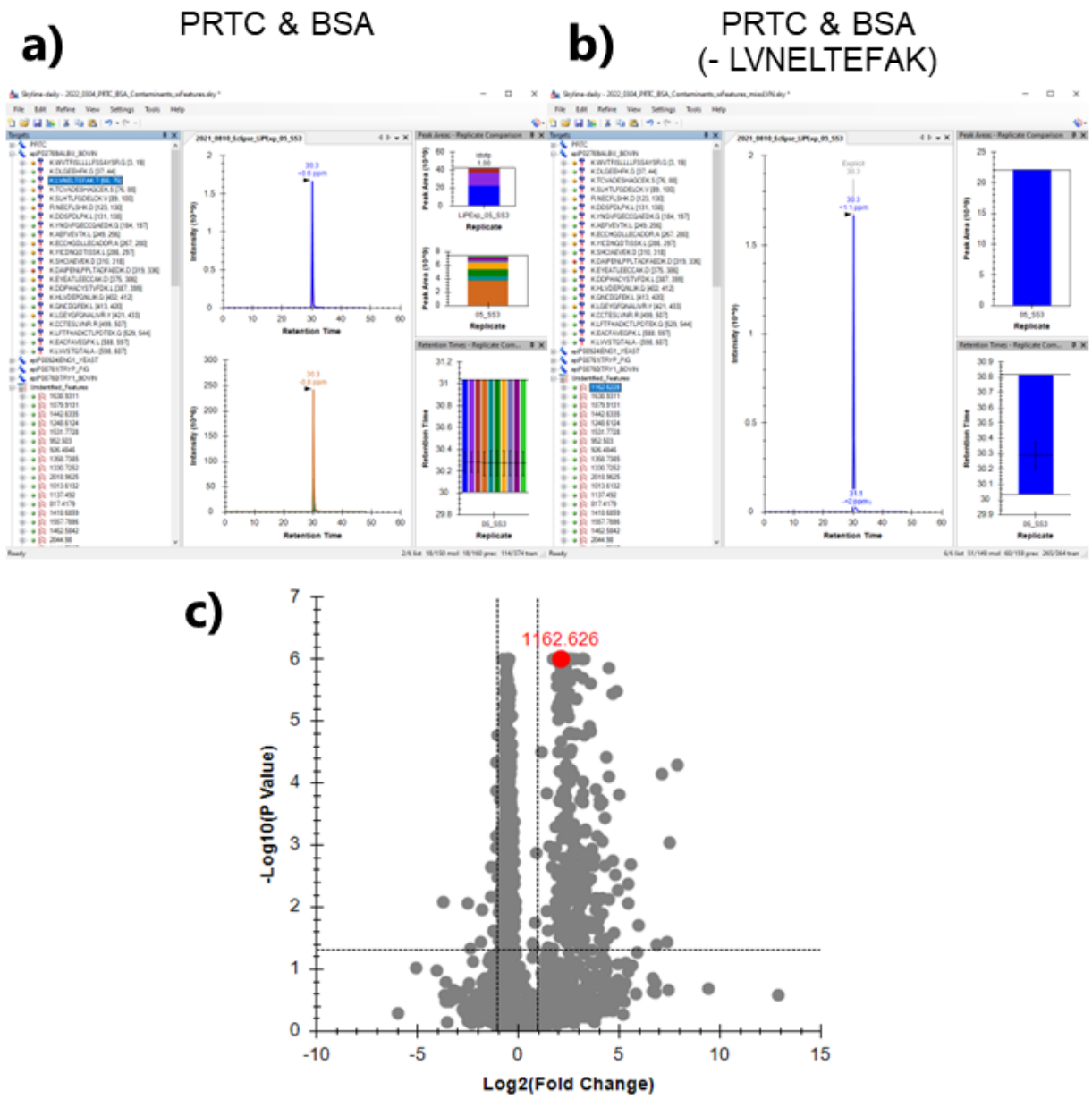


Figure 4.3: PRTC proof of concept with present and missing targeted peptide. Workflow was given the same run and the same target list except (b) was missing one peptide (LVNELTEFAK) compared to (a). The missing peptide was found to be the most intense unidentified feature by the workflow. (c) Volcano plot of spike-in samples (n = 5 for PRTC1 & PRTC2). Two samples were prepared (spike-in PRTC1 = 4 x spike-in PRTC2). The two samples were analyzed 5 times. Spiked-in peptides were around a log2 fold change of 2 as expected.

of the platform and provide insight into developing the platform in other laboratories.

The NPDRR included technical triplicates of a Blank, technical triplicates of a Calibration Sample containing PRTC, and single runs for the other samples. The other samples were the Reference, pH, Spike, and Unknown samples. The NPDRR used the NIST mAb RM 8671 monoclonal antibody as the Reference sample (Figure 4.4).[107–111] The pH sample was the Reference sample that had undergone a pH stress, and the Spike sample contained the mAb and PRTC peptides. The samples that were not analyzed in triplicate had one run with only MS1 acquired and a separate run with MS1 and MS2 acquired. For this example, we used BLK_01, CAL_01, and the MS1 runs for the Reference, pH, Spike, and Unknown samples in the "MAM_Data_01" dataset. For the following figures, The order for the top row of samples is Blank, Calibration, and Reference, and the bottom row is Unknown, Spike, and pH.

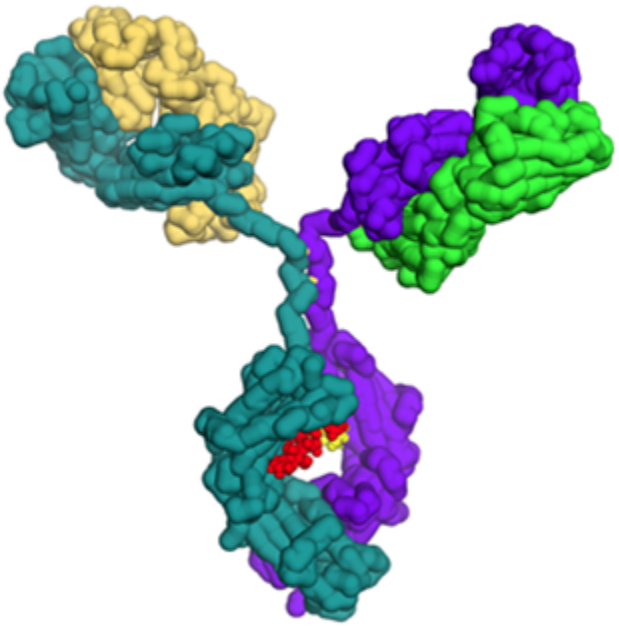


Figure 4.4: National Institute of Standards and Technology (NIST) mAb RM 8671.[106]

Our first check was whether the PRTC peptides were present in the correct samples (Figure 4.5). As expected, the PRTC peaks only appeared in the Calibration and Spike samples, and the amount in the Spike sample was less than the Calibration sample. Second, we examined the presence of trypsin in the samples (Figure 4.6). Trypsin peptides were present in the Reference, Unknown, Spike, and pH samples, which suggested all 4 had proteins that needed to be digested. This was known for the Reference, Spike, and pH samples, but not known for the Unknown sample. The lack of trypsin in the Calibration sample is expected because the Calibration sample contains PRTC peptides, not proteins. Third, we analyzed

the presence of NIST mAb RM 8671 peptides in the samples (Figure 4.7). The mAb peptides were present in the Reference, Spike, and pH samples as expected and were also in the Unknown sample.

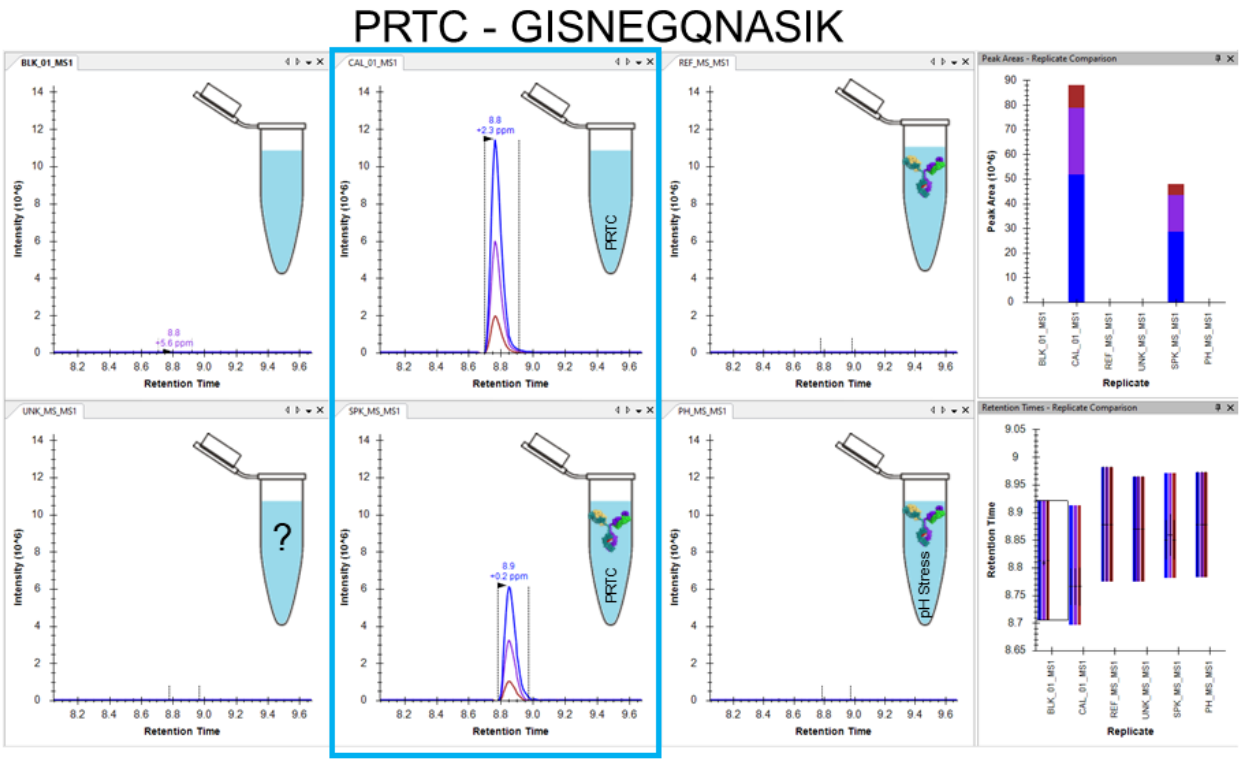


Figure 4.5: MAM Consortium samples with PRTC present. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of the GISNEGQNASIK PRTC peptide in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue box is around the two samples with GISNEGQNASIK peptide peaks. The peaks agree with the identities of the samples because the Calibration and Spike samples contained PRTC.

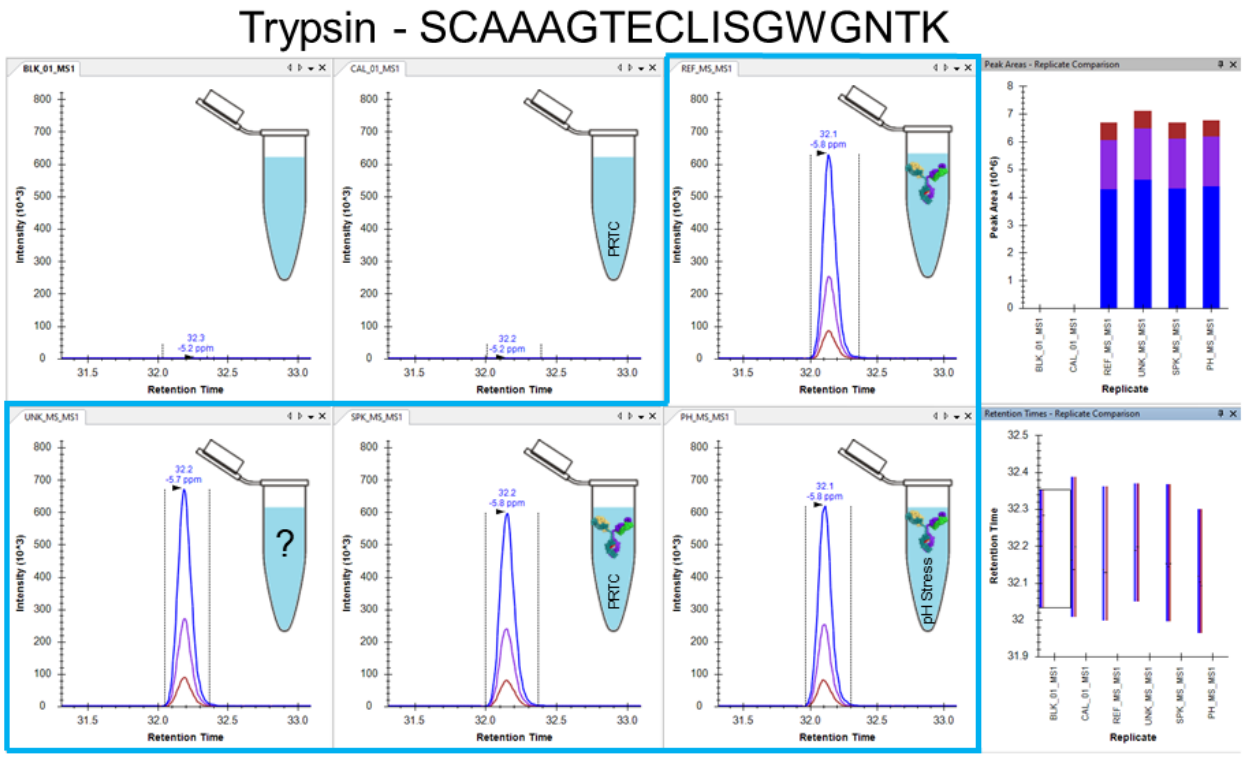


Figure 4.6: MAM Consortium samples with trypsin peptides present. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of the SCAAAGTECLISGWGNTK tryptic peptide in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue outline is around the four samples with SCAAAGTECLISGWGNTK peptide peaks. Trypsin peptides can be seen in the Reference, Unknown, Spike, and pH samples, which suggest all 4 had proteins that needed to be digested.

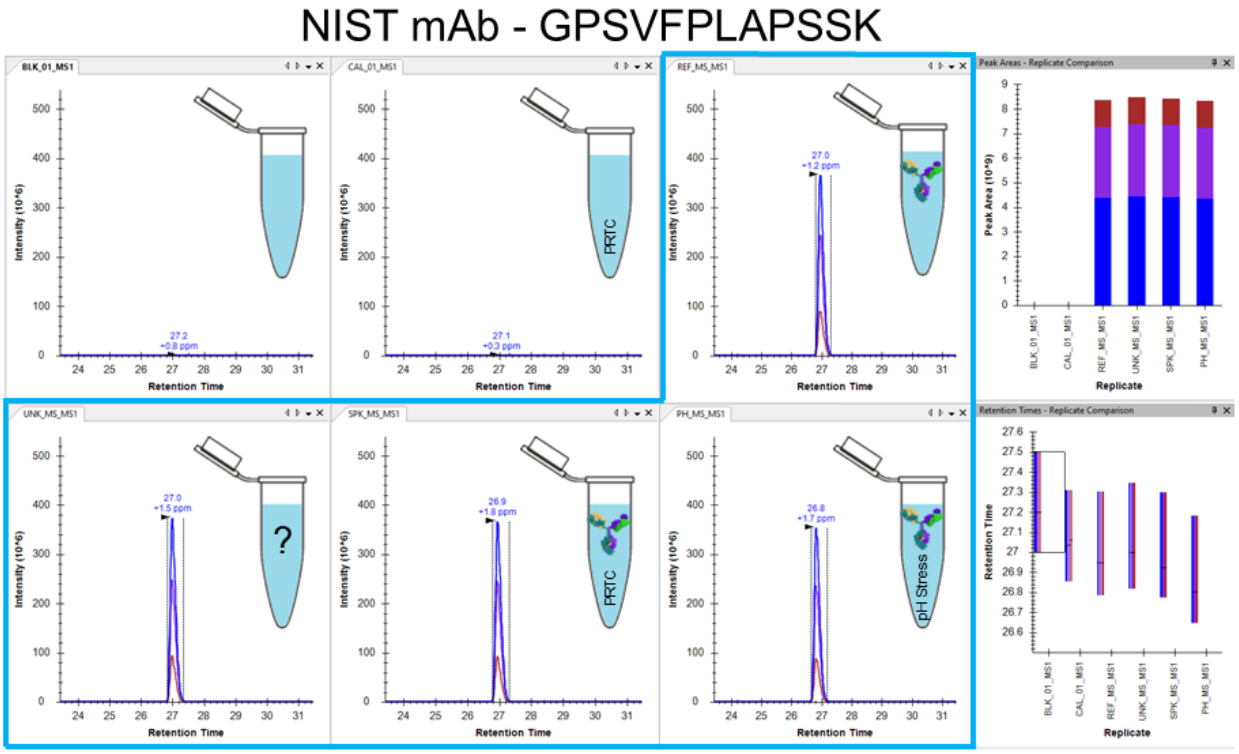


Figure 4.7: NIST mAb peptide example. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of the GPSVFPLAPSSK mAb RM 8671 peptide in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue outline is around the four samples with GPSVFPLAPSSK peptide peaks. mAb peptides can be seen in the Reference, Unknown, Spike, and pH samples, which suggest all 4 had RM 8671 present.

