Computational Methods in Top-down Proteomics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Jihyung Kim

aus Seoul / Südkorea

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.03.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Nico Pfeifer

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Computational Methods in Top-down Proteomics

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Abstract

The study of proteoforms, distinct molecular forms of a protein, is a key to understanding the complexity of biological systems and their underlying molecular mechanisms. For the analysis of proteoforms, top-down proteomics (TDP) based on mass spectrometry (MS) is currently the most powerful analysis technology. It allows intact proteoforms to be directly measured and characterized, preventing loss of information compared to the conventional bottom-up approach. Still, quantitative TDP measurement is an ongoing challenge. Accurate quantification of individual proteoforms is a critical step in identifying alterations in the proteome under various biological conditions. Several quantitative experiment methods for TDP have been introduced, but they still face significant challenges, especially data analysis methods.

In this thesis, we introduce several algorithms to tackle data analysis problems in quantitative TDP. Firstly, as deconvolution is a key first step in TDP data analysis to alleviate the complexity of MS signals in TDP, FLASHDeconv was developed. It is an algorithm for fast and robust spectral deconvolution, employing the novel idea of mass spectra transformation for decharging. FLASHDeconv promises not only unprecedented runtime but also more genuine deconvolution results compared to existing methods.

As FLASHDeconv paved the way for further data analysis steps, FLASHQuant was developed to specifically contribute to quantifying proteoform in a fast and accurate manner. Using a key algorithm of FLASHDeconv allows FLASHQuant to find proteoform features rapidly, and then coeluting proteoforms are resolved and quantified, providing accurate quantification results. Moreover, FLASHQuant showed highly reproducible quantification in both a simple targeted protein dataset and a complex proteome-level dataset.

To satisfy the strong need for a graphical user interface to visualize results from FLASHDeconv and FLASHQuant, we implemented FLASHViewer. It is a web application for visualizing deconvolved or quantified signals. To accommodate the diverse usage for both tools, informative plots are modularized to enable configurable layouts.

All algorithms discussed in this thesis were developed and implemented based on OpenMS, an open-source platform for computational mass spectrometry. As a part of the OpenMS community, the developed tools are all publicly available and platform-independent.

Zusammenfassung

Proteoformen sind die unterschiedlichen molekularen Varianten eines Proteins. Sie sind entscheidend für das Verständnis biologischer Systeme und der Komplexität ihrer zugrunde liegenden molekularen Mechanismen. Die Top-Down-Proteomik (TDP), basierend auf der Massenspektrometrie (MS), ist derzeit die leistungsstärkste Methodik für die Analyse von Proteoformen. Sie ermöglicht es, intakte Proteoformen direkt zu messen und zu charakterisieren, wodurch Informationsverluste im Gegensatz zum herkömmlichen Bottom-Up-Ansatz vermieden werden. Die genaue Quantifizierung einzelner Proteoformen ist ein kritischer Schritt, um Veränderungen im Proteom unter verschiedenen biologischen Bedingungen zu identifizieren. In der Vergangenheit wurden verschiedene experimentelle quantitative Methoden für TDP eingeführt. Diese stehen jedoch weiterhin vor erheblichen Herausforderungen, insbesondere in Bezug auf Datenanalysemethoden.

In dieser Dissertation werden verschiedene Algorithmen vorgestellt, um Schlüsselprobleme bei der Analyse von quantitativen TDP-Daten zu bewältigen. Dekonvolution ist ein entscheidender Schritt, um die Komplexität der MS-Signale in der TDP zu verringern und weitere Analysemethoden zu ermöglichen. Im Rahmen dieser Dissertation wurde FLASHDeconv entwickelt. Es handelt sich hierbei um einen Algorithmus für schnelle und robuste Dekonvolution, welcher die neuartige Idee der Massenspektrum-Transformation zur Ladungsdekonvolution nutzt. FLASHDeconv bietet nicht nur beispiellose Laufzeiten, sondern auch zuverlässigere Dekonvolutionsergebnisse als bestehende Methoden.

FLASHDeconv eröffnet neue Möglichkeiten für konsekutive Datenanalyse. Folglich wurde ebenfalls FLASHQuant entwickelt, um die schnelle und genaue Quantifizierung von Proteoformen zu ermöglichen. Die Integration mit FLASHDeconv ermöglicht es FLASHQuant, Proteoformen schnell zu identifizieren, koeluierende Proteoformen aufzulösen und zu quantifizieren. FLASHQuant weist höchst reproduzierbare Quantifizierungsergebnisse auf - sowohl in einem einfachen, zielgerichteten Proteindatensatz als auch in einem komplexen Proteom-Datensatz.

Um dem Bedarf an einer grafischen Benutzeroberfläche zur Visualisierung der Ergebnisse von FLASHDeconv und FLASHQuant gerecht zu werden, wurde FLASHViewer implementiert. Es handelt sich um eine Webanwendung zur Visualisierung von dekonvolvierten und optional auch quantifizierten Signalen. Zur Unterstützung der vielfältigen Nutzung beider Werkzeuge wurden informative Diagramme modular integriert, um konfigurierbare Layouts zu ermöglichen.

Alle in dieser Dissertation besprochenen Algorithmen wurden im Rahmen von OpenMS, einer Open-Source-Plattform für computergestützte Massenspektrometrie, entwickelt und implementiert. Als Teil der OpenMS-Community sind die entwickelten Werkzeuge öffentlich und plattformunabhängig zugänglich.

Acknowledgments

I would like to express my deep gratitude to Prof. Oliver Kohlbacher for providing me with the opportunity and resources to pursue my doctoral study. Your mentorship and support have been invaluable throughout this journey. From you, I have learned that kindness in leadership is the most important virtue.

My sincere appreciation goes to Dr. Kyowon Jeong for your endless patience, help, and guidance. Your support has been instrumental in shaping both my research and personal growth, and I cannot thank you enough for your dedication.

I want to thank the members of the ABI group from the Kohlbacher lab, especially the OpenMS team, for creating a collaborative and inspiring environment. My special thanks go to Dr. Axel Walter and Tom Müller for their time and effort in proofreading this thesis. I would like to thank Claudia Walter for her patience and support with the challenges I had in bureaucracy.

I am grateful for the A4B project that enabled my research. To my colleagues in the A4B project - Manasi Gaikwad, Siti Nurul Hidayah, and Maša Babović - you have been not just fellow researchers but also dear friends. I extend my appreciation to Prof. Hartmut Schlüter, Prof. Andreas Tholey, Philipp Kaulich, and Dr. Konrad Winkels for our collaboration and their contribution.

To my doctoral study running mate, Yesol, and to Halim, Hanhee, and Yoonwon - thank you for all the countless Zoom calls that kept me motivated, especially during challenging times. I am also thankful to Simon and Xixi for their motivation and encouragement.

I would like to express my appreciation to my family, especially Jiwoo, for their support and motivation. Mom, who has been my first role model as a woman in science - thank you for the inspiration and your unconditional love. Yaki and Nori, I am grateful for the joy you brought to my life. Finally, to Andrew — you have seen me at my best and worst, believing in me when I couldn't. Thank you for your unlimited support and love.

General Remarks

- The research conducted for this thesis was carried out as part of the A4B, Analytics for Biologics, project. A4B consortium is a Europe-wide innovative training network (ITN), funded by the Horizon 2020, Marie Sklodowska-Curie Action ITN 2017, of the European Commission (H2020-MSCA-ITN-2017).
- In accordance with the standard scientific protocol, I will use the personal pronoun ***we*** to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

Motivation

Understanding the intricate interplay of proteins is crucial to disentangle the mysteries of life. Proteins are the primary molecular players that execute and regulate essential processes within cells. Compared to genomes, proteomes are more dynamic and thus provide an accurate reflection of an organism's functional state. It is natural that proteomics, the systematic study of proteins, has been spotlighted during the post-genomic era.

Proteomics has thrived as large-scale studies using high-throughput technologies such as mass spectrometry have enabled the comprehensive analysis of proteins. This has revolutionized our understanding of biological processes and disease mechanisms, which is only possible through advances in computational techniques and data analysis approaches.

The field of proteomics has evolved a long way, starting from the study of proteins to the study of proteoforms. The term proteoform was suggested to accommodate all the different molecular forms of a protein coded from a single gene[1]. These proteoforms can arise from various factors, including genetic variations, alternative splicing, and post-translational modification (PTM). Several studies have already shown that proteoforms are valuable indicators of diseases and prognostic factors[2]. Moreover, efforts and advances in proteoform research have led the non-profit Consortium for Top-Down Proteomics to launch the ambitious Human Proteoform Project in 2021[3].

Among analytical methods of proteomics, liquid chromatography (LC) coupled with mass spectrometry (MS) has been widely adopted as the primary analytical tool. Conventionally, proteins are digested into peptides before being analyzed by LC-MS, known as the bottom-up (BU) approach (Fig. 1.1). With its well-established technolo-

gies, the BU approach has been widely used; however, it faces inherent limitations when it comes to accurately quantifying and characterizing proteoforms. After identification or quantification, peptides should be mapped back to proteins to measure proteins' presence, which is challenging at the proteoform level; proteoforms often share peptides with each other, especially for the ones from the same gene. This is where top-down proteomics (TDP) comes into play.

Intact proteoforms are directly analyzed through MS in the TDP; therefore, preserving all the information within a proteoform and offering a bird-eye-view of a proteome is possible. Since it's an evolving field of research, different terminologies have been used to describe this approach based on the choice of MS experiment methods for analyzing intact proteoforms (i.e., denaturing top-down or native top-down)[4]. In this thesis, we only consider TDP as denaturing top-down proteoform analysis.

Owing to lots of developments in MS instrumentation and separation techniques, TDP has proven its strength and efficacy in proteoform analysis. It started from targeted analysis of a single proteoform and evolved to discovery analysis across multiple MS runs[5–7]. International efforts to estimate the number of human proteoforms[8] and

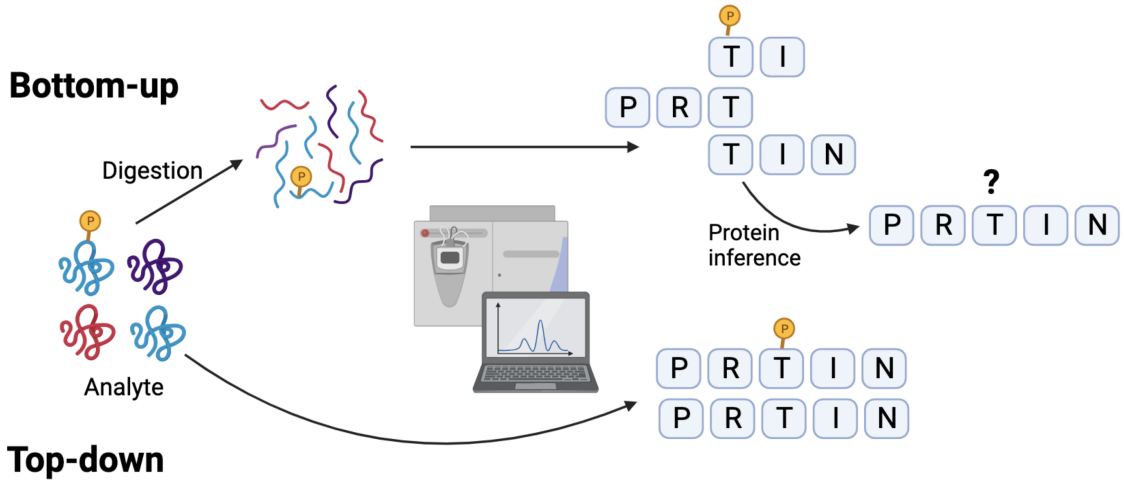


Figure 1.1: Schematic diagram of the difference between bottom-up (BU) and top-down (TD) proteomics. In MS-based proteomics, the BU approach involves enzymatic digestion of proteins into peptides prior to MS analysis. This digestion step enhances the efficiency of MS analysis by generating more manageable analytes. However, it also introduces the challenge of protein inference methods to retrieve protein-level information from peptides. On the other hand, the top-down approach analyzes intact protein without digestion, allowing for direct characterization of proteoforms. Also, keeping the intact protein form enables detailed analysis of PTMs (depicted as yellow-colored P). Created with BioRender.com

suggest a guide for the best practice with TDP[9] are a few examples of milestones in the field, continued through the Human Proteoform Project[3].

Despite the ongoing challenges that remain[10], the field is gaining a lot of momentum. In order to contribute to this progress, we developed three different computational tools for quantitative proteoform analysis in TDP. These tools will be discussed in three main parts of this thesis.

Part 1: FLASHDeconv for the initial step of TDP data analysis

Data analysis is one of the major challenges that need to be addressed in TDP. Improvements in data analysis enabled the identification of more than a thousand proteoforms from human[11] and *E. coli*[12] samples, a major leap forward in qualitative analysis. Yet, significant challenges remain due to the complex nature of proteoforms in MS signals, such as internal ions, chimeric spectra, or ambiguity in PTM localization.

The first step to address any problems in computational analysis (either qualitative or quantitative analyses) is deconvolution, a unique signal processing method that differs from the traditional BU approach. In TDP MS, deconvolution refers to the process of determining the mass of a molecule based on its mass-to-charge ratio (m/z) values, a step to simplifying a mass spectrum/spectra. The high mass of the analyte comes with multiple charge states and isotopes, which leads to multiple signals in MS representing a single proteoform. The heterogeneity of signals is inevitable, making deconvolution precede any further computational methods.

We introduce FLASHDeconv, the tool for high-quality deconvolution in TDP-MS, ensuring ultra-fast runtimes. With the unique idea of transforming a spectrum from m/z values to $log\ m/z$, the deconvolution process is simplified to a constant pattern search. FLASHDeconv detected more genuine mass and minimized mass artifacts than benchmarking tools for both isotopically resolved and unresolved signals. Simple (containing less than six proteins, mass range from 9-68 kDa) and complex (murine myoblast) datasets were used for evaluation.

Part 2: FLASHQuant for the quantitative analysis

A thorough exploration of proteoforms revealing differential expression in diverse conditions offers valuable insights into biological processes and potentially aids in the identification of disease biomarkers. Still, comparative analysis of proteoforms in TDP,

especially at the proteome level, is in its early stages due to low sensitivity on low abundant proteoforms and coeluting proteoforms[7].

Out of all the quantification methods available, label-free quantification (LFQ) is the most widely used due to its simplicity in data preparation and cost-effectiveness. Nevertheless, only a handful of computational methods were introduced to perform LFQ, and most of them are not publicly available or are restricted to specific platforms. Inevitably, many studies rely on deconvolution methods, followed by manual validation or in-house tools for quantification.

We developed FLASHQuant for fast and accurate proteoform quantification, which is equipped with automatic coeluting proteoform resolution. Following feature detection (extracting individual ion chromatograms), FLASHDeconv's algorithm is used for rapid proteoform assignment. Then, coeluting proteoforms are detected and resolved to enhance quantification accuracy. FLASHQuant demonstrates higher reproducibility and closer fold change values to the expected values in both the simple spike-in and complex proteome mixture datasets. Moreover, an algorithm for comparative analysis among different samples is provided. Both algorithms can be executed through a simple Graphical user interface (GUI).

Part 3: GUI to the rescue! FLASHViewer

We have received a tremendous opportunity to work closely with many bench scientists thanks to the project funded by the European Commission. Through our many collaborations, the importance of a GUI has consistently been recognized. A GUI, not only to execute the algorithms but also to show the results in an interactive viewer, was requested.

Also, since the majority of research in TDP still focuses on targeted analysis and manual validation, different types of visualization methods (i.e., bird-eye view in a heatmap or a fragmentation map with protein sequences) have been requested. However, showing all types of visualization at the same time can be confusing to users and raise the hurdle for first-time users. To this end, a visualization with a modularized and configurable layout was needed.

Based on the feedback, we developed FLASHViewer, a web application with configurable layouts for visualizing signals in TDP MS. It was implemented based on the pyOpenMS Streamlit project and can be executed locally or accessed via online `https://abi-services.cs.uni-tuebingen.de/flashviewer`. Using Streamlit, a Python framework for building web applications, fast and easy implementation was possible, and Typescript was used on the side for rapid plotting of figures and

tables. Result files from both FLASHDeconv and FLASHQuant can be used as input, leading to different layout options.

Chapter 2

Background

2.1 Mass spectrometry-based top-down proteomics

For a comprehensive and detailed understanding of proteomes, mass spectrometry (MS) has been used as a core technique to identify and quantify proteins. Technological advancements in MS enabled high-resolution and high-throughput proteomics analyses and influenced the beginning of top-down proteomics (TDP). TDP analyzes the intact proteins through simplified sample preparation steps by avoiding the digestion of proteins into peptides. While having the capacity to provide comprehensive information on proteome at the proteoform level, the inherent high complexity of the analytes in the MS signals requires appropriate separation of the analytes.

Basic experimental techniques in MS-based TDP include sample preparation, separation, and MS analysis. The sample preparation step involves extracting the analyte of interest and preparing it for further analysis. Then, the analyte is subjected to separation techniques to separate the complex protein mixtures based on their physicochemical properties. MS ionizes and detects the separated analytes and outputs information on their mass-to-charge ratios (m/z) and abundances. This section will address a brief overview of the general experimental techniques used in this thesis.

2.1.1 Sample preparation

The sample preparation step aims to convert biological samples suitable for MS analysis. It starts either by collecting raw biological materials or by culturing cells/tissues. Analytes then undergo various extraction, purification, or enrichment steps, depending on the chosen analyte and the used techniques. A universal workflow for any proteome hardly exists due to this variability. Once proteoforms of interest are carefully depleted of contaminants or undesirable components, they are subjected to denaturation, a

process to break down their tertiary structure. Skipping this denaturation step results in the approach called Native MS instead of TDP[4], another emerging field of proteomics. The denaturation step linearizes proteoforms, exposing their residues for further analysis.

In the traditional bottom-up (BU) approach, the next step is for proteins to be digested into peptides; however, proteins are kept intact in TDP. Also, when labeling is required for quantification purposes, various techniques are applied at this stage, either *in vivo* or *in vitro*. As this thesis concentrates on the label-free approach for quantification, no specific labeling techniques will be discussed in detail but roughly described in Section 2.2.3.

2.1.2 Separation techniques

The complexity of the proteome, a significant challenge in TDP, necessitates the separation of proteoforms prior to MS analysis. Compared to peptides, proteoform mixtures are less complex. Still, it has diverse physicochemical properties (e.g., size, charge, and hydrophobicity) and wide dynamic ranges of proteoform expression[6], leading to low sample homogeneity and separation efficiency[7]. Therefore, reducing the sample complexity through separation techniques is critical to enhance sensitivity and broaden coverage.

The most prevalent separation technique is liquid chromatography directly coupled to online mass spectrometry (LC-MS). In this method, a liquid sample (solubilized proteoforms) is injected into the LC system, where it is first solved in a pressurized liquid (mobile phase) and then pumped through a column filled with solid adsorbent material (stationary phase). The physicochemical properties of each proteoform determine how strongly it interacts with the stationary phase, leading to their different time spent in the phase (i.e., different retention times), thus enabling their separation.

Particularly, high-performance liquid chromatography (or even, ultra-high-performance liquid chromatography) has been widely used, in which high pressure is applied to the mobile phase to push through a column, allowing for faster separations and higher resolution. Also, reversed-phase chromatography, with opposite polarity arrangement in the two phases compared to normal phase, is commonly used in TDP[13] due to its compatibility with online MS detection[14]. In this context, "online" means that the chromatographic separation is directly coupled with the MS instrument, resulting in real-time analysis and improved efficiency. This direct coupling is a key factor contributing to the popularity of LC.

To further alleviate the sample complexity, pre-fraction techniques can be applied even before chromatography techniques. This fractionation technique adds another dimensional separation approach and divides analytes into fractions of similar physicochemical properties such as size or isoelectric point. Each fraction is subjected to further separation (i.e., LC), or only a few fractions are selected as an enriched sample for subsequent analysis (e.g., molecules lighter than <30 kDa).

Gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) separation was suggested to be used prior to LC for TDP, which yielded the highest number of characterized proteoforms[9]. GELFrEE is one of the most common pre-fractionation techniques that separates proteins based on their sizes. As the wide protein size range (from smaller than 1 kDa to larger than 30 kDa) increases the spectral interference, larger proteins are often overshadowed by smaller ones in MS detection. Additionally, larger proteins naturally exhibit lower MS signals due to wider distributions of isotopes and charge states. Size-based fractionation enables extracting small or large-sized proteins for further analysis and was also utilized for the dataset in Chapter 4 for enrichment purposes (Section D.1).

2.1.3 Mass spectrometry

A mass spectrometer measures and determines the m/z and abundance of analytes. As it manipulates charged particles, the analyte solutions from LC are required to be ionized first. Here, the setup of the Thermo Fisher Quadrupole-Orbitrap MS[15] will be discussed as it is the instrument used in this thesis (Fig. 2.1).

Electrospray ionization

Electrospray ionization (ESI)[16] is the choice of ionization technique for TDP, allowing the efficient interface with LC and analysis of intact and large molecules including proteins[13]. As the name hints, a liquid sample is sprayed into a high-voltage electric field to be ionized. While droplets of fluid exit the needle-like capillary, a high voltage is applied at the tip, causing the droplets to become charged. At this stage, a droplet carries numerous charged ions in the form of aerosol. Droplets further evaporate to form even smaller droplets, often assisted by heated gas for accelerated evaporation. The positively charged ions in the droplets repel each other so that they push themselves further to the edge of the droplet. When the repulsion between them reaches a critical point, the droplet disintegrates, resulting in gas-phase ions.

Mass analyzer

Ions are filtered, measured, or detected in mass analyzers based on their m/z. Various techniques exist for this purpose, such as time-of-flight, quadrupole, and Orbitrap[17] analyzers. Modern MS instruments often combine these analyzers to achieve different purposes. The type of mass analyzer used is a crucial factor, such that mass spectrometers are commonly named after the specific analyzer they employ. In the case of Quadrupole-Orbitrap MS, the quadrupole serves as a mass filter, and the Orbitrap functions as the mass detector.

Fig. 2.1 illustrates how Quadrupole-Orbitrap MS analyzes ions. By the time ions reach the inlet of the mass filter (quadrupole), uncharged neutral ions are all filtered out.

A quadrupole consists of four parallel metallic rods in a vacuum and space between them, through which ions pass. When voltage is applied to it, ions having m/z within a certain narrow range can maintain stable motion and pass through; otherwise, they move unstably and are filtered out. Thus, a quadrupole analyzer can perform as a detector, but its primary function in this setup is to filter ions based on their m/z. The usage of this mass filter will be discussed in the following subsection.

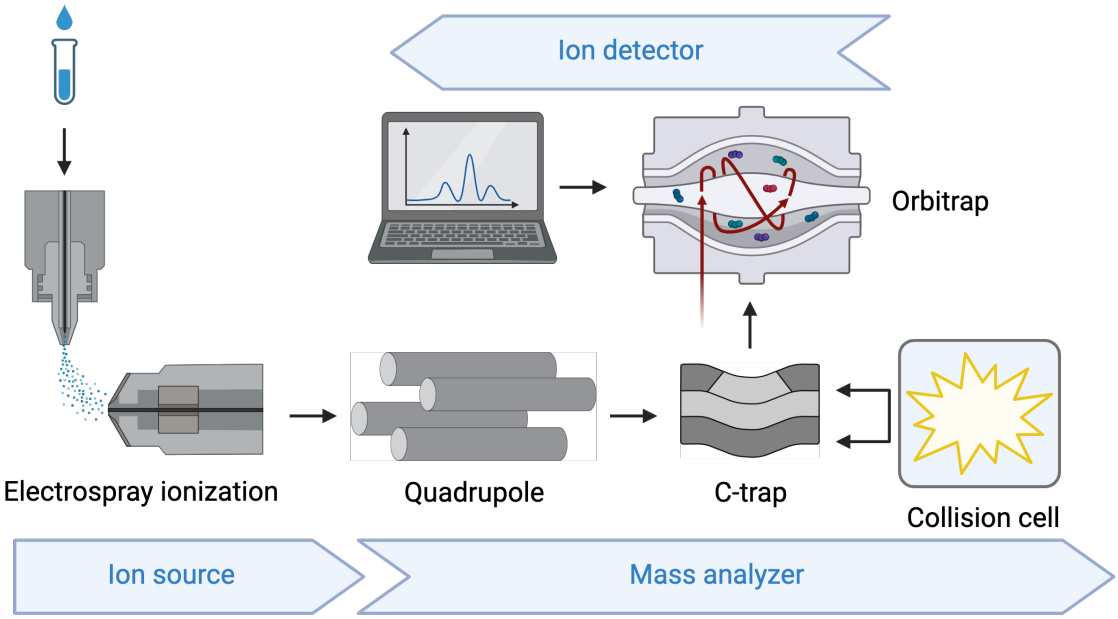


Figure 2.1: Illustration of Quadrupole-Orbitrap MS setup. MS usually consists of three components: ion source (i.e., ESI), mass analyzer, and ion detector. The route from the C-trap to the collision cell is only visited for MS/MS analysis. Created with BioRender.com

Filtered ions are collected (for a specific time range) and stabilized in a *C-trap* before being directed to the Orbitrap for detection. The Orbitrap analyzer traps and detects ions based on their orbital motion. Ions orbit around a central electrode, and the frequency of this orbital motion is proportional to their m/z values. Therefore, ions with different m/z values experience distinct frequencies and are separated. This allows accurate and high-resolution mass measurement. The m/z values and intensity of detected ions are collectively recorded in a mass spectrum, which will be discussed in Section 2.2.

Tandem mass spectrometry

A mass spectrometer cannot only measure intact ions with a wide range (*MS1* or *full scan MS*) but also break up a targeted ion and measure the resulting fragments. This technique, known as *tandem mass spectrometry (MS/MS)* or *MS2*, allows for the structural analysis of molecules, supporting their identification. MS1 is often enough for quantitative analysis; however, for qualitative analysis or evaluation purposes, MS2 is highly valuable. As isobaric molecules are indistinguishable at the MS1 level solely by their masses, their structures (amino acid composition and modifications) become essential for distinguishing them. Especially in TDP, whose characteristics allow detailed analysis of multiple post-translational modification (PTM)s, exhaustive information on protein sequence is crucial to pinpoint the locations of PTMs[18].

The standard method for MS/MS involves two cycles of mass measurement, and Fig. 2.1 will be used to explain the process. The first cycle (MS1) measures and detects all ions across the entire mass range of interest, generating a full scan mass spectrum. From this spectrum, specific m/z values of ions (namely *precursor ions*) are selected for the second cycle, often the most intense n different ion types. In MS2, the analyte enters the mass spectrometer, and only the ions with the selected m/z values survive through the mass filter (here, quadrupole). These precursor ions are then sent to the collision cell, where they are fragmented. The packet of fragments is sent back to the C-trap and stabilized. As in MS1, the Orbitrap analyzer detects the m/z and intensity of the fragmented ions and records them in an MS2 mass spectrum.

A protein sequence can be inferred from the MS2 mass spectrum by examining the pattern of fragment ions, the process called *identification*. It involves calculating the mass differences between the fragment ions and comparing them to the theoretical masses of amino acids.

2.2 Computational mass spectrometry in top-down proteomics

2.2.1 Basic concepts and terminology

The extensive progress in MS surpassed manual analysis capabilities, leading to the inevitable advent of computational methods for data analysis. Consequently, computational mass spectrometry has evolved as an independent scientific field, mainly focusing on identifying, quantifying, or visualizing molecules. Most of the basic ideas were adapted from the conventional BU methods but had to be tuned to accommodate the distinctive characteristics of TDP. This section will exclusively discuss concepts and methods relevant to TDP based on LC-MS for quantitative analysis.

LC-MS map and Mass spectrum

The output from an MS instrument, serving as the primary input for computational MS, consists of time-stacked mass spectra, commonly known as an *LC-MS map* or *peak map* (Fig. 2.2). The instrument records each mass spectrum for a specific interval time, typically lasting less than a second per spectrum. A retention time (RT) of a spectrum represents the analyte's duration to pass a chromatographic column; thus, signals from the same proteoform, for instance, linger in a similar retention time range.

A mass spectrum is made up of so-called *peaks*, each of which has a mass-to-charge ratio (m/z) value and intensity, i.e., two-dimensional information (three-dimensional information for LC-MS map: m/z, intensity, and RT). Therefore, a mass spectrum is often represented as a 2D vertical line graph (Fig. 2.2, upper right).

A raw spectrum from the MS instrument, in reality, does not contain discrete peaks from ions as m/z-intensity pairs. Rather, it shows a continuous distribution of signals referred to as a *profile spectrum*. The signal corresponding to a specific ion is distributed around the actual m/z value of that ion. The width of the distribution around that m/z value (peak width) and the accuracy of distribution (peak's m/z value) depend on the instrument's resolution or setting. This profile spectrum is processed to convert it into discrete peak values (as shown in Fig. 2.2, bottom right), which reduces its complexity. The process is called *peak picking* or *peak centroiding*. In this thesis, we assume that all spectra are centroided.

Isotope patterns

When ions from a specific proteoform with a particular charge state are detected in a mass spectrum, they don't appear as a single peak. Due to the presence of naturally occurring isotopes, they appear as a cluster of peaks, representing their isotope pattern.

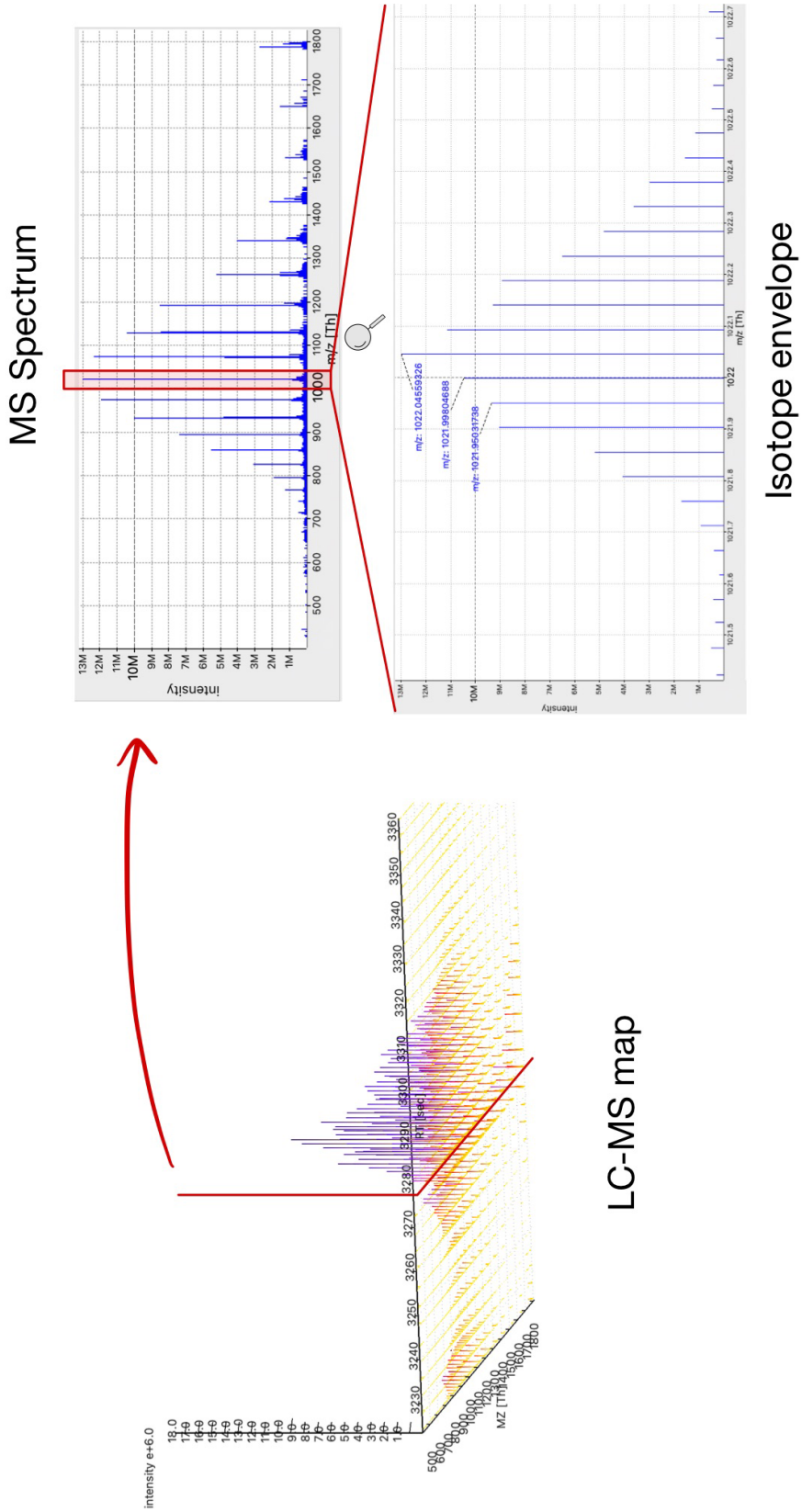


Figure 2.2: Visualization of LC-MS map and mass spectrum. The three-dimensional representation of MS signals on the left depicts a partial LC-MS map from a simple dataset (a six-protein sample from Chapter 4). One retention time from the map is selected to show a mass spectrum in the upper right. The red box on the spectrum is zoomed in to display how many peaks reside in a small region of the spectrum (bottom left). From the isotope envelope, distances between m/z values of the peaks can be used to infer their charge state, 21. See the below sub-section for details on isotopes. LC-MS map and spectra were drawn with TOPPView[19], which will be discussed in Section 2.2.4.

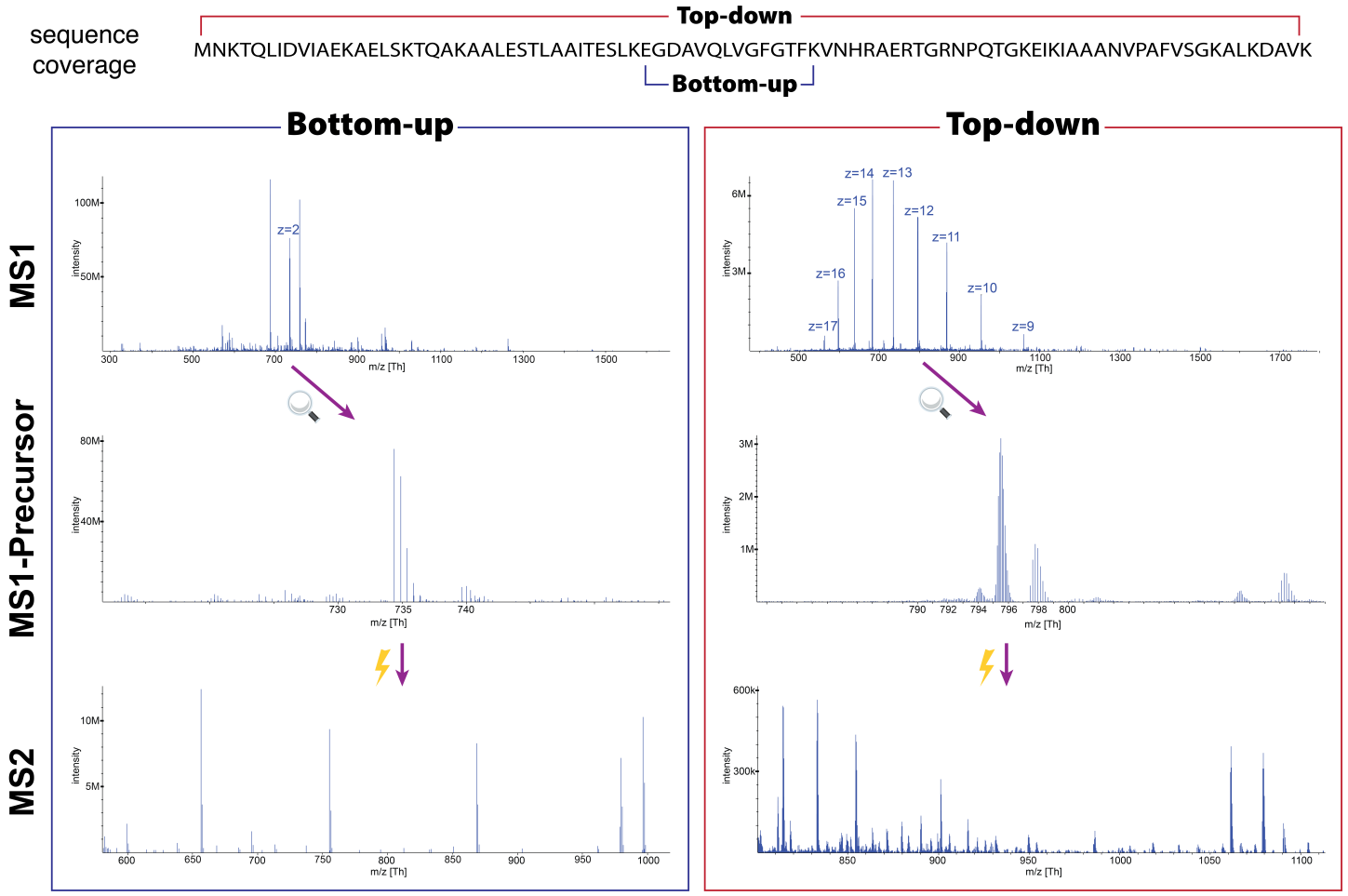


Figure 2.3: A raw signal comparison between BU and TD approaches. The figure highlights distinct differences in the analysis of the DNA-binding protein HU-alpha from *E. coli* (UniProt accession: P0ACF0). In the TD approach, the entire sequence of the intact protein is analyzed (as shown in the top panel), with MS1 spectra showing multiple charge states ranging from +9 to +17 dedicated to a single protein. In contrast, the BU approach covers only portions of the protein (i.e., peptide), and the MS1 spectra display ions with smaller charge state (+2). The precursor ions selected for MS2 analysis also differ (spectra in the middle): in BU, 4–5 isotopic peaks represent the precursor peptide ion, reflecting its relatively low mass, while in TD, more isotopic peaks form a Gaussian shape. In the MS2 spectra within similar m/z ranges (300–400 Th), the BU approach yields sparse fragmentation peaks, whereas the TD approach produces a denser array of peaks. For the TD analysis, the dataset from Chapter 4 (ratio 1/5) was used, and for BU, the public *E. coli* dataset was taken from Doellinger et al.[20].

The peaks in the pattern, known as *isotopic peaks*, are, thus, the occurrence of a molecule's different isotopologues. They have different masses based on the isotope mass shifts, and their intensities are influenced by the abundance of isotopes.

Isotope pattern often exhibits in the form of Gaussian, so-called *isotope envelope* (as shown in Fig. 2.2, lower right) when the molecule is large enough (i.e., proteoform, see the comparison between BU and TDP in Fig. 2.3). Thus, this shape of isotope patterns can be calculated using models like the "averagine" model[21], based on the average elemental composition of a protein molecule, if only the mass of the molecule is known. The averagine model will be utilized throughout this thesis.

