Computational Methods for Four-Dimensional diaPASEF Proteomics Data Analysis

by

Kai Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2026

Doctoral Committee:

Professor Alexey I. Nesvizhskii, Chair
Associate Professor Alla Karnovsky
Professor Kayvan Najarian
Professor Arvind Rao
Professor Brandon T. Ruotolo

Kai Li

lkai@umich.edu

ORCID iD: 0000-0002-0810-2437

Acknowledgements

First and foremost, I would like to thank my mentor, Dr. Alexey I. Nesvizhskii, who is truly the best PI in the world. He is an exceptionally rigorous and thoughtful scientist, and at the same time a great friend with a unique sense of humor. Working under his guidance has been an invaluable experience, and I am deeply grateful for his constant support, trust, and mentorship throughout my PhD.

I would also like to thank Dr. Fengchao Yu, a research faculty member in our lab. We brainstormed together on many of the ideas and results presented in this dissertation. Thank you for being such a supportive co-mentor and collaborator. I am also thankful to all members of the Nesvizhskii lab. We are truly like a big family, working together and supporting one another while contributing to the field of proteomics research.

My journey in proteomics began in 2016, when I joined BGI-Shenzhen as an intern. There, I met Bo Wen, then the director of the mass spectrometry bioinformatics department and now at the University of Washington. He was the first person to introduce me to tandem mass spectrometry–based proteomics. Under his mentorship, I developed PDV, which became my first publication. At the end of 2018, I came to the United States and joined Dr. Bing Zhang's lab at Baylor College of Medicine. I am deeply grateful to Bing for sponsoring me and giving me the opportunity to continue my research. I was also very happy to continue working closely with Bo. During the pandemic, we accomplished a great deal of work together. Those two and a half years were one of the most important stages in my development as a researcher. I sincerely thank everyone in the Bing Zhang lab for their support and collaboration.

I want to give my deepest thanks to my wife, Ning. She gave up opportunities in China to join me here, and I could not have completed this dissertation without her companionship, patience, and unwavering support.

Finally, I would like to thank my parents and my beloved grandmother, who is now 91 years old. I was raised by her in a small and poor village. Looking back on the past 30 years of my life, this journey feels almost magical. It is still unbelievable to me that I am completing my

PhD at the University of Michigan. I am deeply grateful to everyone who taught me, supported me, and believed in me along the way.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) has been commonly used for large-scale proteomics. Early proteomics workflows primarily relied on data-dependent acquisition (DDA), in which a limited number of precursor ions are sequentially selected for isolation and fragmentation based on their intensities. Although DDA enables relatively straightforward interpretation of MS/MS spectra, it suffers from limited reproducibility and incomplete sampling, especially for low-abundance peptides. Data-independent acquisition (DIA) addresses these limitations by systematically fragmenting all precursor ions within predefined, wide isolation windows. However, DIA produces highly multiplexed MS/MS spectra, which substantially increases computational complexity due to the loss of explicit precursor–fragment relationships. Recently, diaPASEF, a DIA acquisition strategy implemented on the Bruker timsTOF platform, has integrated ion mobility separation into DIA workflows, providing an additional analytical dimension that reduces co-elution and improves precursor separation in the same isolation window. As a result, diaPASEF generates four-dimensional proteomics data consisting of mass-to-charge ratio, retention time, ion mobility, and intensity. While diaPASEF offers significant advantages in data quality and acquisition efficiency, effective computational strategies for fully exploiting its four-dimensional data structure remain limited.

A critical step in DIA data analysis is the construction of spectral libraries, which encode precursor properties such as retention time, ion mobility in diaPASEF data, and fragment ion intensities. A spectral library can be built using DDA data acquired from the same or similar samples, which requires additional effort and may lose proteome depth. With the development of machine learning, peptide-centric prediction strategies have been proposed to generate spectral libraries directly from protein sequence databases. Although peptide-centric methods benefit from machine learning–based prediction of peptide properties, they are often impractical for nonspecific digestion or open modification searches, where the candidate search space becomes extremely large. An alternative, spectrum-centric strategy directly derives pseudo-MS/MS

spectra from DIA data, enabling peptide identification using conventional DDA database search engines without prior library constraints. To date, no computational framework has enabled spectrum-centric analysis of diaPASEF data.

In this dissertation, I present diaTracer, a spectrum-centric computational framework for comprehensive four-dimensional diaPASEF proteomics data analysis, fully integrated into the FragPipe platform. diaTracer enables library-free peptide identification without restrictions on digestion specificity or post-translational modification complexity, thereby expanding the applicability of diaPASEF data analysis to a wide range of proteomics workflows.

Chapter 2 describes the core diaTracer algorithms, including four-dimensional feature detection, precursor–fragment clustering, and pseudo-MS/MS spectrum reconstruction. Benchmarking demonstrates performance comparable to established peptide-centric approaches on global proteome datasets. Chapter 3 highlights applications where spectrum-centric analysis offers distinct advantages, including semi-tryptic and open modification searches, phosphorylation analysis, and nonspecific immunopeptidomics workflows, where peptide-centric strategies are often limited.

Chapter 4 extends diaTracer to support diagonal PASEF acquisition methods, including synchro-PASEF and Slice-PASEF, through the introduction of a pseudo-isolation window strategy for reconstructing sliced fragment signals. This chapter also presents algorithmic optimizations implemented in diaTracer 2.0 that improve isotope filtering performance and computational efficiency. Chapter 5 summarizes the contributions of this work and discusses future directions for computational analysis of diaPASEF and related DIA technologies.

Overall, this dissertation establishes a spectrum-centric computational framework that enables flexible, scalable, and unrestricted analysis of four-dimensional diaPASEF proteomics data.

Chapter 1 Introduction

1.1 Mass spectrometry-based bottom-up proteomics

Proteins are the functional products translated from RNA transcripts, which are themselves transcribed from DNA. However, the relationship between genes and proteins is not one-to-one. A single gene can give rise to multiple isoforms through alternative splicing and can undergo extensive post-translational modifications, resulting in distinct proteoforms[1]. It has been estimated that approximately 20,000 protein-coding genes in the human genome can generate more than one million different proteoforms[2]. Therefore, investigations limited to the genomic or transcriptomic level are insufficient to fully characterize biological systems, and direct analysis at the protein level is required.

Proteomics is the large-scale study of proteins aimed at understanding their abundance, structure, function, interactions, and post-translational modifications[3]. Over the past two decades, tandem mass spectrometry (MS/MS)–based proteomics has become a powerful and sensitive approach for analyzing complex protein mixtures[4-6]. Based on the level at which proteins are analyzed, MS/MS-based proteomics can be broadly classified into top-down and bottom-up strategies.

In top-down proteomics, intact proteins are introduced directly into the mass spectrometer and analyzed without prior digestion, allowing direct characterization of proteoforms. In contrast, bottom-up proteomics, which is the focus of this dissertation, analyzes peptides generated from proteolytic digestion of proteins[7]. In a typical bottom-up proteomics workflow, proteins are first enzymatically digested into shorter peptides, most commonly using sequence-specific proteases such as trypsin. The resulting peptide mixture is separated by liquid chromatography to reduce sample complexity[8,9]. Peptides are then ionized, typically by electrospray ionization, and introduced into a tandem mass spectrometer for analysis[10].

At the first level of mass spectrometry (MS1), the mass-to-charge ratios (m/z) of intact peptide precursor ions are measured. Selected precursor ions are subsequently fragmented to generate product ions, which are analyzed in the second level of mass spectrometry (MS2). By

matching specific fragment ion patterns to theoretical spectra derived from protein sequence databases, peptide sequences can be confidently identified. Once peptides are identified, they can be mapped back to their corresponding proteins. Quantitative information can be obtained by measuring the signal intensities of peptide precursor ions or fragment ions, enabling relative or absolute comparisons of protein abundance across samples. Together, these capabilities make bottom-up MS/MS–based proteomics a foundational technology for large-scale protein identification and quantification in biological research.

1.2 Data acquisition

MS/MS–based proteomics data can be acquired using either targeted or untargeted strategies. Targeted proteomics approaches focus on a predefined set of peptides and enable highly sensitive and accurate quantification, making them particularly suitable for clinical and validation-oriented applications[11,12]. In contrast, untargeted proteomics aims to achieve near-complete proteome coverage in a discovery-driven manner and is therefore widely applied in systems biology and large-scale proteomics studies[13]. This dissertation focuses on untargeted bottom-up proteomics.

For untargeted bottom-up proteomics, data acquisition strategies can be broadly grouped into data-dependent acquisition (DDA) and data-independent acquisition (DIA). These two approaches mainly differ in how precursor ions are selected for fragmentation and how MS/MS spectra are acquired.

1.2.1 Data-dependent acquisition

Data-dependent acquisition (DDA) is one of the earliest and most widely used strategies in MS/MS–based proteomics, originally introduced in the early 1990s[14,15]. In a typical DDA workflow, the mass spectrometer first performs a full precursor ion scan (MS1) to measure the m/z and intensities of all ionized peptides. Based on this MS1 survey scan, the n most abundant precursor ions are automatically selected in real time for sequential isolation and fragmentation, generating corresponding MS/MS (MS2) spectra.

This selection strategy represents a compromise between proteome coverage and instrumental limitations. Because mass spectrometers have insufficient scan speed, it is not feasible to acquire MS/MS spectra for every detected precursor ion within a chromatographic time frame[16]. By prioritizing the most intense precursors, DDA reduces spectral complexity and

improves the quality of individual MS/MS spectra. To further increase identification coverage, selected precursor ions are typically placed on a dynamic exclusion list for a period of time, preventing repeated fragmentation of the same ions and allowing lower-abundance precursors to be sampled[16,17].

The MS/MS spectra acquired in DDA experiments can be analyzed using database search algorithms to identify peptide sequences by matching experimental fragment ion patterns to theoretical spectra derived from protein sequence databases[18,19]. Using this approach, DDA workflows routinely identify thousands of proteins and provide quantitative information based on peptide signal intensities.

Despite its widespread use, DDA has several well-recognized limitations. Precursor selection in DDA is inherently stochastic, particularly in complex samples where many peptide species co-elute within a single MS1 scan. Under these conditions, only the most abundant peptides are consistently selected for fragmentation, while low-abundance peptides may be missed[19]. As a result, different subsets of peptides may be identified across replicate analyses, leading to missing values and reduced quantitative reproducibility[20,21]. Furthermore, quantification in DDA is typically based on MS1 chromatographic peak areas, which are sensitive to interference from co-eluting ions, especially in highly complex proteomic samples[22,23]. These limitations have motivated the development of alternative acquisition strategies, such as data-independent acquisition, to improve reproducibility and quantitative consistency.

1.2.2 Data-independent acquisition

In contrast to DDA-based proteomics, data-independent acquisition (DIA) acquires both MS1 and MS2 data without bias toward precursor ion selection, with the goal of capturing comprehensive fragment ion information from all detectable peptides in a sample. Since the early 2000s, a variety of DIA strategies have been proposed that differ in isolation window design, scan order, and data analysis approaches[24-32]. Venable et al. were among the first to formally introduce and name the concept of data-independent acquisition[33].

A major milestone in the development of DIA was reported in 2012 by Gillet et al., who introduced sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS)[25]. SWATH-MS is a representative DIA strategy that takes advantage of the increased scan speed and improved mass resolution available on modern mass

spectrometers, such as the AB Sciex TripleTOF 5600 and Orbitrap-based instruments. In a typical SWATH-MS workflow, the mass spectrometer sequentially cycles through a predefined series of precursor isolation windows, spanning a mass-to-charge range of approximately 400–1200 m/z. Within each isolation window, all precursor ions are co-isolated and fragmented, and the resulting fragment ions are recorded in MS2 scans.

Compared with DDA, DIA-based acquisition strategies such as SWATH-MS provide improved reproducibility and quantitative consistency across samples[25,34-36], as all precursor ions within the specified mass range are systematically fragmented in every acquisition cycle[21], independently of precursor intensity. In principle, the fragmentation of all ionized peptides within defined isolation windows allows comprehensive identification and quantification of the detectable proteome.

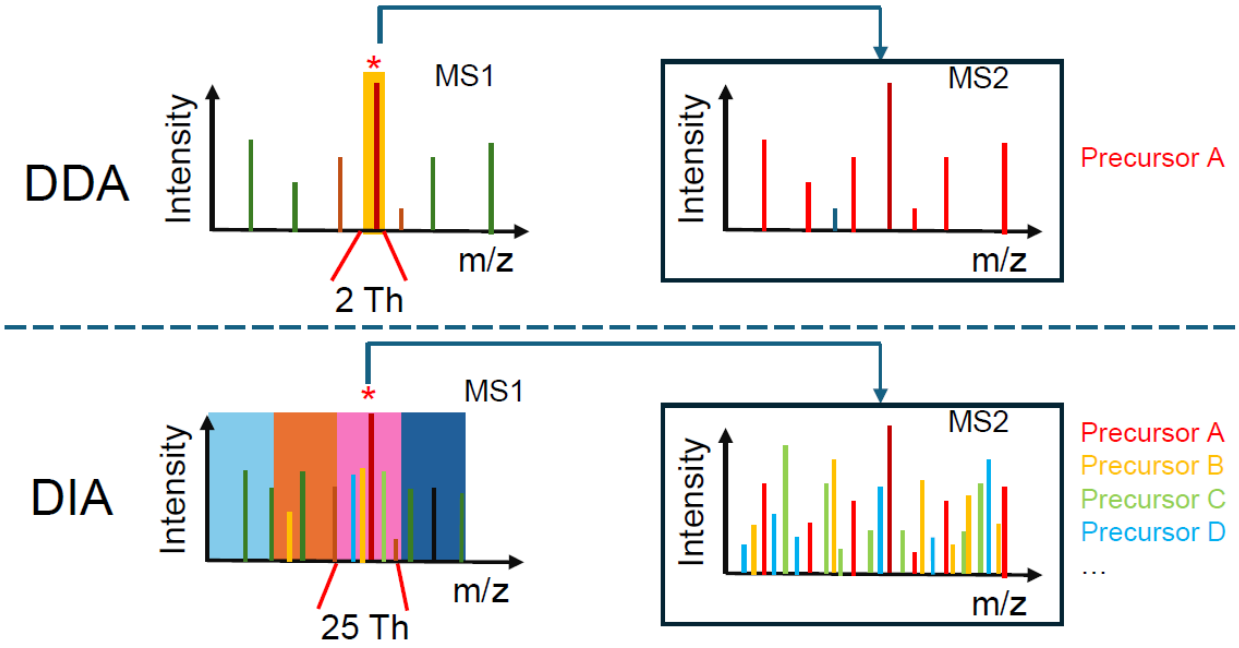


Figure 1-1. Examples of data-dependent acquisition (DDA) and data-independent acquisition (DIA).

In DDA, MS2 spectra are acquired by selectively fragmenting a small number of the most intense precursor ions detected in the MS1 survey scan. As illustrated in the top panel, a narrow isolation window (e.g., 2 Th) centered on a selected precursor ion (Precursor A) is used for fragmentation, resulting in an MS2 spectrum that predominantly contains fragment ions derived from that single precursor. In contrast, DIA systematically fragments all precursor ions within predefined, wide m/z isolation windows, independent of their intensities. As shown in the bottom panel, the MS1 m/z range is partitioned into multiple wide windows (e.g., 25 Th), and all ions within each window are co-isolated and fragmented, producing complex MS2 spectra that contain fragment ions from multiple co-eluting precursors (e.g., Precursors A–D).

Quantification in DIA workflows is typically performed at the MS2 level by extracting fragment ion chromatograms corresponding to peptides of interest. Compared with MS1-based

extracted ion chromatograms used in DDA workflows, MS2-level quantification is generally less susceptible to chemical interference and co-eluting background signals, particularly in complex biological samples. These advantages have contributed to the widespread adoption of DIA as a preferred acquisition strategy for large-scale, reproducible proteomics studies and have motivated continued development of both acquisition methods and computational analysis frameworks[25,37].

1.2.3 Trapped ion mobility spectrometry and the timsTOF/PASEF platform

Trapped ion mobility spectrometry (TIMS) was introduced in the early 2010s as a gas-phase separation method in which ions are held stationary by balancing an electric field against a moving bath gas and are then released in mobility order during a voltage ramp[38]. Early work also demonstrated direct coupling of TIMS to mass spectrometry and established the basic feasibility of performing ion mobility separation immediately upstream of mass analysis[39]. Subsequent methodological studies clarified the physical principles of TIMS, including ion motion, resolving power, and calibration, and helped establish TIMS as a robust platform for reproducible mobility measurements[40].

A major step toward large-scale proteomics came with the development of parallel accumulation–serial fragmentation (PASEF)[41]. In PASEF, ions are accumulated in parallel with the TIMS device and then released sequentially according to their ion mobility, while the quadrupole rapidly switches to isolate different precursor ions as they elute. This synchronization allows multiple precursors to be selected and fragmented during a single TIMS scan, greatly increasing sequencing speed without the sensitivity penalty typically associated with faster acquisition[42]. The PASEF concept was later implemented at the instrument level in the Bruker timsTOF Pro, a quadrupole time-of-flight platform equipped with a dual-TIMS analyzer that separates ion accumulation from mobility analysis so that both processes can occur in parallel. This design improves ion utilization and contributes to the high speed and sensitivity of the platform.

Importantly, the timsTOF/PASEF platform produces data that are four-dimensional. Peptide features are defined not only by retention time and mass-to-charge ratio, but also by ion mobility and signal intensity. In practice, the ion mobility dimension is often reported as reduced inverse mobility, $1/K_0$, and can be converted to collision cross section (CCS), providing an additional descriptor for peptide ions. Because isotope clusters can be localized jointly in

retention time, m/z, ion mobility, and intensity, timsTOF proteomics data are naturally treated as 4D data[42]. This 4D structure underlies both TIMS-aware feature detection and the development of later DIA acquisition methods on the timsTOF platform.

1.2.4 4D PASEF acquisition modes

The first major data-independent acquisition method built on the timsTOF/PASEF platform was diaPASEF. Unlike conventional DIA, which cycles through fixed m/z windows and fragments only a small fraction of the incoming ion beam at any given moment, diaPASEF exploits the characteristic correlation between precursor m/z and ion mobility[43]. The quadrupole isolation windows are therefore defined in joint m/z-ion mobility space and synchronized with ion release from the TIMS device. This design markedly increases ion utilization while preserving the separation power of ion mobility, enabling cleaner fragment ion measurements and improved sensitivity. The original diaPASEF study also showed strong reproducibility and quantitative performance at very low sample amounts, helping establish the method as a sensitive 4D DIA strategy.

More recent PASEF-based DIA methods further refine the geometry of precursor isolation in the m/z-ion mobility plane. Synchro-PASEF uses a continuously moving quadrupole window that follows the precursor distribution diagonally during the TIMS ramp[44]. This continuous synchronization improves sampling efficiency and, importantly, slices precursors in a way that helps associate fragment ions with their originating precursors, thereby supporting interference removal and more specific fragment extraction. Slice-PASEF takes a complementary approach by partitioning precursor space into continuous ion mobility slices, allowing nearly all available precursor ions to be directed to fragmentation[45]. This strategy emphasizes maximal ion utilization and has been reported to improve proteome depth and quantitative precision, particularly for low-input samples. Different diaPASEF acquisition schemes are illustrated in Figure 1-2.

Together, diaPASEF, synchro-PASEF, and Slice-PASEF illustrate how the TIMS dimension can be used not only as an additional axis of separation, but also as a central design principle for acquisition. By coordinating quadrupole isolation with mobility-resolved ion release, these methods partition precursor ions into method-dependent diagonal slices across the m/z-ion mobility plane. As a result, they can improve ion sampling efficiency, precursor specificity, and fragment deconvolution while maintaining short cycle times. These properties

make 4D PASEF acquisition especially attractive for complex and low-input proteomics applications.

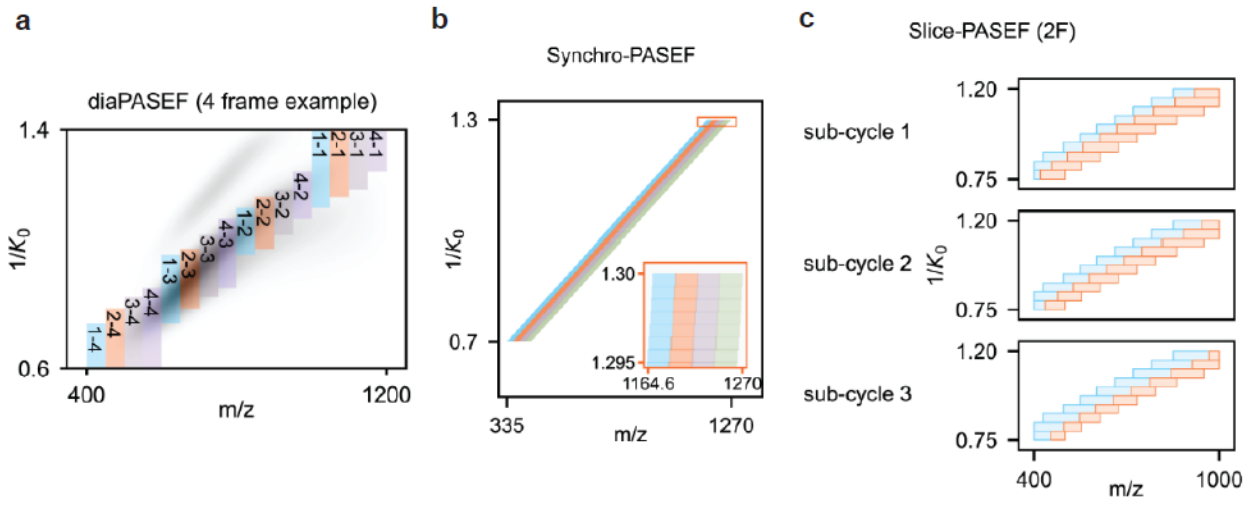


Figure 1-2. Four-dimensional PASEF-based DIA acquisition schemes.

(a) Example of a conventional diaPASEF acquisition using four frames, illustrating two-dimensional isolation windows distributed across the m/z and ion mobility (1/K₀) dimensions. Each frame contains a set of discrete isolation windows. (b) Synchro-PASEF acquisition with four frames. The quadrupole isolation is synchronized with ion mobility elution during each TIMS ramp, resulting in continuous, diagonal isolation across the m/z–ion mobility plane. (c) Slice-PASEF acquisition using two frames and three subcycles. Each subcycle covers the same m/z and ion mobility ranges, but the isolation window boundaries are shifted between subcycles to sample different precursor subsets. *Adapted from Lou et al., Molecular & Cellular Proteomics (2024)[46], under the Creative Commons CC BY 4.0 license.*

1.3 DIA data analysis strategies

DIA has become a widely adopted strategy for large-scale proteome profiling due to its high quantitative reproducibility and reduced between-run missing values across large sample cohorts. By systematically fragmenting all precursor ions within predefined isolation windows, DIA ensures consistent sampling of peptides across experiments. However, this comprehensive acquisition strategy produces highly multiplexed MS/MS data, in which fragment ion signals originate from multiple co-eluting precursors. As a result, DIA data analysis requires specialized computational strategies to separate overlapping fragment ion signals and to achieve confident peptide identification and accurate quantification. To address the complexity of DIA data, Gillet et al. introduced the concept of spectral library-based search to query acquired precursor and fragment ions from information stored in the peptide MS/MS spectral libraries[25]. Typical spectral libraries contain peptide sequence information together with precursor m/z values, charge state, fragment ion m/z values, relative fragment ion intensities, and normalized retention times. For ion mobility–enabled DIA data, such as diaPASEF, spectral libraries may also include ion

mobility values. By providing expected peptide signatures, spectral libraries constrain the search space and enable targeted extraction and scoring of fragment ion chromatograms from DIA data.

A spectral library is usually generated using additional DDA experiments performed on the same or closely related sample types analyzed by DIA. These DDA datasets are searched against protein sequence databases to identify peptides, and confidently identified spectra are collected to build a consensus library. In general, spectral library construction involves three major steps: collection of high-confidence peptide-spectra matches from database search results, integration of spectra corresponding to the same peptide ion into representative consensus spectra with averaged normalized retention times, and quality control to remove unreliable or low-quality entries[47]. During DIA analysis, peptide and protein identification are achieved by scoring the agreement between fragment ion signals observed in the DIA data and peptide query information stored in the spectral library, including peptide sequence, precursor m/z and charge state, fragment ion m/z values, relative intensities, and retention time.

Although DDA-based spectral libraries have enabled robust DIA data analysis, they require additional sample material and instrument time and are subject to the inherent limitations of DDA, including stochastic precursor selection and reduced detection of low-abundance peptides. These limitations can lead to incomplete library coverage and bias toward highly abundant species. To overcome these constraints, DDA library-free DIA analysis strategies have been proposed in recent years, aiming to eliminate the need for additional DDA experiments.

Broadly, DDA library-free DIA analysis approaches can be categorized into peptide-centric and spectrum-centric strategies based on how spectral libraries are generated and utilized. Peptide-centric approaches rely on predicted peptide properties to construct libraries in-silico, whereas spectrum-centric approaches focus on deriving peptide identification directly from DIA data itself.

1.3.1 Peptide-centric

In recent years, advances in machine learning and deep learning have enabled the in-silico generation of spectral libraries. In principle, key peptide properties—including retention time, ion mobility, and MS/MS fragmentation patterns—can be predicted for all possible peptide precursors derived from a protein sequence database under user-defined settings, such as enzymatic digestion rules, missed cleavages, charge states, and post-translational modifications. Multiple studies have demonstrated that computationally predicted fragment ion spectra can

achieve accuracy comparable to that of empirically acquired spectra. As a result, several spectral library prediction tools and peptide-centric DIA analysis tools have been developed in recent years, including Prosit[48], AlphaPepDeep[49], DeepLC[50], and DIA-NN[51].

Despite these advantages, peptide-centric DIA analysis has inherent limitations. Because the set of candidate peptides is specified in advance, the computational search space can grow rapidly under nonspecific digestion immunopeptidomics workflows or open modification searches, leading to increased computational burden and reduced statistical power. In addition, the performance of prediction-based tools is highly dependent on the availability and diversity of training data. For many post-translationally modified peptides, insufficient training data can result in reduced prediction accuracy for retention time, ion mobility, and fragmentation patterns, thereby limiting identification performance in modification-focused experiments. These limitations motivate the development of alternative DIA data analysis strategies that rely less on predefined peptide hypotheses and more directly on information contained within the acquired DIA data.

1.3.2 Spectrum-centric

Spectrum-centric DIA analysis represents an alternative class of computational strategies in which peptide identification is driven primarily by information extracted directly from the acquired DIA data, rather than by querying a predefined set of peptide candidates. In this strategy, DIA data are first processed to identify precursor and fragment ion features, which are then assembled into DDA-like pseudo-MS/MS spectra[52,53]. These reconstructed spectra can subsequently be analyzed using conventional DDA database search engines for peptide identification. Unlike peptide-centric approaches, spectrum-centric methods do not rely on a predefined list of peptide candidates. Instead, they aim to recover precursor–fragment relationships directly from the DIA data by mining signal coherence across multiple dimensions, such as m/z and retention time. By reconstructing pseudo-MS/MS spectra, spectrum-centric workflows enable the application of established peptide identification algorithms without reliance on predefined spectral libraries. Figure 1-3 shows the overview of spectral library generation strategies.

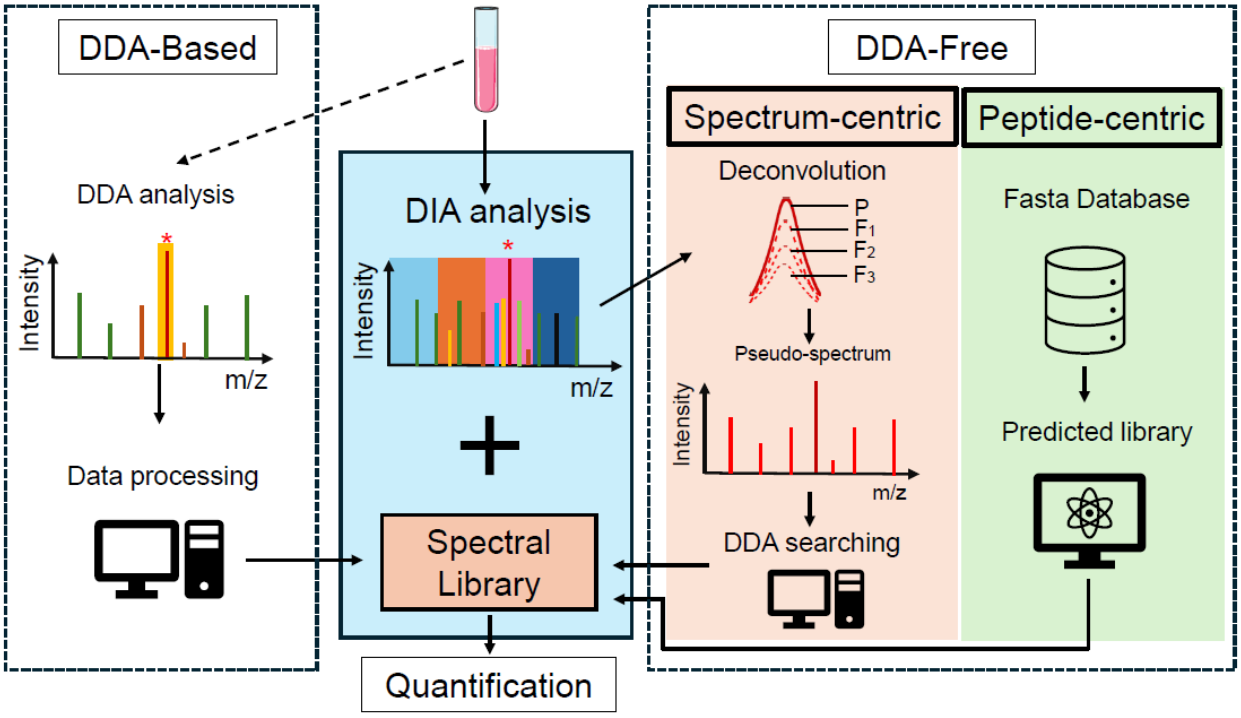


Figure 1-3. Spectral library generation strategies for DIA analysis.

Spectral libraries can be generated using DDA-based approaches from additional DDA experiments or using DDA-free approaches directly from DIA data via spectrum-centric deconvolution or peptide-centric in-silico prediction.

A key advantage of spectrum-centric DIA analysis is its flexibility. Because spectral libraries are constructed after signal detection, spectrum-centric approaches are well suited for applications involving large or poorly defined search spaces, including open modification searches and nonspecific digestion immunopeptidomics. In these contexts, spectrum-centric analysis avoids the rapid expansion of candidate search spaces and reduces reliance on potentially unreliable peptide property predictions required by peptide-centric approaches.

1.4 Motivations for spectrum-centric analysis

Several spectrum-centric DIA analysis tools have been developed and successfully applied to data acquired on mass spectrometry platforms such as Thermo Fisher Scientific Orbitrap and Sciex TripleTOF instruments. Notably, tools such as DIA-Umpire and MSFragger-DIA enable analysis of DIA data directly and subsequent peptide identification using conventional DDA database search engines. However, these methods cannot process diaPASEF data acquired on the Bruker timsTOF platform. The primary limitation arises from the additional ion mobility dimension in diaPASEF data, which introduces significant challenges for detecting

and associating extracted ion chromatogram (XIC) features across multiple dimensions. Although commercial software such as Spectronaut[54] supports direct peptide identification from diaPASEF data, the underlying algorithmic details have not been publicly disclosed to date. As a result, the spectrum-centric analysis of diaPASEF data remains limited in terms of transparency, extensibility, and integration with open computational frameworks.

The primary motivation of this dissertation is to develop a comprehensive spectrum-centric computational framework for four-dimensional diaPASEF proteomics data analysis. By leveraging the additional ion mobility separation provided by the timsTOF platform, such a framework aims to fully utilize the high sensitivity and separation efficiency of diaPASEF data. This capability is particularly important for applications involving nonspecific digestion and open modification searches, where peptide-centric approaches are often constrained by large search spaces and unreliable predictions. Furthermore, given the rapid evolution of timsTOF acquisition technologies, an additional motivation of this work is to extend spectrum-centric analysis to recently introduced diagonal PASEF acquisition methods, including synchro-PASEF and Slice-PASEF.

Chapter 2 diaTracer Enables Spectrum-Centric Analysis for 4D diaPASEF Data

The content of this chapter is adapted from the previous publication by the author in Nature Communications[55].

2.1 Introduction

Liquid chromatography–tandem mass spectrometry (LC–MS/MS)–based data-independent acquisition (DIA) enables reproducible[56] and large-scale proteome profiling but produces highly multiplexed fragment ion spectra that require specialized computational analysis[25,27,33]. As discussed in Chapter 1, ion mobility–enabled DIA methods such as diaPASEF[42] further increase data complexity by adding an additional separation dimension, improving precursor discrimination while introducing new challenges for signal detection and precursor–fragment association.

Existing spectrum-centric DIA analysis frameworks, including DIA-Umpire[52] and MSFragger-DIA[57], have demonstrated robust performance on Orbitrap- and TripleTOF-based DIA data. However, these methods are not directly applicable to diaPASEF data acquired on the Bruker timsTOF platform, primarily due to the need to detect and associate extracted ion chromatogram features across the additional ion mobility dimension. Although commercial software such as Spectronaut[54] supports diaPASEF data analysis, the underlying algorithms are not publicly disclosed, limiting transparency and extensibility.

To address the need for an efficient method for analyzing diaPASEF data, this chapter presents a spectrum-centric computational framework, termed diaTracer, specifically designed for four-dimensional diaPASEF proteomics data analysis. diaTracer is engineered to efficiently process the multi-dimensional feature space defined by mass-to-charge ratio (m/z), retention time, and ion mobility in diaPASEF data, enabling robust detection of precursor and fragment ion features. These features are subsequently assembled into data-dependent acquisition (DDA)-like pseudo-MS/MS spectra that are compatible with conventional DDA peptide identification tools, such as MSFragger. diaTracer is fully integrated into the widely used FragPipe

computational platform. The performance and applicability of diaTracer are demonstrated using diaPASEF datasets derived from triple-negative breast cancer (TNBC) and low-input spatial proteomics studies.

2.2 Methods

2.2.1 Precursor and fragment feature extraction in diaTracer

In conventional DIA data, each mass spectrometry (MS) frame can be viewed as a two-dimensional representation of m/z and intensity. The MS1 precursor mass range is partitioned into a series of isolation windows, and all ions within each window are co-isolated and fragmented to generate corresponding MS2 data. As peptides elute from the liquid chromatography (LC) column, their signals accumulate continuously over retention time (RT) and typically form a bell-shaped extracted ion chromatogram (XIC) with a well-defined apex. Fragment ions originating from the same precursor generally co-elute and therefore exhibit similar XIC shapes aligned in retention time. In diaPASEF, trapped ion mobility (IM) separation introduces an additional dimension. Instead of describing each signal by m/z and intensity alone, diaPASEF measures m/z, ion mobility ($1/K_0$), and intensity within each frame, resulting in a three-dimensional distribution of points per frame in m/z–IM–intensity space, sampled across retention time. To establish precursor–fragment relationships and reconstruct pseudo-MS/MS spectra, diaTracer mines the full four-dimensional structure of the data: m/z, ion mobility, intensity, and retention time.

The first step of diaTracer is to extract candidate signal features from both MS1 and MS2 frames. Because both m/z and $1/K_0$ are continuous-valued, diaTracer discretizes them into integer indices to enable efficient computation. At the beginning of the workflow, the algorithm summarizes the raw data and estimates the smallest observed spacing between neighboring measurements in m/z and IM, denoted as the m/z and IM "resolutions" (m and i). Using these resolutions, each isolation window can be represented as a discrete matrix. For an isolation window spanning a fixed ion mobility range and m/z range, the number of IM bins (W) and m/z bins (N) are determined by the corresponding range divided by the discretization step size. In practice, diaTracer stores the resulting frame-level representation as sparse structures because only a small fraction of grid cells contains nonzero signal.

For computational efficiency and parallelization, diaTracer further divides each isolation window into overlapping m/z sub-windows ("bins") of fixed width. This reduces memory footprint, enables multi-threaded processing, and avoids losing features near sub-window boundaries.

Signals in a single diaPASEF frame, especially low-intensity precursor and fragment signals, can be sparse. Motivated by observations that integrating signal evidence across multiple mobility scans can improve detectability[58], diaTracer aggregates information across a narrow retention time neighborhood rather than processing each frame in isolation. Specifically, when extracting features for a given frame *n*, diaTracer constructs a composite representation by summing data from the last two adjacent frames (typically *n - 2*, *n - 1*, and *n*) within the same isolation window. This aggregation is implemented as a sliding window: when moving from frame *n* to *n + 1*, the contribution from *n - 2* is removed and the contribution from *n + 1* is added. The same strategy is applied at both MS1 and MS2 levels.

After aggregation, diaTracer removes isolated points that lack sufficient local support. Points without enough neighboring nonzero entries within a defined m/z–IM neighborhood are treated as "loners" and discarded. This step substantially reduces noise-driven detections in sparse regions. The remaining signal is then smoothed using a two-dimensional Gaussian kernel to suppress high-frequency noise and to enhance coherent peak-like structures (as shown for an example feature in Figure 2-1). Local maxima are identified in the smoothed m/z–IM intensity surface and used as seeds to define two-dimensional features. Rather than performing full parametric 2D Gaussian fitting, diaTracer traces peak neighborhoods around these maxima and estimates feature centers and widths using intensity-weighted statistics, producing robust peak coordinates in m/z and ion mobility.