After checking the known features, we examined the unidentified features. Many of the unidentified features had peak areas that were unchanged under pH stress (Figure 4.8). The unchanged feature example in Figure 4.8 must have been related to the mAb protein because the feature was present in the same samples as the mAb peptides in Figure 4.7. Unchanged features can generally be ignored as long as they are consistent. For MAM, the features of interest should be changing, new, or missing. Unidentified features with changing peak areas can be determined from observation, by a threshold, or with a statistical test. Since this data did not have replicates to provide statistical significance, the peak areas were checked manually with Skyline. A changing peak area is seen in Figure 4.9 because the Reference, Unknown, and Spike had low levels of the feature, but the pH sample's peak area increased by 6x.

Providing the mass, charge, and retention time of features can be useful for unidentified features with similar attributes. For example, masses can be within a tolerance (2544.1111 vs 2544.1127), but the retention time (40.1 min / 39.1 min) can distinguish between the two features (Figure 4.10). A new peak can also be an intriguing discovery (Figure 4.11). New peaks should to be characterized in a MAM platform because the new peaks could be potential process- or product-related impurities that could impact the efficacy, pharmacokinetics, and safety of the biopharmaceutical.

In addition to monitoring new and changing peaks, analyzing the Unknown sample through these examples was an interesting challenge. Based on the analyses, we believed the Unknown was another Reference sample because there were no PRTC peptides present (Figure 4.5), but the sample did contain trypsin and mAb peptides (Figures 4.6-4.7). The sample did not undergo pH stress because the changing peaks had similar peak areas as the Reference sample (Figures 4.9-4.11). Our hypothesis of the Unknown sample's identity was confirmed by the coordinators of the NPDRR.

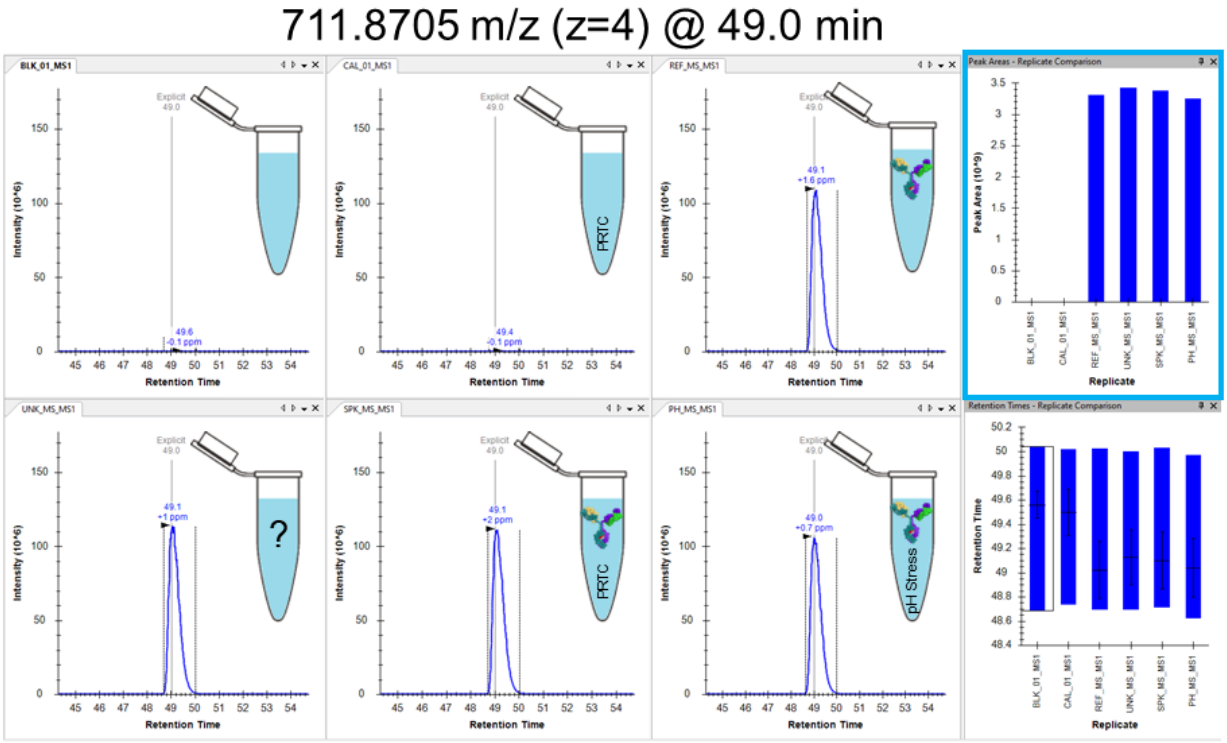


Figure 4.8: Unidentified feature with unchanged peak area. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of an unidentified feature (711.8705 m/z) in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue box shows the unidentified feature peak area is unchanged for the Reference, Unknown, Spike, and pH samples.

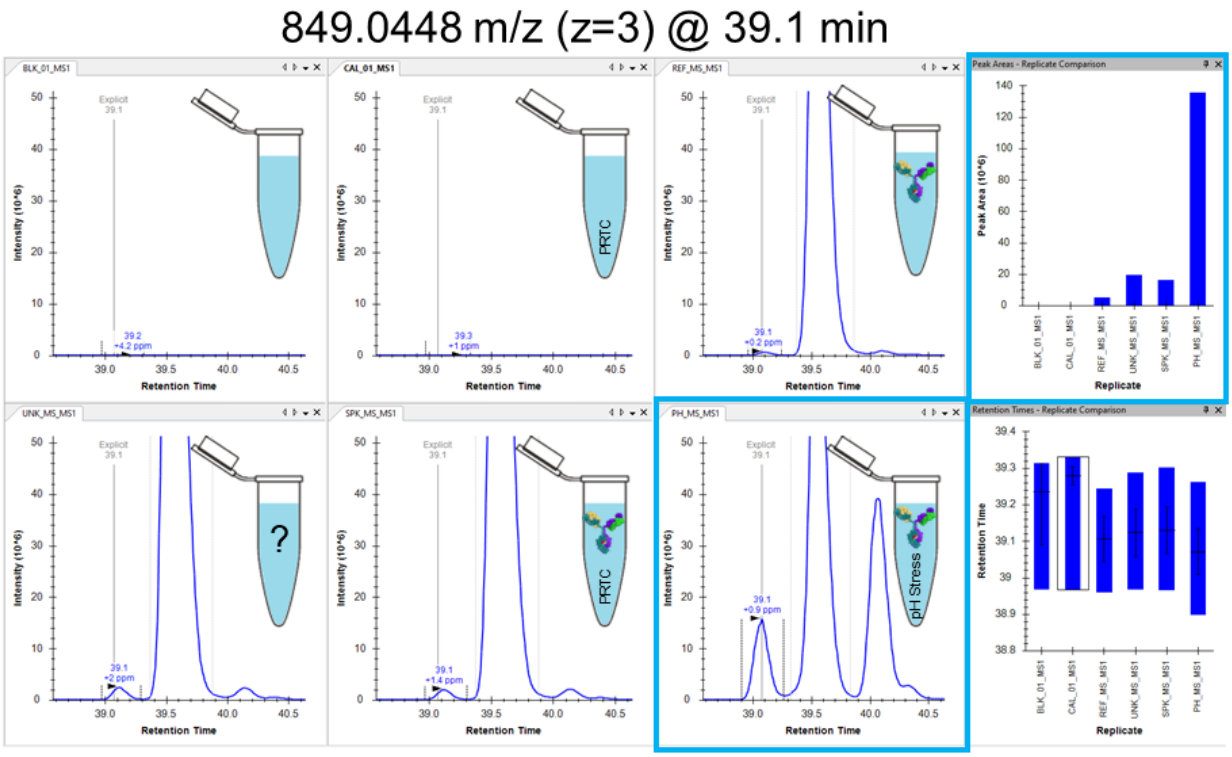


Figure 4.9: Unidentified feature with changing peak area. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of an unidentified feature (849.0448 m/z) in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue boxes show that the unidentified feature peak area is changing for the pH samples.

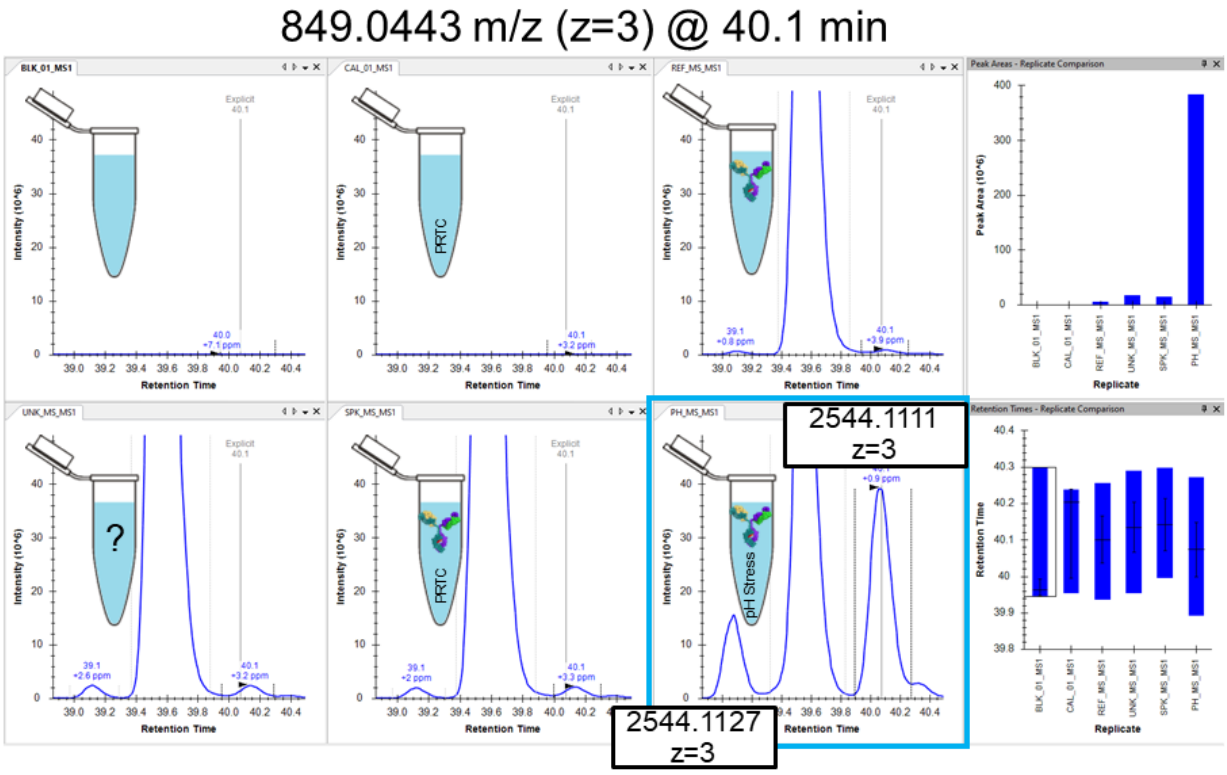


Figure 4.10: Unidentified features with similar mass. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of an unidentified feature (849.0443 m/z) in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue box shows that unidentified features can be distinguished by their retention times (39.1 and 40.1 min) when their mass and charge are similar.

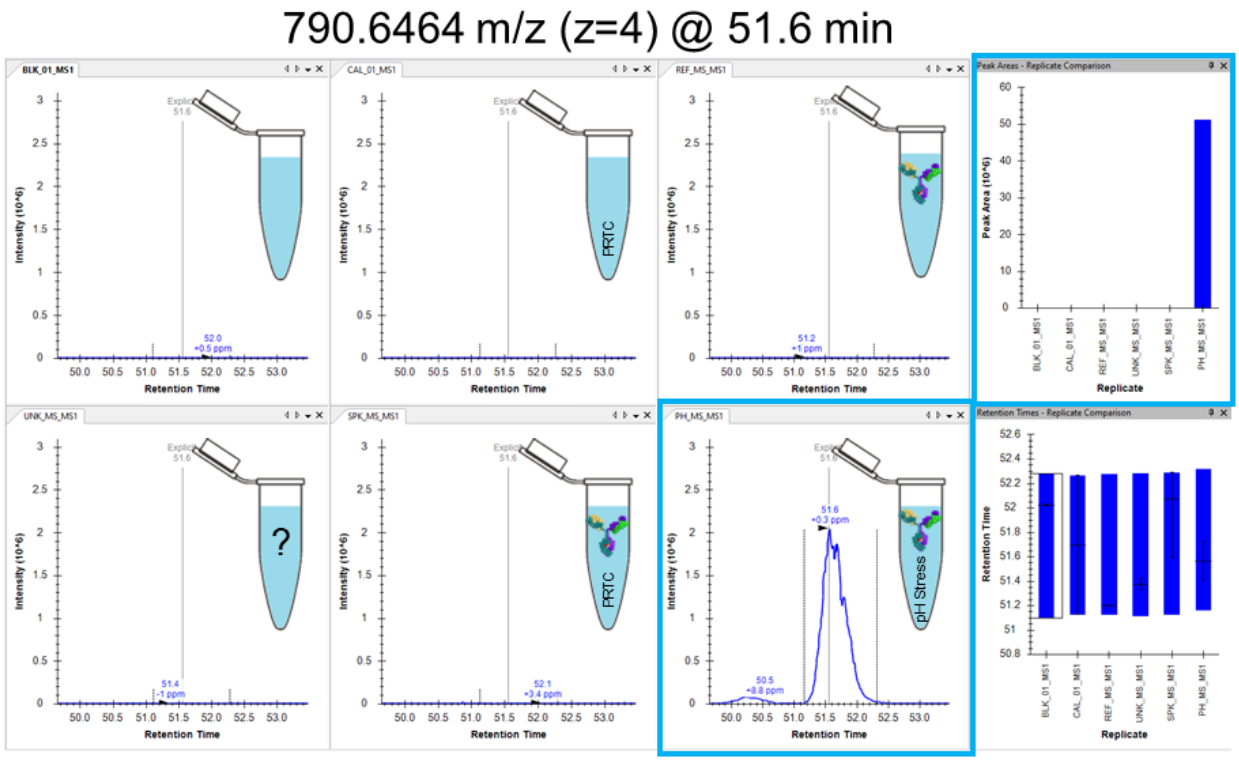


Figure 4.11: New unidentified feature caused by pH stress. The top row of samples is Blank, Calibration, and Reference (L to R). The bottom row is Unknown, Spike, and pH (L to R). On the far right, the top box shows the peak areas of an unidentified feature (790.6464 m/z) in the six samples, and the bottom box shows the retention times of the integrated peaks. The blue boxes show that a new peak can be caused by pH stress.

4.5 Conclusions

In this chapter, we demonstrated that we have developed a workflow that detects persistent MS1 features and enables users to examine unidentified persistent features alongside features that have been assigned a peptide identity. Our pipeline can start from a RAW or converted file (e.g., mzML), detect and summarize the features, match features between runs, create aligned files, and output the transition list for import into Skyline. The unidentified features can be investigated and annotated to improve the knowledge of the sample. This workflow is a promising tool for multi-attribute method (MAM) and other new feature detection (NFD) applications.

