Advancing Proteomic Analyses with Graph-Based Deep Learning: Protein Inference and DIA De Novo Peptide Sequencing

by

Zheng Ma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor(s):	Ali Ghodsi Professor, Department of Statistical and Actuarial Science, University of Waterloo
	Ming Li Professor, Cheriton School of Computer Science, University of Waterloo
Internal Member:	Yang Lu Assistant Professor, Cheriton School of Computer Science, University of Waterloo
	Jeff Orchard Associate Professor, Cheriton School of Computer Science, University of Waterloo
Internal-External Member:	Andrew Doxey Associate Professor, Department of Biology, University of Waterloo
External Member:	Haixu Tang Professor, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I contributed to the model design and implementation of GraphPI and DIANovo, and conceived the research ideas together with my co-authors, under the guidance of Professor Ali Ghodsi and Professor Ming Li. For the theoretical analysis, I conceived the research ideas under guidance from Professor Ali Ghodsi and Professor Ming Li. I designed and executed the experimental protocols, and drafted the initial versions of the manuscripts, later refining them to ensure clarity and precision. My supervisors and co-authors also provided essential guidance, critical feedback, and editorial support throughout the process. Their contributions, particularly in offering supervision, strategic insights, and assistance in securing funding and resources, were invaluable in refining and strengthening the research outcomes.

Abstract

Proteomic analysis plays a central role in unraveling the complex molecular underpinnings of biological systems. However, traditional approaches to protein inference and peptide sequencing have been hampered by challenges such as data complexity, label scarcity, and spectral noise. In this thesis, we leverage advanced deep learning techniques to address these challenges, thereby expanding the efficacy of proteomic analyses.

Our work is organized around three major contributions. First, we introduce GraphPI, a novel protein inference framework that redefines the inference problem as a node classification task within a tripartite graph structure. In GraphPI, proteins, peptides, and peptide-spectrum matches (PSMs) are modeled as interconnected nodes, while edges incorporate features such as peptide identification scores and a specialized peptide-sharing attribute. By harnessing a tailored graph neural network (GNN) architecture inspired by GraphSAGE, our approach effectively aggregates and propagates information across heterogeneous node types. Critically, GraphPI is trained in a semi-supervised manner using pseudo-labels generated from established protein inference methods, combined with hard negative decoy information. This training process not only circumvents the typical bottleneck of limited labeled data but also yields protein scores that generalize across diverse datasets, all while substantially reducing computational overhead relative to Bayesian network–based approaches. Experimental evaluations on multiple benchmark datasets demonstrate that GraphPI delivers competitive accuracy with significant speed improvements, thus paving the way for real-time applications in large-scale proteomic studies.

Second, we present DIANovo, an innovative deep learning method designed to tackle the inherent complexities of Data-Independent Acquisition (DIA) data for de novo peptide sequencing. Unlike conventional de novo approaches that often struggle with the multiplexed nature of DIA spectra, DIANovo incorporates a suite of strategies to manage coelution and spectral noise. Our approach begins by constructing a spectrum graph that captures the mass differences between peaks. Next, a Transformer-based encoder, enhanced with Rotary Positional Embeddings (RoPE), processes the graph by encoding these mass differences along its edges, effectively treating the spectrum graph as fully connected. Furthermore, DIANovo introduces a coelution-aware pretraining stage, where the model is first optimized to predict ion types from coeluting peptides. This pretraining step equips the network with a nuanced understanding of spectral interferences, thereby improving the fidelity of subsequent peptide sequence predictions. In addition, a two-stage decoding strategy is employed: the first stage identifies an optimal path through the spectrum graph, while the second refines this path to generate a final amino acid sequence by filling in mass tags. Comparative analyses against state-of-the-art methods reveal that DIANovo achieves

significant improvements in both amino acid and peptide recall, especially when applied to high-quality narrow-window DIA data obtained from next-generation instruments such as the Orbitrap Astral. Moreover, we investigate whether DIA identifies more peptides than DDA in de novo sequencing by comparing their performance on the same biological sample under varying acquisition modes and parameters. Our results demonstrate that DIA only outperforms DDA when employing narrower isolation windows.

The third component of this thesis presents a comprehensive theoretical analysis that sheds light on the performance limits of peptide identification methods. By linking the signal-to-noise profile to peptide identification accuracy, our study elucidates the inherent trade-offs between Data-Dependent Acquisition (DDA) and DIA strategies. We derive quantitative metrics to predict peptide identification performance under a range of experimental conditions, and these predictions are validated against empirical data. This framework not only explains why Astral DIA data can provide superior peptide coverage in certain scenarios but also delineates the conditions under which peptide identification is most favorable. These insights are crucial for guiding the design of future mass spectrometry experiments and for optimizing computational pipelines in proteomic research.

Collectively, the three contributions of this thesis demonstrate the transformative potential of integrating deep learning with advanced computational frameworks in proteomics. GraphPI and DIANovo both showcase how novel neural network architectures can overcome longstanding challenges in protein inference and de novo peptide sequencing, while the theoretical analysis provides a foundation for understanding and further refining these methodologies. The experimental results across multiple datasets underscore the robustness, efficiency, and generalizability of our approaches, suggesting that deep learning–based strategies will play an increasingly central role in the future of proteomic analysis.

In conclusion, this work not only advances the state-of-the-art in protein and peptide identification but also offers practical solutions for handling large-scale, complex proteomic data. By bridging the gap between theoretical insights and practical implementations, our integrated framework lays the groundwork for enhanced biomarker discovery, more accurate disease diagnosis, and a deeper understanding of biological systems at the molecular level.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Professor Ming Li and Professor Ali Ghodsi, for their unwavering kindness and support. They have been understanding of my personal challenges throughout this journey and have offered me nothing but encouragement. Professor Ali Ghodsi has been aware of my circumstances from the outset and has consistently encouraged me to persevere and complete my program. His unwavering support, along with his invaluable guidance throughout my research, has been instrumental to my progress. Professor Ming Li provided me with a comprehensive introduction to proteomics and has been a steady and reliable source of support throughout my research.

I would also like to sincerely thank my wife, whose steadfast support has been a source of strength both in my academic pursuits and in every aspect of life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins are essential biomolecules that orchestrate virtually every cellular process, including metabolic regulation, signal transduction, and immune responses [5]. The complete set of proteins expressed by a cell, tissue, or organism at a given time—collectively known as the proteome—is highly dynamic and reflects the current state of biological systems [2]. Proteomic analysis is therefore crucial for understanding cellular functions, disease mechanisms, and for identifying potential biomarkers and therapeutic targets [141][24]. Variations in protein expression patterns are frequently associated with diseases such as cancer and neurodegenerative disorders, underscoring the importance of accurate and comprehensive proteomic profiling [48][97].

1.1 Shotgun Proteomics

To navigate the complexity inherent in biological samples, shotgun proteomics has emerged as a pivotal high-throughput approach for comprehensive protein analysis [137]. This method involves the enzymatic digestion of proteins into smaller peptides, which are then analyzed using mass spectrometry to infer the protein content of the sample [146]. Central to this approach is tandem mass spectrometry (MS/MS), where peptides are first ionized and then fragmented to produce spectra representing the mass-to-charge (m/z) ratios of peptide fragments [115]. Peptide identification, the process of determining peptide sequences from MS/MS data, relies on interpreting these spectra [22]. Subsequently, protein identification involves linking the identified peptides back to their parent proteins, which is critical for understanding protein functions and interactions within the biological system [94].

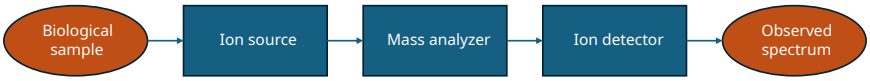


Figure 1.1: Components of mass spectrometer.

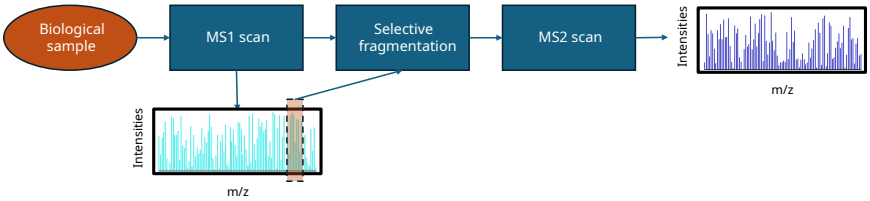


Figure 1.2: Tandem mass spectrometry process.

1.1.1 Tandem Mass Spectrometry

Mass spectrometry (MS) is a fundamental analytical technique that enables the identification and quantification of molecules by measuring their mass-to-charge ratio (m/z). It has revolutionized multiple scientific disciplines, including chemistry, physics, and life sciences, due to its high sensitivity, precision, and capability to analyze complex biological mixtures. A typical mass spectrometer consists of three primary components: an ion source, which ionizes the sample molecules into the gas phase; a mass analyzer, which separates these ions based on their m/z values; and an ion detector, which measures the abundance of detected ions, generating a mass spectrum (Figure 1.1) [1] [28]. This conventional MS setup provides valuable insights into molecular composition but often lacks the structural information required for in-depth analysis of biomolecules such as peptides and proteins.

To overcome this limitation, tandem mass spectrometry (MS/MS) extends conventional MS by incorporating multiple rounds of mass selection and fragmentation, allowing for a more detailed examination of molecular structures. In MS/MS, the first mass spectrometry stage (MS1 scan) records the m/z values of all ionized molecules present in the sample. A subset of these ions, referred to as precursor ions, is then isolated and subjected to fragmentation using dissociation techniques such as collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), or electron-transfer dissociation (ETD) [96][118]. These fragmentation methods break the precursor ions into smaller fragment ions, which are subsequently analyzed in a second mass spectrometry stage (MS2 scan), producing an MS2 spectrum (Figure 1.2). This spectrum provides structural information about the precursor ions, making MS/MS a crucial tool in proteomics, lipidomics, and metabolomics [13][16].

In proteomics, tandem mass spectrometry is widely used for peptide and protein iden-

tification, enabling large-scale characterization of biological samples. The process typically involves digesting proteins into peptides using proteolytic enzymes such as trypsin before MS analysis. The resulting peptide fragments are then analyzed in an MS1 scan, followed by fragmentation and analysis in an MS2 scan. The selection of precursor ions for fragmentation is guided by different acquisition strategies, the most common being Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA). In DDA, only a fixed number of the most abundant precursor ions are selected for fragmentation per scan cycle [85]. In contrast, DIA fragments all precursor ions within a specified m/z range [42][77].

MS/MS has become an indispensable tool in modern biomedical research, enabling the identification of post-translational modifications (PTMs), protein-protein interactions, and biomarker discovery for diseases such as cancer and neurodegenerative disorders [148][113]. Advances in MS-based proteomics have also facilitated breakthroughs in precision medicine, where mass spectrometry is used to identify disease-specific molecular signatures, aiding in the development of targeted therapies [41]. Moreover, MS/MS plays a critical role in metabolomics and lipidomics, allowing researchers to study small molecules and lipids in various biological systems [104][15].

In summary, tandem mass spectrometry significantly enhances the analytical power of traditional MS by providing detailed structural information through ion fragmentation and sequential mass analysis. Its applications extend beyond proteomics to metabolomics, lipidomics, and clinical diagnostics, making it a cornerstone technique in molecular and systems biology.

1.1.2 Identification Modes in Shotgun Proteomics

The identification of proteins from mass spectrometry data typically involves two stages, including peptide sequencing and protein identification. As shown in Figure 1.3, protein identification methods can be broadly categorized into database-dependent and database-free approaches. Database-dependent methods rely on prior knowledge from known protein sequences, while database-free methods infer sequences directly from MS data without requiring a reference database. This section provides an overview of peptide database search, spectral library search, and protein inference (database-dependent approaches), as well as de novo peptide sequencing and de novo protein assembly (database-free approaches).

Database-Dependent Approaches

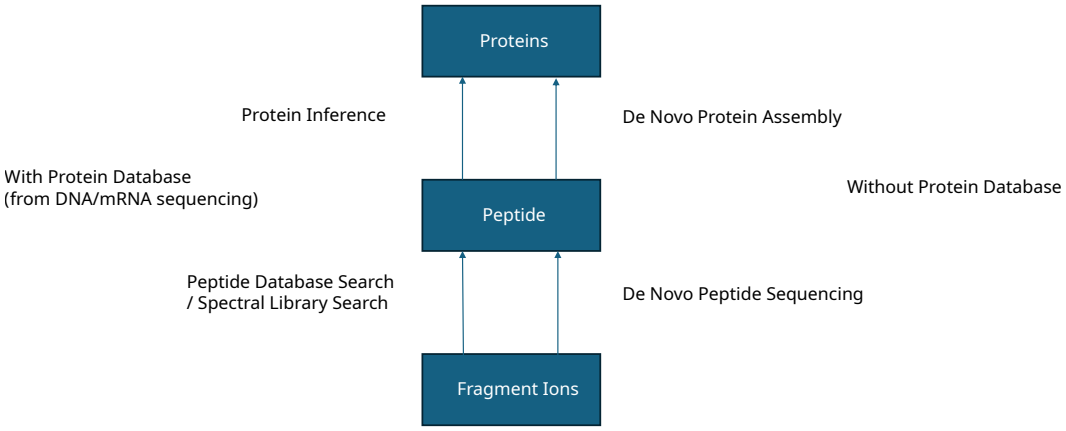


Figure 1.3: Identification modes in shotgun proteomics.

- Peptide Database Search

  Peptide database search is one of the most common computational approaches for peptide identification. In this method, MS/MS spectra are matched against a protein sequence database, which is typically constructed from genome or transcriptome sequencing data [1]. The database is first in silico digested using the cleavage rules of a specific protease (e.g., trypsin) to generate a list of theoretical peptides. Each experimental MS2 spectrum is then compared to these theoretical peptides, and a similarity score is assigned to rank potential matches. The peptide with the highest peptide-spectrum match (PSM) score is selected as the most likely identification [34].

  A key advantage of database searching is the ability to control the false discovery rate (FDR) using the target-decoy search strategy. This strategy involves generating a set of decoy sequences (e.g., reversed or shuffled protein sequences) and searching them alongside the real database [31]. By analyzing the frequency of incorrect identifications from the decoy set, researchers can estimate the FDR and filter results to maintain high confidence identifications. Popular database search tools include SEQUEST [34], Mascot [99], X!Tandem [23], Comet [33], MaxQuant [21], and PEAKS [145].