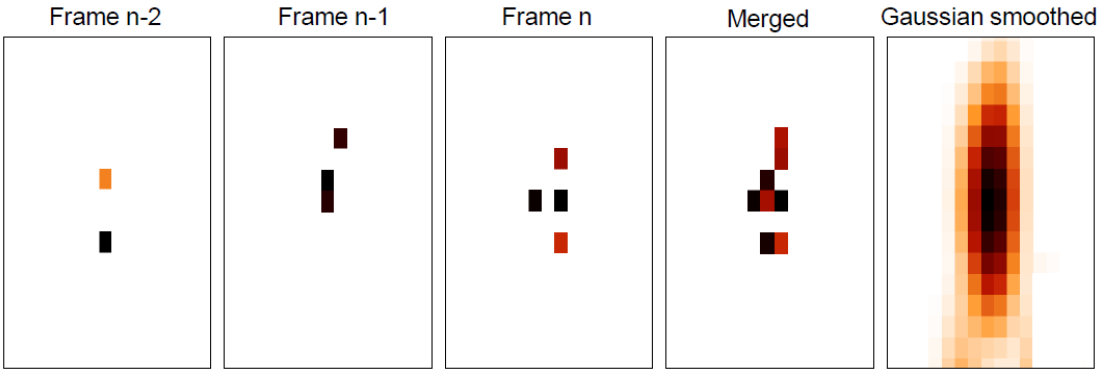


Figure 2-1. Signal enhancement by frame aggregation and Gaussian smoothing.

An example of an extremely sparse signal observed across three consecutive frames (*n–2* to *n*), where individual frames contain few intensity points and the signal center and boundaries are difficult to define. By aggregating neighboring frames, the signal is amplified and becomes more coherent. Subsequent Gaussian smoothing further enhances the signal structure, enabling reliable detection of feature center and extent.

For each detected feature, diaTracer records a set of attributes including the estimated m/z and IM centers, as well as the corresponding m/z and IM boundaries that define the local support region. For every RT frame in which a feature is present, diaTracer constructs a frame-level extracted ion signal by integrating intensities within the feature's m/z–IM support region in the sparse matrix representation. This process yields a one-dimensional XIC along the RT axis that integrates signal information across both m/z and IM dimensions. Feature start and end points are determined using predefined criteria to enforce a coherent, bell-shaped chromatographic profile. When a traced signal is elongated or contains multiple local maxima along retention time, diaTracer applies post-processing to segment it into separate peaks. In the current implementation, this is achieved using long-peak splitting logic followed by Savitzky–Golay smoothing and Z-score–based peak detection[59], enabling reliable separation of merged or extended chromatographic traces into multiple peak candidates.

2.2.2 Isotope filtering and charge assignment in diaTracer

To separate true signals from background noise and assign charge states to precursors, diaTracer performs isotope peak grouping and filtering on the traced MS1 features prior to precursor–fragment association. All traced MS1 peaks are first indexed in a three-dimensional KD-tree[60] using their discrete peak-center coordinates (m/z bin, IM bin, apex frame), enabling efficient range queries in the joint m/z–IM–RT space. For each candidate monoisotopic peak, diaTracer queries neighboring peaks within user-defined tolerances in IM (±ΔIM bins) and apex frame (±Δframe), and then searches in the m/z dimension for putative isotope partners. Candidate isotopes are evaluated for multiple charge hypotheses (z =1-4 by default) by comparing observed isotope spacings to the expected isotope offset $\Delta(m/z) \approx 1.00335/z$, while allowing a parts per million (ppm)-level mass tolerance. For each charge state, up to four isotope peaks (+1 to +4) are selected based on the agreement of normalized intensities with the theoretical isotope envelope predicted from the precursor mass (computed from the peak-center m/z and candidate charge). To ensure the isotope cluster reflects a coherent chromatographic species rather than coincidental co-elution, diaTracer additionally requires strong agreement between isotope and monoisotopic elution profiles: the Pearson correlation is computed between

the smoothed XIC of the monoisotopic feature and each isotope candidate over an overlapping RT segment centered on the apex (using an apex-aligned window and accommodating small apex offsets). Only charge hypotheses that meet minimum quality criteria—specifically, sufficient isotope evidence (at least one isotope peak), a minimum correlation between theoretical and experimental isotope intensity patterns, and a minimum XIC-shape correlation—are retained.

When multiple charge hypotheses pass these criteria, diaTracer applies a tie-breaking strategy that favors (i) higher combined agreement with the theoretical envelope (full-pattern correlation plus isotope-only correlation) and (ii) more complete isotope series (fewer missing isotope peaks). An optional mass-defect filter can be applied as an additional physical constraint to remove charge assignments inconsistent with typical peptide mass-defect ranges[61]. To minimize redundancy from isotope-derived duplicates, diaTracer then performs a second-pass filtering step: once a monoisotopic peak is assigned one or more charge states, peaks consistent with its predicted isotope positions (within the same IM/RT neighborhood and within ppm m/z tolerance) are flagged as isotopes and excluded from the final precursor list. The output of this procedure is a non-redundant set of monoisotopic MS1 precursor features annotated with one or more plausible charge states (multiple charges are retained when ambiguity remains).

2.2.3 Precursor and fragment clustering and pseudo-MS/MS spectrum assembly

After detecting precursor and fragment XIC features and filtering precursor isotopes, diaTracer groups features into precursor–fragment clusters and converts each cluster into one or more pseudo-MS/MS spectra. Clustering is anchored on each monoisotopic MS1 precursor feature within a given diaPASEF isolation window. For a precursor with apex frame index f_{apex} and apex ion mobility IM_{apex}, diaTracer searches for fragment candidates in the pre-indexed MS2 feature matrix within a bounded neighborhood in both retention time (frame index) and ion mobility. Specifically, the RT search interval is defined as $[f_{\mathrm{apex}} - \Delta RT, f_{\mathrm{apex}} + \Delta RT]$, where ΔRT corresponds to the "Delta Apex RT" parameter (default: 3 frames). The IM interval is defined as $[IM_{\mathrm{apex}} - \Delta IM, IM_{\mathrm{apex}} + \Delta IM]$, where ΔIM is the "Delta Apex IM" parameter (default: 0.02 $1/K_0$, converted to integer IM bins via the global IM resolution). Candidate fragments are collected by scanning the corresponding cells of the MS2 feature index across this

(RT, IM) neighborhood, which avoids exhaustive searching over all MS2 features and yields near-constant lookup time per neighborhood cell.

diaTracer then computes the Pearson correlation coefficient (PCC) between the precursor XIC and the fragment XIC to quantify co-elution consistency. In the implementation, the precursor XIC is smoothed and centered around the precursor apex (±7 frames), whereas the fragment XIC uses the raw fragment intensities within the overlapping scan interval. The fragment trace is normalized by its maximum intensity within the overlap region to make the PCC largely shape-driven rather than scale-driven. Only fragments with PCC greater than the user-specified "Corr threshold" (default: 0.3) are retained as verified fragments for pseudo-MS/MS construction. To control spectrum density and downstream database search runtime, only the top N most intense fragment peaks ("RF max", default: 500) are retained in the final pseudo-MS/MS spectrum. The m/z values written into the pseudo-MS/MS spectrum correspond to fragment feature centers, and fragment intensities are taken from each fragment feature's apex intensity, i.e., the maximum integrated signal observed for that fragment feature during tracing.

Finally, diaTracer outputs pseudo-MS/MS spectra as mzML MS2 scans. For each precursor, a pseudo-MS/MS scan is generated for every assigned precursor charge state; when charge cannot be uniquely determined upstream, multiple charge states are retained and yield multiple otherwise identical pseudo-MS/MS scans differing only in precursor charge annotation. Each scan is annotated with the precursor m/z target, precursor isolation range, precursor apex retention time, and precursor apex ion mobility. The precursor intensity and base peak intensity fields are set using the precursor feature's apex intensity, and the scan spectrum contains the ranked fragment list. All generated scans are serialized into a mzML file for subsequent database searching and downstream analysis.

2.2.4 Integration of diaTracer into FragPipe platform

Following precursor and fragment feature detection, isotope filtering, and precursor–fragment clustering, diaTracer generates pseudo-MS/MS spectra in standard mzML format. These spectra are directly compatible with existing database search engines, including MSFragger[62], without requiring any modification to the search workflow. As a result, spectrum-centric diaPASEF data processed by diaTracer can be analyzed using the same identification, filtering, and validation steps routinely applied to DDA experiments, facilitating direct comparison between acquisition strategies and ensuring broad compatibility with established

proteomics pipelines. diaTracer was implemented as a fully integrated module within the FragPipe computational platform, enabling seamless downstream peptide identification and quantification. Figure 2-2 shows a screenshot of diaTracer in FragPipe.

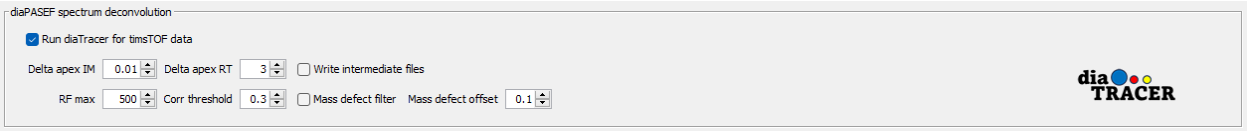


Figure 2-2. Screenshot of diaTracer in FragPipe.

2.2.5 Deep proteome profiling triple-negative breast cancer data analysis

The triple-negative breast cancer (TNBC) dataset from Lapcik et al.[63] contains 12 ddaPASEF runs from a fractionated pooled sample used to generate a DDA-based library and 16 diaPASEF runs from 16 individual TNBC peptide samples. The data were acquired using timsTOF Pro mass spectrometer. We ran FragPipe 22 (pre-release version) with the "DIA_SpecLib_Quant_diaPASEF" workflow to generate a spectral library and perform quantification for the raw .d files. Within diaTracer, "Delta Apex IM" was set to 0.01, "Delta Apex RT" was set to 3, "RF max" was set to 500, and "Corr threshold" was set to 0.3. The mass defect filter was enabled. In the MSFragger database search, the initial precursor and fragment mass tolerances were set to 10 ppm and 20 ppm, respectively. Spectrum deisotoping[64], mass calibration, and parameter optimization[65] were enabled. The isotope error was set to "0/1/2". The reviewed *Homo sapiens* protein sequence database obtained from UniProt (downloaded on November 15, 2023; 20,461 proteins), appended with common contaminants and decoys, was used in the search. Enzyme specificity was set to "stricttrypsin" and the maximum allowed missed cleavages was set to 2. Oxidation of methionine and N-terminal acetylation were set as variable modifications. The maximum number of variable modifications for each peptide was set to 3. The following change was made compared to the default settings of the "DIA_SpecLib_Quant_diaPASEF" workflow: pyro-Glu at peptide N-terminus was added as a variable modification and methylthiolation of cysteine was used as a fixed modification. MSBooster and Percolator were used to predict the RT and MS/MS spectra, and to rescore peptide-spectra matches (PSMs). The final false discovery rate (FDR)-filtered PSMs and the pseudo-MS/MS mzML spectral files (and DDA mzML files in the hybrid workflow) were used by EasyPQP to generate the spectral library. The spectral library was then used with the DIA-NN 1.8.1 quantification module to quantify the library peptide ions in the individual diaPASEF data.

The same settings were employed for the next experiment, unless mentioned otherwise. Spectronaut 18.5 directDIA result and DIA-NN 1.8.1 library-free result were downloaded from the original study. The Spectronaut report file was exported using Spectronaut Viewer 18.5. DIA-NN generated "report.tsv" files were processed using the *iq* R package[66] to count the number of peptides and proteins identified passing the specified confidence thresholds (see Result processing and statistical analysis**)**.

2.2.6 Low-input, spatial proteomics data analysis

We used the dataset from Makhmut et al.[67], which contains 148 microregions from a patient's tonsil sections. In addition, there were 42 diaPASEF runs (from seven different sizes of human tonsil tissue, six replicates each) that were used to build a comprehensive tonsil library. FragPipe with diaTracer results were compared to the DIA-NN results from the original study. Compared to the parameters for the TNBC dataset, carbamidomethylation of cystine is set as a fixed modification. Differential expression analysis was performed using FragPipe-Analyst[68] based on "report.pg_matrix.tsv" and "experiment_annotation.tsv" files generated by FragPipe. The "min percentage of non-missing values globally" and "Min percentage of non-missing values in at least one condition" parameters were set to 0 and 70%, respectively, with all other parameters at default values. "DE Adjusted *p*-value cutoff" was set to 0.05. "DE Log2 fold change cutoff" was set to 1. The imputation type was set to "Perseus-type". The FDR correction type was set to "Benjamini-Hochberg".

Table 2-1 Datasets used for evaluation

	ID	# Runs	DDA library	Instrument
TNBC	PXD047793	16	Yes	timsTOF pro
Low input	PXD042367	148	Yes	timsTOF SCP

2.2.7 Result processing and statistical analysis

To ensure a fair comparison of peptide and protein identification numbers, DIA-NN "report.tsv" files from the diaTracer workflows in FragPipe and from DIA-NN standalone analyses were filtered using the *iq* R package[66] to achieve a 1% FDR at run-specific precursor, global precursor, and global protein levels. Differential expression analyses conducted using FragPipe-Analyst were based on DIA-NN's "report.pg_matrix.tsv" files. Subsequent result processing and plot generation were performed within the RStudio build 402 environment using

the R version 4.3.3 statistical software. The R packages ggplot2, tidyverse, ggrepel, plotly, eulerr, and protti were used in our analysis.

2.3 Results

2.3.1 diaTracer workflow and integration into FragPipe

diaTracer takes diaPASEF raw data (Bruker *.d* files) as input and generates pseudo-MS/MS spectra[52] in mzML format that are compatible with standard DDA-style database searching. The framework is designed to operate directly on four-dimensional diaPASEF data by jointly leveraging m/z, ion mobility, retention time, and intensity information. By incorporating both retention time and ion mobility dimensions during feature detection, diaTracer enhances true signal recovery while suppressing background interference that is difficult to resolve in conventional DIA analyses. Building on our previous observation[58] that aggregating retention time frames with closely aligned ion mobilities improves the fidelity of extracted ion chromatograms (XICs), diaTracer begins by summing intensities across neighboring retention time frames prior to feature detection. This aggregation step increases signal continuity while preserving chromatographic structure, enabling more robust downstream peak detection.

Feature detection is then performed in the m/z–IM space using an adaptive Gaussian-based algorithm applied to the merged frames. This strategy is used consistently for both precursor (MS1) and fragment (MS2) signals, ensuring coherent treatment of all ion features extracted from diaPASEF data. For each detected feature, m/z and ion mobility coordinates are determined based on the apex of the two-dimensional signal distribution. In the MS1 data, isotopic patterns are subsequently grouped to filter precursor features and infer candidate charge states. This reduces redundancy while retaining ambiguity when charge assignment cannot be resolved clearly. To associate fragment ions with their corresponding precursors, diaTracer evaluates Pearson correlation coefficients between precursor and fragment XICs within user-defined ion mobility and retention time tolerances. Fragment features that exceed the correlation threshold are assembled with the precursor into a pseudo-MS/MS spectrum, providing a spectrum-centric representation suitable for downstream peptide identification (see Methods).

diaTracer is fully integrated into the FragPipe computational platform, enabling streamlined, end-to-end analysis of diaPASEF data without the need for external format conversion or custom scripting (Figure 2-3). The pseudo-MS/MS spectra generated by diaTracer

are searched using MSFragger[62] in DDA mode, supporting a wide range of search strategies, including fully tryptic, semi-tryptic, nonspecific, mass-offset, and open modification searches[65,69,70]. This flexibility allows comprehensive interrogation of post-translational modifications and noncanonical peptide species that are difficult to capture using peptide-centric DIA approaches. Peptide spectrum matches are further refined through deep learning-based rescoring with MSBooster[71], followed by statistical validation using Percolator[72] by default, with PeptideProphet[73] available as an alternative. Protein inference is performed using ProteinProphet[74], and false discovery rate control is applied at multiple levels using Philosopher[75]. For workflows that require confident site localization, such as phosphoproteomics, PTMProphet[76] is enabled within the same pipeline.

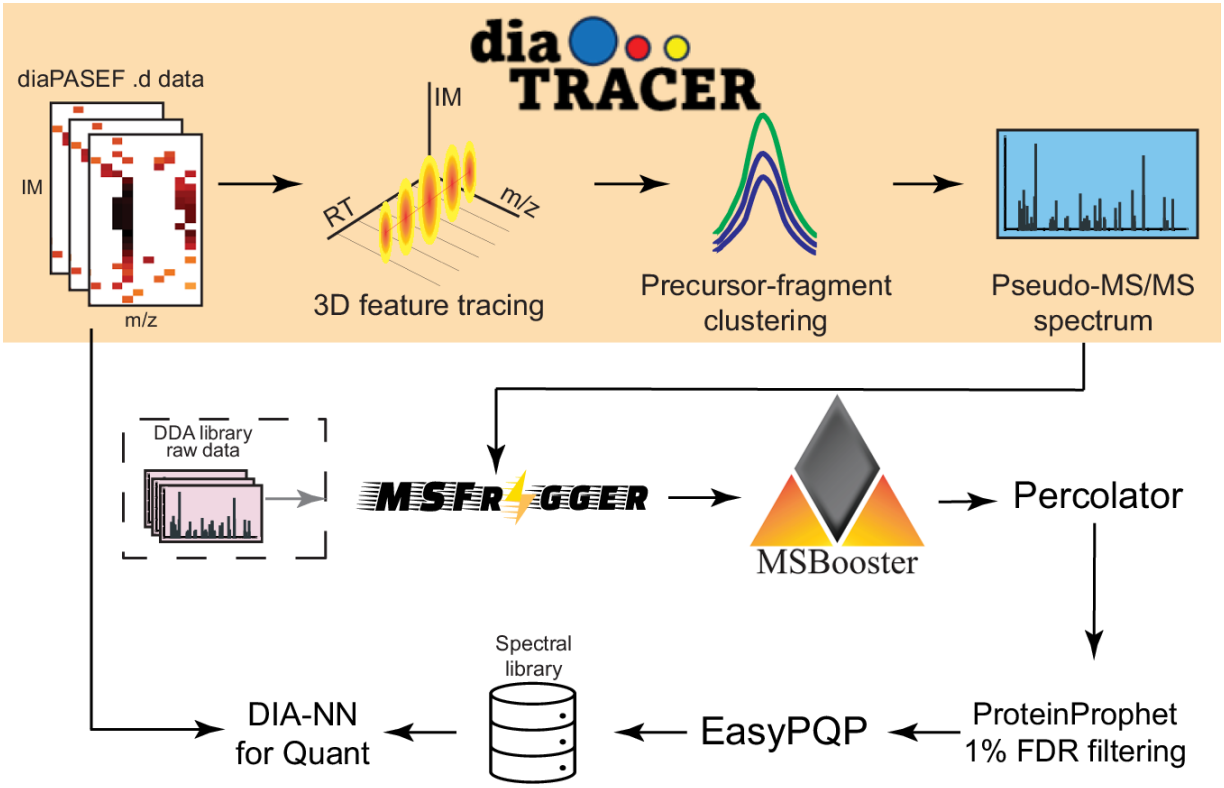


Figure 2-3. Overview of diaTracer and the FragPipe computational platform.

diaTracer applies a 3D feature detection algorithm to detect signals from all possible precursors and fragments in MS1 and MS2 diaPASEF data. Pseudo-MS/MS spectra are generated through precursor-fragment clustering and can be processed as DDA data using MSFragger and FragPipe to build a spectral library directly from the data. A hybrid spectral library can also be generated if DDA data are available. This spectral library is then used to extract quantification using DIA-NN. *Reproduced from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Beyond identification, diaTracer-derived pseudo-MS/MS spectra can be used to construct high-quality spectral libraries using EasyPQP, which are subsequently employed for peptide-

level quantification directly from DIA data using DIA-NN[51,77], with Skyline[78] available as an alternative. When ddaPASEF data are available, hybrid spectral libraries combining diaPASEF- and ddaPASEF-derived spectra can also be generated, further increasing identification depth and quantitative robustness. FragPipe supports both an intuitive graphical user interface and a command-line interface suitable for high-performance computing and cloud-based environments. Visualization of pseudo-MS/MS spectra and peptide-spectrum matches is provided through FragPipe-PDV[79], while downstream statistical and comparative analyses of quantitative matrices can be conveniently performed using FragPipe-Analyst[68] or MSstats[80]. diaTracer is also available for standalone use via a command-line interface. Detailed documentation and descriptions of user-defined parameters are provided in Appendix A.

2.3.2 Performance evaluation using a TNBC dataset

We first evaluated the performance of FragPipe with diaTracer using a dataset representative of deep proteome profiling experiments. We analyzed a publicly available triple-negative breast cancer (TNBC) dataset[63] generated on a Bruker timsTOF Pro platform, which consists of two complementary components: (i) ddaPASEF data acquired from 12 hydrophilic interaction liquid chromatography (HILIC) peptide fractions of a pooled TNBC sample, and (ii) 16 diaPASEF runs acquired from individual tissue lysate samples. In the original study, these data were processed using Spectronaut version 18.5 under multiple strategies, including direct DIA, a DDA-based spectral library, and a hybrid DDA plus direct DIA library. In addition, the diaPASEF data were analyzed using DIA-NN version 1.8.1 in library-free (direct DIA) mode.

Here, we reanalyzed the same dataset using FragPipe with diaTracer. Two FragPipe-based analysis strategies were evaluated: (i) a direct DIA approach using only the 16 diaPASEF runs processed with diaTracer (denoted as “FragPipe” in Figure 2-4), and (ii) a hybrid strategy that combined pseudo-MS/MS spectra derived from the 16 diaPASEF runs with spectra generated from the 12 fractionated ddaPASEF runs (“FragPipe hybrid”). For comparison, the Spectronaut 18.5 results (exported “.sne” files) and DIA-NN 1.8.1 library-free results (exported “report.tsv” files) were obtained directly from the original study.

Across all tools, protein quantification results were filtered using identical criteria: a global protein, global precursor, and run-specific precursor Q-value threshold of 0.01. When considering the 16 diaPASEF runs alone (direct DIA), all methods yielded comparable protein coverage. Under these conditions, FragPipe with diaTracer quantified an average of 9,296

proteins per diaPASEF run (Figure 2-4a), corresponding to 10,341 proteins in total across all runs (Figure 2-4b). In comparison, Spectronaut directDIA quantified fewer proteins, with an average of 8,997 proteins per run and 10,032 proteins in total. DIA-NN library-free analysis achieved slightly higher coverage, quantifying an average of 9,520 proteins per run and 10,628 proteins in total.

As expected, protein coverage increased when ddaPASEF data were incorporated into hybrid library-based workflows. Using the hybrid DDA/DIA strategy, FragPipe again outperformed Spectronaut, quantifying a larger total number of proteins (11,029 vs. 10,552) as well as more proteins per run on average (9,653 vs. 9,021). Notably, FragPipe with the hybrid strategy also quantified the highest number of proteins with complete data (no missing values) or with less than 50% missing values across individual runs (Figure 2-4b), highlighting improved quantitative consistency in addition to increased proteome depth.

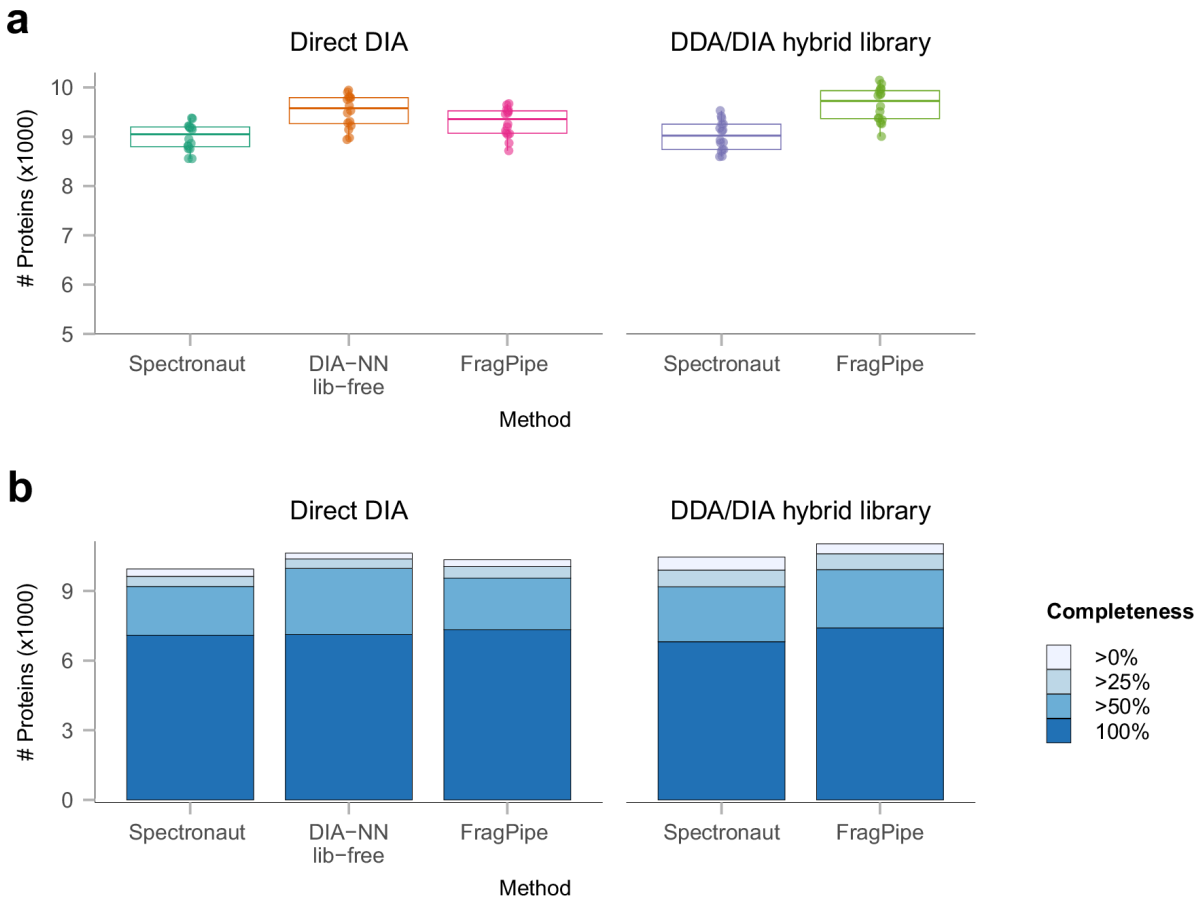


Figure 2-4. Deep proteome profiling using TNBC dataset.

(a) Box plot showing the numbers of quantified proteins across 16 diaPASEF runs from 16 individual TNBC peptide samples using different methods. The lower and upper edges of the box represent the first (Q1) and the third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. The central line represents the median

of the numbers. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. (b) Histogram showing the number of quantified proteins in the TNBC dataset using Spectronaut 18.5 and FragPipe with diaTracer, direct DIA and hybrid DDA/DIA analysis, and using DIA-NN in library-free mode, after application of different non-missing value filters. Shades of blue represent data completeness; darker blues indicate presence in a greater number of samples. *Reproduced from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

2.3.3 Performance evaluation using a low-input, spatial proteomics dataset

We next evaluated the performance of diaTracer using low-input data from a recent spatial proteomics study[67]. In this study, 148 microregions were profiled from four anatomically and biologically distinct regions of a human tonsil—epithelium, germinal center, mantle zone, and T-cell zone—using laser capture microdissection (LCM). Because of the limited protein amounts available from individual microregions, the original study adopted a two-step strategy. First, a comprehensive human tonsil spectral library was generated using 42 diaPASEF runs acquired from samples with high protein input (referred to as "high-input" samples) and processed in the library-free mode using DIA-NN. This high-input spectral library was then applied to quantify proteins in the 148 low-input samples. Notably, the authors did not attempt a direct analysis of the low-input samples without relying on the high-input spectral library.

To enable a direct comparison with the original study, we performed two analyses using diaTracer integrated into FragPipe. In the first analysis, diaTracer was used to generate pseudo-MS/MS spectra from the 46 high-input diaPASEF runs, which were then used to construct a spectral library ("FragPipe high-input Lib"), analogous to the DIA-NN–based workflow in the original study. In the second analysis, we performed a direct, library-free analysis of the 148 low-input samples alone ("FragPipe"), without using any high-input data.

When using the high-input data to construct the spectral library, the diaTracer–FragPipe workflow achieved protein identification numbers that were highly similar to those reported for DIA-NN library-free analysis in the original study (Figure 2-5a; Methods). As expected, leveraging high-input data resulted in increased protein coverage, with 3,349 proteins identified in total (1,992 proteins per sample), compared to 1,602 proteins (1,303 per sample) identified by direct analysis of the low-input samples alone.

However, total protein counts can be misleading in the context of large-cohort spatial proteomics studies, as proteins detected in only a small number of samples contribute limited biological insight. Consistent with the original study, downstream analyses were therefore performed after applying a missing-value filter requiring at least 70% non-missing values across

all samples in at least one of the four anatomical groups. Under this criterion, the use of high-input data yielded 2,175 proteins (1,692 per sample), whereas direct analysis of the low-input samples alone quantified 1,475 proteins (1,265 per sample). Increasing the stringency of the missing-value filter further reduced the relative advantage of the high-input library. When requiring 90% non-missing values, 1,652 proteins (1,398 per sample) and 1,336 proteins (1,182 per sample) were quantified with and without the use of high-input data, respectively. Figure 2-5b shows the overlap between proteins quantified with at least 70% non-missing values in at least one group across the original DIA-NN–based analysis and the diaTracer–FragPipe workflows with and without high-input data. All approaches exhibited a high degree of overlap, indicating that the core proteome was consistently captured. Importantly, the biological conclusions derived from the data were highly consistent across all analyses. Even when relying solely on direct identification from low-input samples, the quantified proteins segregated clearly into the four expected cell types following the same filtering strategy used in the original study (Figure 2-5c).

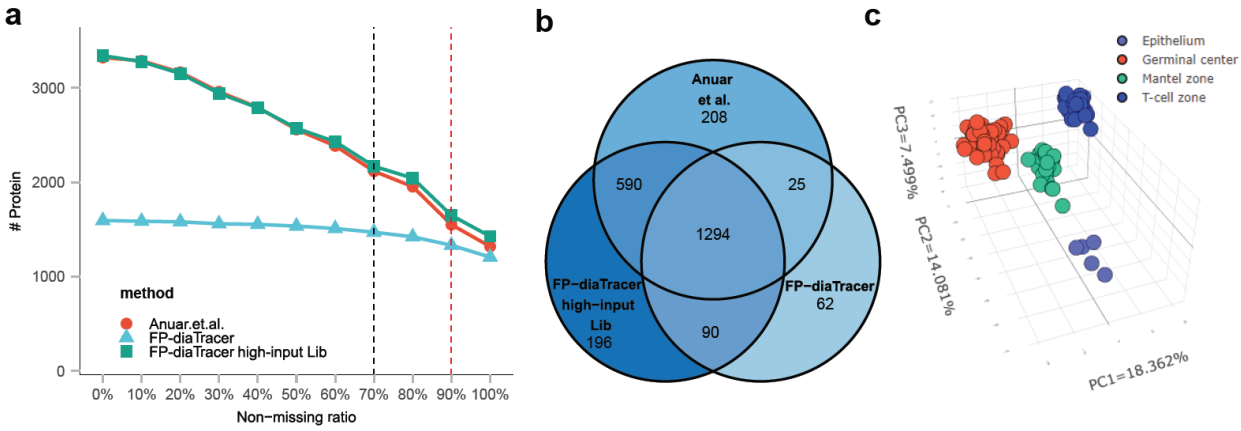


Figure 2-5. Low-input, spatial proteomics data comparison.

(a) Number of quantified proteins after application of non-missing value (in at least one group) filter ranging from 0% to 100%, with line colors representing different methods. Red: results from the original study based on the library built using high-input samples; Green: results based on FragPipe with diaTracer, also using high-input data to build the library ("FragPipe high-input Lib"); Blue: results using diaTracer and FragPipe with low-input data only ("FragPipe"). (b) Venn diagram of quantified proteins between the three methods, with data filtered to keep proteins with at least 70% non-missing values in at least one group. (c) Principal-component analysis (PCA) plot of 148 samples based on 1471 proteins (after missing value filtering and data imputation) quantified using the FragPipe workflow (using low-input data only). *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

We further examined established cell-type markers reported in the original publication, including CD19 (B-cell marker), CD3D (T-cell marker), and CDH1 (epithelial marker). All

markers displayed the expected expression patterns across the corresponding microregions (Figure 2-6a). In addition, a global comparison between mantle zone and T-cell regions revealed statistically significant proteomic differences (Figure 2-6b). Differentially expressed protein lists derived from analyses performed with and without high-input data produced highly similar pathway enrichment results when analyzed using FragPipe-Analyst (Appendix Figure B-1 to 6).

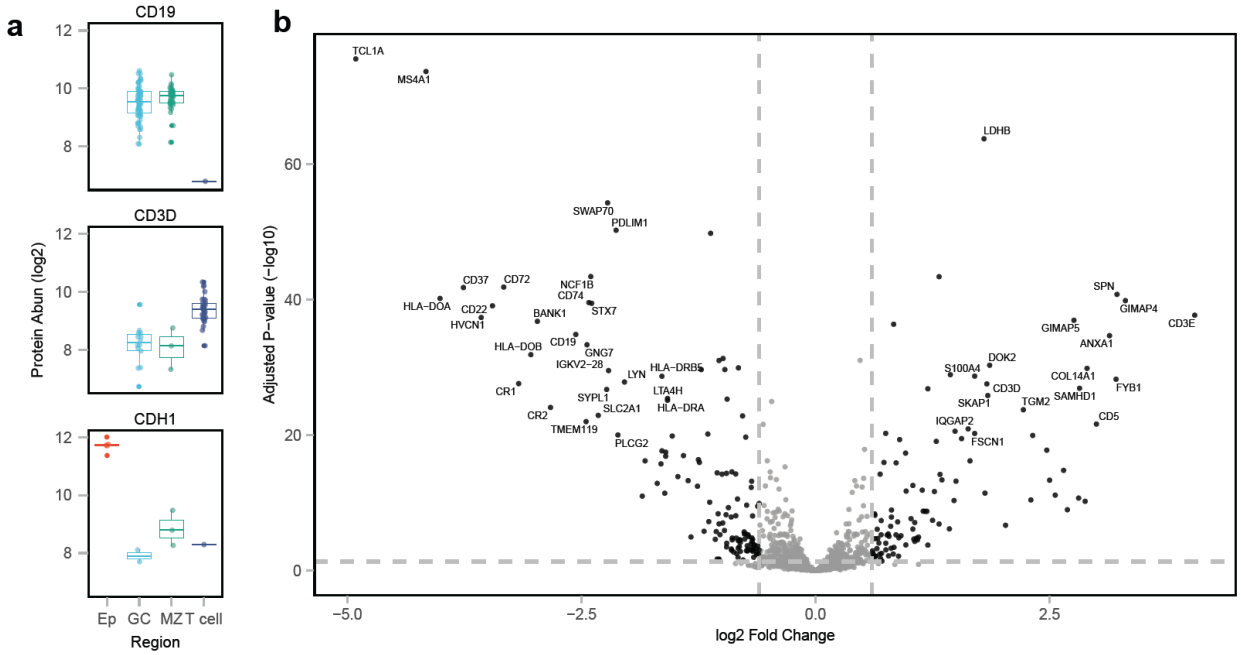


Figure 2-6. Low-input, spatial proteomics data biology analysis.

(a) Log2 transformed protein level abundance distribution of selected cell-type-specific proteins in different regions. In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. (b) Volcano plot showing protein abundance differences between Mantle zone (28 samples) and T-cell zone (34 samples), highlighting tissue-specific proteins (Log2 fold change ≥ 1; adjusted p-value ≤ 0.05). The adjusted p-value is from the moderated t-test followed by the Benjamini-Hochberg procedure. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Taken together, these results demonstrate that FragPipe with diaTracer enables robust identification and quantification of proteins from large cohorts of low-input diaPASEF samples, even in the absence of high-input data. While the availability of high-input samples remains advantageous for maximizing proteome depth, direct spectrum-centric analysis of low-input data alone is sufficient to recover biologically meaningful and reproducible proteomic patterns, particularly when focusing on proteins consistently quantified across a substantial fraction of the cohort.

2.4 Discussion

In this chapter, we presented diaTracer, a computational method and software framework for spectrum-centric analysis of diaPASEF data. diaTracer deconvolves four-dimensional diaPASEF data to generate DDA-like pseudo-MS/MS spectra, which can be searched using database search engines developed for DDA workflows, such as MSFragger. diaTracer is available both as a standalone tool and as a fully integrated component of FragPipe, enabling an end-to-end diaPASEF analysis workflow that includes peptide identification, deep-learning-based rescoring, protein inference, FDR control, post-translational modification (PTM) site localization, spectral library generation, DIA-based quantification, and interactive data visualization.

Across the evaluated datasets, diaTracer integrated within FragPipe demonstrated performance that was comparable to, and in several scenarios competitive with, established diaPASEF analysis tools, including Spectronaut and DIA-NN. In deep proteome profiling experiments, diaTracer achieved similar or higher numbers of protein identifications relative to library-free DIA workflows, while maintaining favorable missing-value characteristics, particularly when hybrid DDA/DIA libraries were used. Notably, diaTracer exhibited strong performance in low-input and spatial proteomics datasets, where direct analysis without reliance on large external spectral libraries still yielded biologically consistent results. These findings indicate that spectrum-centric deconvolution of diaPASEF data, as implemented in diaTracer, can match the sensitivity of state-of-the-art peptide-centric DIA methods while offering greater flexibility in database search strategies.

As with other spectrum-centric DIA approaches, including DIA-Umpire, diaTracer has inherent limitations. First, diaTracer relies on MS1 feature detection as the starting point for fragment grouping and pseudo-MS/MS spectrum construction. Consequently, precursor ions that are extremely weak or absent in MS1 may not be detected, limiting downstream identification. To mitigate this issue, diaTracer aggregates neighboring retention time frames to amplify precursor signals and suppress noise, which substantially improves sensitivity for low-abundance features. Nevertheless, some precursors that are detectable only at the fragment level may remain inaccessible to a strictly MS1-anchored strategy.

Second, diaTracer extracts rich, multi-dimensional feature information, including precursor and fragment m/z values, ion mobility ranges, retention time profiles, and correlation

metrics between precursor and fragment XICs. In the current implementation, this information is primarily used during pseudo-MS/MS spectrum construction. Incorporating this additional information directly into database search scoring and post-search rescoring—similar to strategies employed in tools such as MSFragger-DIA—could further improve discrimination between correct and incorrect peptide-spectrum matches. Extending diaTracer to propagate feature-level metadata into downstream scoring models represents a promising direction for future development.