New feature detection can help uncover unexpected details of complex biological systems, including proteoforms, protein-protein interactions, metabolic responses, and impurities. Because of this outlook, MAM will be a powerful tool in the fields of disease diagnosis, drug discovery, and the development and quality control of biopharmaceuticals. However, one challenge with MAM is interlaboratory non-conformity due to less optimal instrument maintenance and best practices. To address this challenge, laboratories should consider utilizing system suitability controls that monitor NFD parameters for a given process or product. Current technologies are sufficient for implementing MAM in a regulated environment, but continued refinement of best practices for data acquisition and processing is still needed.

Chapter 5

CLOSING REMARKS

5.1 Research Conclusions

In summary, these chapters complement the traditional proteomics field by providing insight into the "Hidden Proteome"–the peptide-like features that are not identified. In Chapter 1, a broad overview of the field is narrowed to provide context for the following chapters. In Chapter 2, I presented *msions* and demonstrated use cases on how *msions* can help investigate the quality of MS data. The *msions* package is a useful tool for detecting issues in data and provides an opportunity to investigate the "Hidden Proteome." In Chapter 3, I explored the potential mechanisms for improved results via high field asymmetric waveform ion mobility spectrometry (FAIMS). Ion mobility techniques also provide insight into the "Hidden Proteome" because features change relative intensities and the additional partially orthogonal separation can assist with characterization of samples. In Chapter 4, I demonstrated that I developed a workflow that detects persistent MS1 features and enables users to examine unidentified persistent features alongside features that have been assigned a peptide identity. These "Hidden Proteome" features can be important in many different scenarios, such as Multi-Attribute Method (MAM) utilization for quality control of biopharmaceuticals or potential analysis of proteoforms, protein-protein interactions, metabolic responses to stimuli, and impurities. Overall, these packages, workflows, and tools will be most beneficial to scientists interested in quality control and developing a deeper understanding of unidentified features. Since *msions* is being implemented in Limelight[1] and the MAM project is being integrated with Skyline,[2] the work should have a friendlier user interface and a broader reach.

5.2 Looking Forward

During my PhD, I became focused on reproducibility and accessibility of research and data. One challenge with reproducibility is even the definition does not appear to be consistent from study to study. In Barba (2019), reproducibility and replicability were found to have 3 categories of usage: 1) The terms were used with no distinction between them. 2) Reproducibility is using the original researcher's data and code to regenerate the results while replicability is arriving at the same conclusion with new data. 3) Reproducibility is arriving at the same results with different data and methods while replicability is using the original author's data and code to arrive at the same results. Categories 2 and 3 are in direct opposition of each other, which leads to confusion.

I appreciate the definitions used in McArthur (2019).[112] In this article, repeatability refers to the same team using the same experimental setup to recreate the results. Replicability refers to a different team with the same experimental setup recreating the results. Reproducibility refers to a different team using a different experimental setup to come to the same conclusion. While repeatability and replicability are important, I believe reproducibility gives the best support for conclusions that are being made. I would appreciate if the scientific community would acknowledge the confusion and come to a consensus because a shared vocabulary would help progress us towards more reproducible science.

With regards to the field of proteomics, I am excited to see the advancements and breakthroughs that will happen in the upcoming years. I believe there could be many opportunities for advancement in technology, automation, and bioinformatics. Two major areas of immediate focus appear to be single-cell proteomics and the integration of multi-omics approaches. Analyzing the protein expression at the single-cell level will allow scientists to identify rare cell populations, elucidate cell heterogeneity, and investigate cellular responses to stimuli and diseases. This advancement will increase our understanding of cellular dynamics and enable personalized medicine approaches. However, the dynamic range and coverage will need to be considered as mentioned in Chapter 2.3. Integrating multi-omics approaches will

also improve our understanding of biological systems. By combining data from different omics levels, researchers can gain deeper insights into the interactions and regulations that govern cellular processes. Proteomics data has consistently been combined with genomics and transcriptomics data. Going forward, I believe we need to find a way to process and integrate additional omics data, such as metabolomics, lipidomics, glycomics, etc.

Another emerging area of interest is artificial intelligence (AI). Machine learning and AI algorithms have already begun to enhance data analysis and interpretation, and Generative AI has the potential to disrupt reading and writing scientific articles. I believe for at least a few years an expert will be needed to monitor the output because some summaries are correct while others are false. After that time period, we may have a chicken-and-the-egg problem because people will be learning from generative AI, which will be learning from us, and the upcoming experts may be unaware if part of an answer is incorrect. The circular feedback of the tool may be difficult to overcome, and I am curious to see how we address that challenge.

BIBLIOGRAPHY

1. Michael Riffle, Michael R. Hoopmann, Daniel Jaschob, Guo Zhong, Robert L. Moritz, Michael J. MacCoss, Trisha N. Davis, Nina Isoherranen, and Alex Zelter. Discovery and Visualization of Uncharacterized Drug-Protein Adducts Using Mass Spectrometry. *Analytical Chemistry*, 94(8):3501–3509, March 2022.

2. Lindsay K. Pino, Brian C. Searle, James G. Bollinger, Brook Nunn, Brendan MacLean, and Michael J. MacCoss. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrometry Reviews*, 39(3):229–244, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mas.21540.

3. Jennifer Griffiths. A Brief History of Mass Spectrometry. *Analytical Chemistry*, 80(15):5678–5683, August 2008.

4. Fred W. McLafferty. A Century of Progress in Molecular Mass Spectrometry. *Annual Review of Analytical Chemistry*, 4(1):1–22, 2011. _eprint: https://doi.org/10.1146/annurev-anchem-061010-114018.

5. Patricia Kahn. From Genome to Proteome: Looking at a Cell's Proteins. *Science*, 270(5235):369–370, October 1995. Publisher: American Association for the Advancement of Science.

6. Marc R. Wilkins, Christian Pasquali, Ron D. Appel, Keli Ou, Olivier Golaz, Jean-Charles Sanchez, Jun X. Yan, Andrew A. Gooley, Graham Hughes, Ian Humphery-Smith, Keith L. Williams, and Denis F. Hochstrasser. From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Bio/Technology*, 14(1):61–65, January 1996. Number: 1 Publisher: Nature Publishing Group.

7. Peter James. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly Reviews of Biophysics*, 30(4):279–331, November 1997. Publisher: Cambridge University Press.

8. Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003. Number: 6928 Publisher: Nature Publishing Group.

9. Yaoyang Zhang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical reviews*, 113(4):2343–2394, April 2013.

10. John R. Yates, Cristian I. Ruse, and Aleksey Nakorchevsky. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering*, 11(1):49–79, 2009. _eprint: https://doi.org/10.1146/annurev-bioeng-061008-124934.

11. John R. Yates III. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, 33(1):1–19, 1998. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-9888%28199801%2933%3A1%3C1%3A%3AAID-JMS624%3E3.0.CO%3B2-9.

12. John R. Yates. Mass Spectral Analysis in Proteomics. *Annual Review of Biophysics and Biomolecular Structure*, 33(1):297–316, 2004. _eprint: https://doi.org/10.1146/annurev.biophys.33.111502.082538.

13. Dirk A. Wolters, Michael P. Washburn, and John R. Yates. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Analytical Chemistry*, 73(23):5683–5690, December 2001. Publisher: American Chemical Society.

14. John B. Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246(4926):64–71, 1989. Publisher: American Association for the Advancement of Science.

15. Matthias Wilm and Matthias Mann. Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry*, 68(1):1–8, January 1996. Publisher: American Chemical Society.

16. Shibdas Banerjee and Shyamalava Mazumdar. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry*, 2012:e282574, March 2012. Publisher: Hindawi.

17. Olga E Petrova and Karin Sauer. High-performance liquid chromatography (HPLC)-based detection and quantitation of cellular c-di-GMP. *Methods in molecular biology (Clifton, N.J.)*, 1657:33–43, 2017.

18. Colin T. Mant, Yuxin Chen, Zhe Yan, Traian V. Popa, James M. Kovacs, Janine B. Mills, Brian P. Tripet, and Robert S. Hodges. HPLC Analysis and Purification of Peptides. *Peptide Characterization and Application Protocols*, 386:3–55, 2007.

19. Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011. Number: 7347 Publisher: Nature Publishing Group.

20. Mikhail E. Belov, Rui Zhang, Eric F. Strittmatter, David C. Prior, Keqi Tang, and Richard D. Smith. Automated Gain Control and Internal Calibration with External Ion Accumulation Capillary Liquid Chromatography-Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance. *Analytical Chemistry*, 75(16):4195–4205, August 2003. Publisher: American Chemical Society.

21. Mikhail E. Belov, Vsevolod S. Rakov, Eugene N. Nikolaev, Michael B. Goshe, Gordon A. Anderson, and Richard D. Smith. Initial implementation of external accumulation liquid chromatography/electrospray ionization Fourier transform ion cyclotron resonance with automated gain control. *Rapid Communications in Mass Spectrometry*, 17(7):627–636, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.955.

22. Anastasia Kalli, Geoffrey T. Smith, Michael J Sweredoski, and Sonja Hess. Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. *Journal of proteome research*, 12(7):3071–3086, July 2013.

23. Lisa E. Kilpatrick and Eric L. Kilpatrick. Optimizing High-Resolution Mass Spectrometry for the Identification of Low-Abundance Post-Translational Modifications of Intact Proteins. *Journal of Proteome Research*, 16(9):3255–3265, September 2017. Publisher: American Chemical Society.

24. Anastasia Kalli and Sonja Hess. Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *PROTEOMICS*, 12(1):21–31, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201100464.

25. Mikhail E. Belov, Gordon A. Anderson, Nicolas H. Angell, Yufeng Shen, Nikola Tolic, Harold R. Udseth, and Richard D. Smith. Dynamic Range Expansion Applied to Mass Spectrometry Based on Data-Dependent Selective Ion Ejection in Capillary Liquid Chromatography Fourier Transform Ion Cyclotron Resonance for Enhanced Proteome Characterization. *Analytical Chemistry*, 73(21):5052–5060, November 2001.

26. Alexander Scherl, Scott A. Shaffer, Gregory K. Taylor, Hemantha D. Kulasekara, Samuel I. Miller, and David R. Goodlett. Genome-Specific Gas-Phase Fractionation

Strategy for Improved Shotgun Proteomic Profiling of Proteotypic Peptides. *Analytical Chemistry*, 80(4):1182–1191, February 2008.

27. R. Guevremont, D. A. Barnett, R. W. Purves, and J. Vandermey. Analysis of a tryptic digest of pig hemoglobin using ESI-FAIMS-MS. *Analytical Chemistry*, 72(19):4577–4584, October 2000.

28. David A. Barnett, Barbara Ells, Roger Guevremont, and Randy W. Purves. Application of ESI-FAIMS-MS to the analysis of tryptic peptides. *Journal of the American Society for Mass Spectrometry*, 13(11):1282–1291, November 2002.

29. David A. Barnett, Luyi Ding, Barbara Ells, Randy W. Purves, and Roger Guevremont. Tandem mass spectra of tryptic peptides at signal-to-background ratios approaching unity using electrospray ionization high-field asymmetric waveform ion mobility spectrometry/hybrid quadrupole time-of-flight mass spectrometry. *Rapid communications in mass spectrometry: RCM*, 16(7):676–680, 2002.

30. Karine Venne, Eric Bonneil, Kevin Eng, and Pierre Thibault. Improvement in peptide detection for proteomics analyses using NanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry. *Analytical Chemistry*, 77(7):2176–2186, April 2005.

31. Valérie Gabelica, Alexandre A. Shvartsburg, Carlos Afonso, Perdita Barran, Justin L.P. Benesch, Christian Bleiholder, Michael T. Bowers, Aivett Bilbao, Matthew F. Bush, J. Larry Campbell, Iain D.G. Campuzano, Tim Causon, Brian H. Clowers, Colin S. Creaser, Edwin De Pauw, Johann Far, Francisco Fernandez-Lima, John C. Fjeldsted, Kevin Giles, Michael Groessl, Christopher J. Hogan Jr, Stephan Hann, Hugh I. Kim, Ruwan T. Kurulugama, Jody C. May, John A. McLean, Kevin Pagel, Keith Richardson, Mark E. Ridgeway, Frédéric Rosu, Frank Sobott, Konstantinos Thalassinos, Stephen J. Valentine, and Thomas Wyttenbach. Recommendations for reporting ion mobility Mass Spectrometry measurements. *Mass Spectrometry Reviews*, 38(3):291–320, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mas.21585.

32. Douglas C. Stahl, Kristine M. Swiderek, Michael T. Davis, and Terry D. Lee. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *Journal of the American Society for Mass Spectrometry*, 7(6):532–540, June 1996. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

33. John R. Yates, Jimmy K. Eng, Ashley L. McCormack, and David. Schieltz. Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in

the Protein Database. *Analytical Chemistry*, 67(8):1426–1436, April 1995. Publisher: American Chemical Society.

34. Scott A. McLuckey. Principles of collisional activation in analytical mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 3(6):599–614, September 1992. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

35. Jesper V. Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, September 2007. Number: 9 Publisher: Nature Publishing Group.

36. Vinzenz Lange, Paola Picotti, Bruno Domon, and Ruedi Aebersold. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 4(1):222, January 2008. Publisher: John Wiley & Sons, Ltd.

37. Amelia C. Peterson, Jason D. Russell, Derek J. Bailey, Michael S. Westphall, and Joshua J. Coon. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics*. *Molecular & Cellular Proteomics*, 11(11):1475–1488, November 2012.

38. Samuel Purvine, Jason-Thomas Eppel*, Eugene C. Yi, and David R. Goodlett. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *PROTEOMICS*, 3(6):847–850, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.200300362.

39. John D. Venable, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, and John R. Yates. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1(1):39–45, October 2004. Number: 1 Publisher: Nature Publishing Group.

40. Matthew C. Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L. Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A. Baker, Mi-Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L. Seymour, Lydia M. Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W. Deutsch, Robert L. Moritz, Jonathan E. Katz, David B. Agus, Michael MacCoss, David L. Tabb, and Parag Mallick. A cross-platform toolkit

for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920, October 2012. Number: 10 Publisher: Nature Publishing Group.

41. Christopher Y. Park, Aaron A. Klammer, Lukas Käll, Michael J. MacCoss, and William S. Noble. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*, 7(7):3022–3027, July 2008. Publisher: American Chemical Society.

42. Sean McIlwain, Kaipo Tamura, Attila Kertesz-Farkas, Charles E. Grant, Benjamin Diament, Barbara Frewen, J. Jeffry Howbert, Michael R. Hoopmann, Lukas Käll, Jimmy K. Eng, Michael J. MacCoss, and William Stafford Noble. Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. *Journal of Proteome Research*, 13(10):4488–4491, October 2014. Publisher: American Chemical Society.

43. Jimmy K. Eng, Tahmina A. Jahan, and Michael R. Hoopmann. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS*, 13(1):22–24, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201200439.

44. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

45. John R. Yates, Jimmy K. Eng, Karl R. Clauser, and Alma L. Burlingame. Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *Journal of the American Society for Mass Spectrometry*, 7(11):1089–1098, November 1996. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

46. Jimmy K. Eng, Bernd Fischer, Jonas Grossmann, and Michael J. MacCoss. A Fast SEQUEST Cross Correlation Algorithm. *Journal of Proteome Research*, 7(10):4598–4602, October 2008. Publisher: American Chemical Society.

47. Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008.

48. Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, 14(5):513–520, May 2017.

49. David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*, 20(18):3551–3567, 1999. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291522-2683%2819991201%2920%3A18%3C3551%3A%3AAID-ELPS3551%3E3.0.CO%3B2-2.

50. Robertson Craig and Ronald C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, June 2004.

51. Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, June 2019. Number: 6 Publisher: Nature Publishing Group.

52. Brian C. Searle, Lindsay K. Pino, Jarrett D. Egertson, Ying S. Ting, Robert T. Lawrence, Brendan X. MacLean, Judit Villén, and Michael J. MacCoss. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, 9(1):5128, December 2018. Number: 1 Publisher: Nature Publishing Group.

53. Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, January 2020.

54. Roland Bruderer, Oliver M. Bernhardt, Tejas Gandhi, Saša M. Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, Olga Vitek, Oliver Rinner, and Lukas Reiter. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues*[S]. *Molecular & Cellular Proteomics*, 14(5):1400–1410, May 2015.