  Despite its widespread use, peptide database search has limitations. It cannot identify novel peptides that are absent from the reference database, which restricts its applicability for studying mutated, modified, or non-canonical proteins [93]. Additionally, it relies on accurate precursor mass measurements, as incorrect mass assignments can lead to missed identifications. [34]

- Spectral Library Search

  Spectral library search is an alternative to traditional peptide database search that uses experimentally derived peptide spectra instead of theoretical sequences. A spectral library is a collection of high-quality MS/MS spectra annotated with known peptide sequences [66]. When analyzing a new sample, the experimental spectra are compared directly to this reference library, and the best-matching spectrum is selected for peptide identification.

  One major advantage of spectral library search is its speed and efficiency. Because the search space is significantly reduced compared to sequence-based database searching, spectral library search is computationally faster and can improve the sensitivity of peptide identification [40]. Moreover, spectral libraries contain rich fragmentation pattern information, allowing for better discrimination between correct and incorrect matches.

  However, spectral library searching also has limitations. It requires high-quality reference spectra, meaning that peptides not represented in the library cannot be identified. Additionally, constructing comprehensive spectral libraries can be labor-intensive and may require extensive prior experimentation. This method is particularly useful for data-independent acquisition (DIA) proteomics, where it often outperforms database searching in identification sensitivity [42]. Common tools for spectral library search include SpectraST [66], Spectronaut [14], and OpenSWATH [105].

- Protein Inference

  Once peptides have been identified using either peptide database search or spectral library search, the next step is protein inference, where peptide sequences are mapped back to their corresponding proteins [94]. Because many peptides can be shared among multiple proteins, determining the exact source protein(s) is often ambiguous. Various computational methods, such as parsimony-based approaches, attempt to infer the minimal set of proteins that can explain the observed peptides.

  Protein inference benefits from having a prior protein sequence database, which allows for robust identification and quantification. However, it suffers from limitations such as ambiguity in peptide-to-protein mapping and incomplete sequence coverage, making it challenging to distinguish protein isoforms or homologous proteins [110].

Database-Free Approaches

- De Novo Peptide Sequencing

  Unlike database-dependent methods, de novo peptide sequencing reconstructs peptide sequences directly from MS/MS spectra, without relying on any existing sequence database [80]. This approach is essential when studying novel proteins, mutations, or post-translational modifications (PTMs) that are not represented in reference databases.

  The core idea of de novo sequencing is to analyze the fragment ion series produced by peptide fragmentation and infer the amino acid sequence step by step [39]. Various computational methods have been proposed to solve this problem, including spectrum graph analysis, dynamic programming, and probabilistic models [8]. More recently, deep learning has significantly improved de novo sequencing accuracy, with models such as DeepNovo [124] achieving state-of-the-art results.

  While de novo sequencing provides unbiased peptide identification, it is challenging due to spectral noise and missing fragment ions. Additionally, de novo sequencing typically has lower confidence than database search methods, and validation is often required using orthogonal techniques [145].

- De Novo Protein Assembly

  While de novo peptide sequencing reconstructs individual peptides, de novo protein assembly attempts to rebuild full protein sequences from identified peptides [111]. This process is significantly more challenging than peptide sequencing due to incomplete peptide coverage, sequencing errors, and ambiguous mappings.

  De novo protein assembly typically involves computational overlapping and stitching of peptides to reconstruct the full-length protein sequence. While it remains an emerging field, it has promising applications in proteogenomics, antibody sequencing, and novel protein discovery. [35]

  Compared to protein inference, which relies on existing databases, de novo protein assembly is entirely database-independent and can uncover previously unknown proteins. However, due to incomplete MS coverage, the assembled sequences often contain gaps or errors, requiring additional validation through RNA sequencing or long-read proteomics [113].

In summary, the identification of peptides and proteins in shotgun proteomics can be performed using database-dependent approaches, such as peptide database search, spectral library search, and protein inference, or database-free approaches, such as de novo peptide sequencing and de novo protein assembly. While database search methods benefit from

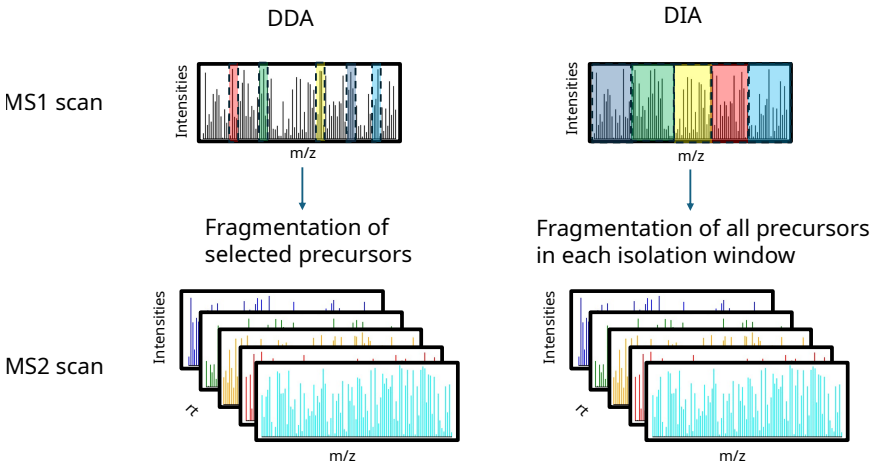


Figure 1.4: DDA vs DIA.

high confidence and controlled FDR, they are limited to known proteins. In contrast, de novo methods allow for the discovery of novel proteins but face challenges such as lower accuracy and increased computational complexity.

1.1.3 Data Dependent Acquisition (DDA) and Data Independent Acquisition (DIA)

Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA) are two widely used strategies for selecting precursor ions in tandem mass spectrometry (MS/MS) experiments. While both methods aim to generate MS2 spectra for peptide identification, they differ in their approaches to precursor ion selection and fragmentation, leading to distinct strengths and weaknesses. The comparison between DDA and DIA is illustrated in Figure 1.4.

DDA is a targeted fragmentation approach in which the mass spectrometer first performs an MS1 scan to detect all ionized peptides in the sample. It then selects a subset of the most intense precursor ions for fragmentation and MS2 analysis, prioritizing those with the highest signal at each scan cycle. This method produces high-quality MS2 spectra with a strong signal-to-noise ratio, making peptide identification relatively straightforward [1]. Additionally, DDA generates a manageable number and size of spectra, reducing computational complexity and making data analysis more efficient. Due to its long-standing use in proteomics, DDA is well-supported by existing protein databases and identification algorithms [28]. However, DDA has notable limitations. Because precursor selection is based

on ion intensity, low-abundance peptides are often underrepresented, leading to incomplete proteome coverage [85]. Furthermore, its stochastic nature means that different runs on the same sample may yield different sets of identified peptides, limiting reproducibility. In highly complex samples, such as clinical or environmental proteomes, DDA may fail to capture all relevant peptides, resulting in missing data [42].

In contrast, DIA is an untargeted fragmentation approach that systematically fragments all precursor ions within predefined m/z windows in each scan cycle. Unlike DDA, which isolates only a few precursor ions per cycle, DIA ensures that all detectable precursors are fragmented, leading to a more comprehensive dataset. This approach provides better reproducibility, as the same peptides are analyzed across multiple runs, making DIA particularly advantageous for large-scale proteomics studies [77]. Moreover, because all precursors are fragmented, DIA achieves improved identification and quantification by minimizing missing values [42]. However, these advantages come at the cost of increased data complexity. Since multiple precursor ions are fragmented simultaneously, MS2 spectra generated in DIA are significantly more complex than those in DDA, requiring sophisticated computational methods for peptide identification [105]. The sheer volume of data produced by DIA also demands higher computational resources, making data processing more challenging [91]. Despite these challenges, DIA is increasingly favored for applications requiring high reproducibility and comprehensive peptide detection.

The protein inference method we propose is theoretically compatible with both DDA and DIA data, as it does not impose restrictions on the acquisition strategy. However, our de novo sequencing project focuses exclusively on the DIA case, taking advantage of its ability to generate extensive peptide coverage. Additionally, our theoretical analysis of peptide identification performance applies to both DDA and DIA, ensuring that our findings remain relevant across different acquisition methods.

1.2 Deep Learning

Deep learning, a subset of machine learning, has emerged as a powerful computational approach for modeling complex patterns and relationships in data. It is characterized by the use of multi-layer artificial neural networks, which are capable of learning hierarchical representations from raw data. Unlike traditional machine learning methods that rely on manually designed features, deep learning enables end-to-end learning, allowing models to automatically extract and refine features through training. This ability to learn complex patterns makes deep learning particularly effective in domains where data is high-dimensional, noisy, or structured in non-trivial ways.

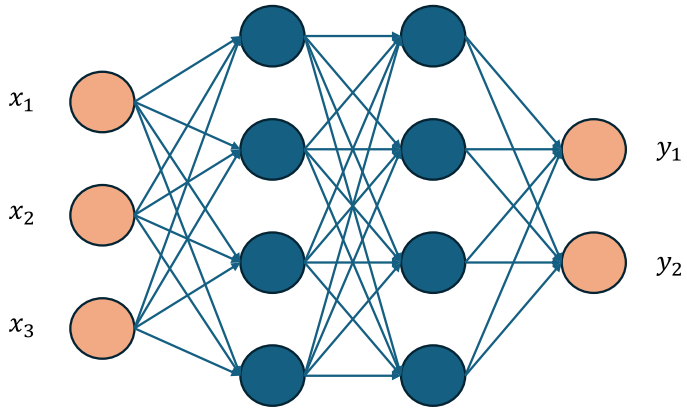


Figure 1.5: A three-layer MLP.

One of the most fundamental deep learning architectures is the multi-layer perceptron (MLP), a fully connected neural network where each layer of neurons applies a learned transformation to the input data before passing it to the next layer, illustrated in Figure 1.5. MLPs are theoretically universal function approximators, meaning that given enough neurons and layers, they can approximate any continuous function [53]. This universal learnability characteristic is crucial in deep learning, as it allows models to capture highly complex data distributions that are difficult to express with traditional rule-based approaches.

In recent years, deep learning has revolutionized a wide range of fields, including computer vision, natural language processing (NLP), speech recognition, drug discovery, and biomedical research [68]. The success of deep learning is largely attributed to three key factors. First, the availability of large-scale datasets has enabled models to learn from vast amounts of information, improving their ability to generalize. Second, advancements in computational hardware, particularly the widespread adoption of graphics processing units (GPUs) and tensor processing units (TPUs), have accelerated neural network training and inference, making deep learning feasible for large-scale applications. Finally, algorithmic innovations, such as improved activation functions, optimization techniques, and novel network architectures, have significantly enhanced model performance and scalability.

In the context of shotgun proteomics, deep learning has become an essential tool for improving peptide and protein identification, quantification, and functional annotation. The complexity of proteomics data arises from the vast combinatorial space of possible peptides, the variability in fragmentation patterns during mass spectrometry (MS), and the challenges in distinguishing signal from noise in MS spectra. Traditional approaches, such as

database searching and heuristic-based scoring functions, often struggle with noisy, incomplete, and ambiguous spectra, limiting their ability to identify peptides with high accuracy [43]. Deep learning methods address these challenges by learning complex fragmentation patterns in mass spectrometry data, allowing for more precise de novo peptide sequencing and improved scoring of peptide-spectrum matches in database searches [124][143]. Additionally, neural networks can infer missing peaks in mass spectrometry data, helping to reconstruct incomplete spectra and improve spectral interpretation. This ability to reconstruct spectra enhances peptide identification and reduces ambiguity in spectral matching [121].

Deep learning is also well-suited for handling the high-dimensional nature of proteomics data, particularly in modeling relationships within peptide fragmentation spectra, protein-peptide-spectrum mappings, and other structured representations of MS data. Architectures such as transformers and graph neural networks (GNNs) have demonstrated effectiveness in capturing these relationships. [81] Furthermore, deep learning enables large-scale proteomics studies by significantly reducing computational costs and increasing the speed of peptide-spectrum analysis. Modern neural networks can efficiently process millions of spectra, facilitating high-throughput proteomics workflows that were previously computationally infeasible [36] [121]. Given these advantages, deep learning is increasingly integrated into proteomics pipelines, enhancing both the accuracy and scalability of peptide and protein identification.

1.2.1 Graph Neural Networks

Graph neural networks (GNNs) have emerged as a powerful class of deep learning models designed to process data represented as graphs. Unlike traditional neural networks, which operate on structured data such as images or sequences, GNNs excel in learning representations from relational data, where entities (nodes) are connected through interactions (edges) [138]. The core idea behind GNNs is message passing, where each node iteratively aggregates information from its neighbors to update its representation. This enables the model to learn structural dependencies within the graph, making GNNs particularly effective for applications in social networks, molecular chemistry, knowledge graphs, and proteomics [149].

In the context of shotgun proteomics, graphs naturally arise when modeling the relationships between proteins, peptides, and mass spectra. These entities form a structured protein-peptide-spectrum graph, where nodes represent different biological components, and edges capture the interactions between them. In an MS2 spectrum, the peaks can be

represented as a spectrum graph, where nodes correspond to the masses of amino acids extending from the N-terminal, and edges indicate possible amino acid sequences connecting two nodes. GNNs provide an effective framework for analyzing these relationships by leveraging graph-based message passing to improve protein inference and peptide identification.

Among the various GNN architectures, GraphSAGE (Graph Sample and Aggregate) [47] is particularly well-suited for large-scale graphs, as it employs neighbor sampling instead of aggregating information from all connected nodes. Instead of considering the entire neighborhood of a node, GraphSAGE randomly samples a fixed number of neighbors and aggregates their information. This approach significantly reduces computational complexity while maintaining representative graph structure learning. The general update rule for GraphSAGE at each node v is:

$$h_v^{(k)} = \sigma\left(W_k \cdot \text{AGGREGATE}\left(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}\right)\right) \tag{1.1}$$

where:

- $h_v^{(k)}$ is the representation of node v at layer k.
- $\mathcal{N}(v)$ denotes the sampled neighborhood of node v.
- W_k is a trainable weight matrix.
- AGGREGATE represents different neighborhood aggregation mechanisms, such as mean pooling, max pooling, or LSTM-based aggregators. Although LSTMs are not permutation-invariant, they can be used when node neighborhoods are ordered (e.g., by structural or learned criteria), enabling more expressive, sequence-based aggregation as explored in GraphSAGE [47].
- σ is a non-linear activation function.