In summary, diaTracer provides a practical and efficient spectrum-centric framework for diaPASEF data analysis that leverages the additional ion mobility dimension to enhance signal detection and reduce interference. By integrating seamlessly into FragPipe, diaTracer enables flexible, reusable, and scalable workflows that bridge DIA acquisition with mature DDA-based identification tools. Future improvements aimed at deeper integration of multi-dimensional feature information into identification and scoring are expected to further enhance performance and broaden the applicability of spectrum-centric analysis for complex diaPASEF datasets.

2.5 Data availability

The raw MS/MS files used in this study can be found through the ProteomeXchange Consortium via the PRIDE partner repository[81] or at the MassIVE repository with the following accession codes:

- Triple Negative Breast Cancer (TNBC) dataset PXD047793: https://www.ebi.ac.uk/pride/archive/projects/PXD047793
- Low-input, spatial proteomics data PXD042367: https://www.ebi.ac.uk/pride/archive/projects/PXD042367
- The diaTracer converted mzML files and FragPipe results generated in this study have been deposited in the MassIVE repository with the identifier MSV000094803: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=85c80cf99442470e9e2aa01830c88120

2.6 Acknowledgements and competing interests

This work was supported in part by National Institutes of Health grants R01-GM-094231 and U24-CA271037.

A.I.N. and F.Y. receive royalties from the University of Michigan for the sale of MSFragger, IonQuant, and diaTracer software licenses to commercial entities. K.L. receives royalties from the University of Michigan for the sale of diaTracer software licenses to commercial entities. All license transactions are managed by the University of Michigan Innovation Partnerships office, and all proceeds are subject to university technology transfer policy. Other authors declare no other competing interests.

2.7 Authors, affiliations, and contributions

Kai Li[1], Guo Ci Teo[2], Kevin L. Yang[1], Fengchao Yu[2] & Alexey I. Nesvizhskii[1,2]

[1]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA

K.L. and A.I.N. developed the diaTracer algorithm. K.L. wrote the software and analyzed the results. F.Y. and G.C.T. assisted with the algorithm and software development. K.L.Y. helped modify MSBooster. F.Y. assisted with the integration of diaTracer into FragPipe. K.L., F.Y., and A.I.N. wrote the manuscript. A.I.N. conceived the study. A.I.N. and F.Y. supervised the study.

Chapter 3 Spectrum-Centric Analysis of Complex diaPASEF Data Using diaTracer

The content of this chapter is adapted from the previous publication by the author in Nature Communications[55].

3.1 Introduction

In Chapter 2, we introduced diaTracer, a spectrum-centric computational framework for the analysis of diaPASEF data and described its integration into the widely used proteomics computational platform FragPipe. We benchmarked diaTracer against existing diaPASEF analysis tools and demonstrated that it achieves competitive peptide and protein identification performance while maintaining high computational efficiency.

As discussed in the Chapter 1, direct peptide identification from data-independent acquisition (DIA) data can also be performed using peptide-centric analysis strategies, including the library-free mode of DIA-NN[51], as well as hybrid approaches such as MSFragger-DIA[57], in which full DIA tandem mass spectrometry (MS/MS) scans are searched against a protein sequence database followed by targeted precursor and fragment peak tracing and rescoring. Both spectrum-centric and peptide-centric strategies benefit from deep learning–based predictions of peptide fragmentation patterns and separation coordinates (retention time and ion mobility), but they differ substantially in how these prediction models are applied. In fully spectrum-centric approaches, as well as in hybrid strategies such as MSFragger-DIA, predictions are generated only for a restricted set of top-scoring peptide candidates derived directly from the data. In contrast, peptide-centric methods such as DIA-NN require in-silico predictions for all possible candidate peptides defined by the protein sequence database and search parameters, leading to a rapid expansion of the search space as analytical complexity increases.

A key advantage of spectrum-centric strategies, such as DIA-Umpire[52] and diaTracer, is their ability to deconvolute DIA data into data-dependent acquisition (DDA)-like pseudo-MS/MS spectra, which enables efficient analysis of datasets requiring large or poorly constrained search spaces. These include studies focusing on post-translational modifications (PTMs),

analyses involving nonspecific digestion (e.g., immunopeptidomics or endogenous human leukocyte antigen (HLA) peptides analysis), semi-enzymatic searches (e.g., N-terminomics), and applications where peptide spectra or retention times are difficult to predict accurately in advance, such as chemoproteomics. In such settings, peptide-centric approaches can become computationally prohibitive or suffer from reduced specificity due to extensive a priori hypothesis generation.

In this chapter, we demonstrate the unique capabilities of diaTracer for analyzing complex diaPASEF datasets using a spectrum-centric strategy. We first focus on the analysis of semi-tryptic peptides in cerebrospinal fluid (CSF) and plasma proteome datasets. Semi-tryptic peptides are of particular interest because they capture endogenous proteolytic processing events that are not accessible through standard tryptic workflows[82,83]. Such peptides can reveal biologically meaningful cleavage patterns associated with aging, neurodegeneration (e.g., Alzheimer's disease)[84], and other pathological processes, providing insights beyond those obtained from protein-level abundance measurements alone[85]. Despite their biological relevance, semi-tryptic searches on diaPASEF data are extremely challenging without a spectrum-centric deconvolution strategy such as that implemented in diaTracer.

In addition to semi-tryptic analysis, we show that diaTracer enables unrestricted PTM discovery through the application of MSFragger's open or mass-offset search modes[62] to diaTracer-generated pseudo-MS/MS spectra. Finally, we demonstrate the application of diaTracer to phosphoproteomics and immunopeptidomics datasets, both of which impose substantial demands on search-space flexibility and spectral interpretability. Together, these examples illustrate how diaTracer extends diaPASEF data analysis beyond conventional workflows and enables a broad range of complex proteomics applications that are difficult or impractical to address using peptide-centric DIA approaches alone.

3.2 Methods

3.2.1 Cerebrospinal fluid data analysis using diaTracer and FragPipe

The cerebrospinal fluid (CSF) dataset from Mun et al.[86] contains 24 ddaPASEF runs used to generate a DDA-based library and 34 diaPASEF runs from 15 patients with Alzheimer's disease (AD) and 19 control subjects. We ran FragPipe with diaTracer using the built-in "DIA_SpecLib_Quant_diaPASEF" workflow. Within diaTracer, "Delta Apex IM" was set to

0.01, “Delta Apex RT” was set to 3, “RF max” was set to 500, and “Corr threshold” was set to 0.3. The mass defect filter was enabled. In the MSFragger database search, the initial precursor and fragment mass tolerances were set to 10 ppm and 20 ppm, respectively. Spectrum deisotoping[64], mass calibration, and parameter optimization[65] were enabled. The isotope error was set to “0/1/2”. The reviewed *Homo sapiens* protein sequence database obtained from UniProt (downloaded on November 15, 2023; 20,461 proteins), appended with common contaminants and decoys, was used in the search. Enzyme specificity was set to “stricttrypsin” and the maximum allowed missed cleavages were set to 2. Carbamidomethylation of cystine was used as a fixed modification. Oxidation of methionine and N-terminal acetylation were set as variable modifications. The maximum number of variable modifications for each peptide was set to 3. MSBooster and Percolator were used to predict the RT and MS/MS spectra, and to rescore peptide-spectrum matches (PSMs). The final FDR-filtered PSMs and the pseudo-MS/MS mzML spectral files (and DDA mzML files in the hybrid workflow) were used by EasyPQP to generate the spectral library. The spectral library was then used with the DIA-NN (version 1.8.1) quantification module to quantify the library peptide ions in the individual diaPASEF data. For semi-tryptic searches, the cleavage parameter was adjusted to “SEMI”. The peptide length was set from 7 to 50 with a peptide mass range of 500–5000 Da. For comparison, DIA-NN (version 1.8.1) as a standalone tool was run in the library-free mode using default settings unless otherwise specified. The precursor m/z range was set to 300–1800. Methionine oxidation was set as a variable modification. Carbamidomethylation of cysteine was used as a fixed modification. A maximum of 1 missed cleavage was allowed. An in-silico predicted spectral library was generated from the same database file as described above except without adding decoys. Quantification results from FragPipe and DIA-NN 1.8.1 were processed using the *iq* R package[66] to count the number of peptides and proteins identified passing the specified confidence thresholds.

To conduct comprehensive PTM searches using the mass-offset and open search mode of MSFragger, FragPipe’s “Mass-Offset-CommonPTMs” and “Open” workflows were used under default settings. In brief, in the mass-offset search, MSFragger takes a list of modification masses as mass offsets and searches the spectrum against peptides within a narrow mass tolerance window placed around each specified mass offset. This approach can be used to sensitively and rapidly detect hundreds of known modifications. In the open search mode,

MSFragger searches the spectrum against the peptides with a wide mass tolerance window, by default from -150 to 500 Da. It is an efficient approach for detecting peptides with any mass shift, including unknown modifications.

3.2.2 Plasma data analysis

The plasma dataset analyzed in this study was obtained from Vitko et al.[87] and consisted of 40 diaPASEF runs acquired on a timsTOF HT mass spectrometer. The dataset included plasma samples from 20 patients diagnosed with stage IV non-small cell lung cancer (NSCLC) and 20 matched control samples. All samples were processed using the Seer Proteograph nanoparticle conjugation kit NP2. Raw diaPASEF data were processed using the diaTracer–FragPipe workflow as described for the CSF dataset. Briefly, diaTracer was used to deconvolute diaPASEF data into pseudo-MS/MS spectra using a spectrum-centric strategy, followed by peptide identification with MSFragger and downstream processing within the FragPipe framework. Semi-tryptic searches were performed to enable detection of endogenous proteolytic peptides that are not accessible using fully tryptic digestion rules. Identified peptides were filtered using standard false discovery rate (FDR) thresholds, and protein-level quantification matrices were generated for downstream statistical analysis.

Differential expression analysis was performed using FragPipe-Analyst[68]. When loading the data, the "data type" parameter was set to DIA, and both the "report.pg_matrix.tsv" file containing protein group–level quantification values and the corresponding "experiment_annotation.tsv" file describing sample group assignments were uploaded to the FragPipe-Analyst web interface. To balance proteome coverage and data completeness in this heterogeneous plasma dataset, the minimum percentage of non-missing values was set to 25% globally and 25% in at least one experimental condition, while all other filtering parameters were retained at their default values. These thresholds were selected to preserve low-abundance but biologically relevant signals while still removing highly sparse protein entries that could confound statistical testing.

Differential expression testing was performed using a moderated statistical framework implemented in FragPipe-Analyst, with p-values adjusted for multiple hypothesis testing using the Benjamini–Hochberg procedure. Proteins with an adjusted p-value below 0.05 and an absolute $\log_2$ fold change greater than 1 were considered significantly differentially expressed. Missing values were imputed using the Perseus-type imputation strategy to enable robust

statistical comparisons between groups. The resulting “DE_result.tsv” file was downloaded and used for all downstream analyses, including visualization, pathway enrichment, and biological interpretation.

3.2.3 Phosphoproteomics data analysis

The phosphorylation-enriched diaPASEF dataset from Oliinyk et al.[88] contained runs acquired using six chromatographic gradient lengths ranging from 7 to 60 min, with each gradient condition analyzed in four technical replicates. This dataset provides a systematic benchmark for evaluating phosphoproteomics performance under varying chromatographic separation conditions.

To analyze these data using FragPipe with diaTracer, we employed the built-in “DIA_SpecLib_Quant_Phospho_diaPASEF” workflow. This workflow extends the standard “DIA_SpecLib_Quant_diaPASEF” pipeline by enabling phosphorylation on serine, threonine, and tyrosine (STY) residues as a variable modification during database searching with MSFragger. In addition, site localization was performed using PTMProphet, allowing confident assignment of phosphorylation sites within identified peptides. Pseudo-MS/MS spectra generated by diaTracer were searched using MSFragger in DDA mode, followed by standard FragPipe processing, including PSM validation, protein inference, and FDR filtering. FragPipe automatically propagated phosphorylation site localization probabilities from the “psm.tsv” files to the quantification outputs generated by DIA-NN, including both “report.tsv” and “report_pr_matrix.tsv”. This ensured that quantitative analyses were directly linked to confidently localized phosphorylation sites.

Phosphopeptide counts were defined as the number of non-redundant phosphorylated peptide sequences passing the global FDR thresholds. When reporting the number of phosphorylation sites, only sites with a PTMProphet localization probability greater than 0.75 were considered confidently localized and included in downstream analyses. To assess quantitative reproducibility, intensity correlations were computed using the R package protti. These correlations were calculated using phosphorylated peptides with localization probabilities greater than 0.75 that were consistently quantified across all four technical replicates of the 7 min gradient experiment.

For qualitative assessment and data inspection, selected PSMs and their corresponding pseudo-MS/MS spectra were visualized using FragPipe-PDV[79], while extracted ion chromatograms (XICs) for representative phosphopeptides were examined using Skyline[78].

3.2.4 HLA immunopeptidomics data analysis

For immunopeptidomics analysis, we used a subset of the diaPASEF human leukocyte antigen (HLA) dataset from Wahle et al.[89] consisting of triplicate measurements from a single donor. We ran FragPipe with diaTracer using the built-in "Nonspecific-HLA-diaPASEF" workflow. Relative to the standard "DIA_SpecLib_Quant_diaPASEF" workflow, the enzyme specificity was set to nonspecific, and the peptide length was restricted to 7–25 amino acids. Carbamidomethylation of cysteine was not specified, consistent with the lack of alkylation in standard HLA sample preparation protocols, while cysteinylation (+119 Da) was included as a variable modification to account for common post-isolation cysteine modifications observed in immunopeptidomics datasets. The isotope error parameter was set to 0/1 to allow for limited precursor mass deviations while maintaining stringent mass accuracy.

Pseudo-MS/MS spectra generated by diaTracer were searched using MSFragger, followed by standard FragPipe processing, including PSM validation and FDR filtering. Identified immunopeptides were subsequently analyzed using NetMHCpan 4.1[90] to predict peptide–major histocompatibility complex (MHC) binding affinities. Only peptides with lengths between 8 and 12 amino acids were considered for binding predictions, as these lengths are most relevant for MHC class I presentation. The six HLA class I alleles corresponding to the donor were provided as input to ensure allele-specific binding predictions. Peptides with percentile ranks below 0.5% were classified as strong binders, while those with ranks below 2% were considered weak binders, following commonly used conventions in immunopeptidomics studies.

For data inspection and validation, PSMs and pseudo-MS/MS spectra were visualized using FragPipe-PDV, which is integrated into FragPipe. In addition, predicted fragment ion intensities and spectral entropy-based similarity scores were generated using MSBooster[71].

Table 3-1 Datasets used for evaluation

	ID	#Runs	DDA library	Instrument
CSF	PXD035249	35	Yes	timsTOF HT
Plasma	PXD047839	40	No	timsTOF HT

Phosphoproteome	PXD033904	24	No	timsTOF HT
HLA	MSV000092557	3	Yes	timsTOF Ultra

3.2.5 Runtime comparison

The runtimes for FragPipe with diaTracer and for the DIA-NN library-free mode analyses were measured on a Linux desktop with an Intel Core i9-13900K CPU (32 logical cores) and 128GB memory. All analyses were configured to utilize all 32 CPUs.

3.3 Results

3.3.1 Comprehensive analysis of cerebrospinal fluid data

To demonstrate the ability of diaTracer to support complex diaPASEF data analyses that extend beyond conventional tryptic workflows, we analyzed a cerebrospinal fluid (CSF) dataset from Mun et al[86]. This dataset comprised 24 fractionated ddaPASEF runs and 34 diaPASEF runs acquired from human CSF samples. CSF represents a biologically informative specimen type, characterized by low protein abundance, a high dynamic range, and extensive endogenous proteolytic processing[83,84]. These properties make CSF particularly well suited for evaluating spectrum-centric DIA strategies that do not rely on restrictive peptide hypotheses.

Using FragPipe with diaTracer, we quantified the 34 diaPASEF samples under three complementary analysis strategies: (i) direct diaPASEF analysis using only the 34 diaPASEF runs, (ii) DIA analysis using a spectral library constructed from the 24 ddaPASEF fractionated runs, and (iii) a hybrid strategy combining both ddaPASEF and diaPASEF data. For reference, we also considered results generated using DIA-NN (v1.8.1) in the library-free mode, as reported in the original study. All results were filtered using consistent global and run-specific FDR criteria to ensure comparability.

Under a standard tryptic search configuration, direct diaPASEF analysis with diaTracer quantified an average of 955 proteins per run, corresponding to 1,023 proteins in total across all samples. DIA-NN library-free analysis quantified fewer proteins per run on average, despite reporting a slightly higher cumulative protein count across the dataset (Figure 3-1a, Appendix Figure C-1). This discrepancy reflects the well-known tradeoff between per-run depth and aggregate identifications in DIA analyses[91], particularly in heterogeneous biofluid samples.

As expected, the use of ddaPASEF-derived spectral libraries improved overall protein-level coverage. The DDA-based library resulted in a 23% increase in the number of proteins

quantified per run relative to direct diaPASEF analysis. Incorporating both ddaPASEF and diaPASEF data into a hybrid library further increased coverage, yielding the highest total number of proteins and precursors across all workflows. These results confirm that diaTracer integrates seamlessly with established library-based DIA strategies while preserving the flexibility of direct analysis.

Interestingly, despite lower protein-level coverage relative to library-based workflows, direct diaPASEF analysis with diaTracer quantified a substantially larger number of peptide precursors per run than the DDA-based library strategy (Figure 3-1b). This observation highlights an important distinction between protein-centric and peptide-centric performance metrics: spectrum-centric deconvolution recovers a richer precursor-level landscape, including peptides that may be absent from or underrepresented in DDA-derived libraries. This property is particularly advantageous in CSF, where extensive proteolysis generates diverse peptide species that are difficult to capture comprehensively using DDA alone.

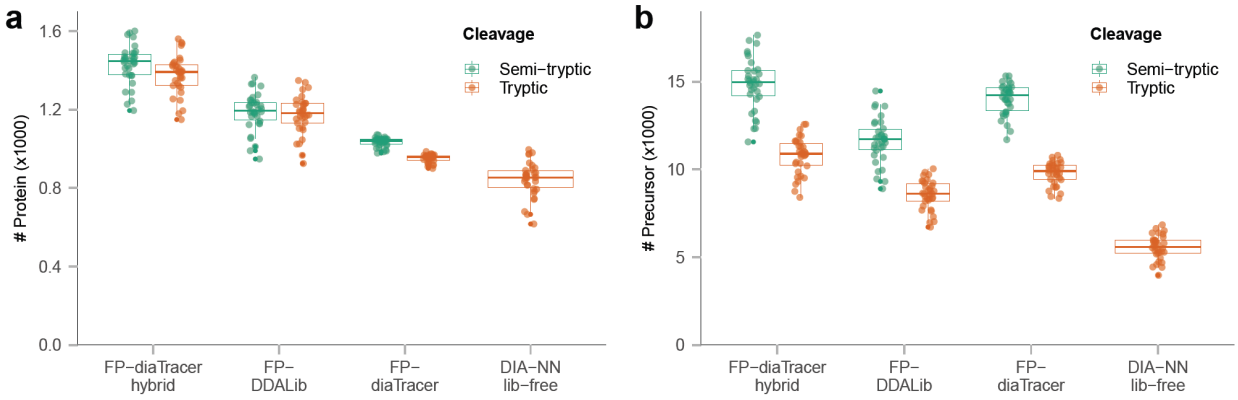


Figure 3-1. CSF data result comparison.

(a) Box plot showing the numbers of quantified proteins using different methods, with colors representing cleavage types (green: trypsin cleavage; orange: allowing semi-tryptic peptides). Each dot represents the number reported for each of the 34 diaPASEF runs from 15 patients with Alzheimer's disease (AD) and 19 control subjects. In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. (b) Box plot showing the numbers of quantified precursors of 34 diaPASEF runs using different methods. The box plots' median, edges, and whiskers are same as those in (a). *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

To further explore this expanded precursor space, we performed semi-tryptic searches on the diaTracer-extracted pseudo-MS/MS spectra. While semi-tryptic searches did not markedly increase the number of quantified proteins per run, they resulted in a dramatic expansion of the

detectable precursor space. Specifically, the average number of quantified precursors per run increased by approximately 39% relative to tryptic searches in the direct diaPASEF workflow (e.g., 9771 and 13,997 precursors per file, on average, in tryptic and semi-tryptic searches for the direct diaPASEF analysis workflow, respectively). This increase reflects the recovery of biologically meaningful endogenous cleavage products beyond canonical tryptic digestion rules.

A further analysis of the additional semi-tryptic peptides revealed a strong enrichment for immunoglobulin chains and secreted proteins, consistent with known properties of CSF[92,93]. Notably, for 158 proteins, we identified semi-tryptic peptides mapping immediately downstream of annotated signal peptide cleavage sites (Appendix Table C-1). These peptides arise from physiological protein maturation rather than nonspecific degradation and are therefore invisible to standard tryptic workflows. A representative example is hemopexin (HPX) shown in Figure 3-2, for which removal of the signal peptide generates mature N-terminal peptides that do not conform to trypsin cleavage specificity. Multiple semi-tryptic peptides derived from HPX and other secreted proteins were consistently detected, providing direct evidence of endogenous proteolytic processing in CSF samples.

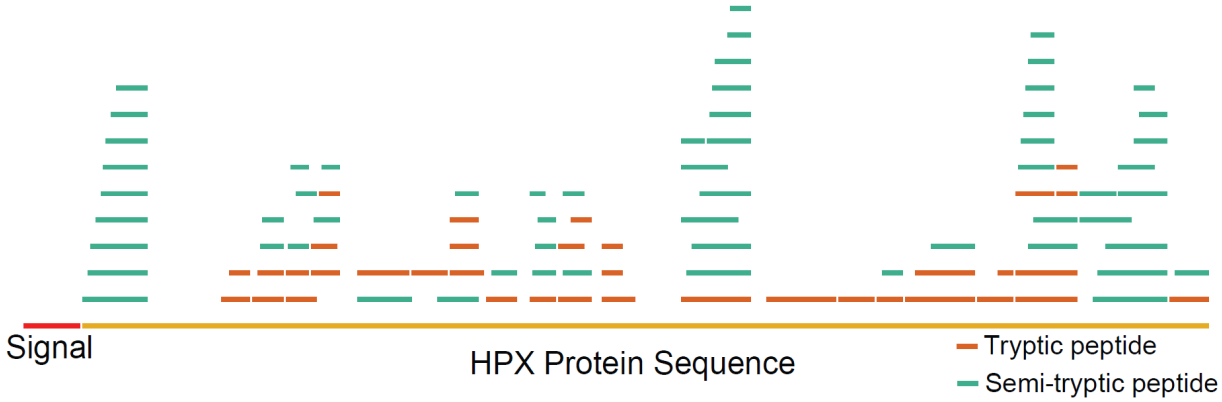


Figure 3-2. HPX protein sequence coverage.

Distribution of identified peptides for protein HPX, with the red segment indicating the signal peptide region at the N-terminal. Orange segments represent tryptic peptides, while green segments represent semi-tryptic peptides. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Beyond semi-tryptic analysis, we further applied the diaTracer-extracted pseudo-MS/MS spectra to perform unrestricted post-translational modification analyses using MSFragger's mass-offset and open search modes. Figure 3-3 displays the most abundant PTMs identified in these data using mass-offset mode, as summarized by PTM-Shepherd[69]. These analyses identified a wide spectrum of mass shifts corresponding to both common chemical artifacts and

biologically relevant PTMs, including phosphorylation and oxidative modifications. Importantly, because diaTracer generates reusable pseudo-MS/MS mzML files, these advanced searches were performed without reprocessing the raw diaPASEF data. The analysis took 341 min (or 10 min per file), from MSFragger search to PTM-Shepherd reports. A similar mass shift histogram (Appendix Figure C-2) was observed from the open search results using the FragPipe's "Open" workflow, and the computational analysis took similar time (317 min). As a result, multiple complex search strategies could be explored rapidly and reproducibly using the same deconvoluted spectra.

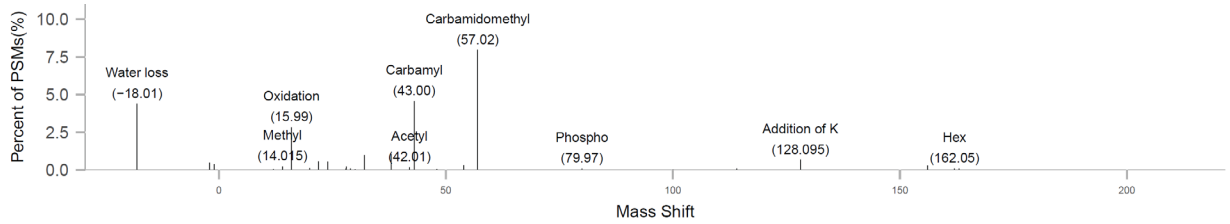


Figure 3-3. Modifications found in mass-offset search.

Modifications identified using the common mass-offset workflow in FragPipe using pseudo-MS/MS spectra generated by diaTracer. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

From a computational perspective, this reuse capability represents a key advantage of the spectrum-centric diaTracer workflow. Once diaPASEF data are converted into pseudo-MS/MS spectra, downstream analyses—including tryptic, semi-tryptic, open, and mass-offset searches—can be executed independently and efficiently. In contrast, peptide-centric DIA approaches typically require regeneration of in-silico spectral libraries whenever search parameters are modified, substantially increasing computational cost and limiting exploratory analyses. We benchmarked the running time of diaTracer and the rest of the FragPipe workflow (Figure 3-4). diaTracer required 661 min to generate pseudo-MS/MS spectra (19.4 min per diaPASEF file, on average), and the rest of FragPipe (tryptic search) required 87 min to complete the analysis (including running DIA-NN for quantification). In contrast, the DIA-NN library-free analysis took 10.5 min to generate the in-silico predicted spectral library and 1,486 min to perform the rest of the analysis, more than twice the time of the diaTracer workflow. Importantly, repeating FragPipe analysis (starting with the existing pseudo-MS/MS mzML files) using the semi-tryptic search took only 110 min.

Collectively, these results demonstrate that diaTracer enables direct diaPASEF analysis of CSF data while also unlocking advanced analytical modes that are impractical or infeasible

with peptide-centric DIA strategies. By capturing a richer precursor-level signal space and supporting flexible downstream searches, diaTracer provides a powerful framework for investigating endogenous proteolysis and post-translational modifications in CSF and other biologically complex sample types.

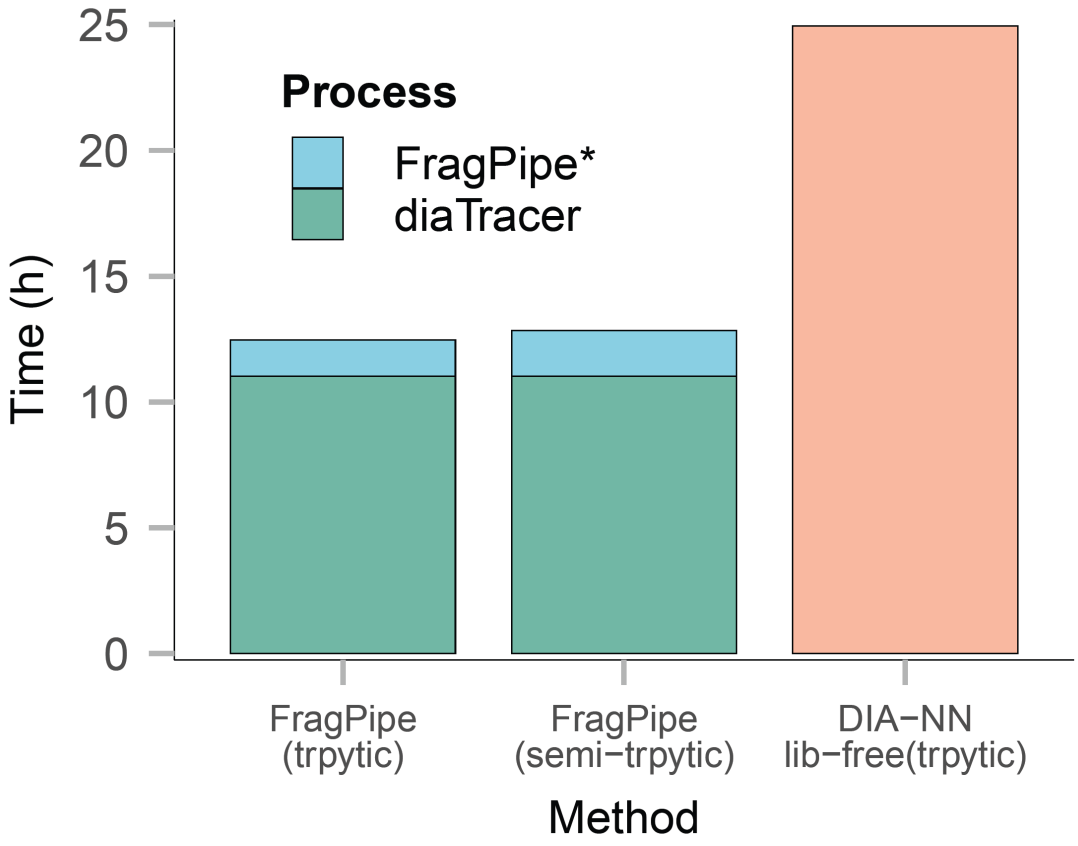


Figure 3-4. Running time comparison.

FragPipe* indicated the FragPipe run time (including DIA-NN for quantification) starting from diaTracer-extracted files. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

3.3.2 Plasma proteomics data analysis

To further demonstrate the utility of diaTracer for complex diaPASEF data analysis, we evaluated its performance on a plasma proteomics dataset from Vitko et al.[87]. Plasma represents one of the most challenging biological matrices due to its extreme dynamic range, high proteolytic activity, and extensive endogenous peptide processing[94,95]. In the original study, plasma samples were processed using Seer's Proteograph Assay Kit, which employs five distinct nanoparticle chemistries (NP1–NP5) to enrich complementary subsets of the plasma proteome. For the present analysis, we focused on 40 diaPASEF runs corresponding to NP2-enriched

plasma samples acquired with a 60 SPD gradient on a timsTOF HT mass spectrometer. The cohort included 20 patients with stage IV non-small cell lung cancer (NSCLC) and 20 non-cancer control samples, enabling downstream differential expression analysis.

Pseudo-MS/MS spectra were generated from the 40 diaPASEF runs using diaTracer, followed by database searching and quantification in FragPipe using both tryptic and semi-tryptic search settings. In parallel, the same dataset was analyzed using the DIA-NN library-free mode, representing a peptide-centric direct DIA strategy.

With a ≥50% non-missing value filter applied, and using a tryptic search, the DIA-NN library-free analysis quantified 2,347 proteins, whereas the diaTracer-based FragPipe workflow quantified 2,922 proteins (Figure 3-5a). At the precursor level, although DIA-NN reported a larger total number of precursors across all runs, the diaTracer workflow yielded substantially more consistently quantified precursors across runs. When requiring a precursor to be quantified in at least 50% of the runs, FragPipe with diaTracer reported 77% more precursors than DIA-NN (Figure 3-5b). This result highlights an important advantage of the spectrum-centric strategy in complex biofluid datasets: improved consistency of peptide quantification across large sample cohorts, which is critical for statistical power in downstream analyses.

As observed for the CSF dataset, enabling semi-tryptic searches primarily increased depth at the peptide level rather than the protein level. Using semi-tryptic search settings, we identified 30,158 peptides, compared with 26,469 peptides identified in the tryptic search. Among these, 7,642 peptides were uniquely detected in the semi-tryptic analysis (Appendix Figure C-3), of which 6,800 (89%) were semi-tryptic. Importantly, these additional identifications were not confined to isolated detections: 95% of the semi-tryptic peptides (6,458 peptides), corresponding to 1,104 proteins, were quantified across the dataset, demonstrating that the additional peptides contributed meaningfully to quantitative analyses.

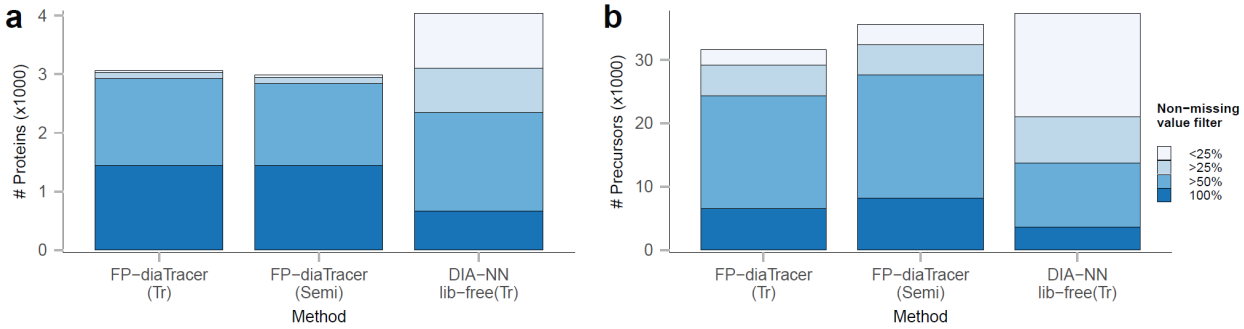


Figure 3-5. Plasma data result comparison.

(a) Histogram showing the number of quantified proteins using diaTracer-based FragPipe workflows (tryptic and semi-tryptic search), DIA-NN library-free mode, colored from deep to light blue, corresponding to different non-missing value filters. Shades of blue represent data completeness; darker blues indicate presence in a greater number of samples. (b) Same as (a) for quantified precursors. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

To assess the biological impact of these additional peptide measurements, we performed differential expression analysis between the NSCLC and control groups using FragPipe-Analyst, comparing results from tryptic and semi-tryptic searches. Under identical filtering criteria, the semi-tryptic analysis identified 69 additional significantly differentially expressed proteins relative to the tryptic analysis alone. In total, 35 proteins were significantly upregulated and 140 proteins were significantly downregulated in the NSCLC group based on the semi-tryptic results (Figure 3-6a).

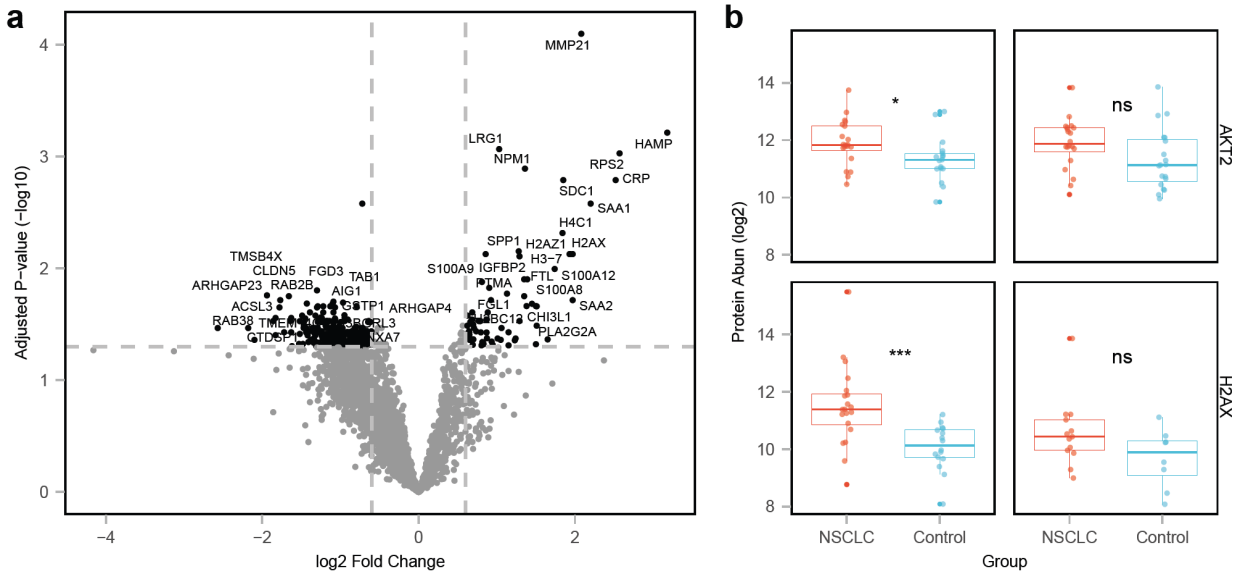


Figure 3-6. Plasma data biology analysis using semi-tryptic search.

(a) Volcano plot comparing protein abundance between stage IV non-small cell lung cancer (NSCLC) and non-cancer control samples, highlighting NSCLC-overexpressed proteins (Log2 fold change ≥ 0.6; adjusted *p*-value ≤ 0.05). The adjusted *p*-value is from the moderated *t*-test followed by the Benjamini-Hochberg procedure. (b) Boxplots of protein abundance distribution of AKT2 (top) and H2AX (bottom) proteins in semi-tryptic (left) and tryptic (right) searches between 40 diaPASEF runs from 20 NSCLC (red) and 20 control (blue) samples. The *p*-values decrease from 0.086 to 0.037 and from 0.089 to 0.00063 for AKT2 and H2AX respectively after using semi-tryptic searches. In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. ns: $p > 0.05$; *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$; ****$p \leq 0.0001$ (two-sided *t*-test). *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Figure 3-6b highlights two representative examples, AKT2 and H2AX, which showed statistically significant upregulation in NSCLC samples only when semi-tryptic searches were

enabled. Both proteins have been previously implicated in NSCLC pathogenesis and tumor biology[96,97]. Investigation at the peptide level revealed distinct mechanisms underlying these observations. For H2AX, the increased protein-level signal was driven by two semi-tryptic peptides that were significantly upregulated in NSCLC samples (Appendix Figure C-4). In contrast, for AKT2, no semi-tryptic peptides were quantified; instead, a single low-confidence tryptic peptide identified in the tryptic search failed to pass FDR filtering in the semi-tryptic analysis. In the absence of this peptide, the remaining evidence supported a statistically significant upregulation of AKT2 in the NSCLC group.