55. Pavel Sinitcyn, Hamid Hamzeiy, Favio Salinas Soto, Daniel Itzhak, Frank McCarthy, Christoph Wichmann, Martin Steger, Uli Ohmayer, Ute Distler, Stephanie Kaspar-Schoenefeld, Nikita Prianichnikov, Şule Yılmaz, Jan Daniel Rudolph, Stefan Tenzer, Yasset Perez-Riverol, Nagarjuna Nagaraj, Sean J. Humphrey, and Jürgen Cox. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology*, 39(12):1563–1573, December 2021. Number: 12 Publisher: Nature Publishing Group.

56. J. Alex Taylor and Richard S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0231%2819970615%2911%3A9%3C1067%3A%3AAID-RCM953%3E3.0.CO%3B2-L.

57. J. Alex Taylor and Richard S. Johnson. Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Analytical Chemistry*, 73(11):2594–2604, June 2001. Publisher: American Chemical Society.

58. V. Dancík, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 6(3-4):327–342, 1999.

59. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.1196.

60. Zhongqi Zhang. De Novo Peptide Sequencing Based on a Divide-and-Conquer Algorithm and Peptide Tandem Spectrum Simulation. *Analytical Chemistry*, 76(21):6374–6383, November 2004. Publisher: American Chemical Society.

61. Ari Frank and Pavel Pevzner. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry*, 77(4):964–973, February 2005. Publisher: American Chemical Society.

62. Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem, and Joachim M. Buhmann. NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing. *Analytical Chemistry*, 77(22):7265–7273, November 2005. Publisher: American Chemical Society.

63. Peter A. DiMaggio and Christodoulos A. Floudas. De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. *Analytical Chemistry*, 79(4):1433–1446, February 2007. Publisher: American Chemical Society.

64. Lijuan Mo, Debojyoti Dutta, Yunhu Wan, and Ting Chen. MSNovo: A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical Chemistry*, 79(13):4870–4878, July 2007. Publisher: American Chemical Society.

65. Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, and Meng-Qiu Dong. pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra. *Journal of Proteome Research*, 9(5):2713–2724, May 2010. Publisher: American Chemical Society.

66. Kyowon Jeong, Sangtae Kim, and Pavel A. Pevzner. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics*, 29(16):1953–1962, August 2013.

67. Bin Ma. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, 2015.

68. Claudia Lindemann, Nikolas Thomanek, Franziska Hundt, Thilo Lerari, Helmut E. Meyer, Dirk Wolters, and Katrin Marcus. Strategies in relative and absolute quantitative mass spectrometry based proteomics. *Biological Chemistry*, 398(5-6):687–699, May 2017. Publisher: De Gruyter.

69. Trina Mouchahoir, John E. Schiel, Rich Rogers, Alan Heckert, Benjamin J. Place, Aaron Ammerman, Xiaoxiao Li, Tom Robinson, Brian Schmidt, Chris M. Chumsae, Xinbi Li, Anton V. Manuilov, Bo Yan, Gregory O. Staples, Da Ren, Alexander J. Veach, Dongdong Wang, Wael Yared, Zoran Sosic, Yan Wang, Li Zang, Anthony M. Leone, Peiran Liu, Richard Ludwig, Li Tao, Wei Wu, Ahmet Cansizoglu, Andrew Hanneman, Greg W. Adams, Irina Perdivara, Hunter Walker, Margo Wilson, Arnd Brandenburg, Nick DeGraan-Weber, Stefano Gotta, Joe Shambaugh, Melissa Alvarez, X. Christopher Yu, Li Cao, Chun Shao, Andrew Mahan, Hirsh Nanda, Kristen Nields, Nancy Nightlinger, Helena Maria Barysz, Michael Jahn, Ben Niu, Jihong Wang, Gabriella Leo, Nunzio Sepe, Yan-Hui Liu, Bhumit A. Patel, Douglas Richardson, Yi Wang, Daniela Tizabi, Oleg V. Borisov, Yali Lu, Ernest L. Maynard, Albrecht Gruhler, Kim F. Haselmann, Thomas N. Krogh, Carsten P. Sönksen, Simon Letarte, Sean Shen, Kristin Boggio, Keith Johnson, Wenqin Ni, Himakshi Patel, David Ripley, Jason C. Rouse, Ying Zhang, Carly Daniels, Andrew Dawdy, Olga Friese, Thomas W. Powers, Justin B. Sperry, Josh Woods, Eric Carlson, K. Ilker Sen, St John Skilton, Michelle Busch, Anders Lund, Martha Stapels, Xu Guo, Sibylle Heidelberger, Harini Kaluarachchi, Sean McCarthy, John Kim, Jing Zhen, Ying Zhou, Sarah Rogstad, Xiaoshi Wang, Jing Fang, Weibin Chen, Ying Qing Yu, John G. Hoogerheide, Rebecca Scott, and Hua Yuan. New Peak Detection Performance Metrics from the MAM Consortium Interlaboratory Study. *Journal of the American Society for Mass Spectrometry*, 32(4):913–928, April 2021. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

70. Sarah Rogstad, Haoheng Yan, Xiaoshi Wang, David Powers, Kurt Brorson, Bazarragchaa Damdinsuren, and Sau Lee. Multi-Attribute Method for Quality Control

of Therapeutic Proteins. *Analytical Chemistry*, 91(22):14170–14177, November 2019. Publisher: American Chemical Society.

71. Paul A. Rudnick, Karl R. Clauser, Lisa E. Kilpatrick, Dmitrii V. Tchekhovskoi, Pedatsur Neta, Nikša Blonder, Dean D. Billheimer, Ronald K. Blackman, David M. Bunk, Helene L. Cardasis, Amy-Joan L. Ham, Jacob D. Jaffe, Christopher R. Kinsinger, Mehdi Mesri, Thomas A. Neubert, Birgit Schilling, David L. Tabb, Tony J. Tegeler, Lorenzo Vega-Montoto, Asokan Mulayath Variyath, Mu Wang, Pei Wang, Jeffrey R. Whiteaker, Lisa J. Zimmerman, Steven A. Carr, Susan J. Fisher, Bradford W. Gibson, Amanda G. Paulovich, Fred E. Regnier, Henry Rodriguez, Cliff Spiegelman, Paul Tempst, Daniel C. Liebler, and Stephen E. Stein. Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses. *Molecular & Cellular Proteomics*, 9(2):225–241, February 2010.

72. Michael J. Sweredoski, Geoffrey T. Smith, Anastasia Kalli, Robert L. J. Graham, and Sonja Hess. LogViewer: A Software Tool to Visualize Quality Control Parameters to Optimize Proteomics Experiments using Orbitrap and LTQ-FT Mass Spectrometers. *Journal of Biomolecular Techniques : JBT*, 22(4):122–126, December 2011.

73. Ze-Qiang Ma, Kenneth O. Polzin, Surendra Dasari, Matthew C. Chambers, Birgit Schilling, Bradford W. Gibson, Bao Q. Tran, Lorenzo Vega-Montoto, Daniel C. Liebler, and David L. Tabb. QuaMeter: Multivendor Performance Metrics for LC–MS/MS Proteomics Instrumentation. *Analytical Chemistry*, 84(14):5845–5850, July 2012. Publisher: American Chemical Society.

74. Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H. Tang, Andreas Römpp, Steffen Neumann, Angel D. Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W. Deutsch. mzML—a Community Standard for Mass Spectrometry Data*. *Molecular & Cellular Proteomics*, 10(1):R110.000133, January 2011.

75. Michael R. Hoopmann, Gregory L. Finney, and Michael J. MacCoss. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Analytical Chemistry*, 79(15):5620–5632, August 2007.

76. Brian C. Searle, Kristian E. Swearingen, Christopher A. Barnes, Tobias Schmidt, Siegfried Gessulat, Bernhard Küster, and Mathias Wilhelm. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature Communications*, 11(1):1548, March 2020. Number: 1 Publisher: Nature Publishing Group.

77. 5 Human Albumin. *Transfusion Medicine and Hemotherapy*, 36(6):399–407, December 2009.

78. Marija Holcar, Maša Kandušer, and Metka Lenassi. Blood Nanoparticles – Influence on Extracellular Vesicle Isolation and Characterization. *Frontiers in Pharmacology*, 12, 2021.

79. David W. Peterson and J. M. Hayes. Signal-to-Noise Ratios in Mass Spectroscopic Ion-Current-Measurement Systems. In David M. Hercules, Gary M. Hieftje, Lloyd R. Snyder, and Merle A. Evenson, editors, *Contemporary Topics in Analytical and Clinical Chemistry: Volume 3*, pages 217–252. Springer US, Boston, MA, 1978.

80. Michaela Scigelova, Martin Hornshaw, Anastassios Giannakopulos, and Alexander Makarov. Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics*, 10(7):M111.009431, July 2011.

81. Harrison Specht, Edward Emmott, Aleksandra A. Petelski, R. Gray Huffman, David H. Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, 22(1):50, January 2021.

82. Alexander Makarov and Eduard Denisov. Dynamics of Ions of Intact Proteins in the Orbitrap Mass Analyzer. *Journal of the American Society for Mass Spectrometry*, 20(8):1486–1495, August 2009.

83. Michael J. MacCoss, Michael J. Toth, and Dwight E. Matthews. Evaluation and Optimization of Ion-Current Ratio Measurements by Selected-Ion-Monitoring Mass Spectrometry. *Analytical Chemistry*, 73(13):2976–2984, July 2001. Publisher: American Chemical Society.

84. Jae C. Schwartz, Xaio-Guang Zhou, and Mark E. Bier. Method and apparatus of increasing dynamic range and sensitivity of a mass spectrometer, November 1996.

85. Shanrong Zhao, Zhan Ye, and Robert Stanton. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, 26(8):903–909, August 2020.

86. Christine C. Wu, Kristine A. Tsantilas, Jea Park, Deanna Plubell, Previn Naicker, Ireshyn Govender, Sindisiwe Buthelezi, Stoyan Stoychev, Justin Jordaan, Gennifer Merrihew, Eric Huang, Edward D. Parker, Michael Riffle, Andrew N. Hoofnagle, and Michael J. MacCoss. Mag-Net: Rapid enrichment of membrane-bound particles enables high coverage quantitative analysis of the plasma proteome, June 2023. Pages: 2023.06.10.544439 Section: New Results.

87. Michael J. MacCoss, Javier Antonio Alfaro, Danielle A. Faivre, Christine C. Wu, Meni Wanunu, and Nikolai Slavov. Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nature Methods*, 20(3):339–346, March 2023. Number: 3 Publisher: Nature Publishing Group.

88. Georgi K. Marinov, Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers, and Barbara J. Wold. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, March 2014. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

89. Ron Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays*, 35(12):1050–1055, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201300066.

90. Alexander Makarov, Eduard Denisov, Alexander Kholomeev, Wilko Balschun, Oliver Lange, Kerstin Strupat, and Stevan Horning. Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Analytical Chemistry*, 78(7):2113–2120, April 2006. Publisher: American Chemical Society.

91. Michael W. Senko, Philip M. Remes, Jesse D. Canterbury, Raman Mathur, Qingyu Song, Shannon M. Eliuk, Chris Mullen, Lee Earley, Mark Hardman, Justin D. Blethrow, Huy Bui, August Specht, Oliver Lange, Eduard Denisov, Alexander Makarov, Stevan Horning, and Vlad Zabrouskov. Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Analytical Chemistry*, 85(24):11710–11714, December 2013. Publisher: American Chemical Society.

92. Ying S. Ting, Jarrett D. Egertson, James G. Bollinger, Brian C. Searle, Samuel H. Payne, William Stafford Noble, and Michael J. MacCoss. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature Methods*, 14(9):903–908, September 2017. Number: 9 Publisher: Nature Publishing Group.

93. Florian Meier, Philipp E. Geyer, Sebastian Virreira Winter, Juergen Cox, and Matthias Mann. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, 15(6):440–448, June 2018. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Data acquisition;Mass spectrometry;Proteomic analysis;Proteomics Subject_term_id: data-acquisition;mass-spectrometry;proteomic-analysis;proteomics.

94. Jesse D. Canterbury, Xianhua Yi, Michael R. Hoopmann, and Michael J. MacCoss. Assessing the dynamic range and peak capacity of nanoflow LC-FAIMS-MS on an ion trap mass spectrometer for proteomics. *Analytical Chemistry*, 80(18):6888–6897, September 2008.

95. Sibylle Pfammatter, Eric Bonneil, Francis P. McManus, Satendra Prasad, Derek J. Bailey, Michael Belford, Jean-Jacques Dunyach, and Pierre Thibault. A Novel Differential Ion Mobility Device Expands the Depth of Proteome Coverage and the Sensitivity of Multiplex Proteomic Measurements*. *Molecular & Cellular Proteomics*, 17(10):2051–2067, October 2018.

96. Alexander S. Hebert, Satendra Prasad, Michael W. Belford, Derek J. Bailey, Graeme C. McAlister, Susan E. Abbatiello, Romain Huguet, Eloy R. Wouters, Jean-Jacques Dunyach, Dain R. Brademan, Michael S. Westphall, and Joshua J. Coon. Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry*, 90(15):9529–9537, August 2018.

97. Devin K. Schweppe, Satendra Prasad, Michael W. Belford, José Navarrete-Perea, Derek J. Bailey, Romain Huguet, Mark P. Jedrychowski, Ramin Rad, Graeme McAlister, Susan E. Abbatiello, Eloy R. Woulters, Vlad Zabrouskov, Jean-Jacques Dunyach, João A. Paulo, and Steven P. Gygi. Characterization and Optimization of Multiplexed Quantitative Analyses Using High-Field Asymmetric-Waveform Ion Mobility Mass Spectrometry. *Analytical Chemistry*, 91(6):4010–4016, March 2019.

98. Gennifer E. Merrihew, Jea Park, Deanna Plubell, Brian C. Searle, C. Dirk Keene, Eric B. Larson, Randall Bateman, Richard J. Perrin, Jasmeer P. Chhatwal, Martin R. Farlow, Catriona A. McLean, Bernardino Ghetti, Kathy L. Newell, Matthew P. Frosch, Thomas J. Montine, and Michael J. MacCoss. A peptide-centric quantitative proteomics dataset for the phenotypic assessment of Alzheimer's disease. *Scientific Data*, 10(1):206, April 2023. Number: 1 Publisher: Nature Publishing Group.

99. Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, November 2007. Number: 11 Publisher: Nature Publishing Group.

100. Edward J. Hsieh, Michael R. Hoopmann, Brendan MacLean, and Michael J. MacCoss. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. *Journal of Proteome Research*, 9(2):1138–1143, February 2010. Publisher: American Chemical Society.

101. M Kösters, J Leufken, S Schulze, K Sugimoto, J Klein, R P Zahedi, M Hippler, S A Leidel, and C Fufezan. pymzML v2.0: introducing a highly compressed and seekable gzip format. *Bioinformatics*, 34(14):2513–2514, July 2018.

102. Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. *Journal of proteome research*, 9(8):4152–4160, August 2010.

103. Richard S Rogers, Nancy S Nightlinger, Brittney Livingston, Phil Campbell, Robert Bailey, and Alain Balland. Development of a quantitative mass spectrometry multi-attribute method for characterization, quality control testing and disposition of biologics. *mAbs*, 7(5):881–890, September 2015. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2015.1069454.

104. BCC Publishing Staff. Global Markets and Manufacturing Technologies for Protein Drugs. Technical Report BIO021F, BCC Research, September 2021.

105. Sarah Rogstad, Anneliese Faustino, Ashley Ruth, David Keire, Michael Boyne, and Jun Park. A Retrospective Evaluation of the Use of Mass Spectrometry in FDA Biologics License Applications. *Journal of the American Society for Mass Spectrometry*, 28(5):786–794, May 2017. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

106. NIST Monoclonal Antibody Reference Material 8671. *NIST*. Last Modified: 2022-12-19T16:56-05:00.

107. Trina Formolo, Mellisa Ly, Michaella Levy, Lisa Kilpatrick, Scott Lute, Karen Phinney, Lisa Marzilli, Kurt Brorson, Michael Boyne, Darryl Davis, and John Schiel. Determination of the NISTmAb Primary Structure. In *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 2. Biopharmaceutical Characterization: The NISTmAb Case Study*, volume 1201 of *ACS Symposium Series*, pages 1–62. American Chemical Society, January 2015. Section: 1.

108. Wenzhou Li, James L. Kerwin, John Schiel, Trina Formolo, Darryl Davis, Andrew Mahan, and Sabrina A. Benchaar. Structural Elucidation of Post-Translational Modifications in Monoclonal Antibodies. In *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 2. Biopharmaceutical Characterization: The NISTmAb Case Study*, volume 1201 of *ACS Symposium Series*, pages 119–183. American Chemical Society, January 2015. Section: 3.