Feature detection

While a mass spectrum reveals m/z-intensity relationships of LC-MS data, a *chromatogram* represents the intensities of signals detected over time (Fig. 2.4, lower left). Summing all the signal intensities at each time point (i.e., *total ion current (TIC)*) provides an overview of the chromatographic separation of the LC-MS experiment, as shown in Fig. 2.4, upper left. However, to correctly detect and measure each proteoform, m/z information should be involved due to the frequent co-elution of multiple proteoforms within a similar time window. The chromatogram that isolates a specific analyte of interest by extracting its m/z values is known as an *extracted ion chromatogram (XIC)*. This is often a unit for quantifying proteoforms (area under the profile, Section 2.2.3) and usually appears as a skewed distribution with a long tail.

Individual ion chromatograms (i.e., *mass traces*) corresponding to a single analyte with a specific charge state are collectively called a *feature*. Thus, the mass traces within a specific feature form an isotope pattern, which correlates with their m/z value and intensity, contingent on their isotopes, as shown in the lower right of Fig. 2.4. *Feature detection* is a method of identifying all peaks that describe these distinct features, commonly used for quantification. Since the expected shape of a feature can already be estimated, the feature is detected by matching a mathematical model to the relevant peaks. For individual proteoform, multiple features should be detected to correctly measure the quantity of it.

2.2.2 Deconvolution

Deconvolution is the first crucial step for TDP data analysis due to the unique nature of its analyte compared to the conventional BU approach (Fig. 2.3). This step aims to resolve the complexity by transforming peaks at the m/z level to masses. As a proteoform (or fragments of it) comes with multiple charge states and isotopes, the high

complexity of MS signals challenges accurate mass determination. Therefore, grouping corresponding peaks into a single mass must precede to reduce data complexity in subsequent computational methods.

Deconvolution can be roughly performed at two levels: feature level or spectrum level. When the input is an LC-MS map, deconvolution determines which features are related and together describe individual proteoforms, similar to feature detection. Spectral deconvolution, on the other hand, determines masses of proteoform or fragment (depending on whether MS1 or MSn) from a mass spectrum.

Since the introduction of THRASH[22], the first deconvolution algorithm for TDP data analysis, various deconvolution algorithms have been developed. However, two key steps have remained unchanged: *decharging* and *deisotoping*. The decharging

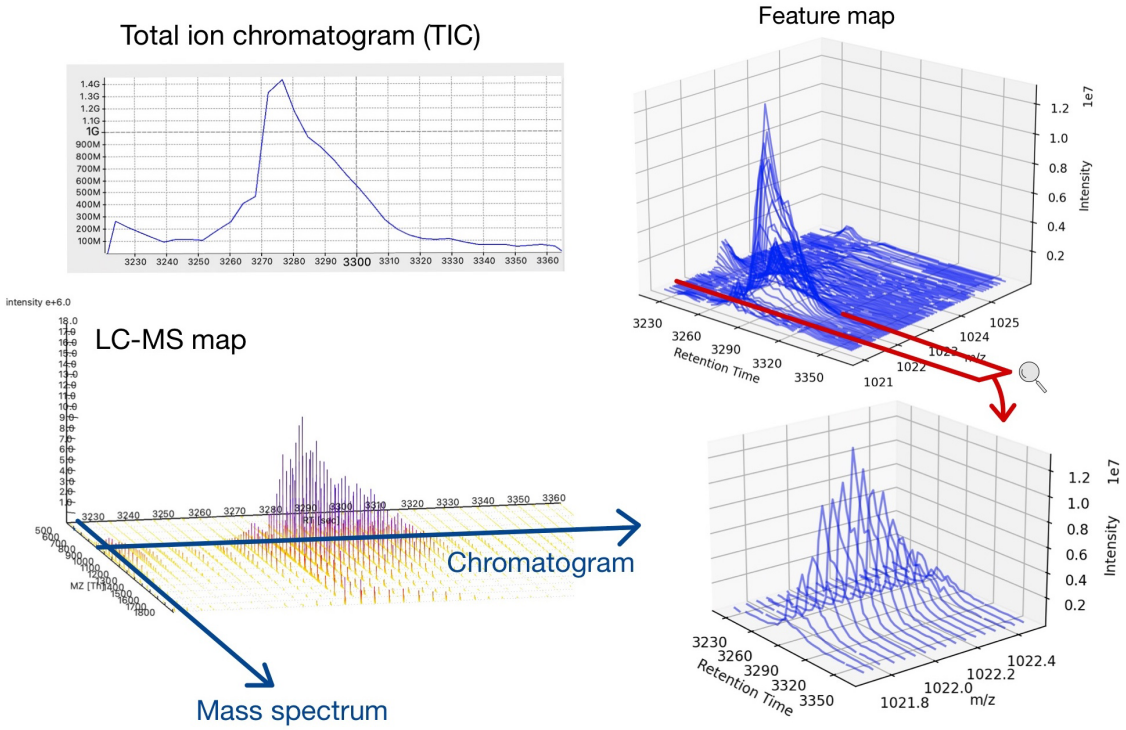


Figure 2.4: From LC-MS map to features. LC-MS maps are commonly interpreted at the chromatogram level for quantification, rather than at the mass spectrum level (lower left). TIC provides a comprehensive chromatographic overview of the experiment, allowing us to observe the eluted time points of the analyte. In the figure, the RT axis of the TIC is the same as that of the LC-MS map. A small m/z region in the LC-MS map has been chosen to show a feature map of detected mass traces (the mass traces are detected by MassTraceExtractor from OpenMS). Despite the narrow m/z range of 1,021-1,026 Th, a considerable number of mass traces were detected (upper right). Within the feature map, an even smaller m/z range was selected to draw a single feature(lower right). Note that the m/z range of a feature corresponds to that of the isotope envelope shown in Fig. 2.2. The same data from Fig. 2.2 was used.

step assigns charge states on each peak, while the deisotoping step places peaks in corresponding isotopic envelopes.

Decharging has been traditionally performed by calculating the m/z distance between two consecutive isotopologues. Given uncharged, the m/z distance between them is estimated constant, the mass difference between ^{13}C and ^{12}C (denoted Δ). Thus, when the observed m/z distance is x, $x = \Delta/z$, so that z can be inferred. This method is effective when the expected charge range is small. However, isotopically resolved data, in which isotopic peaks can be distinguished and separated, is required. High-mass molecules have higher charge states, and, in turn, the distances between their isotopic peaks are very narrow. Thus, for the large molecules (often >30 kDa), MS data becomes inevitably isotopically unresolved even with high-resolution instruments, making it challenging to use this method for decharging.

Alternatively, charge patterns (patterns of the peaks from the same mass but of different charges) can be pre-generated and correlated with peaks to assign charges. A charge pattern involving mass m and consecutive charge states $z, z+1, z+2$, is expressed as $m/z, m/(z+1), m/(z+2)$. This charge pattern is scanned across all the peaks for matching. While particularly valuable for isotopically unresolved data, this method becomes computationally demanding when the mass is unknown, especially when attempting to cover a wide mass range. In Chapter 3, we will introduce a smart idea for decharging utilizing a universal charge pattern that requires minimal computation.

Deisotoping involves matching theoretical isotope envelopes to observed peaks, a process conducted either concurrently with or after the decharging step. In the former approach, isotope envelopes are identified at the m/z level, as in the decharging method using isotope distance. The latter approach is performed at the mass level after the peaks are decharged. Thus, it is more effective in cases of poor isotopic resolution and demonstrates robustness against noise.

2.2.3 Quantification

Quantitative analysis in proteomics allows comprehensive investigation of proteins differentially expressed under various conditions. This investigation can provide valuable insights into biological processes and potentially facilitate the discovery of disease biomarkers.

Quantifying proteoforms is commonly performed by measuring the relative abundances of proteoforms in analytes. As an alternative, an absolute measurement of proteoform abundance can be conducted, which requires internal standards to be

mixed with an analyte in advance of MS analysis. Then, the ratio of intensity between the analyte and the standard is calculated for quantification. Without internal standards, absolute quantification is challenging and tends to have low accuracy due to the varying ionization efficiencies of different proteins.

Proteoforms can be relatively quantified within the same run or across multiple runs (Fig. 2.5). Within the same run, the methodology involves gauging the abundance of ionized proteoforms and subsequently comparing them with each other. However, each proteoform may have different ionization efficiencies and sensitivities, leading to intensity differences that do not necessarily reflect their actual abundances. To address this issue, this method is often performed between isotopically labeled versions of the same proteoform, which are expected to ionize similarly. For instance, different samples are chemically labeled with isotopic tags (which will be explained in the following section) and combined for analysis in a single MS run. The same proteoform, now

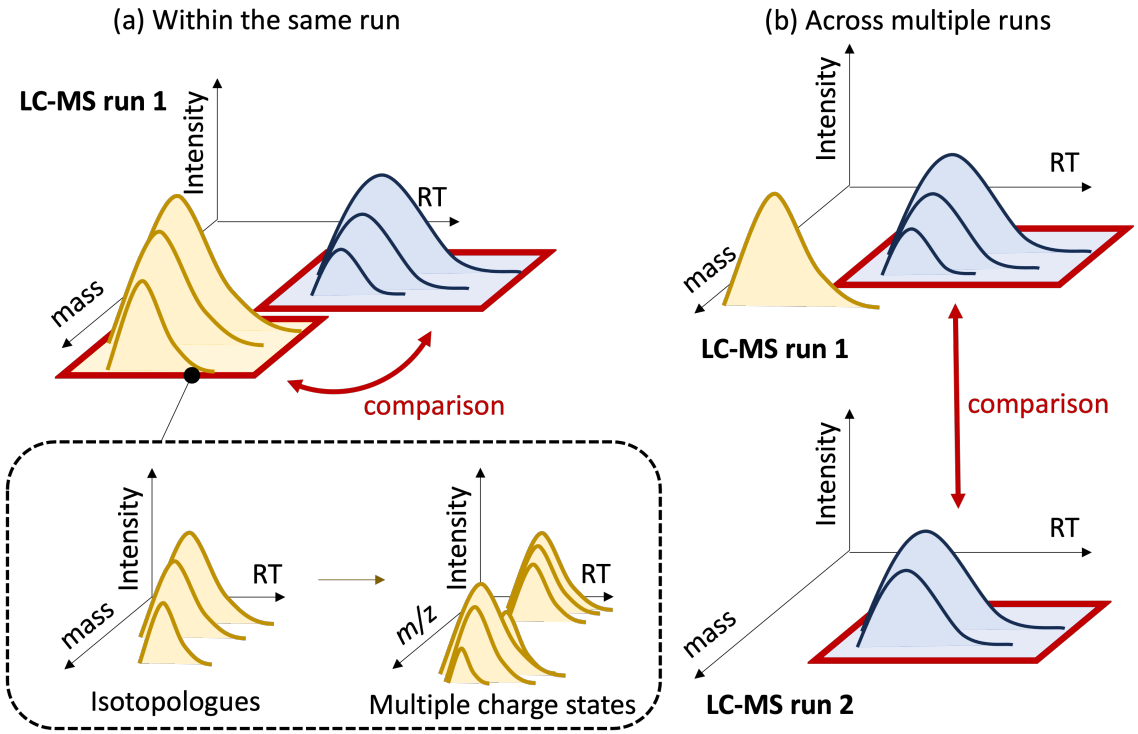


Figure 2.5: Relative quantification methods. Proteoforms can be compared either (a) within the same run or (2) across multiple samples. In the case of within the same run, different proteoforms are compared with each other, whereas, in the across multiple sample scenario, the quantities of the identical proteoforms are compared. Note that quantification values rely on measuring the area under the elution profiles, with multiple elution profiles for each proteoform (including those from isotopologues and multiple charge states) involved in this calculation, as shown in the dashed box.

carrying different masses due to the labels, is detected and quantified for comparison across samples. This method is particularly useful in targeted analyses where the proteoform of interest is predefined. In untargeted analyses, proteoform identification using MS/MS is often necessary to increase confidence in the quantified proteoforms, as they may otherwise exist as anonymous mass features without fragmentation evidence.

On the other hand, when quantifying across multiple runs, the comparison extends to finding proteoforms that are jointly detected across MS runs and then comparing their abundances among the samples. Since this method does not require specific experimental modifications (e.g., isotopic labeling), there is flexibility in the number of involved samples for the analysis. This method often involves addressing potential missing values (i.e., proteoforms) or retention time alignment among samples. We developed FLASHQuant in Chapter 4 for the detection of quantifiable proteoform features within the same run and TopDownConsensusFeatureGroup in Section 4.2.2 for quantification across multiple samples. Unless otherwise stated, relative quantification within the same run will be discussed in the following sections for clarity.

Quantification techniques

Various quantification technologies have been developed, which can be classified into two approaches: label-free and labeling methods. The label-free method, as the name hints, refrains from introducing any labels to the analyte, making it a cost-effective and straightforward approach. On the other hand, labeling methods incorporate isotopic labels or tags into the proteome samples, thereby allowing the combination of differently conditioned samples in a single MS run. This, in turn, reduces the technical variation between conditions. Commonly used labeling methods include metabolic labeling and isobaric chemical labeling[7,23].

Metabolic labeling methods label proteins *in vivo*; that is, an organism is supplied with labeled compounds so that it expresses labeled proteins during cellular metabolism. The most widely used metabolic method, stable isotope labeling by amino acids in cell culture (SILAC)[24], for example, involves a medium containing stable isotope-labeled amino acids. Labeled and unlabeled cultures are combined early in the experimental workflow, allowing them to undergo similar sample processing and be measured in a single run. The abundance ratio between labeled and unlabeled proteoforms is utilized for relative quantification.

Chemical labeling strategies introduce specific chemical tags into proteins, enabling quantification based on their mass shifts. Labeling is performed *in vitro*, in the later stage of the workflow, by applying chemical agents that react with specific amino acids.

Successful usage in BU approach notwithstanding, chemical labeling is still underdeveloped in TDP due to its inherent heterogeneous signals[7]. When chemical labeling is applied, the already high MS signal complexity increases significantly, causing isotope patterns to overlap with others.

Labeling methods were initially devised to improve precision and accuracy in quantification. However, they face challenges related to potential signal overlap between proteoforms from different conditions, along with the necessity of extra algorithms for detecting labeled pairs. The advantages of the label-free method, such as simplicity and (theoretically) unlimited number of samples, come into play and make it the most common approach in proteoform quantification. Let us spotlight the label-free approach more, as this thesis will introduce computational methods for analyzing label-free MS data.

Label-free quantification

In label-free quantification (LFQ), an abundance of proteoform can be measured by two different approaches: spectral counting and MS signal intensity-based. The spectral counting approach regards the number of identified spectra for a given proteoform as its relative abundance. This is based on the premise that the more abundant a proteoform is, the more likely it is to be selected as a precursor ion for fragmentation and a subsequent higher number of identifications. The idea is intuitive and enthralling and showed its potential in TDP[25]. However, this approach is still controversial with its relatively low accuracy and its reliance on the assumption that each proteoform has an equal chance of being selected for fragmentation, which is often not the case (i.e., due to ionization efficiency or utilization of dynamic exclusion list).

The intensity-based approach is predominantly employed in LFQ, and the abundance of a proteoform is inferred from peak intensities or chromatographic peak areas from MS1 scans. Generally, three steps are involved in this approach: feature detection (i.e., deconvolution), intensity calculation, and statistical analysis[7]. Feature detection must precede as multiple m/z values represent a single proteoform (i.e., from multiple charge states and isotopes). Also, coeluting proteoforms can lead to overlapping MS signals or signal suppression, resulting in inaccurate intensity measurements. Depending on the algorithm, the proteoform intensity is calculated either via the summed intensity of all related peaks (often across the entire MS scans) or via the area under the relevant XICs/traces. To take into account the nature of the elution profile, the area under the elution profile method is preferred. Lastly, various statistical analysis meth-

ods can be applied to the quantified proteoforms in order to assess their significance or further reduce false positives.

For large-scale proteomics studies, LFQ is often utilized for quantification across samples, comparing proteoform abundances among multiple LC-MS runs. Proteoforms with matching mass and retention time from multiple samples, namely consensus proteoforms, are detected to calculate the relative abundance of each proteoform across samples. Even though this can improve the accuracy and reliability of quantification, computational challenges arise for correctly aligning retention time across multiple samples, as reproducibility among runs is often low in TDP.

Advances in TDP LFQ computational methods

Several leaps have been made in TDP LFQ studies in the past (nearly) two decades. In the early stage, unmodified and modified intact proteins were quantified and compared to each other to evaluate the importance of PTMs under different biological conditions. Relative quantification between unmodified and modified proteins was shown to be less disturbed by ionization efficiency than that on peptides, according to Pesavento et al[26]. Also, since these proteoforms are often co-eluted, they can be quantified intra-spectrum (i.e., in an averaged spectrum), which minimizes experimental variability. As a quantification value for a proteoform, peak heights from the top five most abundant isotopes with a certain charge state were integrated.

Such a method was proven effective within samples with limited complexity, such as those in studies focusing on specific proteins or proteoforms. For instance, Pesavento et al. demonstrated the quantitative analysis of modified and unmodified human histone H4 proteoforms, effectively distinguishing between different modification states[26]. Similarly, Ayaz-Guner et al. analyzed a target protein complex, cardiac troponin I from a mouse heart sample, comparing it with its mono- and bis-phosphorylated forms[27]. Their work successfully identified and quantified specific phosphorylation sites. Peng et al. performed deep sequencing of tropomyosin protein and isoforms purified from swine hearts, achieving detailed characterization of protein variants and discovering a novel isoform in controlled sample conditions[28]. These studies highlight the effectiveness of such quantitative methods when applied to targeted proteoform analysis.

Advances in deconvolution methods have enabled quantification utilizing multiple charge states for each proteoform within a single LC-MS run. Ansong et al. employed ICR-2LS (currently retired deconvolution software, unpublished), in which all spectra within a run were summed together into one spectrum, and then deconvolved[29].

Relative quantification was performed based on the deconvolved mass and abundance from ICR-2LS.

Some studies used the identified proteoforms from tandem mass spectra as a starting point for quantification. For example, IPQuant took the identified proteoform masses to retrieve XIC positions and then calculated the area under the curve to determine the abundance of each proteoform[30]. Similarly, DiMaggio et al.[31] and Holt et al.[32] quantified highly modified proteins by calculating ratios of fragment ions at the MS2 level and integrating these ratios throughout an LC-MS run to determine the relative proteoform quantity. While this method is valid for targeted analysis, fragmentation efficiency can limit the quantification sensitivity.

In 2014, the first analytical platform for the large-scale LFQ study was published by the Kelleher group[33]. Comparative LFQ analyses, comparing the abundances of jointly detected proteoforms across multiple samples, had been conducted previously, such as the differential mass spectrometry method adapted from BU approach[34]. However, most were limited to targeted proteins and a small number of proteoforms[33]. This new study demonstrated the potential of comparative LFQ in discovery mode on two yeast proteome samples with high reproducibility.

The Informed-Proteomics software package, consisting of algorithms for feature finding, identification, and interactive viewers, showed its potential in LFQ analysis on two breast tumor samples[35]. Besides being open-source software, it is also accompanied by the automated feature align method across samples. Another open-source software with multiple functionalities, Proteoform Suite, was extended to enable LFQ research across multiple samples[36]. Despite its uniqueness in showing "proteoform families," it requires deconvolution and identification results from other software (not publicly available) to be executed.

In fact, numerous LFQ research have applied commercial software provided by MS vendors, including DataAnalysis from Bruker[37], Protein Deconvolution from Thermo Fisher[36], Xtract from Thermo Fisher[33]. Moreover, most of their relative quantification across samples was performed manually or through in-house script, such that replicating the research would be challenging. Moreover, the known problem of overlapping signals in TDP (i.e., co-elution) notwithstanding, an automatic resolution has not been proposed. In Chapter 4, we will discuss the solution to this problem in detail by suggesting the new open-source and fast quantification tool equipped with overlapping signal resolution, FLASHQuant.

2.2.4 OpenMS framework

Many of the aforementioned concepts or algorithms have existed for over a decade and have been implemented in various software tools. Among them, OpenMS[38] stands out with its versatility, serving as both a framework and an integrated software suite for LC-MS data analysis. As an open-source and platform-independent software implemented in C++, OpenMS offers excellent building blocks for new algorithms or workflows. Also, it provides independent tools that can be executed individually accompanied by visualization tools.

In OpenMS, there are five main layers, depending on the usage, that can interplay: the OpenMS core library, The OpenMS PiPeline (TOPP)[39] tools, graphical applications, pyOpenMS[40], and workflows.

- The **OpenMS core library**, with more than 1,300 classes, is written in C++. It has a wide range, from file input/output operations to advanced data processing algorithms such as peak picking and quantitation.
- **TOPP**, previously known as The OpenMS Proteomic Pipeline, tools are developed for specific tasks; thus, each tool has its own input and output file formats, as well as a unique parameter set. They can be used as standalone command line tools or as components of a workflow/pipeline.
- Users can interact with TOPP tools and examine mass spectrometry data through **graphical applications**, such as TOPPView[19] (e.g., Fig. 2.2).
- **pyOpenMS** is a Python library to interface the functionalities of OpenMS. It expands the usage of OpenMS functionality with Python frameworks, such as building a web application (e.g., UmetaFlow GUI[41]).
- **Workflows** enable high-throughput data analysis through the automation and combination of multiple tools in OpenMS. Examples of supported workflow systems are KNIME[42] and nextflow[43].

This thesis discusses three computational tools implemented based on and as a part of OpenMS. The first tool, FLASHDeconv (Chapter 3), initiated a new top-down proteomics branch in OpenMS. FLASHDeconv and FLASHQuant (Chapter 4) were developed as new TOPP tools. Furthermore, FLASHViewer (Chapter 5) was built upon the pyOpenMS Streamlit template, and it now operates as a new web application example within the OpenMS framework.

Chapter 3

Fast and robust algorithm for deconvolution

Adapted with permission from:

FLASHDeconv: ultrafast, high-quality feature deconvolution for top-down proteomics
Kyowon Jeong+, Jihyung Kim+, Manasi Gaikwad, Siti Nurul Hidayah, Laura Heikaus, Hartmut Schlüter, Oliver Kohlbacher
Cell Systems 10(2):213-218 (2020)[44]
+These authors contributed equally

3.1 Introduction

TDP has gained a lot of momentum for in-depth protein characterization and protein species analytics[5,6,9,23,45–47]. In contrast to BU proteomics, in which proteins are enzymatically digested and peptides are actually analyzed, the top-down (TD) approach allows for the analysis of intact proteoforms – distinct protein species arising from the same gene product via splice variants, genomic variation, PTM, degradation, etc[23,48]. However, because of the high mass of the analytes in the mass-spectrometry based top-down approach (TD-MS), a single protein species can result in multiple MS features with different charge states and isotopes, making the signal structure highly complex and, in turn, the accurate determination of proteoform masses challenging[49] (Fig. 3.1). Feature deconvolution (i.e., determination of the intact proteoform masses) is thus an essential step for TD data analysis[50,51]. Note that a mass feature refers to collective

signals indicating a tentative proteoform, including all charge states and isotopologues, in this chapter.

Current MS instrumentation and experimental protocols enable more complex analyses using TDP but yield equally complex and large datasets[5,9,23,45,50,52]. Hence, both the quality and runtime of deconvolution algorithms have become an issue. Current deconvolution algorithms typically have long processing times of hours to days for complex datasets. Quality is another issue of these approaches. Frequent mass artifacts as well as narrow mass and charge ranges reduce the methods' applicability[50,51,53].

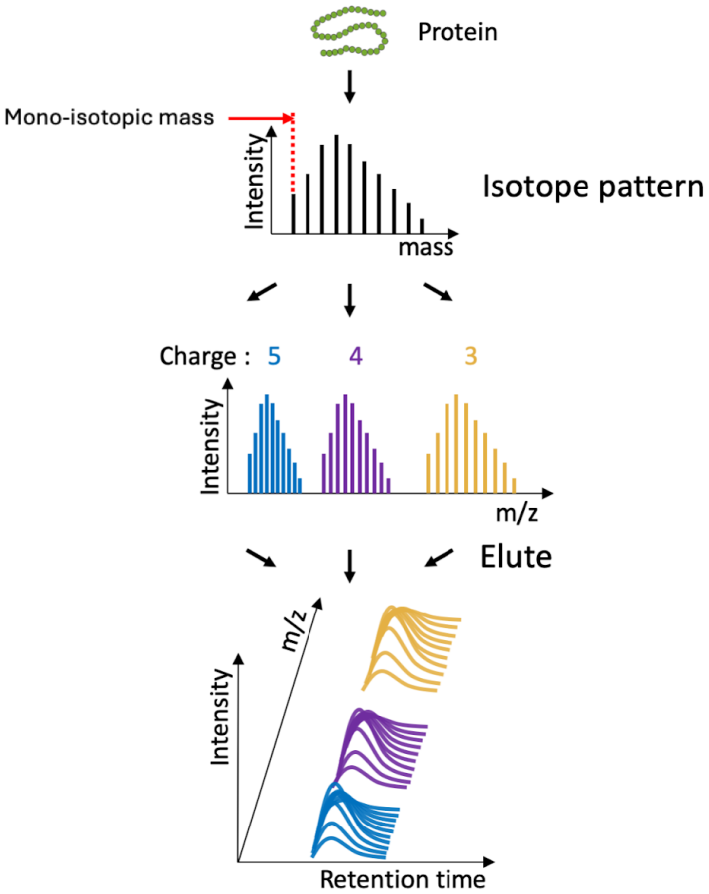


Figure 3.1: Generation of proteoform signals from an analyte in FLASHDeconv. This figure illustrates the generation of the total signal from a single protein species (proteoform) in FLASHDeconv. Due to its large mass, a protein species has many isotopologues, generating a long isotope pattern. We model the generation of the isotope pattern using the averagine model. The figure shows the charge range from 3 to 5. FLASHDeconv assumes the range from 2 to 100 by default. The charged isotopologue ions are eluted at the same retention time range, resulting in coeluted multiple MS features. The set of these MS features is called the total signal from the protein species of mass $\boldsymbol{m}$.

Among mass artifacts, harmonic artifacts, the masses with integer fractions (low harmonic artifacts) or multiples (high harmonic artifacts) of the true masses, are commonly reported in many deconvolution methods. The harmonic artifacts can severely deteriorate the follow-up analyses as they tend to be reproduced through replicates[51].

The main factor limiting mass and charge ranges in deconvolution is the presence of isotopically unresolved peaks in TD-MS spectra[53]. Isotopically unresolved peaks often arise from large (>50 kDa) and highly charged (>50) proteoform ions, as their peaks from distinct isotopologs are not well separated, sometimes even with high-resolution instruments. Such peaks have inaccurate m/z locations and intensities, complicating the deconvolution analysis.

We present FLASHDeconv, an algorithm for high-quality TD deconvolution two orders of magnitude faster than existing tools. FLASHDeconv is capable of analyzing both isotopically resolved and unresolved peaks and allows for wide charge (2–100 by default) and mass ranges (1–100 kDa by default). We also show that FLASHDeconv reports more genuine mass features and substantially fewer mass artifacts than other existing methods.

3.2 Material and methods

3.2.1 FLASHDeconv algorithm

The input to FLASHDeconv is MS data in a HUPO-PSI-compliant mzML file[54]. We used ProteoWizard[55] msconvert tool to convert raw file (Thermo raw file, profile-mode acquired MS data) into mzML file without peak picking or filtration.

The output of FLASHDeconv is the list of (deconvolved) mass features in various formats (i.e., tab-separated values format (tsv), mzML, msalign, etc.). For each mass feature, it reports monoisotopic/average masses, retention time range, apex retention time, summed peak intensity, maximum peak intensity, and the cosine similarity scores for the isotope and charge distributions (see Section 3.2.1 for the scores).

We assume the total signal from a proteoform is generated as illustrated in Fig. 3.1. A single proteoform consists of multiple isotopologues. We assume consecutive isotopologues (i.e., isotopologues whose isotope indices differ by one) are apart from each other by 1.0033 Da (the mass difference between ^{13}C and ^{12}C). Each isotopologue again can have distinct charge states. The range of charge states is 2-100 by default. Finally, the distinctively charged ions from distinct isotopologues are eluted in a specific retention time range along LC.

FLASHDeconv algorithm consists of three sub-algorithms: spectral decharging, deisotoping, and feature finding (Fig. 3.2). The decharging and deisotoping (together called spectral deconvolution) are performed per individual spectrum and output deconvolved masses. Subsequent feature finding then takes the deconvolved masses in all spectra as input and outputs a list of mass features. A mass feature is a virtual trace of an eluted proteoform mass along RT. FLASHDeconv also implements simple yet effective scoring to filter out false positives or harmonic artifacts. Below, we describe each step in more detail.

Decharging and harmonic artifact removal

The decharging step identifies all peaks caused by the same proteoform mass, where the peaks differ only with respect to their charge states. Here, the isotopologues from the same proteoform are treated distinctly. The input to the decharging algorithm is a spectrum (in mzML format) and the output is the list of charge (or equivalently mass) determined peaks.

The major speedup of FLASHDeconv is achieved in this decharging step (Fig. 3.2B). It is based on a simple transformation of mass spectra that transforms peak positions (m/z) to $log\ m/z$ space. The variable (mass-dependent) patterns caused by different charge states of the same proteoform are turned into a mass-independent universal pattern by this transformation. Finding the occurrences of the universal pattern in a transformed mass spectrum can be done very efficiently by a single convolution calculation. Furthermore, with a slight modification, the universal pattern is used to efficiently reduce harmonic artifacts. No intensity-based peak filtration is applied by default since even low-intensity peaks can be important evidence for low-abundance features.

For the input spectrum S, FLASHDeconv computes a transformed spectrum S^* by relocating all peaks in S at their *log-mz transformed* positions. The *log-mz transformed* position of a peak p in S is calculated by subtracting the mass of the charge carrier (e.g., a proton mass) from m/z value of p and then by logarithmizing (natural base) the subtracted value. For example, a peak p of charge q from mass m has m/z value given by

$$(m + qc)/q = m/q + c, \tag{3.1}$$

where c denotes the mass of the charge carrier (a proton). Then its *log-mz transformed* position is given by

$$log(m/q) = log(m) - log(q), \tag{3.2}$$

which is the location of p in the transformed spectrum S^*. Likewise, in the transformed spectrum S^*, a series of peaks $p_1, p_2, ..., p_n$ in S from the same mass m with charges

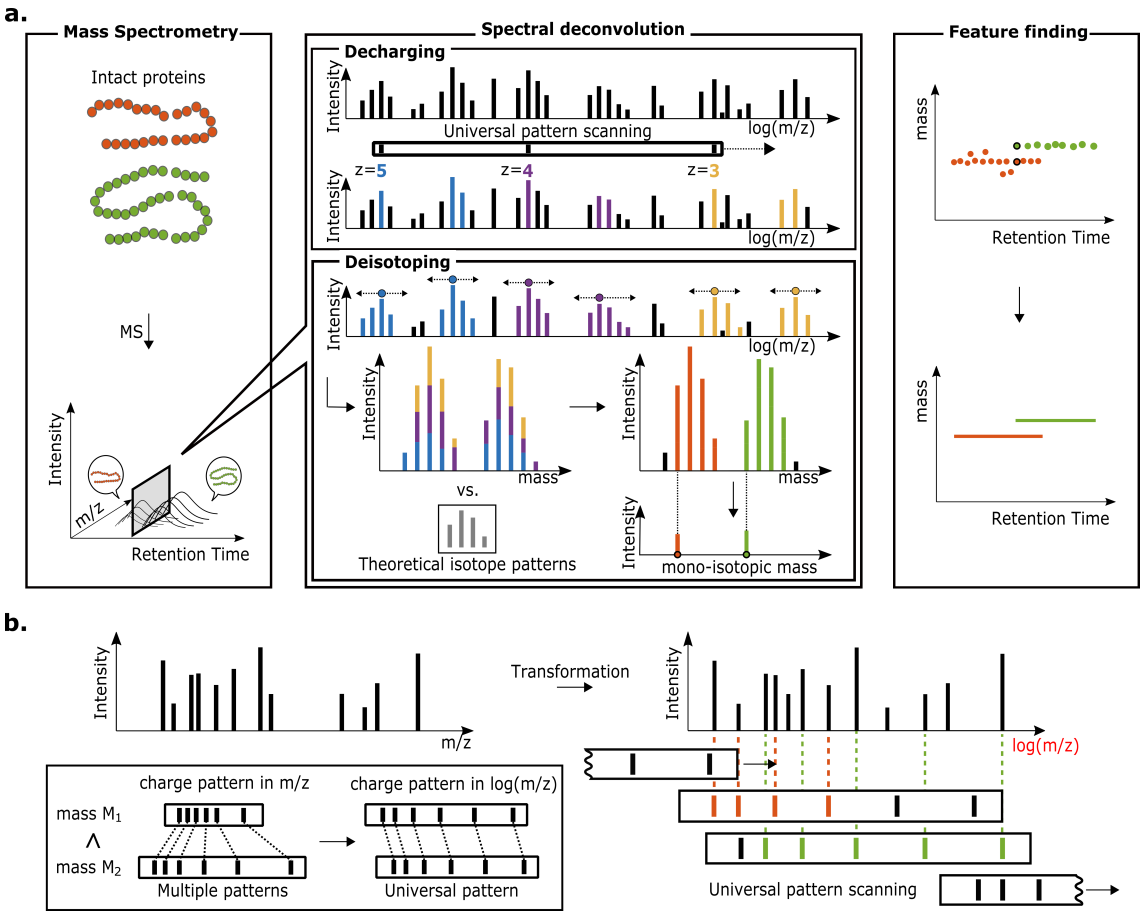


Figure 3.2: FLASHDeconv Algorithm **(A)** FLASHDeconv consists of three separate processing steps: decharging, deisotoping, and feature finding. Decharging and deisotoping are together called spectral deconvolution. Spectral deconvolution finds monoisotopic masses from each input spectrum. Decharging (top panel) is to determine the charges (or equivalent masses) of peaks. The main speed boost of FLASHDeconv is achieved by rapid decharging using universal pattern scanning, as shown in (B). Charge-determined peaks are color-coded (blue, purple, and yellow). Deisotoping (bottom panel) is to determine monoisotopic masses of proteoforms by finding isotope patterns around charge-determined peaks. Around each charge-determined peak, more peaks corresponding to its isotopologs are collected. Then, their intensities are aggregated in the mass space and compared against theoretical isotope patterns. From the matched patterns (red and green color-coded), monoisotopic masses are determined. Feature finding (right panel) then reduces to finding deconvolved masses within a given mass tolerance along the RT direction (done by a mass trace detection algorithm[56]). **(B)** The key idea for rapid decharging: if each peak position is transformed into the log of its neutral m/z value, the charge pattern (the position pattern caused by different charge states of the same proteoform) in the transformed spectrum becomes independent of proteoform masses and hence is universal. Thus, decharging can be done by single sliding this universal pattern along the transformed spectrum to identify matching peaks.

$q_1, q_2, ..., q_n$ are located at their *log-mz transformed* positions given by $log(m)-log(q_1)$, $log(m)-log(q_2)$, ..., $log(m)-log(q_n)$, respectively. One can notice that the distance between any pairs of peaks in this transformed space does not depend on the mass m. We can thus define a universal charge pattern vector

$$U := (-log(q_{min}), -log(q_{min}+1), ..., -log(q_{max})) \tag{3.3}$$

for a given charge range $[q_{min}, q_{max}]$ (2 and 100 by default). Then, the peaks from the mass m should partially match (within tolerance) the universal charge pattern U positioned at $\boldsymbol{log}(m)$.

While the universal pattern U is used to quickly detect peaks from the same mass with distinct charges, we also can define "harmonic charge patterns" to quickly remove low harmonic artifacts, the masses that have integer fractions of true masses. In the algorithmic aspect, low and high harmonic artifacts should be treated separately; while low harmonic artifacts are detected by checking the presence of noisy peaks, high ones are by checking the absence of signal peaks. Thus, low harmonic artifacts are filtered while collecting peaks from a mass, while high ones can be after collecting peaks.

We define the harmonic charge pattern vectors H_r for $r = \frac{1}{2}, \frac{1}{3}, \frac{2}{5}$ by

$$Hr := (-log(q_{min}+r), -log(q_{min}+1+r), ..., -log(q_{max}+r)). \tag{3.4}$$

We take $H_{\frac{1}{2}}$ as an example to see how it is used to remove artifacts having half the true masses. Suppose the mass m is the true mass and its five peaks have charges from 4 to 8. And also suppose we are now examining if a mass $\frac{m}{2}$ is present in the spectrum, without knowing that it is a low harmonic artifact. Note that the universal pattern U is now positioned at $log(\frac{m}{2})$. In the transformed spectrum, the five peaks from the true mass m are located at

$$log(m)-log(4), log(m)-log(5), log(m)-log(6), log(m)-log(7), log(m)-log(8).$$

However, when we test for $\frac{m}{2}$, three locations $log(m)-log(4)$, $log(m)-log(6)$, $log(m)-log(8)$ are considered as the correct peaks having charges 2 to 4, because they can be rewritten by $log(\frac{m}{2})-log(2)$, $log(\frac{m}{2})-log(3)$, $log(\frac{m}{2})-log(4)$. This erroneous decision can be avoided by checking whether peaks are present at the remaining two locations $log(m)-log(5)$ and $log(m)-log(7)$. If they are present, we can consider the mass $\frac{m}{2}$ to be a low mass artifact. The first position $log(m)-log(5)$ is again rewritten by $log(\frac{m}{2})-log(2+\frac{1}{2})$. Likewise, The second position $log(m)-log(7)$ is by $log(\frac{m}{2})-log(3+\frac{1}{2})$. These positions can be detected if we locate $H_{\frac{1}{2}}$ at $log(\frac{m}{2})$,

the same position as U, because $-log(2+\frac{1}{2})$ and $-log(3+\frac{1}{2})$ are elements of $H_{\frac{1}{2}}$. In a similar way, $\frac{m}{3}$ can be detected by using $H_{\frac{1}{3}}$. In the case of $\frac{m}{4}$, $H_{\frac{1}{4}}$, $H_{\frac{2}{4}}$, or $H_{\frac{3}{4}}$ can be used, but we took $H_{\frac{2}{4}}$=$H_{\frac{1}{2}}$, as it is already being used. In the case of $\frac{m}{5}$, we have several options and we took $H_{\frac{2}{5}}$. In this way, we filter out the masses having $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and $\frac{1}{5}$ times the true mass. Further filtrations of (both high and low mass) harmonic artifacts are performed after peak groups are defined below.

FLASHDeconv slides U and H_r along the transformed spectrum S^* to find peaks from the same mass and to avoid harmonic artifacts. We use position relation (detected by U and H_r) as well as intensity relations between peaks. First, we define signal peak pairs and then signal peak triplets. Let $U(x)$ denote the pattern U at the position x in S^* (likewise, $H_r(x)$). Note that $U(x)$ is matched by the peaks from mass $\boldsymbol{exp}(x)$ (within mass tolerance). A pair of peaks p and p' are called comparable if their intensity ratio is between $\frac{1}{4}$ and 4. We call a peak pair p and p' well-aligned if they are comparable and they match any of two consecutive elements of $U(x)$. Then, a well-aligned peak pair p and p' is further called a signal peak pair if no low harmonic peak is present with respect to p and p'. A peak p_h is a low harmonic peak with respect to well-aligned p and p' if

(i) p_h is located between p and p',

(ii) p_h is comparable with p and p', and

(iii) p_h matches to any element of $H_r(x)$.