Together, these results illustrate how the spectrum-centric diaTracer workflow enables the analysis of complex diaPASEF plasma data beyond conventional tryptic constraints. By supporting semi-tryptic searches directly on diaPASEF data, diaTracer captures endogenous proteolytic events that are prevalent in plasma and enhances the detection of disease-associated proteomic changes that may be missed in standard peptide-centric DIA analyses.

3.3.3 Phosphoproteomics data analysis

To further evaluate the capability of diaTracer for handling complex diaPASEF datasets involving large search spaces and site-specific inference, we applied the diaTracer-based direct diaPASEF workflow in FragPipe to a phosphorylation-enriched dataset acquired using six different LC gradient lengths[88]. Phosphoproteomics has substantial analytical challenges due to the frequent co-elution of modified and unmodified peptides[98], and the need for accurate site localization, making it a stringent test case for spectrum-centric DIA analysis.

Across the six gradient conditions, diaTracer required on average 13.5 min to extract signals and generate pseudo-MS/MS spectra for a single 60 min diaPASEF file, demonstrating high computational efficiency even for enriched and information-dense samples. Peptide identification was performed using MSFragger with phosphorylation on serine, threonine, and tyrosine specified as variable modifications. PTMProphet was enabled for confident site localization, and the resulting localization probabilities were propagated through FragPipe into the final quantification matrices generated by DIA-NN.

Using this workflow, we identified 6,844 and 11,281 unique phosphopeptide sequences in the 7 min and 60 min gradient data, respectively (Figure 3-7a). Notably, in contrast to the original study, which reported a plateau in the number of identified phosphopeptides at a 21 min gradient, our diaTracer-based analysis revealed a continued increase in both phosphopeptide and

phosphosite identifications with longer gradient lengths. This trend was also observed for class I phosphorylation sites (localization probability >0.75), indicating that the improved separation afforded by longer gradients was effectively leveraged by the spectrum-centric diaTracer workflow (Appendix Figure C-5).

Quantitative reproducibility was assessed across four technical replicates for 7 min gradient condition. The diaTracer-based direct diaPASEF workflow exhibited high correlation in phosphopeptide intensities between replicates (Figure 3-7b), indicating robust and consistent quantification. We further compared the quantification precision with Spectronaut using coefficient of variation (CV) analysis. The CV distributions (Appendix Figure C-6) showed that the diaTracer–FragPipe workflow achieved quantification precision comparable to that of Spectronaut for these phosphoproteomics data.

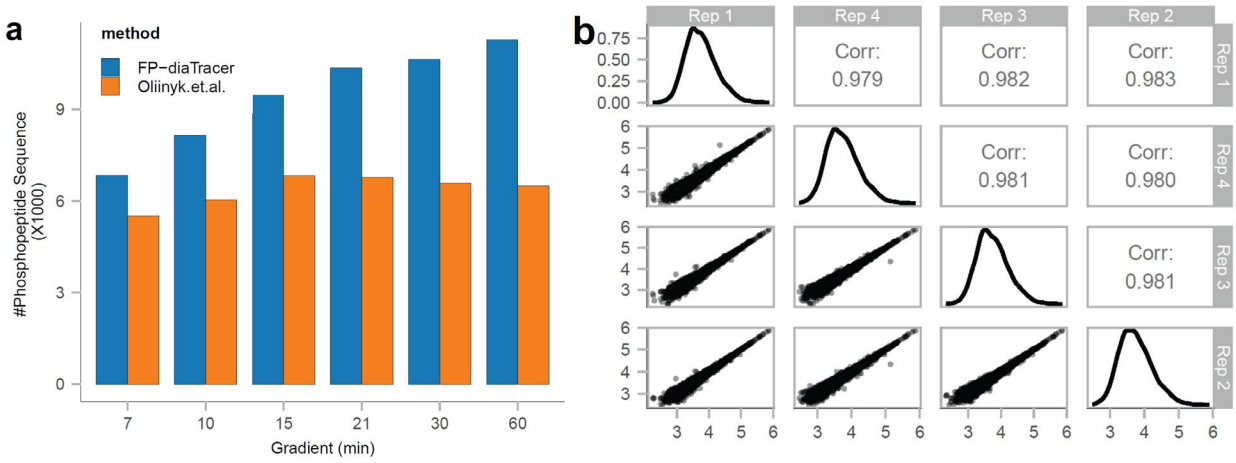


Figure 3-7. Phosphoproteomics data result.

(a) Histogram of quantified phosphorylated peptide sequences across different gradients using FragPipe with diaTracer (blue) and those reported in the original study based on Spectronaut 16 (orange). (b) Quantified phosphorylated peptides intensities and correlations in four replicates in the 7 min gradient time experiment. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Although site localization probabilities are computed during the spectral library generation stage, detailed inspection of chromatographic behavior and site-determining fragment ions can provide additional confidence in challenging cases. In addition to the integrated FragPipe-PDV[79] viewer, the output of the diaTracer–FragPipe workflow is fully compatible with Skyline[78], enabling advanced visualization of precursor and fragment ion signals in diaPASEF data.

As an illustrative example, we examined two co-eluting, isobaric positional phosphopeptide isomers, S(pho)PSPPDGSPAATPEIR and SPSPPDGS(pho)PAATPEIR,

identified and quantified in a 60 min gradient run. diaTracer successfully deconvoluted these peptides into two distinct pseudo-MS/MS spectra, correctly associating precursor and fragment ion signals for each isomer. Both peptides were identified with high site localization confidence (probability >0.75). Despite a near-complete overlap in m/z, the two species were separated by approximately 35 s in retention time and by 0.02 $1/K_0$ in ion mobility (Figure 3-8). Multiple site-determining fragment ions were clearly observed in the corresponding pseudo-MS/MS spectra, underscoring the ability of diaTracer to resolve and correctly localize phosphorylation sites in highly complex diaPASEF data.

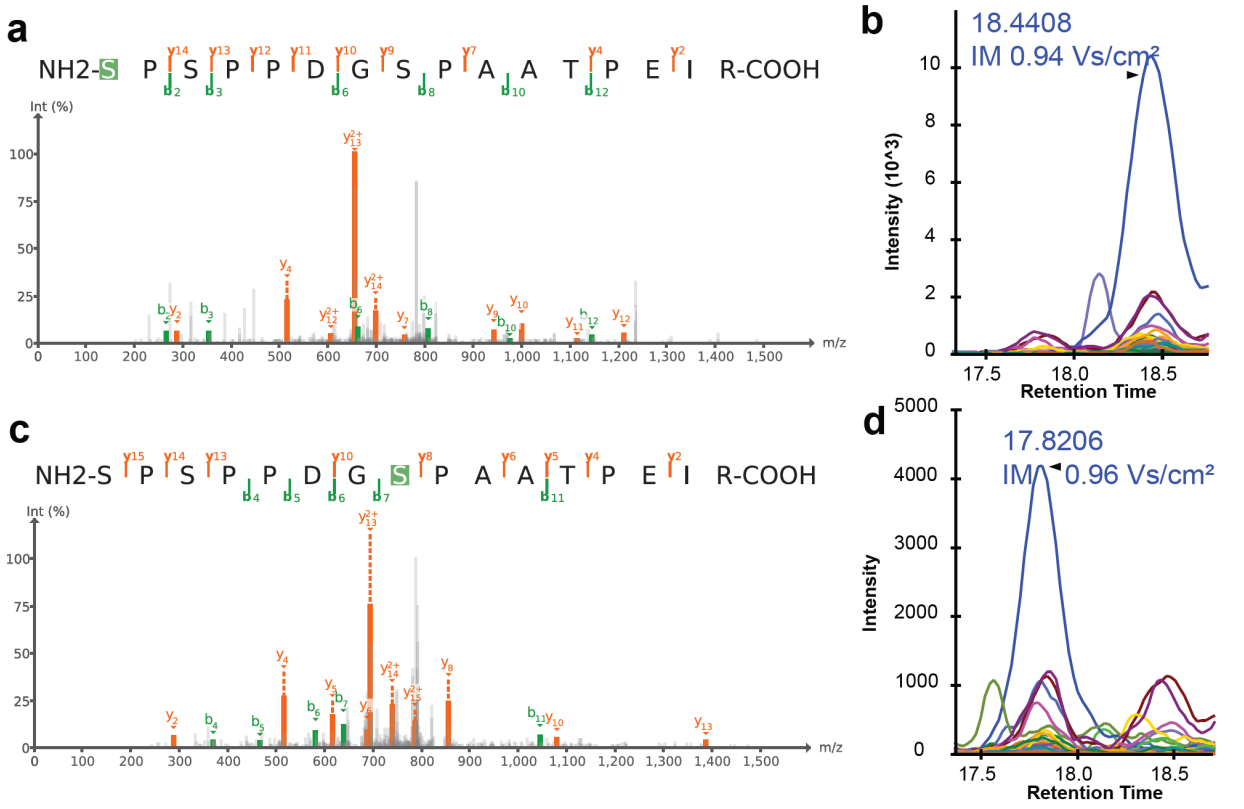


Figure 3-8. Phosphorylated co-eluted isobaric positional isomers.

(a) and (b) PSM and fragment XICs of phosphorylated peptide S(Pho)PSPPDGSPAATPEIR. (c) and (d) PSM and fragment XICs of phosphorylated peptide SPSPPDGS(Pho)PAATPEIR. The spectrum was visualized using FragPipe-PDV. The XICs were generated by Skyline. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

3.3.4 HLA immunopeptidomics data analysis

HLA immunopeptidomics represents one of the most challenging application domains for DIA data analysis due to the nonspecific nature of antigen processing and the resulting explosion of candidate peptide search space. In this context, peptide-centric DIA strategies are not

practically feasible, as they require exhaustive in-silico spectral prediction for all possible peptides generated under nonspecific cleavage rules. In contrast, the spectrum-centric strategy implemented in diaTracer—based on direct extraction of pseudo-MS/MS spectra from diaPASEF data—provides a natural and scalable solution for HLA and endogenous peptidome analysis.

To evaluate the performance of diaTracer for HLA immunopeptidomics, we analyzed a published dataset from Wahle et al.[89]. We selected three technical replicates of an HLA peptidome sample purified from plasma of a healthy donor and acquired using a Whisper40 gradient (31 min) on a timsTOF Ultra instrument. Pseudo-MS/MS spectra were generated using diaTracer and searched in FragPipe using a nonspecific digestion workflow optimized for HLA peptides (see Methods).

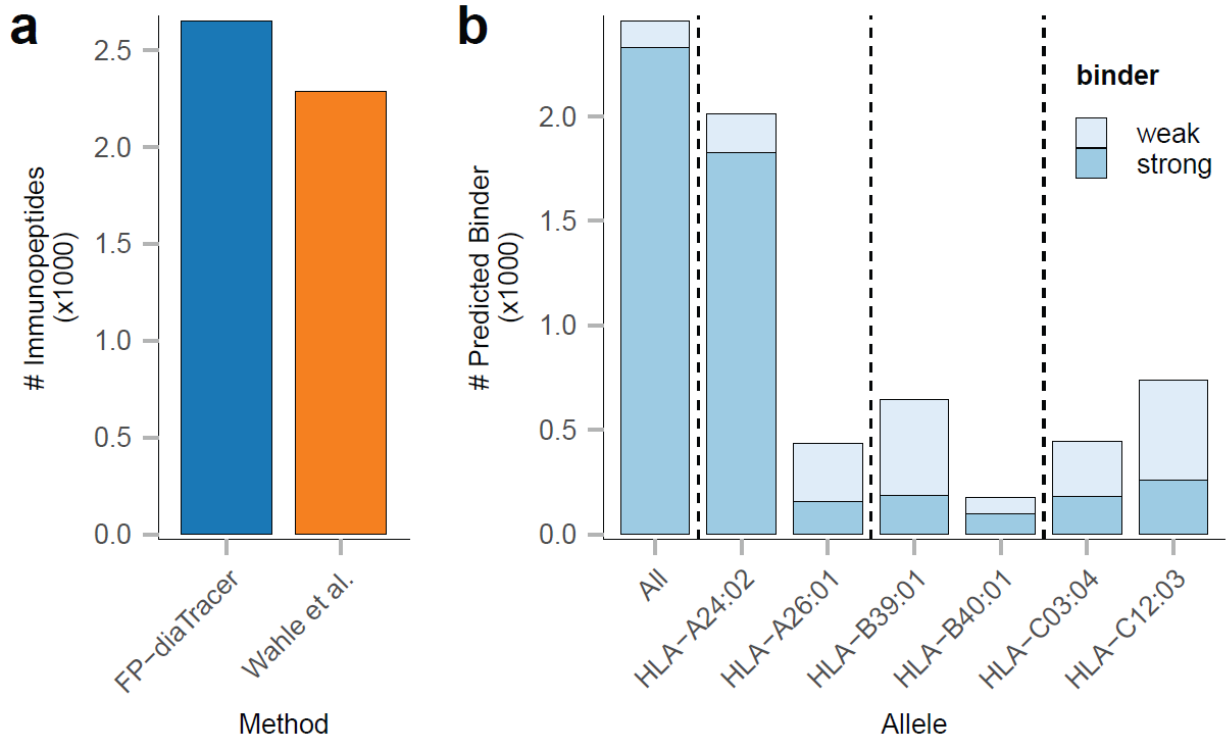


Figure 3-9. Immunopeptidomics HLA results.

(a) Number of quantified immunopeptides obtained using FragPipe with diaTracer and those reported in the original study based on Spectronaut 17. (b) Histogram of predicted binders for all HLA alleles of the corresponding sample donor, colored by binder type (light: weak binder; dark: strong binder). *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Across the three replicates, diaTracer-based analysis identified and quantified 2,651 immunopeptides with lengths ranging from 7 to 14 amino acids (Figure 3-9a), representing an

18% increase compared with the 2,288 peptides reported in the original study using Spectronaut version 17 directDIA. To further assess the biological reasonableness of the identified peptides, we predicted peptide–MHC binding affinities using NetMHCpan-4.1, incorporating the donor-specific HLA allele information provided in the original study. Among the 2,572 peptides of length 8–12, 2,327 (90%) were classified as strong binders (percentile rank <0.5%), and an additional 130 (5%) as weak binders (percentile rank <2%) (Figure 3-9b). Most predicted binders originated from the HLA-A*24:02 allele, consistent with the allele distribution and binding preferences reported previously (Appendix Figure C-7).

We further examined the basic characteristics of the identified immunopeptides. Length and charge state distributions (Figure 3-10) revealed a strong enrichment for canonical 9-mer peptides, predominantly observed as singly or doubly charged ions—features that align well with established properties of MHC class I ligands and with the findings of the original study. We also compared the overlap between immunopeptides identified using diaTracer and those reported previously using an experimental DDA library and a pan-library approach (Appendix Figure C-8). Although more comprehensive libraries are expected to yield higher identification numbers, the diaTracer-based results showed substantial overlap with published datasets, indicating robust and consistent peptide recovery.

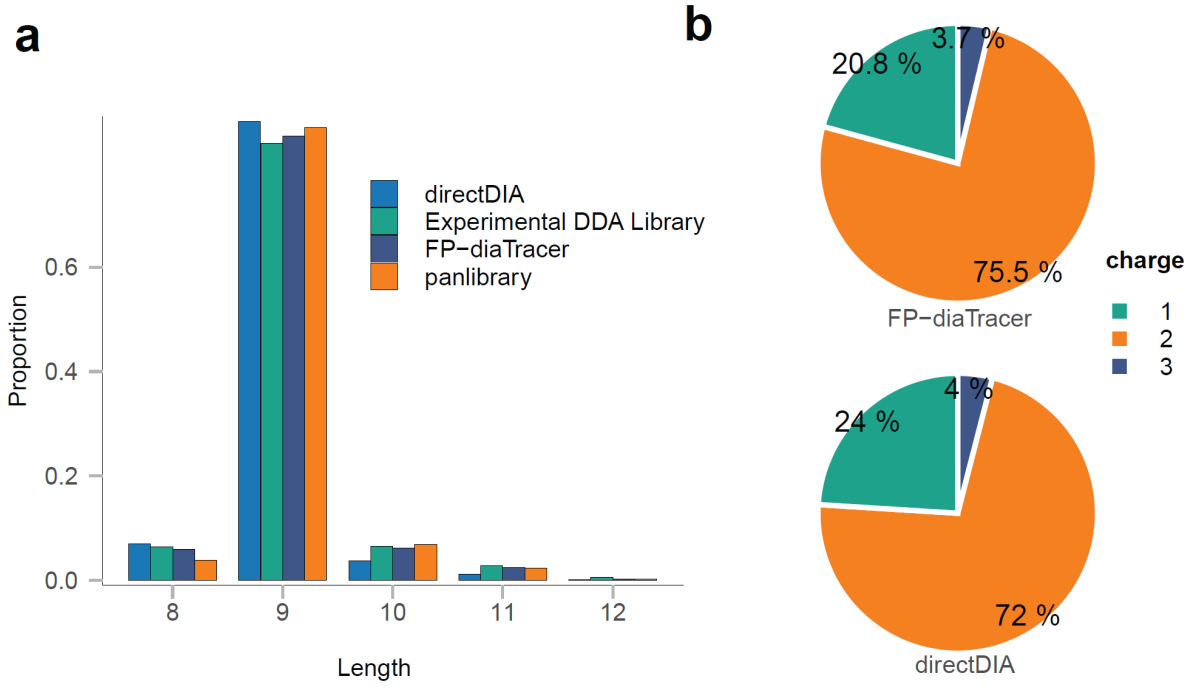


Figure 3-10. Characteristics of quantified HLA peptides.

(a) Length distribution of quantified immunopeptides using FragPipe with diaTracer and that reported in the original study based on Spectronaut 17 directDIA, DDA experimental library, and panlibrary, each with a unique color. (b) Charge state distribution of quantified immunopeptides using FragPipe with diaTracer and that reported in the original study based on Spectronaut 17 directDIA. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Finally, we demonstrate the quality of diaTracer-extracted pseudo-MS/MS spectra using the example peptide VYQHLFTRI, predicted to be a strong HLA binder. Using the integrated FragPipe-PDV viewer, we visualized one representative PSM together with the predicted spectrum generated by MSBooster[71]. The pseudo-MS/MS spectrum exhibited a very high similarity to the predicted spectrum (spectral entropy score of 0.9863; Figure 3-11), underscoring the fidelity of diaTracer-derived spectra for immunopeptidomics applications.

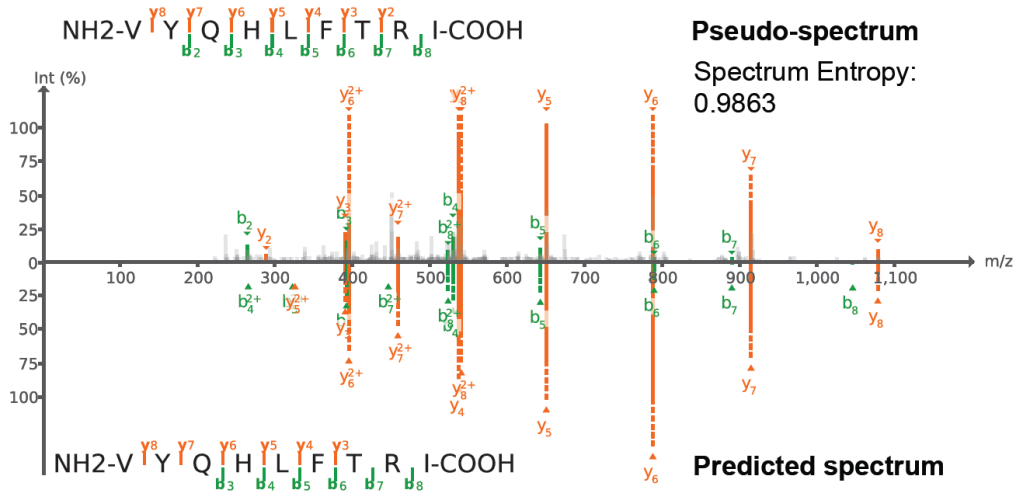


Figure 3-11. One example of HLA peptides.

Pseudo-MS/MS spectrum generated by diaTracer and the predicted spectrum for one of the quantified immunopeptides, VYQHLFTRI. The entropy score between the two spectra is 0.9863. The spectrum was visualized using FragPipe-PDV. *Adapted from Li et al., Nature Communications (2025), under the Creative Commons CC BY 4.0 license.*

Collectively, these results highlight the unique strengths of the spectrum-centric diaTracer workflow for HLA immunopeptidomics, enabling sensitive, biologically consistent, and computationally tractable analysis of complex diaPASEF datasets that are otherwise inaccessible to peptide-centric DIA strategies.

3.4 Discussion

In this chapter, we focused on complex diaPASEF applications that are difficult or impractical to address using peptide-centric DIA analysis strategies and demonstrated how

diaTracer enables these analyses through a direct, spectrum-centric workflow. While Chapter 2 established the general performance, efficiency, and integration of diaTracer within FragPipe for conventional proteomics experiments, the results presented here highlight the conceptual and practical advantages of spectrum-centric analysis when applied to expanded search spaces and unconventional peptide populations.

Recent advances in deep learning have substantially improved the accuracy of peptide property prediction, including fragment ion intensities, retention times, and ion mobilities. These advances strengthen modern peptide-centric DIA tools, which rely on in-silico spectral libraries generated from predicted peptide properties. For standard bottom-up proteomics experiments—such as tryptic digestion of cell lines or tissue samples with limited PTM complexity—peptide-centric strategies are highly effective and computationally efficient. However, these approaches face fundamental limitations when the peptide search space expands dramatically.

In particular, peptide-centric strategies are poorly suited for endogenous peptidomics and immunopeptidomics, where nonspecific proteolytic processing generates an enormous number of candidate peptides. The requirement to generate in-silico predictions for all possible peptides under nonspecific cleavage rules renders such analyses computationally infeasible. Even semi-tryptic searches, which are often biologically informative for plasma proteomics or N-terminomics studies[99,100], substantially increase the number of candidate peptides and quickly become challenging for peptide-centric workflows. The situation is worse in PTM-centric studies. Searches involving multiple variable modifications—or unrestricted PTM discovery via open or mass-offset searches—are generally not tractable within peptide-centric frameworks. Even phosphorylation, one of the most extensively studied PTMs, can become computationally prohibitive when combined with DIA data and expanded search parameters. Moreover, peptides carrying rare PTMs or chemical labels not represented in model training sets pose an additional risk of inaccurate prediction, further limiting identification sensitivity and robustness.

The spectrum-centric strategy implemented in diaTracer, building on the original DIA-Umpire concept and extended here to diaPASEF data, minimizes these limitations by reversing the analysis method. Instead of relying on prior prediction of peptide properties, diaTracer first extracts precursor and fragment ion signals directly from the data and reconstructs pseudo-MS/MS spectra. These spectra can then be searched using mature DDA identification engines,

enabling semi-enzymatic, nonspecific, open, or mass-offset searches without incurring the computational cost of exhaustive peptide prediction.

Across multiple datasets in this chapter, we demonstrated that this strategy enables biologically meaningful analyses that are otherwise inaccessible with peptide-centric DIA. Semi-enzymatic searches substantially increased the number of quantified precursors in CSF and plasma datasets and improved quantitative consistency across samples, leading to the identification of additional differentially expressed proteins. In phosphoproteomics data, the spectrum-centric approach preserved site-localizing fragment information and enabled robust quantification and localization across gradient lengths, even in the presence of co-eluting isobaric positional isomers. In HLA immunopeptidomics, diaTracer enabled direct analysis of diaPASEF data under nonspecific digestion rules, yielding higher peptide identifications than previously reported analyses while maintaining strong agreement with MHC binding predictions and known biological characteristics of HLA ligands.

Importantly, spectrum-centric workflows still benefit from deep learning–based prediction and rescoring. In the diaTracer–FragPipe pipeline, predictions are applied only after peptide-spectrum matching and only for a relatively small number of candidate peptides. As a result, prediction accuracy is less critical for peptide inclusion, increasing robustness for peptides with uncommon PTMs or chemical modifications. This property makes the diaTracer workflow particularly attractive for chemoproteomics applications, such as affinity-based proteome profiling (ABPP)[101,102], where diverse chemical labels are frequently used and accurate prediction of peptide behavior remains challenging.

Computational efficiency is a critical consideration for large-scale diaPASEF studies. In our benchmarks, the total processing time for diaTracer within FragPipe was consistently shorter than the corresponding mass spectrometry acquisition time. Moreover, the conversion of raw diaPASEF data into pseudo-MS/MS spectra can be initiated as soon as individual files are acquired, allowing computational processing to proceed in parallel with data acquisition. Importantly, this conversion step needs to be performed only once. The resulting deconvoluted mzML files can then be reused for multiple downstream analyses, such as testing different database search parameters, applying alternative digestion rules, performing open or mass-offset searches, or using different search engines, without the need to repeat the computationally intensive feature extraction step.

Finally, the spectrum-centric design of diaTracer is inherently compatible with ongoing developments in diaPASEF acquisition technology. Recent acquisition modes, including synchro-PASEF[44] and Slice-PASEF[45], aim to improve precursor–fragment coherence by redesigning quadrupole isolation schemes and timing relationships. These methods further strengthen the assumptions underlying spectrum-centric deconvolution by better preserving the physical relationships between precursors and fragments. While the analyses in this chapter focused on conventional diaPASEF data, these emerging acquisition strategies naturally motivate further methodological extensions of diaTracer, which are addressed in the next chapter.

3.5 Data availability

The raw MS/MS files used in this study can be accessed via the ProteomeXchange Consortium through the PRIDE repository[81] or at the MassIVE repository with the following accession codes:

- Cerebrospinal fluid (CSF) dataset PXD035249: https://www.ebi.ac.uk/pride/archive/projects/PXD035249
- Plasma dataset PXD047839: https://www.ebi.ac.uk/pride/archive/projects/PXD047839
- Phosphoproteomics data PXD033904: https://www.ebi.ac.uk/pride/archive/projects/PXD033904
- HLA data MSV000092557: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=18b344b86e3f44539b35190466c80c73
- The diaTracer converted mzML files and FragPipe results generated in this study have been deposited in the MassIVE repository with the identifier MSV000094803: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=85c80cf99442470e9e2aa01830c88120

3.6 Acknowledgements and competing interests

This work was supported in part by National Institutes of Health grants R01-GM-094231 and U24-CA271037.

A.I.N. and F.Y. receive royalties from the University of Michigan for the sale of MSFragger, IonQuant, and diaTracer software licenses to commercial entities. K.L. receives royalties from the University of Michigan for the sale of diaTracer software licenses to

commercial entities. All license transactions are managed by the University of Michigan Innovation Partnerships office, and all proceeds are subject to university technology transfer policy. Other authors declare no other competing interests.

3.7 Authors, affiliations, and contributions

Kai Li[1], Guo Ci Teo[2], Kevin L. Yang[1], Fengchao Yu[2] & Alexey I. Nesvizhskii[1,2]

[1]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA

K.L. and A.I.N. developed the diaTracer algorithm. K.L. wrote the software and analyzed the results. F.Y. and G.C.T. assisted with the algorithm and software development. K.L.Y. helped modify MSBooster. F.Y. assisted with the integration of diaTracer into FragPipe. K.L., F.Y., and A.I.N. wrote the manuscript. A.I.N. conceived the study. A.I.N. and F.Y. supervised the study.

Chapter 4 Supporting Diagonal PASEF Acquisition Modes with an Optimized diaTracer Framework

4.1 Introduction

The Bruker timsTOF mass spectrometer platform, which integrates trapped ion mobility (IM) separation with the parallel accumulation-serial fragmentation (PASEF) technique, has demonstrated excellent performance for data-independent acquisition (DIA) workflows, particularly in the diaPASEF mode[43]. In previous chapters, we introduced diaTracer[55], a spectrum-centric computational framework designed to analyze diaPASEF data by detecting extracted ion chromatogram (XIC) features and reconstructing data-dependent acquisition (DDA)-like pseudo-tandem mass spectrometry (MS/MS) spectra. Conceptually related to the DIA-Umpire[52] strategy, diaTracer enables the application of conventional DDA database search engines to DIA data and supports large search-space analyses, including open modification, semi-enzymatic, and post-translational modification (PTM)-focused searches. When integrated into the FragPipe computational platform, diaTracer demonstrated strong and competitive performance relative to peptide-centric tools such as DIA-NN[51], particularly in discovery-oriented proteomics workflows that benefit from flexible database search strategies.

Recent advances in diaPASEF acquisition strategies have introduced diagonal scanning schemes, including synchro-PASEF[44] and Slice-PASEF[45]. These methods synchronize quadrupole isolation with ion mobility separation or divide the ion mobility dimension into discrete slices across frames, enabling near-continuous precursor coverage and substantially increasing fragment ion sampling efficiency. As a result, these acquisition modes are increasingly attractive for large-scale proteomics studies and high-throughput experimental designs.

However, diagonal diaPASEF acquisitions fundamentally alter the structure of the resulting MS/MS data. In contrast to conventional diaPASEF, where fragment ions from a precursor are largely confined to a single isolation window and frame, diagonal acquisitions distribute fragment ions across multiple mass-to-charge ratios (m/z) and ion mobility slices. The

traditional concept of fixed, vertical isolation windows is no longer applicable, and precursor–fragment relationships are encoded across dimensions rather than directly through isolation window boundaries.

This change poses a significant challenge for spectrum-centric DIA analysis, which relies on coherent precursor–fragment relationships within well-defined isolation windows to reconstruct pseudo-MS/MS spectra. Without algorithmic adaptation, diagonal diaPASEF data cannot be directly processed using existing spectrum-centric workflows developed for conventional diaPASEF acquisitions. At the same time, the increasing scale of diaPASEF-based studies—often encompassing hundreds to thousands of samples[103]—places growing demands on computational efficiency, memory usage, and data reduction.

In this chapter, we present diaTracer 2.0, an updated and extended version of the diaTracer framework that addresses both the methodological and computational challenges associated with modern diaPASEF data. diaTracer 2.0 introduces performance improvements through optimized data structures, enhanced feature extraction strategies, and refined precursor isotope filtering, enabling faster and more scalable processing of large diaPASEF datasets. In addition, we extend the spectrum-centric method to diagonal diaPASEF acquisitions by introducing a pseudo-isolation window concept, which restores coherent precursor–fragment grouping in the absence of conventional vertical isolation windows. Using multiple public datasets, we demonstrate that diaTracer 2.0 enables robust spectrum-centric analysis of synchro-PASEF and Slice-PASEF data while substantially improving processing speed and scalability, thereby broadening the applicability of spectrum-centric DIA analysis to new timsTOF acquisition methods.

4.2 Methods

4.2.1 Algorithm optimization in diaTracer 2.0

In diaTracer 2.0, the isotope grouping strategy was refined to improve the completeness and robustness of isotopic peak association. Specifically, a loose m/z tolerance was applied during the isotope filtering process, allowing a broader set of potential isotopic peaks from each precursor to be considered and removed from subsequent processing. This refinement enables more consistent grouping of first-order and higher-order isotope peaks with their corresponding monoisotopic precursor features, particularly for lower-intensity signals.

To address the increasing computational demands imposed by large-scale diaPASEF studies, diaTracer 2.0 introduces optimizations to its multi-threaded execution model. Profiling analyses of diaTracer 1.0 revealed that overall CPU utilization was frequently limited not by numerical computation but by excessive task scheduling overhead and thread context switching. In diaTracer 1.0, fragment-ion peak extraction was parallelized by subdividing each isolation window into fine-grained computational units defined by four-dimensional bounds (m/z_{lo}, m/z_{hi}, IM_{lo}, IM_{hi}). While the ion mobility bounds were fixed by the isolation window geometry, the m/z dimension was further partitioned according to the number of available threads. These small units were processed independently within the isolation-window loop.

Although this design enabled parallel execution, it resulted in frequent creation and destruction of short-lived tasks. Because individual computational units were relatively small, worker threads often completed their assigned work quickly and remained idle while awaiting new tasks. This behavior led to suboptimal CPU utilization and reduced throughput, particularly for long-gradient diaPASEF acquisitions where the total number of frames is large.

In diaTracer 2.0, the logical definition of computational units was preserved, but their execution strategy was fundamentally reorganized. Instead of repeatedly creating short-lived tasks within each isolation-window loop, we restructured processing such that each worker thread is assigned a fixed m/z subrange and executes continuously across all frames associated with a given isolation window. For each isolation window, m/z subranges are defined at the beginning and assigned to individual threads. Each thread then processes its assigned m/z region sequentially across the entire retention time dimension, maintaining local state and intermediate data structures throughout execution. This design substantially reduces synchronization overhead and allows threads to perform sustained numerical computation without interruption. To further improve load balancing, the estimated computational workload associated with each m/z subrange is calculated in advance. Subranges are then sorted in descending order of expected cost, and threads are scheduled accordingly so that the most computationally intensive regions are processed first. This strategy minimizes idle time toward the end of execution and improves overall throughput and scalability.

A second major performance optimization in diaTracer 2.0 targets the identification of local maxima during two-dimensional peak extraction in the (m/z, $1/K_0$) plane. In diaTracer, peak extraction is performed on intensity maps represented as sparse matrices. The procedure

consists of three main steps: (i) removal of isolated singleton signals ("loners"), (ii) smoothing of the remaining intensity map using a two-dimensional Gaussian kernel, and (iii) detection of local maxima in the smoothed map, which serve as seed points for downstream two-dimensional peak tracing. In diaTracer 1.0, local maxima were detected by scanning the entire smoothed intensity matrix. While straightforward, this approach becomes computationally expensive in diaPASEF applications because Gaussian smoothing partially densifies the matrix. In diaTracer 2.0, this full-matrix scan is replaced by a seeded local-maximum search that leverages sparsity information preserved prior to smoothing.

Specifically, the non-loner points identified during step (i) are retained as a compact set of candidate seed locations. These seeds are used to explore local neighborhoods on the smoothed matrix. This strategy is based on the observation that true maxima in the smoothed intensity map must lie within a bounded distance of at least one non-loner point, because convolution with a finite Gaussian kernel can only propagate signal within a limited spatial radius. For each non-loner seed coordinate, diaTracer evaluates candidate maxima within a neighborhood whose extent is determined by the Gaussian kernel support and the maximum allowed peak radius along the m/z and ion mobility dimensions.

By replacing dense full-matrix scanning with a seed-driven neighborhood search, diaTracer 2.0 reduces the computational complexity to scale with the number of non-loner informative signal points rather than the total number of grid cells in the full matrix.

4.2.2 Diagonal diaPASEF data processing in diaTracer 2.0

In contrast to conventional diaPASEF acquisitions, synchro-PASEF continuously fragments precursor ions across both the m/z and ion mobility dimensions without predefined vertical isolation windows. Although this diagonal scanning strategy substantially improves acquisition efficiency, fragment ions derived from a given precursor are distributed across multiple slices without a clear isolation-window relationship. As a result, synchro-PASEF data cannot be directly analyzed using traditional spectrum-centric workflows that rely on well-defined isolation windows.

To address this challenge, diaTracer 2.0 introduces the concept of virtual (pseudo) isolation windows, which approximate the functional role of conventional quadrupole isolation windows. These pseudo-isolation windows are defined along the m/z axis and are subsequently linked to the set of diagonal isolation slices that intersect each window. This mapping is

established prior to fragment feature extraction, enabling MS2 scans to be aggregated into window-specific fragment ion maps suitable for downstream spectrum-centric analysis.

As in diaTracer 1.0, the m/z axis is discretized onto a fixed grid to support efficient indexing and fast overlap queries. The pseudo-isolation window width (default: 30 Th) is converted into an integer number of grid bins, and a series of consecutive windows is generated to span the targeted precursor m/z range. The total number of pseudo-windows is computed from the lower and upper bounds of the acquisition range and the selected window width, with a final partial window included if necessary to ensure complete coverage. Each pseudo-isolation window is represented by a four-parameter range $(m/z_{\mathrm{lo}}, m/z_{\mathrm{hi}}, IM_{\mathrm{lo}}, IM_{\mathrm{hi}})$. While the m/z bounds are defined deterministically during window construction, the ion mobility bounds are estimated from the diagonal acquisition geometry, as described below. Synchro-PASEF acquisitions consist of a collection of two-dimensional isolation slices defined in the $(m/z, IM)$ plane, typically organized by frame and indexed within each frame. For each pseudo-isolation window, diaTracer iterates over all available diagonal isolation slices and identifies those that sufficiently intersect the pseudo-window along the m/z dimension. Specifically, a diagonal isolation slice is considered to intersect a pseudo-window if its m/z range overlaps the pseudo-window by at least a predefined fraction of the pseudo-window width (default: 20%).

All diagonal slices meeting this overlap criterion are assigned to the pseudo-isolation window and recorded as (frame index, slice index) pairs. This procedure establishes a relationship between each pseudo-isolation window and the set of diagonal slices that contribute fragment ion information to it. To improve computational efficiency, diaTracer utilizes the ordering of diagonal slices by m/z within each frame: when iterating through slices in descending m/z order, the search terminates early once the lower m/z bound of a slice exceeds the pseudo-window's upper bound adjusted by the overlap margin, thereby avoiding unnecessary comparisons.

After identifying diagonal isolation slices that intersect each pseudo-isolation window, diaTracer determines the corresponding ion mobility range by combining the ion mobility spans of all assigned slices. Pseudo-isolation windows that do not intersect with any diagonal slices—most commonly near the boundaries of the acquisition range—are excluded to avoid generating empty windows and unnecessary computation. Each remaining pseudo-isolation window, together with its associated diagonal slices, serves as the basic unit for fragment ion aggregation.

Fragment ion signals from all slices assigned to the same pseudo-window are stacked and processed using the standard diaTracer fragment feature detection workflow. Because these pseudo-windows recreate a windowed structure like conventional diaPASEF acquisitions, all downstream steps—including precursor feature detection, precursor–fragment correlation, clustering, and pseudo-MS/MS spectrum generation—can be applied without additional modifications.

Slice-PASEF acquisitions involve fewer ion mobility slices per cycle and broader ion mobility coverage within each slice compared with synchro-PASEF. Consequently, the fragmentation pattern of Slice-PASEF more closely resembles that of conventional diaPASEF, where fragment ions are largely confined within discrete isolation regions rather than continuously distributed across the m/z and ion mobility dimensions. For this reason, Slice-PASEF data were processed using the standard diaTracer workflow without major algorithmic changes. Only minor adjustments were made to accommodate differences in acquisition settings, such as the ion mobility range handling and the frame organization. Following these adjustments, fragment ion extraction, feature detection, precursor–fragment association, and pseudo-MS/MS spectrum generation were carried out using the same spectrum-centric procedures as for conventional diaPASEF data.