109. John E. Schiel, Michael J. Tarlov, Karen W. Phinney, Oleg V. Borisov, and Darryl L. Davis. A Global Partnership Advancing Biopharmaceutical Development: Summary and Future Perspectives. In *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 3. Defining the Next Generation of Analytical and Biophysical Techniques*, volume 1202 of *ACS Symposium Series*, pages 415–431. American Chemical Society, January 2015. Section: 15.

110. Robert G. Brinson, John P. Marino, Frank Delaglio, Luke W. Arbogast, Ryan M. Evans, Anthony Kearsley, Geneviève Gingras, Houman Ghasriani, Yves Aubin, Gregory K. Pierens, Xinying Jia, Mehdi Mobli, Hamish G. Grant, David W. Keizer, Kristian Schweimer, Jonas Ståhle, Göran Widmalm, Edward R. Zartler, Chad W. Lawrence, Patrick N. Reardon, John R. Cort, Ping Xu, Feng Ni, Saeko Yanaka, Koichi Kato, Stuart R. Parnham, Desiree Tsao, Andreas Blomgren, Torgny Rundlöf, Nils Trieloff, Peter Schmieder, Alfred Ross, Ken Skidmore, Kang Chen, David Keire, Darón I. Freedberg, Thea Suter-Stahel, Gerhard Wider, Gregor Ilc, Janez Plavec, Scott A. Bradley, Donna M. Baldisseri, Mauricio Luis Sforça, Ana Carolina de Mattos Zeri, Julie Yu Wei, Christina M. Szabo, Carlos A. Amezcua, John B. Jordan, and Mats Wikström. Enabling adoption of 2D-NMR for the higher order structure assessment of monoclonal antibody therapeutics. *mAbs*, 11(1):94–105, January 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2018.1544454.

111. Jeffrey W. Hudgens, Elyssia S. Gallagher, Ioannis Karageorgos, Kyle W. Anderson, James J. Filliben, Richard Y.-C. Huang, Guodong Chen, George M. Bou-Assaf, Alfonso Espada, Michael J. Chalmers, Eduardo Harguindey, Hui-Min Zhang, Benjamin T. Walters, Jennifer Zhang, John Venable, Caitlin Steckler, Inhee Park, Ansgar Brock, Xiaojun Lu, Ratnesh Pandey, Arun Chandramohan, Ganesh Srinivasan Anand, Sasidhar N. Nirudodhi, Justin B. Sperry, Jason C. Rouse, James A. Carroll, Kasper D. Rand, Ulrike Leurs, David D. Weis, Mohammed A. Al-Naqshabandi, Tyler S. Hageman, Daniel Deredge, Patrick L. Wintrode, Malvina Papanastasiou, John D. Lambris, Sheng Li, and Sarah Urata. Interlaboratory Comparison of Hydrogen–Deuterium Exchange Mass Spectrometry Measurements of the Fab Fragment of NISTmAb. *Analytical Chemistry*, 91(11):7336–7345, June 2019. Publisher: American Chemical Society.

112. Sally L. McArthur. Repeatability, Reproducibility, and Replicability: Tackling the 3R challenge in biointerface science and engineering. *Biointerphases*, 14(2):020201, March 2019.

Appendix A

GENERATING PROSIT LIBRARIES FOR ENCYCLOPEDIA

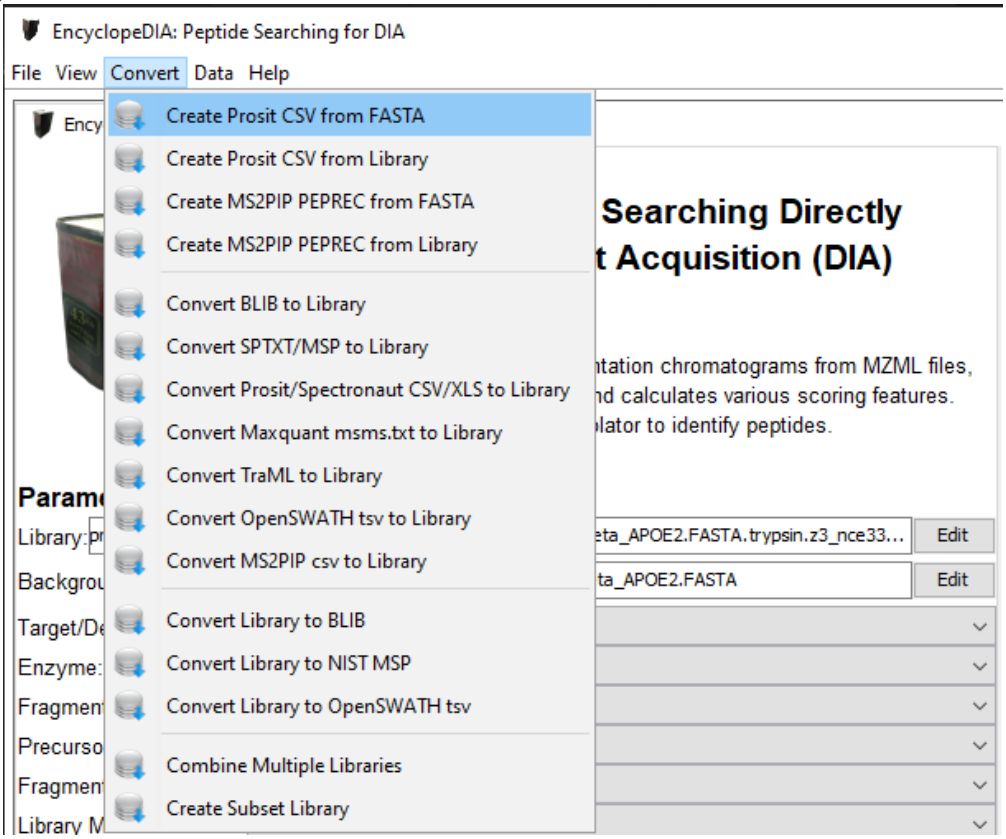
1. Generate .fasta file

2. Convert .fasta file to Prosit .csv input

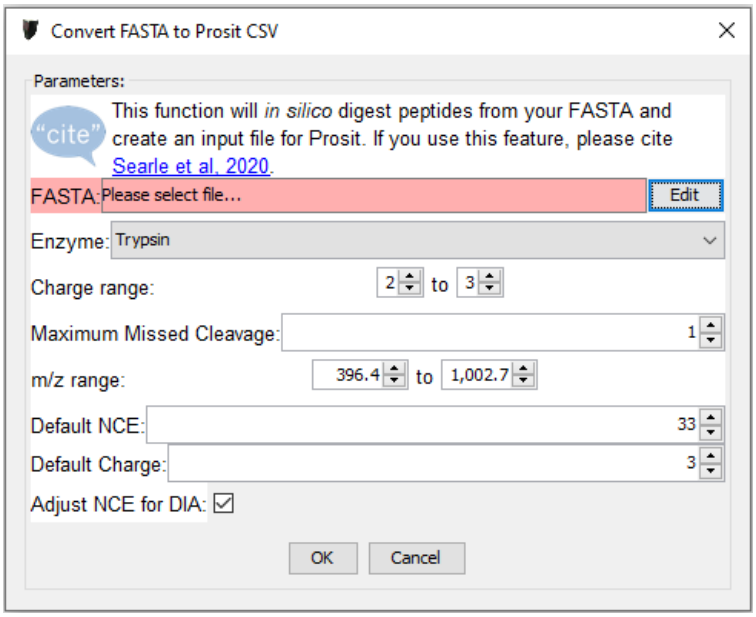
   (a) Open EncyclopeDIA (this appendix was created with v2.12.30)

      i. Confirm Enzyme = Trypsin and Fragmentation = CID/HCD (B/Y)

   (b) Click the "Convert" tab and select "Create Prosit CSV from FASTA"



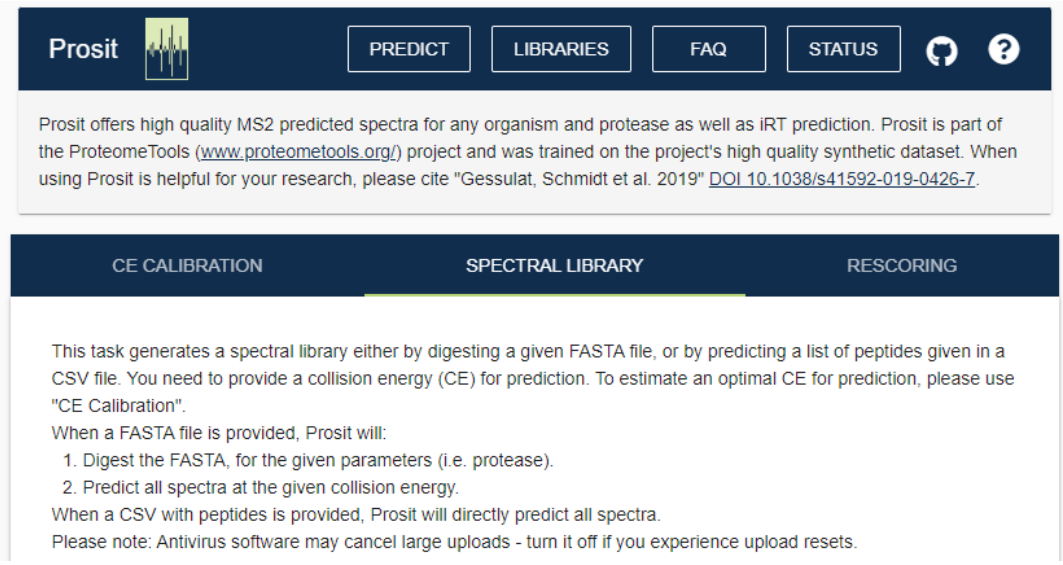
(c) Add the .fasta file and continue with the default settings and select “OK”



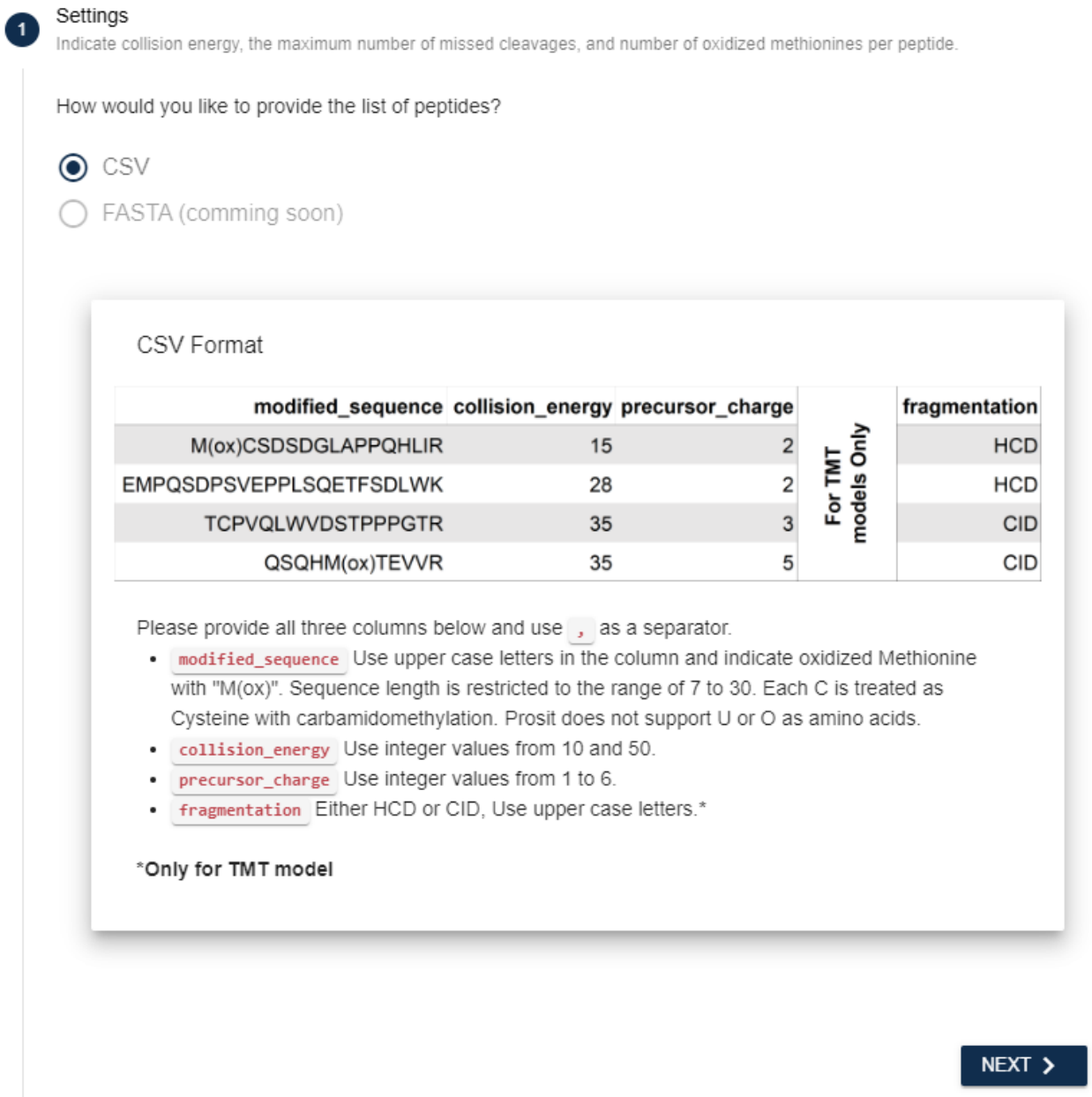
3. Upload .csv to Prosit for spectral library generation

(a) Navigate to Prosit: `https://www.proteomicsdb.org/prosit/`

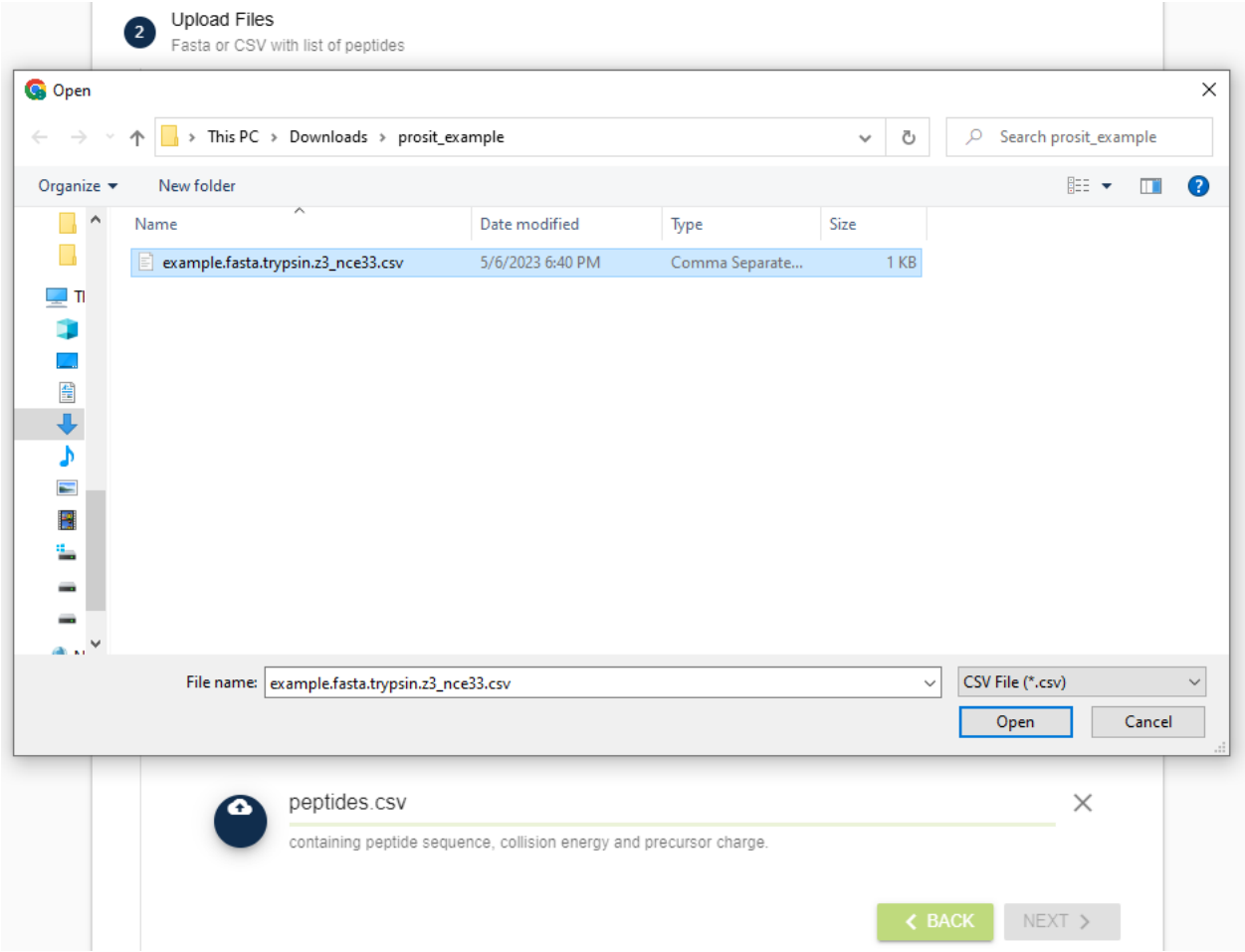
(b) Change to the “Spectral Library” tab