Finally, three peaks p_1, p_2, and p_3 form a signal peak triplet if both (p_1, p_2), and (p_2, p_3) are signal peak pairs. The presence of a signal peak triplet is strong evidence that they are from the same proteoform mass (in practice, the number of consecutive signal peaks is a user input parameter whose default value is three).

Therefore, while sliding U and H_r along the transformed spectrum, FLASHDeconv stops at the location x wherever a signal peak triplet (p_1, p_2, p_3) is found. Then, it recruits all peaks matched to $U(x)$ including p_1, p_2, and p_3. The set of peaks that matched to $U(x)$ is called a peak group. A peak matched to an element $-log(c)$ in $U(x)$ now has the assigned charge of c.

Even though the method described above efficiently collects peak groups and determines charges to their peaks, still multiple peak groups may represent either low or high harmonic artifacts. Also, quite often a single peak has multiple assigned charges because they are matched to $U(x)$ with different x values. Then, a single peak has membership in multiple peak groups, often those representing harmonic artifacts. To resolve this, we assign an intensity score (described in Section 3.2.1) to each peak

group. Briefly, the intensity score measures the total signal peak intensity subtracted by the total harmonic (both high and low) peak intensity. Then, for each peak, we collect all peak groups containing the peak. Only the peak group with the highest intensity score is allowed to retain the peak. The peak charge is updated accordingly and thus is the peak mass. At this point, the set of peak groups is called the candidate peak group set. Therefore, the candidate peak group set contains all the charge-determined peaks but does not include peaks whose charges are not determined. The next deisotoping algorithm takes this candidate peak group set as well as the charge undetermined peaks in the original spectrum as its inputs.

Before describing the deisotoping algorithm, we define the intensity score of a peak group, the notion used above. To do so, we should measure the intensity sum of low and high harmonic peaks. Low harmonic peaks are already defined above. Intuitively, a high harmonic peak is a peak such that its assigned charge and other peaks' in the peak group do not make consecutive integers while its intensity and others' are similar. To rigorously define high harmonic peaks, first, we sort the peaks in ascending order of their assigned charges. Denote the i-th peak in the sorted peak list as p_i. For each peak p_i, we take the peaks p_{i-1} and p_{i+1}. The peak p_i is called a high harmonic peak if p_i is comparable but is not well-aligned with p_{i-1} and p_{i+1}. For all collected signals, low harmonic, and high harmonic peaks in the peak group, the intensity score of a peak group is defined as its total signal peak intensity subtracted by its total low and high harmonic peak intensities.

Deisotoping

The next step, deisotoping, is performed by finding theoretical isotope patterns (derived from the averagine model[21]) around the charge-determined peaks (Fig. 3.2A). It takes the candidate peak group set and the original spectrum as its inputs and outputs the list of deconvoluted monoisotopic masses.

Briefly, the algorithm runs per peak group in the candidate peak group set. Each peak group contains charge-determined peaks. We examine if charge-undetermined peaks in the original spectrum represent the isotopologues of a charge-determined peak. After collecting all such peaks, they are added to the peak group, forming an extended peak group. An extended peak group consists of peaks from the same proteoform with distinct isotope indices and charges.

The observed isotope distribution is obtained from this extended peak group and is compared against a theoretical isotope distribution. The theoretical distribution is generated from the averagine model[21] and called an average distribution. If the

isotope pattern from the extended peak group is similar enough to the average isotopic pattern, the monoisotopic and average masses are calculated. Otherwise, the group is discarded. A detailed description of the deisotoping algorithm can be found in Section C.2.1.

Feature Finding by Mass Trace Detection

Once all spectra have been deconvolved as described above (and if there are multiple spectra), FLASHDeconv searches for (mass) features along the RT. This idea goes back to the notion of 'mass trace detection' frequently used in quantitative mass spectrometry. We use the term feature here to describe all the signals arising from the same proteoform. Integrating these signals over retention time and summing them up across charge states/isotopes reduces the complexity of the dataset and results in a feature primarily characterized by its retention time, mass, and intensity. Thus, feature detection is based on the notion of a mass trace in the deconvolved spectra.

We use a robust mass trace detection algorithm[56] based on a Gaussian kernel density estimation of the local mass error and as implemented in OpenMS. For mass tolerance, the input tolerance to FLASHDeconv was used. The "re-estimate mass tolerance" function was deactivated.

Scoring and Filtration

To reduce false-positive masses, we filter out extended peak groups as well as mass features based on their per-charge intensity distribution. Here, we explain the scoring method with an extended peak group, but the same procedure is applied to mass features.

For an extended peak group, let the charge intensity for c be the summed intensity over the peaks of charge c in the peak group (regardless of their isotope indices). We say a charge is intense if its charge intensity is higher than 10% of the maximum charge intensity. Firstly, if the maximum number of intense charges that continuously increment by one is less than three, we filter this extended peak group. This simple filtration dramatically reduces the number of high harmonic artifacts.

Then, the charge intensities are fitted to a Gaussian function using Caruana's algorithm[57]. If the fitness is good, we retain the peak group; otherwise, we discard it. Before the fitting, we find the smallest and largest charges such that their intensities exceed 2% of the maximum charge intensity. The charge intensities between the smallest and largest charges are used as the input to Caruana's algorithm. If Caruana's algorithm yields a negative variance, the extended peak group is also disregarded.

Otherwise, the cosine similarity between the charge intensities and the fitted Gaussian is calculated. If the cosine similarity is less than 0.6 (user-specified parameter), the group is discarded.

3.2.2 Dataset description

We benchmarked FLASHDeconv with three less complex samples and a more complex sample. The three low-complexity samples consisted of individual proteins (Cyto, Bovine Cytochrome C and Fil, Filgrastim) and a standard mix of six proteins (PIP, Pierce Intact Protein Standard Mix; see Section C.1 for details). They were acquired on a Quadrupole-Orbitrap at a resolution of 70,000 (Cyto and Fil) and 17,500 (PIP). Because of their low masses (12,223.21 Da for Cytochrome C and 18,786.68 Da for Filgrastim) and high resolution, the spectra in the Cyto and Fil datasets contain isotopically resolved peaks. The PIP dataset was acquired at a lower resolution to assess the deconvolution quality for unresolved isotope patterns. For the complex dataset, we used a murine myoblast sample (unfractionated and generated by Thermo Q-Exactive HF hybrid quadrupole-Orbitrap mass spectrometer at a resolution of 120,000) published previously[36]. This dataset contains both isotopically resolved and unresolved peaks because of the wide mass range.

3.2.3 Availability

The accession number for the Cyto, Fil, and PIP triplicate datasets reported in this paper is MassIVE: MSV000084001 (`https://massive.ucsd.edu`). The datasets are also available under the digital object identifier `https://doi.org/10.25345/C59D26`.

FLASHDeconv is implemented in C++ as a part of OpenMS[38] and available as platform-independent open-source software under a BSD three-clause license at `https://OpenMS.org/FLASHDeconv`.

3.3 Results

For comparison, Xtract[58], ReSpect (Positive Probability Ltd), and Promex[35] were tested against FLASHDeconv for selected datasets: Xtract and Promex for Cyto, Fil, and myoblast, and ReSpect for PIP and myoblast (Xtract and Promex are more optimized for an isotopically resolved signal, whereas ReSpect is optimized for an isotopically unresolved signal). The detailed tool versions and parameters are found in Section C.3.1.

3.3.1 Analysis on the simple datasets

Fig. 3.3 shows the results for the simple protein datasets. In Fig. 3.3A, it is shown that FLASHDeconv is 60–90 times faster than comparable tools. Per spectrum, the deconvolution took less than 10 ms (excluding spectrum loading time). Fig. 3.3B shows the deconvolved spectra (deconvolved masses and their intensities, merged for all of the RT range) for the Cyto and Fil datasets. All three tools found a signal around the input protein masses, showing similar mass accuracies (∼3 Da error). However, Promex reported additional low harmonic artifacts for both datasets (see also Fig. C.1 and C.2 for the same analysis on the technical replicates).

In the isotopically unresolved PIP dataset, both FLASHDeconv and ReSpect found all six input masses within a 3 Da error (ranging from 9 to 70 kDa; see Fig. 3.3C, C.1, C.2, and C.4). FLASHDeconv and ReSpect showed better mass accuracy for small and large masses, respectively. The features from each tool were also mapped to the RT-mass plane in Fig. 3.3D (feature map). The feature map from ReSpect shows that ReSpect reports thousands of features in addition to those directly relatable to the input masses. However, when we examined the mass and RT relations of the features, we found that at least 50% of them were classified as mass artifacts or isotopolog artifacts (the masses with incorrectly determined isotope indices). For mass artifacts, harmonic and charge-off-by-one artifacts[51] were considered. A charge-off-by-one artifact appears when an off-by-one true charge is assigned to a peak (see Section C.3.2).

The results on the low-complexity samples demonstrate that FLASHDeconv accurately identifies the proteoform masses with high specificity and sensitivity for both isotopically resolved and unresolved datasets.

3.3.2 Analysis on the complex dataset

The advantage of FLASHDeconv is even more pronounced on the complex dataset. Fig. 3.4A shows that FLASHDeconv was significantly faster (up to 190 times) than the other tools; it took only 7 min to process the whole dataset (approximately 100 ms/spectrum).

FLASHDeconv also found the largest number of features among all tools (∼27,000 features; see Fig. 3.4B). To assess whether mass or isotopolog artifacts contribute to this large feature count, we again examined the masses and RT relations among the features as above (see Section C.3.2). For FLASHDeconv, only 13% of the detected features actually turned out to be artifacts. For other tools, the percentages were between 60% and 80%. These results collectively demonstrate the high sensitivity and specificity of FLASHDeconv.

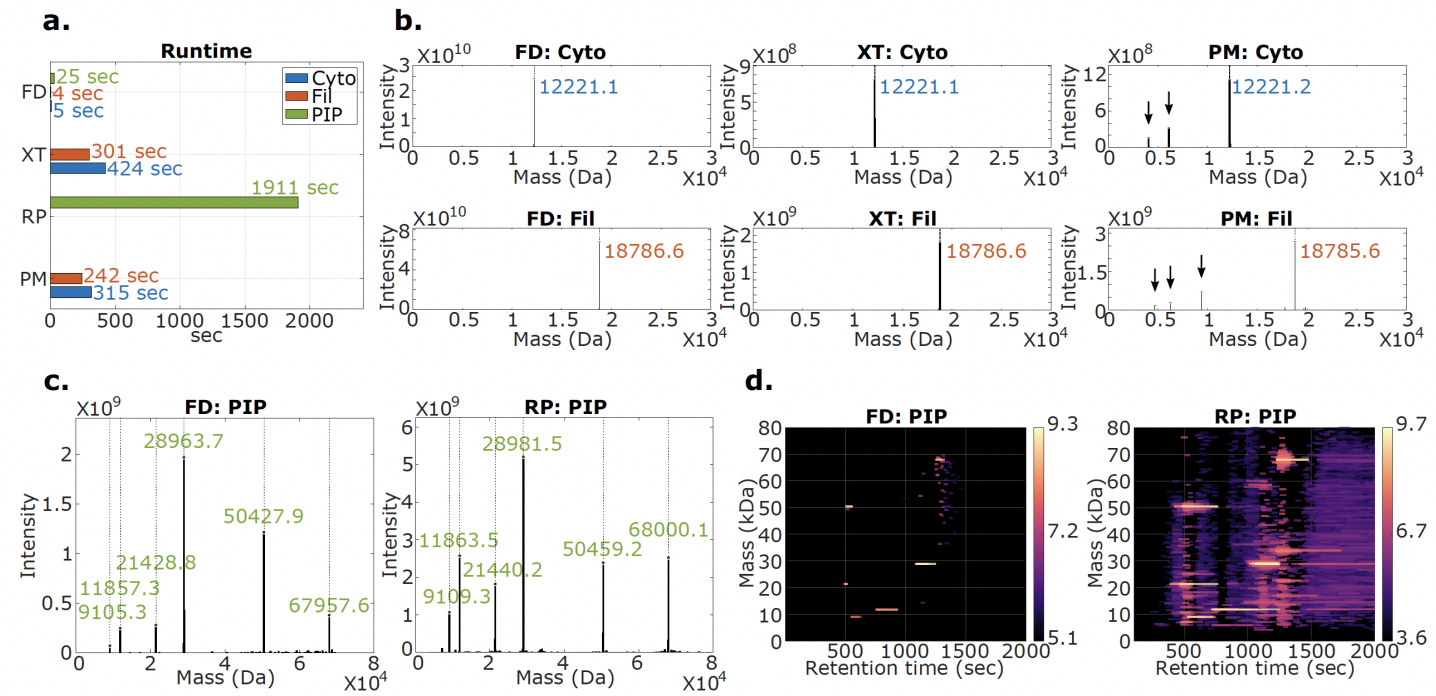


Figure 3.3: Analysis of simple datasets. **a.** Runtime comparison for the simple datasets (Cyto, Fil, and PIP). FD stands for FLASHDeconv; XT for Xtract; RP for ReSpect; and PM for Promex. FD was up to 90 times faster than the other methods. **b.** Deconvolved spectra generated by FD, XT, and PM for isotopically resolved pure protein datasets(Cyto - upper panel; Fil - lower panel). The dotted black lines indicate the expected exact masses (12,223.21 Da and 18,786.68 Da), and the colored digits specify the reported masses within 3 Da from exact input monoisotopic masses. For intensity estimates, The values in the SumIntensity column were used for FD, Sum_Intensity for XT, and Abundance for PM. All three methods found the expected masses, but PM reported additional high-intensity low-harmonic mass artifacts (e.g., $\frac{1}{2}$ and $\frac{1}{3}$ input masses; marked with arrows). Fig. C.1 and C.2 show that these harmonic artifacts are reproduced through replicates. **c.** Deconvolved spectra from FD and RP for the PIP dataset (isotopically unresolved). The colored digits specify the reported masses within 3 Da from input monoisotopic (for FD) or average (for RP) masses. For intensity estimation, Sum_Intensity was used for RP. FD and RP found all six input protein masses. FD and RP showed better mass accuracy for small and large masses, respectively. The expected *exact* monoisotopic masses are 9,105.3, 11,858.0, 21,429.8, 28,963.7, 50,429.8, and 67,959.4 Da. The expected *average* masses are 9,111.5, 11,865.5, 21,442.6, 28,981.3, 50,459.7, and 68,001.2 Da. **d.** Identified features mapped on the RT-mass plane (feature maps) from FD and RP for the PIP dataset. Except for the expected masses, RP also reported thousands of masses, out of which 50% were identified as mass or isotopologue artifacts (see Section C.3.2 for artifact detection). See also Fig. C.1 and C.2 for more results from the replicates.

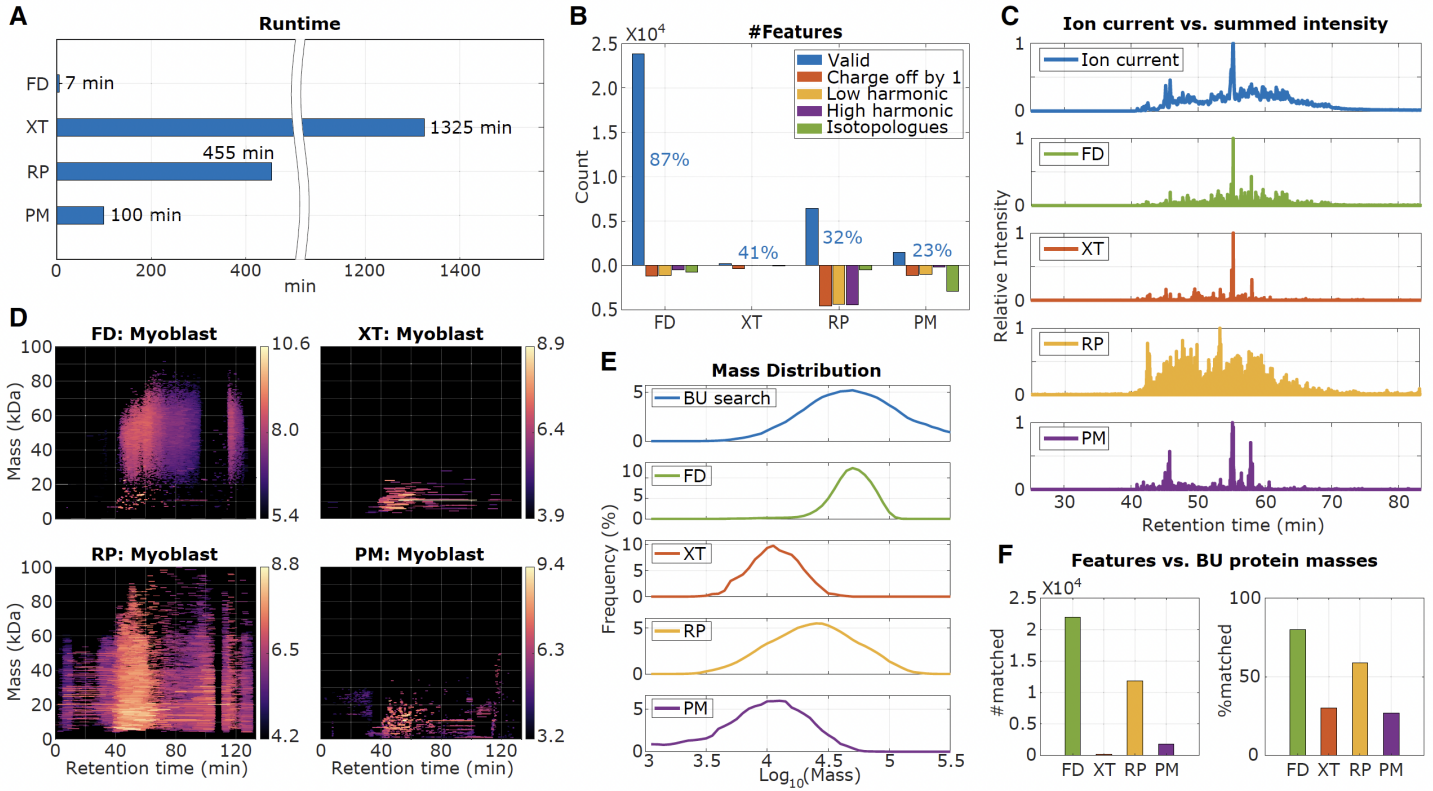


Figure 3.4: Analysis of Complex Murine Myoblast Dataset. **(A)** Runtime comparison for the murine myoblast dataset. FLASHDeconv (FD) was drastically faster (14–190 times) than the other methods. **(B)** The numbers of valid features (i.e., not detected as artifacts) and detected mass and isotopolog artifacts from each tool (see Section C.3.2). The blue digits show the portion of the valid features. FD found almost 27,000 features, and only 13% of them were artifact features. The other tools generated a significantly higher portion (60–80%). **(C)** For each tool, feature intensities were summed across all masses and plotted along the RT direction (see Section C.3.3 for the details). For reference, the TIC chromatogram was also drawn on top. FD and the TIC showed an almost perfect fit, demonstrating that FD features cover most of the eluted ions. **(D)** The feature maps for each tool. For the 0- to 20- kDa region, all tools produced intense features (see Figure S7 for rounded monoisotopic mass overlaps in the low mass region). Xtract (XT) and Promex (PM) reported only a few features for the 20- to 100- kDa region because most of the corresponding peaks would be isotopically unresolved. **(E)** The mass distribution of each tool as compared with the protein mass distribution obtained from a BU search (of the same myoblast sample). The distribution from FD is similar to the distribution from BU-identified protein mass. **(F)** The number and portion of the feature masses matched to the BU-identified protein masses for each tool (one modification per protein is allowed; see Tables C.1 and C.2). FD exhibits a much higher consistency with BU results than other tools.

3.3.3 Evaluation of the proteoforms from the complex dataset

To see this from a different perspective, we compared the total ion current (TIC) chromatogram (from the raw spectrum file) with the intensities summed over masses (from the reported features) in Fig. 3.4C (see Section C.3.3). FLASHDeconv showed the best fit to the TIC among all methods. Good agreement between the TIC and the summed-up feature signal basically indicates that a large portion of the recorded signal could be assigned to the detected features. Poor agreement can occur when parts of the signal are not detected (causing a sensitivity drop) and/or false-positive features are erroneously included (causing a specificity drop).

The feature maps in Fig. 3.4D and the feature mass distributions in Fig. 3.4E together show that FLASHDeconv features are centered between 40 and 60 kDa, whereas the others between 0 and 20 kDa or 0 and 40 kDa (see Fig. C.4 for the rounded monoisotopic mass overlap between tools in the low mass region). Not surprisingly, Xtract and Promex exclusively reported masses lower than 20 kDa, as most masses larger than 20 kDa would be represented by isotopically unresolved peaks.

Next, we compared our feature masses with the protein masses identified by BU searches for the same sample (also reported by Schaffer et al.[36]). The protein masses identified by the BU approach can provide orthogonal evidence of the proteoforms' presence. The distribution of these BU-identified protein masses (Fig. 3.4E, top panel) was similar to the FLASHDeconv feature mass distribution, suggesting that the high masses reported by FLASHDeconv are indeed present in the sample. For more rigorous evidence, we matched our feature masses against the BU-identified protein masses (Fig. 3.4F; see Section C.3.4). As BU-identified protein masses only represent unmodified proteins, we allowed a single modification per protein for matching. About 22,000 (80%) feature masses were matched, confirming that most of our feature masses are likely to be genuine proteoform masses. The other tools showed substantially lower agreement than FLASHDeconv (see Tables C.1 and C.2).

In Fig. 3.5, it is demonstrated that FLASHDeconv has comparable reproducibility with other tools, even if it yields substantially fewer mass artifacts (Fig. 3.4B) that inflate reproducibility than others.

3.4 Discussion

We presented FLASHDeconv that can carry out high-quality deconvolution for diverse types of (simple and complex and isotopically resolved and unresolved) TD-MS datasets at an unprecedented speed. With its high speed and versatility, FLASHDeconv

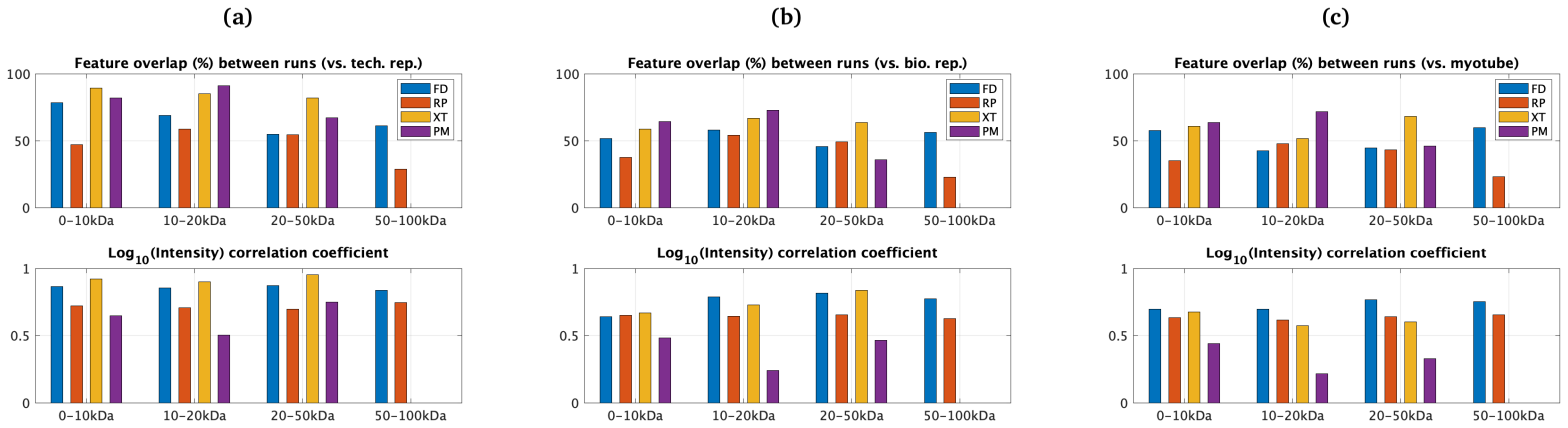


Figure 3.5: Reproducibility analysis for complex samples, Related to Fig. 3.4. **(a)** Myoblast vs. technical replicate. **(b)** vs. biological replicate. **(c)** vs. myotube.
To evaluate the reproducibility, we analyzed the technical and biological replicates of the myoblast dataset by all tools. A myotube dataset was analyzed as well (all datasets are published in Schaffer et al.[36]). The reproducibility was measured by the portion of features in the myoblast dataset that overlap with those in other datasets. We also measured the Pearson's correlation coefficient between the log of intensities of overlapping features. Low-intensity correlation is an indication of coincidental feature overlaps or of overlaps between mass artifacts (as the artifacts have inaccurate intensities as compared to true masses). Two features in distinct datasets are declared to be overlapping if they are within 5 minutes in retention time and within 10 ppm mass tolerance, allowing up to three isotope index errors. The overlapping portion and correlation coefficients are drawn for different mass ranges (0-10, 10-20, 20-50, and 50-100 kDa).
For all tools, the technical replicate resulted in the best reproducibility both in terms of overlapping portion and log intensity correlation. For low-mass regions (<20 kDa) containing isotopically resolved peaks, Xtract (XT) and Promex (PM) showed high overlap portions. FLASHDeconv (FD) showed higher overlap than ReSpect (RP) in most cases. In terms of log intensity correlation, Xtract and FLASHDeconv showed high correlations (Promex showed low-intensity correlations for all cases). For the middle mass range (20-50 kDa) where both isotopically resolved and unresolved peaks exist, Xtract showed the highest overlap portions for all datasets. FLASHDeconv and Xtract showed comparable intensity correlations. But fair comparison in this mass range is not easy as Xtract and Promex yield only features from isotopically resolved peaks while FLASHDeconv and ReSpect report those from both resolved and unresolved peaks. For the higher mass region (>50 kDa), Xtract and Promex did not report any feature. FLASHDeconv showed better reproducibility than ReSpect for this region. Overall, FLASHDeconv showed comparable reproducibility even if dense mass artifacts in other tools (shown in Fig. 3.4b) are supposed to inflate reproducibility.

could significantly increase the potential of various TD analyses including native-MS and TD-BU integrated studies[46,59]. The very fast processing times for spectral deconvolution of FLASHDeconv could be used for real-time mass deconvolution on MS instruments, enabling smarter and more efficient MS/MS fragmentation methods for TD-MS than are currently being used.

We demonstrated that FLASHDeconv is highly resistant to mass artifacts and capable of processing both isotopically resolved and unresolved peaks. As such, FLASHDeconv would boost the quality of overall TD-MS analysis including proteoform identification and characterization. Also, FLASHDeconv has a relatively small number of parameters compared with other algorithms. Most of the parameters are directly determined by experimental or instrumental setups and do not require a deep understanding of the algorithm itself. Thus, tuning of parameters for optimal results can be readily achieved, and the results are expected to be robust.

The current version of FLASHDeconv only performs the deconvolution of precursor ion (i.e., MS1) spectra. The key idea of log m/z transformation for quick decharging can be applied to product ion (i.e., MS/MS) spectra. However, the scoring function and other details should be modified to incorporate signal characteristics of MS/MS spectra[60].

Further development would also include exact quantification of mass features. Even if FLASHDeconv reports summed peak intensity for each mass feature, some signal peaks, in particular the low-intensity ones, may not be counted for the intensity calculation. In addition, the RT span and feature shape should be taken into account.

Finally, the accuracy of monoisotopic mass estimation could be enhanced, particularly for large molecules. An effective signal processing method for isotopically unresolved peaks could be developed such as improved baseline elimination and peak picking. We also could use exact isotope patterns of proteins (in the place of averagine-derived ones) to raise mass accuracy when protein chemical compositions are available, as in targeted studies[61,62].

Chapter 4

Quantification algorithm for proteoform analysis

Adapted with permission from:

FLASHQuant: a fast algorithm for proteoform quantification in top-down proteomics
Jihyung Kim+, Kyowon Jeong+, Philipp T. Kaulich, Konrad Winkels, Andreas Tholey, Oliver Kohlbacher
Analytical Chemistry (2024)[63]
+ These authors contributed equally

4.1 Introduction

In the past decade, research in TDP has seen substantial improvements in protein separation, MS techniques, and bioinformatic software and demonstrated its potential to elucidate the important role of proteoforms in biomedical processes[10,13,64,65]. Technical advances have enabled researchers to move forward from qualitative analysis to quantitative analysis on proteoform studies[7,66,67].

Quantitative analysis is crucial in proteomics, as it opens the door for comparative studies of proteins in different biological functions or for biomarker discovery[7,68,69]. In principle, three general strategies can be employed to quantify proteoform abundances: LFQ, metabolic labeling, and chemical labeling. Metabolic labeling has seldom been employed in TDP, mainly due to inherent challenges such as interfering isotope patterns between light/heavy labeled proteins[24,70,71]. In contrast, chemical labeling-based

strategies such as isobaric labeling have been successfully applied, but further improvements are still needed regarding the labeling procedures as well as the bioinformatics data interpretation[72,73].

Therefore, the most widely applied approach in TDP is LFQ[33,66,67], which performs a relative quantification by direct comparison between MS runs. It has advantages over the labeling method, such as low costs (no need for expensive labeling reagents), fewer experimental steps, and omission of increased sample complexity (i.e., introduced in chemical labeling approaches due to incomplete or over-labeling).

Nevertheless, the analysis of TDP-LFQ data still is a major bottleneck. Existing data analysis tools for MS1 level quantification (including chromatographic feature detection and intensity calculation) still need further developments and improvements regarding reproducibility, overlapping signal resolution, or usability.

LFQ of proteoforms in most cases still relies on deconvolution-centric software, such as Protein Deconvolution (Thermo Fisher Scientific)[36,74], DataAnalysis (Bruker Daltonics)[75–77], TopPIC suite[60,78,79], and ProMex[35,80]. As the primary goal of deconvolution tools is to detect proteoforms' masses rather than quantify them, chromatographic information needed for quantification is widely neglected on these platforms. This results in limited peak coverage and compromises reproducibility in quantification. Therefore, many LFQ studies have been performed using in-house scripts to retrieve chromatographic data based on deconvoluted proteoform masses.

Despite its significance, the problem of overlapping signals in LFQ studies has often been overlooked[12,22]. Overlapping signals occur when the distinct proteoforms interfere with each other (i.e., sharing m/z values), primarily due to the complexity of the samples and the co-elution of multiple proteoforms from the LC within a narrow retention window. TDP, which deals with larger analytes compared to bottom-up proteomics or metabolomics, exacerbates this issue; the larger size of analytes (proteoforms) leads to broader isotopic envelopes and wider charge state ranges compared to those of peptides, contributing to the dataset's high complexity. As a result, co-eluted proteoforms of different charge states can occupy similar or overlapping m/z ranges even when they have highly distinct masses[12,81], introducing possible quantification bias, particularly in the above-mentioned deconvolution-centric approaches. Thus, the overlapping signal issue should be addressed for accurate LFQ in TDP, which necessitates advanced data analysis strategies.

Several TDP-LFQ data analysis algorithms have been developed but are not freely available to users, e.g., QMT[33] and IPQuant[30], and commercial platforms such as ProSightPD (Proteinaceous, Inc., Evanston, USA).

Here, we introduce FLASHQuant, a fast and robust tool for MS1-level LFQ analysis in TDP. FLASHQuant incorporates an individual chromatogram extraction procedure and a conflicting feature (i.e., overlapping proteoform signals) resolution method using a non-negative least squares solver. In our evaluation, FLASHQuant demonstrated its effectiveness compared to the widely used ProSightPD, TopFD[81] (from the TopPIC Suite), and FLASHDeconv[44]. Also, since quantitative analysis inevitably accompanies multiple MS runs, we implemented ConsensusFeatureGroupDetector to align quantified proteoforms across multiple MS runs. As part of OpenMS[38], both FLASHQuant and ConsensusFeatureGroupDetector are publicly available as platform-independent open-source software with a graphical user interface at `https://OpenMS.org/FLASHQuant`.

4.2 Materials and methods

4.2.1 FLASHQuant algorithm

FLASHQuant comprises four main stages: m/z trace extraction, feature group assembly, conflict resolution, and quantity calculation (Fig. 4.1). An m/z trace is equivalent to the mass trace described in Section 2.2.1, representing an individual ion chromatogram with a specific charge state and isotope. Feature group refers to the group of m/z traces of different charge states from a single putative proteoform and its isotopes. For accurate quantification, the detected feature groups are refined while the quantities are calculated.

m/z trace extraction

The feature group detection stage starts with extracting m/z traces from the LC-MS spectra using MassTraceExtractor[56]. MassTraceExtractor is a self-consistent kernel density estimator that connects peaks having m/z values within the tolerance (defined by the instrument's mass accuracy) along the retention time. Thus, each extracted m/z trace represents a chromatographic signal having a specific charge state and isotopic composition over an elution window.

MassTraceExtractor is performed using two algorithms in OpenMS: MassTraceDetection and ElutionPeakDetection. MassTraceDetection takes centroid LC-MS spectra as input and connects the peaks within m/z tolerance that continuously appear across the retention time. FLASHQaunt takes five ppm for this m/z tolerance as a default value to only gather peaks from the same analyte (of distinct charge state and isotope) into a single m/z trace. Since distances between isotopes from an intact protein with a large charge state are relatively smaller than those from peptide or metabolite, a small

m/z deviation should be employed to distinguish those isotopes. After MassTraceDetection effectively separates and groups spectral peaks from distinct compounds with different m/z values, ElutionPeakDetection further separates them along the retention time direction by analyzing the shape of each m/z trace. For each m/z trace, its abundance is determined by its total area, and its retention time range is truncated with its full width at half maximum. This truncation is done since the elution profile often holds long tails with low intensities. Note that different m/z traces may represent the same (proteoform) molecule but with different isotopes or charges.

Feature group assembly

The extracted m/z traces are deconvolved and assembled into feature groups using a fast and robust spectral deconvolution algorithm FLASHDeconv[44]. Both charge and isotope deconvolution are applied to m/z traces to determine the masses. In this step,

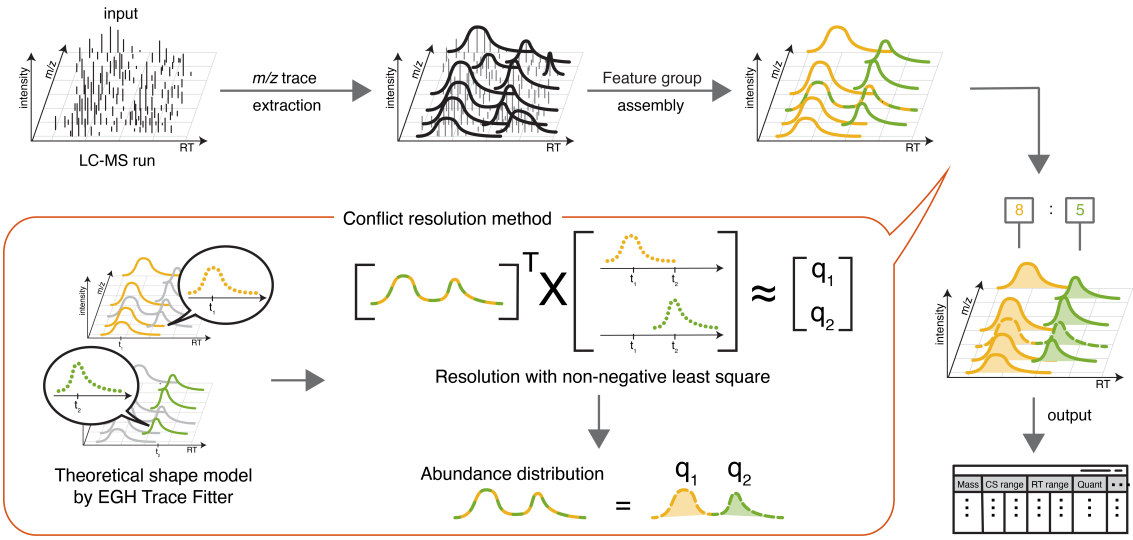


Figure 4.1: Illustration of FLASHQuant algorithm. From a centroided LC-MS1 full scan, FLASHQuant extracts m/z traces, denoted as black curved lines at the top-center. Detected m/z traces are then assembled into feature groups, representing putative proteoforms. Feature groups (depicted as yellow and green) often share m/z traces (the mixed-colored line), which should be resolved for accurate quantification. The theoretical shapes of each feature group are estimated using its non-shared m/z traces by fitting them against the EGH function. Based on the estimated shapes (having retention time and intensity values), the proportion of each shape (q_1 and q_2) in the shared trace is estimated with a non-negative least square solver. Then, the shared m/z trace quantity is distributed between the corresponding feature groups. The feature group abundances are finally calculated as the summed area of m/z traces in each feature group. The output of FLASHQuant is given as a tsv format file by default.

an m/z trace may be assigned to multiple feature groups representing the overlapping signal (the shared m/z trace) from distinct proteoform ions (the feature groups).

During its deconvolution to find monoisotopic masses, FLASHDeconv automatically groups the spectral peaks from different masses. Since the core algorithm of FLASHDeconv takes a spectrum as input, we first convert the m/z traces into a series of spectra along the retention time direction. To convert m/z traces along retention time into spectra, we first bin the retention time with a fixed bin size of the minimum retention time span of the m/z traces. Then, for each bin, all m/z traces whose retention time range overlaps with the bin are first collected. Then, the collected traces are projected into a single spectrum, where peak intensities are determined by the abundances of the projecting traces. If different bins share exactly the same set of m/z traces, only one bin out of them was processed to avoid redundancy.

Each generated spectrum is then deconvolved using the FLASHDeconv algorithm so that m/z traces are collected into feature groups. Each feature group has its (monoisotopic) mass, abundance (the summation of its member m/z traces), and cosine score (a fit score to the theoretical isotope patterns) determined by FLASHDeconv. The retention time range of a feature group is given by the union of the ranges of its member m/z traces. Although exactly the same spectra are not generated from the above conversion, still many (in particular, the ones from consecutive bins) would yield the same or quite similar feature groups that are likely to be from the same proteoform. Next, we attempt to detect such feature groups and merge them into a single one. To this end, first, the feature groups are sorted in descending order of their abundances. Then, we iterate over the sorted feature groups (denoted by F) and perform the following merging process (i) - (iv):

(i) For each iteration for the feature group f, all the feature groups whose masses are within 10 Da from the mass of f and whose abundances are less than the abundance of f are selected.

(ii) Then only the ones whose retention time ranges overlap with that of f by more than 50% are retained. The remaining feature groups are denoted by $\bar{F}$.

(iii) From $\bar{F}$, we further discard the feature groups f' such that the mass difference between f' and f is larger than 3 Da and no more than 50% of the m/z traces in f' overlap with those in f. Note that $\bar{F}$ contains f itself.

(iv) Generate a merged feature group $\bar{f}$ by taking all m/z traces in feature groups within $\bar{F}$. FLASHDeconv is again employed to calculate the mass, abundance, and cosine score for $\bar{f}$. Only if the cosine score of $\bar{f}$ is higher than f, f is replaced

by $\bar{f}$ in the original feature group set F. Otherwise, f is unchanged (and thus, the above steps have no effect).