4.2.3 Experimental datasets

Two publicly available datasets obtained from the PRIDE[81] repository and the Japan ProteOme STandard Repository (jPOST)[104] were used to evaluate the performance and scalability of diaTracer 2.0 across multiple diaPASEF acquisition strategies.

The dataset JPST003679[105] was used to systematically benchmark diaTracer performance under different diaPASEF acquisition schemes, including the conventional diaPASEF, Slice-PASEF (1-frame and 4-frame variants), synchro-PASEF, and thin-diaPASEF. Thin-diaPASEF represents an optimized diaPASEF acquisition strategy characterized by a narrower ion mobility range ($1/K_0$= 0.7–1.3) compared with the default diaPASEF setting ($1/K_0$= 0.6–1.6), thereby increasing precursor density within each frame while reducing ion mobility coverage.

In this dataset, 100 ng of trypsin-digested HEK293 cell lysate was analyzed in five technical replicates using a 50-minute liquid chromatography (LC) gradient on a timsTOF HT mass spectrometer (Bruker Daltonics). For thin-diaPASEF, the quadrupole isolation window was set to 25 Th. The conventional diaPASEF method was designed using the Python package for

DIA with automated isolation design software py_diAID[106]. Slice-PASEF acquisition methods were configured following the parameters reported by Szyrwiel et al.[45], including both single-frame (Slice-PASEF-1F) and four-frame (Slice-PASEF-4F) variants. These methods employed an ion mobility range of $1/K_0$= 0.75–1.2 with a ramp time of 100 ms. Synchro-PASEF acquisitions were configured according to the protocol described by Skowronek et al.[44], using an ion mobility range of $1/K_0$= 0.7–1.3, a ramp time of 100 ms, an isolation window width of 25 Th, and four synchronized quadrupole scans per cycle.

Five conventional diaPASEF replicates downloaded from the Japan ProteOme STandard Repository with the identifier JPST003679 were additionally used to benchmark improvements in precursor isotope filtering introduced in diaTracer 2.0.

To evaluate computational performance and scalability, the dataset PXD017703[43] was used. It includes 200 ng HeLa cell lysate samples acquired using diaPASEF methods with multiple gradient lengths, including 200 SPD, 100 SPD, and 60 SPD, on a timsTOF Pro instrument. The diaPASEF acquisition scheme employed 32 isolation windows of 25 Th each, spanning from 400 to 1,200 Th.

4.2.4 Data analysis

All datasets were analyzed using FragPipe with the built-in **"**DIA_SpecLib_Quant_diaPASEF**"** workflow with diaTracer-enabled spectrum-centric processing. Pseudo-MS/MS spectra were generated from diaPASEF raw data using diaTracer (version 2.0 and 1.1.5) with parameters selected based on empirical optimization described in Chapter 2. Specifically, the ion mobility and retention time tolerances for precursor–fragment association were controlled by setting "Delta Apex IM" to 0.01 ($1/K_0$ units) and "Delta Apex RT" to 3 frames. The maximum number of fragment peaks retained per pseudo-MS/MS spectrum ("RF max") was set to 500, and the minimum Pearson correlation coefficient required for precursor–fragment association ("Corr threshold") was set to 0.3. The mass defect filter was enabled.

Database searching of the diaTracer-generated pseudo-MS/MS spectra was performed using MSFragger[62] version 4.4. Initial precursor and fragment mass tolerances were set to 10 ppm and 20 ppm, respectively. Spectrum deisotoping, mass calibration, and parameter optimization were enabled, and the isotope error was allowed to range from 0 to 2 to accommodate residual mass assignment uncertainty. Searches were conducted against the

reviewed *Homo sapiens* UniProt protein database (downloaded November 15, 2023; 20,461 entries), supplemented with common contaminants and reversed decoy sequences for false discovery rate (FDR) estimation. Enzyme specificity was set to strict trypsin, allowing up to two missed cleavages. Oxidation of methionine and protein N-terminal acetylation were included as variable modifications, with a maximum of three variable modifications permitted per peptide. Following database searching, peptide-spectrum matches (PSMs) were rescored using MSBooster[71] for deep-learning-based scoring, followed by Percolator[72]. PSMs were filtered to a 1% FDR at the PSM level. The final set of FDR-filtered PSMs, together with the corresponding pseudo-MS/MS mzML files produced by diaTracer, were used as input for EasyPQP to construct spectral libraries. These libraries contained precursor m/z, charge state, normalized retention time, ion mobility values, and fragment ion annotations derived from the pseudo-MS/MS spectra.

Protein and peptide quantification was performed using DIA-NN[51] version 2.04, which provides native support for diagonal diaPASEF acquisition schemes, including synchro-PASEF and Slice-PASEF. To ensure fair and consistent comparisons across workflows and software versions, DIA-NN output "report.tsv" files were post-processed and filtered to achieve a 1% FDR at run-specific precursor, global precursor, and global protein levels using the *iq* R package[66]. This filtering strategy was applied uniformly across all datasets and acquisition modes. All downstream statistical analyses and data visualizations were performed in RStudio (Build 402) using R version 4.3.3. The R packages ggplot2, tidyverse, ggrepel, plotly, eulerr, and protti were used for data manipulation, visualization, overlap analysis, and quantitative performance assessment.

4.3 Results

4.3.1 Improved isotope filtering and computational performance in diaTracer 2.0

Pseudo-MS/MS spectra generated by diaTracer enable advanced database search strategies for diaPASEF data, including mass-offset and open modification searches. During the analysis of open search results generated from diaTracer version 1.1.5, we observed that a substantial fraction of PSMs were assigned to isotopic peaks rather than monoisotopic precursors. This observation indicates that isotopic peaks were not sufficiently filtered during

pseudo-MS/MS generation, leading to redundant spectra, inflated search spaces, and unnecessarily large pseudo-MS/MS mzML files.

To address this limitation, diaTracer 2.0 introduces a refined precursor isotope filtering strategy that improves the completeness and accuracy of isotopic peak association prior to pseudo-MS/MS spectrum construction. Using five replicate diaPASEF runs from the dataset JPST003679[105], we identified the proportion of PSMs assigned to the first, second, and third isotopic peaks in open modification searches performed by FragPipe. Compared with diaTracer version 1.1.5, the pseudo-MS/MS spectra generated by diaTracer 2.0 showed a reduction in isotope-derived PSMs: assignments to the first, second, and third isotopes were reduced by 70.3%, 64.5%, and 38.2%, respectively (Figure 4-1a).

Improved isotope filtering directly translated into a more compact and efficient representation of pseudo-MS/MS spectra. As a result, the average size of the pseudo-MS/MS mzML files generated by diaTracer 2.0 was reduced by approximately 39% compared with those produced by version 1.1.5 (Figure 4-1b). This reduction substantially decreases disk storage requirements and lowers the computational burden associated with downstream database searching, rescoring, and spectral library construction.

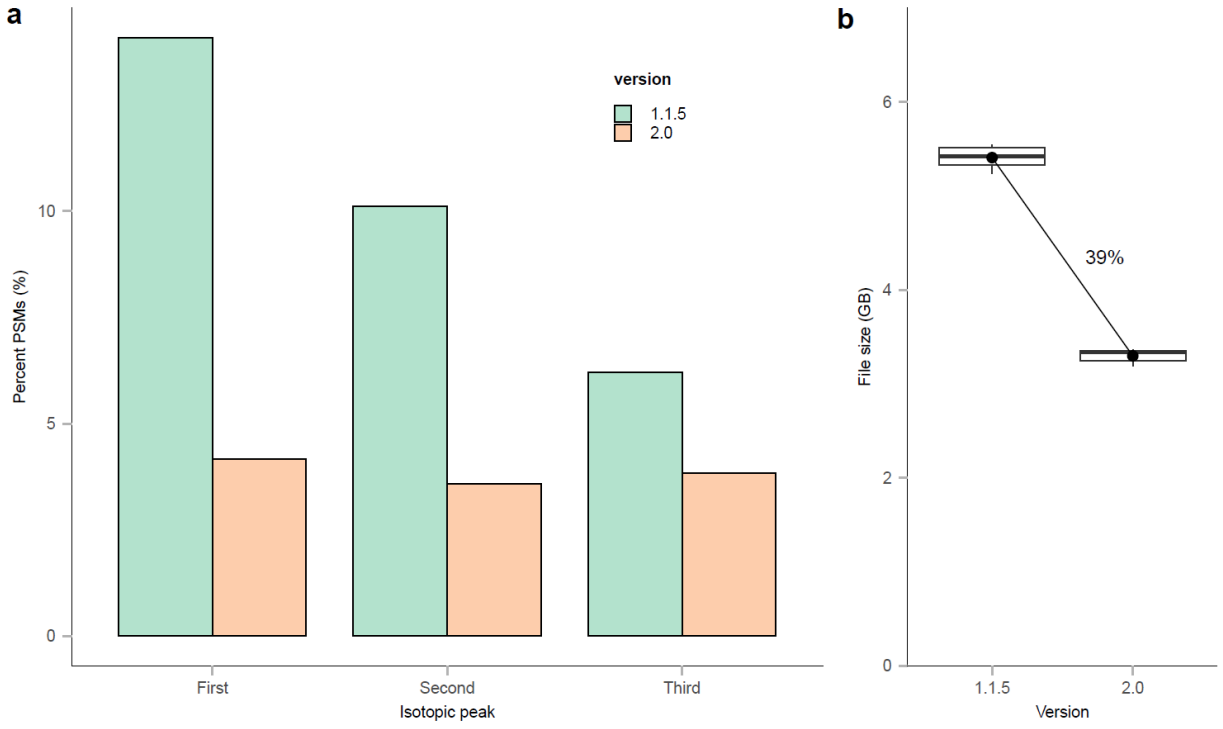


Figure 4-1. Improved precursor isotope filtering in diaTracer 2.0.

(a) Percentage of PSMs assigned to the first, second, and third isotopic peaks in open modification searches using pseudo-MS/MS spectra generated by diaTracer version 1.1.5 (green) and diaTracer version 2.0 (orange). (b) File sizes of pseudo-MS/MS mzML spectra generated by diaTracer 1.1.5 and 2.0. Numbers indicate the average percentage reduction in file size achieved by diaTracer 2.0 relative to version 1.1.5.

In addition to improved isotope handling, diaTracer 2.0 was optimized for computational efficiency to meet the growing demands of large-scale diaPASEF studies, which involve hundreds of samples analyzed within a single study. These optimizations include redesigned internal data structures, reduced task scheduling overhead, improved multithreading strategies, and a more efficient seeded local-maximum detection algorithm for feature extraction (details in Methods).

To benchmark these performance improvements, we analyzed the data acquired from 200 ng HeLa samples using short- and long-gradient acquisition methods (200 SPD, 100 SPD, and 60 SPD) from the dataset PXD017703[43]. For each dataset, we measured the wall-clock time required to extract features and generate pseudo-MS/MS spectra. Across all gradient lengths, diaTracer 2.0 demonstrated reductions in runtime relative to version 1.1.5, with the most pronounced improvements observed for longer gradient acquisitions. Specifically, processing times were reduced by 60%, 55.5%, and 48.3% for the 60 SPD, 100 SPD, and 200 SPD datasets, respectively (Figure 4-2).

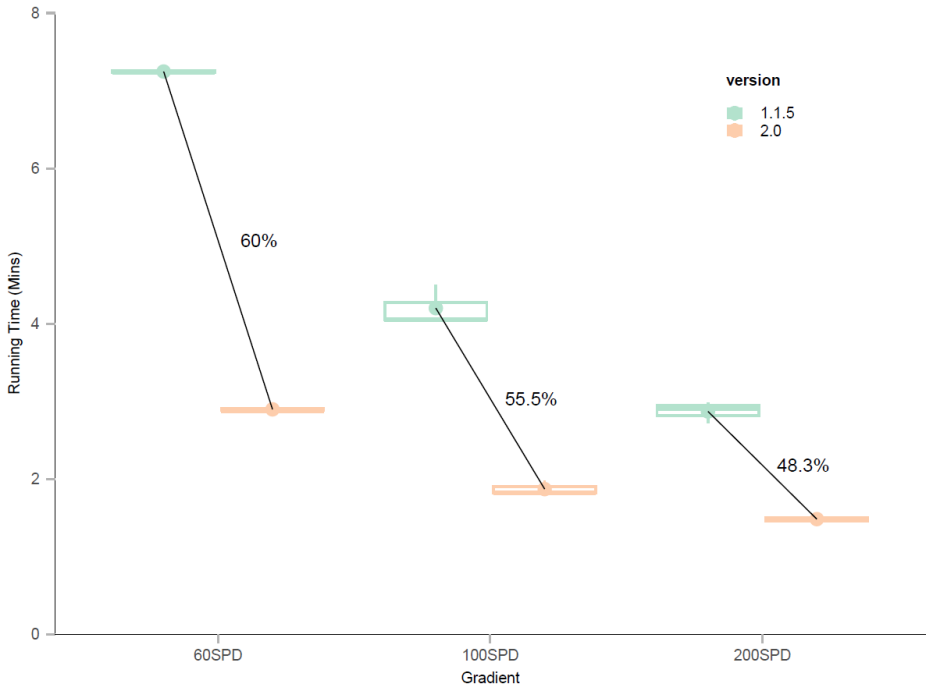


Figure 4-2. Computational performance benchmark of diaTracer 2.0.

Comparison of running times required to generate pseudo-MS/MS spectra using diaTracer version 1.1.5 (green) and diaTracer version 2.0 (orange) across diaPASEF datasets acquired with different gradient complexities (60 SPD, 100 SPD, and 200 SPD). Numbers indicate the average percentage reduction in runtime achieved by diaTracer 2.0 relative to version 1.1.5, demonstrating substantial performance gains, particularly for long-gradient acquisitions.

Together, these results demonstrate that diaTracer 2.0 achieves both higher-quality pseudo-MS/MS spectra and markedly improved computational performance. The combination of more effective isotope filtering and significantly reduced runtimes enhances the scalability of spectrum-centric diaPASEF analysis and enables efficient processing of large cohorts, positioning diaTracer 2.0 as a practical and robust solution for high-throughput and large-scale PASEF-based proteomics studies.

4.3.2 Spectrum-centric analysis of synchro-PASEF data using diaTracer

In synchro-PASEF acquisitions, fragment ions originating from a single precursor are distributed into multiple ion mobility slices as the quadrupole selection is synchronized with the ion mobility separation. While this acquisition strategy improves sampling efficiency, it disrupts conventional precursor–fragment relationships that are required for spectrum-centric data analysis. Without reconstruction, fragment ion signals from a given precursor are fragmented across multiple diagonal isolation slices and cannot be directly assembled into DDA-like spectra.

To address this challenge, diaTracer 2.0 introduces the concept of pseudo-isolation windows, which approximate the role of conventional vertical quadrupole windows in diagonal diaPASEF data. By grouping diagonal isolation slices that intersect a given pseudo-window in the m/z–$1/K_0$ plane, diaTracer reconstructs fragment ion signals into coherent fragment ion maps (Figure 4-3). Fragment features are then detected within each pseudo-isolation window and associated with precursor features based on retention time, ion mobility, and chromatographic correlation. This reconstruction restores the precursor–fragment relationships necessary for spectrum-centric analysis and enables the generation of high-quality, DDA-like pseudo-MS/MS spectra from synchro-PASEF data.

Following pseudo-MS/MS generation, the reconstructed spectra can be processed using the standard FragPipe workflow, including database searching, rescoring, spectral library construction, and DIA-based quantification.

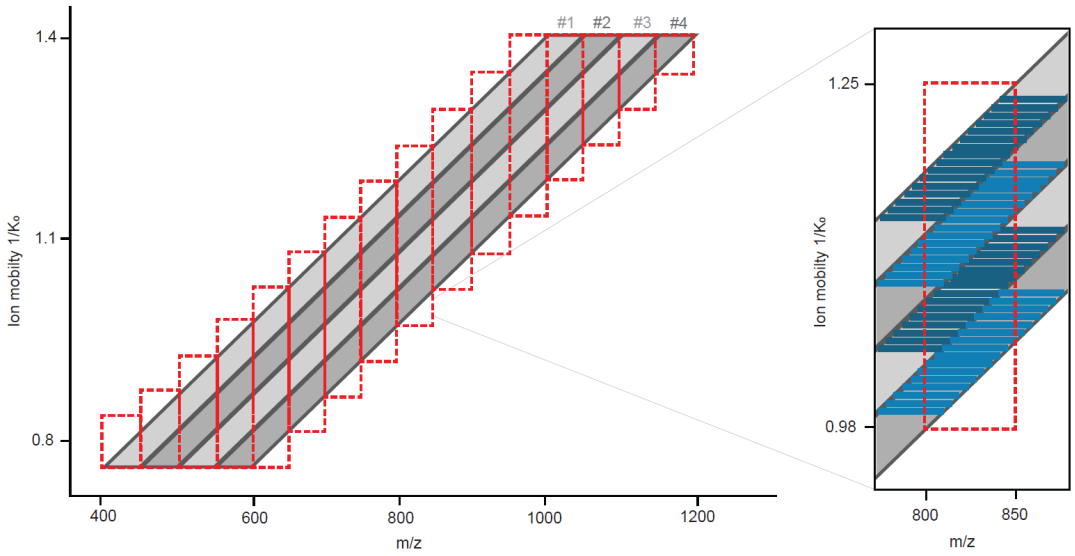


Figure 4-3. Pre-processing of synchro-PASEF data in diaTracer.

Schematic illustration of the pseudo-isolation window strategy used to process synchro-PASEF data. Red vertical boxes indicate manually defined pseudo-isolation windows introduced to reconstruct fragment ion signals that are sliced across the m/z and $1/K_0$ dimensions. The right panel shows a single pseudo-isolation window. Each blue line represents an individual MS/MS slice acquired during synchro-PASEF. Only MS/MS slices whose m/z ranges overlap with the pseudo-isolation window by at least the defined threshold (20% by default) are retained and aggregated for downstream spectrum-centric analysis.

4.3.3 Comparison of different diaPASEF acquisition strategies using FragPipe

In addition to synchro-PASEF, Slice-PASEF represents another recently proposed diagonal diaPASEF acquisition strategy, in which the ion mobility dimension is divided into a limited number of discrete slices per cycle. The updated diaTracer framework supports spectrum-centric analysis of Slice-PASEF data alongside synchro-PASEF and conventional diaPASEF, enabling a unified and unbiased comparison of different acquisition strategies within the same FragPipe-based analysis pipeline.

We applied FragPipe with diaTracer 2.0 to the benchmark dataset reported by Konno et al.[105], in which thin-diaPASEF was introduced and systematically compared with conventional diaPASEF (implemented via py_diAID[106]), synchro-PASEF, and Slice-PASEF (1-frame and 4-frame variants). In this dataset, 100 ng of HEK293 cell digest was analyzed in five technical replicates per acquisition strategy using a 50-minute LC gradient. For all methods, diaTracer 2.0 was used to generate pseudo-MS/MS spectra, which were subsequently processed by FragPipe to perform identification and quantification.

Across the evaluated acquisition strategies, conventional diaPASEF yielded the highest number of precursor-level quantifications (130,749), followed closely by thin-diaPASEF

(127,939), synchro-PASEF (94,932), Slice-PASEF-1F (62,219), and Slice-PASEF-4F (61,452) (Figure 4-4a). A similar trend was observed at the protein level (Figure 4-4b), with thin-diaPASEF producing the highest average number of protein identifications (8,342), followed by conventional diaPASEF (8,061), synchro-PASEF (7,323), Slice-PASEF-4F (6,279), and Slice-PASEF-1F (6,237).

To further assess quantitative robustness, we compared the numbers of quantified precursors and proteins under different coefficients of variation (CV) filtering thresholds across technical replicates. While thin-diaPASEF consistently maintained higher overall proteome coverage across CV cutoffs, Slice-PASEF-1F demonstrated superior reproducibility, retaining a large fraction of quantified precursors and proteins at strict CV thresholds (Figure 4-4c and d). This suggests that the simpler slicing geometry of Slice-PASEF-1F may reduce run-to-run variability, with the cost of reduced identification and quantification depth.

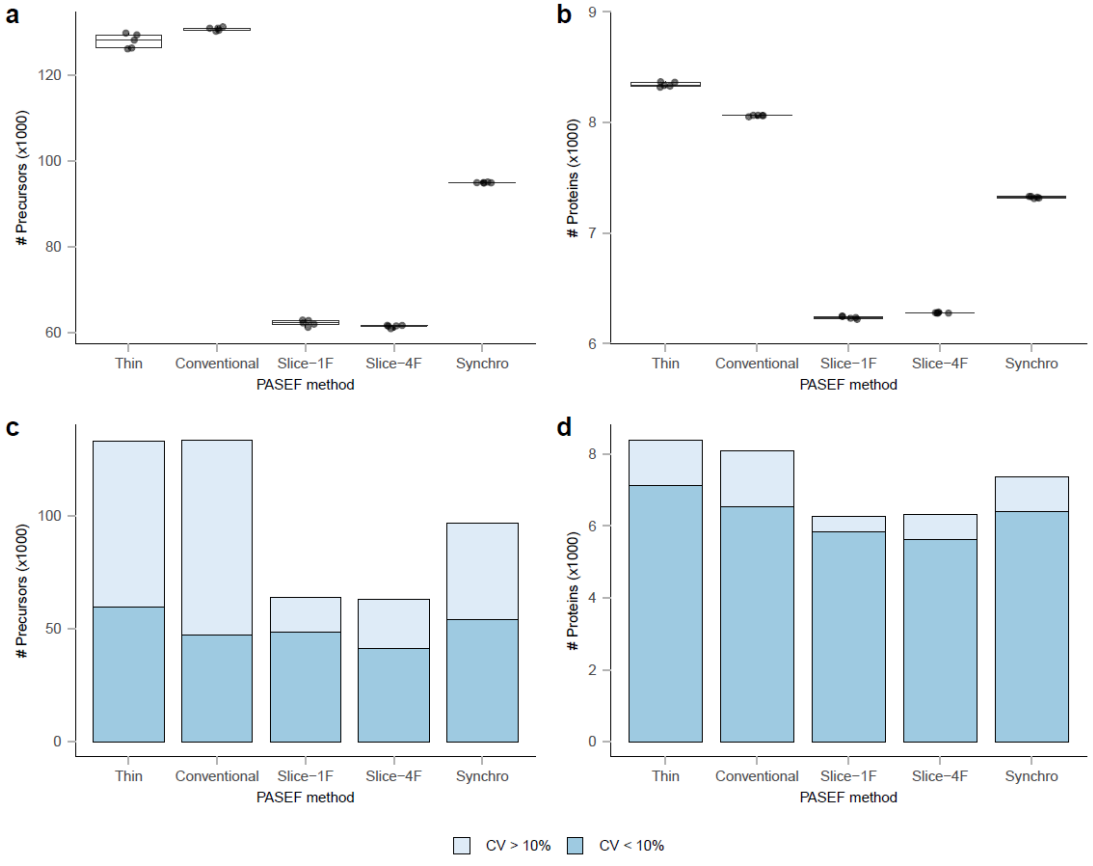


Figure 4-4. Performance comparison of different diaPASEF acquisition strategies analyzed by FragPipe.

Quantitative performance benchmarking of five diaPASEF acquisition methods—thin-diaPASEF, conventional diaPASEF implemented via py-diAID, Slice-PASEF 1-frame (Slice-1F), Slice-PASEF 4-frame (Slice-4F), and synchro-PASEF—using dataset JPST003679. (a) Number of quantified precursors. (b) Number of quantified

proteins. (c) Number of quantified precursors after coefficient of variation (CV) filtering under the conditions shown in (a). (d) Number of quantified proteins after CV filtering under the conditions shown in (b). All results are based on five technical replicates of 100 ng HEK293 cell digests acquired with a 50-minute gradient.

Importantly, the trends observed in diaTracer-based analyses closely mirrored those reported in the original study (Table 4-1), confirming that spectrum-centric analysis of synchro-PASEF and Slice-PASEF data preserves the relative performance characteristics of different diaPASEF acquisition strategies. Collectively, these results demonstrate that diaTracer-integrated FragPipe enables robust, consistent, and unbiased spectrum-centric analysis across conventional and diagonal diaPASEF methods, facilitating direct methodological comparisons within a single computational framework.

Table 4-1. Numbers of quantified precursors and proteins reported in the original study.

Values marked with an asterisk (*) were taken directly from the main text of the original study and were generated using DIA-NN version 1.9.2 without additional filtering.

	FragPipe		Konno et al. *	
	Precursors	Proteins	Precursors	Proteins
Thin-diaPASEF	127,939	8,342	126,918	8,734
diaPASEF	130,749	8,061	136,212	8,636
Slice-PASEF 1F	62,219	6,237	82,638	7,748
Slice-PASEF 4F	61,452	6,279	88,258	7,958
Synchro-PASEF	94,932	7,323	94,156	8,225

4.4 Discussion

diaPASEF has demonstrated excellent performance across a wide range of quantitative proteomics applications by combining data-independent acquisition with ion mobility separation, thereby improving ion utilization, sensitivity, and quantitative reproducibility. To address the analytical complexity introduced by this four-dimensional data structure, we previously introduced diaTracer, a spectrum-centric framework that deconvolutes diaPASEF signals into DDA-like pseudo-MS/MS spectra. Through its integration with the FragPipe ecosystem, diaTracer enables a complete and flexible workflow for diaPASEF data analysis, supporting conventional database search strategies while preserving the advantages of DIA-based acquisition.

In this chapter, we extend the scope of diaTracer to support diagonal diaPASEF acquisition strategies, including synchro-PASEF and Slice-PASEF. These methods enhance

fragment ion sampling efficiency by continuously or discretely partitioning precursor space across the m/z and ion mobility dimensions. However, they fundamentally alter the MS2 data structure by eliminating the fixed, vertical isolation windows that are assumed by traditional spectrum-centric approaches. To overcome this challenge, we introduce a pseudo-isolation window strategy that reconstructs fragmented MS/MS signals from diagonal acquisition geometries and restores coherent precursor–fragment relationships. Importantly, this reconstruction is performed as a preprocessing step, allowing the downstream diaTracer workflow—feature detection, precursor–fragment association, and pseudo-MS/MS generation—to remain unchanged. Using this approach, diaTracer enables spectrum-centric analysis of diagonal diaPASEF data in a manner that is both conceptually consistent and computationally efficient.

Beyond expanding methodological coverage, diaTracer 2.0 addresses practical challenges associated with the increasing scale of diaPASEF-based proteomics studies. Modern experiments frequently involve hundreds to thousands of samples, placing growing demands on computational efficiency, memory usage, and data reduction. While diaTracer 1.0 already demonstrated competitive performance compared with existing tools, diaTracer 2.0 introduces redesigned data structures, improved isotope filtering, and optimized peak extraction algorithms that substantially reduce runtime and spectral redundancy. These improvements are particularly impactful for long-gradient diaPASEF data.

Collectively, these advances position diaTracer 2.0 as a scalable and versatile solution for spectrum-centric analysis of both conventional and diagonal diaPASEF data. Its seamless integration within the FragPipe workflow enables unified, end-to-end analysis of diverse PASEF acquisition strategies within a single computational framework. As diaPASEF methods continue to evolve and are increasingly adopted in large cohort studies and time-sensitive clinical proteomics applications, the flexibility, robustness, and computational efficiency of diaTracer 2.0 provide a strong foundation for future methodological and translational developments.

4.5 Data availability

The raw MS/MS files used in this study can be found through the ProteomeXchange Consortium via the PRIDE partner repository[81], Japan ProteOme STandard Repository[104], or at the MassIVE repository with the following accession codes:

- Different acquisition strategies evaluation data JPST003679:

https://repository.jpostdb.org/entry/JPST003679

- Speed performance benchmark data PXD017703: https://www.ebi.ac.uk/pride/archive/projects/PXD017703

4.6 Acknowledgements and competing interests

This work was supported in part by National Institutes of Health grants R01-GM-094231 and U24-CA271037.

A.I.N. and F.Y. receive royalties from the University of Michigan for the sale of MSFragger, IonQuant, and diaTracer software licenses to commercial entities. K.L. receives royalties from the University of Michigan for the sale of diaTracer software licenses to commercial entities. All license transactions are managed by the University of Michigan Innovation Partnerships office, and all proceeds are subject to university technology transfer policy. Other authors declare no other competing interests.

4.7 Authors, affiliations, and contributions

Kai Li[1], Guo Ci Teo[2], Fengchao Yu[2] & Alexey I. Nesvizhskii[1,2]

[1]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA

K.L. and A.I.N. developed the diaTracer algorithm. K.L. wrote the software and analyzed the results. F.Y. and G.C.T. assisted with the algorithm and software development. F.Y. assisted with the integration of diaTracer into FragPipe. K.L., F.Y., and A.I.N. wrote the manuscript. A.I.N. conceived the study. A.I.N. and F.Y. supervised the study.

Chapter 5 Conclusions and Future Directions

5.1 Conclusions

In this dissertation, I developed a spectrum-centric computational framework for the analysis of four-dimensional diaPASEF proteomics data, addressing key methodological and computational challenges associated with modern data-independent acquisition (DIA) workflows. By combining algorithmic innovation with scalable software design, this work expands the analytical reach of ion mobility–enabled proteomics and provides practical solutions for both routine quantitative studies and complex discovery-oriented applications.

The primary contribution of this work is diaTracer, a spectrum-centric tool that deconvolutes diaPASEF data into data-dependent acquisition (DDA)-like pseudo-tandem mass spectrometry (MS/MS) spectra. By modeling the four dimensions of diaPASEF data—mass-to-charge ratios (m/z), retention time, ion mobility, and intensity—diaTracer reconstructs coherent precursor–fragment relationships that are obscured by DIA multiplexing. This design enables direct compatibility with established DDA database search engines and downstream statistical frameworks, effectively bridging the gap between DIA acquisition and DDA-centric identification.

In Chapter 2, I established the core methodology of diaTracer and demonstrated its seamless integration into the FragPipe computational ecosystem. Extensive benchmarking across multiple biological datasets showed that spectrum-centric diaPASEF analysis achieves identification depth and quantitative performance comparable to, and in many cases exceeding, peptide-centric DIA strategies. Importantly, these results were obtained without reliance on in-silico spectral library prediction, highlighting the robustness of the spectrum-centric approach in scenarios where prior assumptions about peptide composition, digestion specificity, or modification state are limited or unreliable.

In Chapter 3, I demonstrated that the advantages of spectrum-centric analysis become particularly important in the context of complex diaPASEF datasets. Applications such as semi-tryptic analysis, open and mass-offset searches, phosphoproteomics, and human leukocyte

antigen (HLA) immunopeptidomics are difficult for peptide-centric methods because they generate an extremely large number of candidate peptides and rely heavily on accurate predictions. Using diaTracer, I showed that such analyses can be performed efficiently and reproducibly on diaPASEF data, enabling the detection of biologically meaningful proteolytic events, improved site localization of post-translational modifications, and sensitive identification of endogenous and immunopeptides. These results underscore the value of spectrum-centric DIA analysis for discovery-driven proteomics, where the goal extends beyond protein quantification to mechanistic and functional insight.

In Chapter 4, I extended diaTracer to address recent advances in diaPASEF acquisition, including synchro-PASEF and Slice-PASEF, which employ diagonal scanning strategies to improve sampling efficiency. These acquisition modes fundamentally alter the structure of MS2 data and challenge existing spectrum-centric assumptions. To overcome this limitation, I introduced a pseudo-isolation window strategy that reconstructs fragmented MS/MS information and restores precursor–fragment coherence. Together with substantial algorithmic optimizations—such as improved isotope filtering, optimized peak detection, and redesigned multithreading—diaTracer 2.0 enables efficient, scalable spectrum-centric analysis of both conventional and diagonal diaPASEF data. Benchmarking against published datasets confirmed that diaTracer preserves the relative performance characteristics of different acquisition strategies while substantially reducing computational processing time and data redundancy.

A central strength of this work is the adoption of a spectrum-centric analysis strategy, which substantially broadens the analytical scope of diaPASEF data beyond the limitations of peptide-centric strategies. By reconstructing DDA-like pseudo-MS/MS spectra directly from DIA data, diaTracer enables unrestricted database search workflows, including semi-enzymatic and nonspecific digestion, open and mass-offset modification searches, and comprehensive post-translational modification characterization. In peptide-centric workflows, such analyses are often computationally infeasible or methodologically constrained because they require exhaustive in-silico prediction of all candidate peptides and modifications.

Beyond expanding the search space, this spectrum-centric design also provides important advantages in practical scalability and reusability. Modern proteomics studies increasingly involve large cohorts, multiple experimental conditions, and iterative hypothesis testing, placing growing demands on computational efficiency, storage, and analytical flexibility. By generating

reusable pseudo-MS/MS spectra and decoupling signal extraction from downstream peptide identification, diaTracer enables rapid reanalysis under different search settings—such as changing enzyme specificity, modification lists, or mass-offset definitions—without requiring repeated processing of the raw diaPASEF data. This separation of concerns substantially reduces computational overhead and facilitates exploratory and iterative analyses.

Integration of diaTracer into the FragPipe ecosystem further amplifies these advantages by providing seamless access to advanced peptide identification, deep-learning-based rescoring, protein inference, rigorous false discovery rate (FDR) control, spectral library construction, DIA-based quantification, and data visualization. Together, these capabilities establish diaTracer and FragPipe as a unified, extensible, and discovery-oriented framework for diaPASEF data analysis, particularly well-suited for complex biological questions that extend beyond conventional tryptic protein quantification.

In conclusion, this dissertation establishes diaTracer as a flexible, efficient, and versatile framework for spectrum-centric analysis of diaPASEF data. The methods and insights presented here lay a strong foundation for future developments in computational proteomics and contribute to the continued evolution of DIA as a powerful and general-purpose strategy for biological and biomedical discovery.

5.2 Future directions

Although spectrum-centric strategies for DIA analysis offer distinct advantages for large-scale and discovery-oriented proteomics, peptide-centric methods retain complementary strengths. During benchmarking of diaTracer (spectrum-centric) against DIA-NN[51] (peptide-centric), we observed that certain peptide signals identified by peptide-centric approaches were absent from diaTracer-extracted results. These observed differences highlight important opportunities for further methodological development.

One contributing factor is signal intensity. Spectrum-centric methods rely on data-driven feature extraction and therefore require detectable and spatially coherent signals across m/z, ion mobility, and retention time. Extremely low-abundance features, particularly those appearing in only a single frame or represented by a small number of weak signal points, may fall below extraction thresholds and be discarded as noise. This limitation is worse at the MS2 level, where fragment ion signals are relatively weaker. As a result, pseudo-MS/MS spectra generated by spectrum-centric approaches may be incomplete, potentially reducing identification sensitivity

even when some informative fragment ions are present. In contrast, peptide-centric methods leverage predicted spectral libraries as hypotheses and directly query raw data at predicted m/z, retention time, and ion mobility coordinates. This targeted strategy allows peptide-centric tools to recover weak signals that would otherwise be excluded during feature extraction.

A second limitation arises from the correlation-based clustering strategy currently used in diaTracer to associate precursor and fragment features. Correlation metrics are effective for high-quality, well-shaped chromatographic signals but can fail when fragment intensities are low or when precursor and fragment elution profiles are imperfectly aligned. Relaxing correlation thresholds increases sensitivity but also introduces additional noise, highlighting a fundamental tradeoff. Peptide-centric methods reduce this issue by restricting candidate fragment ions based on predicted peptide spectra, effectively filtering out unrelated noise even when correlation is weak. However, this advantage is strongly dependent on the accuracy of prediction models and scoring algorithms, and peptide-centric tools often struggle to maintain well-calibrated false discovery rates under complex or expanded search conditions.

Despite these challenges, peptide-centric approaches are known to suffer from difficulties in controlling the true FDR[107], particularly in large or heterogeneous search spaces. Motivated by these observations, the following directions outline a roadmap for improving DIA analysis by extending diaTracer and exploring hybrid strategies.

5.2.1 Continued algorithmic development of diaTracer

As diaPASEF continues to gain adoption, additional acquisition variants and experimental designs are likely to emerge. Ongoing development of diaTracer will therefore focus on expanding support for the diaPASEF family while addressing current limitations in feature extraction.

At present, pseudo-MS/MS spectrum generation in diaTracer is anchored to MS1 precursor feature detection, which serves as the entry point for precursor–fragment association. Although rare, it is possible for high-quality fragment ion features to exist in the absence of a detectable MS1 precursor signal. To improve coverage in such cases, a future extension will incorporate an MS2-only clustering strategy, enabling high-confidence fragment features without matched MS1 signals to seed pseudo-MS/MS spectrum construction.

In addition, diagonal diaPASEF acquisition strategies such as synchro-PASEF and Slice-PASEF introduce deterministic relationships between precursor and the distribution of fragment

ions across adjacent ion mobility slices. These structured slicing patterns encode valuable precursor–fragment and fragment–fragment relationships that are not yet fully utilized in diaTracer 2.0. Future versions will incorporate these geometric constraints to improve fragment grouping, refine pseudo-MS/MS spectra, and enhance robustness against noise.

5.2.2 Development of a hybrid spectrum- and peptide-centric strategy

Spectrum-centric methods excel in discovery-driven applications—such as nonspecific digestion, open modification searches, and endogenous peptidomics—because they avoid enumeration of candidate peptides prior to signal extraction. However, their sensitivity to extremely low-abundance features remains limited. Peptide-centric methods, by contrast, leverage prior knowledge encoded in predicted libraries and are more tolerant of weak or partially observed signals.

A promising direction is therefore the development of a hybrid strategy that integrates spectrum-centric and peptide-centric principles. In such a framework, diaTracer would continue to perform comprehensive signal extraction, preserving its ability to support large and poorly defined search spaces. Following this initial extraction, peptide-centric indexing could be applied: candidate peptides generated from user-defined search parameters would be indexed by m/z, retention time, and ion mobility, allowing extracted MS1 and MS2 features to be mapped back to possible peptide hypotheses.