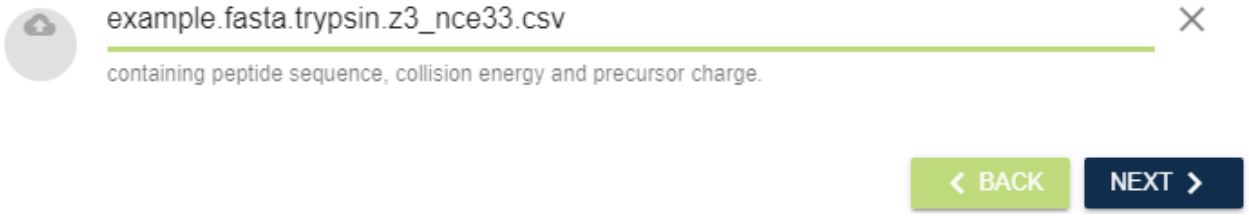
(c) Scroll down and click “NEXT >”



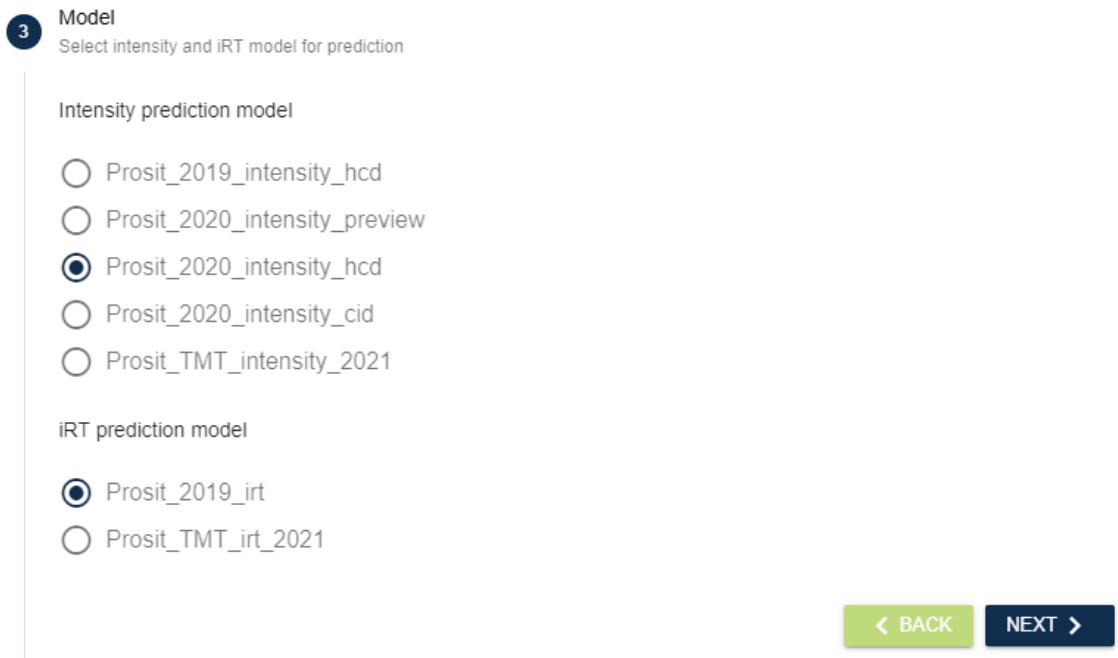
(d) Click on “peptides.csv” and add your .csv file by pressing “Open”



(e) When the .csv file is uploaded (button will turn blue), click “NEXT >”

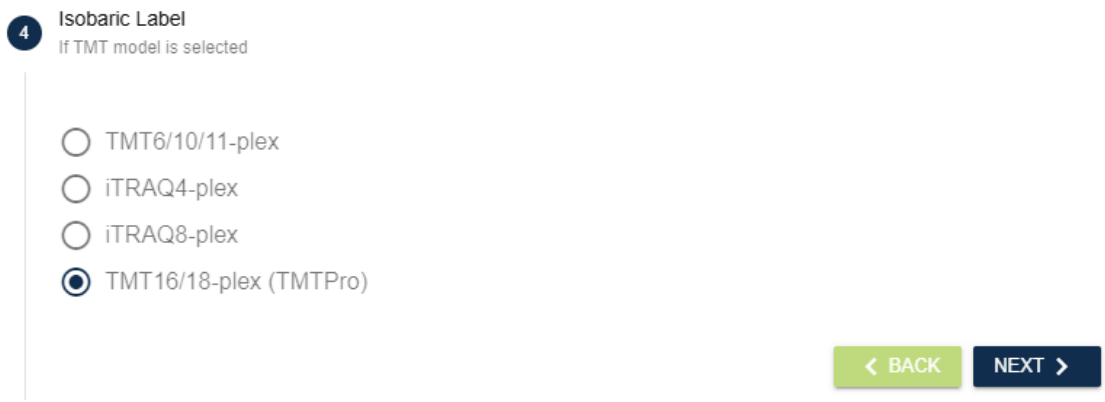


(f) Choose the models you want to use to build the predicted libraries, and press “NEXT >”

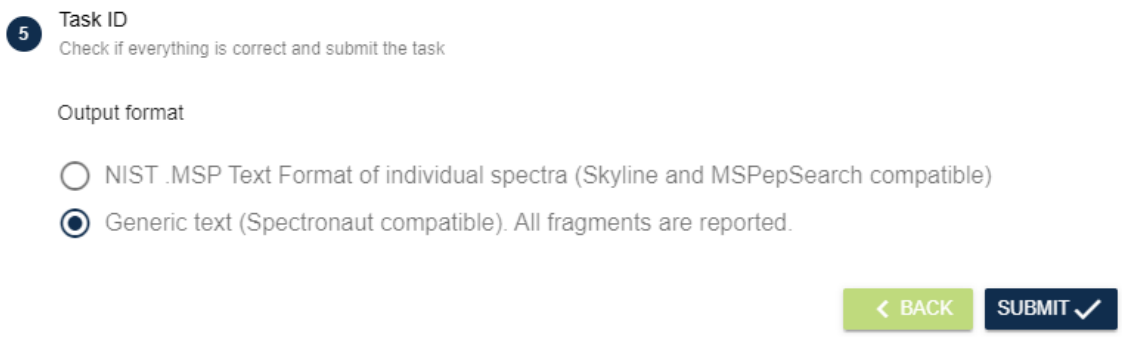


(g) *I*f a TMT model is selected in previous step, choose kit type

This step was skipped with the models chosen above



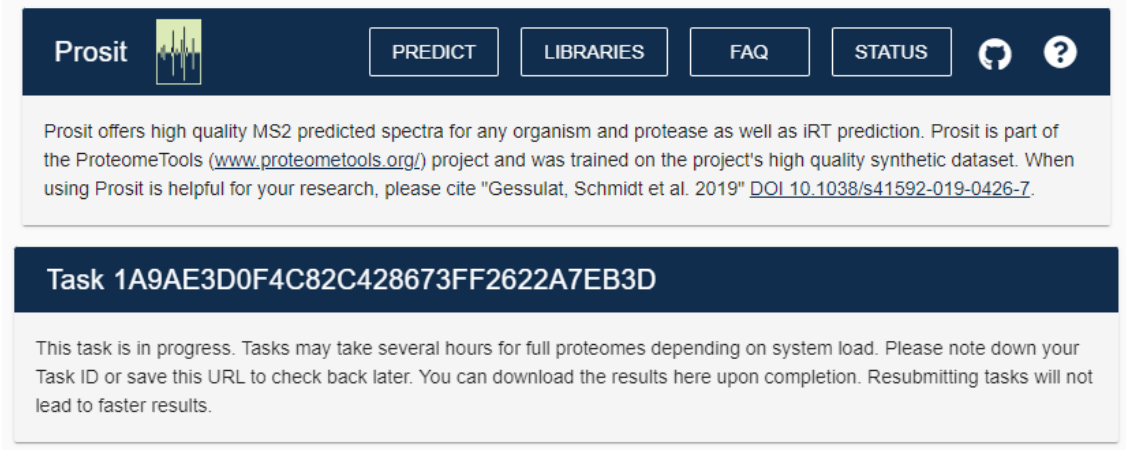
(h) Choose output format



(i) **BEFORE PRESSING "SUBMIT ✓"**, record the file and parameters you chose. The generated library will have a generic name of "myPrositLib.csv"

(j) Click "SUBMIT ✓"

(k) A new page will be shown with a unique task ID. Save the task ID or URL. It could be useful to record it in the same place as the information above



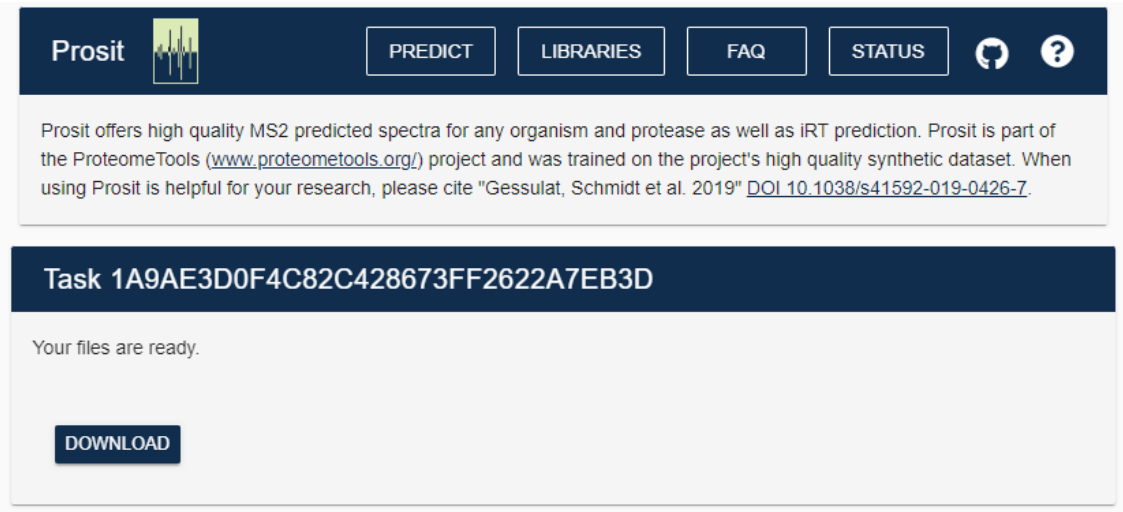
(l) Check back to download the results

YOU WILL NOT BE NOTIFIED WHEN THE JOB IS COMPLETE

The results are stored for 14 days. A task will need to be re-submitted if a file needs to be downloaded after that time

4. Download Prosit output

   (a) Click “Download” when the files are ready

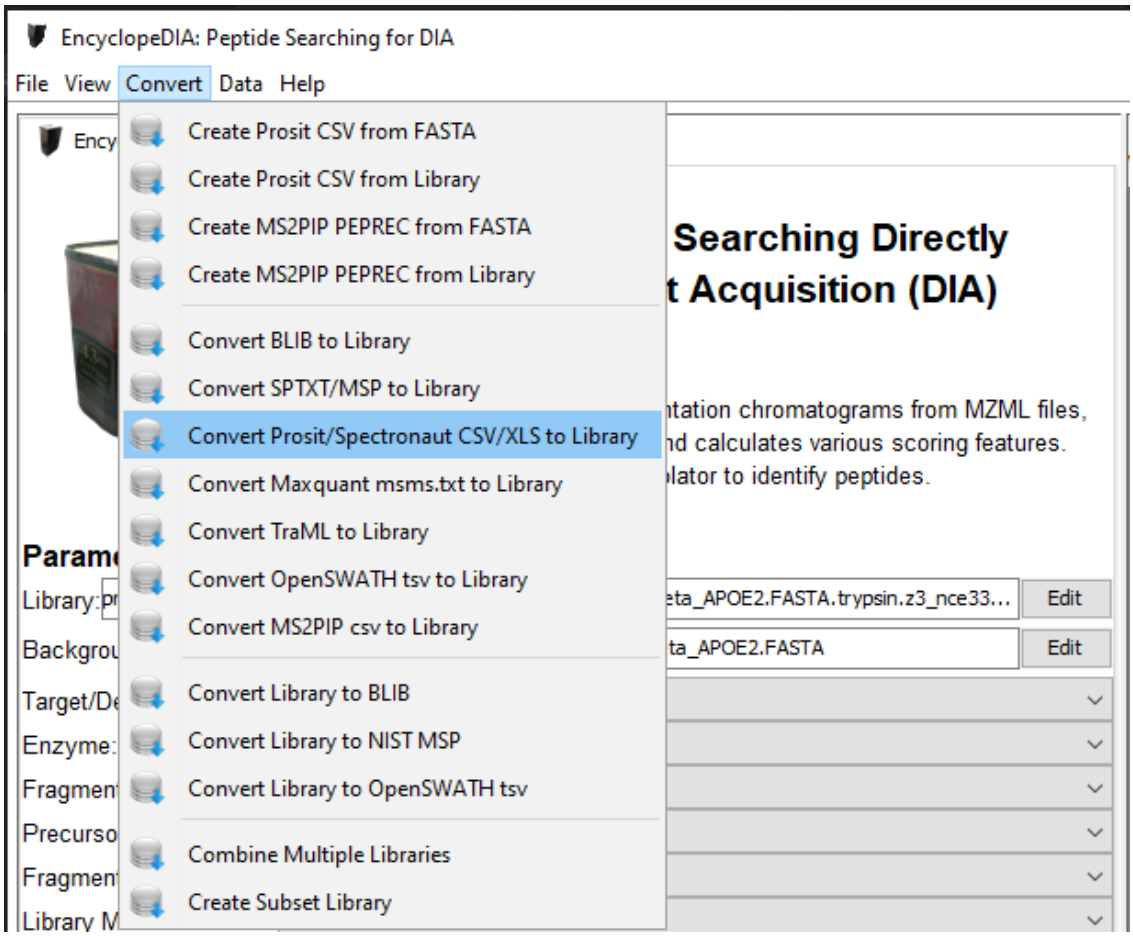


   (b) Extract the file from the zipped folder “download.zip”

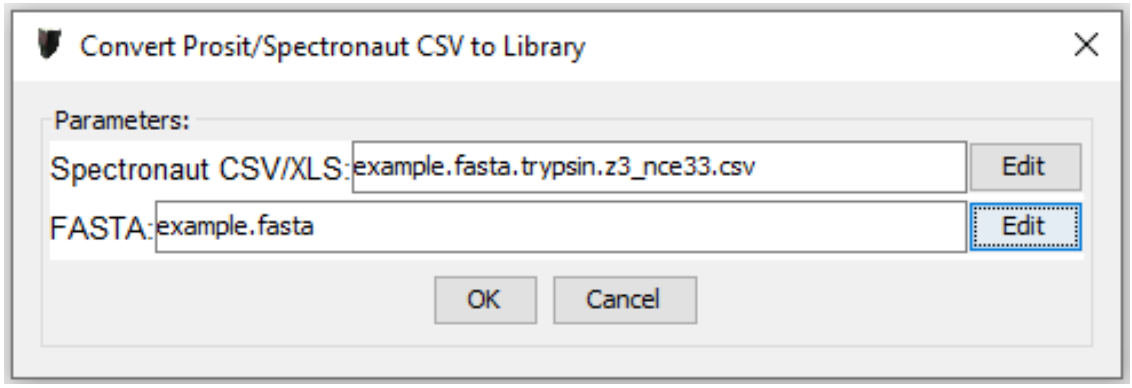
   (c) Rename the “download.zip” folder and extracted “myPrositLib.csv” file to aid in future identification

5. Convert Prosit .csv to EncyclopeDIA .dlib

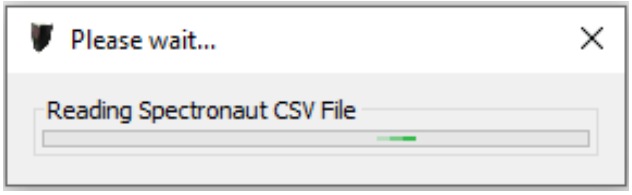
   (a) Click the "Convert" tab and select "Convert Prosit/Spectronaut CSV/XLS to Library"



(b) For the "Spectronaut CSV/XLS" parameter, choose the new .csv file obtained from Prosit. For the "FASTA" parameter, choose the original .fasta file used to generate the .csv uploaded to Prosit



(c) Press "OK" and wait for the file to be converted.