After this merging process is done for all features in F, the conflict resolution is performed as described below.

Conflict resolution method

This stage aims to separate shared m/z traces of different feature groups in the feature group set F. The shared m/z traces are mainly from overlapping signals but also from errors in feature group detection when an m/z trace is assigned to an incorrect feature group. While different in origins, both cases can be addressed by comparing the m/z trace elution profiles within the feature groups containing the shared m/z trace. Therefore, this stage only concerns feature groups in conflict with others. Moreover, since sharing of the m/z traces between different feature groups is highly localized, we divided the m/z traces in each feature group into subgroups (called features) with the same charge states and used the features as the basic unit for the conflict resolution. Since each feature consists of the m/z traces from the same (putative) mass of the same charge, they are highly localized in the m/z dimension.

For fast conflict resolution, it is important to quickly detect conflicting features (i.e., features sharing conflicting m/z trace) and find clusters of conflicting features. Veit[82] suggested the idea of these conflicts being modeled as an undirected graph $G = (V, E)$, where nodes v in V represent features and two nodes v and v' are connected by an edge when their representing features have a conflict. By finding the connected components of this graph, one can quickly find the clusters of conflicting features.

For each cluster, the conflict resolution method is applied for each edge of the cluster, similar to the scheme already established and proved in BUP[83]. The key idea of this scheme is to reconstruct shared m/z traces using the unshared m/z traces. In theory, all the m/z traces' elution profiles in a feature (and further those in a feature group) should be the same, and the traces shared by different features should have the shape of a weighted summation of the unshared traces from the conflicting features. To get the theoretical elution profile of a feature (from hereon, theoretical shape), first, the unshared traces are collected from each feature. Then, the theoretical shape is modeled based on an exponential-Gaussian hybrid function[84] using the EGHTraceFitter algorithm in OpenMS. If no unshared m/z trace exists for a feature, we track the feature group that contains the feature, find the most abundant m/z trace in that feature group, and use it to model the theoretical shape.

Upon finding the theoretical shapes from the conflicting features, the conflict resolution can be formulated by a distribution problem in which we want to distribute the quantity of the shared m/z trace to the features. This distribution problem can be described as a non-negative least square problem.

Let M be a matrix composed of theoretical shapes A,

$$M = \left[A_0, A_1, ..., A_n\right] \tag{4.1}$$

where n is the number of theoretical shapes, thus the number of corresponding features. Y denotes the shape of the shared m/z trace, and the lengths of each As are adjusted to match the length of Y by zero padding or truncation. We compute a vector Q of length n, $q_n \in Q$, that solves:

$$M \times \left[q_0, q_1, ..., q_n\right]^\top = Y, \text{ subject to } q \geq 0 \tag{4.2}$$

q refers to the proportions of features to describe the quantity of shared m/z trace. Each q is multiplied by the quantity of the m/z trace and distributed over the corresponding features.

Quantity calculation

After all shared m/z traces are resolved, the abundance of each feature group is calculated by summing areas under all corresponding m/z trace curves. Along with the quantity, FLASHQuant outputs information on feature groups, including their monoisotopic masses, retention time range, charge range, and cosine score.

4.2.2 ConsensusFeatureGroupDetection

ConsensusFeatureGroupDetection allows users to easily attain jointly detected feature groups among multiple LC-MS runs (i.e., technical replicates) within mass and retention time tolerance.

Starting from the most abundant feature group, ConsensusFeatureGroupDetector searches for feature groups within tolerances (mass and retention time) in all given input files (tab-separated vector (TSV) format) until exhaustion. When specific column names from input files are provided as parameters for mass, apex retention time, and quantity, this tool can be used on result files from quantification software other than FLASHQuant (e.g., TopFD).

4.2.3 FLASHQuantWizard

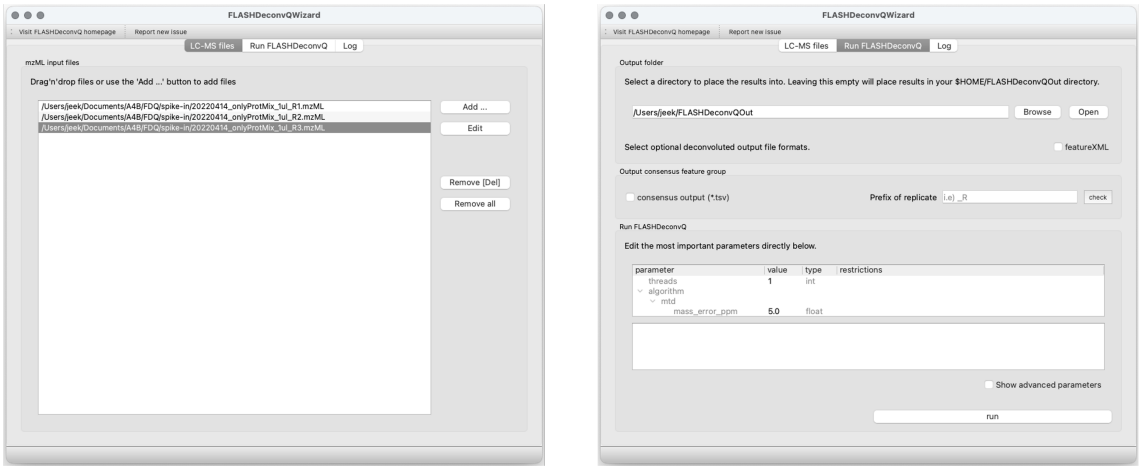


Figure 4.2: Screenshots of FLASHQuantWizard. Instead of a command line tool, FLASHQuant can be run with GUI. FLASHQuant will be executed per input file, and then consensus feature group detection among replicates can be optionally executed as well.

We provide a GUI, FLASHQuantWizard, for users to conveniently run FLASHQuant (Fig. 4.2). The implementation of FLASHDeconvWizard is heavily inspired by the SwathWizard in OpenMS. SwathWizard is an effective pipeline GUI tool that assists OpenSWATH[85] proteomics data analysis. Some of the SwathWizard's useful segments are shared with FLASHQuantWizard, such as LC-MS file loader and displaying logs.

The main segment of FLASHQuantWizard contains three widgets: Output selection, Output consensus feature group, and Run FLASHQuant. In the first widget, the Output selection widget, users can specify an output directory or ask for additional output files in featureXML format. Also, the next widget, the Output consensus feature group widget, offers additional consensus feature group output files (from here on, consensus files) among replicates upon request. When the input files contain replicates generated under multiple experimental conditions, each requiring separate consensus files, FLASHQuantWizard can distinguish these replicates. For this, the shared substring between replicates' file names should be given by users. Otherwise, all input files are used to output a consensus file. The last widget, Run FLASHQuant, allows for parameter adjustment of FLASHQuant.

4.2.4 Dataset description and runtime

We evaluated the performance of FLASHQuant using three datasets with different complexity: **PIPMix**, **SpikeIn**, and **ProteomeMix** (Fig. 4.3). PIPMix is the least complex

dataset consisting of single LC-MS runs with a six-protein mixture (Pierce Intact Protein Standard Mix from Thermo Fisher). To generate the SpikeIn dataset, different amounts of the PIPMix (relatively diluted 1, 1/2, 1/3, 1/5, 1/7, and 1/10 with 20 ng/µl as 1) were spiked into an *E. coli* lysate, each of which was subject to an LC-MS run. The ProteomeMix dataset contains varying proportions of proteins from a human Caco-2 cell lysate (1/5, 1/2, 1, 2, and 5) with a constant concentration for *E. coli* lysate.

Prior to MS analysis, proteoforms were separated with a 90-minute linear gradient on an Ultimate 3000 nano-UHPLC system equipped with a reversed-phase C4 column coupled online to a Fusion Lumos Tribrid mass spectrometer. MS1 spectra were acquired at high resolution (120,000), and within a cycle time of 4 s, the most intense ions were fragmented by collision-induced dissociation (see Section D.1 for the details on data generation).

4.2.5 Availability

The mass spectrometry proteomics data have been deposited at the ProteomeXchange Consortium via the MassIVE repository. They are publicly available under the accession number MSV000091923 and DOI `https://doi.org/doi:10.25345/C5XW4861V`. FLASHDeconv is implemented in C++ as a part of OpenMS. FLASHQuant original code has been deposited and is publicly available at `https://github.com/JeeH-K/OpenMS/tree/feature/FLASHQuant`. Also, `https://OpenMS.org/FLASHQuant` website contains binary installers for all platforms and a link to the GitHub repository with the latest version source code.

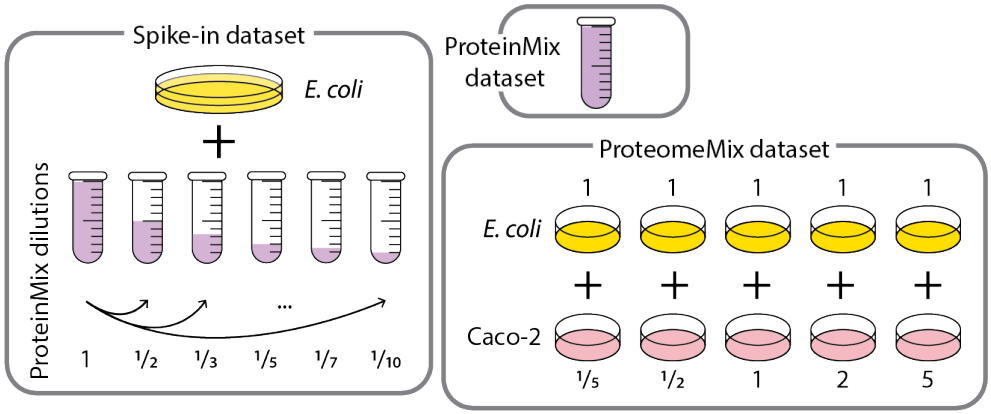


Figure 4.3: Illustration of the datasets used for the analyses.

4.3 Results

4.3.1 Runtime comparison

We benchmarked FLASHQuant against ProSightPD ("ProSightPD Hi Res. Feature Detector" node, version 4.2 in Proteome Discoverer 3.0, Thermo Fisher Scientific), TopFD (TopPIC Suite version 1.7.1), and FLASHDeconv in all comparisons (see Section D.2.1 for the details on parameters). The runtimes of four tools were measured on a desktop PC with an Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz CPU and 64 GB RAM. FLASHQuant had a significant advantage in terms of runtimes compared to other tools, taking only 26 minutes for the PIPMix and SpikeIn datasets combined and 14 minutes for the ProteomeMix dataset (Fig. 4.4); 1.5 times faster than FLASHDeconv, > 4 times faster than TopFD, and > 17 times faster than ProSightPD. FLASHDeconv showed comparable runtimes for all tests, but FLASHQuant showed slightly faster speeds for all tests.

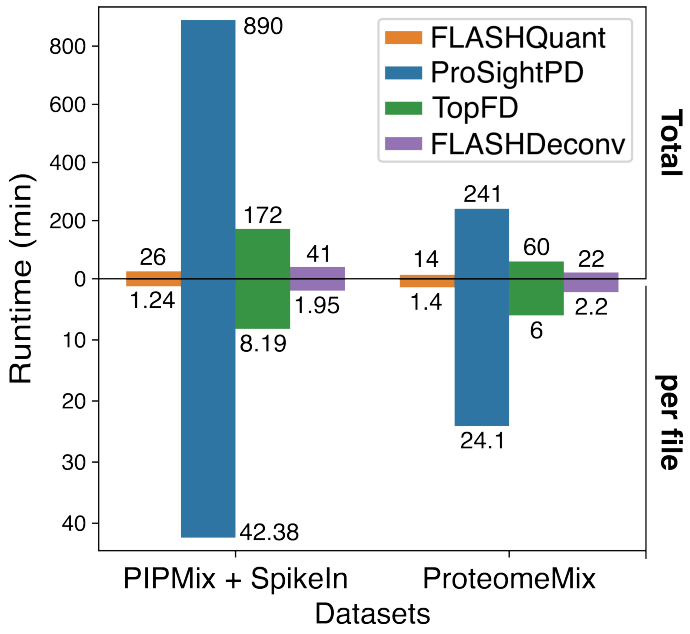


Figure 4.4: Runtime comparison among four tools. FLASHQuant showed the shortest run time out of all benchmarking tools. The PIPMix and SpikeIn datasets were executed together as they share the same parameters. They consist of 21 files, while the ProteomeMix dataset contains 10 files. Runtime per file varies between the datasets due to the highly variable parameters (charge and mass) per individual dataset in the PIPMix and SpikeIn datasets. FLASHQuant showed significantly shorter (4-34 times) run times than TopFD and ProSightPD per file.

4.3.2 Quantification sensitivity evaluation with SpikeIn dataset

As the first validity check on proteoform detection and quantification sensitivity, we compared the performance among the four tools with the SpikeIn dataset. Fig. 4.5 shows the relative fold changes of four spiked-in proteins, as separately depicted in four columns, from all dilutions and technical replicates in the SpikeIn dataset (see Fig. D.2 for the total number of results). Due to the chosen MS setting, only four spiked-in proteins of mass below 28 kDa were detected out of six proteins. Each quantity of proteins was normalized by the respective TIC value of the LC-MS run. Then, the fold changes were computed based on the average quantity of spiked-in protein at 20 ng/µl concentration, denoted as "1" in Fig. 4.5.

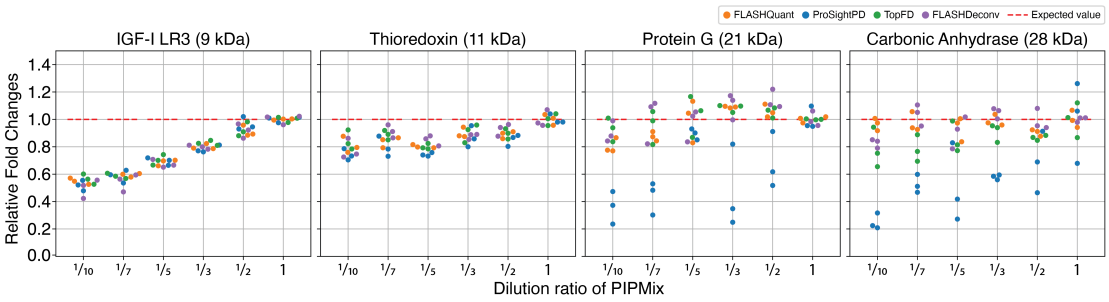


Figure 4.5: Relative fold changes comparison for the SpikeIn dataset. The four spiked-in proteins are shown in four columns, with the expected relative fold change values as red dashed lines. The x-axis in each column displays three pairs of dots per each tested tool, representing relative fold change values from three replicates. FLASHQuant showed the least biased quantification across all samples, demonstrating its high quantification accuracy and reproducibility. The large deviations from the expected value in IGF-I LR3 protein were due to the distorted raw signals (see Fig. D.1 for the details).

All tools succeeded in detecting all four proteins across dilution samples (down to a 1/10 ratio) and triplicates, proving their limit of detection ranges. However, FLASHQuant showed the highest quantification accuracy, as relative fold change differences from the expected value were the smallest by the average value of 0.133, while 0.308 for ProSightPD, 0.143 for TopFD, and 0.14 for FLASHDeconv (see D.2.2 for the paired t-test for the significance values comparison). The fold change values of ProSightPD are more dispersed from the expected values than other tools, especially for the larger (21 and 28 kDa) and lower abundance proteins than others (9 and 11 kDa).

Moreover, when we calculated the coefficient of variation (CV) of replicates with feature group quantity, FLASHQuant and FLASHDeconv reported low median CV values of ∼0.08 from the four proteins, showing high reproducibility, while ProSightPD delivered higher values (∼0.29) (Fig. 4.6). TopFD also reported mostly low CV val-

ues, with medians of ∼0.1, but slightly higher (0.13) for the largest protein (28 kDa). Similar to the fold change analysis, FLASHQuant demonstrated its strength in reproducibility for the larger proteins by delivering smaller CV values than others.

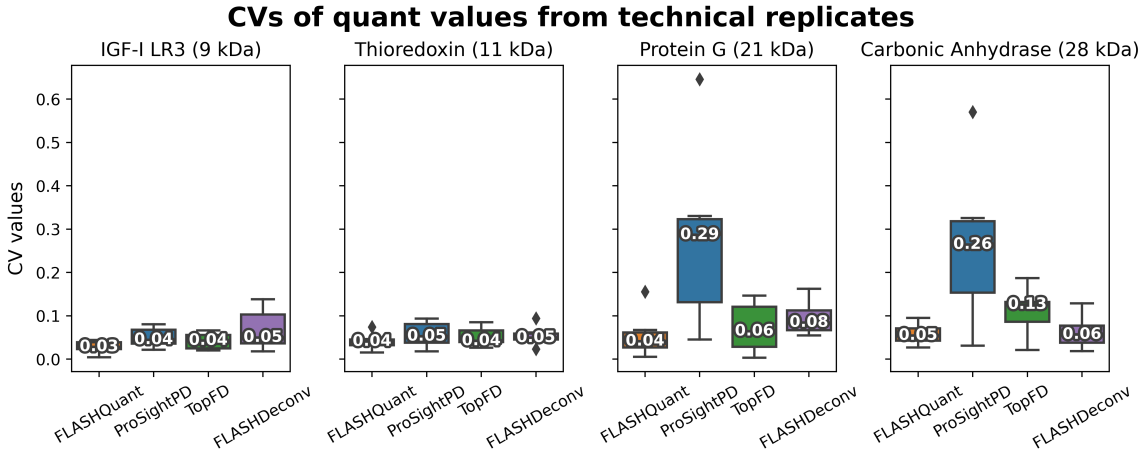


Figure 4.6: CV values from technical replicates of the SpikeIn dataset. Boxplot of CV values calculated from quantification values of technical replicates per spiked-in proteins. The digits on the boxplot indicate their median CV values. FLASHQuant shows consistently low CV values from all four proteins, while ProSightPD results in high CV values in larger molecules. Note that Carbonic Anhydrase is not only the largest protein but also has the lowest abundance among the four proteins.

4.3.3 FLASHQuant delivers a strong connection to the identification

In order to check the validity of the quantified feature groups, we attempted to explain the masses of the output feature groups from the simplest dataset available, PIPMix (see Fig. D.3 for the total number of results). This is achieved by identifying proteoforms from the tandem mass spectrometry (MS/MS) spectra in the dataset and comparing their masses to the feature groups reported by four tools.

We began by detecting, for each tool, consensus feature groups among all output feature groups from technical triplicates. Consensus feature groups refer to the feature groups detected across all the runs in the entire dataset. The default parameter settings of ProSightPD, a mass tolerance of 100 ppm and a retention time tolerance of 8 min (used for all consensus feature group detection from hereon), were applied to this process. For the detection of consensus feature groups in FLASHQuant, TopFD, and FLASHDeconv, we utilized the ConsensusFeatureGroupDetector. The masses of consensus feature groups were then compared to the proteoform masses identified by either TopPIC or ProSightPD Search within 20 ppm mass tolerance (see Table D.2 for the detailed number of identified proteoforms). ProSightPD Search is different from

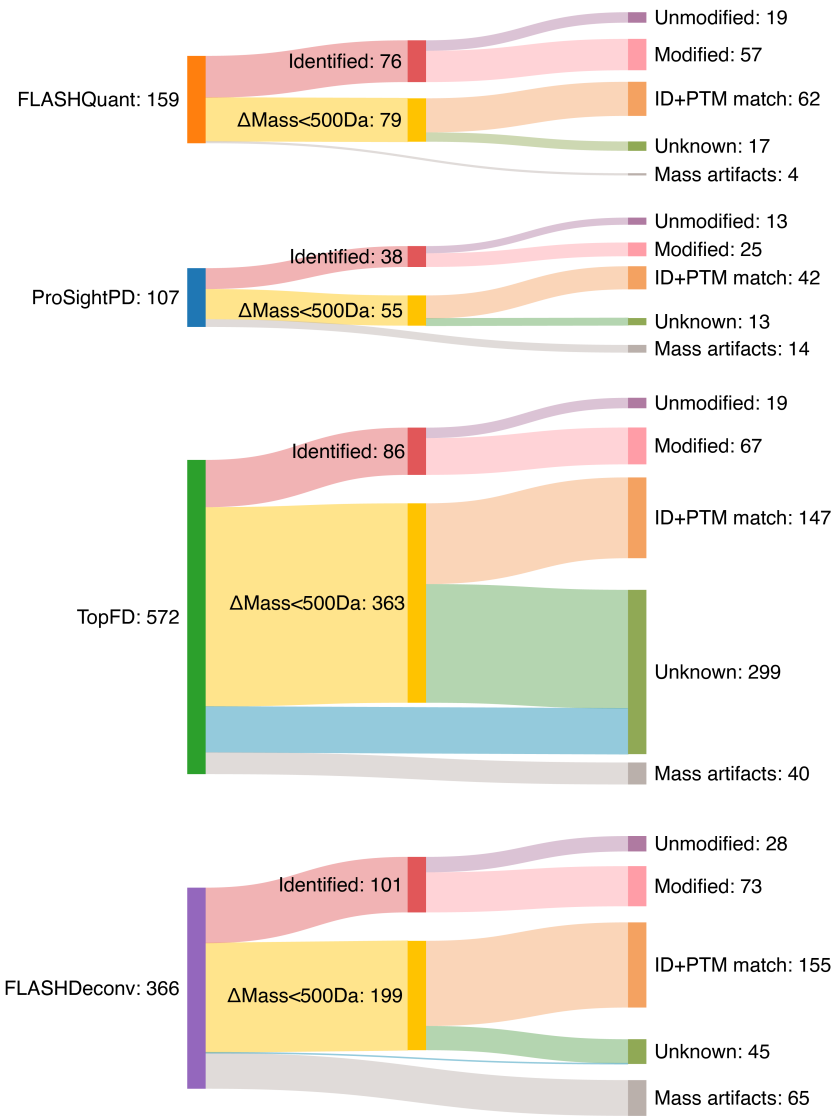


Figure 4.7: Sankey diagram of the number of consensus feature groups from the PIPMix dataset. The Sankey diagram shows the number and type of consensus feature groups. The identified type represents consensus feature groups having the same mass as identified proteoforms and has two flows depending on whether they matched unmodified or modified identified proteoforms. The rest feature groups with a mass matched to an identified proteoform within 500 Da mass tolerance are collected as the Δmass<500Da type. From this type, when the mass difference is matched to PTM (i.e., has a Unimod accession number, excluding amino-acid substitutions), feature groups are considered to be possibly true, thus labeled as ID+PTM match type. Blue-colored flows contributing to the Unknown type indicate highly likely false positives, feature groups having masses that are largely off compared to the identified proteoforms (Δ mass≥500Da). Note that FLASHQuant and ProSightPD only have unknown type feature groups from Δ mass<500Da type (without blue-colored flows). The mass artifacts type includes isotopologues and harmonics. Diagram created using SankeyMATIC.

the ProSightPD LFQ (ProSightPD Feature Detector) and consists of the "ProSightPD 4.2 Annotated Proteoform Search" node and the "ProSightPD 4.2 Subsequence Search" node. Thus, identification and quantification were separately executed (see D.2.1 and D.2.3 for details on parameters and execution, respectively).

Fig. 4.7 shows the Sankey diagrams of each tool's consensus feature groups, depicting the portion of feature groups that can be explained with identification results. FLASHQuant and FLASHDeconv found 159 and 366 consensus feature groups. TopFD resulted in the most consensus feature groups while ProSightPD the least (572 vs. 107), showing almost 5 times the difference.

The types of consensus feature groups in Fig. 4.7 can be categorized into three cases: likely true (see 4.7, the identified type plus ID+PTM type), definitely false (the mass artifact type), and unknown. The likely true ones provide a measure of the sensitivity; the higher the number, the more sensitive. FLASHDeconv outperformed the others by reporting the most likely true consensus feature groups. To assess specificity, the number of mass artifacts can be used, and FLASHQuant had the lowest number. It is worth highlighting that FLASHQuant also provides the highest percentage of likely

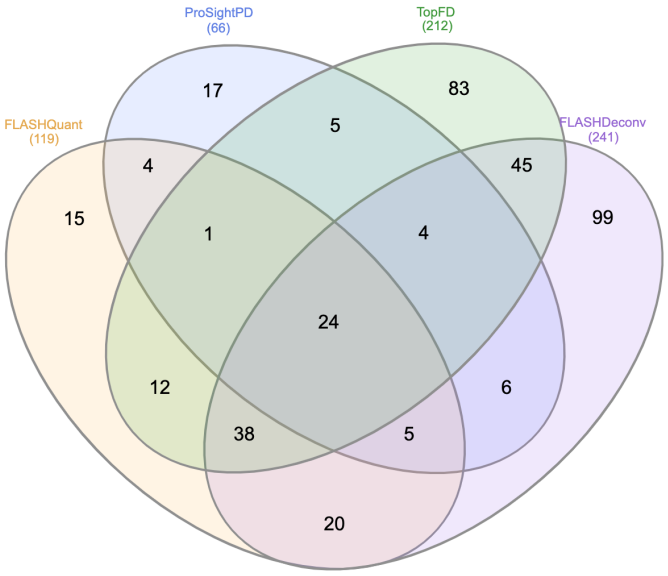


Figure 4.8: Venn diagram of the likely true feature groups from the PIPMix dataset. The Venn diagram illustrates the overlaps among the likely true feature groups detected by four tools. Rounded monoisotopic masses of the feature groups were used for this analysis. For FLASHQuant and ProSightPD, approximately 95% of their results were also verified by at least one other tool, indicating a high level of validation. Diagram created using InteractiVenn.

true ones, at 87%, and only 10% belonged to the unknown type, whereas more than half of the TopFD results were of the unknown type, confirming FLASHQuant's strong connection to the identification results.

The overlaps of likely true ones among the four tools are depicted as a Venn diagram in Fig. 4.8, indicating that many FLASHQuant and ProSightPD results are validated by other tools. The mass distributions between total consensus feature groups and likely true ones were compared in Fig. 4.9. While FLASHQuant, ProSightPD, and FLASHDeconv demonstrated similar trends between them, TopFD had predominantly abundant distribution in the <500 Da mass range, accounting for the large number of

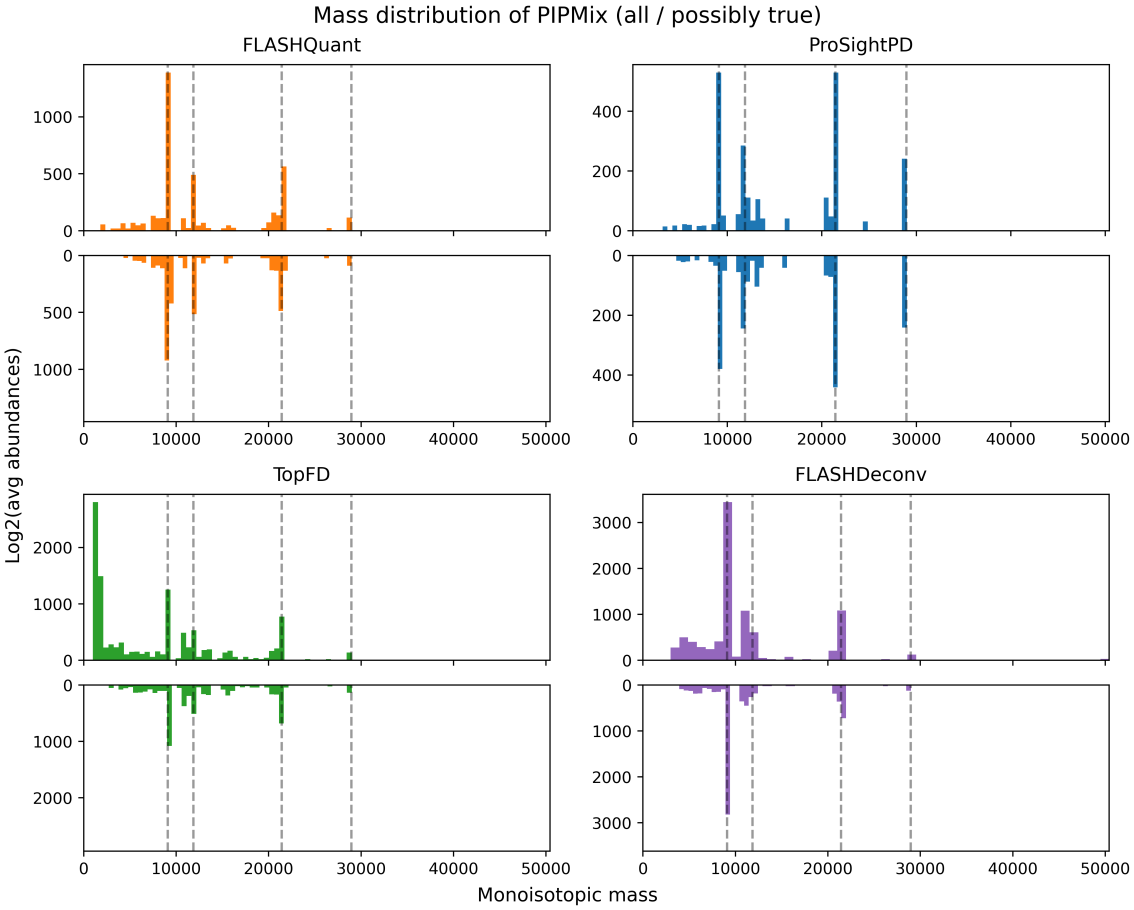


Figure 4.9: Mass distributions of the consensus feature groups from the PIPMix dataset. In each mirror plot, the upper one depicts the mass distribution of all consensus feature groups, while the lower one contains only likely true feature groups. The dataset's expected masses (four proteins) are marked as dashed lines. Similar trends between mirror plots were observed from FLASHQuant, ProSightPD, and FLASHDeconv, whereas TopFD includes high abundance distribution on the small mass range (>500 Da) only on the upper plot. This accounts for the large numbers of the unknown type feature groups reported from TopFD in Fig. 4.7.

unknown types in Fig. 4.7. As the detection of feature groups on the MS1 level does not necessarily lead to identifications on the MS/MS level (i.e., due to incomplete data acquisition, insufficient fragmentation, or database search limitations), the unknown ones may not necessarily imply false detection of features for quantification. In fact, those from FLASHQuant had plausible isotope patterns (see Fig D.4 for examples), indicating a possibility of multiple PTMs or truncation. Nonetheless, it is unexpected to find a large number of features smaller than 500 Da (equivalent to the mass of 5-9 amino acids) in this well-controlled PIPMix dataset, which contains only six intact protein species of large masses.

4.3.4 Performance on proteome-wide quantification

FLASHQuant performs rather conservative feature group detection and quantification, leading to fewer consensus feature groups than other tools at the proteome level (See Fig. D.5 for the total number of results). However, the quantification variances from FLASHQuant were relatively small, while its quantification reproducibility was high (See Fig. 4.10). Also, when we matched the masses of those consensus feature groups against the identified proteoform masses (by either TopPIC or ProSightPD Search), 84.8% consensus feature groups were matched for FLASHQuant, while 70.2% for ProSightPD, 55.6% for TopFD, and 76% for FLASHDeconv. These are the same trends as the above analysis of the PIPMix dataset, with more percentages of consensus feature groups from FLASHQuant being validated by identified proteoforms than from other tools.

To test the quantification accuracy, we evaluated relative fold changes of human Caco-2 proteoforms in the ProteomeMix dataset (Fig. 4.10A). Among the reported feature groups from each LC-MS run, those with a mass within 20 ppm tolerance to identified human Caco-2 proteoform masses were selected for further analysis. Relative fold changes were calculated as in the SpikeIn dataset analysis, based on the average quantity at the 1:1 mixing ratio, then logarithmized. The average difference between the measured and expected fold changes is the smallest in FLASHQuant (Fig. 4.10B), with median values of 0.45. Fig. 4.10A shows that the values of all ratio samples in ProSightPD have larger variances than in other tools, presenting a similar trend as in SpikeIn dataset analysis.

Quantification reproducibility between replicates was demonstrated through a linear regression between each quantity of the jointly detected feature groups (Fig 4.10C and D.6). Note that jointly detected feature groups represent the feature groups detected across replicates per sample, whereas consensus feature groups used in the

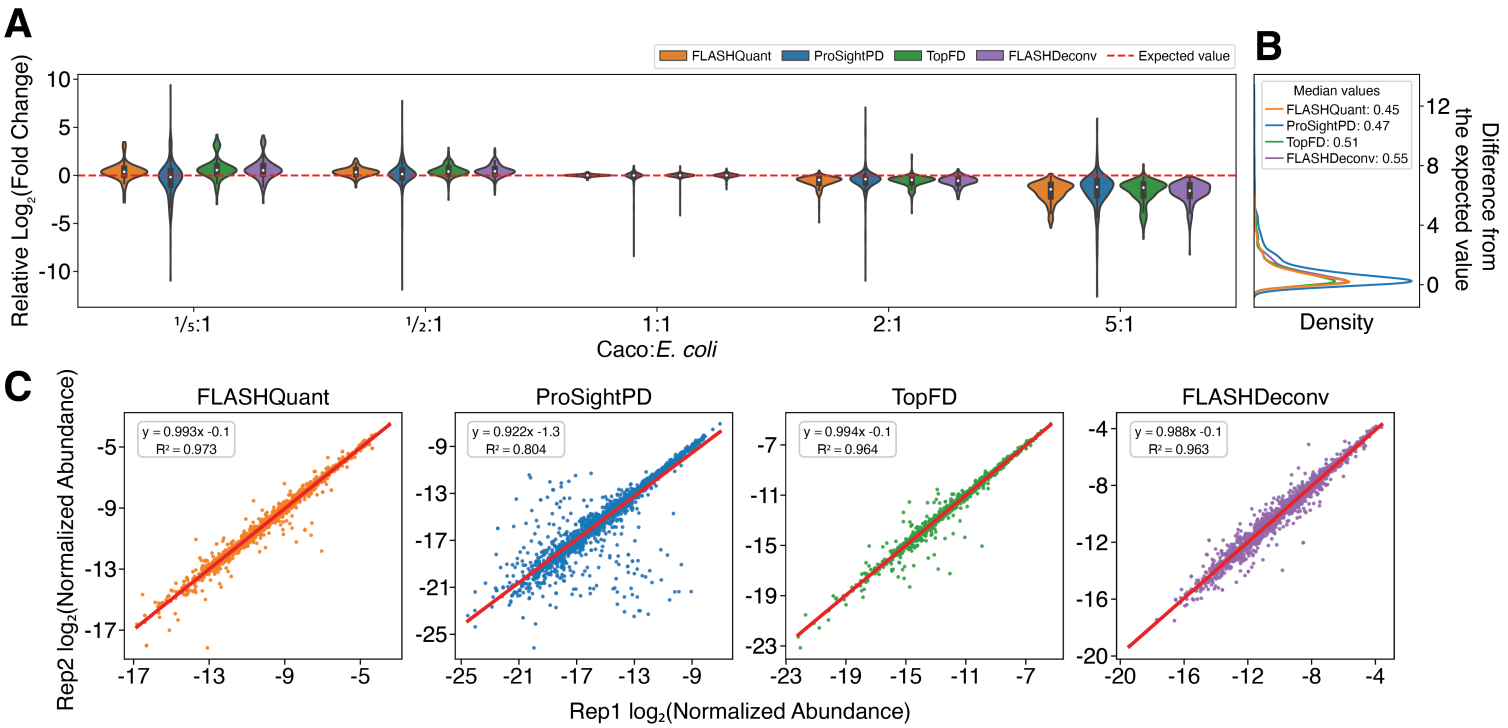


Figure 4.10: Data analysis of the ProteomeMix dataset. **(A)** Relative fold changes comparison for the ProteomeMix dataset. The consensus feature groups of all samples (5 different mixing ratios and two replicates each) that were matched to the human Caco-2 proteoforms (identified by either TopPIC or ProSightPD Search) were employed in this analysis. The violin plots depict the relative logarithmized fold change values per mixture ratio. Their differences from the expected value (the red dashed line in (A)) are shown in **(B)** as in the density plot. FLASHQuant shows less variance to the expected values with minimum median than other tools. The number of consensus feature groups used for this analysis is 134 for FLASHQuant, 262 for ProSightPD, 129 for TopFD, and 147 for FLASHDeconv. **(C)**Quantification reproducibility of the jointly detected feature groups matched against identified proteoforms for the ProteomeMix dataset. The jointly detected feature groups between technical replicates were taken, and those with masses matched to the identified proteoforms (either by TopPIC or ProSightPD) within 20 ppm tolerance are plotted in the figure. Reproducibility between technical replicates has been evaluated by calculating linear regression between each normalized quantity (depicted as a red line). FLASHQuant and TopFD demonstrated high reproducibility with a higher R^2 value and regression slope closer to 1 than other tools.

previous analysis are from the entire dataset. Between the replicates of each sample, we collected the jointly detected feature groups (same mass and retention time tolerance as consensus feature groups) and then evaluated the similarity of their normalized quantities (based on respective TICs). FLASHQuant, TopFD, and FLASHDeconv showed high reproducibility with R^2 values of >0.95 and regression slopes of >0.98, while ProSightPD resulted in less concentrated values to the linear regression line, with R^2 values of 0.7 and regression slopes of 0.86 (Fig. D.6). When only the jointly detected feature groups matching the identified proteoforms were considered (Fig 4.10C), all tools showed high reproducibility with R^2 values of >0.8 and regression slopes of >0.9. However, slightly closer-to-one R^2 and regression slope values were observed in FLASHQuant and TopFD than the others in all samples.

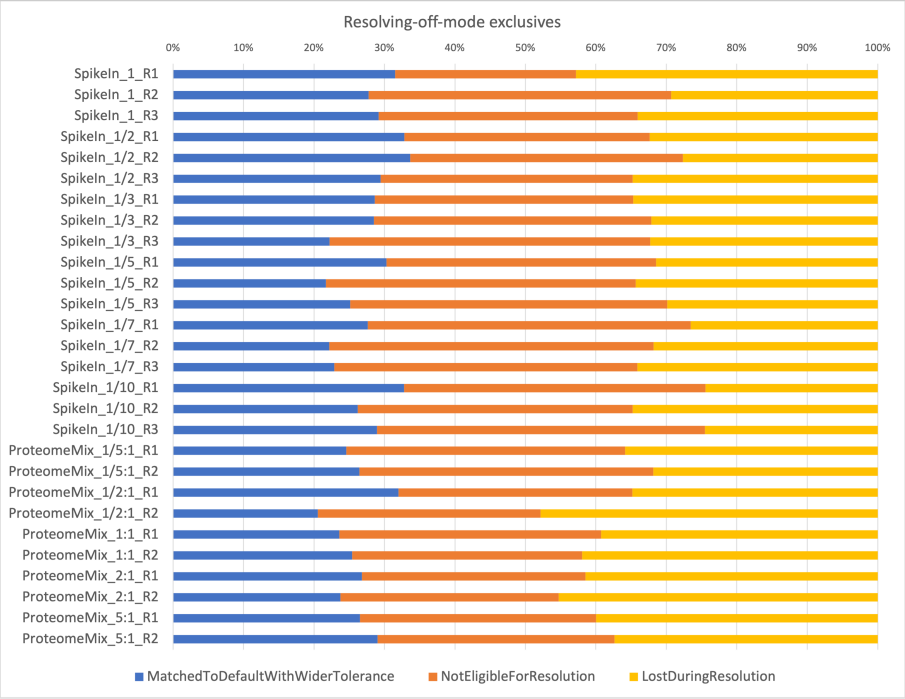


Figure 4.11: Barplot of resolving-off-mode exclusives analysis. "Resolving-off-mode" exclusives were compared with FLASHQuant "default-mode" results to determine where they could be originated. 20-35 % of the exclusives were matched to FLASHQuant results (blue bars) when mass and retention time tolerances were widened (10 Da and 6 min). These exclusives may have been redundant and removed in FLASHQuant. Mostly, the exclusives appear to have been eliminated during the conflict resolution method (red and yellow bars). Red bars indicate that a large portion of the exclusives were not eligible for resolution (i.e., harmonics or composed of only shared m/z traces) and, thus, were removed before the resolution method. The exclusives that were likely to be removed during the conflict resolution were marked in yellow bars.