For candidates supported by partial evidence, such as a small number of fragment features or weak precursor signals, the algorithm could revisit raw data within localized retention time and ion mobility neighborhoods to recover signals missed during the initial extraction pass. This approach leverages the assumption that even very low-abundance peptides must exhibit some consistent signal structure across dimensions. Moreover, run-specific retention time and ion mobility information derived from confidently identified peptides could be used to refine predictions through transfer learning, enabling more precise extraction and scoring in subsequent iterations.

5.2.3 Toward a unified identification and quantification framework

The methodological advances outlined above naturally lead to a broader goal: the development of a comprehensive DIA analysis framework that tightly integrates identification and quantification rather than treating them as loosely coupled, sequential steps. While current

DIA workflows in FragPipe—including those enabled by diaTracer—typically generate a spectral library that is subsequently passed to an external quantification engine, this separation limits the ability to fully utilize the rich feature-level information extracted during spectrum-centric analysis.

Spectrum-centric processing inherently produces detailed, multi-dimensional descriptors for both precursor and fragment features, including chromatographic profiles, ion mobility distributions, inter-feature correlations, isotopic structure, and confidence metrics derived from pseudo-MS/MS spectrum reconstruction. Much of this information is either discarded or reduced to simple identifiers when exported as a conventional spectral library. Existing DIA quantification tools generally accept only a narrow subset of inputs—such as m/z values, retention times, and fragment ion lists—preventing the incorporation of additional evidence that could improve quantitative accuracy and statistical validation.

A unified framework would retain and propagate this information throughout the entire analysis pipeline. In such a system, pseudo-MS/MS spectra generated during identification would not merely serve as static library entries, but as anchors for targeted, feature-aware signal extraction during quantification. MS1 and MS2 signals could be re-extracted using adaptive boundaries informed by feature geometry, ion mobility consistency, and precursor–fragment relationships established during the identification stage. Rather than treating each fragment independently, quantification could model groups of correlated fragments and their joint agreement with precursor behavior.

Importantly, this integrated approach would enable the use of identification-derived confidence metrics directly in quantitative scoring and false discovery rate estimation. For example, precursor-level confidence, fragment consistency scores, isotope coherence, and ion mobility alignment could be incorporated into probabilistic models for both detection and quantification. Such joint modeling has the potential to improve robustness against interference, reduce missing values, and provide better-calibrated confidence estimates, particularly in complex or low-input datasets.

Finally, such a framework would position spectrum-centric DIA analysis as a truly end-to-end solution, capable of supporting both hypothesis-driven quantification and open-ended discovery within a single computational environment. By unifying signal extraction, identification, rescoring, quantification, and statistical validation, this approach would maximize the informational yield of DIA data and provide a flexible foundation for future methodological innovation. As DIA acquisition strategies and experimental designs continue to evolve, an

integrated identification–quantification platform will be critical for translating increasingly complex raw data into reliable biological insight.

Appendices

Appendix A: diaTracer Manual

A.1 Introduction

diaTracer is a computational tool that enables spectrum-centric analysis of Bruker's diaPASEF data-independent acquisition proteomics data, facilitating direct ("spectral-library free") peptide identification and quantification. diaTracer processes diaPASEF raw mass spectrometry data (.d files) and performs three-dimensional (m/z, retention time, ion mobility) peak tracing and feature detection, groups precursor and fragment signals, and generates "pseudo-MS/MS" spectra (in mzML format). These pseudo-MS/MS spectra can then be processed as DDA spectra using MSFragger or any other search engine. diaTracer supports analysis of any diaPASEF proteomics data, including data requiring semi-tryptic (e.g. N-terminomics) or nonspecific (e.g. HLA immunopeptidomics) searches, searches allowing for chemical (e.g. chemical proteomics) or biological modifications (e.g. phosphoproteomics). diaTracer is fast, making direct DIA analysis of large sample cohorts possible. Furthermore, diaTracer enables unrestricted identification of post-translational modifications from diaPASEF data using open/mass offset searches.

diaTracer is available as a standalone tool and can also be run as part of the FragPipe (https://fragpipe.nesvilab.org/) computational platform.

A.2 System requirements

- diaTracer is written in Java and requires at least Java version 11.
- diaTracer requires the "ext" folder from the latest MSFragger (https://msfragger.arsci.com/upgrader/) for Bruker data parse.
- The latest FragPipe (https://github.com/Nesvilab/FragPipe/releases/latest) is optional for whole analysis.

A.3 License

diaTracer is available freely for academic research, non-commercial or educational purposes under academic license (https://msfragger-upgrader.nesvilab.org/diatracer/).

Other uses require a commercial license that can be obtained by visiting Fragmatics (https://www.fragmatics.com/) or emailing at info@fragmatics.com.

A.4 Run diaTracer

- In FragPipe: Download the latest FragPipe (https://github.com/Nesvilab/FragPipe/releases/latest) and follow the tutorial.
- Standalone usage in command line interface:
  Run java -jar diaTracer-1.1.4.jar to get options.

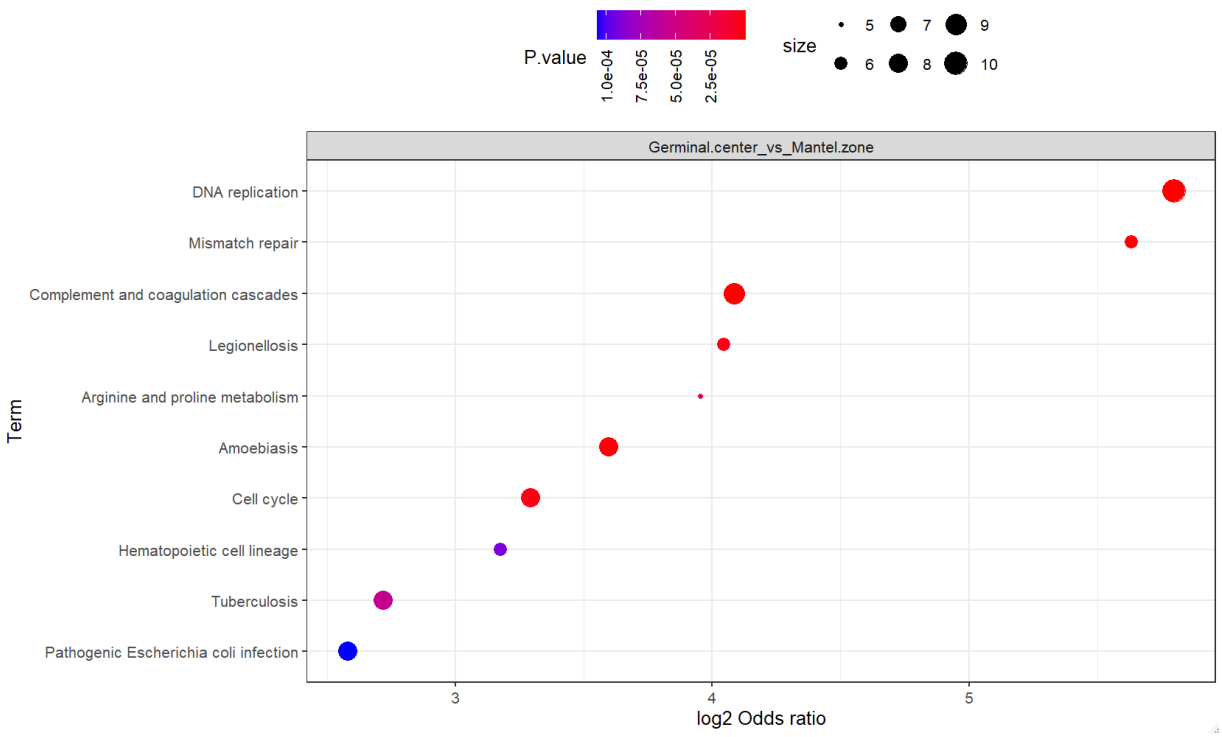
```
java -jar diaTracer-1.1.4.jar --dFilePath <.d file path> --threadNum <thread number>

options:
 -d,--dFilePath <arg>           .d file path
 -dI,--deltaApexIM <arg>        Ion mobility delta range for ms1 and ms2
                                match. default 0.01
 -dR,--deltaApexRT <arg>        Apex scan delta range for ms1 and ms2
                                match. default 3
 -h,--help                      Help
 -help                          Help
 -mC,--ms1MS2Corr <arg>         MS1 and MS2 correlation threshold.
                                default: 0.3
 -mF,--massDefectFilter <arg>   Apply mass defect filter. 1: apply; 0: not
                                apply. default: 1
 -mO,--massDefectOffset <arg>   Mass defect offset. default: 0.1
 -r,--writeInter <arg>          write inter files . 1: write; 0: not
                                write. default: 0
 -rM,--RFMax <arg>              Top N peaks in the spectrum. default: 500
 -t,--threadNum <arg>           thread number
 -w,--workDir <arg>             work directory
```

Example: java -jar diaTracer-1.1.4.jar --dFilePath ./20200505_Evosep_100SPD_SG06-16_MLHeLa_100ng_py8_S2-C1_1_2731.d --workDir ./ --writeInter 0 --deltaApexIM 0.01 --deltaApexRT 3 --ms1MS2Corr 0.3 --massDefectFilter 0 --massDefectOffset 0.1 --RFMax 500 --threadNum 12

After running the above command, a mzML file named 20200505_Evosep_100SPD_SG06-16_MLHeLa_100ng_py8_S2-

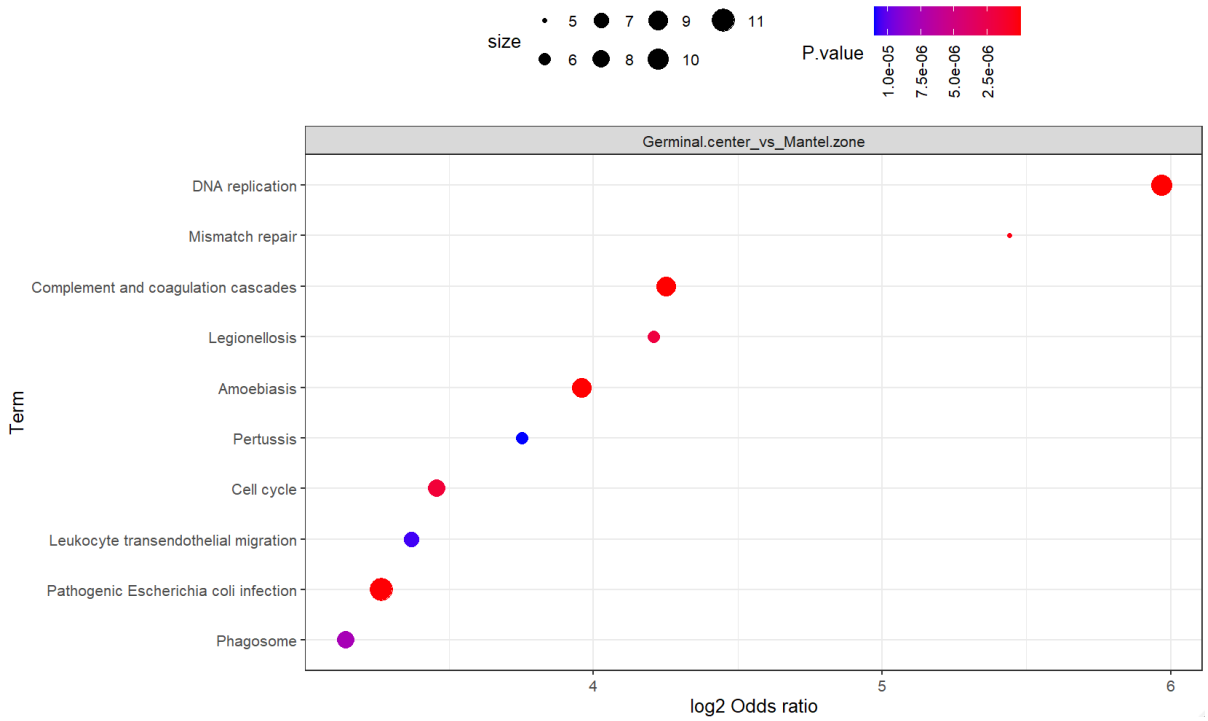
C1_1_2731_diaTracer.mzML will be generated in the same path of .d file. The running time will be around 10 minutes using 12 threads.

Appendix B: Supplement Data for Chapter 2



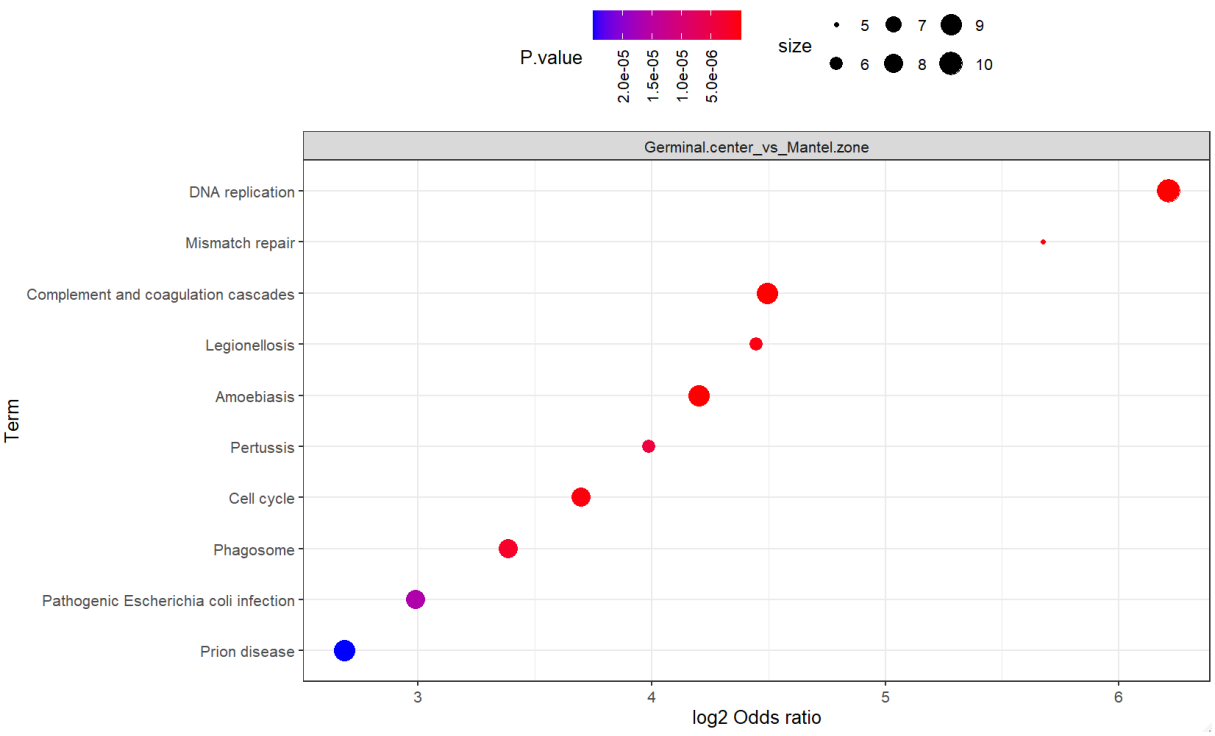
Appendix Figure B-1. GC group pathway enrichment analysis using results from Makhmut et al.[67].

Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the GC group using result from Makhmut et.al.



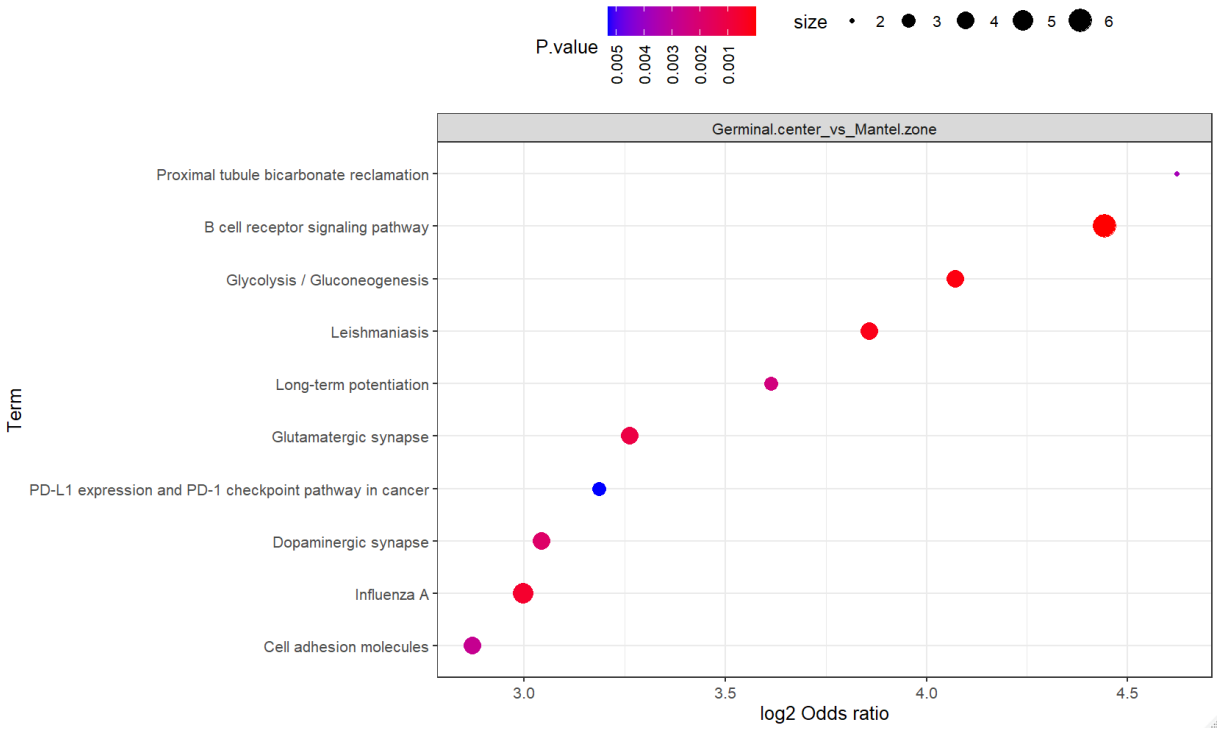
Appendix Figure B-2. GC group pathway enrichment analysis using results from FP-diaTracer high-input library.

Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the GC group using result from FP-diaTracer high-input library.



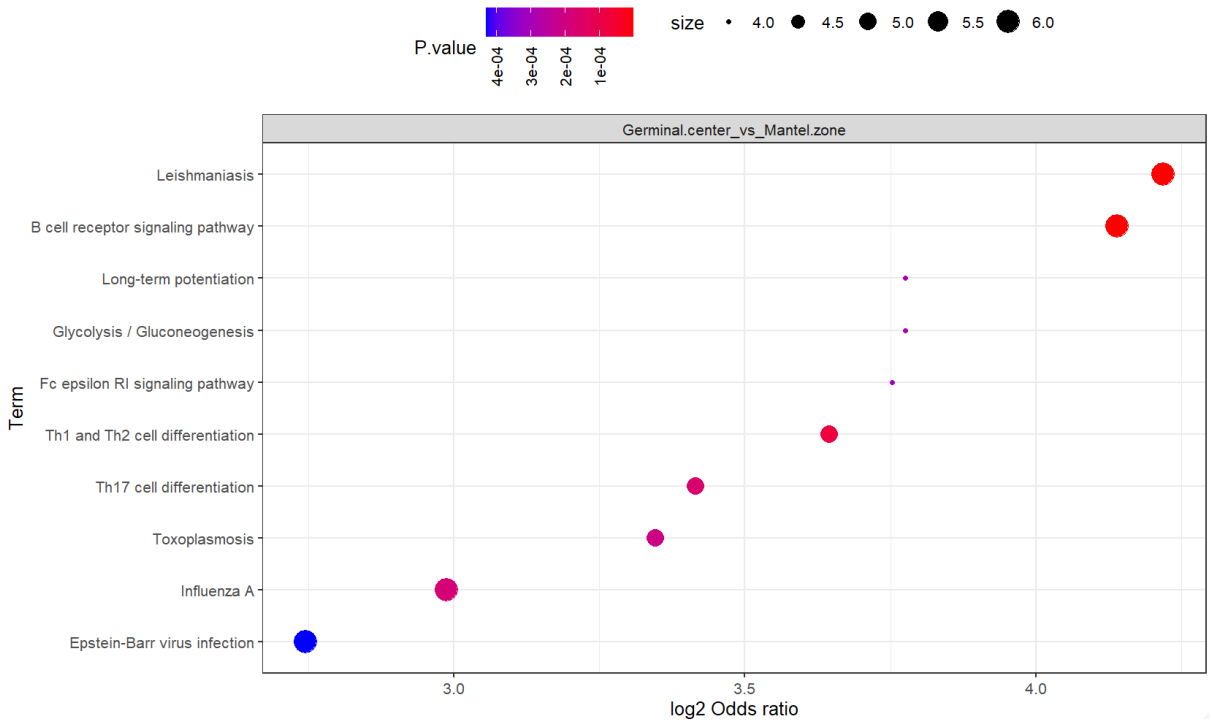
Appendix Figure B-3. GC group pathway enrichment analysis using results from FP-diaTracer.

Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the GC group using result from FP-diaTracer.



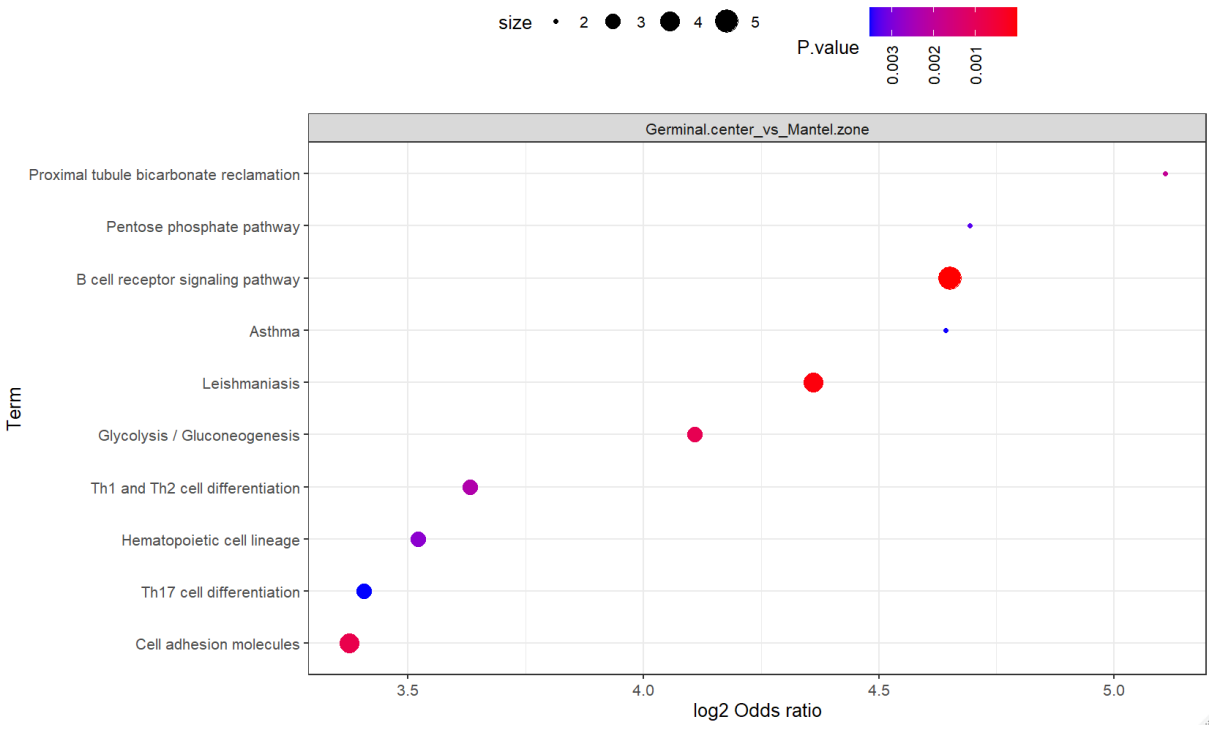
Appendix Figure B-4. MZ group pathway enrichment analysis using results from Makhmut et al.

Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the MZ group using result from Makhmut et.al.



Appendix Figure B-5. MZ group pathway enrichment analysis using results from FP-diaTracer high-input library.

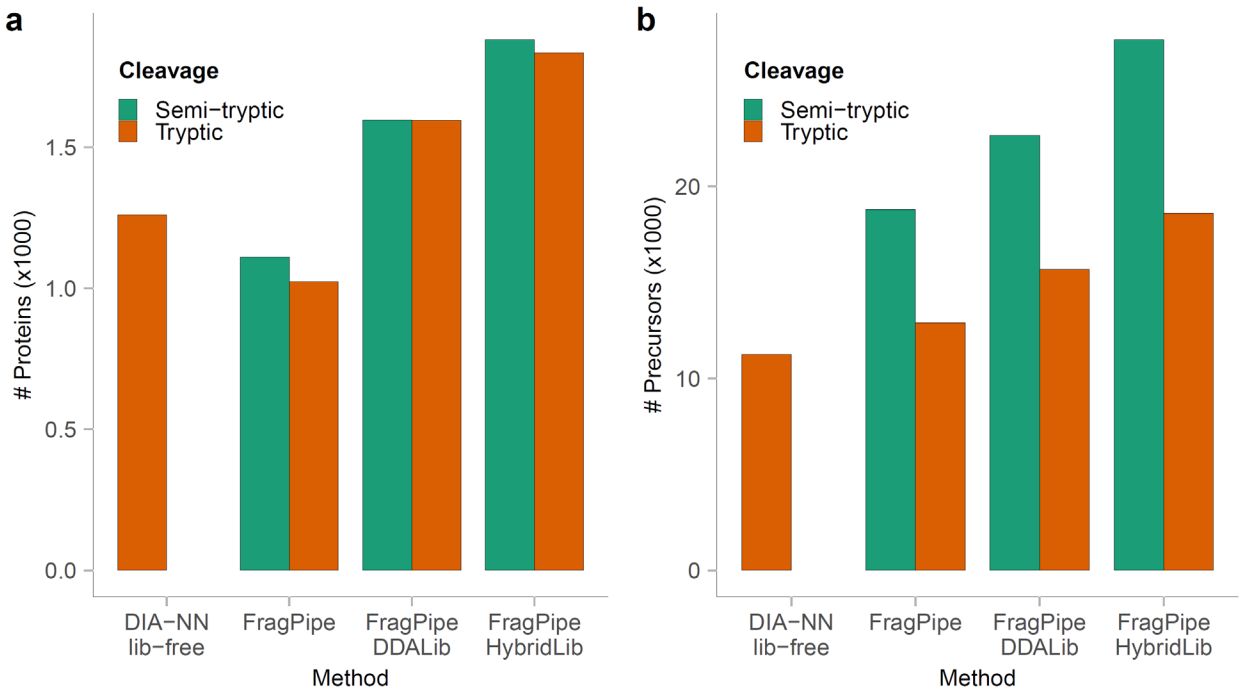
Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the MZ group using result from FP-diaTracer high-input library.



Appendix Figure B-6. MZ group pathway enrichment analysis using results from FP-diaTracer.

Pathway enrichment analysis based on KEGG database of the low-input dataset between GC (germinal center) and MZ (mantle zone) groups showing higher enrichment in the MZ group using result from FP-diaTracer.

Appendix C: Supplement Data for Chapter 3



Appendix Figure C-1. Quantification numbers comparison is CSF dataset.

Performance evaluation of tryptic, semi-tryptic, and mass offset search using a CSF dataset of the 34 diaPASEF runs from 15 patients with Alzheimer's disease (AD) and 19 control subjects. a) Bar plot showing the total number of quantified proteins using different methods, with colors representing cleavage types (green: strict trypsin cleavage; orange: semi-tryptic cleavage). b) Bar plot showing the total number of precursors quantified in the CSF dataset using different methods.

Appendix Table C-1. 158 proteins containing semi-tryptic peptides mapping to the region immediately following the signal peptides in the N-terminal portion of the corresponding protein.

ProteinID	Entry Name	Gene Names	Length	Subcellular location [CC]	stop_pos
O00451	GFRA2_HUMAN	GFRA2 GDNFRB RETL2 TRNR2	464	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Lipid-anchor, GPI-anchor {ECO:0000250}.	21
O00533	NCHL1_HUMAN	CHL1 CALL	1208	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Single-pass type I membrane protein {ECO:0000250}. Note=Soluble forms produced by cleavage/shedding also exist. {ECO:0000250}.; SUBCELLULAR LOCATION: [Processed neural cell adhesion molecule L1-like protein]: Secreted, extracellular space, extracellular matrix {ECO:0000250}.	24
O14514	AGRB1_HUMAN	ADGRB1 BAI1	1584	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:12074842, ECO:0000269|PubMed:26838550}; Multi-pass membrane protein {ECO:0000255}. Cell projection, phagocytic cup {ECO:0000250|UniProtKB:Q3UHD1}. Cell junction, focal adhesion {ECO:0000250|UniProtKB:Q3UHD1}. Cell projection, dendritic spine {ECO:0000250|UniProtKB:C0HL12}. Postsynaptic density {ECO:0000250|UniProtKB:Q3UHD1}.; SUBCELLULAR LOCATION: [Vasculostatin-120]: Secreted {ECO:0000269|PubMed:15782143, ECO:0000269|PubMed:22330140}.; SUBCELLULAR LOCATION: [Vasculostatin-40]: Secreted {ECO:0000269|PubMed:22330140}.	30
O14594	NCAN_HUMAN	NCAN CSPG3 NEUR	1321	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:25326458}.	22
O14773	TPP1_HUMAN	TPP1 CLN2 GIG1 UNQ267/PRO304	563	SUBCELLULAR LOCATION: Lysosome {ECO:0000269|PubMed:19941651}. Melanosome {ECO:0000269|PubMed:12643545}. Note=Identified by mass spectrometry in melanosome fractions from stage I to stage IV. {ECO:0000269|PubMed:12643545}.	19
O15240	VGF_HUMAN	VGF	615	SUBCELLULAR LOCATION: [Neurosecretory protein VGF]: Secreted {ECO:0000269|PubMed:19194657}. Cytoplasmic vesicle, secretory vesicle {ECO:0000269|PubMed:19194657}. Note=Stored in secretory vesicles and then secreted, NERP peptides colocalize with vasopressin in the storage granules of hypothalamus.	22
O43405	COCH_HUMAN	COCH COCH5B2 UNQ257/PRO294	550	SUBCELLULAR LOCATION: Secreted, extracellular space, extracellular matrix {ECO:0000269|PubMed:12843317, ECO:0000269|PubMed:22610276}.	24
O95390	GDF11_HUMAN	GDF11 BMP11	407	SUBCELLULAR LOCATION: Secreted {ECO:0000305}.	24
P00738	HPT_HUMAN	HP	406	SUBCELLULAR LOCATION: Secreted.	18
P00747	PLMN_HUMAN	PLG	810	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:10077593, ECO:0000269|PubMed:14699093}. Note=Locates to the cell surface where it is proteolytically cleaved to produce the active plasmin. Interaction with HRG tethers it to the cell surface.	19
P00751	CFAB_HUMAN	CFB BF BFD	764	SUBCELLULAR LOCATION: Secreted.	25
P01008	ANT3_HUMAN	SERPINC1 AT3 PRO0309	464	SUBCELLULAR LOCATION: Secreted, extracellular space.	32
P01009	A1AT_HUMAN	SERPINA1 AAT PI PRO0684 PRO2209	418	SUBCELLULAR LOCATION: Secreted. Endoplasmic reticulum. Note=The S and Z allele are not secreted effectively and accumulate intracellularly in the endoplasmic reticulum.; SUBCELLULAR LOCATION: [Short peptide from AAT]: Secreted, extracellular space, extracellular matrix.	24
P01011	AACT_HUMAN	SERPINA3 AACT GIG24 GIG25	423	SUBCELLULAR LOCATION: Secreted.	23
P01023	A2MG_HUMAN	A2M CPAMD5 FWP007	1474	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:6203908}.	23
P01024	CO3_HUMAN	C3 CPAMD1	1663	SUBCELLULAR LOCATION: Secreted.	22
P01033	TIMP1_HUMAN	TIMP1 CLGI TIMP	207	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:1730286, ECO:0000269|PubMed:24635319, ECO:0000269|PubMed:3010309, ECO:0000269|PubMed:3839290, ECO:0000269|PubMed:3903517, ECO:0000269|PubMed:8541540}.	23
P01178	NEU1_HUMAN	OXT OT	125	SUBCELLULAR LOCATION: Secreted.	19
P01210	PENK_HUMAN	PENK	267	SUBCELLULAR LOCATION: Cytoplasmic vesicle, secretory vesicle, chromaffin granule lumen {ECO:0000250|UniProtKB:P01211}. Secreted {ECO:0000250|UniProtKB:P01211}.	24
P01597	KV139_HUMAN	IGKV1-39	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
P01602	KV105_HUMAN	IGKV1-5	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
P01619	KV320_HUMAN	IGKV3-20	116	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
P01721	LV657_HUMAN	IGLV6-57	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
P01768	HV330_HUMAN	IGHV3-30	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
P02647	APOA1_HUMAN	APOA1	267	SUBCELLULAR LOCATION: Secreted.	18
P02649	APOE_HUMAN	APOE	317	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:2498325, ECO:0000269|PubMed:30333625}. Secreted, extracellular space {ECO:0000269|PubMed:8340399}. Secreted, extracellular space, extracellular matrix {ECO:0000269|PubMed:9488694}. Extracellular vesicle {ECO:0000269|PubMed:26387950}. Endosome, multivesicular body {ECO:0000269|PubMed:26387950}. Note=In the plasma, APOE is associated with chylomicrons, chylomicrons remnants, VLDL, LDL and HDL lipoproteins (PubMed:1911868, PubMed:8340399). Lipid poor oligomeric APOE is associated with the extracellular matrix in a calcium- and heparan-sulfate proteoglycans-dependent manner (PubMed:9488694). Lipidation induces the release from the extracellular matrix (PubMed:9488694). Colocalizes with CD63 and PMEL at exosomes and in intraluminal vesicles within multivesicular endosomes. {ECO:0000269|PubMed:1911868, ECO:0000269|PubMed:26387950, ECO:0000269|PubMed:8340399, ECO:0000269|PubMed:9488694}.	18
P02654	APOC1_HUMAN	APOC1	83	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:2835369}.	26
P02655	APOC2_HUMAN	APOC2 APC2	101	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:3525527}.	22
P02656	APOC3_HUMAN	APOC3	99	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:18201179, ECO:0000303|PubMed:22510806}.	20
P02671	FIBA_HUMAN	FGA	866	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:19296670, ECO:0000269|PubMed:9628725}.	19
P02745	C1QA_HUMAN	C1QA	245	SUBCELLULAR LOCATION: Secreted.	22
P02749	APOH_HUMAN	APOH B2G1	345	SUBCELLULAR LOCATION: Secreted.	19
P02765	FETUA_HUMAN	AHSG FETUA PRO2743	367	SUBCELLULAR LOCATION: Secreted.	18
P02766	TTHY_HUMAN	TTR PALB	147	SUBCELLULAR LOCATION: Secreted. Cytoplasm.	20
P02790	HEMO_HUMAN	HPX	462	SUBCELLULAR LOCATION: Secreted.	23
P03952	KLKB1_HUMAN	KLKB1 KLK3	638	SUBCELLULAR LOCATION: Secreted.	19
P04196	HRG_HUMAN	HRG	525	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:21215706}.	18
P04216	THY1_HUMAN	THY1	161	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Lipid-anchor, GPI-anchor {ECO:0000250}.	19