By learning node representations through sampled message passing, GraphSAGE generalizes to unseen nodes, making it highly effective for large and dynamic graphs.

Biological graphs, such as protein-peptide-spectrum graphs, exhibit heterogeneity, meaning they contain multiple node types and edge types, each with different semantic meanings. Standard GNNs typically assume homogeneous graphs, where all nodes and edges follow the same message passing scheme. However, in real-world proteomics applications, different types of interactions, such as protein-peptide bindings and peptide-spectrum matches, require distinct aggregation strategies.

GraphSAGE can be extended to handle heterogeneous graphs by assigning different message passing schemes to different types of nodes and edges. Specifically, GraphSAGE can adopt type-specific aggregation functions where each node type has a separate aggregation function to process neighborhood information differently. This advantage makes GraphSage suitable our protein inference problem.

1.2.2 Transformers

The Transformer architecture has become a foundational model in deep learning, particularly in domains involving sequential and structured data. Originally introduced for natural language processing tasks [131], the Transformer relies on a mechanism known as self-attention, which allows the model to capture dependencies between elements in a sequence regardless of their distance. This characteristic gives Transformers a significant advantage over traditional recurrent neural networks (RNNs), which model sequences sequentially and suffer from limitations in long-range dependency modeling.

At the heart of the Transformer is the scaled dot-product attention mechanism, illustrated in Figure 1.6(a). Given a query matrix Q, key matrix K, and value matrix V, attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1.2}$$

where

- Q is the query matrix
- K is the key matrix
- V is the value matrix
- d_k is the dimensionality of the key vectors

This mechanism computes the relevance of each token in the sequence with respect to all others using a similarity score between queries and keys. The resulting scores are normalized and used to compute a weighted sum of the value vectors, enabling the model to selectively focus on different parts of the input.

To enhance expressiveness, the Transformer uses multi-head attention (Figure 1.6(b)), where multiple attention mechanisms (or "heads") operate in parallel. Each head learns

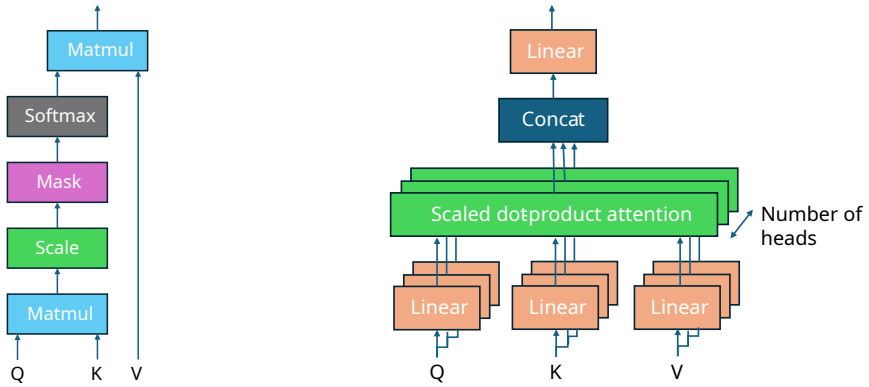


Figure 1.6: (a) Scaled dot-product attention, and (b) multi-head attention.

to focus on different aspects of the input, and their outputs are concatenated and linearly transformed. This design enables the model to capture richer and more diverse patterns in the data, improving its capacity to model complex relationships.

Beyond their application to sequences, Transformers can be adapted to graph-structured data, leading to a class of models known as graph Transformers. A notable example is Graphormer [139], which modifies the attention mechanism to incorporate edge information and graph topology. In Graphormer, structural priors such as shortest path distances, edge types, and centrality encodings are embedded directly into the attention matrix, allowing the model to learn both node features and connectivity patterns in a unified attention-based framework. Unlike traditional GNNs that rely on local neighborhood aggregation, graph Transformers offer global receptive fields, enabling every node to attend to every other node with awareness of the graph's structural context.

The overall encoder-decoder structure of the Transformer is depicted in Figure 1.7. The encoder is composed of stacked layers that include multi-head self-attention and feed-forward sublayers, and it maps the input sequence to a series of contextualized representations. The decoder follows a similar structure but includes an additional cross-attention sublayer that allows it to attend to the encoder outputs. The decoder generates outputs step-by-step using causal self-attention, which ensures that each position can only attend to previous positions, preventing information leakage. Both encoder and decoder layers are equipped with residual connections and layer normalization, which stabilize training and improve convergence.

This encoder-decoder architecture is particularly well-suited for de novo peptide se-

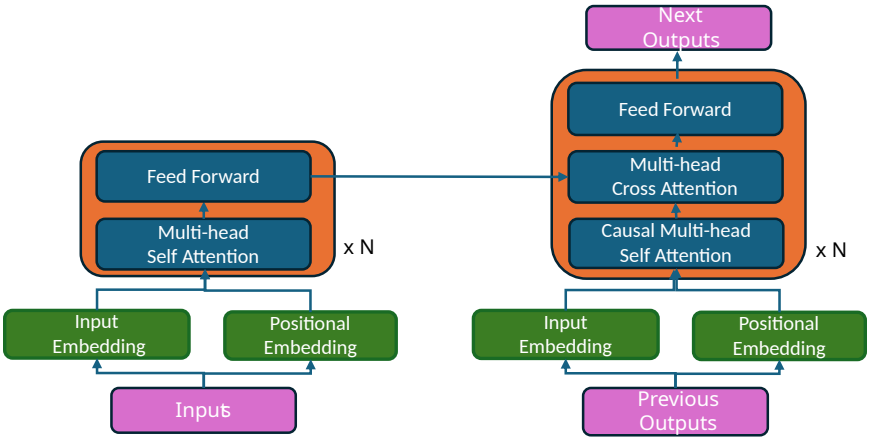


Figure 1.7: The encoder-decoder Transformer architecture.

quencing, where the goal is to predict a peptide sequence directly from a given mass spectrum. In this setting, the spectrum can be treated as the encoder input, and the peptide sequence is generated by the decoder, analogous to machine translation. The attention mechanism allows the model to align fragment ions in the spectrum with specific amino acids during decoding, leading to more accurate sequence reconstruction from complex and noisy spectra.

1.3 Thesis Outline

Mass spectrometry–based proteomics has become an indispensable tool for characterizing the proteome. However, the interpretation of MS/MS data remains computationally challenging due to several inherent limitations. For protein inference, a fundamental difficulty arises from the fact that many peptides can be shared among multiple proteins, leading to ambiguity in assigning peptides to their true source proteins. Moreover, the presence of false or uncertain peptide-spectrum matches (PSMs), especially for low-abundance peptides, introduces additional noise into the inference process. These challenges are further exacerbated by the scarcity of high-quality labeled datasets for training machine learning models, which limits the performance and generalizability of traditional inference approaches. On the other hand, de novo peptide sequencing, especially in the context of Data-Independent Acquisition (DIA), presents a different set of difficulties. DIA leads to highly multiplexed spectra, where signals from multiple coeluting peptides are superimposed, making it difficult to distinguish which fragment ions belong to which precursor.

Traditional de novo sequencing algorithms are not designed to handle this level of complexity, often resulting in reduced accuracy and incomplete sequence recovery. Furthermore, the high dimensionality and noise in DIA spectra pose significant barriers to effective interpretation. These technical barriers have limited the effectiveness of DIA in de novo contexts, despite its advantages in proteome coverage and reproducibility.

This thesis aims to address the core challenges of protein inference and DIA-based de novo sequencing through the development of novel deep learning–based computational frameworks. These methods are designed to be both data-efficient and scalable, leveraging modern neural architectures to better model the intricate structure of MS/MS data. The work is organized into three major contributions.

Chapter 2 presents GraphPI, a graph neural network–based framework that formulates the protein inference problem as a node classification task over a tripartite graph. In this graph, proteins, peptides, and PSMs are represented as distinct node types, with edges encoding both structural and statistical relationships, such as shared sequences and PSM confidence scores. To capture the heterogeneous nature of this graph, GraphPI extends the GraphSAGE architecture to support type-specific message passing and aggregation functions. In addition, GraphPI adopts a semi-supervised learning paradigm, where pseudo-labels are generated from existing protein inference tools and augmented with decoy-derived hard negatives. This allows the model to learn effectively even in the absence of large manually annotated datasets. The result is a highly generalizable and computationally efficient inference method that achieves strong performance across multiple benchmark datasets, while significantly reducing the computational overhead associated with Bayesian or combinatorial inference frameworks.

Chapter 3 focuses on DIANovo, a deep learning framework for de novo peptide sequencing from DIA data. DIANovo introduces several architectural and algorithmic innovations to tackle the multiplexed and noisy nature of DIA spectra. First, it constructs a spectrum graph in which each node represents a peak, and edges are defined based on possible mass differences between amino acids. A Transformer-based encoder, enhanced with Rotary Positional Embeddings (RoPE), processes this graph by encoding mass differences along its edges. To cope with signal interference from coeluting peptides, DIANovo includes a coelution-aware pretraining stage that trains the model to identify ion types using signals from mixed peptide sources. This pretraining stage enables the model to develop a better understanding of spectral structure, improving its downstream sequencing performance. During inference, DIANovo employs a two-stage decoding strategy: the first stage finds a high-confidence path through the spectrum graph, and the second refines this path to produce the final amino acid sequence. Empirical evaluation shows that DIANovo significantly outperforms state-of-the-art de novo sequencing tools in both amino

acid–level and peptide–level recall, particularly on narrow-window DIA datasets obtained from next-generation instruments such as the Orbitrap Astral. In addition, we investigate the comparative performance of DIA and DDA for de novo sequencing on matched samples, revealing that DIA provides superior peptide coverage only under specific acquisition configurations, such as narrower isolation windows.

Chapter 4 presents a theoretical framework for analyzing peptide identification performance under varying experimental conditions. The goal is to quantify how factors such as signal-noise profile, ion intensity distributions, and peptide length affect the likelihood of correct peptide identification. We derive estimated p-value of peptide identification by performing spectrum-peptide matching with XCorr[34] on simulated spectra. These predictions are then validated using real data acquired from various mass spectrometers and experimental parameters. Our findings not only explain empirical trends—such as why Orbitrap Astral [46] improve de novo sequencing—but also provide guidance for optimizing acquisition parameters to maximize identification yield. This analysis serves as a foundation for better experiment design and offers a principled understanding of the trade-offs involved in MS/MS data acquisition and interpretation.

Collectively, the three contributions of this thesis illustrate how deep learning can be harnessed to address the key computational bottlenecks in modern proteomics. By rethinking protein inference as a graph problem and adapting sequence modeling architectures to the structure of DIA data, we introduce robust and scalable solutions that extend the frontier of peptide and protein identification. Moreover, our theoretical analysis provides a quantitative lens through which to interpret empirical performance and improve acquisition strategies. Together, these efforts advance the state of the art in computational proteomics, enabling more accurate, efficient, and interpretable analysis of large-scale MS/MS datasets. Ultimately, this work contributes to the broader goal of making high-throughput proteomic analysis more accessible and reliable for applications in biomarker discovery, disease diagnostics, and systems biology.

Chapter 2

GraphPI: Efficient Protein Inference with Graph Neural Networks

*

Understanding the proteins in a biological sample is crucial for unraveling their functions and roles in biological systems. Protein inference, which involves identifying proteins through peptides detected in tandem mass spectrometry (MS/MS) experiments, is fundamental to proteomics. Accurate protein identification is essential for applications such as discovering biomarkers, identifying drug targets, and annotating protein functions, which are vital for advancing personalized medicine and therapeutic strategies [6] [27] [52].

In an MS/MS experiment, proteins are initially digested with some proteolytic enzyme, like trypsin, into peptides. The peptide mixture is then passed through a mass spectrometer, generating MS1 spectra. One prominent peptide, or a group of peptides, at a time is selected from the MS1 spectra and further fragmented into fragment ions, and mass spectrometer will capture the m/z values and intensities of these fragment ions, resulting in a specific mass spectrum signature for each peptide, named the MS2 spectrum. Next, the acquired MS2 spectra are matched to a peptide database to detect which peptides are present in the sample. Finally, the peptide profile is used to predict which proteins are more likely to produce the observed peptide set.

*This chapter is reproduced with permission from Journal of Proteome Research, 2024, American Chemical Society. DOI: https://doi.org/10.1021/acs.jproteome.3c00845.

The article can be accessed via ACS Article on Request, https://pubs.acs.org/articlesonrequest/AOR-GESZT64SEDG6BQKSFSTX?_gl=1*j2intf*_ga*MTg3NjgwOTE1Ni4xNzUyMTAzMTM1*_ga_XP5JV6H8Q6*czE3NTIxMDMxMzQkbzEkZzAkdDE3NTIxMDMxMzQkajYwJGwwJGgw.

However, high dynamic range of protein abundance, as well as limitations in digestion and mass spectrometry often lead to ambiguous peptide identification result [100] [109]. In addition, the existence of shared peptides and one-hit proteins further complicates the problem [100]. Shared peptides are those that can be derived from multiple (degenerate) proteins. Those peptides introduce ambiguity in associating detected peptides with their respective proteins. One-hit proteins are those that are backed by single-peptide evidence, making it difficult to confidently infer the presence of this protein. The issue is amplified as some peptides are more prone to detection, skewing protein identification toward those producing such peptides. These elements together complicate the protein inference process, creating an intricate, complex landscape for protein identification.

Various approaches have been developed to address the challenges associated with protein inference, with limited success due to the fundamental challenges mentioned above. ProteinProphet [95], an early innovator in this field, employs a heuristic-based probabilistic model to estimate protein probabilities, accounting for factors such as shared peptides and the quantity of peptides identified for each protein. PIA [128], a rule-based algorithm, conducts inference by identifying the minimal set of proteins that most accurately accounts for the observed Peptide-Spectrum Matches (PSMs).