4.3.5 Resolving overlapping proteoforms boosts the quantification accuracy

FLASHQuant experienced a boost in both SpikeIn and ProteomeMix datasets using the conflict resolution method. When FLASHQuant was executed without the conflict resolution method ("resolving-off-mode" from hereon vs. “default-mode”), still a large portion of reported feature groups (86-88 % for SpikeIn and 80-86 % for ProteomeMix) overlapped with the default-mode (see Table D.3 for the numbers of overlapped or exclusive feature groups). These overlapped feature groups (within mass and retention time tolerances) from the two modes even showed similar dynamic ranges. The exclusive ones from the resolving-off-mode were mostly eliminated during or after the conflict resolution method in the default-mode, as most of their m/z traces were incorrectly assigned to them (Fig. 4.11).

The conflict resolution method not only removed possible errors in the feature group assignment but also alleviated the quantification errors. From the ProteomeMix dataset, the exclusive ones from the resolving-off-mode have larger relative fold change differences from the expected values compared to the default-mode (Fig. 4.12A). Also,

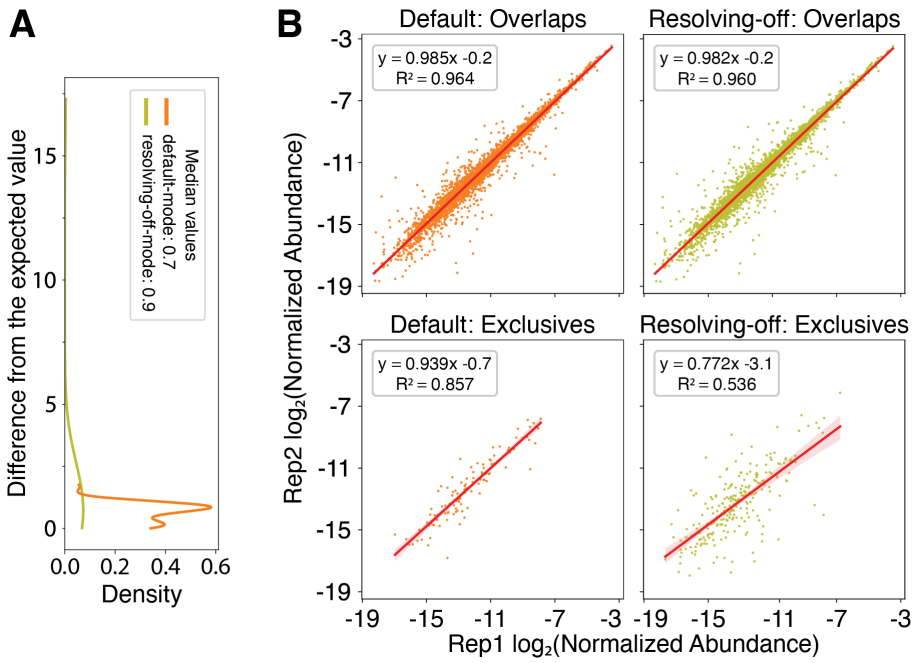


Figure 4.12: Comparison between the default-mode and resolving-off-mode FLASHQuant with the ProteomeMix dataset. **(A)** Analogue of Fig. 4.10B for the relative fold change differences from the expected value with the default-mode and resolving-off-mode exclusives, showing the conflict resolution method offers higher quantification accuracy. **(B)** Analogue of Figure 4.10C for the reproducibility comparison between the default-mode and resolving-off-mode. Overlaps between two modes show high reproducibility, while exclusives of the default-mode outperformed exclusives of the resolving-off-mode.

reproducibility between technical replicates was higher for the ones from the default-mode (Fig. 4.12B for the ProteomeMix and D.7 for the SpikeIn).

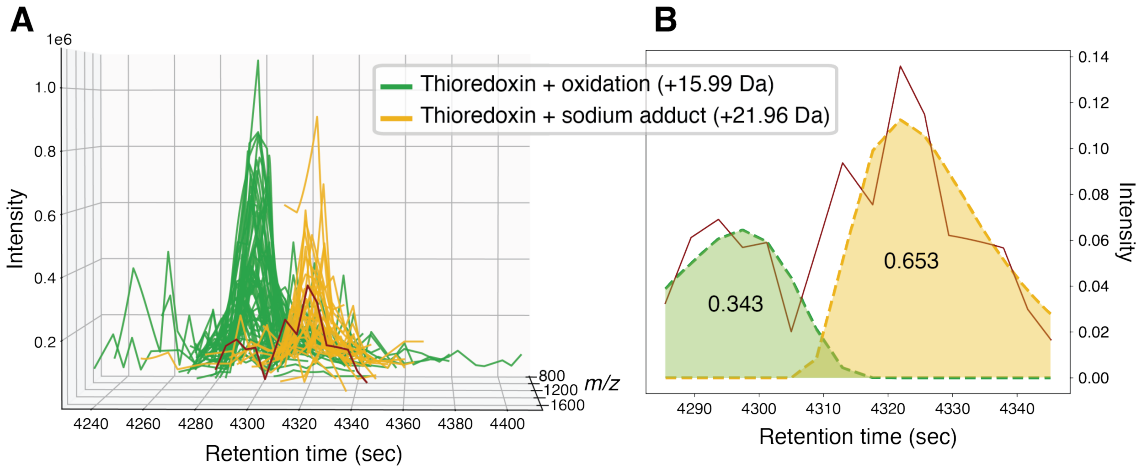


Figure 4.13: An example of how the conflict resolution method works in FLASHQuant. Two proteoforms from the SpikeIn dataset with a 6 Da mass difference were indicated as green and yellow. One of their shared mz traces is picked and colored in dark red as an example. **(A)** All raw unique m/z traces of each proteoform. **(B)** Modeled theoretical trace shapes for each proteoform. The black digits on the theoretical shapes refer to the resolved abundance ratio for each proteoform.

A detailed example of how the conflict resolution method performed on proteoforms is illustrated in Fig. 4.13, which shows the raw signals (i.e., m/z traces) from two proteoforms of the target protein Thioredoxin (11,858.04 Da) in the SpikeIn dataset (ratio $1/2$). The green traces are Thioredoxin with an oxidation (+15.99 Da), while the yellow with a sodium adduct (+21.96 Da). Due to a difference of only 6 Da in mass between the two proteoforms, they inevitably shared some m/z traces, depicted as dark red. However, with the conflict resolution method, the abundance of the shared signal was successfully distributed to the two proteoforms, with 34.3% attributed to the oxidized proteoform and 65.3% to the proteoform with the sodium adduct (the remaining 0.4% corresponds to the portion unexplained by both proteoform feature groups, e.g., noisy component).

4.4 Discussion

Here, we introduce FLASHQuant, a robust quantification tool designed explicitly for MS1-level LFQ data analysis in TDP. One of the key features of FLASHQuant is its automatic conflicting resolution method, which addresses the commonly overlooked yet critical problem of overlapping signals. By effectively resolving overlapping sig-

nals, FLASHQuant ensures highly accurate quantification and offers remarkable reproducibility among technical replicates. Benefiting from the ultrafast and robust FLASHDeconv algorithm, FLASHQuant achieves a rapid runtime of ∼1.5 min per LC-MS run (containing over 1,800 MS1 spectra).

In this paper, we evaluated the quantification performance of four tools, comparing several important metrics, including reproducibility, quantification accuracy, and specificity. Our results demonstrated that ProSightPD performed unsatisfactorily in reproducibility and quantification accuracy (Fig. 4.5 and 4.10), yielding biased results. More significant fold change and the CV value differences were observed in larger proteins (21 kDa and 28 kDa, with lower abundances), while smaller proteins exhibited fewer variations. TopFD, while showing better reproducibility and accuracy than ProSightPD, exhibited low specificity, with a higher number of quantified artifacts (Fig. 4.7). On the other hand, both FLASHQuant and FLASHDeconv demonstrated similarly strong performance in reproducibility and accuracy across a wide range of samples. This is not surprising due to their shared deconvolution algorithm, which is highly relevant to quantification performance. However, FLASHQuant stood out for its high specificity, yielding fewer mass artifacts than FLASHDeconv, proving its strength in delivering reliable results.

In addition to FLASHQuant, we also provided two accompanying tools, ConsensusFeatureGroupDetector and FLASHQuantWizard, to enhance the user experience. We have implemented a Python script to visualize the overlapping signals, such as in Fig. 4.13. In the near future, we plan to develop a simple web application that provides interactive visualization of quantified feature groups (which will be shared at `https://openms.org/flashquant/`). This application will greatly assist users in visually examining the raw signals, which is essential in LFQ. Moreover, given the strong reliance of FLASHQuant on FLASHDeconv and both tools being part of OpenMS, we are actively working on integrating the results from both tools, which will offer users a comprehensive approach that combines the strengths of both tools.

It is worth noting that while FLASHQuant excels in accuracy and reproducibility, the other tools outperformed it in terms of sensitivity (Fig. D.2, D.3, and D.5). This trade-off can be attributed to FLASHQuant's use of strict scoring thresholds, which prioritize specificity to reduce false positives - a common issue in proteoform-level findings. In Fig. 4.7, the increased numbers from TopFD and FLASHDeconv largely come from ID+PTM matches or unknown, where the possibility of false positives still remains. Despite reporting fewer total results, FLASHQuant achieved the highest percentage of likely true proteoforms (90%) with the fewest mass artifacts. Also, given that the PIPMix dataset contains only four target proteins, the lower number of

detections by FLASHQuant should not be viewed negatively, as a similar trend was seen with ProSightPD with a large reduction in the number of results compared to other datasets. While an ideal evaluation would involve Receiver Operating Characteristic (ROC) curves to balance sensitivity and specificity, it is impractical to control such metrics across different tools, as each is optimized for different levels of sensitivity and specificity. Therefore, direct comparisons based solely on sensitivity metrics may not fully reflect FLASHQuant's robust performance.

Furthermore, in order to enhance the reproducibility of FLASHQuant further, a retention time alignment method could be applied. This alignment method would reduce variation in retention times across technical replicates, thereby improving the consistency and comparability of quantification results. Additionally, we acknowledge the importance of incorporating statistical functions in conjunction with FLASHQuant results. To address this need, we are exploring the possibility of integrating statistical methods directly into FLASHQuant or providing users with R scripts for performing comprehensive downstream statistical analysis on the FLASHQuant results.

Chapter 5

Web application for visualizing proteoform signals

5.1 Introduction

Top-down proteomics (TDP) based on mass spectrometry (MS) has emerged and has advanced in various aspects, including separation, MS techniques, and data analysis software. Among available software tools, only a few incorporate visualization features and are freely available, such as ProSight Lite[86], Proteoform Suite[52], MASH Explorer[87], TopMSV[88], and TopDownApp[89]. As the prevailing trend in software development leans towards web applications, the recently developed tools adopted this option. For example, TopMSV is the first web application for visualizing TDP-MS data and also supports executing TopPIC[60]. Additionally, TopDownApp, a modularized tool for generating data analysis workflows, has been recently introduced, allowing users to analyze and visualize high-throughput TDP-MS datasets. Including these two, all visualization software offers a predefined layout for the visualization, focusing on their specific strengths.

As TDP studies still require manual validation and distinct types of visualization differ, a configurable layout can enhance usability. Implementing this configurable layout necessitates the modularization of visualization components. Once components are modularized, the potential for extending various component types emerges, enabling the significant expansion of the viewer's functionality. This approach facilitates the development of different figures and plots upon user request, without compromising runtime efficiency based on the chosen layout configuration. For optimal speed, users have the flexibility to selectively include only the necessary components, avoiding

unnecessary computational overhead. Furthermore, the versatility of the viewer is increased, as it can accommodate an extended range of applications.

The development of modularized and extendable visualization software is best realized through open-source software in a developer-friendly environment, which encourages community contributions with lower barriers. For this, Python is the programming language of choice, accompanied by the Streamlit framework for web application development. Streamlit simplifies web application development without requiring front-end expertise, making it popular for many scientific web applications[90,91]. OpenMS[38], a powerful framework for MS data analysis, recently initiated the Streamlit project for building web applications using pyOpenMS and the TOPP command line interface tools[41,92]. With their template and deployment resources, this serves as a cornerstone for various MS data analysis tools.

Here, we present FLASHViewer, a web application for visualizing proteoform signals with a user-configurable layout. Specialized visualization components tailored to TDP-MS analysis, such as protein sequence view and internal fragment map, are implemented in a modular manner. Powered by Vue.js (Vue from hereon) for efficient rendering of visualization components, FLASHViewer ensures fast interactivity within the viewer. FLASHDeconv[44] (from Chapter 3) was mainly employed for input, and we showcase the scalability of FLASHViewer through the usage of FLASHQuant (from Chapter 4). FLASHViewer is open-source and platform-independent software that can be executed either locally or remotely (`https://abi-services.cs.uni-tuebingen.de/flashviewer/`). Also, all visualization components can be saved as images (Scalable Vector Graphics (SVG) format) for publications.

5.2 Methods

5.2.1 Architecture of FLASHViewer

FLASHViewer is developed based on the OpenMS Streamlit template[92] and is written in Python and TypeScript (Fig. 5.1). For front-end development, except for the Viewer page, FLASHViewer utilizes the Streamlit Python framework (version 1.34.0). Also, most of the back-end, including data parsing and preparation, occurs on the Streamlit side, leveraging libraries such as pyOpenMS and pandas for data processing. User inputs are managed using Streamlit's predefined functionalities, including interface elements such as markdowns, tabs, forms, file uploads, alerts, etc. FLASHViewer employs Streamlit-Pages (version 0.4.5), a third-party package, for multi-page app support, allowing selective pages to be displayed for each tool (i.e., different pages are

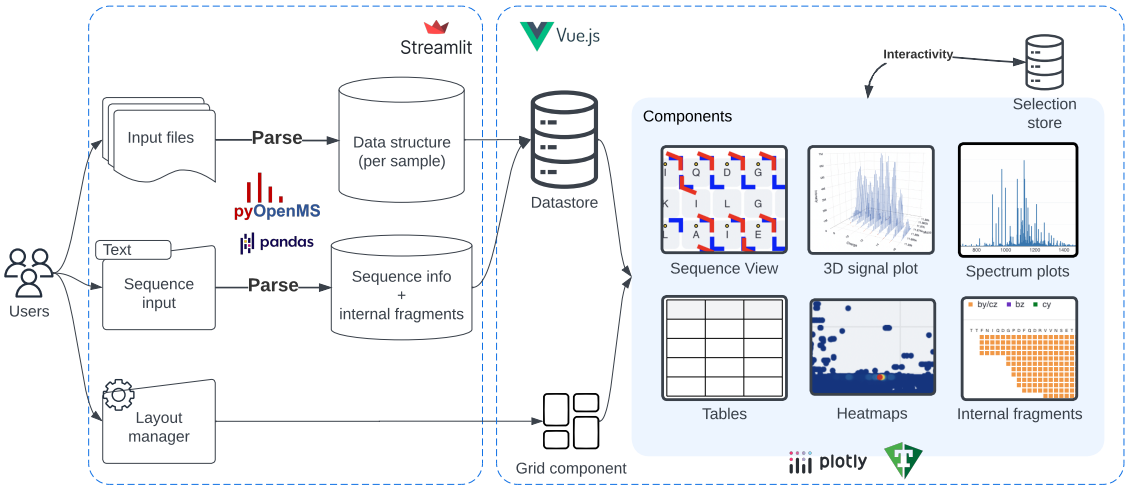


Figure 5.1: An architecture diagram to describe the data flow. This figure is created in Lucidchart `www.lucidchart.com`.

shown for FLASHDeconv and FLASHQuant). Streamlit's session management, caching (cache_data) and session state (session_state), is used to store and persist data between sessions and reruns within the browser, enhancing performance by avoiding repeated callbacks and enabling data sharing across multiple pages.

To enhance performance in rendering complex plots and figures, we developed the Viewer page using Vue with TypeScript. Data processed on the Streamlit side is converted into JavaScript Object Notation (JSON) format and sent to the TypeScript side through Streamlit's custom components (Streamlit.components.v1). The Vue framework speeds up the interactive plot and table rendering. Libraries such as Plotly.js, Tabulator, and Vuetify were utilized for visualization components. The TypeScript code is compiled into JavaScript for deployment, ensuring compatibility with web browsers, and Node.js is used for package management. This integration of Streamlit and Vue, along with efficient data flow between Python and TypeScript, allows FLASHViewer to provide a responsive and interactive user experience in a multi-page application.

FLASHViewer runs on a Streamlit server, either locally or remotely. For this, TypeScript codes from Vue are compiled and built into JavaScript to be integrated into the Streamlit application. Two options are available to run FLASHViewer locally: use the publicly available source code or Windows installation file. The online version of FLASHViewer was deployed using Docker and can be accessed on the university server (`https://abi-services.cs.uni-tuebingen.de/flashviewer/`).

FLASHViewer is an open-source, platform-independent software freely available at `https://github.com/JeeH-K/FLASHViewer`. The source codes for FLASHViewer's

Vue component can be found at `https://github.com/drewbudd/openms-streamlit-vue-component` and is bonded to the main GitHub repository as a submodule. Additionally, the OpenMS Streamlit app offers a repository to deploy its multiple apps together at `https://github.com/OpenMS/streamlit-deployment`, which allows FLASHViewer to be easily deployed for online access. A detailed description of the implementation of FLASHViewer will be discussed in the following sections.

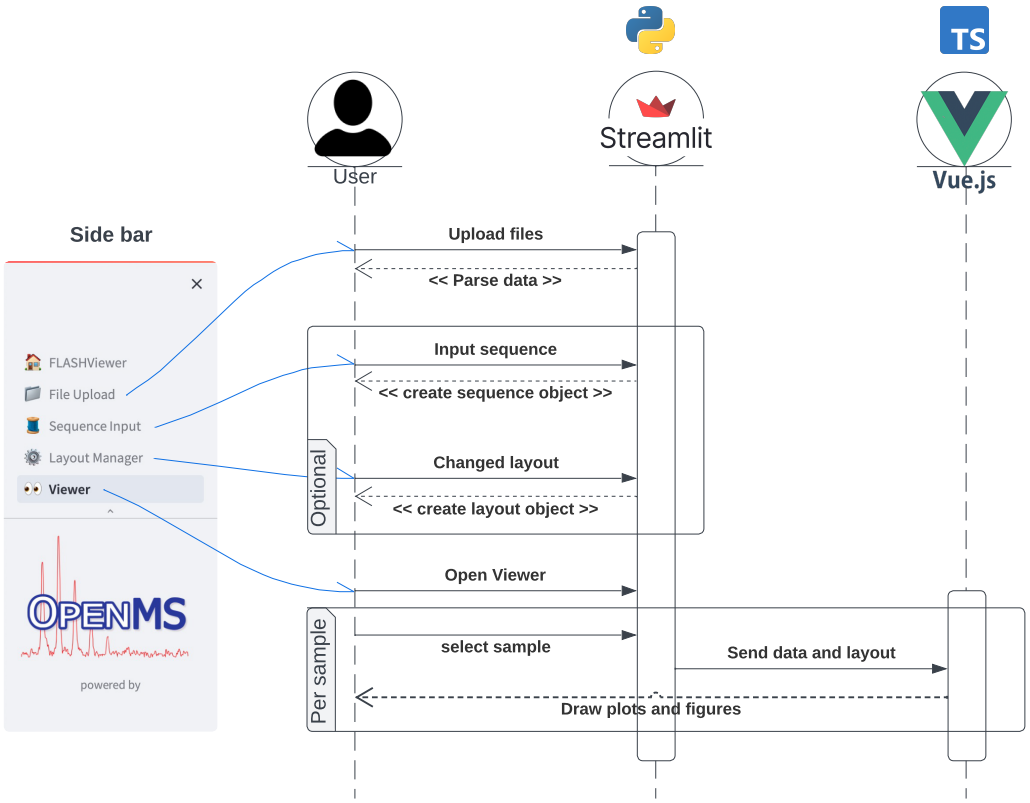


Figure 5.2: The UML sequence diagram of FLASHViewer. The screenshot on the left shows available pages in FLASHViewer's sidebar. All the pages in FLASHViewer are developed using Streamlit and written in Python. File uploading is required, but the sequence input page and the layout manager page are optional. If the user skips these two pages, all the available components will be displayed on the viewer page, except for the disabled viewer components for sequences (sequence viewer and internal fragments). When entering the viewer page, the part displaying the figures and plots was written in TypeScript using Vue for speed. This figure is created in Lucidchart `www.lucidchart.com`.

5.2.2 User interaction in FLASHViewer

Fig. 5.2 illustrates how users interface with FLASHViewer using a Unified Modeling Language (UML) sequence diagram. Users mostly interact with the front-end side of Streamlit, except for the Viewer page. Three different pages, the File Upload, Sequence Input, and Layout Manager pages, require user input. Those inputs are parsed and saved in Streamlit Session State (i.e., in the browser). The Viewer page takes the saved session data and sends it to the Vue side to prepare and display figures and plots to display (see Section 5.2.3 for details).

The pages within FLASHViewer can be navigated through the sidebar (Fig. 5.2, on the left). On the main page of FLASHViewer, users can switch between two tools, FLASHDeconv and FLASHQuant. When switching between these tools, distinct sets of pages become available in the sidebar. We will primarily focus on the pages within the FLASHDeconv mode, considering that this set encompasses the pages available in FLASHQuant mode.

5.2.3 Functionalities of FLASHViewer

Input files

The input for FLASHViewer comes from the output of FLASHDeconv; two mzML format output files for each sample: annotated and deconvolved output files (Fig. 5.3). FLASHViewer allows uploading multiple result files from different samples. The viewer page can display multiple samples in one scroll or toggle between each sample, depending on the user setting. The default setting is toggling, if not specified on the layout manager page.

Once uploaded, files are parsed into pandas dataframe format for each sample using pyOpenMS and pandas. Each row in a dataframe corresponds to scan-level information, including details such as MS-level, scan number, retention time, deconvolved masses, precursor information (relevant only for MS2 scans; precursor scan number and mass), and specific data per deconvolved mass (e.g., charge ranges, isotope ranges, and corresponding signal and noise peaks).

At the time of writing, FLASHDeconv is being used as the core deconvolution software for the input source, but other deconvolution software can be used if the right parser is implemented. This was demonstrated using FLASHQuant as the input source. For FLASHQuant, one consensusXML file for each sample is needed, and a respective parser was implemented.

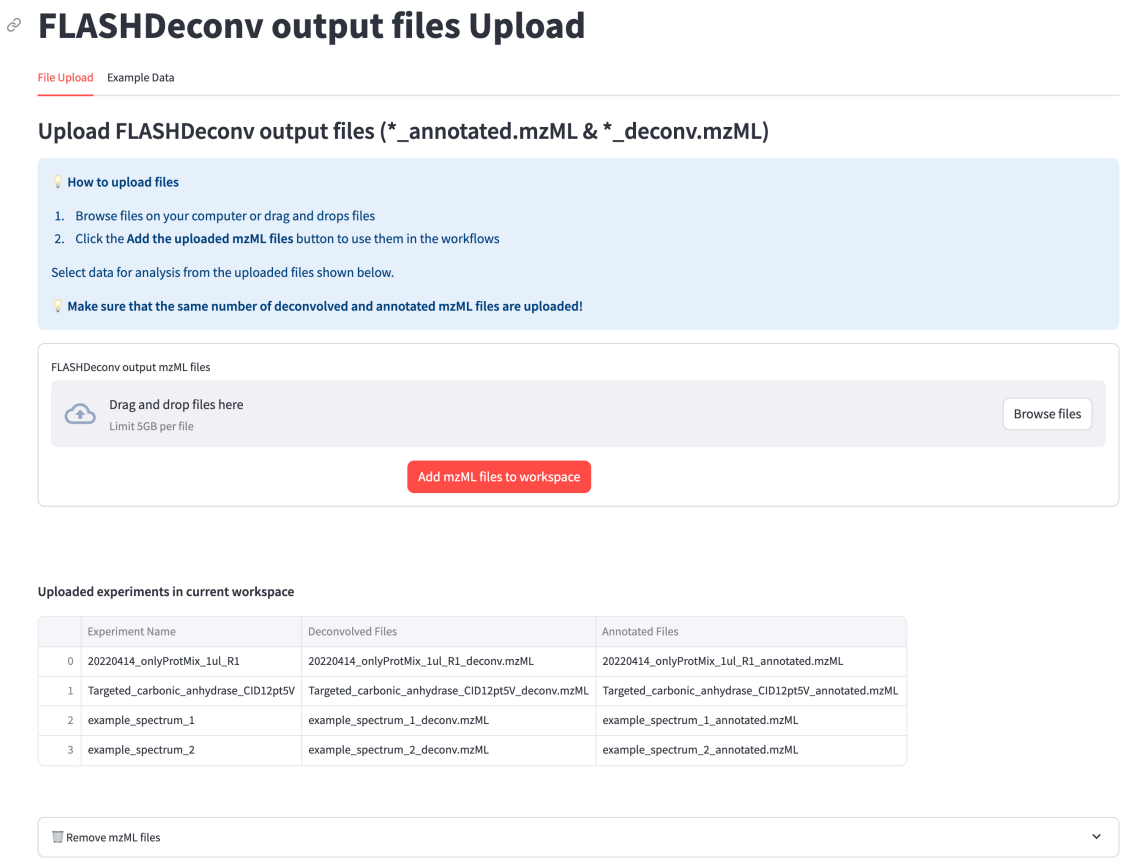


Figure 5.3: An example screenshot of the File Upload page. Once files are uploaded, they are listed in the table with the corresponding experiment names. Multiple samples (experiments) can be uploaded at once.

Displaying a proteoform sequence

Visualization of proteoform sequence coverage and fragment ion map is popularly demanded for MS/MS analysis. FLASHDeconv offers deconvolution at the MS2 level since the 2.0 beta version. This allows for a simple matching analysis between deconvolved fragment masses and theoretical proteoform masses. FLASHViewer has two components in the Viewer, the sequence view and the internal fragment map, to visualize this matching analysis (see Section 5.3.1 for details and figures).

The sequence view component is designed to visualize the possible matches between the theoretical fragment masses from a given protein sequence and deconvolved fragment masses. Once an MS2 scan is selected on the scan table, the sequence view updates and shows the corresponding matches. The core design was inspired by ProSight Lite[86], a popular GUI for visualizing fragment matching between a sequence and a peak list.

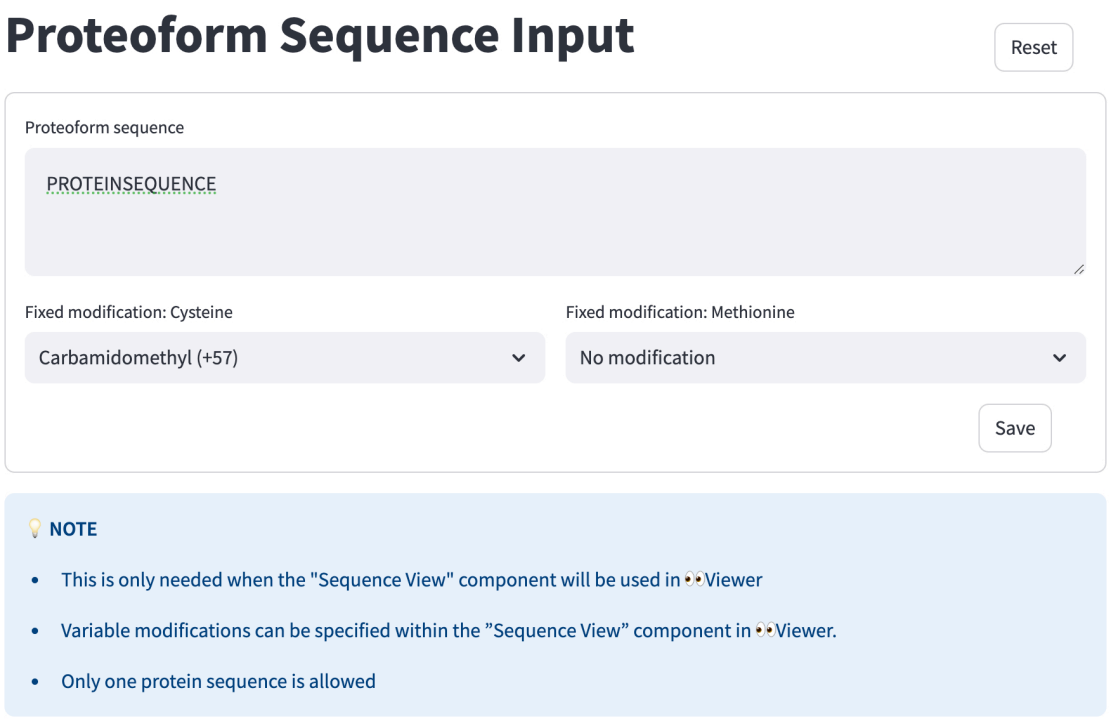


Figure 5.4: An example screenshot of the Sequence Input page.

While the sequence view supports visualizing terminal fragment ions, the internal fragment map is specifically designed to focus on visualizing internal fragment ions. Terminal fragment ions, originating from the conventional cleavages, have the amino or carboxy terminus; on the contrary, internal ions are derived from multiple cleavages, containing no protein terminus[93]. Given the frequent occurrence of internal ions in TDP due to the larger molecules[94], having a dedicated visualization component specifically designed for them is crucial. This component employs a straightforward matching method, comparing theoretical ion masses derived from the given protein sequence (on the Sequence Input page) with deconvolved MS2 fragment masses.

Input from users is required to activate these two viewer components; otherwise, these components will not appear as options on the Layout Manager page. On the Sequence Input page (Fig. 5.4), users can put a proteoform sequence and select fixed modifications (for cysteine and methionine). pyOpenMS is used to parse these inputs into theoretical fragment masses.

Configurable layout with modularized components

The configurable layout allows users to tailor the visualization app to their unique needs and preferences (e.g., targeted vs. discovery analysis). It is especially advantageous in TDP, where manual validation still remains prevalent. Achieving this flexibility relies on modularized visualization components that users can selectively choose and arrange according to their preferences. Also, this design streamlines the implementation of new components. To mitigate potential confusion, particularly for new users, a default layout with all available components is used if users have not specified any configuration.

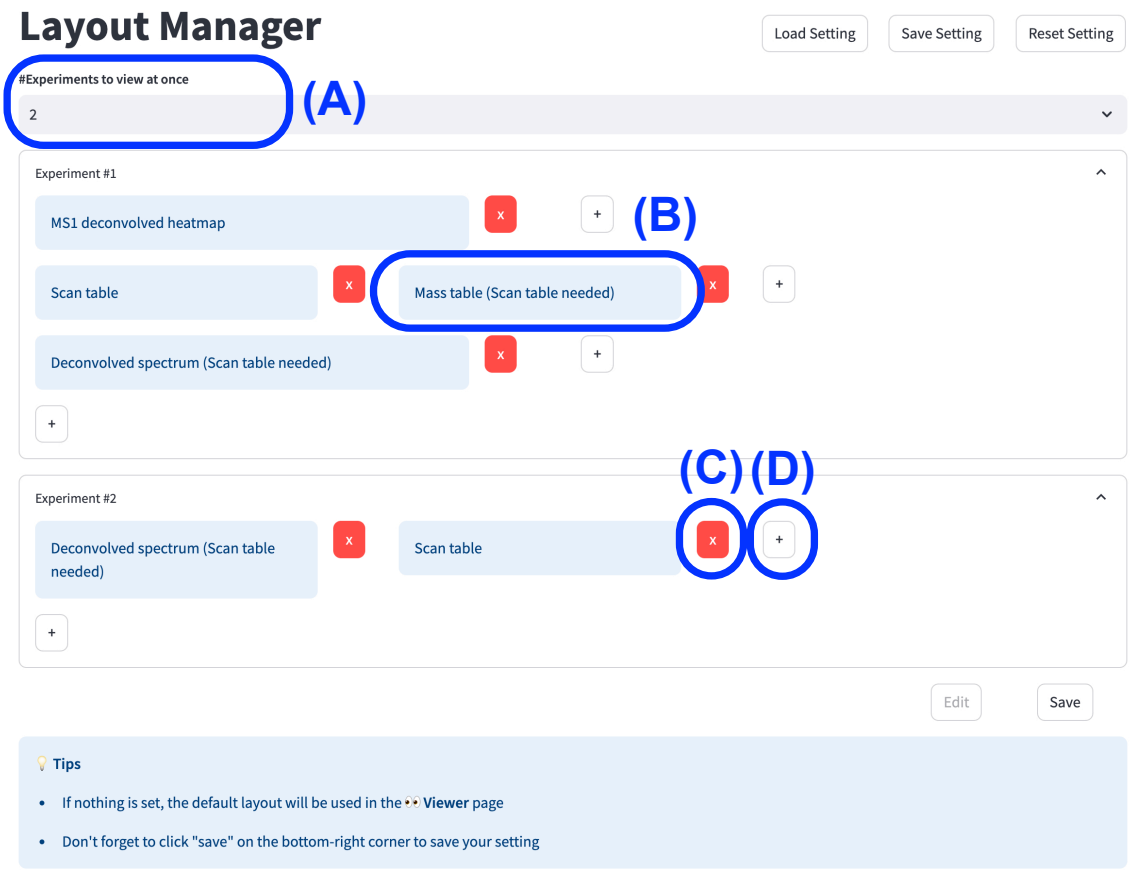


Figure 5.5: An example screenshot of the Layout Manager page. **(A)** It was selected to display two samples (experiments) in one scroll. The layouts for the first and second experiments were chosen differently, with four components for Experiment #1 and two for Experiment #2. **(B)** If a component type has a dependency on another type (e.g., a mass table requires a scan table), this is indicated in the component type name. **(C)** Clicking the red 'x' mark will delete the selected component, and **(D)** the '+' mark will add a new row or column.

The layout of the Viewer page can be configured on the Layout Manager page (Fig. 5.5). It allows users to choose the number of samples to display in one scroll (up to five). Such configurability facilitates smooth comparative analysis among samples (e.g., comparison between technical replicates). Users have the flexibility to decide which samples are displayed where on the Viewer page, and it is also possible to display a single sample multiple times.

As shown in Fig. 5.5, FLASHViewer allows displaying different component types tailored to each experiment location. For example, the first location contains a heatmap, two tables, and a deconvolved spectrum view, while the second location only has a table and a deconvolved spectrum view. Each location is even customizable with distinct row and column layouts; e.g., the first location has two rows of one column and one row of two columns, while the second location adopts a layout of one row with two columns. Furthermore, the final layout configuration can be saved in a JSON format file and loaded later (Fig. 5.5 on the upper right).

Interactive viewer with TypeScript

An interactive viewer handling complex mass spectrometry data requires fast and effective visualization rendering to provide a seamless user experience. Especially when multiple graphic components are to be rendered within a web application, the use of JavaScript becomes crucial, as it is the primary scripting language supported by web browsers. In implementing the Viewer page of FLASHViewer, TypeScript, a superset of JavaScript, is employed along with the popular GUI framework, Vue, to enhance development structure and maintainability.

Fig. 5.1 illustrates the data flow from users to the visualization components. When users enter the Viewer page, Streamlit sends two pieces of data to Vue: layout configuration and parsed dataframes. The layout configuration from the Layout Manager is sent to the grid component in Vue. The grid component is only in charge of placing the requested components on the page according to the configuration. Each component then reads the data from the Datastore and draws the corresponding figure. This design was possible owing to the modularized components. The parsed dataframes from Streamilt are serialized and sent to Vue. There, they are deserialized to native JavaScript objects and stored in the Datastore.

Available component options at the time of writing are described in Table 5.1. Interaction between the components was allowed through the selection store (i.e., state in Vue). Given the modular nature of all components, a centralized storage mechanism was essential for data sharing among them. For instance, the mass table

Component options	Used library	Dependency
MS1 raw heatmap	Plotly	
MS1 deconvolved heatmap		
Annotated spectrum		Scan table
Deconvolved spectrum		Scan table
3D signal plot		Scan table Mass table
Scan table	Tabulator	
Mass table		Scan table
Sequence view	Vuetify	Mass table
Internal fragment map		Mass table

Table 5.1: Available visualization components and the libraries used to implement them. The Dependency column indicates which components are required for the corresponding component to function as intended. Examples of each component option are addressed in detail in the following Section 5.3.1.

depends on the selected row in the scan table. Once a row in the scan table is selected, its corresponding row number (i.e., scan number) is saved in the selection store. Then, the mass table listens for the row number change in the selection store and updates its content accordingly with the deconvolved masses from the selected scan.

5.2.4 Data description

The datasets from Chapter 4 were used for the example screenshots and testing runtimes. Details about the dataset generation can be found in Section 4.2.4.