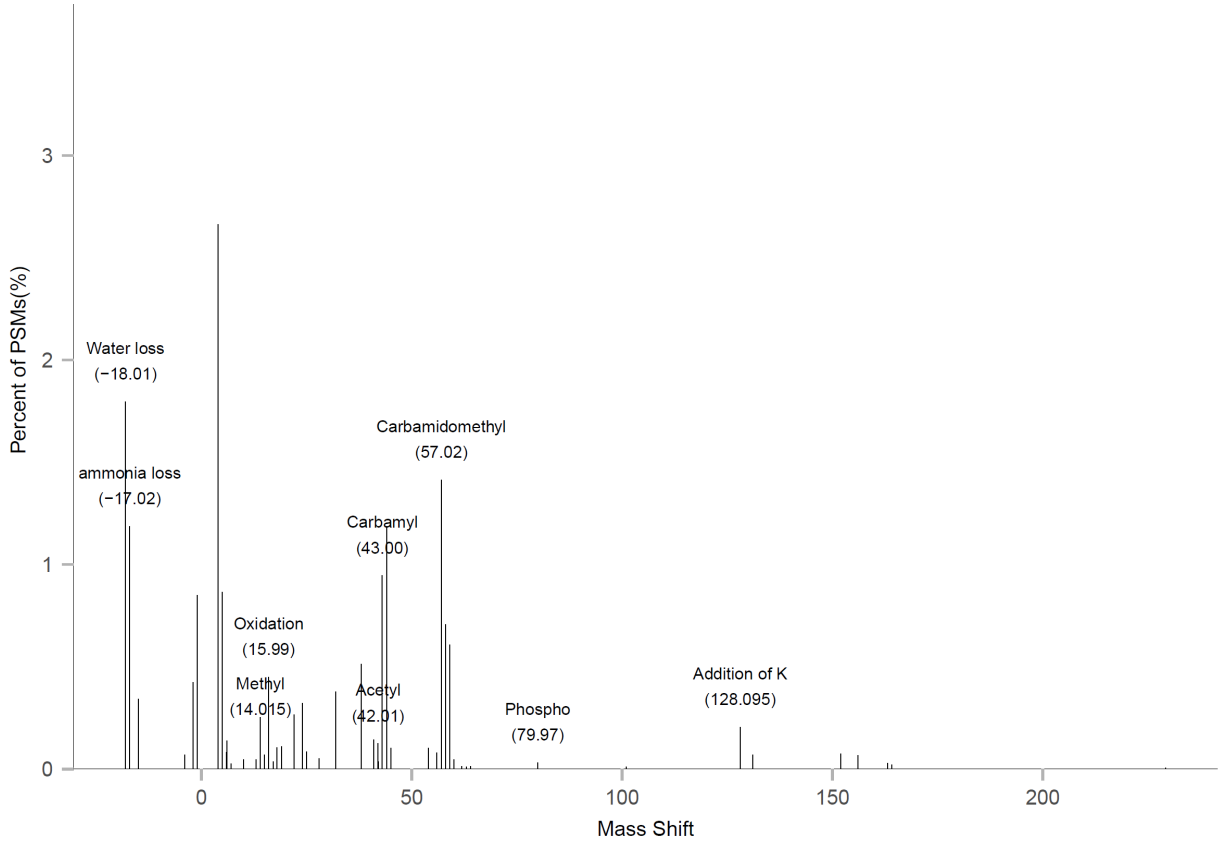
P05067	A4_HUMAN	APP A4 AD1	770	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:10383380, ECO:0000269|PubMed:20580937, ECO:0000269|PubMed:2649245, ECO:0000305|PubMed:25122912}; Single-pass type I membrane protein {ECO:0000269|PubMed:30630874, ECO:0000305|PubMed:10383380, ECO:0000305|PubMed:25122912}. Membrane {ECO:0000269|PubMed:2900137, ECO:0000305|PubMed:22584060}; Single-pass type I membrane protein {ECO:0000269|PubMed:2900137, ECO:0000269|PubMed:30630874, ECO:0000305|PubMed:22584060}. Perikaryon {ECO:0000269|PubMed:10341243}. Cell projection, growth cone {ECO:0000269|PubMed:10341243}. Membrane, clathrin-coated pit {ECO:0000269|PubMed:20580937}. Early endosome {ECO:0000269|PubMed:20580937}. Cytoplasmic vesicle {ECO:0000269|PubMed:20580937, ECO:0000269|PubMed:25122912}. Note=Cell surface protein that rapidly becomes internalized via clathrin-coated pits. Only a minor proportion is present at the cell membrane; most of the protein is present in intracellular vesicles (PubMed:20580937). During maturation, the immature APP (N-glycosylated in the endoplasmic reticulum) moves to the Golgi complex where complete maturation occurs (O-glycosylated and sulfated). After alpha-secretase cleavage, soluble APP is released into the extracellular space and the C-terminal is internalized to endosomes and lysosomes. Some APP accumulates in secretory transport vesicles leaving the late Golgi compartment and returns to the cell surface. APP sorts to the basolateral surface in epithelial cells. During neuronal differentiation, the Thr-743 phosphorylated form is located mainly in growth cones, moderately in neurites and sparingly in the cell body (PubMed:10341243). Casein kinase phosphorylation can occur either at the cell surface or within a post-Golgi compartment. Associates with GPC1 in perinuclear compartments. Colocalizes with SORL1 in a vesicular pattern in cytoplasm and perinuclear regions. {ECO:0000269|PubMed:10341243, ECO:0000269|PubMed:20580937}.; SUBCELLULAR LOCATION: [C83]: Endoplasmic reticulum {ECO:0000269|PubMed:14527950}. Golgi apparatus {ECO:0000269|PubMed:14527950}. Early endosome {ECO:0000269|PubMed:14527950}.; SUBCELLULAR LOCATION: [C99]: Early endosome {ECO:0000269|PubMed:14527950}.; SUBCELLULAR LOCATION: [Soluble APP-beta]: Secreted {ECO:0000269|PubMed:10656250, ECO:0000269|PubMed:2649245}.; SUBCELLULAR LOCATION: [Amyloid-beta protein 40]: Cell surface {ECO:0000269|PubMed:16154999}.; SUBCELLULAR LOCATION: [Amyloid-beta protein 42]: Cell surface {ECO:0000269|PubMed:11689470, ECO:0000269|PubMed:16154999}. Note=Associates with FPR2 at the cell surface and the complex is then rapidly internalized. {ECO:0000269|PubMed:11689470}.; SUBCELLULAR LOCATION: [Gamma-secretase C-terminal fragment 59]: Nucleus {ECO:0000269|PubMed:11544248}. Cytoplasm {ECO:0000269|PubMed:11544248}. Note=Located to both the cytoplasm and nuclei of neurons. It can be translocated to the nucleus through association with APBB1 (Fe65) (PubMed:11544248). In dopaminergic neurons, the phosphorylated Thr-743 form is localized to the nucleus (By similarity). {ECO:0000250|UniProtKB:P12023, ECO:0000269|PubMed:11544248}.	17
P05156	CFAI_HUMAN	CFI IF	583	SUBCELLULAR LOCATION: Secreted, extracellular space. Secreted {ECO:0000269|PubMed:6327681}.	18
P05546	HEP2_HUMAN	SERPIND1 HCF2	499		19
P06312	KV401_HUMAN	IGKV4-1	121	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
P06865	HEXA_HUMAN	HEXA	529	SUBCELLULAR LOCATION: Lysosome.	22
P07333	CSF1R_HUMAN	CSF1R FMS	972	SUBCELLULAR LOCATION: Cell membrane; Single-pass type I membrane protein.	19
P07711	CATL1_HUMAN	CTSL CTSL1	333	SUBCELLULAR LOCATION: Lysosome {ECO:0000250|UniProtKB:P06797}. Apical cell membrane {ECO:0000250|UniProtKB:P06797}; Peripheral membrane protein {ECO:0000250|UniProtKB:P06797}; Extracellular side {ECO:0000250|UniProtKB:P06797}. Cytoplasmic vesicle, secretory vesicle, chromaffin granule {ECO:0000250|UniProtKB:P25975}. Secreted, extracellular space {ECO:0000250|UniProtKB:P06797}. Secreted {ECO:0000250|UniProtKB:P06797}. Note=Localizes to the apical membrane of thyroid epithelial cells. Released at extracellular space by activated dendritic cells and macrophages. {ECO:0000250|UniProtKB:P06797}.; SUBCELLULAR LOCATION: [Isoform 2]: Nucleus {ECO:0000269|PubMed:15099520}. Note=Translation initiation at downstream start sites allows the synthesis of isoforms that are devoid of a signal peptide and do not transit through the endoplasmic reticulum to localize to the nucleus (PubMed:15099520). Nuclear location varies during the cell cycle, with higher levels during S phase (PubMed:15099520). {ECO:0000269|PubMed:15099520}.	17
P07996	TSP1_HUMAN	THBS1 TSP TSP1	1170	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:101549, ECO:0000269|PubMed:14568985, ECO:0000269|PubMed:6777381}. Cell surface {ECO:0000269|PubMed:6777381}. Secreted, extracellular space, extracellular matrix {ECO:0000269|PubMed:18285447, ECO:0000269|PubMed:6341993}. Endoplasmic reticulum {ECO:0000250|UniProtKB:P35441}. Sarcoplasmic reticulum {ECO:0000250|UniProtKB:P35441}. Note=Secreted by thrombin-activated platelets and binds to the cell surface in the presence of extracellular Ca(2+) (PubMed:6777381, PubMed:101549). Incorporated into the extracellular matrix (ECM) of fibroblasts (PubMed:6341993). The C-terminal region in trimeric form is required for retention in the ECM (PubMed:18285447). Also detected in the endoplasmic reticulum and sarcoplasmic reticulum where it plays a role in the ER stress response (By similarity). {ECO:0000250|UniProtKB:P35441, ECO:0000269|PubMed:6341993, ECO:0000269|PubMed:6777381}.	18
P08138	TNR16_HUMAN	NGFR TNFRSF16	427	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:3022937}; Single-pass type I membrane protein {ECO:0000305}. Perikaryon {ECO:0000250|UniProtKB:Q9Z0W1}. Cell projection, growth cone {ECO:0000250|UniProtKB:Q9Z0W1}. Cell projection, dendritic spine {ECO:0000250|UniProtKB:Q9Z0W1}.	28
P08174	DAF_HUMAN	CD55 CR DAF	381	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 2]: Cell membrane; Lipid-anchor, GPI-anchor.; SUBCELLULAR LOCATION: [Isoform 3]: Secreted {ECO:0000269|PubMed:16503113}.; SUBCELLULAR LOCATION: [Isoform 4]: Secreted {ECO:0000269|PubMed:16503113}.; SUBCELLULAR LOCATION: [Isoform 5]: Secreted {ECO:0000269|PubMed:16503113}.; SUBCELLULAR LOCATION: [Isoform 6]: Cell membrane {ECO:0000305|PubMed:16503113}; Lipid-anchor, GPI-anchor {ECO:0000305|PubMed:16503113}.; SUBCELLULAR LOCATION: [Isoform 7]: Cell membrane {ECO:0000305|PubMed:16503113}; Lipid-anchor, GPI-anchor {ECO:0000305|PubMed:16503113}.	34
P08253	MMP2_HUMAN	MMP2 CLG4A	660	SUBCELLULAR LOCATION: [Isoform 1]: Secreted, extracellular space, extracellular matrix {ECO:0000305|PubMed:2834383}. Membrane. Nucleus. Note=Colocalizes with integrin alphaV/beta3 at the membrane surface in angiogenic blood vessels and melanomas. Found in mitochondria, along microfibrils, and in nuclei of cardiomyocytes.; SUBCELLULAR LOCATION: [Isoform 2]: Cytoplasm. Mitochondrion.	29
P08493	MGP_HUMAN	MGP MGLAP GIG36	103	SUBCELLULAR LOCATION: Secreted.	19
P08571	CD14_HUMAN	CD14	375	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:1698311, ECO:0000269|PubMed:2462937, ECO:0000269|PubMed:3385210}; Lipid-anchor, GPI-anchor {ECO:0000269|PubMed:1698311, ECO:0000269|PubMed:2462937, ECO:0000269|PubMed:3385210}. Secreted {ECO:0000269|PubMed:25497142, ECO:0000269|PubMed:2779588}. Membrane raft {ECO:0000269|PubMed:16880211}. Golgi apparatus {ECO:0000269|PubMed:16880211}. Note=Secreted forms may arise by cleavage of the GPI anchor. {ECO:0000269|PubMed:2462937, ECO:0000269|PubMed:2779588, ECO:0000269|PubMed:3385210}.	19
P08603	CFAH_HUMAN	CFH HF HF1 HF2	1231	SUBCELLULAR LOCATION: Secreted.; SUBCELLULAR LOCATION: Note=(Microbial infection) In the mosquito midgut, localizes to P.falciparum (NF54 strain) macrogamete and young zygote cell membranes. {ECO:0000269|PubMed:23332154}.	18

P09603	CSF1_HUMAN	CSF1	554	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:1531650, ECO:0000269|PubMed:3264877}; Single-pass type I membrane protein {ECO:0000269|PubMed:1531650, ECO:0000269|PubMed:3264877}.; SUBCELLULAR LOCATION: [Processed macrophage colony-stimulating factor 1]: Secreted, extracellular space.	32
P09871	C1S_HUMAN	C1S	688		15
P0DP58	LYNX1_HUMAN	LYNX1	116	SUBCELLULAR LOCATION: Cell membrane {ECO:0000255}; Lipid-anchor, GPI-anchor {ECO:0000255}. Cell projection, dendrite {ECO:0000250|UniProtKB:P0DP60}. Endoplasmic reticulum {ECO:0000250|UniProtKB:P0DP60}. Note=Detected in Purkinje cells soma and proximal dendrites. {ECO:0000250|UniProtKB:P0DP60}.	20
P10451	OSTP_HUMAN	SPP1 BNSP OPN PSEC0156	314	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:25326458, ECO:0000269|PubMed:36213313, ECO:0000269|PubMed:37453717}.	16
P10645	CMGA_HUMAN	CHGA	457	SUBCELLULAR LOCATION: [Serpinin]: Secreted {ECO:0000250|UniProtKB:P26339}. Cytoplasmic vesicle, secretory vesicle {ECO:0000250|UniProtKB:P26339}. Note=Pyroglutaminated serpinin localizes to secretory vesicle. {ECO:0000250|UniProtKB:P26339}.; SUBCELLULAR LOCATION: Cytoplasmic vesicle, secretory vesicle {ECO:0000250|UniProtKB:P10354}. Cytoplasmic vesicle, secretory vesicle, neuronal dense core vesicle {ECO:0000250|UniProtKB:P10354}. Secreted {ECO:0000269|PubMed:25326458, ECO:0000269|PubMed:37453717}. Note=Associated with the secretory granule membrane through direct interaction to SCG3 that in turn binds to cholesterol-enriched lipid rafts in intragranular conditions. In pituitary gonadotropes, located in large secretory granules. {ECO:0000250|UniProtKB:P10354}.	18
P10909	CLUS_HUMAN	CLU APOJ CLI KUB1 AAG4	449	SUBCELLULAR LOCATION: [Isoform 1]: Secreted {ECO:0000269|PubMed:11123922, ECO:0000269|PubMed:17260971, ECO:0000269|PubMed:17412999, ECO:0000269|PubMed:17451556, ECO:0000269|PubMed:2387851, ECO:0000269|PubMed:24073260, ECO:0000269|PubMed:2780565, ECO:0000269|PubMed:3154963, ECO:0000269|PubMed:8292612, ECO:0000269|PubMed:8328966}. Note=Can retrotranslocate from the secretory compartments to the cytosol upon cellular stress. {ECO:0000269|PubMed:17451556}.; SUBCELLULAR LOCATION: [Isoform 4]: Cytoplasm {ECO:0000269|PubMed:24073260}. Note=Keeps cytoplasmic localization in stressed and unstressed cell. {ECO:0000269|PubMed:24073260}.; SUBCELLULAR LOCATION: [Isoform 6]: Cytoplasm {ECO:0000269|PubMed:24073260}. Note=Keeps cytoplasmic localization in stressed and unstressed cell. {ECO:0000269|PubMed:24073260}.; SUBCELLULAR LOCATION: Nucleus {ECO:0000269|PubMed:12551933, ECO:0000269|PubMed:19137541}. Cytoplasm {ECO:0000269|PubMed:12551933, ECO:0000269|PubMed:17689225, ECO:0000269|PubMed:19137541, ECO:0000269|PubMed:20068069, ECO:0000269|PubMed:22689054, ECO:0000269|PubMed:24073260}. Mitochondrion membrane; Peripheral membrane protein; Cytoplasmic side {ECO:0000269|PubMed:17689225}. Cytoplasm, cytosol {ECO:0000269|PubMed:17451556, ECO:0000269|PubMed:22689054, ECO:0000269|PubMed:24073260}. Microsome {ECO:0000269|PubMed:22689054}. Endoplasmic reticulum {ECO:0000269|PubMed:16113678, ECO:0000269|PubMed:22689054}. Mitochondrion {ECO:0000269|PubMed:16113678, ECO:0000269|PubMed:22689054}. Mitochondrion membrane {ECO:0000269|PubMed:16113678, ECO:0000269|PubMed:17689225}. Cytoplasm, perinuclear region {ECO:0000250|UniProtKB:P05371}. Cytoplasmic vesicle, secretory vesicle, chromaffin granule {ECO:0000250}. Note=Secreted isoforms can retrotranslocate from the secretory compartments to the cytosol upon cellular stress (PubMed:17451556). Detected in perinuclear foci that may be aggresomes containing misfolded, ubiquitinated proteins (PubMed:20068069). Detected at the mitochondrion membrane upon induction of apoptosis (PubMed:17689225). Under ER stress, a immaturely glycosylated pre-secreted form retrotranslocates from the endoplasmic reticulum (ER)-Golgi network to the cytoplasm to localize in the mitochondria through HSPA5 interaction (PubMed:22689054). ER stress reduces secretion (PubMed:22689054). Under the stress, minor amounts of non-secreted forms accumulate in cytoplasm (PubMed:24073260, PubMed:22689054, PubMed:17451556). Non-secreted forms emerge mainly from failed translocation, alternative splicing or non-canonical initiation start codon (PubMed:24073260, PubMed:12551933). {ECO:0000269|PubMed:12551933, ECO:0000269|PubMed:17451556, ECO:0000269|PubMed:17689225, ECO:0000269|PubMed:20068069, ECO:0000269|PubMed:22689054, ECO:0000269|PubMed:24073260}.	22
P13591	NCAM1_HUMAN	NCAM1 NCAM	858	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 2]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 3]: Cell membrane; Lipid-anchor, GPI-anchor.; SUBCELLULAR LOCATION: [Isoform 4]: Cell membrane {ECO:0000305}; Lipid-anchor, GPI-anchor {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 5]: Secreted.; SUBCELLULAR LOCATION: [Isoform 6]: Secreted {ECO:0000305}.	19
P13987	CD59_HUMAN	CD59 MIC11 MIN1 MIN2 MIN3 MSK21	128	SUBCELLULAR LOCATION: Cell membrane; Lipid-anchor, GPI-anchor. Secreted. Note=Soluble form found in a number of tissues.	25
P16035	TIMP2_HUMAN	TIMP2	220	SUBCELLULAR LOCATION: Secreted.	26
P16519	NEC2_HUMAN	PCSK2 NEC2	638	SUBCELLULAR LOCATION: Cytoplasmic vesicle, secretory vesicle. Secreted {ECO:0000269|PubMed:28719828}. Note=Localized in the secretion granules.	25
P19320	VCAM1_HUMAN	VCAM1	739	SUBCELLULAR LOCATION: [Vascular cell adhesion protein 1]: Cell membrane {ECO:0000305|PubMed:12878595}; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Soluble Vascular Cell Adhesion Molecule-1]: Secreted {ECO:0000269|PubMed:12878595, ECO:0000269|PubMed:36127634}.	24
P20062	TCO2_HUMAN	TCN2 TC2	427	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:3782074, ECO:0000269|PubMed:8443384}.	18
P20333	TNR1B_HUMAN	TNFRSF1B TNFBR TNFR2	461	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 2]: Secreted.; SUBCELLULAR LOCATION: [Tumor necrosis factor-binding protein 2]: Secreted.	22
P22692	IBP4_HUMAN	IGFBP4 IBP4	258	SUBCELLULAR LOCATION: Secreted.	21
P23142	FBLN1_HUMAN	FBLN1 PP213	703	SUBCELLULAR LOCATION: Secreted, extracellular space, extracellular matrix.	29
P23515	OMGP_HUMAN	OMG OMGP	440	SUBCELLULAR LOCATION: Cell membrane; Lipid-anchor, GPI-anchor.	24
P24593	IBP5_HUMAN	IGFBP5 IBP5	272	SUBCELLULAR LOCATION: Secreted.	20
P27797	CALR_HUMAN	CALR CRTC	417	SUBCELLULAR LOCATION: Endoplasmic reticulum lumen {ECO:0000269|PubMed:10358038, ECO:0000269|PubMed:11149926}. Cytoplasm, cytosol {ECO:0000269|PubMed:11149926}. Secreted, extracellular space, extracellular matrix {ECO:0000305}. Cell surface {ECO:0000269|PubMed:10358038}. Sarcoplasmic reticulum lumen {ECO:0000250|UniProtKB:P28491}. Cytoplasmic vesicle, secretory vesicle, Cortical granule {ECO:0000250|UniProtKB:Q8K3H7}. Cytolytic granule {ECO:0000269|PubMed:8418194}. Note=Also found in cell surface (T cells), cytosol and extracellular matrix (PubMed:10358038). During oocyte maturation and after parthenogenetic activation accumulates in cortical granules. In pronuclear and early cleaved embryos localizes weakly to cytoplasm around nucleus and more strongly in the region near the cortex (By similarity). In cortical granules of non-activated oocytes, is exocytosed during the cortical reaction in response to oocyte activation (By similarity). {ECO:0000250|UniProtKB:P28491, ECO:0000250|UniProtKB:Q8K3H7, ECO:0000269|PubMed:8418194}.	17
P30101	PDIA3_HUMAN	PDIA3 ERP57 ERP60 GRP58	505	SUBCELLULAR LOCATION: Endoplasmic reticulum {ECO:0000269|PubMed:23826168}. Endoplasmic reticulum lumen {ECO:0000250|UniProtKB:P11598}. Melanosome {ECO:0000269|PubMed:12643545, ECO:0000269|PubMed:17081065}. Note=Identified by mass spectrometry in melanosome fractions from stage I to stage IV (PubMed:12643545). {ECO:0000269|PubMed:12643545}.	24
P36222	CH3L1_HUMAN	CHI3L1	383	SUBCELLULAR LOCATION: Secreted, extracellular space {ECO:0000269|PubMed:9492324}. Cytoplasm {ECO:0000250}. Cytoplasm, perinuclear region {ECO:0000250}. Endoplasmic reticulum {ECO:0000250}.	21
P36980	FHR2_HUMAN	CFHR2 CFHL2 FHR2 HFL3	270	SUBCELLULAR LOCATION: Secreted.	18

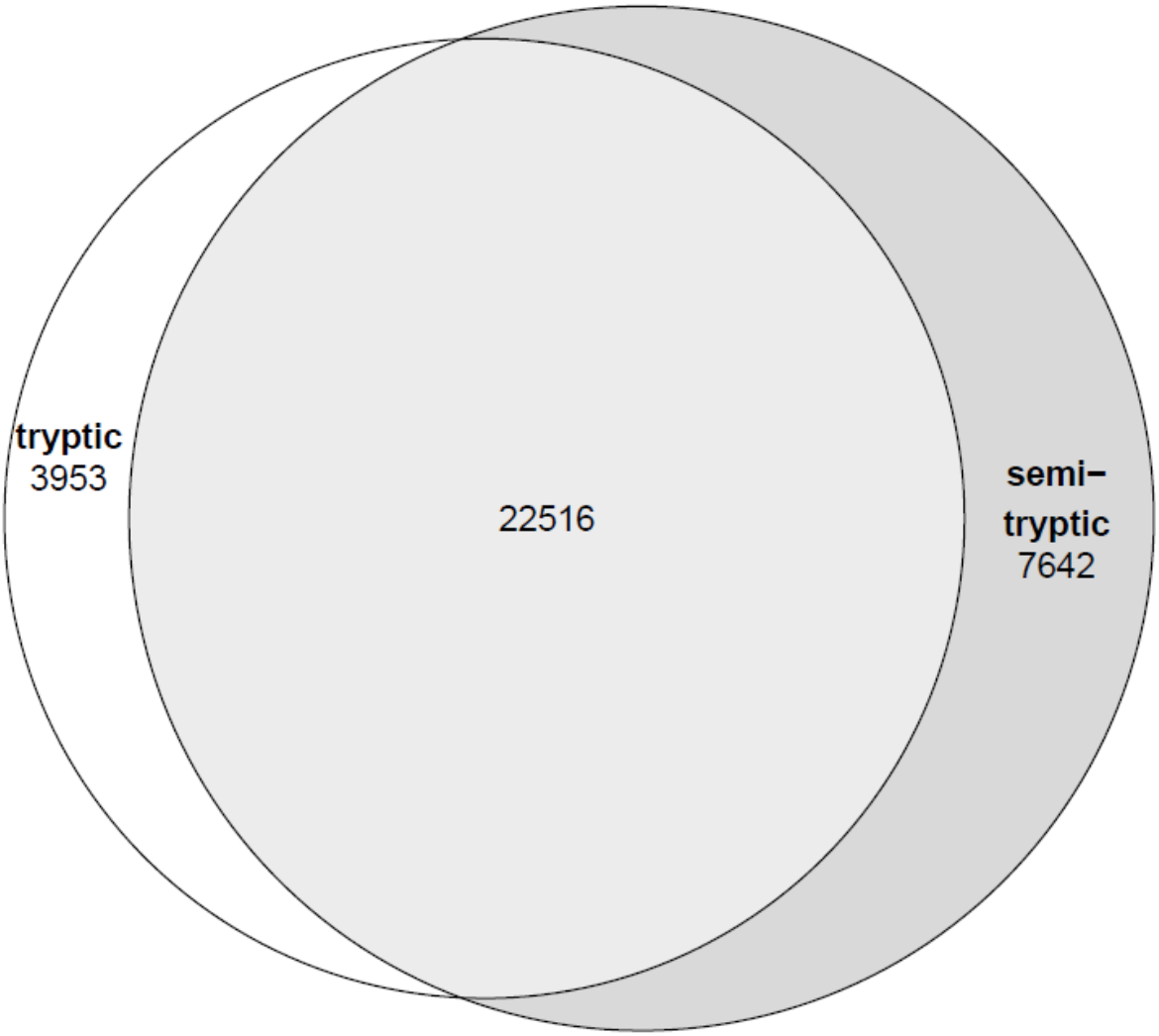
P41222	PTGDS_HUMAN	PTGDS PDS	190	SUBCELLULAR LOCATION: Rough endoplasmic reticulum {ECO:0000269|PubMed:9065498}. Nucleus membrane {ECO:0000269|PubMed:9065498}. Golgi apparatus {ECO:0000269|PubMed:9065498}. Cytoplasm, perinuclear region {ECO:0000269|PubMed:9065498}. Secreted {ECO:0000269|PubMed:9065498}. Note=Detected on rough endoplasmic reticulum of arachnoid and menigioma cells. Localized to the nuclear envelope, Golgi apparatus, secretory vesicles and spherical cytoplasmic structures in arachnoid trabecular cells, and to circular cytoplasmic structures in meningeal macrophages and perivascular microglial cells. In oligodendrocytes, localized to the rough endoplasmic reticulum and nuclear envelope. In retinal pigment epithelial cells, localized to distinct cytoplasmic domains including the perinuclear region. Also secreted.	22
P49908	SEPP1_HUMAN	SELENOP SELP SEPP1	381	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:8142465}. Note=Passes from plasma into the glomerular filtrate where it is removed by endocytosis mediated by LRP2 in the proximal tubule epithelium. {ECO:0000250|UniProtKB:P70274}.	19
P51693	APLP1_HUMAN	APLP1	650	SUBCELLULAR LOCATION: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [C30]: Cytoplasm. Note=C-terminally processed in the Golgi complex.	38
P54108	CRIS3_HUMAN	CRISP3	245	SUBCELLULAR LOCATION: Secreted. Note=In neutrophils, localized in specific granules.	20
P55290	CAD13_HUMAN	CDH13 CDHH	713	SUBCELLULAR LOCATION: Cell membrane; Lipid-anchor, GPI-anchor.	22
P61916	NPC2_HUMAN	NPC2 HE1	151	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:11125141, ECO:0000269|PubMed:15937921, ECO:0000269|PubMed:19723497, ECO:0000269|PubMed:21315718}. Endoplasmic reticulum {ECO:0000269|PubMed:19723497}. Lysosome {ECO:0000269|PubMed:15937921, ECO:0000269|PubMed:19723497, ECO:0000305|PubMed:11125141}. Note=Interaction with cell-surface M6PR mediates endocytosis and targeting to lysosomes. {ECO:0000305|PubMed:11125141}.	19
P78324	SHPS1_HUMAN	SIRPA BIT MFR MYD1 PTPNS1 SHPS1 SIRP	504	SUBCELLULAR LOCATION: Membrane; Single-pass type I membrane protein.	30
P80108	PHLD_HUMAN	GPLD1 PIGPLD1	840	SUBCELLULAR LOCATION: Secreted.	23
P81172	HEPC_HUMAN	HAMP HEPC LEAP1 UNQ487/PRO1003	84	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:11034317}.	24
Q12907	LMAN2_HUMAN	LMAN2 C5orf8	356	SUBCELLULAR LOCATION: Endoplasmic reticulum-Golgi intermediate compartment membrane {ECO:0000269|PubMed:10444376}; Single-pass type I membrane protein {ECO:0000269|PubMed:10444376}. Golgi apparatus membrane {ECO:0000269|PubMed:10444376}; Single-pass membrane protein {ECO:0000269|PubMed:10444376}. Endoplasmic reticulum membrane {ECO:0000269|PubMed:10444376}; Single-pass type I membrane protein {ECO:0000269|PubMed:10444376}.	44
Q13508	NAR3_HUMAN	ART3 TMART	389	SUBCELLULAR LOCATION: Cell membrane; Lipid-anchor, GPI-anchor.	26
Q13740	CD166_HUMAN	ALCAM MEMD	583	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:15048703, ECO:0000269|PubMed:15294938, ECO:0000269|PubMed:16352806, ECO:0000269|PubMed:23169771, ECO:0000269|PubMed:24740813, ECO:0000269|PubMed:24945728, ECO:0000269|PubMed:7760007}; Single-pass type I membrane protein {ECO:0000305}. Cell projection, axon {ECO:0000250|UniProtKB:Q61490}. Cell projection, dendrite {ECO:0000250|UniProtKB:Q61490}. Note=Detected at the immunological synapse, i.e, at the contact zone between antigen-presenting dendritic cells and T-cells (PubMed:15294938, PubMed:16352806). Colocalizes with CD6 and the TCR/CD3 complex at the immunological synapse (PubMed:15294938). {ECO:0000269|PubMed:15294938, ECO:0000269|PubMed:16352806}.; SUBCELLULAR LOCATION: [Isoform 3]: Secreted {ECO:0000269|PubMed:15496415}.	27
Q14118	DAG1_HUMAN	DAG1	895	SUBCELLULAR LOCATION: [Alpha-dystroglycan]: Secreted, extracellular space.; SUBCELLULAR LOCATION: [Beta-dystroglycan]: Cell membrane {ECO:0000269|PubMed:18764929}; Single-pass type I membrane protein. Cytoplasm, cytoskeleton. Nucleus, nucleoplasm {ECO:0000269|PubMed:18764929}. Cell membrane, sarcolemma {ECO:0000250}. Postsynaptic cell membrane {ECO:0000250}. Note=The monomeric form translocates to the nucleus via the action of importins and depends on RAN. Nuclear transport is inhibited by Tyr-892 phosphorylation. In skeletal muscle, this phosphorylated form locates to a vesicular internal membrane compartment. In muscle cells, sarcolemma localization requires the presence of ANK2, while localization to costameres requires the presence of ANK3. Localizes to neuromuscular junctions (NMJs) in the presence of ANK2 (By similarity). In peripheral nerves, localizes to the Schwann cell membrane. Colocalizes with ERM proteins in Schwann-cell microvilli. {ECO:0000250}.	29
Q14624	ITIH4_HUMAN	ITIH4 IHRP ITIHL1 PK120 PRO1851	930	SUBCELLULAR LOCATION: Secreted.	28
Q15084	PDIA6_HUMAN	PDIA6 ERP5 P5 TXNDC7	440	SUBCELLULAR LOCATION: Endoplasmic reticulum lumen {ECO:0000269|PubMed:15466936}. Cell membrane {ECO:0000269|PubMed:15466936}. Melanosome {ECO:0000269|PubMed:12643545, ECO:0000269|PubMed:17081065}. Note=Identified by mass spectrometry in melanosome fractions from stage I to stage IV (PubMed:12643545). {ECO:0000269|PubMed:12643545}.	19
Q15904	VAS1_HUMAN	ATP6AP1 ATP6IP1 ATP6S1 VATPS1 XAP3	470	SUBCELLULAR LOCATION: Endoplasmic reticulum membrane {ECO:0000269|PubMed:27231034}; Single-pass type I membrane protein {ECO:0000305}. Endoplasmic reticulum-Golgi intermediate compartment membrane {ECO:0000269|PubMed:27231034}. Cytoplasmic vesicle, secretory vesicle, synaptic vesicle membrane {ECO:0000250|UniProtKB:O54715}; Single-pass type I membrane protein {ECO:0000305}. Cytoplasmic vesicle, clathrin-coated vesicle membrane {ECO:0000250|UniProtKB:O54715}; Single-pass type I membrane protein {ECO:0000305}. Note=Not detected in trans-Golgi network. {ECO:0000269|PubMed:27231034}.	41
Q16288	NTRK3_HUMAN	NTRK3 TRKC	839	SUBCELLULAR LOCATION: Membrane; Single-pass type I membrane protein.	31
Q16553	LY6E_HUMAN	LY6E 9804 RIGE SCA2 TSA1	131	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250|UniProtKB:Q64253}; Lipid-anchor, GPI-anchor {ECO:0000250|UniProtKB:Q64253}.	20
Q16610	ECM1_HUMAN	ECM1	540	SUBCELLULAR LOCATION: Secreted, extracellular space, extracellular matrix.	19
Q16769	QPCT_HUMAN	QPCT	361	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:18486145}.	28
Q5JZY3	EPHAA_HUMAN	EPHA10	1008	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 3]: Cell membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 2]: Secreted {ECO:0000305}.	33
Q6EMK4	VASN_HUMAN	VASN SLITL2 UNQ314/PRO357/PRO1282	673	SUBCELLULAR LOCATION: Membrane {ECO:0000305|PubMed:15247411}; Single-pass type I membrane protein {ECO:0000305|PubMed:15247411}. Secreted {ECO:0000269|PubMed:15247411}.	23
Q7Z5A7	TAFA5_HUMAN	TAFA5 FAM19A5 UNQ5208/PRO34524	132	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:15028294, ECO:0000269|PubMed:29453251}.	43
Q86YD3	TMM25_HUMAN	TMEM25 UNQ2531/PRO6030	366	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 4]: Cell membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 2]: Secreted {ECO:0000305}.; SUBCELLULAR LOCATION: [Isoform 3]: Secreted {ECO:0000305}.; SUBCELLULAR LOCATION: Late endosome {ECO:0000250|UniProtKB:Q9DCF1}. Lysosome {ECO:0000250|UniProtKB:Q9DCF1}.	26
Q8NFT8	DNER_HUMAN	DNER BET UNQ262/PRO299	737	SUBCELLULAR LOCATION: Cell membrane; Single-pass type I membrane protein. Note=Present on the membrane of dendrites and cell bodies but excluded from axonal membrane. Also found in early endosomes in the somatodendritic region (By similarity). {ECO:0000250}.	34
Q8NFY4	SEM6D_HUMAN	SEMA6D KIAA1479	1073	SUBCELLULAR LOCATION: [Isoform 1]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 2]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 3]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 4]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 5]: Cell membrane; Single-pass type I membrane protein.; SUBCELLULAR LOCATION: [Isoform 7]: Cytoplasm.	20

Q8NI22	MCFD2_HUMAN	MCFD2 SDNSF	146	SUBCELLULAR LOCATION: Endoplasmic reticulum-Golgi intermediate compartment {ECO:0000269|PubMed:12717434}. Endoplasmic reticulum {ECO:0000269|PubMed:12717434}. Golgi apparatus {ECO:0000269|PubMed:12717434}.	26
Q8TCZ2	C99L2_HUMAN	CD99L2 MIC2L1 UNQ1964/PRO4486	262	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Single-pass type I membrane protein {ECO:0000250}; Extracellular side {ECO:0000250}. Cell junction {ECO:0000250}. Secreted {ECO:0000269|PubMed:25326458}.	25
Q8TER0	SNED1_HUMAN	SNED1	1413	SUBCELLULAR LOCATION: Secreted, extracellular space, extracellular matrix {ECO:0000269|PubMed:33724335}. Note=Forms microfibrils within the extracellular matrix and colocalizes with fibronectin (FN1). {ECO:0000250|UniProtKB:Q70E20}.	24
Q8WXD2	SCG3_HUMAN	SCG3 UNQ2502/PRO5990	468	SUBCELLULAR LOCATION: Cytoplasmic vesicle, secretory vesicle {ECO:0000250|UniProtKB:P47868}. Cytoplasmic vesicle, secretory vesicle membrane {ECO:0000250|UniProtKB:P47868}; Peripheral membrane protein {ECO:0000250}. Secreted {ECO:0000269|PubMed:12098761, ECO:0000269|PubMed:25326458}. Note=Associated with the secretory granule membrane through direct binding to cholesterol-enriched lipid rafts. {ECO:0000250|UniProtKB:P47868}.	19
Q92765	SFRP3_HUMAN	FRZB FIZ FRE FRP FRZB1 SFRP3	325	SUBCELLULAR LOCATION: Secreted {ECO:0000305}.	32
Q92876	KLK6_HUMAN	KLK6 PRSS18 PRSS9	244	SUBCELLULAR LOCATION: Secreted. Nucleus, nucleolus. Cytoplasm. Mitochondrion. Microsome. Note=In brain, detected in the nucleus of glial cells and in the nucleus and cytoplasm of neurons. Detected in the mitochondrial and microsomal fractions of HEK-293 cells and released into the cytoplasm following cell stress.	16
Q969P0	IGSF8_HUMAN	IGSF8 CD81P3 EWI2 KCT4	613	SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Single-pass membrane protein.	27
Q96FE7	PI3IP1_HUMAN	PIK3IP1 HGFL	263	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:19088825}; Single-pass type I membrane protein {ECO:0000255}.	21
Q96GW7	PGCB_HUMAN	BCAN BEHAB CSPG7 UNQ2525/PRO6018	911	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:25326458, ECO:0000269|PubMed:37453717}.; SUBCELLULAR LOCATION: [Isoform 1]: Secreted, extracellular space, extracellular matrix.; SUBCELLULAR LOCATION: [Isoform 2]: Membrane; Lipid-anchor, GPI-anchor.	22
Q96IY4	CBPB2_HUMAN	CPB2	423	SUBCELLULAR LOCATION: Secreted.	22
Q96KN2	CNDP1_HUMAN	CNDP1 CN1 CPGL2 UNQ1915/PRO4380	507	SUBCELLULAR LOCATION: Secreted {ECO:0000269|PubMed:12473676}.	26
Q96PX8	SLIK1_HUMAN	SLITRK1 KIAA1910 LRRC12 UNQ233/PRO266	696	SUBCELLULAR LOCATION: Membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}. Secreted {ECO:0000269|PubMed:19640509}. Synapse {ECO:0000250|UniProtKB:Q810C1}.	17
Q9BZR6	RTN4R_HUMAN	RTN4R NOGOR UNQ330/PRO526	473	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:12426574, ECO:0000269|PubMed:12694398, ECO:0000269|PubMed:12839991, ECO:0000269|PubMed:16712417, ECO:0000269|PubMed:18411262}; Lipid-anchor, GPI-anchor {ECO:0000269|PubMed:12694398}. Membrane raft {ECO:0000269|PubMed:12694398}. Cell projection, dendrite {ECO:0000250|UniProtKB:Q99PI8}. Cell projection, axon {ECO:0000250|UniProtKB:Q99PI8}. Perikaryon {ECO:0000250|UniProtKB:Q99M75}. Note=Detected along dendrites and axons, close to synapses, but clearly excluded from synapses. {ECO:0000250|UniProtKB:Q99PI8}.	26
Q9H2A7	CXL16_HUMAN	CXCL16 SCYB16 SRPSOX UNQ2759/PRO6714	254	SUBCELLULAR LOCATION: Cell membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}. Secreted. Note=Also exists as a soluble form.	29
Q9NP84	TNR12_HUMAN	TNFRSF12A FN14	129	SUBCELLULAR LOCATION: Membrane; Single-pass type I membrane protein.	27
Q9NQX5	NPDC1_HUMAN	NPDC1	325	SUBCELLULAR LOCATION: Membrane {ECO:0000305}; Single-pass membrane protein {ECO:0000305}.	34
Q9P0K1	ADA22_HUMAN	ADAM22 MDC2	906	SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:27066583}; Single-pass type I membrane protein {ECO:0000305}. Cell projection, axon {ECO:0000250|UniProtKB:Q9R1V6}.	25
Q9P2S2	NRX2A_HUMAN	NRXN2 KIAA0921	1712	SUBCELLULAR LOCATION: Presynaptic cell membrane {ECO:0000250|UniProtKB:Q9CS84}; Single-pass type I membrane protein {ECO:0000255}.	28
Q9UBV2	SE1L1_HUMAN	SEL1L TSA305 UNQ128/PRO1063	794	SUBCELLULAR LOCATION: Endoplasmic reticulum membrane {ECO:0000269|PubMed:16186509}; Single-pass type I membrane protein {ECO:0000269|PubMed:16186509}.	21
Q9UHI8	ATS1_HUMAN	ADAMTS1 KIAA1346 METH1	967	SUBCELLULAR LOCATION: Secreted, extracellular space, extracellular matrix {ECO:0000250}.	49
Q9ULB1	NRX1A_HUMAN	NRXN1 KIAA0578	1477	SUBCELLULAR LOCATION: Presynaptic cell membrane {ECO:0000250|UniProtKB:Q9CS84}; Single-pass type I membrane protein {ECO:0000305}.	30
Q9UNN8	EPCR_HUMAN	PROCR EPCR	238	SUBCELLULAR LOCATION: Membrane; Single-pass type I membrane protein.	17
Q9Y3B3	TMED7_HUMAN	TMED7 CGI-109	224	SUBCELLULAR LOCATION: Endoplasmic reticulum membrane; Single-pass type I membrane protein. Golgi apparatus, cis-Golgi network membrane; Single-pass type I membrane protein. Endoplasmic reticulum-Golgi intermediate compartment membrane; Single-pass type I membrane protein. Cytoplasmic vesicle, COPI-coated vesicle membrane {ECO:0000250}; Single-pass type I membrane protein {ECO:0000250}. Cytoplasmic vesicle, COPII-coated vesicle membrane {ECO:0000250}; Single-pass type I membrane protein {ECO:0000250}. Note=Cycles between compartments of the early secretatory pathway.	34
Q9Y4C0	NRX3A_HUMAN	NRXN3 C14orf60 KIAA0743	1643	SUBCELLULAR LOCATION: Presynaptic cell membrane {ECO:0000250|UniProtKB:Q9CS84}; Single-pass type I membrane protein {ECO:0000255}.	27
P01703	LV140_HUMAN	IGLV1-40	118	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
P01714	LV319_HUMAN	IGLV3-19	112	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
P01717	LV325_HUMAN	IGLV3-25	112	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
P01718	LV327_HUMAN	IGLV3-27	113	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
P01762	HV311_HUMAN	IGHV3-11	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
P01767	HV353_HUMAN	IGHV3-53	116	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
Q01459	DIAC_HUMAN	CTBS CTB	385	SUBCELLULAR LOCATION: Lysosome.	38
Q5FWE3	PRRT3_HUMAN	PRRT3 UNQ5823/PRO19642	981	SUBCELLULAR LOCATION: Membrane {ECO:0000305}; Multi-pass membrane protein {ECO:0000305}.	27
Q99727	TIMP4_HUMAN	TIMP4	224	SUBCELLULAR LOCATION: Secreted.	29
Q9NS98	SEM3G_HUMAN	SEMA3G	782	SUBCELLULAR LOCATION: Secreted {ECO:0000250}.	22
A0A075B6H8	KVD42_HUMAN	IGKV1D-42	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A075B6Q5	HV364_HUMAN	IGHV3-64	118	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
A0A075B6S4	KVD17_HUMAN	IGKV1D-17	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A075B6S9	KV137_HUMAN	IGKV1-37	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A087WSY6	KVD15_HUMAN	IGKV3D-15	115	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
A0A0A0MRZ8	KVD11_HUMAN	IGKV3D-11	115	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	20
A0A0A0MS15	HV349_HUMAN	IGHV3-49	119	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0A0MT36	KVD21_HUMAN	IGKV6D-21	114	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19