In addition to these methods, Bayesian networks have emerged as a highly efficacious technique for protein inference. Fido [109] was the first to implement Bayesian networks in this domain, incorporating simple yet rational assumptions. Specifically, for proteins sharing peptides, the method gives preference to those with independent evidence. Concurrently, the "explain away" effect reduces the scores for proteins without distinct evidence. For instance, if two proteins share a common peptide with no other evidence, their scores should be equal due to symmetry. If one protein gains unique peptide evidence, its score should be elevated due to this new evidence. Concurrently, the score of the protein without unique evidence should be lowered, as the unique peptide evidence for the first protein effectively "explains away" the shared peptide, reducing the likelihood of the second protein being present. Epifany [100] utilizes a Bayesian network akin to Fido's, but incorporates additional priors to regularize the inference process and employs a rapid approximation inference algorithm to enhance computational efficiency. Nevertheless, these techniques are constrained by the inherent limitations of Bayesian networks, which include high computational costs in terms of time and memory, and susceptibility to prior probabilities [3] [92].

With the advent of neural network advancements, deep learning techniques have delivered promising results in biomedical research, including, but not limited to, prediction of peptide properties from tandem mass spectra [45], peak detection [150], peptide database search [26], and de novo sequencing of peptides [124] [101] [123]. However, given that deep learning methods usually require a massive amount of labeled data for training, it is

challenging to apply them in the field of protein inference. The scarcity of labeled data in this field, possibly because of the prohibitive costs of accurately annotating the proteins in a biological sample, presents a considerable barrier.

In recent years, a limited number of deep learning-based methods have been proposed to address the protein inference problem. For instance, Barista [114] trains a binary classification network using decoy proteins (in silico-generated by shuffling or reversing real sequences) as negative labels and real proteins as positive labels. Thereby the classification scores can be directly adopted as the protein scores. However, this methodology can inadvertently infer an inaccurate classification boundary, as not all assumed positive proteins are actually present in a given sample. Secondly, the strategy also runs the risk of overfitting decoy proteins, which are conventionally employed to discern truly present proteins based on a predefined FDR threshold [31]. Moreover, the neural network design of Barista [114] does not allow the feature of one protein to affect another, making it unable to take advantage of the "explain away" effect of a Bayesian network.

Contrastingly, DeepPep [60] leverages self-supervised learning, using peptide scores to circumvent the problem of protein label scarcity. Specifically, it utilizes peptide identification scores as labels and trains a model to predict these scores based on all proteins. By sequentially removing each protein from the input, the model can infer the contribution of each protein to each peptide. The final protein score is an aggregation of each protein's contribution to all potential peptides based on their identification scores. Nonetheless, this approach suffers from computational inefficiency, as it necessitates iterating over all proteins for each peptide.

Moreover, the self-supervised objective introduces a misalignment between the training and testing objectives, further diminishing its effectiveness. Therefore, despite the potential of existing deep learning methods, they remain somewhat marginalized within the protein inference field due to their subpar performance relative to Bayesian networks. For instance, DeepPep [60] underperforms by over 25% compared to the best performing Bayesian method.

In this study, we present GraphPI, a novel deep learning-based framework to address the protein inference problem. Drawing inspiration from Bayesian network techniques, we design a protein-peptide-spectrum graph structure with uniquely crafted node and edge features.

A graph consists of nodes (vertices) and edges (connections). In our case, the nodes represent proteins, peptides, and Peptide-Spectrum Matches (PSMs), forming a tripartite structure. Edges connect proteins to peptides, and peptides to PSMs, reflecting the relationships formed during the protein digestion and mass spectrometry process. This enables

us to perceive candidate proteins as interconnected entities rather than isolated individuals. The protein inference problem can be then formulated as a node classification problem with protein scores generated directly from the node classification scores. To process the protein-peptide-spectrum tripartite graph, we employ Graph Neural Networks (GNNs)[61], which learn node representations by recursively aggregating information from neighbors. However, standard GNNs are not well-suited to handle the heterogeneous nature of our graph, which consists of different types of nodes and edges. To address this issue, we develop a tailored GNN architecture based on GraphSAGE[47], designed to effectively manage this heterogeneity. To mitigate the label scarcity issue and enhance computation efficiency, our model is primarily trained on a set of large and unlabeled public protein datasets from ProteomeXchange (https://proteomecentral.proteomexchange.org/) in a semi-supervised learning setting, using pseudo-labels generated by an existing protein inference algorithm as the base model. We further refine these labels by introducing hard negative decoy protein information, allowing the model to surpass capabilities of the base model and produce improved test results. Finally, we perform self-training to further enhance our model's performance by iteratively refining labels based on their confidence scores. Figure 2.1 presents the overall pipeline of GraphPI. Contrary to alternative approaches which necessitate the execution of training or fine-tuning processes for each individual dataset, our analysis indicates that the data pertaining to peptide identification exhibits considerable normalization across all test datasets. This standardization facilitates the application of a single model, which can be trained and subsequently assessed universally, thus circumventing the issue of overfitting. Additionally, this method enhances computational efficiency by mitigating the need for repetitive training processes for disparate datasets.

To the best of our knowledge, we are the first to apply GNNs and a semi-supervised training scheme to the protein inference problem, and the experiments demonstrate that our approach achieves competitive performance across diverse test datasets. Additionally, we leverage the inherently parallelizable structure of neural networks, leading to considerably faster computations in comparison to existing methods. Moreover, our model, pre-trained on a wealth of publicly available datasets, is adept at performing inference instantaneously during real-time applications. As a result, the improvements in efficiency facilitate our model to handle protein inference tasks, seamlessly scaling to accommodate large protein datasets.

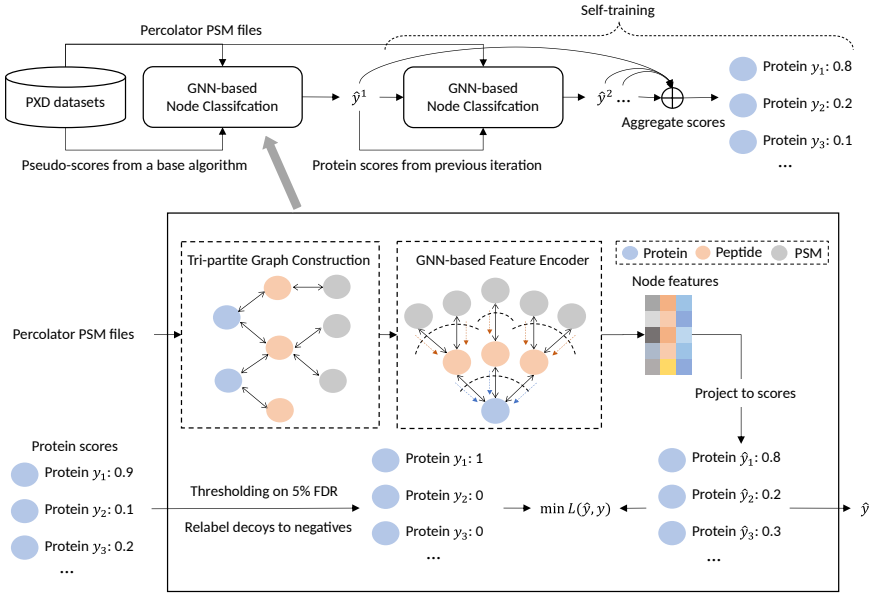


Figure 2.1: The architecture of our protein inference algorithm leveraging GNN for node classification. The process begins with the input of Percolator [65] PSM files into the GNN-based Node Classification module. This module utilizes pseudo-scores from an established protein algorithm to guide the initial classification. Within the module, a tripartite graph is constructed, linking proteins, peptides, and PSMs, as indicated in the detailed view. The algorithm then employs a self-training strategy over multiple iterations to refine the protein scores, as illustrated in the top sequence. Finally, the iterative process yields aggregated confidence scores for each protein, denoted by y_i, which are presented on the right. These resulting scores reflect the cumulative learning and adjustment from the iterative self-training, yielding a robust set of protein identifications.

2.1 Methods

2.1.1 Overview

We present a deep-learning framework, GraphPI, for protein inference that avoids the need for labeled protein datasets, and offers improved computational efficiency compared to existing methods. Our approach leverages a semi-supervised binary classification model that is trained on a set of protein datasets, which are pseudo-labeled by an existing protein inference method. Each dataset is represented as a tripartite graph and encoded using a GNN network that accommodates the heterogeneity of node and edge types. Following

training on the pseudo-labels, we implement a self-training procedure that refines the labels based on protein probability scores followed by a fresh retraining, and we iterate this process for multiple rounds. The final protein scores are calculated as an ensemble of the models from all rounds. To ensure the generalizability of our model to real-world data, we utilize a variety of experimental datasets from published biological research. The following sections delineate our methodology in three parts: 1) Construction of Tri-partite Graph: the conversion of protein datasets into tripartite graphs; 2) GNN Model Architecture: the elaboration of our model's architecture for encoding these graphs into latent representations amenable to deep learning; 3) Training: a detailed exposition of our training paradigm. This paradigm leverages the power of semi-supervised learning with an iterative self-training mechanism that begins with pseudo-labels and evolves to deliver refined, reliable inferences.

2.1.2 Construction of Tri-partite Graph

For the protein inference problem, understanding the relationship between proteins, peptides, and their associated PSMs is paramount since the identification of each protein is closely tied to its constituent peptides, which in turn is validated by the confidence of their PSMs. Given the inherent interconnectedness of these components, a graphical formulation naturally emerges as an apt choice.

By representing the data as a graph, we can capture the complex dependencies among proteins, peptides, and PSMs through GNNs, which learn robust node representations by recursively aggregating information from neighboring nodes. Then, we leverage the learned protein node representation, which includes information propagated through connected peptides and PSMs, to generate the protein score.

In the following paragraphs, the details of the graph are presented, focusing primarily on the nodes and edges, and their associated features.

Nodes in the Tri-partite Graph

The foundation of our graph lies in its nodes, which are categorized into three specific types: proteins, peptides, and PSMs. We prioritize PSMs over spectra as a node type due to their inherent adaptability in incorporating supplementary features. For instance, peptide search engines such as Percolator [65] rely on a suite of PSM-centric features to compute peptide identification scores. In this study, we primarily utilize the PSM features from Percolator. In our model, these PSM features are harnessed not just for peptide

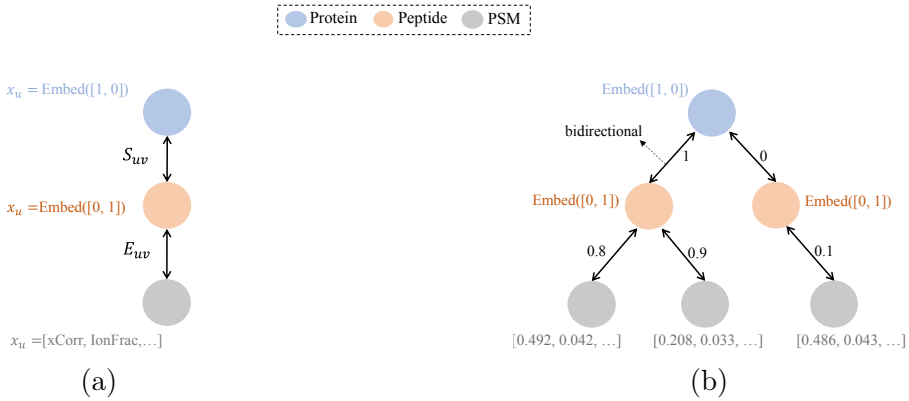


Figure 2.2: (a): The schema of the bidirectional tri-partite graph, with S_{uv} and E_{uv} denote as the edge attribute for (protein, peptide) and (peptide, PSM) node pairs respectively, and x_u is an example of the node feature vector for each type of the three nodes (one-hot embedding for protein and peptide nodes, and database search engine features for PSM nodes). (b): An illustrative example of the tripartite graph containing one protein surrounded by its peptide and PSM nodes within 2-hop.

identification but are further refined to aid in the prediction of protein scores. By doing so, we facilitate an end-to-end training process where peptide identification scores are optimized in alignment with the ultimate goal of predicting protein scores. Moreover, the volume of PSM nodes corresponding to a peptide is deemed valuable; a higher number of matches between peptides and spectra, especially with elevated identification scores, often indicates a heightened likelihood of peptide presence.

While PSM nodes come with pre-existing features from a database search engine, the nodes representing peptides and proteins lack such attributes. Nonetheless, GNNs necessitate node features. Possible strategies for feature assignment range from designating unique one-hot vectors for each node to allotting universal features (e.g., a vector with all ones) across all nodes. However, a unique one-hot vector for every node tends to restrict the graph learning to be transductive, inhibiting its generalization to datasets with varying node counts and structures. On the other hand, uniform features for peptide and protein nodes might compromise the GNN's ability to differentiate between messages from diverse node types. To circumvent these challenges while maintaining inductive capabilities, we allocate two separate one-hot vectors for protein and peptide nodes. Additionally, learnable embedding layers are applied on top of the one-hot vectors to make them dimensionally equivalent to the PSM node features. This strategy enables our model to distinguish between different types of nodes while retaining the capacity to adapt to diverse datasets.

Connecting the Dots: Edges in the Tri-partite Graph

While nodes represent distinct entities, the true essence of their relationships is captured through edges. Peptides and PSMs are interconnected in a one-to-many relationship, with each PSM associated with a unique peptide-spectra pair. To enable our model to selectively control the flow of information from a specific PSM node to a peptide node, we choose the peptide identification score as the edge weight for each (peptide, PSM) pair, which is defined as E_{uv} for a given source node u and a target node v in the remaining sections. In doing so, a lower identification score signals a weaker presence of a peptide in the sample mixture, thereby indicating that the corresponding PSM node is less reliable for computing the score of its parent protein nodes.