5.3 Results

5.3.1 Visualization components

Fig. 5.6 shows example screenshots of the Viewer page with six visualization components. The deconvolved MS1 heatmap (located on the top) offers an overview of the selected sample on the mass and RT plane, with colors denoting mass intensities (m/z and RT plane for the MS1 raw heatmap). The scan table allows users to explore deconvolved and annotated signals in detail on a per-scan basis. The mass table displays deconvolved proteoform masses for a selected MS1 scan or deconvolved fragment

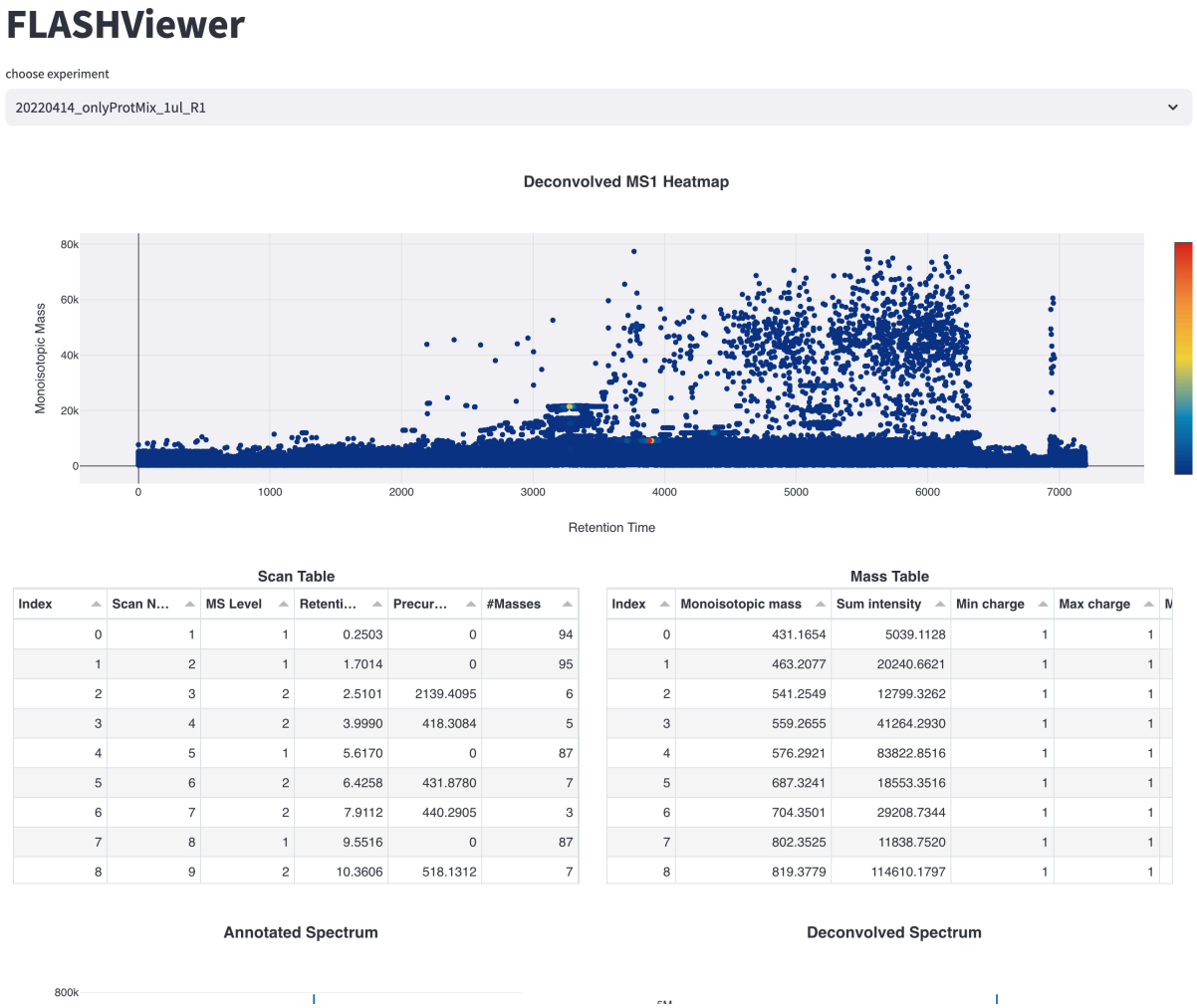


Figure 5.6: An example screenshot of the Viewer page. The input file was one of the replicates of the PIPMix dataset from Chapter 4. Three components out of six components are shown here: an MS1 deconvolved heatmap, a scan table, and a mass table. An MS2 scan was selected in the scan table, but capturing the selected scan row in the screenshot was impossible.

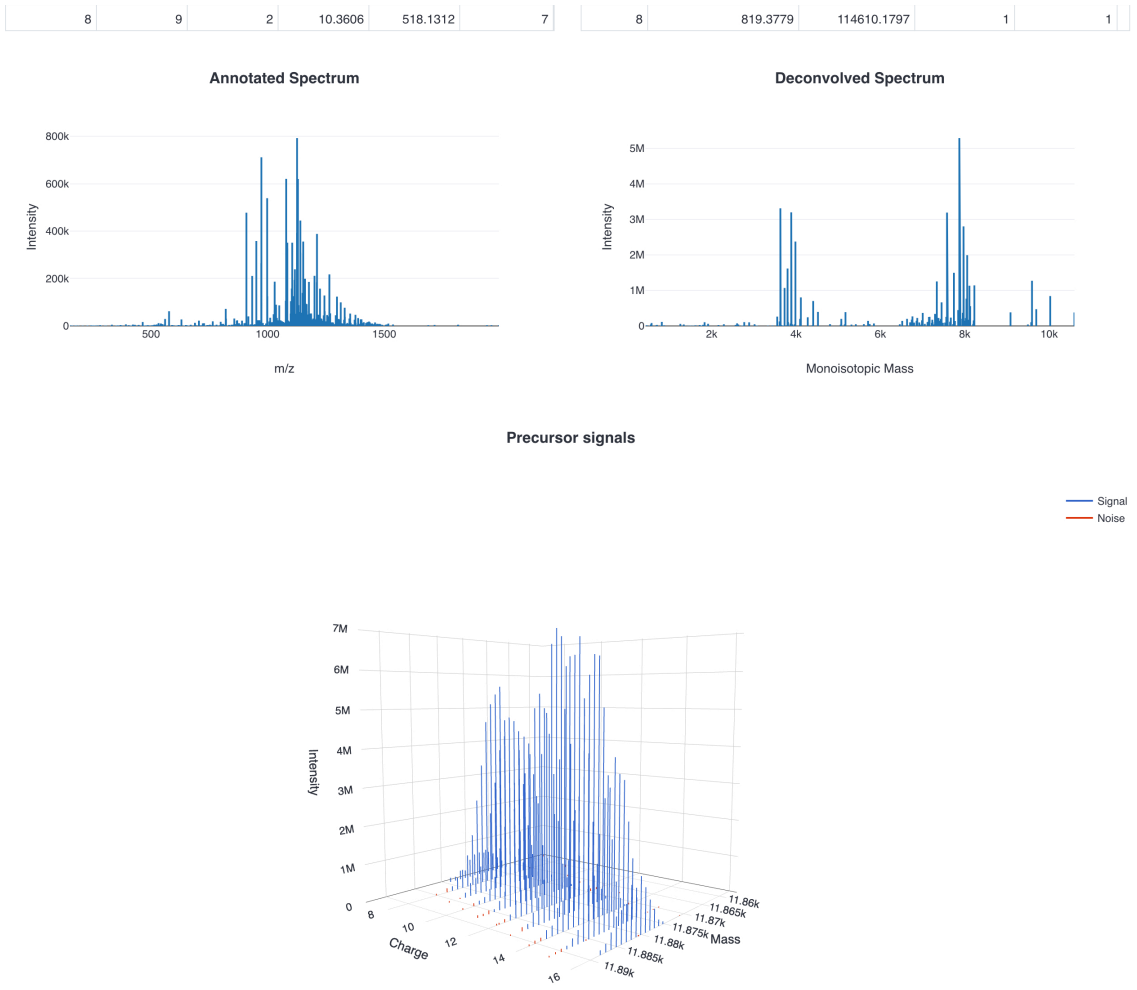


Figure 5.6 (cont.)**:** An example screenshot of the Viewer page: an annotated spectrum, a deconvolved spectrum, and a 3D signal plot. Upon selecting the MS2 scan in the scan table, the mass table and the deconvolved spectrum display fragment masses, while corresponding *m/z* values of the masses were depicted in the annotated spectrum. At the moment, no mass was selected in the mass table, so the precursor signal of the MS2 scan is shown in the 3D signal plot.

masses for a selected MS2 scan. For each mass, the mass table delivers information on charge and isotope ranges, summed intensity, and scores. The deconvolved spectrum plot visualizes these masses, while the annotated spectrum plot depicts the associated m/z values.

The 3D signal plot offers an advanced view of each mass in the charge, mass (i.e., deconvolved isotopologues), and intensity dimension. When a proteoform mass is selected in the mass table of an MS1 scan, it displays the mass's signal and noise peaks in blue and red color, respectively. In the case of an MS2 scan selection, it shows the signal and noise peaks of the corresponding precursor mass. Further, as a fragment mass is selected in the mass table, it narrows the focus to display only the relevant peaks of that specific mass.

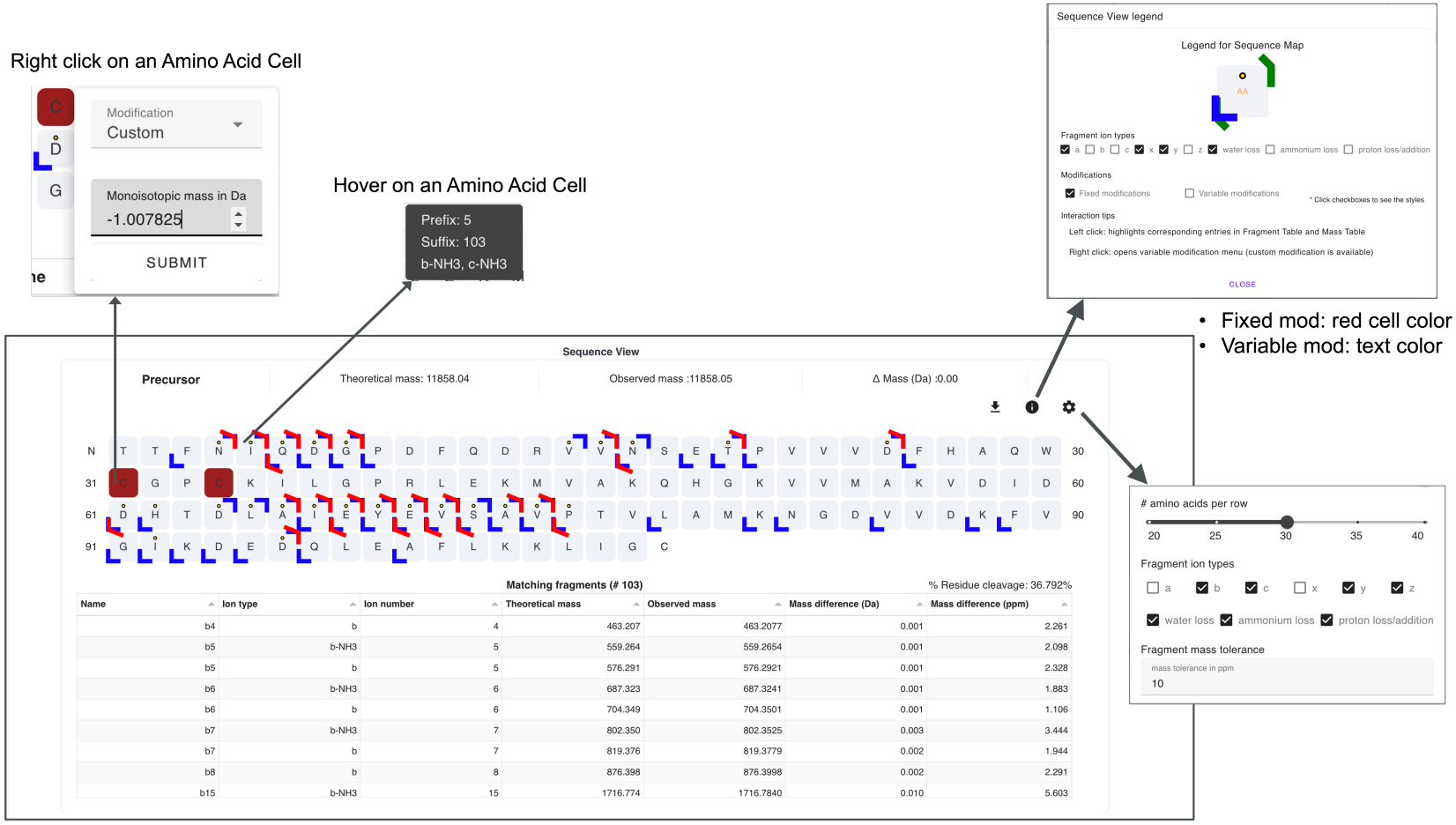
The heatmaps function as independent components, neither affecting nor being affected by other components. In contrast, the scan table component influences all other components within the viewer. By clicking on a row in the scan table, the selected scan can update the mass table, annotated spectrum, deconvolved spectrum, precursor signal 3D plot, sequence view, and internal fragment map. In the case of the mass table, when one row is selected, the corresponding mass can trigger the 3D plot to display its signal and noise peaks. This interactivity only happens when related components are configured in the layout. To maintain the modularity of the components, we have minimized interactions between them.

All the heatmaps, spectrum plots, and 3D plots support dragging, zooming (also, rotation for the 3D plot), and saving in SVG file format. The tables can be saved as comma-separated values (CSV) files by clicking the table title, and sorting is possible by each column.

Sequence View

The sequence view consists of two panels: the sequence map and the matching fragments table (Fig. 5.7). In the sequence map, the sequence from the Sequence Input page is displayed, along with marks indicating matches. The matching fragments table offers information about the matches, including the ion type and number, the theoretical and observed masses of the fragment, and the mass differences between them in Da or ppm. Note that the observed fragment mass is equivalent to the deconvolved mass from the mass table.

Users can tailor the sequence view as needed by clicking the gear icon in the upper right corner of the sequence map. For example, adjusting fragment ion types and mass tolerance will update both panels accordingly in real time. Clicking the "i" icon next



Name	Ion type	Ion number	Theoretical mass	Observed mass	Mass difference (Da)	Mass difference (ppm)
b4	b	4	463.207	463.2077	0.001	2.261
b5	b-NH3	5	559.264	559.2654	0.001	2.098
b5	b	5	576.291	576.2921	0.001	2.328
b6	b-NH3	6	687.323	687.3241	0.001	1.883
b6	b	6	704.349	704.3501	0.001	1.106
b7	b-NH3	7	802.350	802.3525	0.003	3.444
b7	b	7	819.376	819.3779	0.002	1.944
b8	b	8	876.398	876.3998	0.002	2.291
b15	b-NH3	15	1716.774	1716.7840	0.010	5.603

Figure 5.7: An example screenshot of the sequence view component. It demonstrates how the sequence view is displayed on the Viewer page and the different functionalities of the sequence view. The truncated sequence of the Human Thioredoxin protein is matched to a deconvolved MS2 scan from the same data used in Fig. 5.6. The red-colored cells on the sequence indicate that variable modifications have been added to them. Different fragment types are colored in distinctive colors: green for a/x, blue for b/y, and red for c/z. In this example, the b/y and c/z ion types have been selected (panel on the bottom-right), so the blue and red markers are shown on the sequence. Yellow dots on the residue cell imply that the fragment masses have been matched to the theoretical masses after water loss, ammonium loss, or proton loss/addition. To check this information, users can click on the "i" icon on the top-right corner of the sequence view to open the legend menu (top-right panel on the figure).

to the gear icon will inform users what each mark on the sequence map means. The sequence view supports saving the figure in SVG format (by clicking the download icon next to the "i" icon) and the matching fragments table in CSV format (by clicking the table title). A right-click on a residue cell within the sequence map opens a popup menu, enabling the addition of PTM mass shifts to the theoretical sequence (Fig. 5.7 top right corner). A few pre-defined PTM sets for each amino acid type are available, and users can also include custom mass shifts. By clicking on the residue cell with a marker indicating a matched fragment, the associated rows in the mass table and matching fragment table are highlighted. This helps users to review the detailed information easily.

Internal fragment map

Two visualization options are provided, stacked and overlay presentation types, as shown in Fig. 5.8. To enhance user flexibility, the settings menu allows real-time adjustment of the mass tolerance for matching.

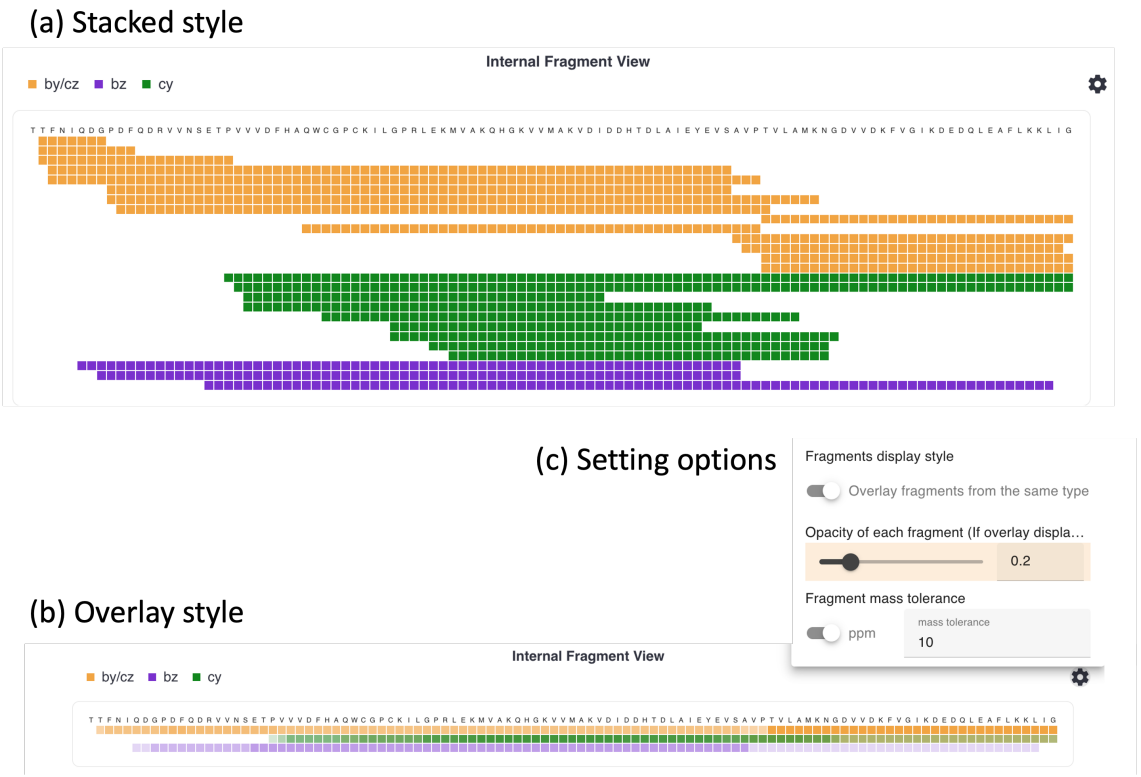


Figure 5.8: An example screenshot of the internal fragment map component. This component has two types of visualizing internal fragments: (a) Stacked and (b) Overlay. Two types can be switched with the toggle button in the settings menu in (c).

5.3.2 Runtime evaluation

FLASHViewer was installed on a Windows desktop with AMD Ryzen 5 5600 (6 core, 12 thread) CPU, AMD Radeon RX 6600 XT GPU, and 16 GB RAM for the runtime evaluation. Firefox 121.0 was used for execution and measuring the runtime. The proteome-level dataset from Chapter 4 was used, and two replicate files of the sample with the largest raw mzML files (5:1 ratio, see Table D.1 for details) were chosen for the evaluation. These files, with an average size of 373.1 MB and 6107 scans (1983 MS1 and 4124 MS2 spectra), were deconvolved with FLASHDeconv. The default parameter was used except for the charge range of 2-50 and mass range of 1-30 kDa. On average, 84,577 masses from MS1 and 177,430 from MS2 were detected. Ultimately, an average size of 39.1 MB deconvolved output file and 283.9 MB annotated output file were used as input for FLASHViewer.

Parsing the two files took less than 14 seconds. The most time-consuming part was preparing and rendering the Viewer page. Given the default Viewer page setting (including six viewer components as shown in Fig. 5.6), it took an average of 40.17 seconds for each replicate sample to fully load the Viewer page after clicking the link to the page. However, without the heatmap component (which contains the most data points), it took only 27.69 seconds. After loading was completed, the Viewer page was fully interactive.

5.4 Discussion

FLASHViewer is an open-source and platform-independent web application for TDP-MS data visualization, equipped with a configurable layout. The modularized visualization components support users in configuring the viewer according to their preferences, with nine components currently implemented. Furthermore, the viewer allows for simultaneous visualization of multiple datasets in a single scroll, facilitating comparative analysis between samples. FLASHViewer demonstrated its scalability by accommodating two tools and can be enhanced through further contributions from the community. Given the FLASHViewer implementation with Python and Vue, no advanced coding skills are expected to contribute.

To enhance FLASHViewer, we plan to incorporate the capability to execute tools directly within the application. As FLASHDeconv and FLASHQuant are parts of OpenMS, their integration into pyOpenMS for execution is feasible. Also, additional parsers are foreseen to be implemented to support other tools, particularly those without dedicated viewer options.

Chapter 6

Conclusion and Outlook

This thesis presents novel computational methods for MS data analysis in the rising field of TDP. Particularly, we aimed to converge the presented methods to enable the quantitative analysis of proteoforms. Our contributions are three-fold.

First, the development of FLASHDeconv, a rapid and robust deconvolution tool, served as a foundational step for subsequent data analysis processes. A novel idea for peak de-charging, that involves peak transformation and universal pattern matching, was incorporated and significantly advanced the runtime. The strength of this tool lies not only in its speed but also in its quality. From the evaluation, FLASHDeconv reported more genuine mass features with fewer artifacts compared to the benchmarking tools. Moreover, it succeeded in processing a broad range of proteoform masses, enabling the analysis of both isotopically resolved and unresolved datasets. FLASHDeconv has already been combined into OpenMS as a TOPP tool since the OpenMS 3.0 release, starting a new TDP branch in OpenMS.

Second, FLASHQuant was proposed as a quantification-dedicated tool for MS1-level comparative analyses. The challenge of overlapping signals in MS, the innate issue in TDP, has often been overlooked. To enhance quantification accuracy, FLASHQuant automatically resolves signal overlaps by employing the theoretical elution profiles of individual proteoforms. Our demonstrations underscored FLASHQuant's high quantification accuracy and reproducibility, extending to proteome-wide analyses, through fold-change comparisons. Manual validation confirmed the tool's capability to distinctly detect and quantify proteoforms with even minor mass differences. Furthermore, its fast runtime, inheriting the efficiency of the FLASHDeconv deconvolution algorithm, ensures rapid data processing. When pairing FLASHQuant with consensus feature group detection, comparative analyses across multiple samples are simplified, offering a comprehensive solution.

Third, a web-based GUI tool named FLASHViewer is provided to visualize results from FLASHDeconv and FLASHQuant. To improve user experience and ensure future scalability, the concept of a configurable layout was implemented in Python and Typescript. Nine interactive visualization components are currently available and can be saved as images or in a file for publication purposes. Notably, FLASHViewer supports the display of multiple samples in a single scroll, facilitating comparative analyses. This versatile tool is accessible both locally and remotely.

This thesis deeply engages with OpenMS, aiming to establish a dedicated TDP branch within such an expansive platform. This initiative yields mutual benefits: enriching TDP with diverse data analysis platforms and drawing a new user base to OpenMS. The success of this effort has already been proven, with the integration of FLASHDeconv into MASH Explorer[87], a comprehensive software environment that accommodates various computational methods focused on TDP.

Data analysis has posed a bottleneck for TDP, with limited options offered, and suffers prolonged runtimes, particularly for proteome-wide experiments. The developments of new software have been hindered by the lack of a gold standard dataset, which serves as a crucial benchmarking point with an expected output. Unfortunately, such a dataset is absent at the proteome level, except for a protein mixture with a small number of targeted proteins. The PIP (Pierce Intact Protein Standard Mix) dataset is mainly employed as a protein mixture, consists of differently-sized proteoforms, and is typically purchased for quality control of MS instruments. This PIP sample was also used in our accuracy evaluation of both FLASHDeconv and FLASHQuant (in Chapters 3 and 4, respectively). To support the forthcoming software development, all generated datasets are shared through the publicly accessible repository, MassIVE. To cope with the absence of a proteome-level gold standard dataset, we attempted to use the BU dataset results as a benchmark in Chapter 3. However, it cannot be a concrete answer sheet, as TDP has distinct strengths compared to BU. Likewise, in Chapter 4, the identified proteoforms from MS/MS were used as benchmarking results; however, these cannot provide a definitive evaluation due to the disparate information carried out by MS1 and MS2 level data. Unidentified proteoforms may result from low fragmentation efficiency, but this does not imply their inaccuracy in quantification at the MS1 level. Evaluating results with inherent ambiguity presents challenges. The silver lining here is the ongoing project by Consortium for Top-Down Proteomics, a non-profit international community for TDP, to generate the gold standard dataset. We anticipate that this initiative will accelerate the development and implementation of TDP-dedicated software.

Bibliography

[1] Lloyd M Smith and Neil L Kelleher. Proteoform: a single term describing protein complexity. *Nature methods*, 10(3):186–187, 2013. 1

[2] Katrina Carbonara, Martin Andonovski, and Jens R Coorssen. Proteomes are of proteoforms: embracing the complexity. *Proteomes*, 9(3):38, 2021. 1

[3] Lloyd M Smith, Jeffrey N Agar, Julia Chamot-Rooke, Paul O Danis, Ying Ge, Joseph A Loo, Ljiljana Paša-Tolić, Yury O Tsybin, Neil L Kelleher, and Consortium for Top-Down Proteomics. The human proteoform project: defining the human proteome. *Science advances*, 7(46):eabk0734, 2021. 1, 3

[4] Frederik Lermyte, Yury O Tsybin, Peter B O'Connor, and Joseph A Loo. Top or middle? up or down? toward a standard lexicon for protein top-down and allied mass spectrometry approaches. *Journal of the American Society for Mass Spectrometry*, 30(7):1149–1157, 2019. 2, 8

[5] Timothy K Toby, Luca Fornelli, and Neil L Kelleher. Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry*, 9:499–519, 2016. 2, 25, 26

[6] Bifan Chen, Kyle A Brown, Ziqing Lin, and Ying Ge. Top-down proteomics: ready for prime time? *Analytical chemistry*, 90(1):110–127, 2017. 8, 25

[7] Kellye A Cupp-Sutton and Si Wu. High-throughput quantitative top-down proteomics. *Molecular omics*, 16(2):91–99, 2020. 2, 4, 8, 19, 20, 41

[8] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, et al. How many human proteoforms are there? *Nature chemical biology*, 14(3):206–214, 2018. 2

[9] Daniel P Donnelly, Catherine M Rawlins, Caroline J DeHart, Luca Fornelli, Luis F Schachner, Ziqing Lin, Jennifer L Lippens, Krishna C Aluri, Richa Sarin, Bifan Chen, et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature methods*, 16(7):587–594, 2019. 3, 9, 25, 26

[10] Jake A Melby, David S Roberts, Eli J Larson, Kyle A Brown, Elizabeth F Bayne, Song Jin, and Ying Ge. Novel strategies to address the challenges in top-down proteomics. *Journal of the American Society for Mass Spectrometry*, 32(6):1278–1294, 2021. 3, 41

[11] Kenneth R Durbin, Luca Fornelli, Ryan T Fellers, Peter F Doubleday, Masashi Narita, and Neil L Kelleher. Quantitation and identification of thousands of human proteoforms below 30 kda. *Journal of proteome research*, 15(3):976–982, 2016. 3

[12] Kyowon Jeong, Maša Babović, Vladimir Gorshkov, Jihyung Kim, Ole N Jensen, and Oliver Kohlbacher. Flashida enables intelligent data acquisition for top-down proteomics to boost proteoform identification counts. *Nature Communications*, 13(1):4407, 2022. 3, 42

[13] Kyle A Brown, Jake A Melby, David S Roberts, and Ying Ge. Top-down proteomics: challenges, innovations, and applications in basic and clinical research. *Expert review of proteomics*, 17(10):719–733, 2020. 8, 9, 41

[14] Zhaorui Zhang, Si Wu, David L Stenoien, and Ljiljana Paša-Tolić. High-throughput proteomics. *Annual review of analytical chemistry*, 7:427–454, 2014. 8

[15] Annette Michalski, Eugen Damoc, Jan-Peter Hauschild, Oliver Lange, Andreas Wieghaus, Alexander Makarov, Nagarjuna Nagaraj, Juergen Cox, Matthias Mann, and Stevan Horning. Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular & cellular proteomics*, 10(9), 2011. 9

[16] 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. 9

[17] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of mass spectrometry*, 40(4):430–443, 2005. 10

[18] Jennifer S Brodbelt. Deciphering combinatorial post-translational modifications by top-down mass spectrometry. *Current Opinion in Chemical Biology*, 70:102180, 2022. 11

[19] Marc Sturm and Oliver Kohlbacher. Toppview: an open-source viewer for mass spectrometry data. *Journal of proteome research*, 8(7):3760–3763, 2009. 13, 23

[20] Joerg Doellinger, Andy Schneider, Marcell Hoeller, and Peter Lasch. Sample preparation by easy extraction and digestion (speed)-a universal, rapid, and detergent-free protocol for proteomics based on acid extraction. *Molecular & Cellular Proteomics*, 19(1):209–222, 2020. 14

[21] Michael W Senko, Steven C Beu, and Fred W McLaffertycor. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995. 15, 32

[22] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, 2000. 16, 42

[23] Leah V Schaffer, Robert J Millikin, Rachel M Miller, Lissa C Anderson, Ryan T Fellers, Ying Ge, Neil L Kelleher, Richard D LeDuc, Xiaowen Liu, Samuel H Payne, et al. Identification and quantification of proteoforms by mass spectrometry. *Proteomics*, 19(10):1800361, 2019. 19, 25, 26

[24] Leonie F Waanders, Stefan Hanke, and Matthias Mann. Top-down quantitation and characterization of silac-labeled proteins. *Journal of the American Society for Mass Spectrometry*, 18(11):2058–2064, 2007. 19, 41

[25] Lucía Geis-Asteggiante, Suzanne Ostrand-Rosenberg, Catherine Fenselau, and Nathan J Edwards. Evaluation of spectral counting for relative quantitation of proteoforms in top-down proteomics. *Analytical chemistry*, 88(22):10900–10907, 2016. 20

[26] James J Pesavento, Craig A Mizzen, and Neil L Kelleher. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone h4. *Analytical chemistry*, 78(13):4271–4280, 2006. 21

[27] Serife Ayaz-Guner, Jiang Zhang, Lin Li, Jeffery W Walker, and Ying Ge. In vivo phosphorylation site mapping in mouse cardiac troponin i by high resolution top-down electron capture dissociation mass spectrometry: Ser22/23 are the only sites basally phosphorylated. *Biochemistry*, 48(34):8161–8170, 2009. 21

[28] Ying Peng, Xin Chen, Han Zhang, Qingge Xu, Timothy A Hacker, and Ying Ge. Top-down targeted proteomics for deep sequencing of tropomyosin isoforms. *Journal of proteome research*, 12(1):187–198, 2013. 21

[29] Charles Ansong, Si Wu, Da Meng, Xiaowen Liu, Heather M Brewer, Brooke L Deatherage Kaiser, Ernesto S Nakayasu, John R Cort, Pavel Pevzner, Richard D Smith, et al. Top-down proteomics reveals a unique protein s-thiolation switch in salmonella typhimurium in response to infection-like conditions. *Proceedings of the National Academy of Sciences*, 110(25):10153–10158, 2013. 21

[30] Si Wu, Joseph N Brown, Nikola Tolić, Da Meng, Xiaowen Liu, Haizhen Zhang, Rui Zhao, Ronald J Moore, Pavel Pevzner, Richard D Smith, et al. Quantitative analysis of human salivary gland-derived intact proteome using top-down mass spectrometry. *Proteomics*, 14(10):1211–1222, 2014. 22, 42

[31] Peter A DiMaggio, Nicolas L Young, Richard C Baliban, Benjamin A Garcia, and Christodoulos A Floudas. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Molecular & Cellular Proteomics*, 8(11):2527–2543, 2009. 22

[32] Matthew V Holt, Tao Wang, and Nicolas L Young. High-throughput quantitative top-down proteomics: histone h4. *Journal of the American Society for Mass Spectrometry*, 30(12):2548–2560, 2019. 22

[33] Ioanna Ntai, Kyunggon Kim, Ryan T Fellers, Owen S Skinner, Archer D Smith IV, Bryan P Early, John P Savaryn, Richard D LeDuc, Paul M Thomas, and Neil L Kelleher. Applying label-free quantitation to top down proteomics. *Analytical chemistry*, 86(10):4961–4968, 2014. 22, 42

[34] Matthew T Mazur, Helene L Cardasis, Daniel S Spellman, Andy Liaw, Nathan A Yates, and Ronald C Hendrickson. Quantitative analysis of intact apolipoproteins in human hdl by top-down differential mass spectrometry. *Proceedings of the National Academy of Sciences*, 107(17):7728–7733, 2010. 22

[35] Jungkap Park, Paul D Piehowski, Christopher Wilkins, Mowei Zhou, Joshua Mendoza, Grant M Fujimoto, Bryson C Gibbons, Jared B Shaw, Yufeng Shen, Anil K Shukla, et al. Informed-proteomics: open-source software package for top-down proteomics. *Nature methods*, 14(9):909–914, 2017. 22, 34, 42

[36] Leah V Schaffer, Jarred W Rensvold, Michael R Shortreed, Anthony J Cesnik, Adam Jochem, Mark Scalf, Brian L Frey, David J Pagliarini, and Lloyd M Smith. Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. *Journal of proteome research*, 17(10):3526–3536, 2018. 22, 34, 38, 39, 42, 103

[37] Ziqing Lin, Liming Wei, Wenxuan Cai, Yanlong Zhu, Trisha Tucholski, Stanford D Mitchell, Wei Guo, Stephen P Ford, Gary M Diffee, and Ying Ge. Simultaneous quantification of protein expression and modifications by top-down targeted proteomics: A case of the sarcomeric subproteome*[s]. *Molecular & Cellular Proteomics*, 18(3):594–605, 2019. 22

[38] Julianus Pfeuffer, Chris Bielow, Samuel Wein, Kyowon Jeong, Eugen Netz, Axel Walter, Oliver Alka, Lars Nilse, Pasquale Domenico Colaianni, Douglas McCloskey, et al. Openms 3 enables reproducible analysis of large-scale mass spectrometry data. *Nature methods*, 21(3):365–367, 2024. 23, 34, 43, 64

[39] Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. Topp—the openms proteomics pipeline. *Bioinformatics*, 23(2):e191–e197, 2007. 23

[40] Hannes L Röst, Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. pyopenms: a python-based interface to the openms mass-spectrometry algorithm library. *Proteomics*, 14(1):74–77, 2014. 23

[41] Eftychia E Kontou, Axel Walter, Oliver Alka, Julianus Pfeuffer, Timo Sachsenberg, Omkar S Mohite, Matin Nuhamunada, Oliver Kohlbacher, and Tilmann Weber. Umetaflow: an untargeted metabolomics workflow for high-throughput data processing and analysis. *Journal of Cheminformatics*, 15(1):52, 2023. 23, 64

[42] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007. 23

[43] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017. 23

[44] Kyowon Jeong, Jihyung Kim, Manasi Gaikwad, Siti Nurul Hidayah, Laura Heikaus, Hartmut Schlüter, and Oliver Kohlbacher. Flashdeconv: ultrafast, high-quality feature deconvolution for top-down proteomics. *Cell Systems*, 10(2):213–218, 2020. 25, 43, 44, 64

[45] Aneika C Leney and Albert JR Heck. Native mass spectrometry: what is in the name? *Journal of the American Society for Mass Spectrometry*, 28(1):5–13, 2017. 25, 26

[46] Huilin Li, Hong Hanh Nguyen, Rachel R Ogorzalek Loo, Iain DG Campuzano, and Joseph A Loo. An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. *Nature chemistry*, 10(2):139–148, 2018. 40

[47] Owen S Skinner, Nicole A Haverland, Luca Fornelli, Rafael D Melani, Luis HF Do Vale, Henrique S Seckler, Peter F Doubleday, Luis F Schachner, Kristina Srzentić, Neil L Kelleher, et al. Top-down characterization of endogenous protein complexes with native proteomics. *Nature chemical biology*, 14(1):36–41, 2018. 25

[48] Lloyd M Smith and Neil L Kelleher. Proteoforms as the next proteomics currency. *Science*, 359(6380):1106–1107, 2018. 25

[49] Wenxuan Cai, Huseyin Guner, Zachery R Gregorich, Albert J Chen, Serife Ayaz-Guner, Ying Peng, Santosh G Valeja, Xiaowen Liu, and Ying Ge. Mash suite pro: a comprehensive software tool for top-down proteomics. *Molecular & Cellular Proteomics*, 15(2):703–714, 2016. 25

[50] Michael T Marty, Andrew J Baldwin, Erik G Marklund, Georg KA Hochberg, Justin LP Benesch, and Carol V Robinson. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical chemistry*, 87(8):4370–4376, 2015. 25, 26

[51] Marshall Bern, Tomislav Caval, Yong J Kil, Wilfred Tang, Christopher Becker, Eric Carlson, Doron Kletter, K Ilker Sen, Nicolas Galy, Dominique Hagemans, et al. Parsimonious charge deconvolution for native mass spectrometry. *Journal of proteome research*, 17(3):1216–1226, 2018. 25, 26, 27, 35, 102

[52] Anthony J Cesnik, Michael R Shortreed, Leah V Schaffer, Rachel A Knoener, Brian L Frey, Mark Scalf, Stefan K Solntsev, Yunxiang Dai, Audrey P Gasch, and Lloyd M Smith. Proteoform suite: Software for constructing, quantifying, and visualizing proteoform families. *Journal of proteome research*, 17(1):568–578, 2018. 26, 63

[53] Philip D Compton, Leonid Zamdborg, Paul M Thomas, and Neil L Kelleher. On the scalability and requirements of whole protein mass spectrometry. *Analytical chemistry*, 83(17):6868–6874, 2011. 26, 27

[54] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, et al. mzml—a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10(1), 2011. 27

[55] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008. 27

[56] Erhan Kenar, Holger Franken, Sara Forcisi, Kilian Wörmann, Hans-Ulrich Häring, Rainer Lehmann, Philippe Schmitt-Kopplin, Andreas Zell, and Oliver Kohlbacher. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Molecular & cellular proteomics*, 13(1):348–359, 2014. 29, 33, 43

[57] Richard A Caruana, Roger B Searle, Thomas Heller, and Saul I Shupack. Fast algorithm for the resolution of spectra. *Analytical chemistry*, 58(6):1162–1167, 1986. 33

[58] Vlad Zabrouskov, Michael W Senko, Yi Du, Richard D Leduc, and Neil L Kelleher. New and automated msn approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry*, 16(12):2027–2038, 2005. 34

[59] Ioanna Ntai, Richard D LeDuc, Ryan T Fellers, Petra Erdmann-Gilmore, Sherri R Davies, Jeanne Rumsey, Bryan P Early, Paul M Thomas, Shunqiang Li, Philip D Compton, et al. Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Molecular & Cellular Proteomics*, 15(1):45–56, 2016. 40

[60] Qiang Kou, Likun Xun, and Xiaowen Liu. Toppic: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, 32(22):3495–3497, 2016. 40, 42, 63

[61] Zachery R Gregorich, Ying Peng, Wenxuan Cai, Yutong Jin, Liming Wei, Albert J Chen, Susan H McKiernan, Judd M Aiken, Richard L Moss, Gary M Diffee, et al. Top-down targeted proteomics reveals decrease in myosin regulatory light-chain phosphorylation that contributes to sarcopenic muscle dysfunction. *Journal of proteome research*, 15(8):2706–2716, 2016. 40

[62] Henrique dos Santos Seckler, Luca Fornelli, R Kannan Mutharasan, C Shad Thaxton, Ryan Fellers, Martha Daviglus, Allan Sniderman, Daniel Rader, Neil L Kelleher, Donald M Lloyd-Jones, et al. A targeted, differential top-down proteomic methodology for comparison of apoa-i proteoforms in individuals with high and low hdl efflux capacity. *Journal of proteome research*, 17(6):2156–2164, 2018. 40

[63] Jihyung Kim, Kyowon Jeong, Philipp T Kaulich, Konrad Winkels, Andreas Tholey, and Oliver Kohlbacher. Flashquant: a fast algorithm for proteoform quantification in top-down proteomics. *Analytical Chemistry*, 2024. 41

[64] Zachery R Gregorich and Ying Ge. Top-down proteomics in health and disease: Challenges and opportunities. *Proteomics*, 14(10):1195–1210, 2014. 41

[65] Jessica L Nickerson, Venus Baghalabadi, Subin RCK Rajendran, Philip J Jakubec, Hammam Said, Teresa S McMillen, Ziheng Dang, and Alan A Doucette. Recent advances in top-down proteome sample processing ahead of ms analysis. *Mass Spectrometry Reviews*, 42(2):457–495, 2023. 41

[66] Jake T Kline, Michael W Belford, Jingjing Huang, Joseph B Greer, David Bergen, Ryan T Fellers, Sylvester M Greer, David M Horn, Vlad Zabrouskov, Romain Huguet, et al. Improved label-free quantification of intact proteoforms using field asymmetric ion mobility spectrometry. *Analytical Chemistry*, 2023. 41, 42

[67] Jan Leipert, Philipp T Kaulich, Max K Steinbach, Britta Steer, Konrad Winkels, Christine Blurton, Matthias Leippe, and Andreas Tholey. Digital microfluidics and magnetic bead-based intact proteoform elution for quantitative top-down nanoproteomics of single c. elegans nematodes. *Angewandte Chemie International Edition*, 62(28):e202301969, 2023. 41, 42

[68] Pavel Bouchal, Monika Dvorakova, Alexander Scherl, Spiros D Garbis, Rudolf Nenutil, and Borivoj Vojtesek. Intact protein profiling in breast cancer biomarker discovery: Protein identification issue and the solutions based on 3 d protein separation, bottom-up and top-down mass spectrometry. *Proteomics*, 13(7):1053–1058, 2013. 41

[69] Jiang Zhang, Moltu J Guy, Holly S Norman, Yi-Chen Chen, Qingge Xu, Xintong Dong, Huseyin Guner, Sijian Wang, Takushi Kohmoto, Ken H Young, et al. Top-down quantitative proteomics identified phosphorylation of cardiac troponin i as a candidate biomarker for chronic heart failure. *Journal of proteome research*, 10(9):4054–4065, 2011. 41

[70] Timothy S Collier, Prasenjit Sarkar, Balaji Rao, and David C Muddiman. Quantitative top-down proteomics of silac labeled human embryonic stem cells. *Journal of the American Society for Mass Spectrometry*, 21(6):879–889, 2011. 41

[71] Timothy W Rhoads, Christopher M Rose, Derek J Bailey, Nicholas M Riley, Rosalynn C Molden, Amelia J Nestler, Anna E Merrill, Lloyd M Smith, Alexander S Hebert, Michael S Westphall, et al. Neutron-encoded mass signatures for quantitative top-down proteomics. *Analytical chemistry*, 86(5):2314–2319, 2014. 41

[72] Konrad Winkels, Tomas Koudelka, and Andreas Tholey. Quantitative top-down proteomics by isobaric labeling with thiol-directed tandem mass tags. *Journal of Proteome Research*, 20(9):4495–4506, 2021. 42

[73] Dahang Yu, Zhe Wang, Kellye A Cupp-Sutton, Yanting Guo, Qiang Kou, Kenneth Smith, Xiaowen Liu, and Si Wu. Quantitative top-down proteomics in complex samples using protein-level tandem mass tag labeling. *Journal of the American Society for Mass Spectrometry*, 32(6):1336–1344, 2021. 42

[74] Donatien Lefebvre, François Fenaille, Déborah Merda, Kevin Blanco-Valle, Cécile Feraudet-Tarisse, Stéphanie Simon, Jacques-Antoine Hennekinne, Yacine Nia, and François Becher. Top-down mass spectrometry for trace level quantification of staphylococcal enterotoxin a variants. *Journal of Proteome Research*, 21(2):547–556, 2022. 42

[75] Wenxuan Cai, Jianhua Zhang, Willem J de Lange, Zachery R Gregorich, Hannah Karp, Emily T Farrell, Stanford D Mitchell, Trisha Tucholski, Ziqing Lin, Mitch Biermann, et al. An unbiased proteomics method to assess the maturation of human pluripotent stem cell-derived cardiomyocytes. *Circulation research*, 125(11):936–953, 2019. 42

[76] Jake A Melby, Willem J de Lange, Jianhua Zhang, David S Roberts, Stanford D Mitchell, Trisha Tucholski, Gina Kim, Andreas Kyrvasilis, Sean J McIlwain, Timothy J Kamp, et al. Functionally integrated top-down proteomics for standardized assessment of human induced pluripotent stem cell-derived engineered cardiac tissues. *Journal of proteome research*, 20(2):1424–1433, 2021.