A0A0B4J1V0	HV315_HUMAN	IGHV3-15	119	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0B4J1V6	HV373_HUMAN	IGHV3-73	119	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0B4J1Y9	HV372_HUMAN	IGHV3-72	119	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0B4J2D9	KVD13_HUMAN	IGKV1D-13	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A0C4DH32	HV320_HUMAN	IGHV3-20	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0C4DH34	HV428_HUMAN	IGHV4-28	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0C4DH38	HV551_HUMAN	IGHV5-51	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0C4DH42	HV366_HUMAN	IGHV3-66	116	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
A0A0C4DH55	KVD07_HUMAN	IGKV3D-7	119	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	23
A0A0C4DH69	KV109_HUMAN	IGKV1-9	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A0C4DH72	KV106_HUMAN	IGKV1-6	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A0C4DH73	KV112_HUMAN	IGKV1-12	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	22
A0A0J9YXX1	HV5X1_HUMAN	IGHV5-10-1	117	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
O14498	ISLR_HUMAN	ISLR UNQ189/PRO215	428	SUBCELLULAR LOCATION: Secreted {ECO:0000305}.	18
P0DP04	HV43D_HUMAN	IGHV3-43D	118	SUBCELLULAR LOCATION: Secreted {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}. Cell membrane {ECO:0000303|PubMed:20176268, ECO:0000303|PubMed:22158414}.	19
Q5BLP8	NICOL_HUMAN	NICOL1 C4orf48	95	SUBCELLULAR LOCATION: Secreted {ECO:0000250|UniProtKB:Q3UR78}.	34
A6NLU5	VTM2B_HUMAN	VSTM2B	285	SUBCELLULAR LOCATION: Membrane {ECO:0000305}; Single-pass type I membrane protein {ECO:0000305}.	28

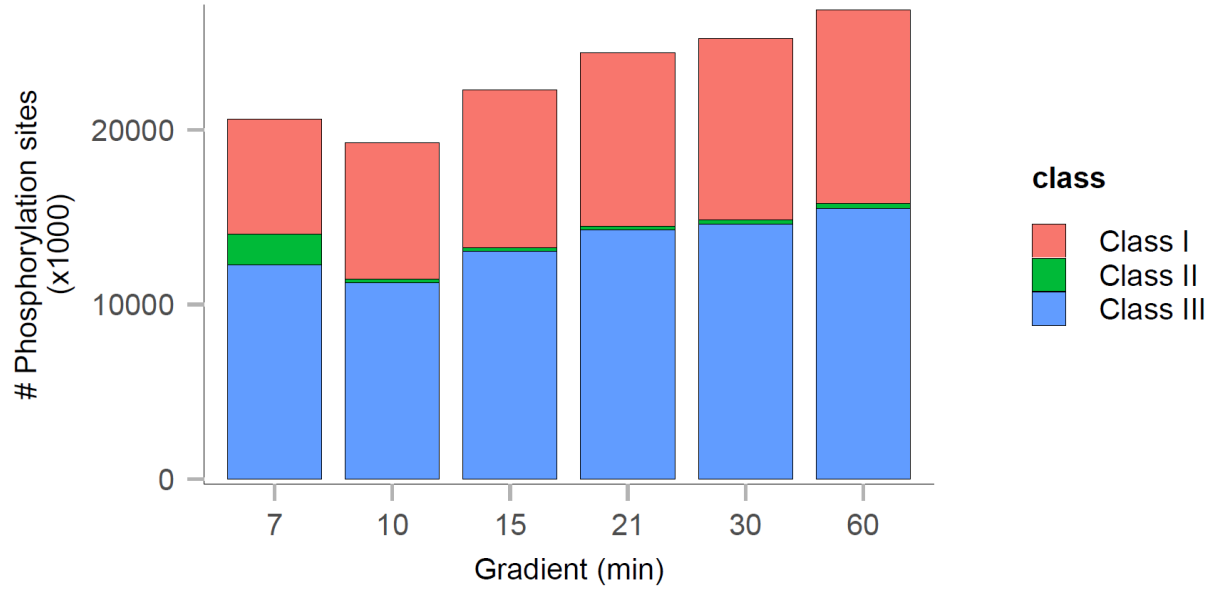


Appendix Figure C-2. Modifications identified using the open search workflow in FragPipe using pseudo-MS/MS spectra generated by diaTracer in CSF dataset.



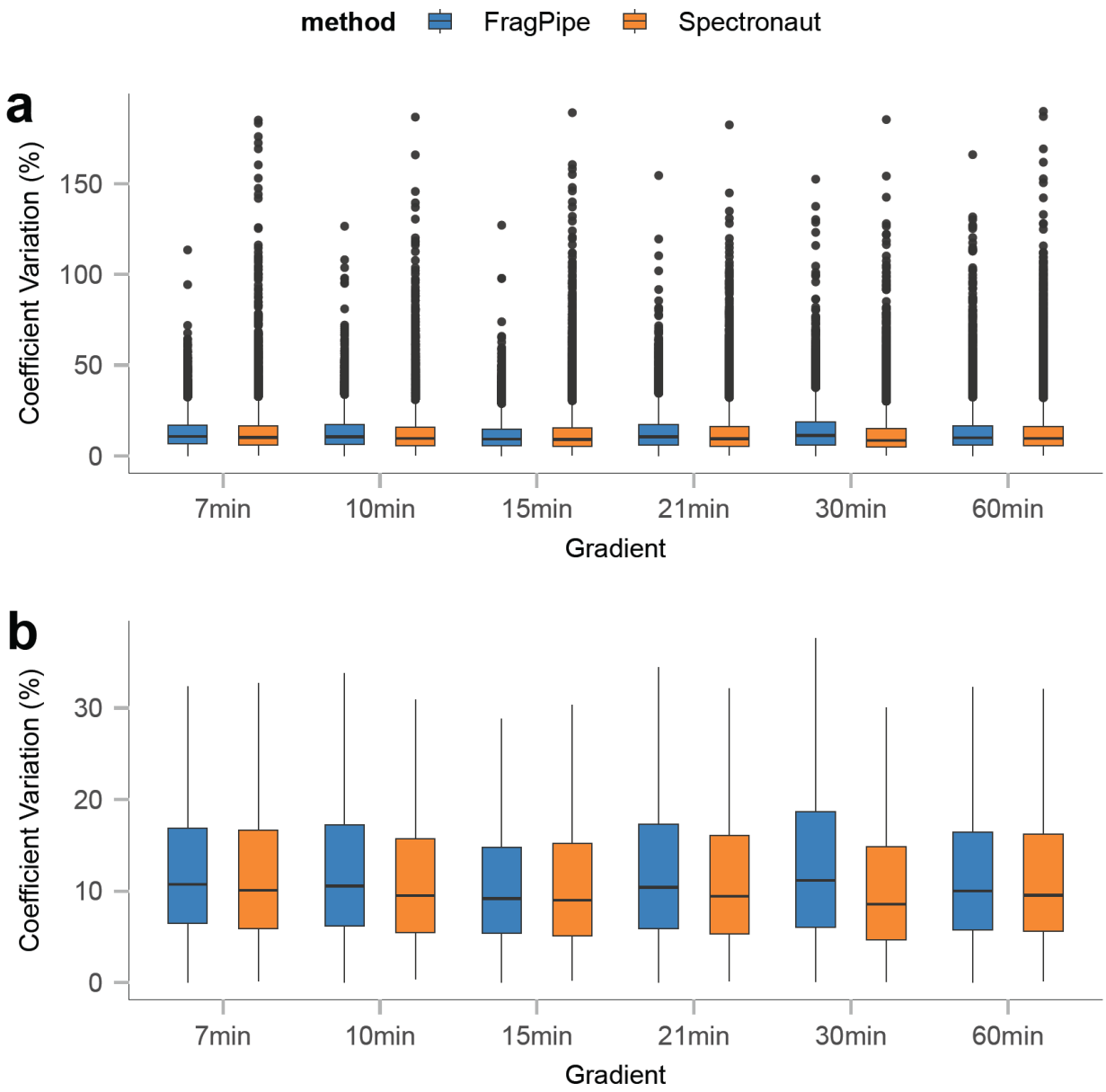
Appendix Figure C-3. Comparison of identified peptides between tryptic search and semi-tryptic search in plasma dataset.

Appendix Figure C-4. Wood's plot of protein H2AX showing quantified tryptic and semi-tryptic (with star) peptides.



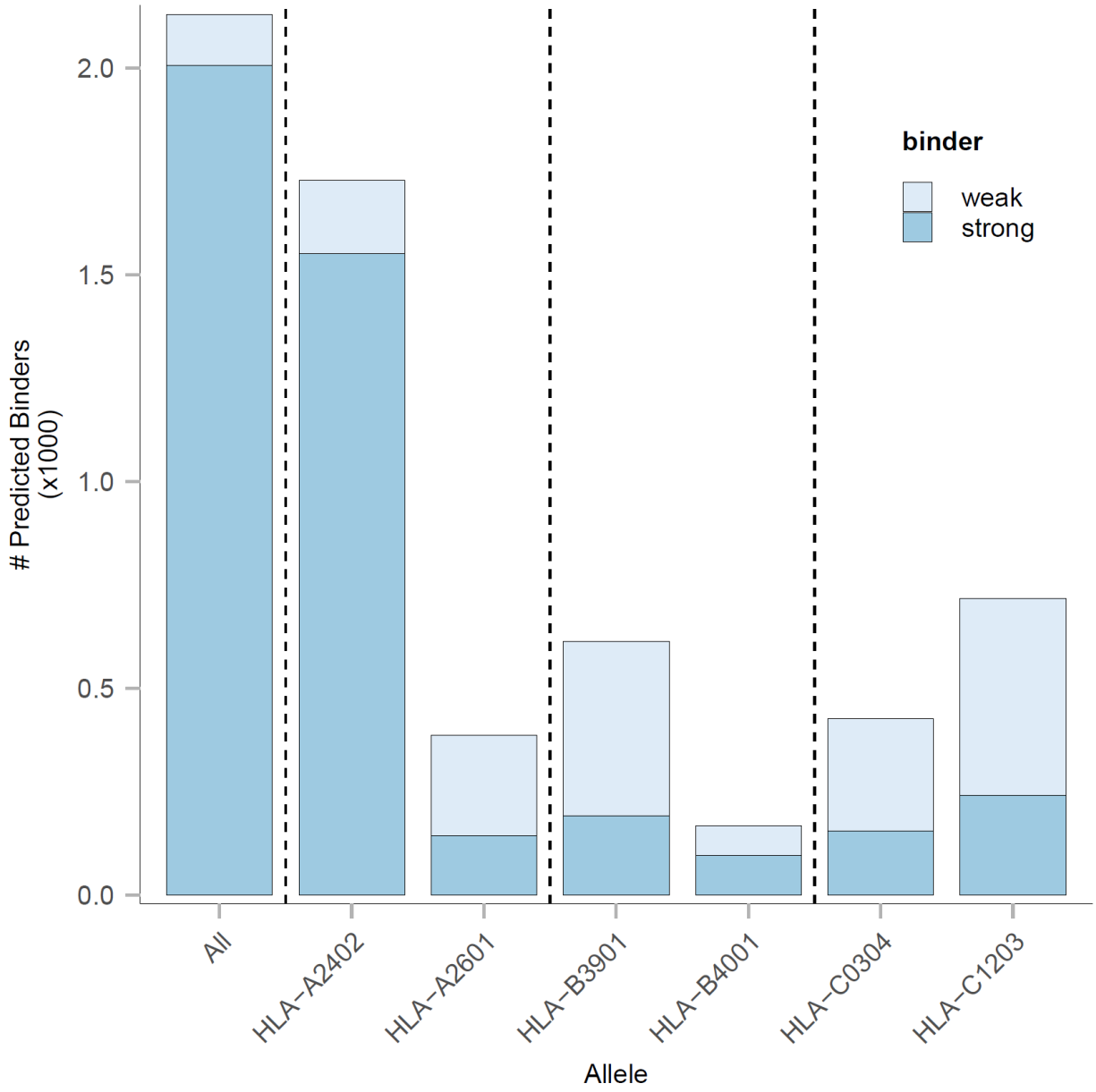
Appendix Figure C-5. Results of the phosphoproteomics dataset.

Total unique phosphorylation sites identified by FragPipe (pink: class I sites, localization probability > 0.75; green: class II sites, localization probability > 0.5 and <=0.75; blue: class III sites, localization probability <= 0.5).

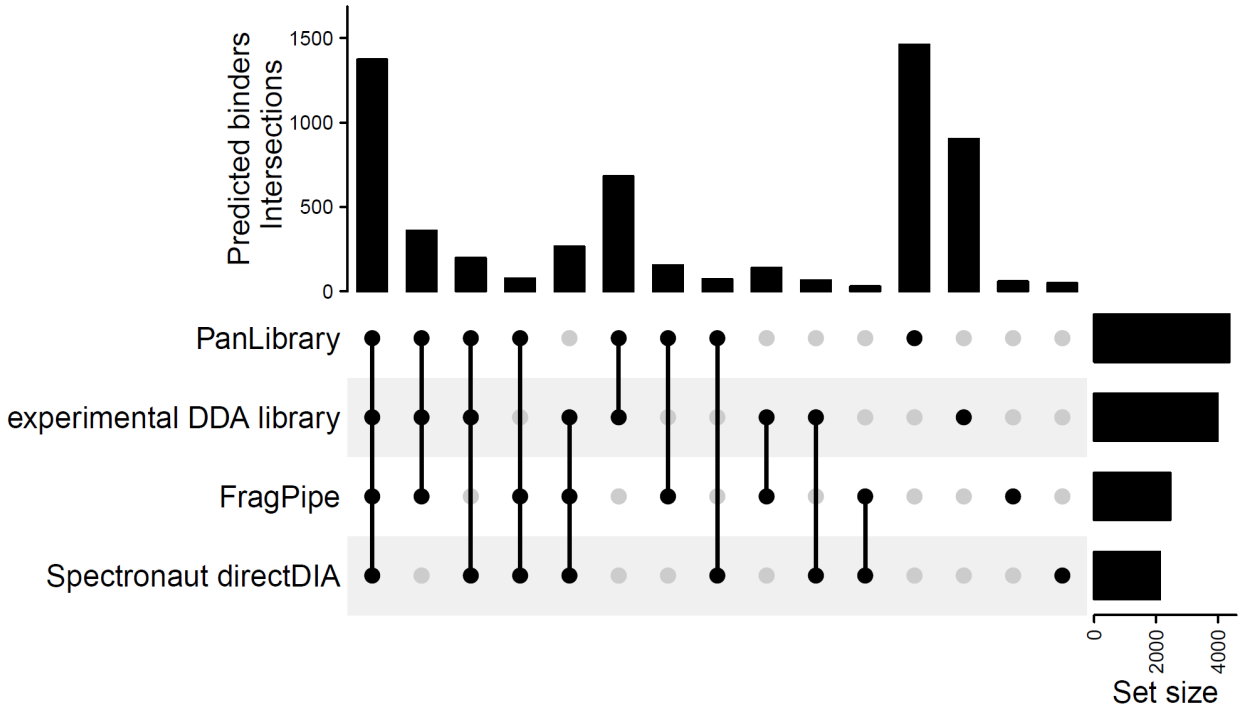


Appendix Figure C-6. CV based on phosphorylated precursors.

Box plots showing coefficient variation (CV) based on phosphorylated precursor (phosphorylation probability >= 0.75) from 7-60min gradients, with colors representing processing method (blue: FragPipe with diaTracer; orange: Spectronaut). There are 4 single-shot DIA runs from 4 replicates for each gradient. a) Boxplot of full range of CVs with outliers; b) Boxplot of CVs without outliers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The central line represents the median of the numbers. The interquartile range (IQR) is the box between Q1 and Q3. Data points outside this range are considered outliers and are shown as individual dots.



Appendix Figure C-7. Histogram of predicted binders of Spectronaut directDIA result from the Wahle et al. study for all HLA alleles of the corresponding sample donor, colored by binder type (light: weak binder; dark: strong binder).



Appendix Figure C-8**.** Predicted binders overlapped with the Wahle et al. study, including the Spectronaut directDIA, experimental DDA library, and panlibrary based results.

Bibliography

1 Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat Commun* **12**, 3810 (2021). https://doi.org:10.1038/s41467-021-24030-x

2 Aebersold, R. *et al.* How many human proteoforms are there? *Nat Chem Biol* **14**, 206-214 (2018). https://doi.org:10.1038/Nchembio.2576

3 Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol* **17**, 121-127 (1999). https://doi.org:Doi 10.1016/S0167-7799(98)01245-1

4 Han, X., Aslanian, A. & Yates, J. R., 3rd. Mass spectrometry for proteomics. *Curr Opin Chem Biol* **12**, 483-490 (2008). https://doi.org:10.1016/j.cbpa.2008.07.024

5 Zhang, Y. Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem Rev* **113**, 2343-2394 (2013). https://doi.org:10.1021/cr3003533

6 Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface* **5**, 3-20 (2012).

7 Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **314**, 65-66 (2006). https://doi.org:10.1126/science.1133987

8 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003). https://doi.org:10.1038/nature01511

9 Yamashita, M. & Fenn, J. B. Electrospray Ion-Source - Another Variation on the Free-Jet Theme. *J Phys Chem-Us* **88**, 4451-4459 (1984). https://doi.org:DOI 10.1021/j150664a002

10 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **246**, 64-71 (1989). https://doi.org:DOI 10.1126/science.2675315

11 Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **9**, 555-566 (2012). https://doi.org:10.1038/Nmeth.2015

12 Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. *Cell* **138**, 795-806 (2009). https://doi.org:10.1016/j.cell.2009.05.051

13 Sobsey, C. A. *et al.* Targeted and Untargeted Proteomics Approaches in Biomarker Development. *Proteomics* **20** (2020). https://doi.org:10.1002/pmic.201900029

14 Stahl, D. C., Swiderek, K. M., Davis, M. T. & Lee, T. D. Data-controlled automation of liquid chromatography tandem mass spectrometry analysis of peptide mixtures. *J Am Soc Mass Spectr* **7**, 532-540 (1996). https://doi.org:Doi 10.1016/1044-0305(96)00057-8

15 Yates, J. R., Eng, J. K., Mccormack, A. L. & Schieltz, D. Method to Correlate Tandem Mass-Spectra of Modified Peptides to Amino-Acid-Sequences in the Protein Database. *Anal Chem* **67**, 1426-1436 (1995). https://doi.org:DOI 10.1021/ac00104a020

16 Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-473 (2001). https://doi.org:DOI 10.1146/annurev.biochem.70.1.437
17 Kohli, B. M., Eng, J. K., Nitsch, R. M. & Konietzko, U. An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments. *Rapid Commun Mass Sp* **19**, 589-596 (2005). https://doi.org:10.1002/rcm.1827
18 Meyer, J. G. Fast Proteome Identification and Quantification from Data-Dependent Acquisition-Tandem Mass Spectrometry (DDA MS/MS) Using Free Software Tools. *Method Protocol* **2** (2019). https://doi.org:10.3390/mps2010008
19 Bateman, N. W. *et al.* Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Mol Cell Proteomics* **13**, 329-338 (2014). https://doi.org:10.1074/mcp.M112.026500
20 Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res* **10**, 1785-1793 (2011). https://doi.org:10.1021/pr101060v
21 Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* **14**, 1400-1410 (2015). https://doi.org:10.1074/mcp.M114.044305
22 Bondarenko, P. V., Chelius, D. & Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74**, 4741-4749 (2002). https://doi.org:10.1021/ac0256991
23 Zhu, W., Smith, J. W. & Huang, C. M. Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol* **2010**, 840518 (2010). https://doi.org:10.1155/2010/840518
24 Purvine, S., Eppel, J. T., Yi, E. C. & Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847-850 (2003). https://doi.org:10.1002/pmic.200300362
25 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111 016717 (2012). https://doi.org:10.1074/mcp.O111.016717
26 Plumb, R. S. *et al.* UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom* **20**, 1989-1994 (2006). https://doi.org:10.1002/rcm.2550
27 Panchaud, A. *et al.* Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem* **81**, 6481-6488 (2009). https://doi.org:10.1021/ac900888s
28 Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics* **9**, 2252-2261 (2010). https://doi.org:10.1074/mcp.M110.001537
29 Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* **10**, 744-746 (2013). https://doi.org:10.1038/nmeth.2528
30 Moseley, M. A. *et al.* Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *J Proteome Res* **17**, 770-779 (2018). https://doi.org:10.1021/acs.jproteome.7b00464

31 Koopmans, F., Ho, J. T. C., Smit, A. B. & Li, K. W. Comparative Analyses of Data Independent Acquisition Mass Spectrometric Approaches: DIA, WiSIM-DIA, and Untargeted DIA. *Proteomics* **18** (2018). https://doi.org:10.1002/pmic.201700304
32 Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods* **15**, 440-448 (2018). https://doi.org:10.1038/s41592-018-0003-5
33 Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**, 39-45 (2004). https://doi.org:10.1038/nmeth705
34 Muntel, J. *et al.* Advancing Urinary Protein Biomarker Discovery by Data-Independent Acquisition on a Quadrupole-Orbitrap Mass Spectrometer. *J Proteome Res* **14**, 4752-4762 (2015). https://doi.org:10.1021/acs.jproteome.5b00826
35 Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol Cell Proteomics* **16**, 2296-2309 (2017). https://doi.org:10.1074/mcp.RA117.000314
36 Barkovits, K. *et al.* Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition. *Mol Cell Proteomics* **19**, 181-197 (2020). https://doi.org:10.1074/mcp.RA119.001714
37 Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T. & Bruce, J. E. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J Proteome Res* **11**, 1621-1632 (2012). https://doi.org:10.1021/pr2008175
38 Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int J Ion Mobil Spectrom* **14** (2011). https://doi.org:10.1007/s12127-011-0067-8
39 Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev Sci Instrum* **82**, 126106 (2011). https://doi.org:10.1063/1.3665933
40 Michelmann, K., Silveira, J. A., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry. *J Am Soc Mass Spectrom* **26**, 14-24 (2015). https://doi.org:10.1007/s13361-014-0999-4
41 Meier, F. *et al.* Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J Proteome Res* **14**, 5378-5387 (2015). https://doi.org:10.1021/acs.jproteome.5b00932
42 Meier, F. *et al.* Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* **17**, 2534-2545 (2018). https://doi.org:10.1074/mcp.TIR118.000900
43 Meier, F. *et al.* diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. *Nat Methods* **17**, 1229-+ (2020). https://doi.org:10.1038/s41592-020-00998-0
44 Skowronek, P. *et al.* Synchro-PASEF Allows Precursor-Specific Fragment Ion Extraction and Interference Removal in Data-Independent Acquisition. *Mol Cell Proteomics* **22**, 100489 (2023). https://doi.org:10.1016/j.mcpro.2022.100489
45 Lukasz Szyrwiel, L. S., Markus Ralser, Vadim Demichev. Slice-PASEF: fragmenting all ions for maximum sensitivity in proteomics. *bioRxiv* (2022). https://doi.org:https://doi.org/10.1101/2022.10.31.514544

46 Lou, R. & Shui, W. Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023. *Mol Cell Proteomics* **23**, 100712 (2024). https://doi.org:10.1016/j.mcpro.2024.100712
47 Zhang, F. F., Ge, W. G., Ruan, G., Cai, X. & Guo, T. N. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics* **20** (2020). https://doi.org:10.1002/pmic.201900276
48 Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **16**, 509-+ (2019). https://doi.org:10.1038/s41592-019-0426-7
49 Zeng, W. F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat Commun* **13**, 7238 (2022). https://doi.org:10.1038/s41467-022-34904-3
50 Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat Methods* **18**, 1363-1369 (2021). https://doi.org:10.1038/s41592-021-01301-5
51 Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41-+ (2020). https://doi.org:10.1038/s41592-019-0638-x
52 Tsou, C. C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258-+ (2015). https://doi.org:10.1038/Nmeth.3255
53 Li, Y. Y. *et al.* Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat Methods* **12**, 1105-1106 (2015). https://doi.org:10.1038/nmeth.3593
54 Bruderer, R. *et al.* New targeted approaches for the quantification of data-independent acquisition mass spectrometry. *Proteomics* **17** (2017). https://doi.org:10.1002/pmic.201700021
55 Li, K., Teo, G. C., Yang, K. L., Yu, F. & Nesvizhskii, A. I. diaTracer enables spectrum-centric analysis of diaPASEF proteomics data. *Nat Commun* **16**, 95 (2025). https://doi.org:10.1038/s41467-024-55448-8
56 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016). https://doi.org:10.1038/nature19949
57 Yu, F. C. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications* **14** (2023). https://doi.org:10.1038/s41467-023-39869-5
58 Avtonomov, D. M., Polasky, D. A., Ruotolo, B. T. & Nesvizhskii, A. I. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Anal Chem* **90**, 2369-2375 (2018). https://doi.org:10.1021/acs.analchem.7b04999
59 Brakel, J. P. G. v. (2014).
60 Bentley, J. L. Multidimensional Binary Search Trees Used for Associative Searching. *Commun Acm* **18**, 509-517 (1975). https://doi.org:Doi 10.1145/361002.361007
61 Tsou, C. C., Tsai, C. F., Teo, G. C., Chen, Y. J. & Nesvizhskii, A. I. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics* **16**, 2257-2271 (2016). https://doi.org:10.1002/pmic.201500526

62 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-+ (2017). https://doi.org:10.1038/Nmeth.4256
63 Lapcik, P. *et al.* A hybrid DDA/DIA-PASEF based assay library for a deep proteotyping of triple-negative breast cancer. *Sci Data* **11**, 794 (2024). https://doi.org:10.1038/s41597-024-03632-2
64 Teo, G. C., Polasky, D. A., Yu, F. C. & Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *Journal of Proteome Research* **20**, 498-505 (2021). https://doi.org:10.1021/acs.jproteome.0c00544
65 Yu, F. C. *et al.* Identification of modified peptides using localization-aware open search. *Nature Communications* **11** (2020). https://doi.org:10.1038/s41467-020-17921-y
66 Pham, T. V., Henneman, A. A. & Jimenez, C. R. iq: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics. *Bioinformatics* **36**, 2611-2613 (2020). https://doi.org:10.1093/bioinformatics/btz961
67 Makhmut, A. *et al.* A framework for ultra-low-input spatial tissue proteomics. *Cell Systems* **14**, 1002-+ (2023). https://doi.org:10.1016/j.cels.2023.10.003
68 Hsiao, Y. *et al.* Analysis and Visualization of Quantitative Proteomics Data Using FragPipe-Analyst. *J Proteome Res* **23**, 4303-4315 (2024). https://doi.org:10.1021/acs.jproteome.4c00294
69 Geiszler, D. J. *et al.* PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results. *Molecular & Cellular Proteomics* **20** (2021). https://doi.org:10.1074/mcp.TIR120.002216
70 Polasky, D. A. *et al.* MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics. *Molecular & Cellular Proteomics* **22** (2023). https://doi.org:10.1016/j.mcpro.2023.100538
71 Yang, K. L. *et al.* MSBooster: improving peptide identification rates using deep learning-based features. *Nature Communications* **14** (2023). https://doi.org:10.1038/s41467-023-40129-9
72 Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923-925 (2007). https://doi.org:10.1038/Nmeth1113
73 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002). https://doi.org:10.1021/ac025747h
74 Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003). https://doi.org:10.1021/ac0341261
75 Leprevost, F. D. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* **17**, 869-870 (2020). https://doi.org:10.1038/s41592-020-0912-y
76 Shteynberg, D. D. *et al.* PTMProphet: Fast and Accurate Mass Modification the Trans-Proteomic Pipeline. *Journal of Proteome Research* **18**, 4262-4272 (2019). https://doi.org:10.1021/acs.jproteome.9b00205
77 Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nature Communications* **13** (2022). https://doi.org:10.1038/s41467-022-31492-0

78 MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010). https://doi.org:10.1093/bioinformatics/btq054
79 Li, K., Vaudel, M., Zhang, B., Ren, Y. & Wen, B. PDV: an integrative proteomics data viewer. *Bioinformatics* **35**, 1249-1251 (2019). https://doi.org:10.1093/bioinformatics/bty770
80 Kohler, D. *et al.* MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. *Journal of Proteome Research* **22**, 1466-1482 (2023). https://doi.org:10.1021/acs.jproteome.2c00834
81 Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**, D442-D450 (2019). https://doi.org:10.1093/nar/gky1106
82 Mackmull, M. T. *et al.* Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker. *Nat Struct Mol Biol* **29**, 978-+ (2022). https://doi.org:10.1038/s41594-022-00837-0
83 Kamalian, A. *et al.* Unveiling proteomic and peptide-level modifications in cerebrospinal fluid and plasma in normal cognitive aging. *Commun Med-London* **5** (2025). https://doi.org:10.1038/s43856-025-01183-0
84 Wang, B. *et al.* Structural Proteomic Profiling of Cerebrospinal Fluids to Reveal Novel Conformational Biomarkers for Alzheimer's Disease. *J Am Soc Mass Spectr* (2023). https://doi.org:10.1021/jasms.2c00332
85 Demir, F. *et al.* Proteolytic profiling of human plasma reveals an immunoactive complement C3 fragment. *Embo J* **44**, 7721-7758 (2025). https://doi.org:10.1038/s44318-025-00598-8
86 Mun, D. G. *et al.* Four-dimensional proteomics analysis of human cerebrospinal fluid with trapped ion mobility spectrometry using PASEF. *Proteomics* **23** (2023). https://doi.org:10.1002/pmic.202200507
87 Vitko, D. *et al.* timsTOF HT Improves Protein Identification and Quantitative Reproducibility for Deep Unbiased Plasma Protein Biomarker Discovery. *Journal of Proteome Research* **23**, 929-938 (2024). https://doi.org:10.1021/acs.jproteome.3c00646
88 Oliinyk, D. & Meier, F. Ion mobility-resolved phosphoproteomics with dia-PASEF and short gradients. *Proteomics* **23**, e2200032 (2023). https://doi.org:10.1002/pmic.202200032
89 Wahle, M. *et al.* IMBAS-MS Discovers Organ-Specific HLA Peptide Patterns in Plasma. *Mol Cell Proteomics* **23**, 100689 (2024). https://doi.org:10.1016/j.mcpro.2023.100689
90 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449-W454 (2020). https://doi.org:10.1093/nar/gkaa379
91 Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* **14**, 921-+ (2017). https://doi.org:10.1038/Nmeth.4398
92 Zougman, A. *et al.* Integrated analysis of the cerebrospinal fluid peptidome and proteome. *Journal of Proteome Research* **7**, 386-399 (2008). https://doi.org:10.1021/pr070501k

93 Guldbrandsen, A. *et al.* In-depth Characterization of the Cerebrospinal Fluid (CSF) Proteome Displayed Through the CSF Proteome Resource (CSF-PR). *Molecular & Cellular Proteomics* **13**, 3152-3163 (2014). https://doi.org:10.1074/mcp.M114.038554
94 Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* **13** (2017). https://doi.org:10.15252/msb.20156297
95 Anderson, N. L. & Anderson, N. G. The human plasma proteome - History, character, and diagnostic prospects. *Molecular & Cellular Proteomics* **1**, 845-867 (2002). https://doi.org:10.1074/mcp.R200007-MCP200
96 Liu, T. *et al.* AKT2 drives cancer progression and is negatively modulated by miR-124 in human lung adenocarcinoma. *Respiratory Research* **21** (2020). https://doi.org:10.1186/s12931-020-01491-0
97 Matthaios, D., Hountis, P., Karakitsos, P., Bouros, D. & Kakolyris, S. H2AX a promising biomarker for lung cancer: a review. *Cancer Invest* **31**, 582-599 (2013). https://doi.org:10.3109/07357907.2013.849721
98 Srinivasan, A., Sing, J. C., Gingras, A. C. & Röst, H. L. Improving Phosphoproteomics Profiling Using Data-Independent Mass Spectrometry. *Journal of Proteome Research* (2022). https://doi.org:10.1021/acs.jproteome.2c00172
99 Chang, C. H., Chang, H. Y., Rappsilber, J. & Ishihama, Y. Isolation of Acetylated and Unmodified Protein N-Terminal Peptides by Strong Cation Exchange Chromatographic Separation of TrypN-Digested Peptides. *Molecular & Cellular Proteomics* **20** (2021). https://doi.org:10.1074/mcp.TIR120.002148
100 Klein, T., Eckhard, U., Dufour, A., Solis, N. & Overall, C. M. Proteolytic Cleavage-Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chem Rev* **118**, 1137-1168 (2018). https://doi.org:10.1021/acs.chemrev.7b00120
101 Kuljanin, M. *et al.* Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nat Biotechnol* **39**, 630-641 (2021). https://doi.org:10.1038/s41587-020-00778-3
102 Desai, H. S. *et al.* SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State- Dependent Redox-Sensitive Cysteines. *Molecular & Cellular Proteomics* **21** (2022). https://doi.org:10.1016/j.mcpro.2022.100218
103 Lileikyte, G. *et al.* Proteomic analysis of serum samples after cardiac arrest: Rationale and design of a TTM-trial substudy. *Resusc Plus* **25** (2025). https://doi.org:10.1016/j.resplu.2025.101014
104 Okuda, S. *et al.* jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res* **45**, D1107-D1111 (2017). https://doi.org:10.1093/nar/gkw1080
105 Konno, R. *et al.* Thin-diaPASEF: diaPASEF for maximizing proteome coverage in single-shot proteomics. *DNA Res* **32** (2025). https://doi.org:10.1093/dnares/dsaf019
106 Skowronek, P. *et al.* Rapid and In-Depth Coverage of the (Phospho-)Proteome With Deep Libraries and Optimal Window Design for dia-PASEF. *Mol Cell Proteomics* **21**, 100279 (2022). https://doi.org:10.1016/j.mcpro.2022.100279
107 Wen, B. *et al.* Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics. *Nat Commun* **16**, 9815 (2025). https://doi.org:10.1038/s41467-025-64928-4