An edge is formed between a protein and a peptide if the peptide appears in the protein. Instead of setting all edge weights to 1, which simply indicates connectivity between peptide and protein nodes, we could incorporate certain prior knowledge into the design. To this end, we integrate a peptide-sharing feature into our graph design by creating a specialized edge attribute S_{uv} between a peptide u and a protein v, which is defined as

$$S_{uv} = \begin{cases} \frac{1}{|C|}, \text{ if } v = \arg max_k f(k) \ \forall k \in C, \text{where } C = \{k | A_{uk} = 1\} \\ 0, \text{ if } v \neq \arg max_k f(k) \end{cases} \tag{2.1}$$

Here, A is the bipartite adjacency matrix for peptide and protein, and $A_{uv} = 1$ indicates a connection between peptide u and protein v. C is the node index set of the protein nodes that connect to the peptide u, and $|C|$ represents the size of the set C. $f(\cdot)$, a score indicating if a protein is connected to many high-scoring peptide, is defined as

$$f(k) = \sum_{l \in D} (\mathbb{1}_{d_l > \epsilon}), \text{where } D = \{l | A_{lk} = 1\} \tag{2.2}$$

where d_l is the maximum peptide identification scores of peptide l (i.e., each score associates with one PSM), and $\epsilon \in [0, 1]$ is a hyper-parameter. $\mathbb{1}_{d_l > \epsilon}$ is an indicator function, which is 1 when $d_l > \epsilon$, and 0 otherwise. D is the node index set of the peptide nodes that connect to the protein k. The peptide-sharing feature discounts a peptide's relevance to its parental proteins if it is connected to multiple proteins. Elaborating, for a peptide that connects to multiple proteins, a surrogate score (i.e., $f(k)$ in Eq.(2.2)) is ascertained for each protein, gauged by the count of top-scoring peptides tethered to it. Subsequent to this, solely the edge weight for the protein with the highest surrogate score is retained, with all others being nullified. The penultimate step involves diminishing the edge weight of this top-scoring protein by a factor of $\frac{1}{|C|}$ as laid out in Eq.(2.1). The intuition here is

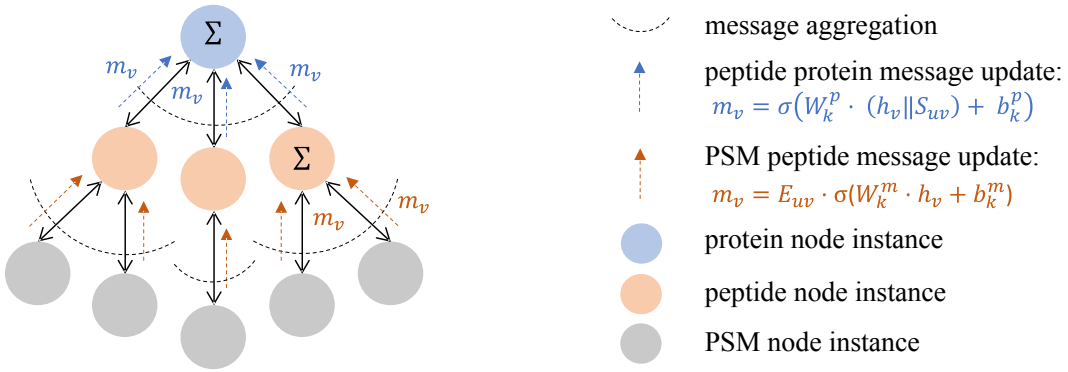


Figure 2.3: An illustrative example of the message passing operation of GNN, where the message update function of each edge type is processed differently.

that we aim to further mitigate the contribution of a peptide to its parental protein if it is not unique evidence of the protein. In practice, the threshold ϵ is set at 0.9, a decision stemming from our intention to prioritize peptides with high confidence. Note that in this feature design, although the edge weight between a peptide and a weakly connected protein is set to 0, it is only used as a feature in the subsequent model, and the edge still exists. A schematic depiction of the tri-partite graph is available in Figure 2.2a, and an example of the tripartite graph containing the related information of one protein is presented in Figure 2.2b.

2.1.3 GNN Model Architecture

Using the tripartite graph as input, we can naturally formulate the protein score generation as a protein node prediction task, utilizing GNNs as the foundational model architecture. Our design draws inspiration from GraphSAGE, a network that elevates message-passing operations through a customized aggregation function. This offers greater flexibility compared to GCNs (Graph Convolutional Neural Networks [61]), which rely solely on a linear transformation of node features and a subsequent mean aggregation from neighboring nodes. Notably, our model augments GraphSAGE by addressing the inherent heterogeneity of the tripartite graph. This refined architecture is underpinned by two core principles: 1) distinct edge types propagate messages uniquely. 2) different node types update their hidden representations differently.

Formally, given the tri-partite graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$, where the set of nodes $\mathcal{V}$ is partitioned into three disjoint subsets: $\mathcal{V}_{\text{pep}}$, $\mathcal{V}_{\text{pro}}$, and $\mathcal{V}_{\text{psm}}$, representing the nodes for peptides, pro-

teins, and PSMs, respectively. We define the message passing operation between a source node $v \in \mathcal{V}$ and a target node $u \in \mathcal{V}$ as

$$a_u^{(k)} = \sum_{v \in \mathcal{N}_u} m_v^{(k-1)} \tag{2.3}$$

$$h_u^{(k)} = \sigma(W_k \cdot (h_u^{k-1} \parallel a_u^{(k)}) + b_k) \tag{2.4}$$

where h_u^k denotes the output feature representation of node u at the k-th GNN layer ($h_u^0 = x_u$, the feature of node u), $\mathcal{N}_u$ represents the set of neighboring nodes of u, W_k and b_k are the learnable weights and biases, σ is a ReLU function, and $\parallel$ denotes a vector concatenation. The message m_v propagating from node v to u is uniquely defined for different edge pairs. For node pairs comprising peptides and proteins, i.e., $(u \in \mathcal{V}_{pro}, v \in \mathcal{V}_{pep})$ or vice versa, the message m_v at the k-th layer is computed as

$$m_v^{(k)} = \sigma(W_k^p \cdot (h_v^{(k-1)} \parallel S_{uv}) + b_k^p) \tag{2.5}$$

where S_{uv} represents the edge attribute associated with the (u, v) pair, i.e., the feature that penalizes the peptide having non-unique evidence for the target protein. W_k^p and b_k^p are the weights and biases for the peptide-protein pairs. Conversely, the message between peptides and PSMs, i.e., $(u \in \mathcal{V}_{psm}, v \in \mathcal{V}_{pep})$ or vice versa, is defined as

$$m_v^{(k)} = E_{uv} \cdot \sigma(W_k^m \cdot h_v^{(k-1)} + b_k^m) \tag{2.6}$$

where E_{uv} corresponds to the identification score of the source PSM node, W_k^m and b_k^m are the weights and biases for the peptide-PSM pairs. See Figure 2.3 for a visual illustration of the proposed GNN message passing operation.

It's noteworthy that our model employs a unique set of learnable weights and biases for different edge pair types, capturing the varied distributions across node types. For example, the number of PSM nodes serves as important prior information, as a peptide connected to more PSMs suggests a higher likelihood of occurrence. Similarly, the count of peptides linked to a protein indicates the probability of its target protein's presence. By using specific transformations for these node distributions, our model captures information intrinsic to each node type.

Moreover, we use different message functions for pairs of nodes that have different edge attributes. For the edge attribute between peptides and PSMs, which reflects the probability that a PSM originates from a given peptide, we selectively filter out PSMs that are unlikely to have originated from the peptide directly based on the edge score (shown in Eq. (2.6)). On the other hand, for the edge attribute between proteins and peptides, we treat the edge attribute as a feature rather than a filter (shown in Eq. (2.5)), as it provides relatively weak prior information. In this way, we let the neural network learn the importance of peptides in relation to a given protein primarily in a data-driven way.

2.1.4 Training

We train our model under a self-training scheme, which is a classic semi-supervised learning technique that iteratively uses the label generated by the trained model as the training label for the next round of training procedure. This approach hinges on the premise that the efficacy of a new model iteration is closely tied to the quality of the pseudo protein labels from its predecessor.

In a typical self-training setting, a model is initially trained using a small labeled dataset and subsequently applied to a large unlabeled dataset to generate pseudo labels. The same model is then trained on a combination of labeled and pseudo-labeled data, with unlabeled data iteratively added to the training set. However, in our problem setting, we lack labeled data initially, while having access to other protein inference models to generate pseudo labels for the unlabeled data. As a result, we adopt a self-training variation similar to the work [98], in which an existing benchmark model provides the initial labeled dataset.

In this study, our model undergoes training within a binary classification framework. The training loss, binary cross entropy, is mathematically defined as:

$$L = \frac{1}{|\mathcal{V}_{pro}|} \sum_{i \in \mathcal{V}_{pro}} [\hat{y}_i \log y_i + (1 - \hat{y}_i) \log (1 - y_i)] \tag{2.7}$$

In this formulation, y_i represents the predicted label, derived from the GNN output representation h_i of protein i through a trainable one-layer feedforward neural network followed by a sigmoid layer, denoted as $f_o(\cdot) : R^d \rightarrow R$. The term $\hat{y}_i$ corresponds to the actual label of protein i, and $|\mathcal{V}_{pro}|$ is the number of proteins used for training. The classification process predominantly focuses on protein nodes.

Given the typical unavailability of ground truth labels for proteins, our approach employs pseudo labels generated by thresholding the scores from a benchmark model. The threshold is determined based on the False Discovery Rate (FDR); specifically, proteins exceeding a defined FDR threshold (e.g., 0.05 in our experiments) are categorized as positive samples, with the rest deemed negative. Additionally, decoy proteins are explicitly labeled as negative, owing to their inherent absence in biological samples. After the initial training round based on the labels provided by the benchmark model, we replace the previous pseudo labels with the new ones generated by our latest model and retrain our deep learning model accordingly in subsequent self-training rounds.

In each round i (including the initial round, where the training is based on labels provided by the selected benchmark model), the self-learning procedure outputs a learned

model ϕ_i and returns a list of trained models. Analogous to ensemble learning, we perform an average aggregation of all learned models to obtain the optimal protein score. Specifically, the final protein score for a given protein x is computed as follows:

$$score(x) = \frac{1}{t}\sum_{i=1}^{t}\phi_i(x) \tag{2.8}$$

where t denotes the total number of self-training rounds. In our experiments, we set t to 10 rounds.

We employ self-training on several public datasets. Once trained, the model is directly applied to each test dataset to produce the respective protein scores, eliminating the need for re-training. In contrast to methods demanding distinct training or fine-tuning for individual datasets, our findings underscore a pronounced consistency in peptide identification data across all examined datasets. Such uniformity allows us to deploy a single, universally adaptable model, mitigating the problem of overfitting. Moreover, this approach increases computational efficiency by reducing the need for redundant training procedures for different datasets.

2.1.5 Implementation

Experimental Setting

In our experiments, we select Epifany [100] as the base model to generate the initial pseudo labels due to its relative computational efficiency compared to other models and competing performance, as demonstrated in their original paper.

We adopted the Adam optimizer with a learning rate of 0.001. The model is composed of six GNN layers, with node and edge hidden dimensions set to 100. In addition, our model is trained on a single Nvidia RTX4090 graphics card over 1000 epochs, and the parameters leading to the best validation loss are stored.

The software versions and configurations in our experiments are all listed in Table S3.

Trainging Datasets

The training datasets are public protein datasets downloaded from ProteomeXchange. To avoid over-fitting to a specific benchmark dataset used in our experiment, those datasets are selected randomly, and processed with Comet [32] search and Percolator [65] for peptide

database search, with respect to their experimental specifications. For convenience, we selected only human data, but this does not affect our generalizability.

The following are the datasets used for training: PXD004789 [64], PXD005388 [84], PXD006640 [133], PXD010319 [84], PXD022881 [12], PXD023034 [144], PXD023593 [78], PXD025701 [88], PXD026991 [38], PXD030330 [57], PXD030448 [54], PXD032035 [74], PXD032284 [75], PXD034012 [17], PXD035125 [49], PXD036171 [117], and PXD039272 [56], the links and search parameters of which are listed in Table S2.

Test Datasets

We evaluated our algorithm on the following MS/MS datasets: iPRG2016 [119], UPS2 [4], 18Mix [62], Yeast [102] and Hela-3T3 [107]. All datasets except Hela-3T3 provides us with ground truth labels for evaluating entrapment FDR, while for Hela-3T3 we evaluate based on two-species FDR. Among these datasets, iPRG2016 was specifically curated to test protein inference algorithms on proteins that share peptides. 18Mix and Yeast also contain a small portion of peptide-sharing proteins. Hela-3T3 provide us with human hela or mouse 3t3 cells, with large amount of shared peptides. Table 2.1 offers the summary statistics of these datasets, while a detailed description of these datasets are provided in the Material S1.

Dataset	**# True Proteins**	**# Contaminate Proteins**	**# Identified**
iPRG2016 A	191	1,191	179
iPRG2016 B	191	1,191	187
iPRG2016 AB	382	1000	368
Yeast	4,265	6,330	551
UPS2	48	48	23
18Mix	18	1,802	13
Hela	20,419	17,202	1335
3T3	17,202	20,419	799

Table 2.1: Number of true proteins and contaminate proteins of each test dataset, along with numbers of protein identified at 5% FDR by GraphPI.

To process the tandem MS data, we first converted and centroided the raw files with msConvert [18]. For UPS2, we do not need to generate decoy proteins since they are already provided by the fasta file. For other datasets, the provided fasta file was used to generate a decoy database through shuffling of amino acid with the OpenMS [106]

DecoyDatabase tool. Then, spectra were searched using Comet allowing 10 ppm precursor mass tolerance, 0.01 Da fragment mass tolerance, and one missed cleavage for fully tryptic peptides (for 18Mix, the precursor and fragment mass tolerance is set to 1.005 Da since it comes from a low resolution instrument). Carbamidomethylation(C) is selected as the fixed modification except Yeast (which does not undergo alkylation), and oxidation(M) the variable one. Then, we extracted additional features from the Comet search, and feed them into Percolator to obtain better peptide scores.

Benchmark Methods

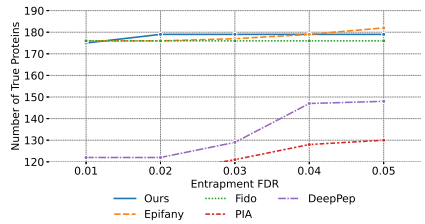
We used the following four popular and representative methods to compare against our model: Epifany [100], Fido [109], PIA [128], and DeepPep [60]. They are either based on parsimony, probability, Bayesian, or deep learning approaches, covering the majority of the approaches to this problem.

Leveraging PSM features from Percolator, each method derives a probability score for individual proteins. The probability output of each model is used to rank the proteins. Groups of identically connected proteins are treated as one single protein group during inference. When we are referring to the number of proteins, such groups contribute only one per group, instead of contributing once per protein.