[77] Trisha Tucholski, Wenxuan Cai, Zachery R Gregorich, Elizabeth F Bayne, Stanford D Mitchell, Sean J McIlwain, Willem J de Lange, Max Wrobbel, Hannah Karp, Zachary Hite, et al. Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics. *Proceedings of the National Academy of Sciences*, 117(40):24691–24700, 2020. 42

[78] Rachele A Lubeckyj, Abdul Rehman Basharat, Xiaojing Shen, Xiaowen Liu, and Liangliang Sun. Large-scale qualitative and quantitative top-down proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with nanograms of proteome samples. *Journal of The American Society for Mass Spectrometry*, 30(8):1435–1445, 2019. 42

[79] Elijah N McCool, Tian Xu, Wenrong Chen, Nicole C Beller, Scott M Nolan, Amanda B Hummon, Xiaowen Liu, and Liangliang Sun. Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. *Science Advances*, 8(51):eabq6348, 2022. 42

[80] Mowei Zhou, Naomi Uwugiaren, Sarah M Williams, Ronald J Moore, Rui Zhao, David Goodlett, Irena Dapic, Ljiljana Pasa-Tolic, and Ying Zhu. Sensitive top-down proteomics analysis of a low number of mammalian cells using a nanodroplet sample processing platform. *Analytical chemistry*, 92(10):7087–7095, 2020. 42

[81] Abdul Rehman Basharat, Yong Zang, Liangliang Sun, and Xiaowen Liu. Topfd: A proteoform feature detection tool for top–down proteomics. *Analytical Chemistry*, 2023. 42, 43

[82] Johannes Veit. *Efficient Workflows for Analyzing High-Performance Liquid Chromatography Mass Spectrometry-Based Proteomics Data*. PhD thesis, Universität Tübingen, 2019. 46

[83] Yeon Choi, Kyowon Jeong, Sanghee Shin, Joon Won Lee, Young-suk Lee, Sangtae Kim, Sun Ah Kim, Jaehun Jung, Kwang Pyo Kim, V Narry Kim, et al. Ms1-level proteome quantification platform allowing maximally increased multiplexity for silac and in vitro chemical labeling. *Analytical chemistry*, 92(7):4980–4989, 2020. 46

[84] Kevin Lan and James W Jorgenson. A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. *Journal of Chromatography A*, 915(1-2):1–13, 2001. 46

[85] Hannes L Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, Ben C Collins, Johan Malmström, Lars Malmström, et al. Openswath enables automated, targeted analysis of data-independent acquisition ms data. *Nature biotechnology*, 32(3):219–223, 2014. 48

[86] Ryan T Fellers, Joseph B Greer, Bryan P Early, Xiang Yu, Richard D LeDuc, Neil L Kelleher, and Paul M Thomas. Prosight lite: graphical software to analyze top-down mass spectrometry data. *Proteomics*, 15(7):1235–1238, 2015. 63, 68

[87] Zhijie Wu, David S Roberts, Jake A Melby, Kent Wenger, Molly Wetzel, Yiwen Gu, Sudharshanan Govindaraj Ramanathan, Elizabeth F Bayne, Xiaowen Liu, Ruixiang Sun, et al. Mash

explorer: a universal software environment for top-down proteomics. *Journal of proteome research*, 19(9):3867–3876, 2020. 63, 80

[88] In Kwon Choi, Tianze Jiang, Sreekanth Reddy Kankara, Si Wu, and Xiaowen Liu. Topmsv: A web-based tool for top-down mass spectrometry data visualization. *Journal of the American Society for Mass Spectrometry*, 32(6):1312–1318, 2021. 63

[89] Mathias Walzer, Kyowon Jeong, David L. Tabb, and Juan Antonio Vizcaíno. Topdownapp: An open and modular platform for analysis and visualisation of top-down proteomics data. *PROTEOMICS*, n/a(n/a):2200403, 2023. 63

[90] JM Nápoles-Duarte, Avratanu Biswas, Mitchell I Parker, JP Palomares-Baez, MA Chávez-Rojo, and LM Rodríguez-Valdez. Stmol: A component for building interactive molecular visualizations within streamlit web-applications. *Frontiers in Molecular Biosciences*, 9:990846, 2022. 64

[91] Chanin Nantasenamat, Avratanu Biswas, JM Nápoles-Duarte, Mitchell I Parker, and Roland L Dunbrack Jr. Building bioinformatics web applications with streamlit. In *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development*, pages 679–699. Elsevier, 2023. 64

[92] Axel Walter and developers. Openms streamlit template. `https://github.com/OpenMS/streamlit-template`, 2023. Accessed: 2023-12-26. 64

[93] Kenneth R Durbin, Owen S Skinner, Ryan T Fellers, and Neil L Kelleher. Analyzing internal fragmentation of electrosprayed ubiquitin ions during beam-type collisional dissociation. *Journal of the American Society for Mass Spectrometry*, 26(5):782–787, 2015. 69

[94] Carter Lantz, Muhammad A Zenaidee, Benqian Wei, Zachary Hemminger, Rachel R Ogorzalek Loo, and Joseph A Loo. Clipsms: an algorithm for analyzing internal fragments resulting from top-down mass spectrometry. *Journal of proteome research*, 20(4):1928–1935, 2021. 69

[95] George A Khoury, Richard C Baliban, and Christodoulos A Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports*, 1(1):90, 2011. 103

[96] Christian Treitz, Brice Enjalbert, Jean-Charles Portais, Fabien Letisse, and Andreas Tholey. Differential quantitative proteome analysis of escherichia coli grown on acetate versus glucose. *Proteomics*, 16(21):2742–2746, 2016. 109

[97] DM Wessel and UI Flügge. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical biochemistry*, 138(1):141–143, 1984. 109

[98] Liam Cassidy, Andreas O Helbig, Philipp T Kaulich, Kathrin Weidenbach, Ruth A Schmitz, and Andreas Tholey. Multidimensional separation schemes enhance the identification and molecular characterization of low molecular weight proteomes and short open reading frame-encoded peptides in top-down proteomics. *Journal of proteomics*, 230:103988, 2021. 110

[99] John C Tran and Alan A Doucette. Multiplexed size separation of intact proteins in solution phase for mass spectrometry. *Analytical chemistry*, 81(15):6201–6209, 2009. 111

Abbreviations

BU bottom-up. 1–3, 8, 12, 15, 20, 22, 25, 38, 40, 80, 103

CSV comma-separated values. 75, 77

CV coefficient of variation. 51, 52, 61, 121

ESI Electrospray ionization. 9

GELFrEE Gel-eluted liquid fraction entrapment electrophoresis. 9, 110, 111

GUI Graphical user interface. 4, 48, 68, 71, 80

JSON JavaScript Object Notation. 65, 71

LC liquid chromatography. 1, 8, 9, 42, 43, 48, 49, 51, 56, 61, 97, 98, 111

LC-MS liquid chromatography coupled to mass spectrometry. 8, 12, 15, 16, 21, 23, 94

LFQ label-free quantification. 4, 20–22, 41–43, 60, 61, 93

MS mass spectrometry. 1–4, 7–10, 12, 16–20, 22, 25–27, 40–43, 48, 49, 51, 56, 61, 63, 64, 78–80, 93, 97, 98, 110, 111

MS/MS tandem mass spectrometry. 11, 19, 40, 52, 56, 68, 80, 111

PTM post-translational modification. 1, 3, 11, 25, 53, 56, 77

RT retention time. 12, 28, 33, 35, 40, 72, 100, 102

SVG Scalable Vector Graphics. 64, 75, 77

TD top-down. 25, 27, 40

TD-MS mass-spectrometry based top-down approach. 25, 27, 38, 40, 93

TDP top-down proteomics. 2–4, 7–9, 11, 12, 15, 20–22, 25, 26, 41–43, 60, 63, 64, 69, 70, 78–80, 93

TIC total ion current. 15, 37, 38, 51, 58

TOPP The OpenMS PiPeline. 23

XIC extracted ion chromatogram. 15, 20, 22

Appendix A

Contributions

All ideas, approaches, and results presented in this work were developed and discussed by myself (JK) and my supervisor Prof. Dr. Oliver Kohlbacher (OK). The following co-workers contributed to the different projects:

- Dr. Kyowon Jeong (KJ)
- Dr. Manasi Gaikwad (MG)
- Dr. Siti Nurul Hidayah (SH)
- Dr. Laura Heikaus (LH)
- Prof. Dr. Hartmut Schlüter (HS)
- Dr. Philipp T. Kaulich (PK)
- Dr. Konrad Winkels (KW)
- Prof. Dr. Andreas Tholey (AT)
- Andrew Almaguer (AA)
- Dr. Axel Wlater (AW)
- Dr. Wonhyeuk Jung (WJ)

Chapter 3: Fast and robust algorithm for deconvolution

OK conceived the idea of the fast deconvolution algorithm for TD-MS datasets. KJ and **JK** developed and implemented the FLASHDeconv algorithm. MG, SH, LH, and HS designed the experiment and performed the sample preparation and data acquisition. OK and HS led the project and provided resources. KJ and **JK** wrote the manuscript with input from all authors. All authors commented on and approved the paper.

Chapter 4: Quantification algorithm for proteoform analysis

OK conceived the idea of the proteoform LFQ algorithm for the TDP MS datasets, and

KJ proposed the conflict resolution method. **JK** and KJ developed and implemented the FLASHQuant algorithm. PK, KW, and AT designed the experiment and performed the sample preparation and LC-MS experiments. All authors analyzed the data. **JK**, KJ, and PK wrote the manuscript with input from all authors, which was read, commented on, and approved by all authors.

Chapter 5: Web application for visualizing proteoform signals

The project was designed by **JK**, KJ, and OK. **JK**, AA, AW, and KJ developed and implemented FLASHViewer. WJ reviewed the FLASHViewer and provided ideas for improvements. The manuscript was initially written by JK and revised by KJ, AA, AW, WJ, and OK.

Appendix B

Publications

Accepted manuscripts

Kim, J. *, Jeong, K.*, Kaulich, P. T., Winkels, K., Tholey, A., Kohlbacher, O. (2024) "FLASHQuant: a fast algorithm for proteoform quantification in top-down proteomics." *Analytical Chemistry*
*Co-first authors

Pfeuffer, J., Bielow, C., Wein, S., Jeong, K., Netz, E., Walter, A., Alka, O., Nilse, L., Colaianni, P. D., McCloskey, D., **Kim, J.**, Rosenberger, G., Bichmann, L., Walzer, M., Veit, J., Boudaud, B., Bernt, M., Patikas, N., Pilz, M., ... Sachsenberg, T. (2024). "OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data." *Nature Methods, 1-3.*

Jeong, K., Kaulich, P. T., Jung, W., **Kim, J.**, Tholey, A., Kohlbacher, O. (2023). "Precursor deconvolution error estimation: The missing puzzle piece in false discovery rate in top-down proteomics." *Proteomics, 2300068.*

Jeong, K., Babović, M., Gorshkov, V., **Kim, J.**, Jensen, O. N., Kohlbacher, O. (2022). "FLASHIda enables intelligent data acquisition for top-down proteomics to boost proteoform identification counts." *Nature Communications, 13(1), 4407.*

Jeong, K., **Kim, J.**, Kohlbacher, O. (2022). "Chapter 11: Mass Deconvolution of Top-Down Mass Spectrometry Datasets by FLASHDeconv." *Proteoform Identification: Methods and Protocols (pp. 145-157). New York, NY: Springer US.*

Jeong, K.*, **Kim, J.** *, Gaikwad, M., Hidayah, S. N., Heikaus, L., Schlüter, H., Kohlbacher, O. (2020). "FLASHDeconv: ultrafast, high-quality feature deconvolution for top-down proteomics." *Cell Systems, 10(2), 213-218.*
*Co-first authors

Manuscripts in preparation

Müller, T., Siraj, A., Walter, A., **Kim, J.** , Wein, S., von Kleist, J., Feroz, A., Pilz, M., Jeong, K., Sing, J., Charkow, J., Röst, H., Sachsenberg, T. "OpenMS WebApps: Building User-Friendly Solutions for MS Analysis."

Appendix C

Supplemental information: FLASHDeconv

C.1 Dataset generation

C.1.1 Cyto (Bovine Cytochrome C) and Fil (Filgrastim) dataset acquisition

For the Cyto dataset, Bovine Cytochrome C (P62894 UniProt accession number) was purchased from Sigma-Aldrich (Darmstadt, Germany). For the Fil dataset, Filgrastim was kindly provided by CinnaGen Co. HPLC-grade H_2O, acetonitrile (ACN), and formic acid (FA) were purchased from Merck (Darmstadt, Germany).

Each protein was dissolved with 0.1% FA in H_2O to the concentration of 0.5 µg/µL. Reversed-phase LC analysis was performed with an ultra-pressure liquid-chromatography (UPLC) system (ACQUITY, Waters, Manchester, UK). 0.1 % FA in H_2O was used as mobile phase A, and 0.1% FA in ACN was used as mobile phase B. 1 µL of the sample was loaded onto a reversed-phase column (monolithic Proswift RP-4 Analytical 1 x 50 mm; Thermo Scientific, Bremen, Germany) and washed for 1 min with 2% mobile phase B. The proteins were eluted at a constant flow rate of 0.1 µL/min with a linear gradient increasing to 25% B in 5 min, 40% B in 5 min and 70% B in 3 min and hold in 70% B for 1 min. The column temperature was set to 30℃.

The eluting proteins were ionized via electrospray ionization and analyzed with Quadrupole-Orbitrap MS (Q-Exactive, Thermo Scientific, Bremen, Germany) in a positive mode with 20.0 eV in-source collision-induced dissociation (CID). Full scan mass spectra were acquired at a mass range of 500 – 4,000 m/z, with a resolution of 70,000 and an Automatic Gain Control(AGC) target of 3,000,000.

C.1.2 PIP (Pierce Intact Protein Standard Mix) dataset acquisition

Pierce Intact Protein Standard Mix containing six standard proteins, namely Protein G, Protein AG, IGF-I LR3, Thioredoxin, Carbonic Anhydrase II, and Exo Klenow, was purchased from Thermo Scientific (Bremen, Germany). The lyophilized powder of the mixture was reconstituted in 100 µL HPLC-grade water to a final concentration of 0.76 µg/µL. LC-MS analysis was performed as mentioned above with 0.1% FA water as mobile phase A, 0.1% FA ACN as mobile phase B, and the column temperature was set to 30℃. 1 µL of the sample was loaded onto a reversed-phase column (monolithic Proswift RP-4 Analytical 1 x 250 mm, Thermo Scientific, Bremen) and washed for 2 min with 5% mobile phase B. The proteins were eluted at a constant flow rate of 0.2 µL/min with a linear gradient increasing to 25% B in 5 min and 45% B in the next 27 min. The MS data was obtained on a Quadrupole-Orbitrap mass spectrometer operated in positive mode with 20.0 eV in-source CID at a mass range of 500 to 3,000 m/z, MS1 only at 17,500 resolution with 5 micro-scans and AGC target of 3,000,000.

C.2 FLASHDeconv Algorithm

C.2.1 Deisotoping algorithm in detail

Given the candidate peak group set, the deisotoping is done in a greedy fashion from the peak group with the smallest mass to the one with the largest mass. Denote Δ=1.0033 Da (the mass difference between ^{13}C and ^{12}C). Suppose we are processing a peak group P of mass m. First, we examine if there exist peak groups of mass $m-\Delta$ or $m+\Delta$, denoted as $P^{-1}$and P^{+1}. If $P^{-1}$and $P^{+1}$are absent, we discard P, because P is unlikely to be a part of an isotope distribution. Otherwise, only if the intensity score of P is higher than those of $P^{-1}$and P^{+1}, P is retained. This is to start the deisotoping algorithm from the highest intensity isotopologue.

If P is retained, then we process for each peak p in P as follows. Assume p has charge q and m/z value t (which gives $qt = m$ within mass tolerance). In the original spectrum S, we search for peaks around the m/z value t and collect peaks that are likely to be isotopologues of the peak p. First, we search in the right direction from t. Denote the first found peak by p_1 and its m/z by x_1. If p_1 is an (heavy) isotopologue of p, its m/z should be $t + n\Delta/q$ for some $n = 0, 1, 2, ...$, within mass tolerance. The n value can be calculated by

$$[(x_1 - t)q/\Delta] \tag{C.1}$$

where [.] denotes rounding to the nearest integer. Denote this n value by n_1. If $n_1 = 1$, then the isotope index of p_1 should be larger than p by 1 (if they are isotopologues). If n_1 is too large (e.g., n_1>2), then p_1 is unlikely to be an isotopologue of p; p_1 is likely to be there by coincidence. Also, if x_1 and $t + n_1\Delta/q$ are not within mass tolerance, p_1 should be excluded. Thus, we collect the peak p_1 if and only if $n_1 \leq 2$ and $t + n_1\Delta/q$ and $x1$ are within mass tolerance. If p_1 is collected, we assign the charge q to p_1 and keep searching for p_2 in the right direction. Otherwise, the right direction search stops. Like p_1, p_2 of m/z x_2 is collected if and only if $n_2 - n_1 \leq 2$ and $t + n_2\Delta/q$ and x_2 are within mass tolerance, where n_2 is given by

$$[(x_2 - t)q/\Delta]. \tag{C.2}$$

If p_2 is collected, we assign the charge q to p_2 and keep searching for p_3 in the right direction. Otherwise, the search stops. If the right direction search stops, we go back to p and search in the left direction. This two-direction processing ensures that light isotopologues with low abundance are included in isotope pattern matching.

The collection process is done per peak in the peak group. On finishing the process, all collected peaks are added to the peak group. This augmented peak group is called the extended peak group. After forming the extended group, we determine the isotope indices of all peaks within and then determine its monoisotopic mass.

To determine the isotopic indices, we take the highest intensity peak in the extended peak group. Denote its mass by M. The (tentative) isotope index of a peak of mass m is calculated by

$$[(m - M)/\Delta]. \tag{C.3}$$

The tentative isotope indices are calculated for all peaks in the extended peak group. Then, many peaks can have negative indices. Thus, we shift all the indices so that the minimum isotope index becomes zero. We found that using the mass of the highest intensity peak results in accurate final masses in particular for isotopically unresolved regions. The isotope indices at this point are, however, still tentative and readjusted further below.

Since we have isotope indices for all peaks, the intensity for each isotope index i can be calculated by summing up peaks' intensities of isotope index i. Denote this aggregated intensity by I_i. If the maximum isotope index in the extended peak group is n, The observed isotope distribution can be written as a vector

$$I := (I_0, I_1, ..., I_n). \tag{C.4}$$

This distribution is now matched against the average isotope distribution. Denote the average isotope distribution by

$$J = (J_0, J_1, ..., J_l), \tag{C.5}$$

where l is the maximum isotope index thereof. We calculate cosine similarity between two vectors I and J. If two vectors have different lengths, we append zeros at the end of the shorter one to make them have the same length (in most cases, $n < l$).

Sometimes, the cosine similarity is higher if we shift the observed isotope distribution vector I. This happens when the isotopologues with low abundances are not detected or noisy peaks are present before the monoisotopic mass. Thus, we give a large isotope index offset δ ranging from $-n$ to n. Then, shift the vector I by δ. The shifted vector is then used to calculate cosine similarity with J as above. The offset δ^* that maximizes the cosine similarity is chosen, and all isotope indices are shifted by the offset δ^* (giving the final isotope indices). If the resulting maximum cosine similarity is less than 0.75 (user-specified), the extended peak group is filtered out. Otherwise, the monoisotopic mass of this extended peak group is determined by $M - i\Delta$, where M is the mass and i is the isotope index of its highest intensity peak. Also, if we denote the mass difference between the average and monoisotopic masses of the average isotope pattern by D, the average mass of the extended peak group is given by $M - i\Delta + D$. The intensity of an extended peak group is given by the summation of all peak intensities within. The extended peak groups are subject to scoring and filtration described in Section 3.2.1.

An extended peak group corresponds to a deconvolved (monoisotopic) mass at the spectrum level and can be represented by a single peak whose intensity equals the extended peak group intensity and position equals the monoisotopic mass. The deconvolved spectrum is composed of peaks representing all extended peak groups.

Lastly, we describe how FLASHDeconv uses information from spectra close in retention time (similar to Xtract and ReSpect). Given a spectrum S at RT $\boldsymbol{rt}$ and an RT window $\boldsymbol{\Delta rt}$ (user-specified), we retain all the average masses of the extended peak groups obtained from spectra at RTs between $\boldsymbol{rt}$ - $\boldsymbol{\Delta rt}$and $\boldsymbol{rt}$. When processing the spectrum S, these masses are used to generate additional peak groups for the candidate peak groups; for these masses, the peak groups are formed and added even if there is no signal peak triplet. The next deisotoping step is equally applied to these additional peak groups.

C.3 Result

C.3.1 Tool versions and parameters

FLASHDeconv

FLASHDeconv was used with default parameters for all datasets. By default, the mass range is set to 1-100 kDa and the charge range is 2-100. The isotope cosine threshold is set to 0.75, and the charge intensity cosine threshold 0.6. Mass tolerance is 10 ppm. The RT window is automatically set to 1% of total gradient time per dataset.

Xtract

We used Xtract in Thermo BioPharma Finder 3.1. "Intact Protein Analysis" experiment type was used for all datasets. We used the processing method "Default SW Xtract" with the following modifications. "Output Mass Range" was set to 1-100 kDa for all datasets. In the case of "Charge Range", 2-100 was used for the simple (Cyto, Fil, and PIP) datasets, but 2-50 for the myoblast dataset, because Xtract often stopped working with the 2-100 range for the myoblast dataset. The "Automatic Sliding Window Parameter Values" option was turned on for the simple datasets. But for the myoblast dataset, the option was deactivated and "Target Avg Spectrum Width" was set to 3 minutes.

ReSpect

We used ReSpect in Thermo BioPharma Finder 3.1. "Intact Protein Analysis experiment type was used for all datasets. We used the processing method "Default SW ReSpect" with the following modifications. The mass ranges (both "Output Mass Range" and "Model Mass Range") were set to 1-100 kDa and "Charge Range" to 2-100 for all datasets. "Target Mass" was set to 50 kDa. "Automatic Sliding Window Parameter Values" option was turned on for the simple datasets. But for the myoblast dataset, the option was deactivated, and "Target Avg Spectrum Width" was set to 3 minutes.

Promex

We used Promex version 1.0.7017. For all datasets, the mass range was set to 1-100 kDa, and the charge range was set to 2-60 (the maximum charge for Promex is 60).

C.3.2 Mass and isotopologue artifact detection

Given a set of features F, to test if a feature f in F with mass m is a mass artifact or isotopologue artifact, we first collected all features in F such that the RT overlap between the feature and f exceeds 80% of the RT span of f. Secondly, the features having lower intensity than f are filtered out. Denote the remaining feature set by F_f. To determine, for example, if f is a low harmonic artifact, we tested if there exists a feature in F_f that has a mass of qm for $q = 2, ..., 100$ (within 10 ppm tolerance) allowing up to 10 isotope error. If such a feature exists, we declared f as a low harmonic artifact feature. More formally, f is a low harmonic artifact if there exists a feature in F_f whose mass equals $q(m+k\Delta)$ within 10 ppm tolerance for $q = 2, ..., 100$ and $k = -10, ..., 10$, where Δ denotes the mass difference between ^{13}C and ^{12}C. For high harmonic artifacts, the mass term $(m+k\Delta)/q$ was used in the place of $q(m+k\Delta)$.

Charge-off-by-one artifact features[51] occur when a charge $q+1$ or $q-1$ is assigned to a charge q peak. They mainly arise when the charge assignment is based on the peak m/z distance between isotopologues because the measurement of the distance for highly charged peaks is often inaccurate, not to mention the isotopically unresolved peaks. But even when the charge assignment is based on the charge pattern, the artifact can be generated if two peaks with consecutive charge states are close to each other (e.g., below 4 Th), due to peak m/z variations. Thus, to detect the charge-off-by-one artifacts, we chose the charge states q^* (between 2 and 100) such that

$$m/q^* - m/(q^*+1) < 4.$$

For all such q^*, we tested if features having masses

$$(m+k\Delta)^*(q^*+1)/q^*$$

or

$$(m+k\Delta)^*(q^*-1)/q^*$$

are present in F_f (10 ppm mass tolerance allowed) for $k = -10, ..., 10$. If present, we declared f to be a charge-off-by-one artifact.

Lastly, to see if f has other features corresponding to isotopologue artifacts, we simply checked if features having masses $m-\Delta$ or $m+\Delta$, within 10 ppm tolerance, are present in F_f. If present, f is considered to be one of the isotopologues. We repeat the above procedure for each feature in the list F.

C.3.3 Calculation of the ion current

The ion current in Fig. 3.4C was drawn by plotting the total intensity of each spectrum along the retention time. The summed intensities from FLASHDeconv, Xtract, ReSpect, and Promex were drawn as follows. Each tool outputs feature intensities and feature apex retention times (Promex does not output the apex retention time, but it can be calculated with other information). We binned the retention time with a bin size of one minute. For each bin, all features whose apex retention time is within the bin are collected and their intensities are aggregated. The figure was drawn by plotting the aggregated intensities along the time bins.

The intensities were taken from the SumIntensity column for FLASHDeconv, Sum_ Intensity for XTract and ReSpect, and Abundance for Promex. In FLASHDeconv, the SumIntensity of a feature is the summed intensity of the deconvolved masses (i.e., the extended peak groups) constituting the feature.

C.3.4 Matching feature masses against protein masses identified by bottom-up searches

Feature masses from complex datasets were matched to the protein masses identified by BU search, from the same myoblast sample. To collect the protein masses to be matched, we took the UniProt (downloaded in April 2019) entries identified by BU search in Schaffer et al[36]. The (unmodified) protein masses were then calculated using their amino acid sequences. For the matches with modifications, 18 relatively abundant modification types found by PTMCuration[95] and two protein N-terminal modifications (Met-loss and Met-loss+Acetyl) were considered (see Table C.2). We performed two matchings: one with zero modification and the other with up to one modification per protein.

Matching between a feature mass and the set of all theoretical proteoform masses (subject to the above modification criteria) was done by using the multimap data structure. For each possible proteoform mass, the multimap takes the mass as key and the proteoform(s) having the mass as value. A feature is said to be matched to a proteoform if the key (i.e., mass) of the proteoform is within 10 ppm tolerance from the feature mass, allowing one isotope error.

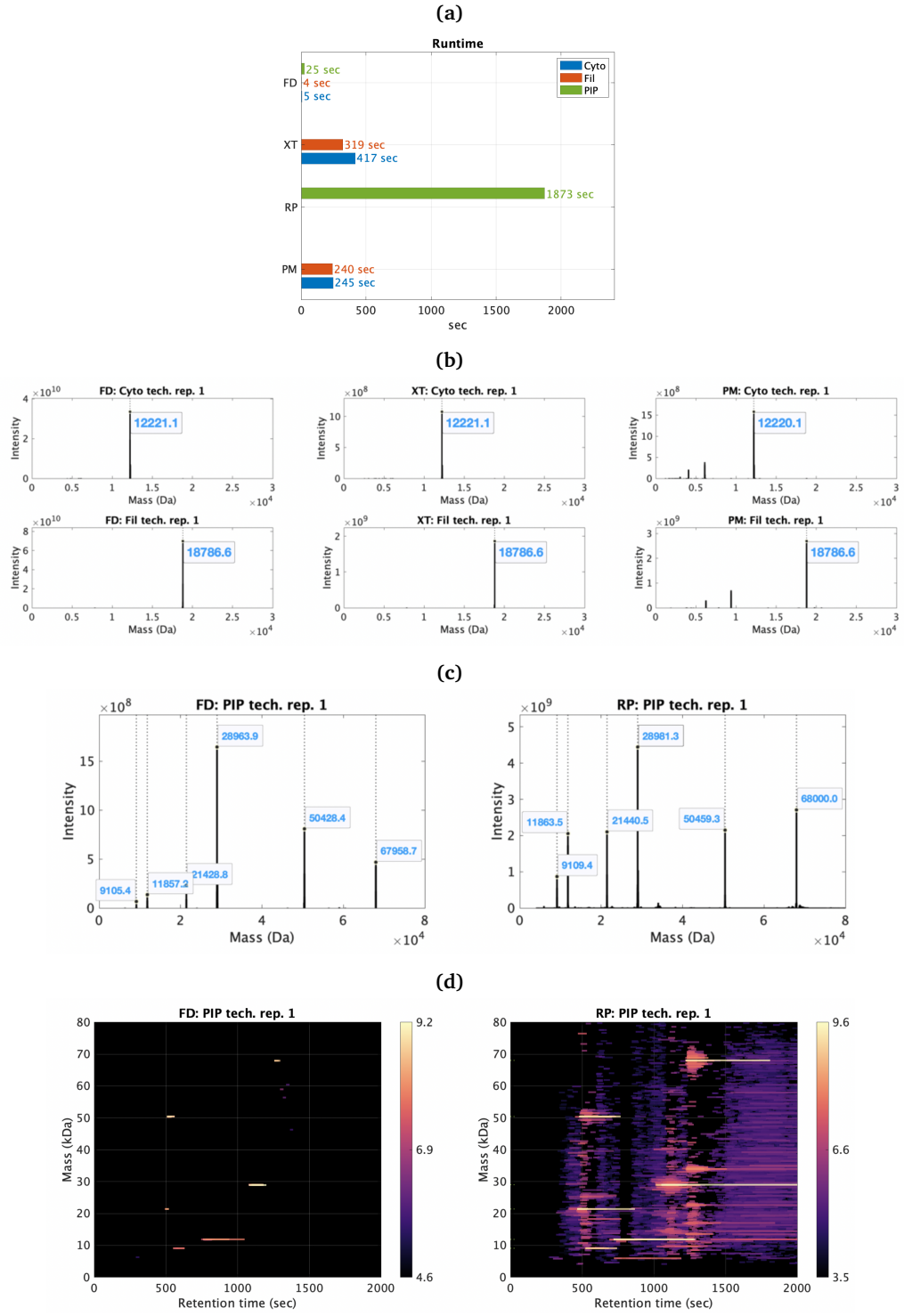


Figure C.1: Analogue of Fig. 3.3 for the simple datasets (Cyto, Fil, and PIP) technical replicate 1. **(a)** Runtime comparison between tools. **(b)** Deconvolved spectra from the Cyto and Fil (isotopically resolved) datasets. **(c)** Deconvolved spectra from the PIP (isotopically unresolved) dataset. **(d)** Features maps from the PIP dataset.

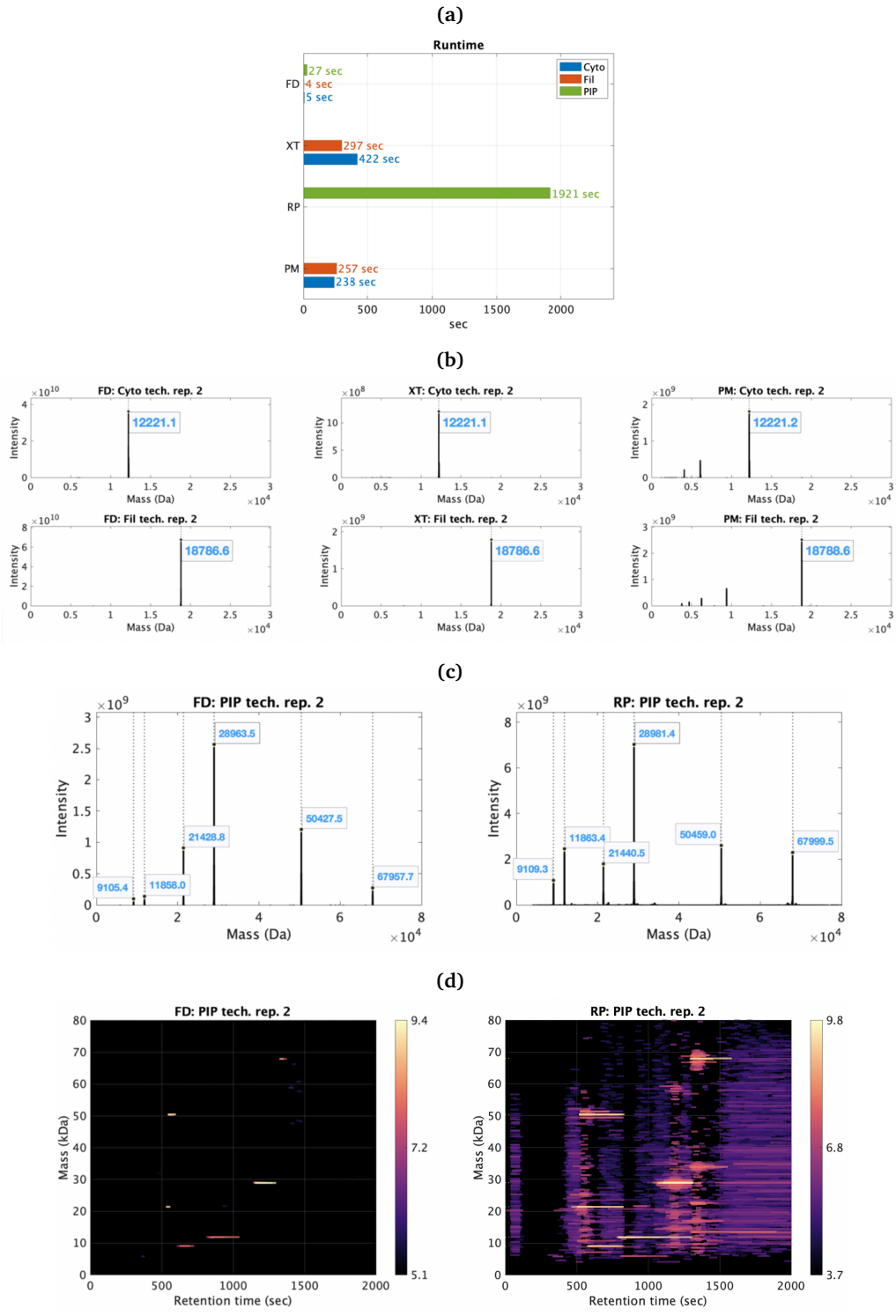


Figure C.2: Analogue of Fig. 3.3 for the simple datasets (Cyto, Fil, and PIP) technical replicate 2.