(d) Once complete, the pop-up box will disappear, and the console will say "Finished reading {example.fasta.trypsin.z3_nce33.csv}"

Appendix B

DIA DATA ANALYSIS WITH ENCYCLOPEDIA

MSConvert & EncyclopeDIA GUIs

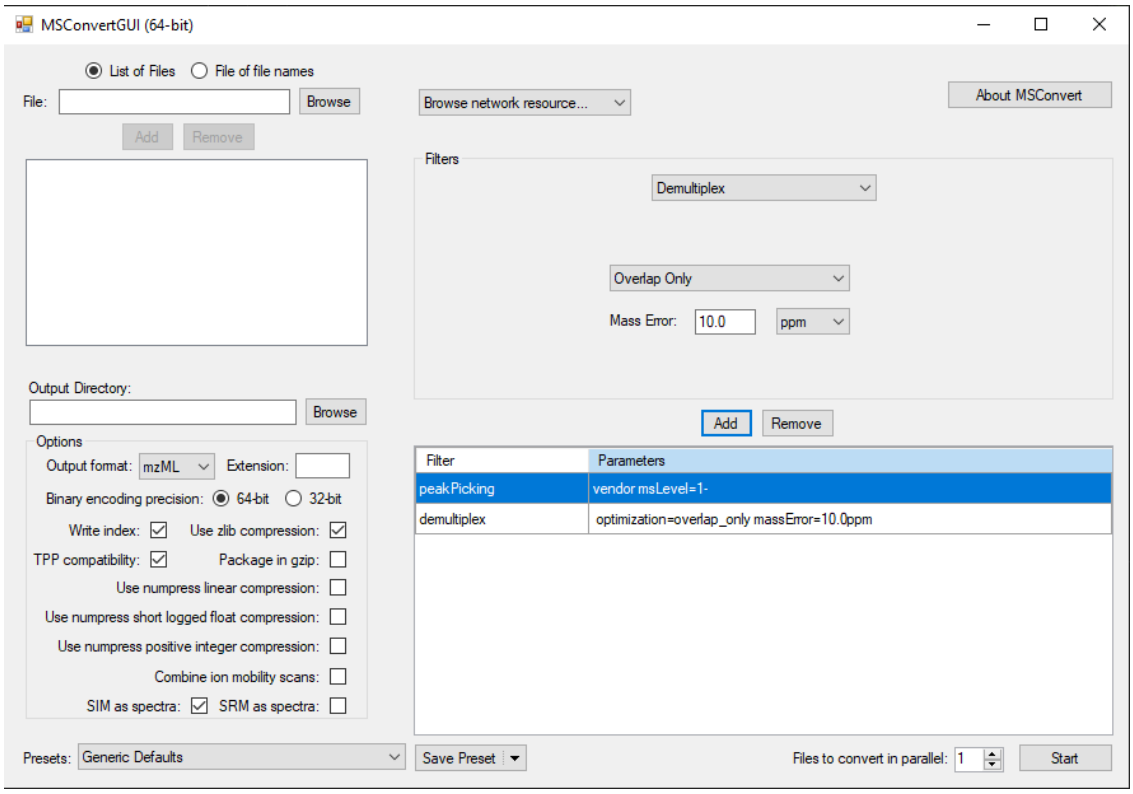
1. Convert .raw files to .mzML with MSConvert

   (a) Have “peakPicking” as the first filter

   (b) Check that the “Output format” is “mzML”

   (c) If the MS1 was acquired as a SIM scan, check the box for “SIM as spectra”

   (d) If the experiment used overlapping windows, add a “demultiplex” filter



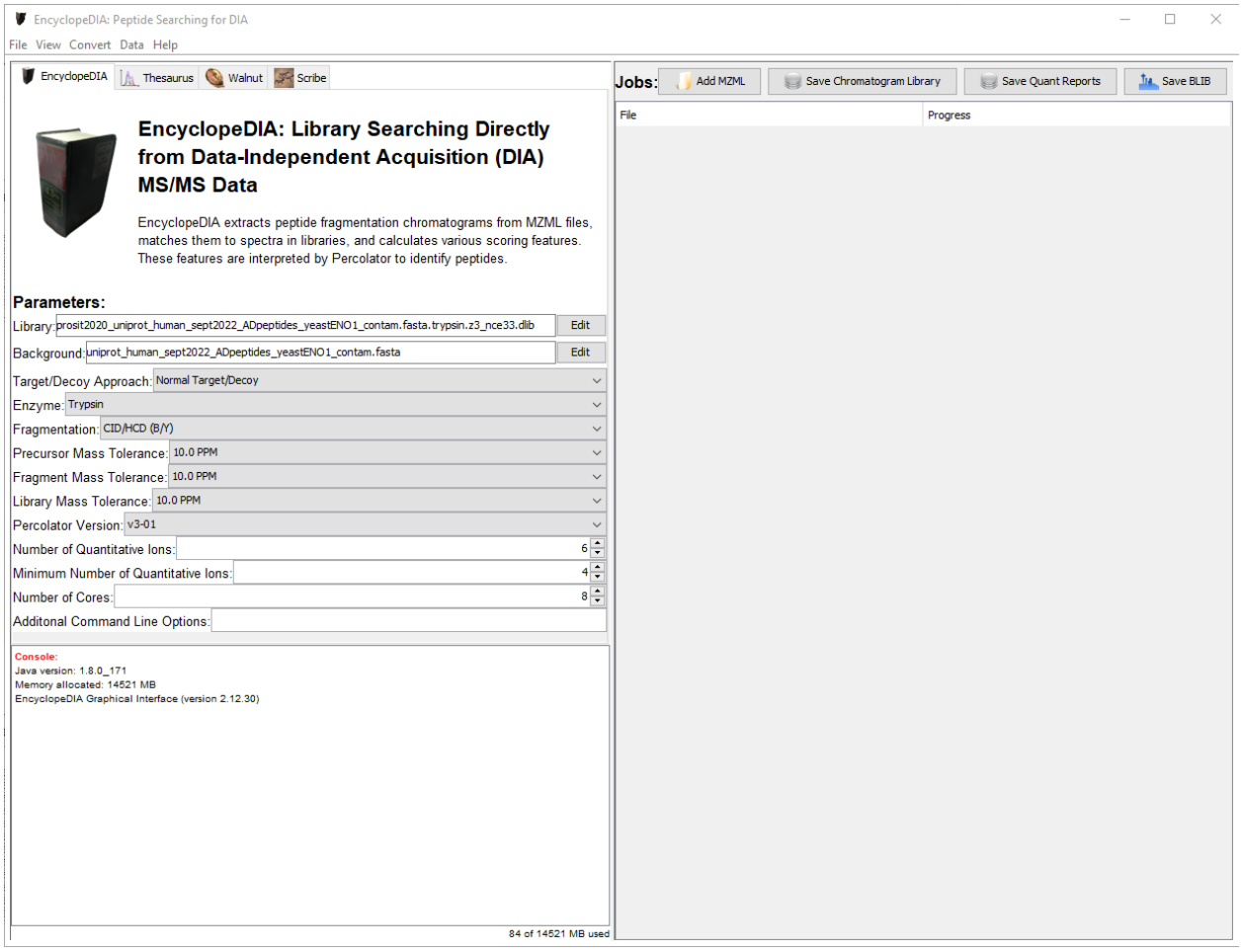
(e) Add files using the “Browse” button. If you only choose 1 file, you must press the “Add” button to place it in the list of files to be converted

2. Create a chromatogram library .elib with EncyclopeDIA

(a) Add the generated Prosit library (see Appendix A) to the “Library” parameter by pressing “Edit”

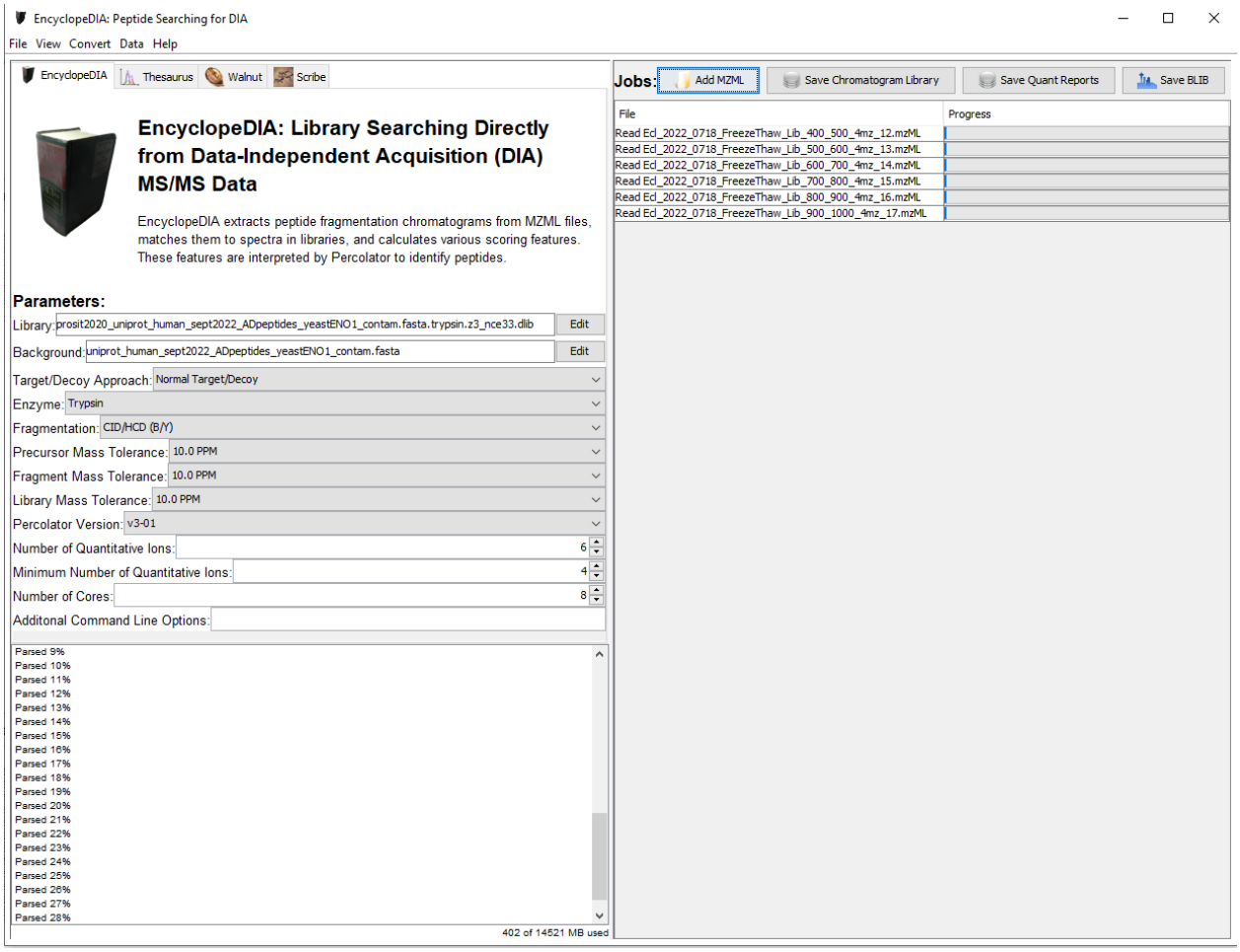
(b) Add the .fasta file (most likely used to generate Prosit library) to the “Background” parameter by pressing “Edit”

(c) Check that the other parameters match the experimental setup

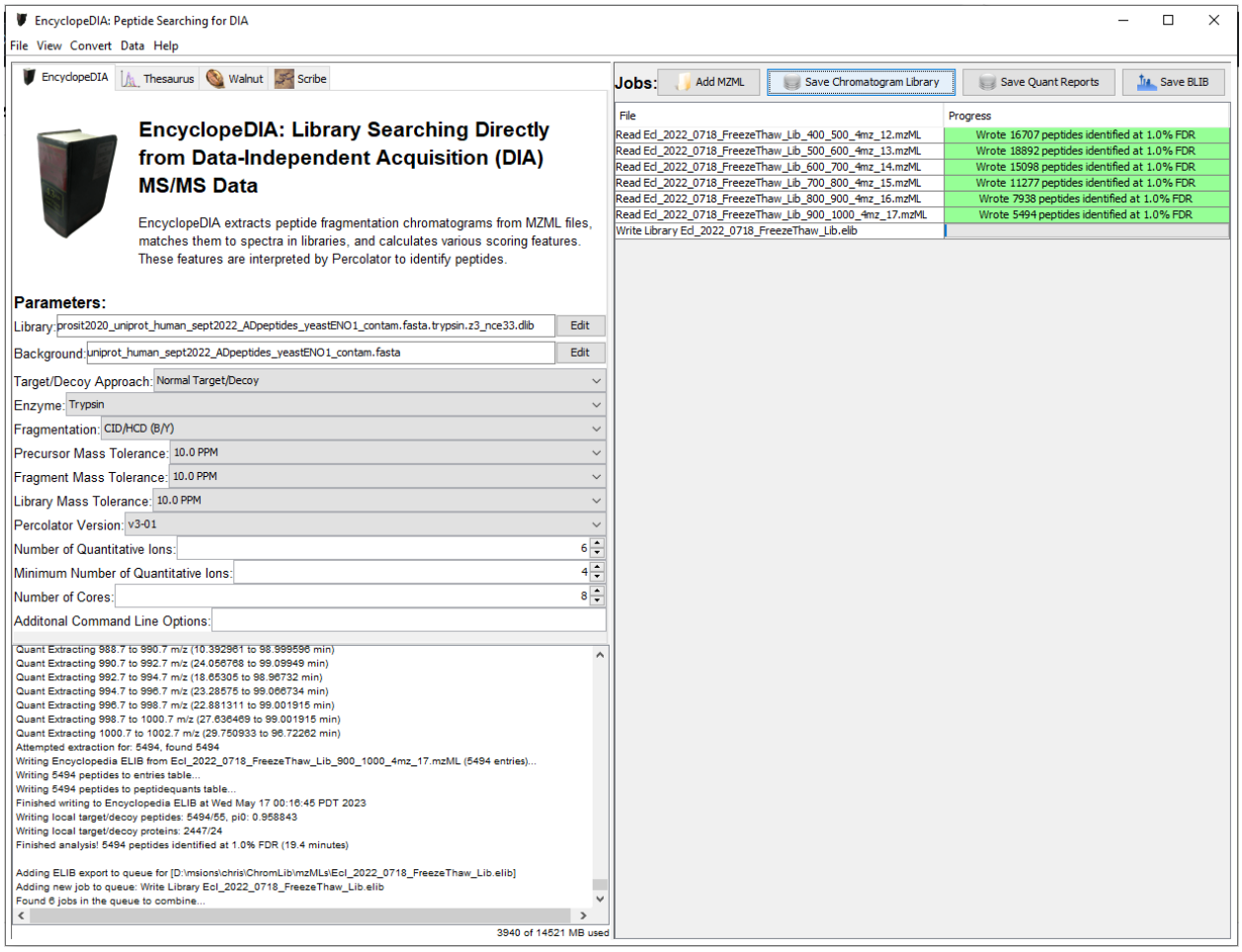


(d) Click “Add MZML” near the top of the right side, and select the files for the gas-phase fractionated library