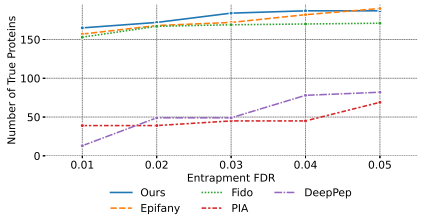
2.2 Results and Discussion

2.2.1 Comparison Study

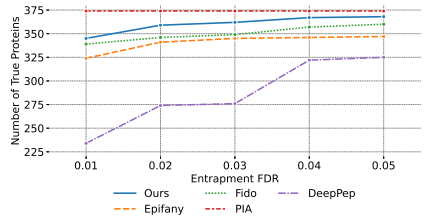
The performances of all methods are evaluated by the receiver operator characteristic (ROC) curve, which plots the number of true positive proteins (i.e. the number of ground truth proteins) as a function of entrapment FDR (the portion of contaminate proteins in the identified list, where "contaminate" proteins refers to real proteins in the database, but known not to exist in the biological sample). The curve is plotted by varying the FDR threshold above which a protein will appear in the identified list. Given that the test datasets all have ground truth attached (except Hela-3T3, where we implement a two-species approach), we can evaluate based on empirical FDR instead of decoy FDR. Since we are mostly interested in the performance of the methods when the FDR is small, we plot the curve within the entrapment FDR range of $[0.01, 0.05]$. The results are shown in Figure 2.4.



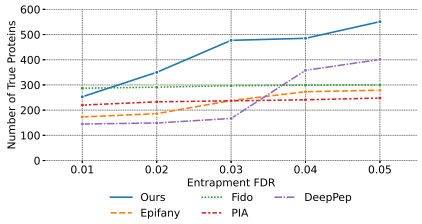
(a) iPRG2016 A



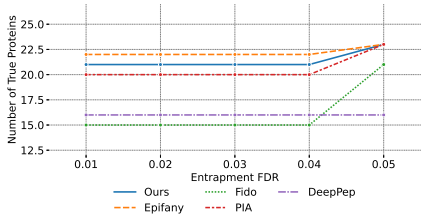
(b) iPRG2016 B



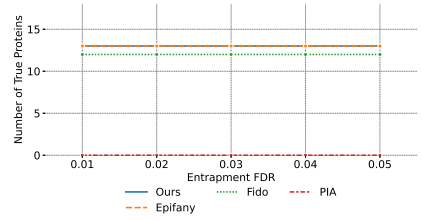
(c) iPRG2016 AB



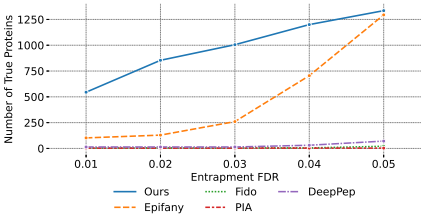
(d) Yeast



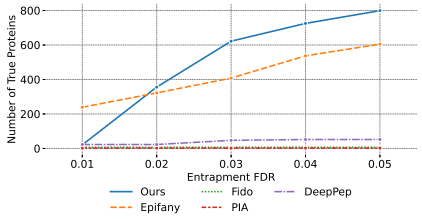
(e) UPS2



(f) 18Mix



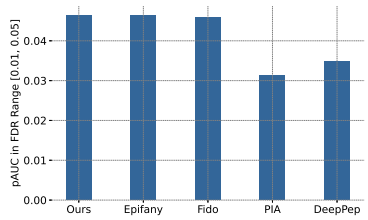
(g) Hela



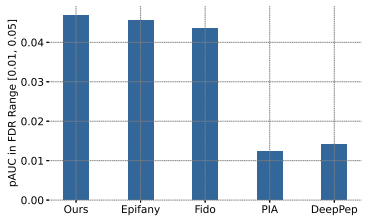
(h) 3T3

Figure 2.4: ROC curve (entrapment FDR vs. number of true proteins) of various models on the benchmark datasets: (a) iPRG2016 A, (b) iPRG2016 B, (c) iPRG2016 AB, (d) Yeast, (e) UPS2, (f) 18Mix, (g) Hela, and (h) 3T3.

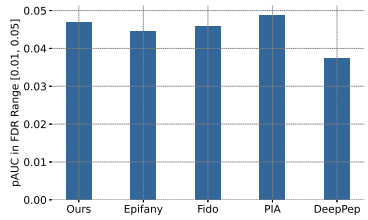
Overall, our method shows competing performance across all datasets. For both the iPRG2016 A and B datasets, GraphPI is the only one that can compete with Epifany in



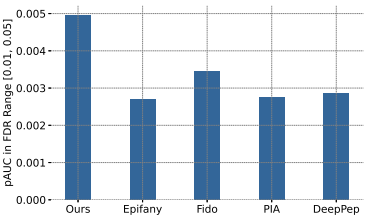
(a) iPRG2016 A



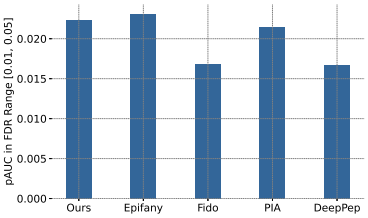
(b) iPRG2016 B



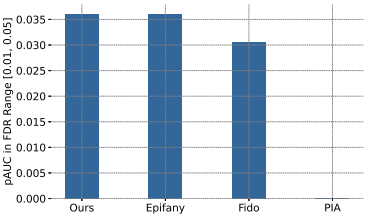
(c) iPRG2016 AB



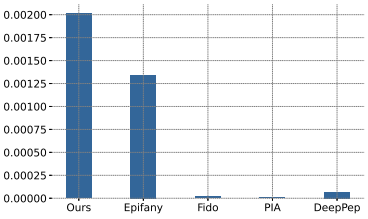
(d) Yeast



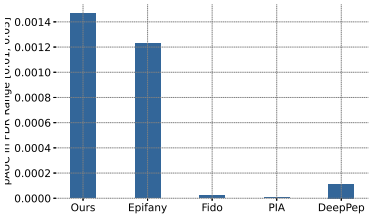
(e) UPS2



(f) 18Mix



(g) Hela



(h) 3T3

Figure 2.5: pAUC (partial AUC) score of various models on the benchmark datasets: (a) iPRG2016 A, (b) iPRG2016 B, (c) iPRG2016 AB, (d) Yeast, (e) UPS2, (f) 18Mix, (g) Hela, and (h) 3T3.

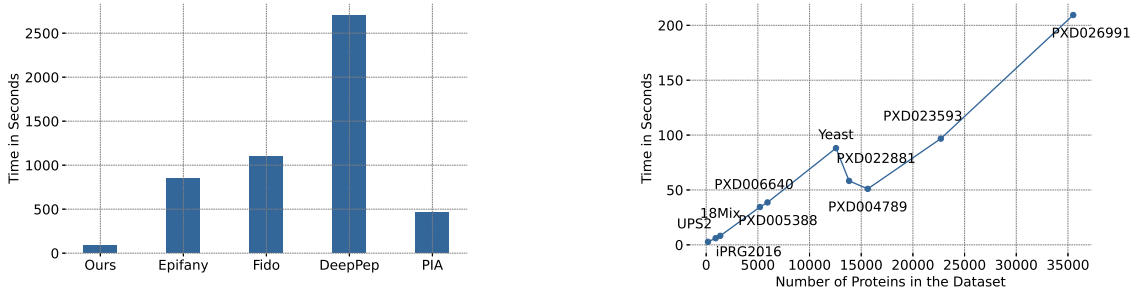
terms of identification performance. In addition, our ROC curves ascend faster than any other method in these two datasets, showing improved performance in the low FDR range. In iPRG2016 AB, GraphPI ranks the second throughout the whole FDR range, only falling behind PIA, which performs extraordinarily well in this dataset. In the Yeast dataset, our method again shows overall best performance. In the UPS2 dataset, GraphPI acquires identification performance similar to that of Epifany, surpassing Fido, PIA, and DeepPep. With the 18Mix dataset, we rank the top along with Epifany. Additionally, we tested our model on the Hela-3T3 [107] dataset, where we take the samples of 100% Hela cells and 100% 3T3 cells, and evaluate the performance based on a two-species library strategy. Our model achieves the top most performance in the comparison, only falling behind Epifany on the 3T3 dataset at 1% FDR. Moreover, only GraphPI and Epifany can achieve adequate performance in these two datasets,largely due to the high prevalence of shared peptides, which adds complexity to the analysis. It is worth noting that, while Epifany can produce state-of-the-art performance in iPRG2016 A and B, it falls short in other datasets like AB and Yeast, but our approach retains competitive and consistent performance across all datasets.

We also compared the pAUC based on FDR from 1% to 5% of test methods, summarized in Figure 2.5. pAUC computes the area under our ROC curve (between 1% and 5% FDR), relative to the perfect curve (a horizontal line with all groundtruth proteins identified at any FDR). A higher pAUC value indicates we identify more proteins within the FDR range of interest. The results are consistent with our previous analysis, with our method achieving the top performance in iPRG2016 A, B, Yeast, 18Mix, Hela, and 3T3, being second in iPRG2016 AB and UPS2.

2.2.2 Computational Efficiency

To demonstrate the computational efficiency of GraphPI, we report the runtimes in minutes on the Yeast dataset, which is the largest dataset in our test data. The runtimes are measured on an Intel Core i7 8700K processor. Every algorithm is executed with 12 threads to take advantage of parallelism. GPU acceleration is disabled during inference time to make the comparison fair.

Our method shows a significant advantage when compared to other models, especially Bayesian methods, like Fido or Epifany. On the Yeast dataset, our method takes only 88 seconds to run, while Epifany takes over 14 minutes. Partially this is due to the inherent speed advantage of neural networks over Bayesian networks. On the other hand, Epifany and Fido need to run a grid search of their parameters α, β, and γ for every test dataset, while our approach does not need such a procedure, further improving efficiency.



(a) (b)

Figure 2.6: (a) Inference time of the benchmarked methods on the Yeast dataset. (b) Scaling of inference time for our model on datasets of different sizes.

Additionally, the design of the GNN model architecture guarantees that the runtime scales linearly with respect to the dataset size (number of proteins to be evaluated), eliminating the undesired effect of higher-order scaling. Figure 2.6b shows the runtime of our method plotted against the number of proteins in the dataset, demonstrating the linear scaling.

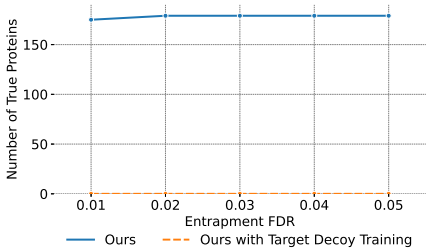
The efficiency of our method enables its application in large-scale datasets, which might be impossible for other methods due to computational constraints.

2.2.3 Extended Investigation

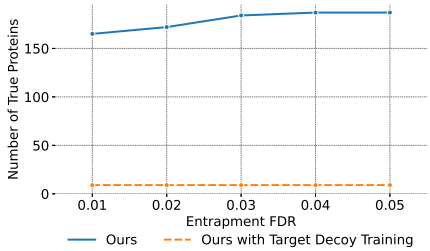
Experiments on Target-Decoy Labels

Noticing that Barista [114] is removed from the latest version of Crux [83], we create a comparable model using our GNN network as the foundational architecture. We trained this model on each test dataset, designating decoy proteins as negative samples and the remainder as positive. Adhering to Barista's documented procedures, we employed 5-fold cross-validation, averaging protein scores across the five models. We've termed this model "Ours with Target Decoy Training".

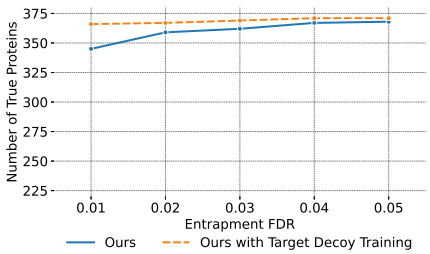
The ROC curves of our GraphPI and the target-decoy setting is plotted in Figure 2.7. It is evident that while this training setting yields favorable outcomes in the iPRG 2016 AB and UPS2 datasets, it underperforms in datasets featuring shared peptides among proteins (such as iPRG2016 A and B). This limitation stems from the fact that decoy proteins typically do not share peptides with target proteins, leading to a training gap where the



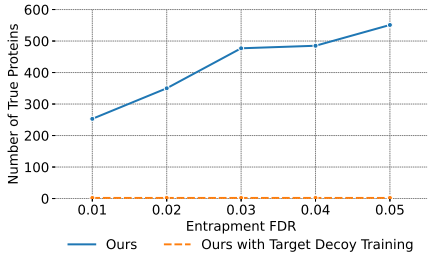
(a) iPRG2016 A



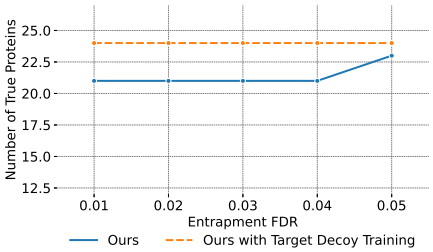
(b) iPRG2016 B



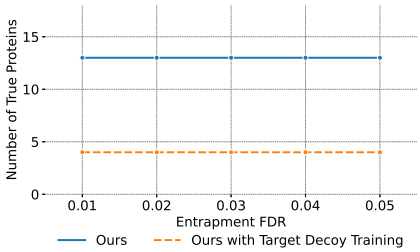
(c) iPRG2016 AB



(d) Yeast



(e) UPS2



(f) 18Mix

Figure 2.7: ROC curve (entrapment FDR vs. number of true proteins) of GraphPI and GraphPI with target-decoy training on the benchmark datasets: (a) iPRG2016 A, (b) iPRG2016 B, (c) iPRG2016 AB, (d) Yeast, (e) UPS2, and (f) 18Mix.

model lacks exposure to examples necessitating differentiation between proteins that share peptides. Our approach effectively addresses this issue by incorporating pseudo-labels from Epifany which additionally imposes penalization on degenerate proteins.