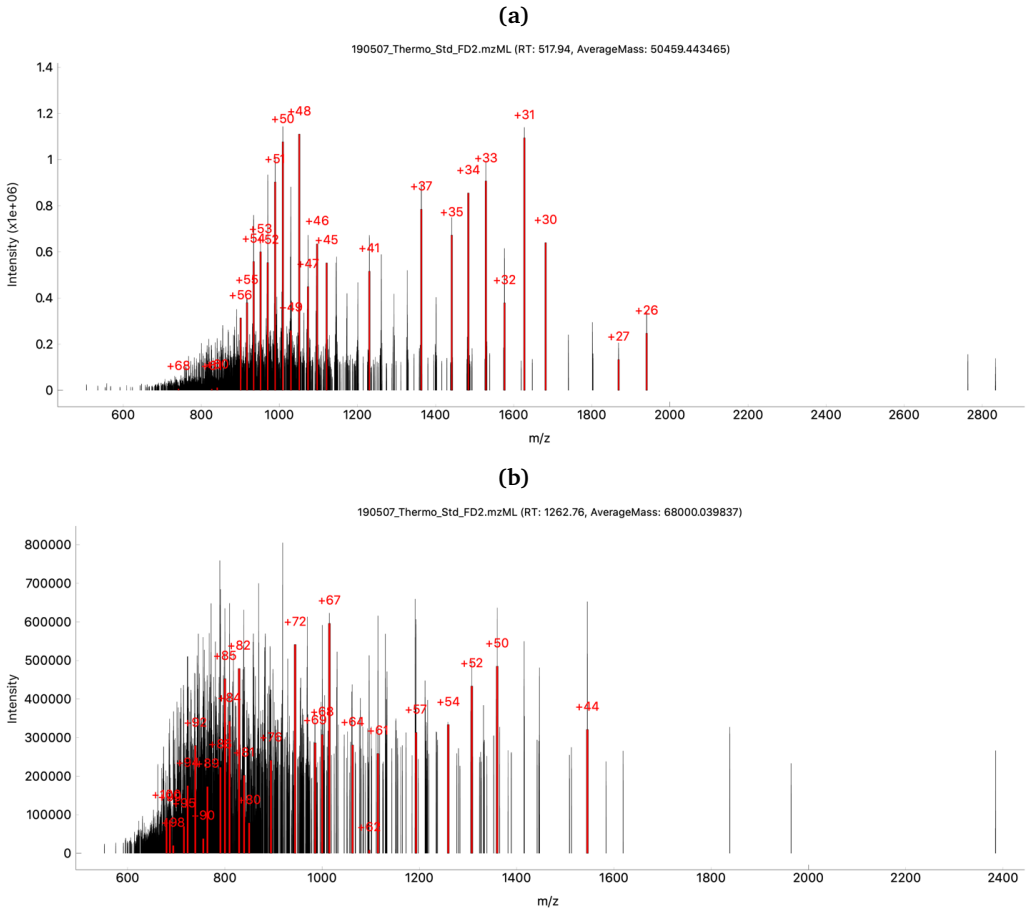


Figure C.3: Annotated spectrum examples for high-mass proteins (50,429.8, and 67,959.4 Da). The annotated peaks (red-colored peaks) from two high-mass proteins in the PIP dataset. The red numbers specify determined charge states. **(a)** The spectrum for the protein of mass 50,429.8 Da. The reported mass by FLASHDeconv was 50427.9 Da. **(b)** The spectrum for 67,959.4 Da. The reported mass by FLASHDeconv was 67957.6 Da.

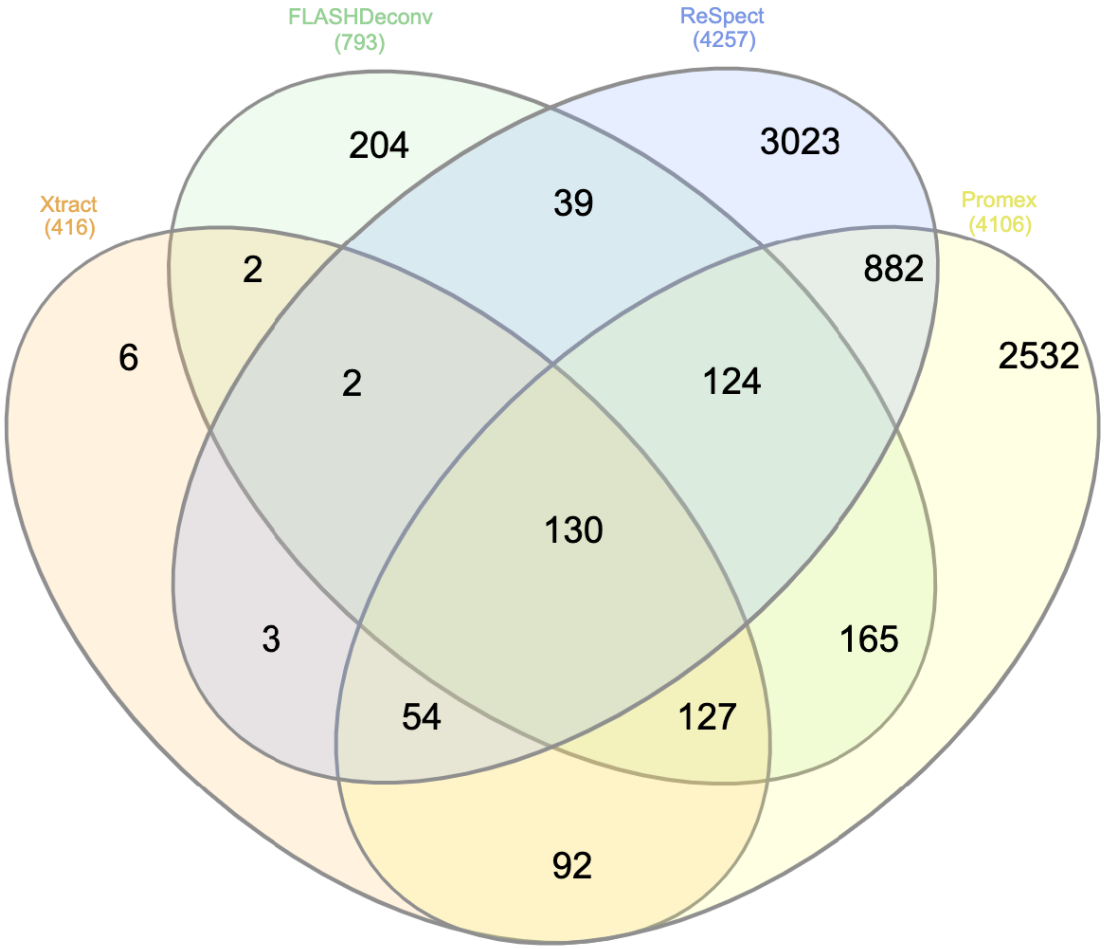


Figure C.4: Venn diagram for the overlap of the rounded monoisotopic masses between tools in the low mass region for the myoblast data set, Related to Fig. 3.4. For each tool, the features of mass less than 20 kDa and of RT 40-80 min were collected. For ReSpect, monoisotopic masses were calculated from average masses using the averagine model. About 75% of FLASHDeconv masses overlapped with the other tools. The Venn diagram was drawn using InteractiVenn (`http://www.interactivenn.net`).

Tool	# features	Unmod	1 PTM
FD	27,307	2,514 (9%)	21,949 (80%)
XT	594	13 (2%)	176 (30%)
RP	20,258	1,208 (6%)	11,870 (59%)
PM	6,653	155 (2%)	1,800 (27%)

Table C.1: This table compares the number/portion of features that are matched to the protein masses from BU searches for the myoblast dataset (also shown in Fig. 3.4F). The second column gives the total number of reported features per tool. The third and fourth columns show the numbers (and portions) of the matched features when no PTM and 1 PTM were allowed, respectively (see C.3.4 for details).

PSI-MS name	Monoisotopic mass
Phospho	79.966331
Acetyl	42.010565
Amidated	-0.984016
Oxidation	15.994915
Methyl	14.015650
LRGG	383.228103
Glu->pyro-Glu	-18.010565
Gln->pyro-Glu	-17.026549
Carboxy	43.989829
Palmitoyl	238.229666
Myristoyl	210.198366
ADP-Ribosyl	541.061110
Farnesyl	204.187801
Nitrosyl	28.990164
GeranylGeranyl	272.250401
Formyl	27.994915
Deamidated	0.984016
Sulfo+amino (Interim name)	94.967714
Met-loss	-131.040485
Met-loss+Acetyl	-89.029920

Table C.2: The list of modifications considered in the feature-BU identified protein mass matching, Related to Fig. 3.4

Appendix D

Supplemental information: FLASHQuant

D.1 Dataset generation

D.1.1 Human Caucasian colon adenocarcinoma cultivation

The cultivation of human Caucasian colon adenocarcinoma (Caco-2) cells was maintained as per the European Collection of Authenticated Cell Cultures recommendation. The cells were grown at 37 °C with 5% CO_2 in RPMI-1640 medium (25 mM HEPES, 2 mM L-glutamine, 13 nM phenol red) supplemented with 10% (v/v) fetal bovine serum and 1% (v/v) penicillin (10,000 U/ml). After reaching 90-100% confluence, the cells were passaged using TrypLE^{TM} Express enzymes to detach the cells. Prior to harvesting, the cells were washed three times with PBS buffer (centrifugation at 200xg for 5 min at 25 °C). The cells were stored at -80 °C prior to cell lysis.

D.1.2 *Escherichia coli* cultivation

Escherichia coli (strain MG1655) was cultured at 37 °C in an M9 minimal medium containing 15 mM glucose, as described previously[96]. In brief, bacterial cells were grown to reach an OD_{600} of 1. The culture was divided into 50 ml aliquots and centrifuged (3,000×g, 5 min, 25 °C) to pellet the cells. The cells were washed twice in MilliQ water and stored at -80 °C prior to cell lysis.

D.1.3 Methanol-Chloroform-Water precipitation

Methanol-chloroform-water precipitation was performed according to Wessel and Flügge[97]. In brief, 150 µl of the sample was mixed with 600 µl methanol, 150 µl

chloroform, and 450 µl MilliQ water. The mixture was centrifuged (14,000×g, 20 min, 25 °C), and the upper phase was removed. 600 µl of methanol was added, mixed thoroughly, and centrifuged. The supernatant was removed, and the protein pellet was washed twice with 600 µl methanol prior to air drying.

D.1.4 Generation of the SpikeIn sample

For the generation of the SpikeIn sample, small *E. coli* proteins (<20 kDa) were enriched by solid-phase extraction, as described previously[98]. In brief, *E. coli* cells were lysed in 8 M guanidium hydrochloride, 1× cOmplete protease inhibitor (Promega) by freeze-thaw cycling. The sample was heated for 10 min at 70 °C and acidified with 5% formic acid (FA). After centrifugation (20 min, 21,100×g, 4 °C), the supernatant was transferred to an activated (methanol) and equilibrated (5% FA) C18 solid-phase extraction cartridge (3 cc 200 mg Waters, Eschborn, Germany). The proteins were washed twice with 5% FA prior to elution of the small proteins with 70% and 100% acetonitrile. The proteins were vacuum-dried and resuspended in MS loading buffer. Protein concentration was determined by the Pierce BCA protein assay kit (Thermo Fisher Scientific).

To obtain the SpikeIn samples, the Pierce™ Intact Protein Standard Mix (Thermo Fisher Scientific, six proteins ranging from 9-68 kDa (human IGF-I LR3 (P05019, 40-118): 9,105.3482 Da; human Thioredoxin (Q99757, 60-166): 11,858.04393 Da; *Streptococcus dysgalactiae* Protein G (P06654, 223-413): 21,429.75915 Da; bovine carbonic anhydrase (P00921, full length): 28,963.6881 Da; *Streptococcus* Protein AG (P02976, P19909, chimeric): 50,429.84641 Da; *Escherichia coli* Exo Klenow (P00582, 324-928): 67,959.42515 Da), resuspended in MS loading buffer) was added to the small protein enriched *E. coli* lysate. The concentration of *E. coli* proteins was kept constant (160 ng/µl), while the final concentration of the protein mixture was varied (20 ng/µl, 10 ng/µl, 6.67 ng/µl, 4 ng/µl, 2.86 ng/µl or 2 ng/µl).

D.1.5 Generation of the ProteomeMix sample

E. coli cells and Caco-2 cells were lysed in 1% sodium dodecyl sulfate, 10 mM TRIS (pH 8.8), and 1× cOmplete protease inhibitor (Promega, Madison, USA) by ultrasonication. Protein concentration was determined by the Pierce BCA protein assay kit (Thermo Fisher Scientific, Bremen, Germany). Approximately 500 µg of proteins were purified by methanol-chloroform-water precipitation and subjected to GELFrEE fractionation (8% tris-acetate cartridge) according to the manufacturer's protocol (Expedeon, Eching, Germany). In brief, the samples were mixed with 30 µl 5× sample buffer, 8 µl 1 M

dithiothreitol, 112 µl MilliQ and incubated 10 min at 50 °C (1,400 rpm) prior to fractionation. To enrich proteoforms smaller than approximately 30 kDa, the first GELFrEE fraction was used from the Caco-2 or *E. coli*sample, respectively[99]. The fractions were purified by chloroform-methanol-water precipitation, and the proteins were dissolved in MS loading buffer (3% acetonitrile, 0.1% trifluoroacetic acid (TFA)). The Caco-2 and *E. coli*samples were analyzed by LC-MS/MS to determine total ion counts (TICs) and were diluted with MS loading buffer to yield approximately the same intensity of TICs.

In order to obtain the proteome mixture, Caco-2 and *E. coli*proteins were mixed in five different ratios (from 1:5 to 5:1), keeping the concentration of *E. coli*proteins constant and varying only the amount of Caco-2 proteins (Table D.1).

Ratio (Caco-2: *E. coli*)	**Caco-2 sample**	***E. coli* sample**	**MilliQ**
5:1	50 µl	10 µl	-
2:1	20 µl	10 µl	30 µl
1:1	10 µl	10 µl	40 µl
0.5:1	5 µl	10 µl	45 µl
0.2:1	2 µl	10 µl	48 µl

Table D.1: Generation of the ProteomeMix. The ProteomeMix data set was generated in five ratios that varied only the amount of Caco-2 protein while keeping the concentration of *E. coli* proteins constant.

D.1.6 LC-MS/MS analysis

Proteoforms were separated using an Ultimate 3000 nano-UHPLC system (Thermo Fisher Scientific) equipped with an analytical reversed-phase C4 column (Accucore, 50 cm×75 µm, 2.6 µm, 150 Å, Thermo Fisher Scientific). A precolumn (C4 PepMap 300, 5 µm, 300 Å, Thermo Fisher Scientific) in forward-flush mode was used for sample loading. Eluent A was 0.05% FA, and eluent B was 80% acetonitrile and 0.04% FA. The separation was performed over a 90 min gradient from 15-55% B: 0-5 min 4% B, 5-7 min 4-15% B, 7-97 min 15-55% B, 97-99 min 55-98% B, 99-110 min 98% B, 110-110.1 min 98-4% B, 110.1-120 min 4% B.

The UHPLC system was coupled online to a Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific). MS1 spectra (400-1800 m/z) were acquired with a resolution of 120,000, 246 ms maximum injection time, 200% normalized AGC target, and 4 microscans. Within a cycle time of 4 s, the most intense ions were selected (charge states: 4-50 and undetermined charge states with dynamic exclusion enabled

(n=2, 60 s) for fragmentation with collision-induced dissociation (25%). The settings for MS2 spectra were 60,000 resolution, 250 ms maximum injection time, and 1,000% normalized AGC target. The acquisition was performed in intact protein mode (ion-routing multipole pressure of 2 mTorr), and lock-mass was enabled (445.12003 m/z). The RF value was set to 30%, and source-induced dissociation was 15 V.

D.2 Result

D.2.1 Tool parameters

We used FLASHQuant with its default parameters, including the m/z tolerance of 10 ppm, the mass tolerance of 3 Da, and the isotope cosine threshold of 0.85. Only mass and charge ranges were modified for each dataset based on the characteristic of the target analyte; the mass range and the charge range were set to 1-70 kDa and 2-100, respectively, for the SpikeIn and the PIPMix datasets, and 1-30 kDa and 2-50 for the ProteomeMix dataset. Since the Pierce Intact Protein Standard Mix from Thermo Fisher includes a protein with a mass of 68 kDa, we decided to broaden the mass and charge ranges for SpikeIn and PIPMix datasets. In runtime analyses, eight threads were given for multi-threading.

The same parameter settings were applied to FLASHDeconv execution. One additional parameter, -max_ms_level, was set to 1 to limit deconvolution at the MS1 level.

ProSightPD (“ProSightPD Hi Res. Feature Detector” node, version 4.2 in Proteome Discoverer 3.0, Thermo Fisher Scientific) was executed with the default "PSPD LFQ for HI HI data processing workflow" parameter settings, except for the mass and charge ranges. The mass and charge ranges were consistent with FLASHQuant. For consensus feature group detection, the default “PSPD 1 percent FDR Consensus workflow” parameter settings were used, including a mass tolerance of 100 ppm and a retention time tolerance of 8 min. To match the eight threads used for FLASHQuant runtime analyses, the “Max. Number of Processing Workflows in Parallel Execution” and Max. Number of Consensus Workflows in Parallel Execution” configuration parameters in Proteome Discoverer were set to four.

For TopFD (from TopPIC Suite 1.7.1 version), the default parameters were used except for the maximum mass and maximum charge, which were set to the equivalent values used for FLASHQuant. As TopFD does not have parameters for minimum mass or minimum charge state, we excluded the results with mass $<$ 1000 Da or Rep_Charge of 1 for all analyses. Same as FLASHQuant, eight threads were allowed for multi-

threading. Unlike FLASHDeconv, TopFD doesn't allow the MS level to be limited; thus, both MS1 and MS/MS spectra were used as input. However, only results from MS1 spectra were used for analyses.

TopPIC (from TopPIC Suite 1.7.1 version) and ProSightPD Search were utilized to analyze MS2 spectra from all datasets and to identify proteoforms' masses. Again, both tools are executed with their default parameters except for the mass and charge ranges, consistent with FLASHQuant.

D.2.2 Significance values comparison for the SpikeIn dataset

In Fig. 4.5A, FLASHQuant outperformed the other tools in quantification accuracy, showing the smallest average fold change differences to the expected value (0.1333), compared to ProSightPD (0.3079), TopFD (0.1426), and FLASHDeconv (0.1399). To statistically assess this outperformance, we performed a paired t-test using all relative fold change values from the SpikeIn dataset analysis, with a sample size of 72 per tool (4 target masses $\times$ 6 experiments $\times$ 3 replicates). The null hypothesis was that FLASHQuant did not outperform the other tools, and the significance level was set to 0.01.

The paired t-test results indicated a significant difference between FLASHQuant and ProSightPD, with a p-value of 0.0000005221. However, as anticipated based on the small average differences between FLASHQuant and TopFD/FLASHDeconv, the differences were not statistically significant. The p-values were 0.1553 for TopFD and 0.1665 for FLASHDeconv, indicating that while the sample averages for TopFD and FLASHDeconv were slightly higher than FLASHQuant, the differences were not substantial enough to reject the null hypothesis.

D.2.3 Identification results for filtering feature groups

With the PIPMix dataset, we aimed to explain all the reported feature groups as much as possible. Thus, we first conducted an open search with TopPIC to gather a possible modification list. In total, 12 modifications were chosen based on the most frequent mass shifts (Disulfide, Dehydro, Didehydro, Amidated, Pro->pyro-Glu, Trp->Oxolactone, Methyl+Deamidated, and Dioxidation) and four common modifications suggested by TopPIC (Acetyl, Phopho, Oxidation, and Methyl). Using these 12 modifications as variable modifications, TopPIC was executed once more to collect identified proteoform masses. The six protein sequences given by Thermo Fisher were taken as a database. For ProSightPD Search, variable modifications needed to be annotated per site on each sequence. Therefore, instead of using the FASTA format database, the XML

format database equipped with a modification list was downloaded from UniProt. As some of the UniProt proteins had a slightly different sequence compared to the protein sequences given by Thermo Fisher, we utilized the "Database Manager" to modify the input sequences manually. The numbers of identified proteoforms are in Table. D.2.

The database for the ProteomeMix dataset was generated by combining the *E. coli*database (Swiss-Prot, taxon ID 83333, downloaded from UniProt in July 2023) with the human database (Swiss-Prot, taxon ID 9606, downloaded from UniProt in July 2023). The dataset and database for ProteomeMix are far larger than PIPMix, so TopPIC could not be run twice. Also, database search with variable modifications was extremely time-consuming (more than a day for one file), such that only an open search was performed. The XML format of the equivalent database (also including modifications) was downloaded from UniProt and employed for ProSightPD Search.

	# Identified proteoforms
ProSightPD Search	62
TopPIC (Rep1)	143
TopPIC (Rep2)	146
TopPIC (Rep3)	164

Table D.2: The number of identified proteoforms by ProSightPD Search and TopPIC for the PIPMix dataset. This table shows the number of identified proteoforms that were used to validate consensus feature groups. The masses of all identified proteoforms (both ProSightPD Search and TopPIC) were matched to the masses of consensus feature groups within 20 ppm mass tolerance.

Files	Default-mode total	Default-mode overlap	Default-mode exclusives	Resolving-off-mode total	Resolving-off-mode overlap	Resolving-off-mode exclusives
SpikeIn_1_R1	1462	1419	43	1641	1419	222
SpikeIn_1_R2	1440	1405	35	1596	1405	191
SpikeIn_1_R3	1393	1357	36	1542	1357	185
SpikeIn_1/2_R1	1464	1419	45	1626	1419	207
SpikeIn_1/2_R2	1431	1400	31	1599	1400	199
SpikeIn_1/2_R3	1378	1339	39	1546	1339	207
SpikeIn_1/3_R1	1444	1406	38	1605	1406	199
SpikeIn_1/3_R2	1401	1364	37	1557	1364	193
SpikeIn_1/3_R3	1366	1333	33	1522	1333	189
SpikeIn_1/5_R1	1381	1353	28	1528	1353	175
SpikeIn_1/5_R2	1347	1315	32	1513	1315	198
SpikeIn_1/5_R3	1331	1296	35	1487	1296	191
SpikeIn_1/7_R1	1330	1302	28	1483	1302	181
SpikeIn_1/7_R2	1315	1287	28	1463	1287	176
SpikeIn_1/7_R3	1296	1268	28	1447	1268	179
SpikeIn_1/10_R1	1322	1289	33	1469	1289	180
SpikeIn_1/10_R2	1262	1231	31	1395	1231	164
SpikeIn_1/10_R3	1269	1243	26	1402	1243	159
ProteomeMix_1/5:1_R1	1221	1193	28	1380	1193	187
ProteomeMix_1/5:1_R2	1302	1259	43	1482	1259	223
ProteomeMix_1/2:1_R1	1344	1295	49	1545	1295	250
ProteomeMix_1/2:1_R2	1272	1229	43	1482	1229	253
ProteomeMix_1:1_R1	1363	1301	62	1589	1301	288
ProteomeMix_1:1_R2	1326	1278	48	1569	1278	291
ProteomeMix_2:1_R1	1431	1365	66	1652	1365	287
ProteomeMix_2:1_R2	1337	1284	53	1591	1284	307
ProteomeMix_5:1_R1	1545	1472	73	1800	1472	328
ProteomeMix_5:1_R2	1474	1427	47	1775	1427	348

Table D.3: The number of detected feature groups reported by FLASHQuant default-mode and resolving-off-mode and their overlaps

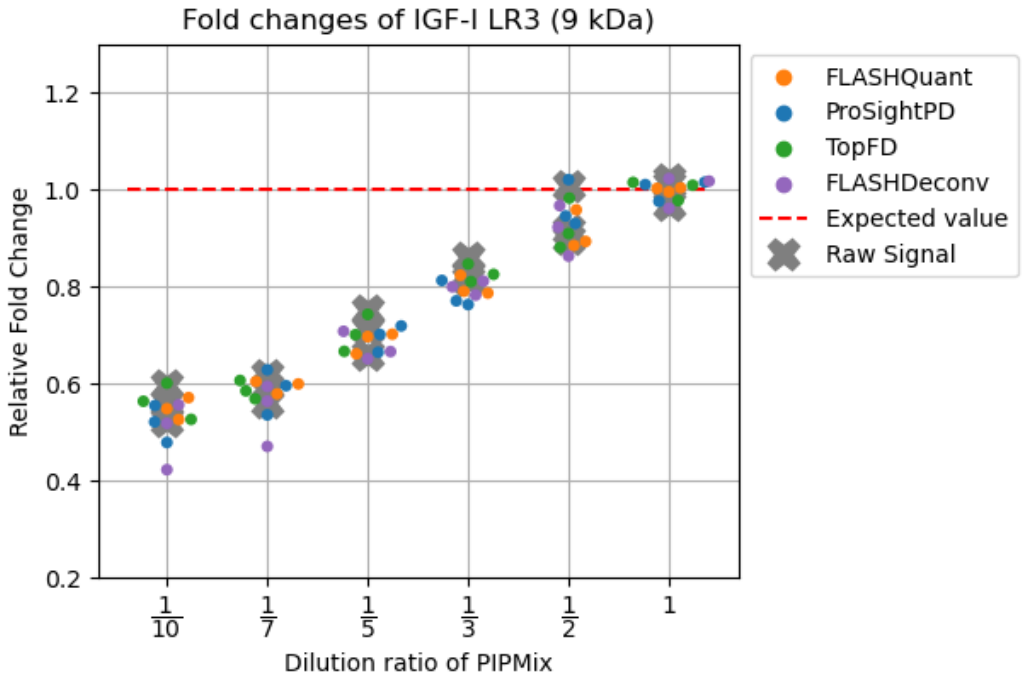


Figure D.1: Relative fold changes comparison between the quantified results and raw signal for the IGF-I LR3 protein from the SpikeIn dataset. Large deviations from the expected value (the red dashed line) were observed for the most abundant protein, IGF-I LR3, consistently across all tools. Given that this is the most abundant protein, we did not anticipate the quantification results to be incorrect but rather suspected distortion in the raw signal. To investigate, we examined the raw signal by extracting all peaks from the main charge state 8 of the protein within the corresponding retention time range (3865 ±50 seconds) and aggregating their intensities. Relative fold change values from these summed intensities, when compared to the relative fold change values across all tools (depicted as grey X markers), demonstrate the distorted raw signal contributing to the unexpected quantification results.

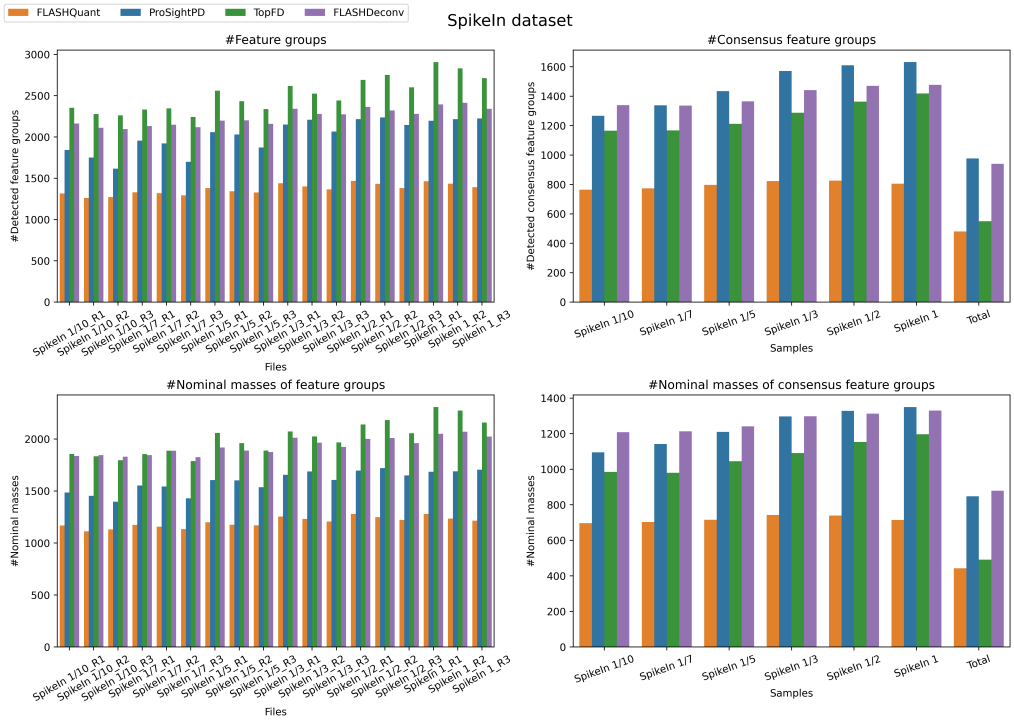


Figure D.2: The numbers of all feature groups, consensus feature groups, and nominal masses from the SpikeIn dataset. The bar plots show the number of detected feature groups per file (upper left), consensus feature groups per dilution and total consensus features among the dataset (upper right), nominal masses of detected feature groups (lower left), and nominal masses of consensus feature groups (lower right).

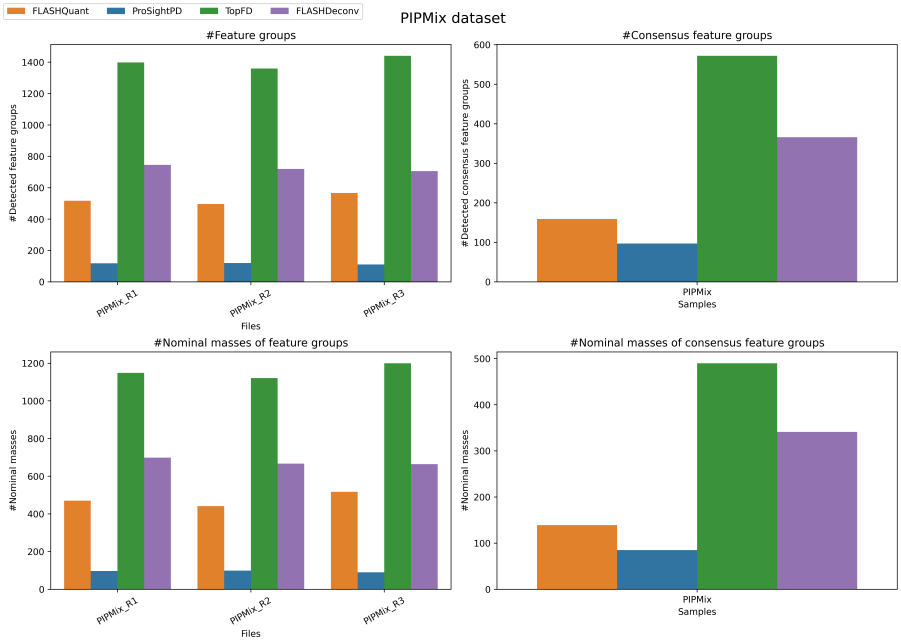
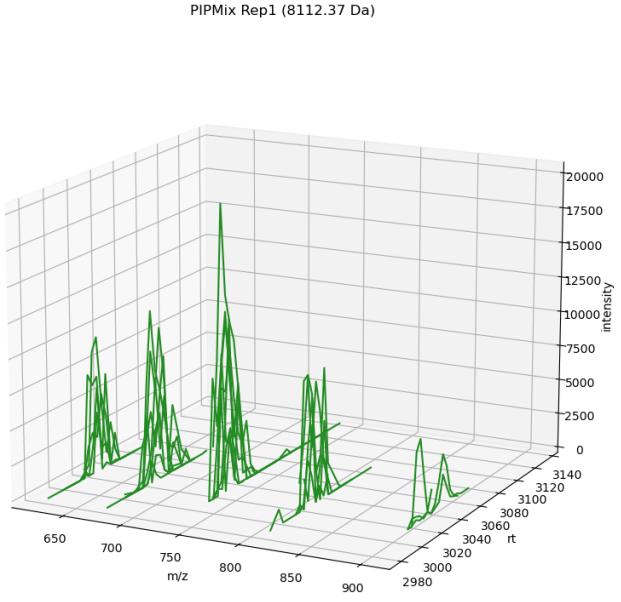


Figure D.3: Analogue of Fig. D.2 for the numbers of all feature groups, consensus feature groups, and nominal masses from the PIPMix dataset.

(a) (Proteoform with the mass 8112.37 Da



(b) Proteoform with the mass 11814.02 Da

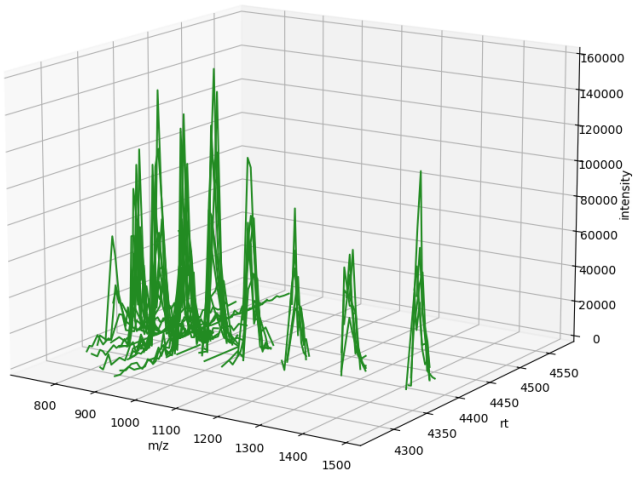


Figure D.4: Examples of raw m/z traces of FLASHQuant consensus feature group from the PIPMix dataset that did not match against identified masses and included in the unknown type from Fig. 4.7

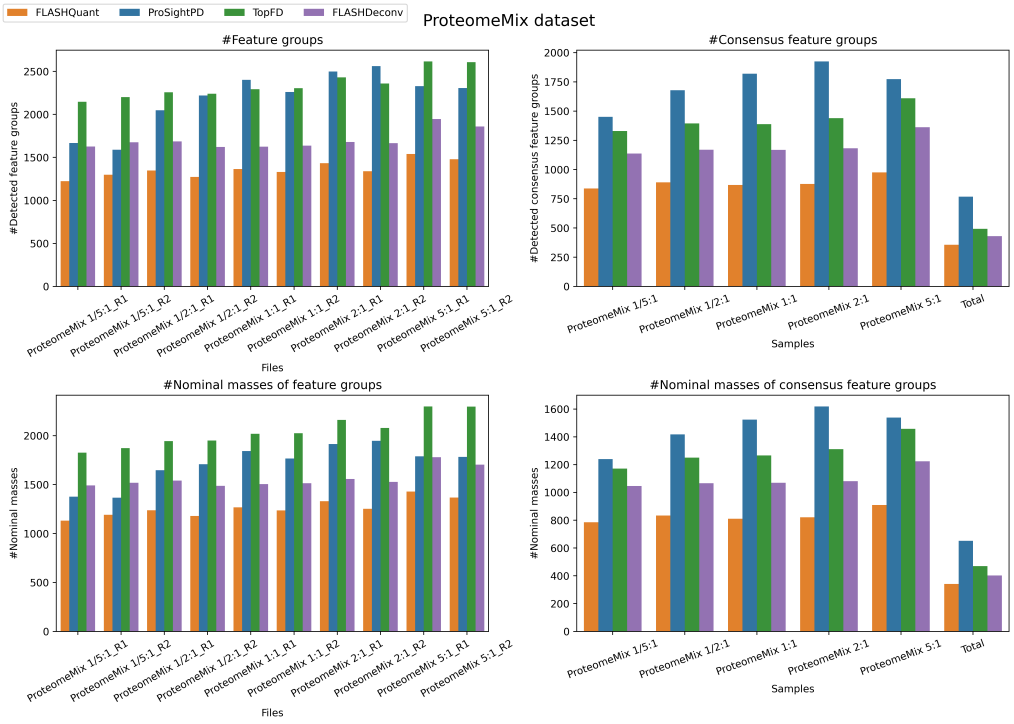


Figure D.5: Analogue of Fig. D.2 for the numbers of all feature groups, consensus feature groups, and nominal masses from the ProteomeMix dataset.

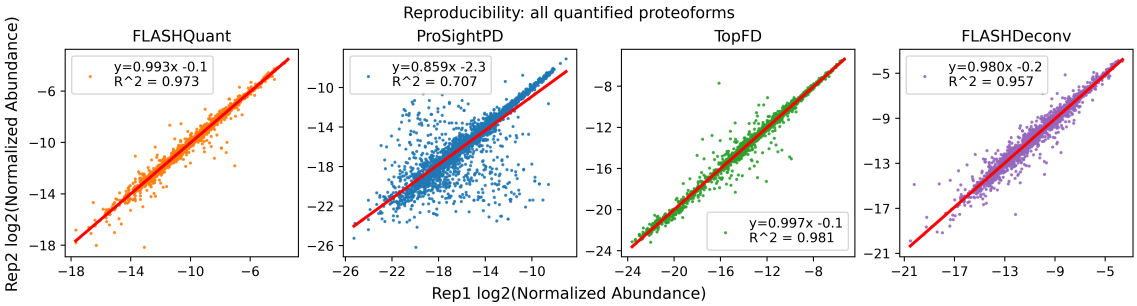


Figure D.6: Analogue of Fig. 4.10C for the quantification reproducibility of all detected consensus feature groups from the ProteomeMix dataset

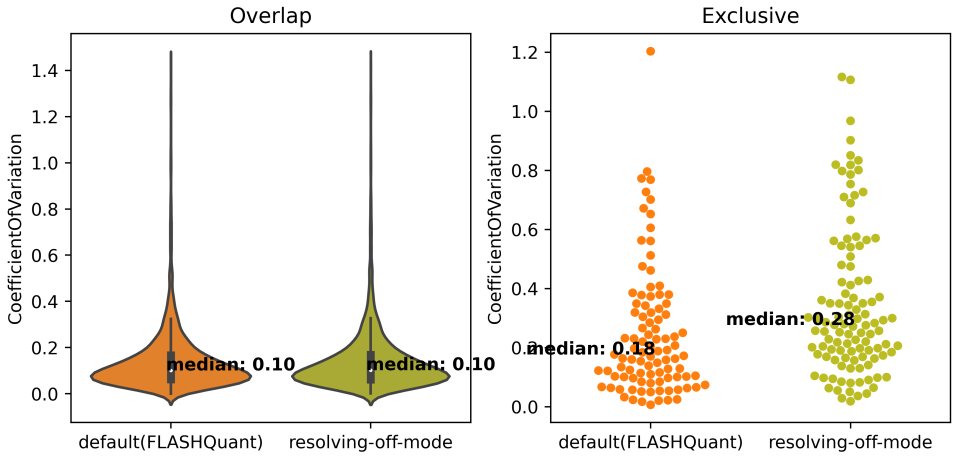


Figure D.7: Reproducibility comparison between the default-mode and resolving-off-mode with the SpikeIn dataset. CV values from technical replicates of the SpikeIn dataset were used for this analysis. While overlaps between the two modes show low median CV values of 0.1 (violin plots on the left), default-mode exclusives have a lower median CV value than the resolving-off-mode (swarm plots on the right).