(e) Click “Open” and the mzML files will appear under the “Jobs” buttons

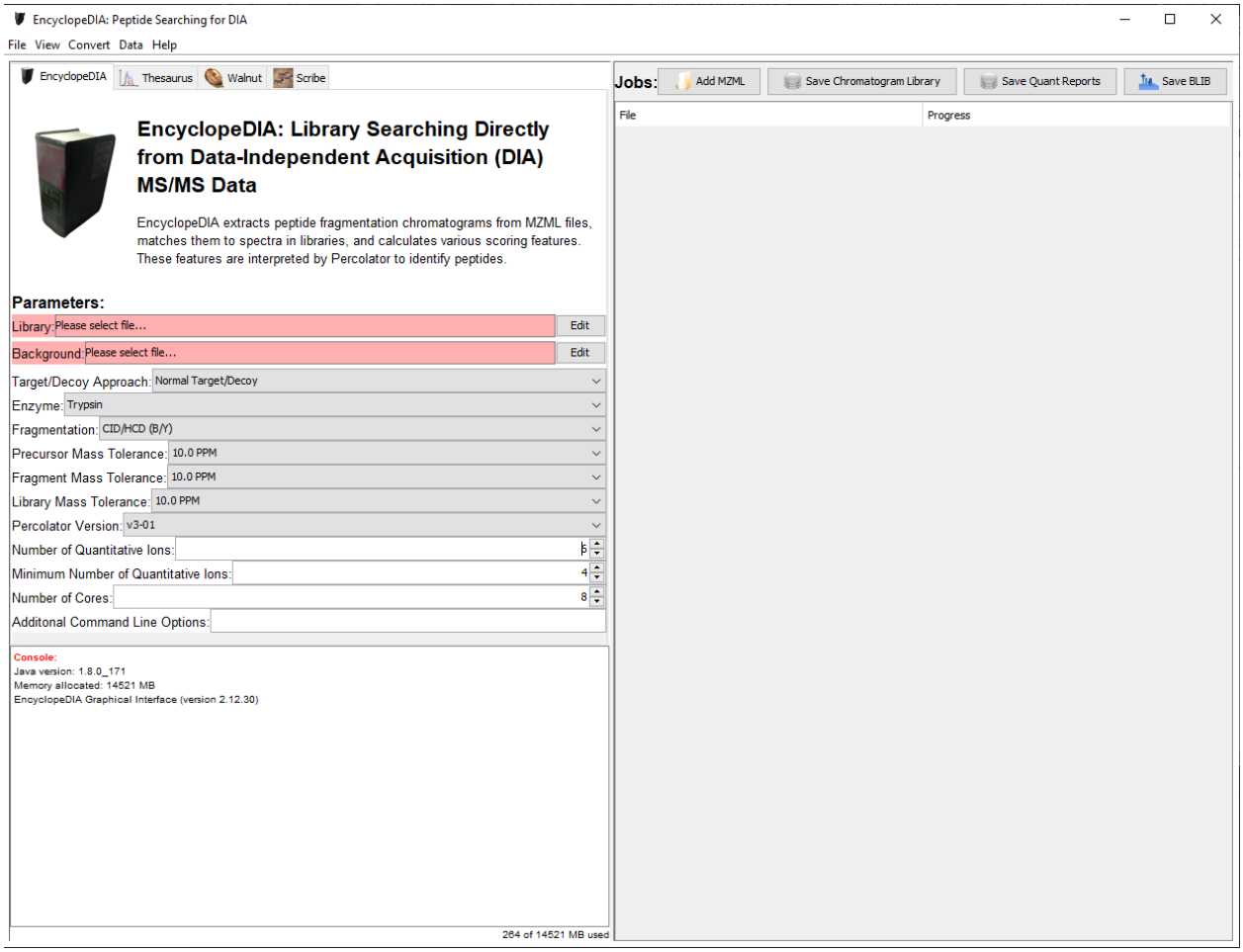


(f) When the gas-phase fractionated files have finished, click "Save Chromatogram Library" and give the ELIB a descriptive filename



3. Search wide-window data with a chromatogram library

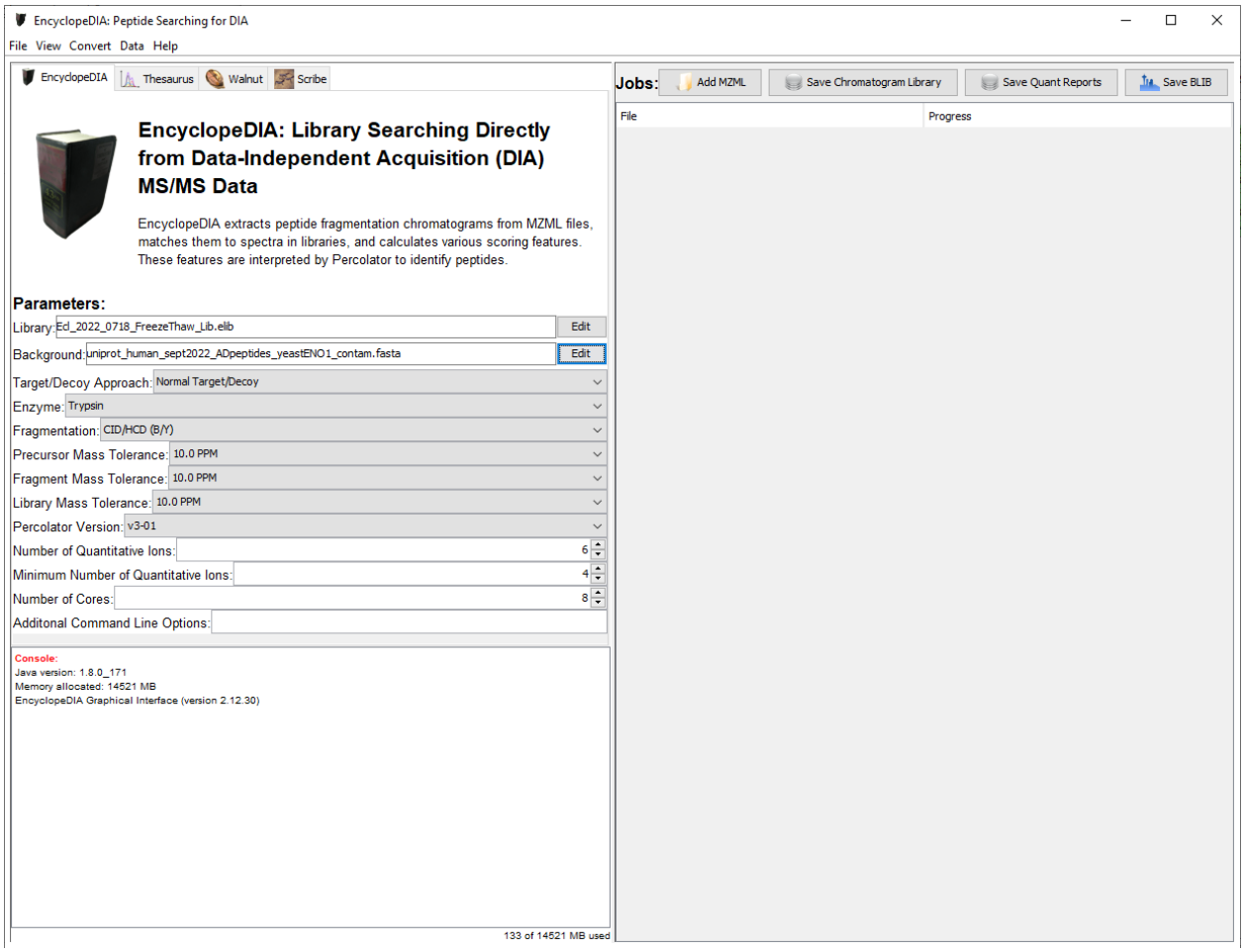
   (a) Close and re-open the EncyclopeDIA GUI to clear EncyclopeDIA's cache/history



(b) Add the newly generated .elib file from Step 2f to the "Library" parameter by pressing "Edit"

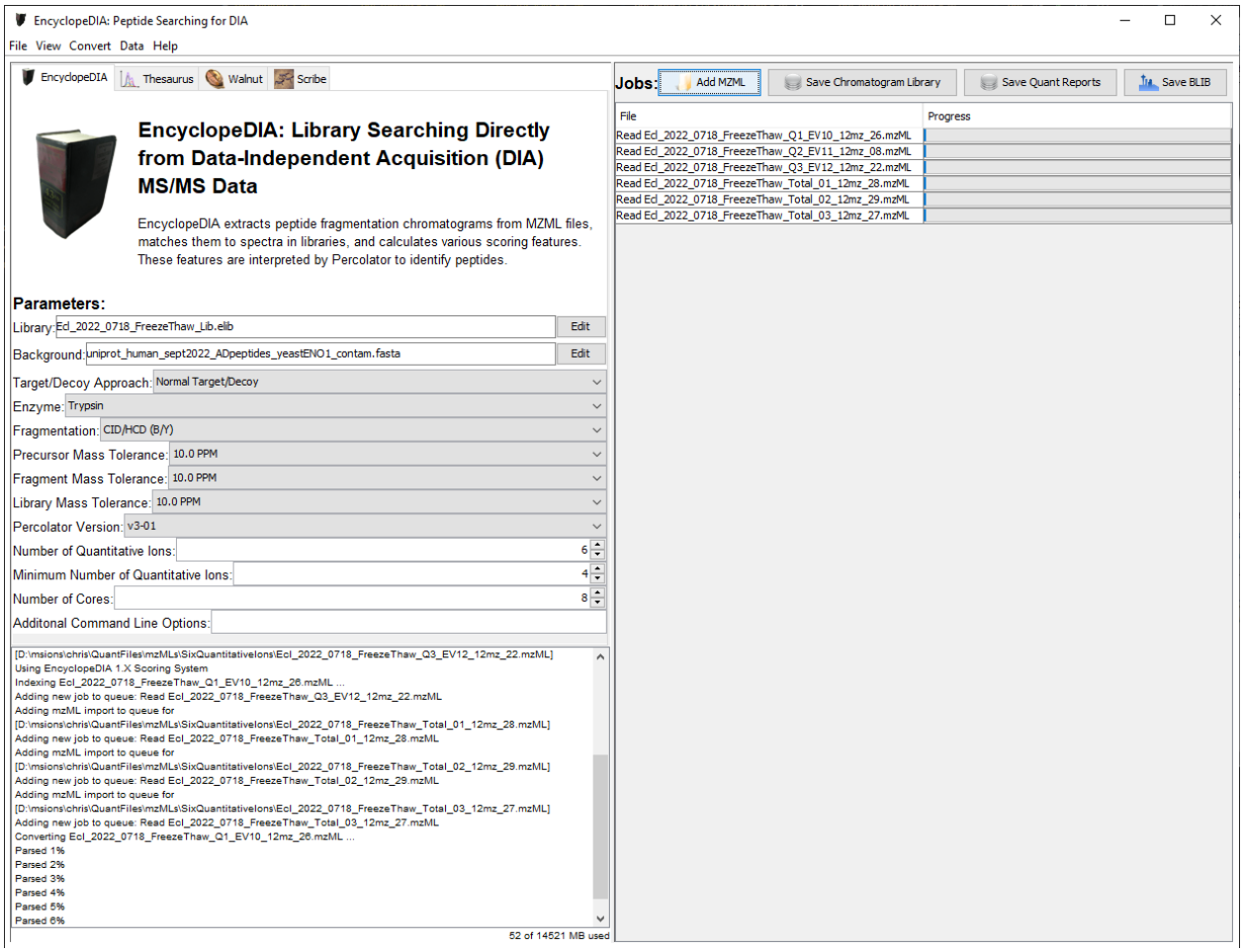
(c) Add the same .fasta file to the "Background" parameter by pressing "Edit"

(d) Check that the other parameters match the experimental setup

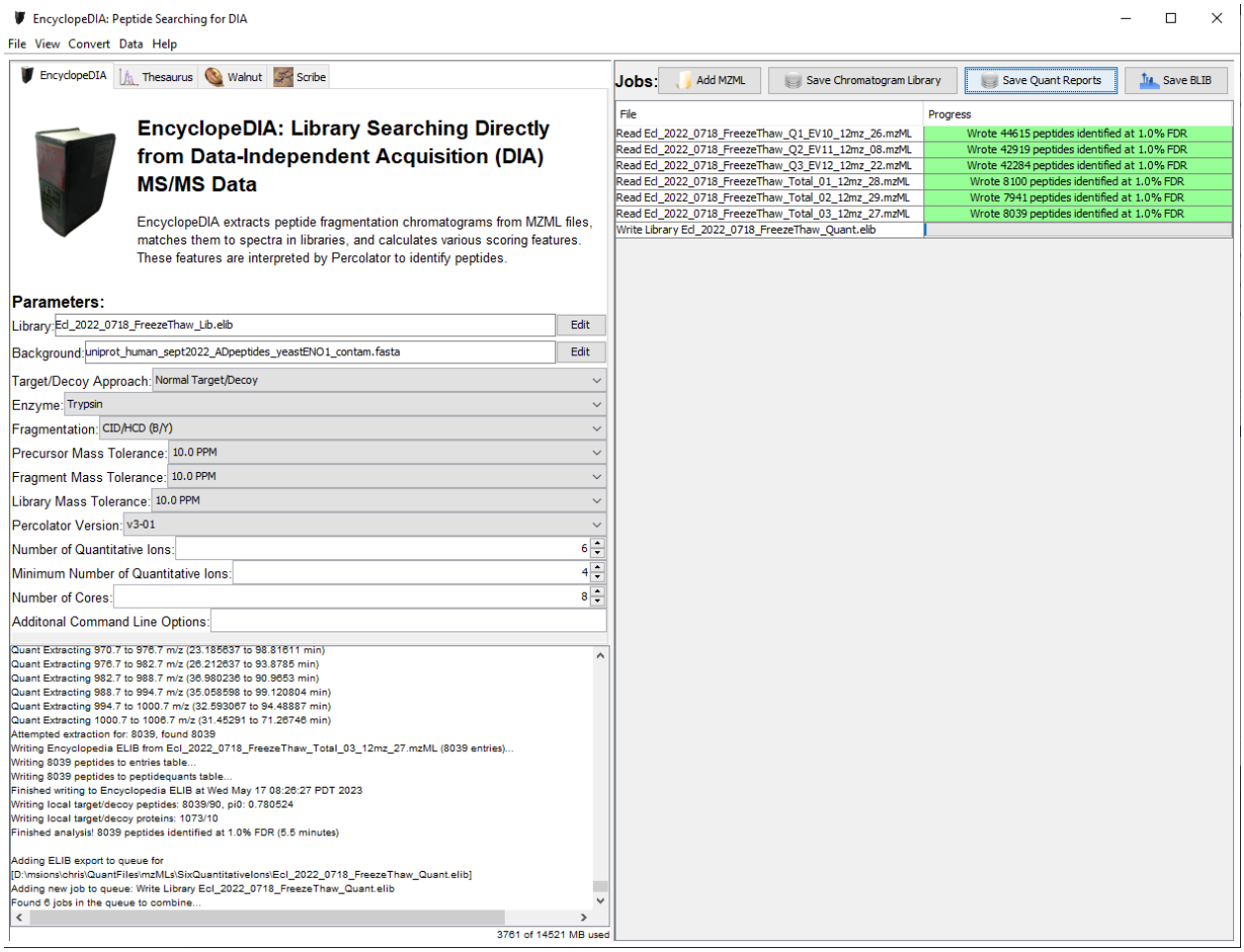


(e) Click “Add MZML” near the top of the right side, and select the wide-window sample files that were acquired with the narrow-window library files

(f) Click “Open” and the mzML files will appear under the “Jobs” buttons



(g) When the sample files have finished, click "Save Quant Reports" and give this quant ELIB a descriptive filename



VITA

Danielle A. Faivre was born in Houston, Texas, and grew up in Washington, Illinois. She graduated from the University of Notre Dame with a Bachelor of Science in Honors Biochemistry in 2017. While at ND, she fell in love with mass spectrometry in Norm Dovichi's lab and enjoyed being part of the marching band. Outside of the lab, Dani likes to rock climb and dance. Her favorite dance fitness format is Oula, and she was certified as an instructor in November 2022.