Analysis on Epifany Results

The benefit of GraphPI, a data-driven deep-learning-based approach, is its ability to learn the contribution of each peptide to its parental protein through the data distribution of a common set of datasets. In contrast, Epifany relies on prior probabilities and makes strong assumptions about the data distribution, which can lead to inaccurate protein scores. We provide two visual examples in the following paragraphs, supported by explanations that demonstrate the limitations of Epifany in dealing with certain types of proteins.

Protein prior distribution presents an important role: In situations where a peptide has multiple siblings, Epifany tends to downweight the contribution of the peptide score and place greater reliance on the prior score of the protein, particularly in cases where there is no other supporting evidence (i.e., unique peptides connecting to the protein). This can lead to inaccuracies in protein inference, especially when the prior score of the protein is high and the dataset contains many peptides with shared proteins. We demonstrate an example of this phenomenon in Figure 2.8a, where the highlighted protein is a decoy protein that should not have received a high score, yet Epifany assigns a score that is close to its protein prior score, which is around 0.7. This is attributed to the fact that the prior score of proteins is optimized by the grid search program of Epifany based on Decoy FDR, which has the risk of overfitting the distribution of decoy rather than the actually contaminated proteins.

Epifany has strong penalization on degenerate proteins: We use Figure 2.8b as an example for illustration: Specifically, the highlighted protein present in the figure is supposed to belong to the true protein group (having higher protein scores). However, Epifany assigns a relatively lower score to it regardless of the high-score peptides that connect to it. The reason behind this is that Epifany assigns a strong prior to peptides that have connections to multiple proteins, having the purpose of lowering their contribution to each parental protein. Specifically, Epifany sets the probability of a peptide that a certain number of proteins can generate to $1/N$, where N is the number of proteins that the peptide is connected to. This prior probability affects the posterior probability score of the peptide, leading to a significant penalty in the contribution of the peptide to the calculation of each parental protein score. This approach gives Epifany superior performance in iPRG2016 A and B where most true proteins do not have shared peptides connected, but the performance on other datasets are lacking.

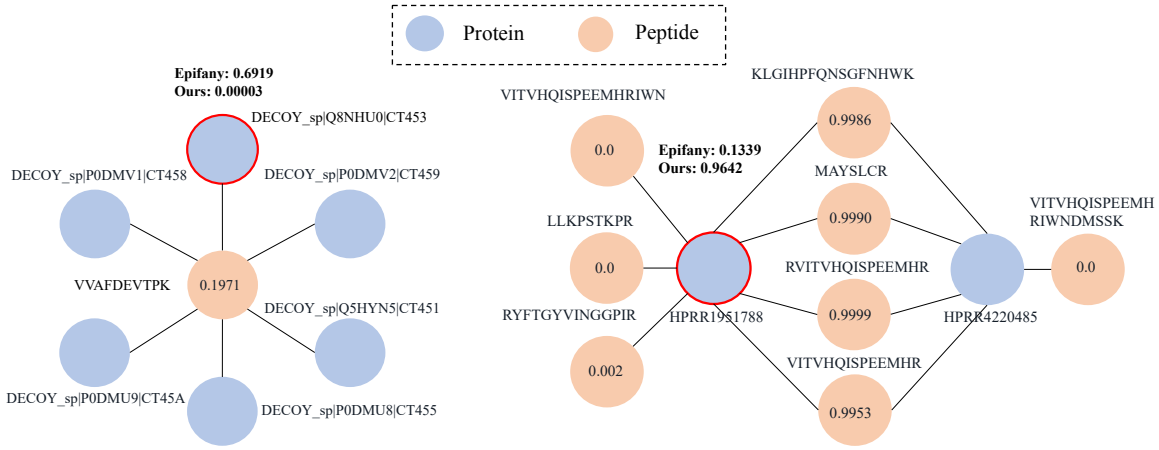


Figure 2.8: In (a)&(b), the orange circle represents a peptide with its peptide identification score inside, while the blue circle represents a protein. The scores generated by both Epifany and our model are given for the highlighted protein in each graph, respectively. Accordingly, (a) shows the bipartite graphs for the selected proteins and their connected peptides from dataset "PXD005388". The protein prior is set to 0.7 based on the grid search program of Epifany. The highlighted protein is a decoy protein, which should has a lower score. (b) shows the bipartite graph for proteins selected from the dataset "iPRG2016 AB". The highlighted protein is a true protein, which should have a higher score.

Learning beyond Epifany Scores

Some may posit that GraphPI, which is trained on Epifany's pseudo-labels, might merely mimic Epifany's outputs without surpassing its performance. We argue that the divergence between GraphPI and Epifany can be distilled into three primary distinctions.

Generalization over overfitting: As elucidated earlier in the case study, Epifany is prone to overfitting, particularly given its optimization strategy on priors anchored in Decoy FDR. In contrast, the training of GraphPI harnesses a set of public protein datasets. By building upon Epifany's results generated from diverse datasets, GraphPI inherently promotes broader generalization capabilities. As shown in both Figure 2.8a and Figure 2.8b, our method is able to produce a relatively lower score for the decoy protein and a higher score for the true protein, whereas the scores generated from Epifany is either inflated or deflated primarily due to its overfitting of prior probabilities on decoys of individual

datasets.

Discerning label incorporation: In our label-generation mechanisms, decoy proteins, known to be artificial constructs and absent from biological samples, are consistently tagged as negative, irrespective of the scoring of Epifany. This information enables GraphPI to learn patterns surpassing those identified solely by Epifany. Supporting this claim, a comparison of GraphPI's performance with and without the presence of decoy proteins is shown in Table 2.2. It is evident that the absence of decoy proteins significantly hampers our model's performance.

Setting	**True Positives**
Normal Train Data	187
Without Decoy	176
Decoys within 5% FDR as Positive Samples	183

Table 2.2: Number of true positive (TP) proteins identified by our algorithm, with decoy proteins, without decoy proteins, or with decoys within 5% FDR labeled as positive proteins. The numbers are acquired under 5% entrapment FDR, from iPRG2016 B dataset, using only pseudo-labeled training (without self-training) to demonstrate the performance difference.

Nonetheless, an argument could be made that incorporating decoy proteins merely expands the training set, which in turn leads to enhanced performance. To address this concern, we conducted another experiment, where decoy proteins within a 5% FDR threshold are labeled as positive (mimicking their incorrect identification by Epifany), and the remaining decoy proteins are labeled as negative. In this scenario, positive and negative sample labels are derived by thresholding Epifany's scores based on a 5% FDR, resulting in labels that precisely match those obtained from Epifany. The results in Table 2.2 show that while integrating decoy data as hard negatives does enhance the model's efficacy, the presence of a few decoy proteins in the positive class can impair performance relative to our standard configuration.

Tailored GNN architecture, graph design, and self-training Another aspect of GraphPI's enhanced capability is its GNN architecture, specifically tailored for a tripartite graph encompassing proteins, peptides, and PSMs. This tailored structure adeptly captures the complex relationships within proteomic data, reducing the likelihood of overfitting to

inaccuracies in Epifany's initial labels. Complementing this, GraphPI employs iterative self-training to refine these pseudo-labels from Epifany. This process allows our model to continually learn and adjust, potentially rectifying inaccuracies in the initial labels.

Accuracy of FDR Estimation

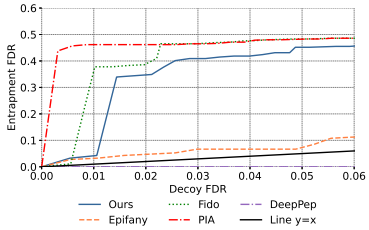
We provided an additional analysis on the relationship between entrapment FDR and decoy FDR. Generally speaking, our method provides less conservative FDR estimates than Epifany, while being more conservative than other methods in our comparison.

Figure 2.9 illustrates the relationship between decoy FDR (our estimated FDR) and entrapment FDR. Generally, Epifany provides the most conservative FDR estimate in iPRG2016 A and B, while our model lies between Epifany and Fido. In iPRG2016 AB, Yeast and 18Mix dataset, our model yields the most accurate FDR estimate. In hela, GraphPI offered the second best estimate, after DeepPep, while being after DeepPep and Epifany in 3t3. Every model gets irregular FDR estimate in UPS2, making this dataset an outlier.

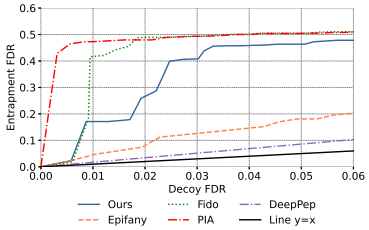
We can observe that a higher entrapment FDR at a consistent decoy FDR level doesn't necessarily imply subpar identification performance. Instead, it indicates that decoy proteins are ranked after both positive and entrapment ones. As illustrated in Figure 2.10, at a specific decoy FDR, our model identifies a greater number of proteins than Epifany in the iPRG2016 B dataset, leading to a higher entrapment FDR. Such disparities can be readily calibrated.

2.3 Conclusion

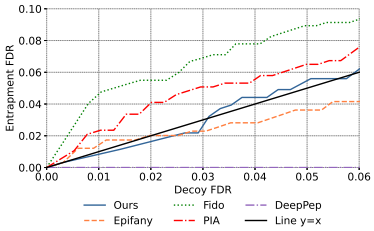
In this study, we present GraphPI, a deep-learning framework designed to tackle the protein inference problem in proteomics. Our method conceptualizes proteins, peptides, and PSMs as interconnected entities within a protein-peptide-spectrum graph. By formulating the protein inference problem as a node classification task, we designed a GNN architecture inspired by GraphSAGE to handle the heterogeneous nature of the tri-partite graph. Recognizing the hurdle of limited labeled data in proteomics, we leverage large, unlabeled public protein datasets in a semi-supervised learning setting, utilizing pseudo-labels generated by existing protein inference algorithms. We further refined these labels by incorporating hard negative decoy protein information and employing self-training to iteratively improve the performance of the model. The experimental results demonstrated



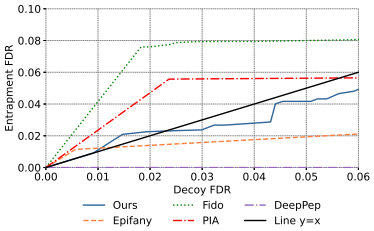
(a) iPRG2016 A



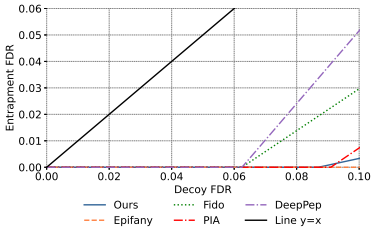
(b) iPRG2016 B



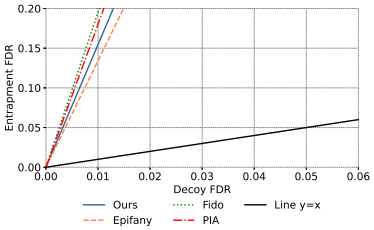
(c) iPRG2016 AB



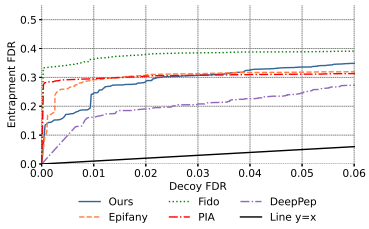
(d) Yeast



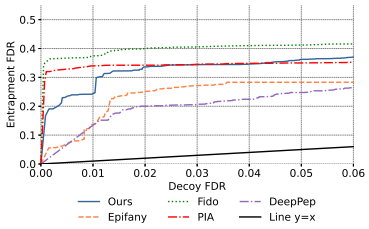
(e) UPS2



(f) 18Mix

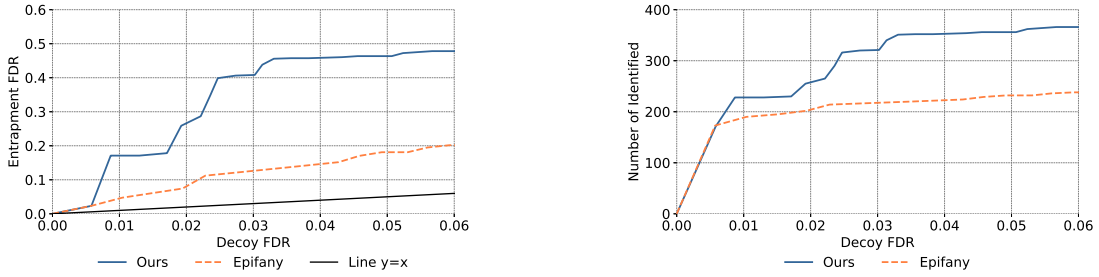


(g) Hela



(h) 3T3

Figure 2.9: Relationship between decoy FDR and entrapment FDR for different models, demonstrating the accuracy of FDR estimate. The dashed line is a straight line from (0, 0) to (1, 1), demonstrating the perfect FDR estimate.



(a) (b)

Figure 2.10: (a) shows the relationship between decoy FDR and entrapment FDR for iPRG2016 B, (b) shows the number of identified proteins under different decoy FDR values.

that our approach achieved superior performance across diverse test datasets. The use of GNNs and the semi-supervised training scheme contribute to significant improvements in protein inference accuracy. Furthermore, our method exhibited enhanced computational efficiency by leveraging the inherent parallelizability of neural networks. The universal applicability of GraphPI, trained on a common set of peptide identification data, eliminated the need for repetitive training processes on different datasets. In prospect, the realm of protein inference holds vast research potential. Future research directions could explore incorporating additional features and information, such as protein-protein interactions or post-translational modifications, to further enhance the performance of protein inference methods.

Chapter 3

Disentangling the Complex Multiplexed DIA Spectra in De Novo Peptide Sequencing

De novo peptide sequencing plays a crucial role in proteomics, enabling the identification of novel peptides, post-translational modifications, and mutations absent from existing protein databases[76][124][39][147]. This capability is crucial for personalized immunotherapy, guiding targeted treatments by identifying unique neoantigens. It is also essential for studying species with unsequenced genomes, where database searches are not feasible [37] [10] [134].

Compared to database-driven methods, de novo sequencing offers the advantage of discovering new peptide sequences independently of pre-existing data, enables capturing the full diversity of proteomes, especially in complex and dynamic biological systems [134] [7] [9]. In traditional Data-Dependent Acquisition (DDA) methods, where abundant peptides are selected to fragment[140] [87], deep learning-based approaches such as DeepNovo[124], Casanovo[142], PepNet[73] and GraphNovo[82] have significantly improved performance, making them valuable and efficient tools for biological research.

Despite their promising results, traditional DDA methods suffer from limitations such as biased sampling and inconsistent detection of low-abundance peptides, leading to incomplete proteome coverage [87] [132] [42]. Data-Independent Acquisition (DIA) [42][70] addresses these limitations by fragmenting all ionized peptides within a predefined mass range, providing a more comprehensive and unbiased snapshot of the proteome [132].

DIA data fundamentally differ from DDA data in that DIA produces highly multiplexed

spectra containing fragment ions from multiple coeluting peptides, where fragments of multiple peptides coexist in the same spectrum [105] [30]. However, these characteristics come with significant challenges for de novo peptide sequencing. DIA spectra are inherently noisier due to the simultaneous fragmentation of all peptides, complicating the distinction between signal peaks and noise. The concurrent fragmentation leads to overlapping fragmentation patterns, making it difficult to assign fragment ions to specific peptides. The mixed signals from multiple coeluting peptides increase the complexity of spectral interpretation compared to DDA, where each spectrum typically associates with a single precursor ion. Furthermore, the need to process and analyze these complex, noisier spectra increases computational demands. In addition, DIA provides chromatogram information that offers a temporal profile of peptide ions, and the similarities between these chromatograms could help trace the source of the corresponding ion.

Recent advancements such as DeepNovo-DIA [123] and Transformer-DIA [29] have harnessed the power of deep learning to address the challenges inherent to de novo sequencing in DIA, while approaches such as PepNet [73] have sought to develop unified models capable of handling both DDA and DIA data modalities. These models employ neural networks to analyze precursor and fragment ions across multiple dimensions, including mass-to-charge ratio (m/z), retention time, and intensity, thus effectively handling the complexity of highly multiplexed spectra. For instance, DeepNovo-DIA integrates Ion-CNN and Spectrum-CNN [69] with a long short-term memory (LSTM) [50] network to capture the three-dimensional structure of fragment ions and their interrelationships. Similarly, Transformer-DIA adopts a Transformer-based [131] architecture, using an encoder-decoder framework to predict peptide sequences by iteratively generating each subsequent amino acid based on prior outputs. Despite their efficacy, both models are constrained by their focus on encoding only a portion of the spectrum, potentially missing important signals related to future amino acids.

We propose DIANovo, a framework specifically designed to address the complexities of DIA data introduced by coelution. Concretely, we first set to tackle the sheer size of highly-multiplexed DIA spectrum. Our encoder encodes the spectrum graph, like GraphNovo[82]. To reduce memory consumption, we implements an automated edge generation process with rotary positional embeddings [116], which primarily focuses on learning the relative mass differences between graph nodes while the inter-node amino acid information is learned in a task-specific manner. To further reduce computational overhead, we leverage FlashAttention 2 [25], a linear-memory Transformer architecture designed for efficient computation. We also integrate dilated convolutional neural networks (CNNs) [129] [67] to directly process chromatograms, enabling efficient encoding of time series information. Time-series embeddings are fed into the Transformer encoder, where the self-attention

mechanism explicitly learns the similarities of the chromatograms, helping us to distinguish signal ions from those of coeluting peptides and noise. Moreover, we introduce a coelution-aware pretraining step that further improves model performance by incorporating coeluting peptides alongside the target peptide. This involves pretraining a model to predict ion types from coeluting peptides, with the resulting embeddings used as features in subsequent training stages. The coelution information helps the model better differentiate between signal and noise, leading to more accurate target peptide predictions. To the best of our knowledge, this is the first work to leverage coeluting peptide information in de novo sequencing, addressing a gap overlooked by previous methods.

Additionally, our study seeks to assess the realistic performance of DIA de novo sequencing. We report amino acid and peptide recalls compared with DIA-NN [26] search outcomes, considering only peptides unique to the test set. Our findings indicate that de novo peptide sequencing using DIA on older-generation instruments like the Q Exactive[86] or Fusion [136] is suboptimal, yielding relatively lower peptide recall. However, the Orbitrap Astral [46] significantly outpaces older-generation in acquisition speed, enabling the use of narrow-window data-independent acquisition (nDIA). The Astral narrow-window DIA method [46], which offers more consistent fragmentation patterns and fewer missing fragmentations, even in case of high coelution level, yields better results.

Our extensive experiments across various datasets, including older-generation and Astral data, highlight the robustness and effectiveness of our proposed method. We demonstrated that our model consistently outperformed the baselines DeepNovo-DIA and Transformer-DIA. Sensitivity-to-coelution-number analysis further reveals that our algorithm maintains stable performance even with any number of coeluting peptides. These findings underscore the potential of our method for robust and accurate de novo peptide sequencing in the challenging DIA setting.

Finally, we address a key question: can de novo sequencing in DIA mode detect more peptides than DDA mode? We present a comparison between DDA and DIA data acquired from the same biological sample, demonstrating that with older-generation mass spectrometers, DIA surpasses DDA in peptide detection when using smaller isolation windows. However, as the isolation windows widen, DIA loses this advantage and falls behind DDA. With Astral data, DIA can consistently detect more peptides than DDA because of the narrow DIA method enabled, showcasing the superior performance of next-generation mass spectrometry.

3.1 Methods

Our model architecture is an encoder-decoder Transformer operating on the spectrum graph, including spectrum graph construction, dilated CNN (convolutional neural network) layers to process the chromatograms, RoPE (rotary positinal embedding) [116] integration to encode mass differences in the graph, coelution-aware pretraining to provide better understanding of the multiplexed spectra, and the two-stage decoding, where we predict the optimal path through the spectrum graph in stage 1, and refine the path to generate the final peptide sequence with mass tag filling in stage 2, in order to alleviate the sequence memorization problem commonly associated with de novo sequencing algorithms, where the model relies too much on previous seen peptide sequences during training. The model architecture is shown in Figure 3.1.

3.1.1 Feature Extraction

In a tandem mass spectrometry experiment, peptide precursors are first analyzed in their intact form, generating first-stage scans (MS1 scans). The peptides are then fragmented, and the resulting fragments are analyzed again, producing second-stage scans (MS2 scans).

The feature construction for the encoder is designed to capture the chromatogram characteristics in DIA, in order to better distinguish signal peaks from coeluting fragment ions and noise. For each PSM (peptide-spectrum match), we collect five neighboring MS2 spectra along the retention time (RT) dimension, centered around the RT peak. This selection is crucial as it differentiates DIA data from DDA. The neighboring spectra provide a chromatogram for each observed peak in the spectrum, and the relationship among these chromatograms improves the robustness of peptide sequence predictions. [26]

For MS1 spectra, alignment with corresponding MS2 spectra is needed. To achieve this, MS1 intensity values are linearly interpolated based on their RT values to match the MS2 spectra exactly, ensuring that the features derived from MS1 and MS2 spectra are directly comparable.

During preprocessing, each spectrum is binned into fixed-size 0.01 Th intervals, enabling the construction of chromatograms for each m/z value. Once the chromatograms are generated through binning, the spectra are reconstructed as a list of (m/z, chromatogram) pairs for neural network processing.

In addition to the primary spectral data, our methodology involves the calculation of eight supplementary features to enrich the feature set used for encoding, adapted from

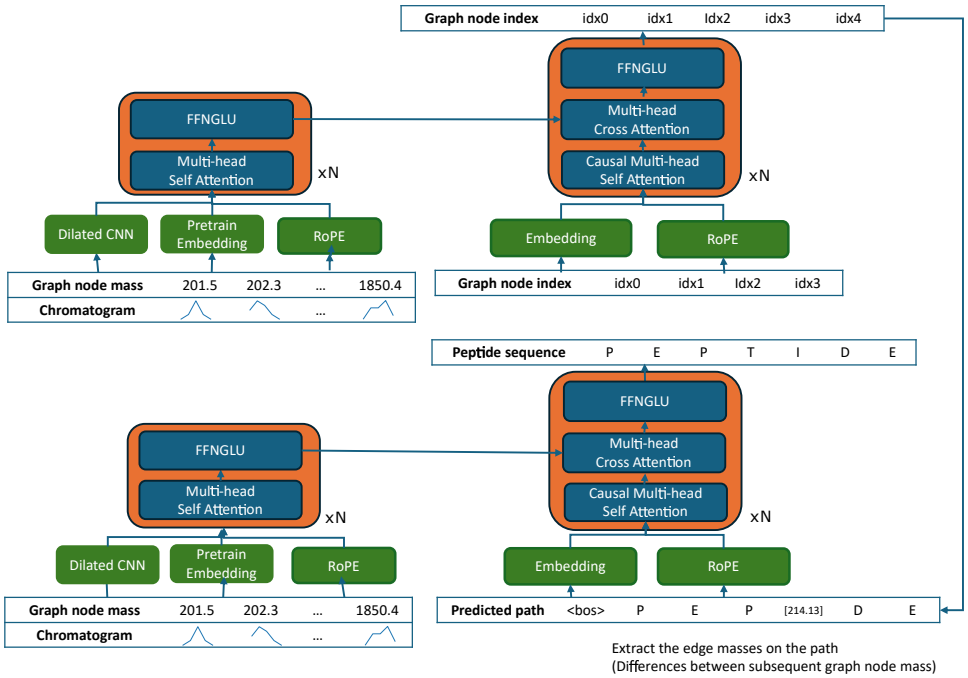


Figure 3.1: The model structure of Our entire workflow. On the top is the optimal path task, generating a series of node indices, which are transformed into the optimal path. The mass values in the optimal path are then translated to the corresponding amino acids when a single match is found. On the button is the sequence generation task. It takes the generated optimal path as input and outputs the amino acid sequence to replace mass tags. In the figure, FFNGLU refers to Feed-Forward Network with Gated Linear Units, commonly used in Transformers and deep learning architectures to enhance expressiveness and efficiency. CNN refers to Convolutional Neural Networks, and RoPE refers to Rotary Positional Embeddings.

GraphNovo[82] and Novor[79]. These features are derived from each spectrum within the ± 50 Th neighboring window and are designed to capture various statistical and intensity-based characteristics of the spectral peaks. Below is a list of the 8 additional features about the spectrum in our model.

1. Normalized Mass Over Charge: We compute the mass-to-charge ratio, normalized by a predefined upper limit of $3500Th$, to standardize this value across different spectra. Computed as $e^{\frac{m/z}{UPPER\_LIMIT}}$.

2. Relative Intensity: Each peak's intensity is expressed as a fraction of the intensity

of the most abundant peak in the spectrum, facilitating comparisons across variable signal strengths. Computed as $\frac{I}{I_{max}}$, where I_{max} is the intensity of most abundant peak in the spectrum.

3. Rank: Each peak is assigned a rank based on its abundance relative to other peaks, sorted from the most to the least abundant. Computed as $\frac{R}{N}$, where R is the target peak's rank in terms of intensity.

4. Half Rank: This metric assigns a rank to a peak after the intensities of all peaks are halved, highlighting the relative stability of peak abundances. Computed as $\frac{R_{half}}{N}$, where R_{half} is the target peak's rank with half its intensity.

5. Local Significance: Using the hyperbolic tangent function, we scale the intensity of each peak relative to the minimum intensity within the neighboring window, emphasizing peaks with significant local variance. Computed as $tanh(\frac{I}{2(I_{min}-1)})$, where I_{min} is the lowest intensity within 50 Th window.

6. Local Rank: Similar to the global rank, but restricted to peaks within the ± 50 Th m/z range, providing a localized perspective on peak significance. Computed as $\frac{R_{local}}{N_{local}}$, similar to rank.

7. Local Half Rank: This is computed like the half rank but limited to the local window, offering insights into the comparative dynamics of local peaks. Computed as $\frac{R_{half\_local}}{N_{local}}$.

8. Local Relative Intensity: We measure each peak's intensity relative to the most intense peak within the neighboring window, allowing for an assessment of local peak dominance. Computed as $\frac{I}{I_{local\_max}}$.

Post feature extraction, we compile the data into two structured tensors for each m/z value in combined spectrum:

- A $[5, 1]$ tensor representing the chromatogram data across the five neighboring MS2 spectra.

- A $[5, 8]$ tensor that encapsulates the eight computed features across the same neighboring spectra.

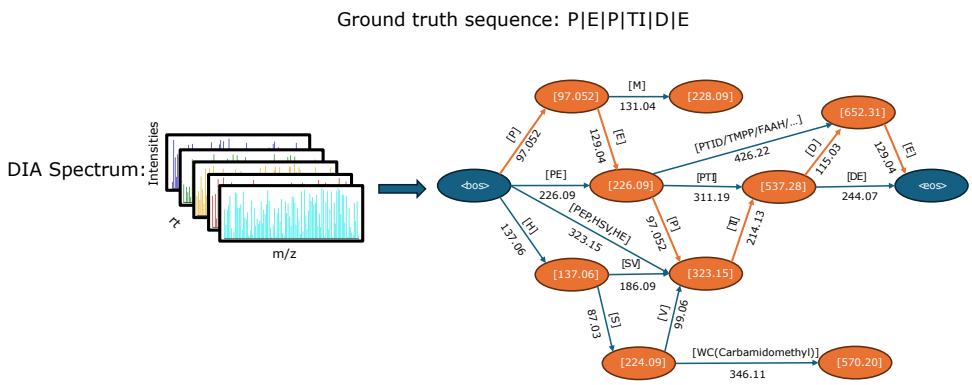


Figure 3.2: An example of a spectrum graph is shown, where the bottom value on each edge represents the mass difference between nodes, encoded by RoPE, and the top value indicates the corresponding amino acid sequence. Only a subset of nodes and edges is plotted for clarity, whereas, in a complete spectrum graph, all possible forward connections would be present.

Although these manually engineered features enhance model performance in our current training setup, it is important to acknowledge that our training dataset is relatively small. As we scale to larger datasets, such features may hinder the model's ability to learn optimal representations and potentially lead to reduced performance [68]. This is an important consideration that warrants further investigation in future work.

Finally, the spectrum graph is constructed by transforming every peak in the original spectrum from m/z to N-terminal residual masses. Each peak in the original spectrum is converted into 6 peaks in spectrum graph, taking into account 6 types of ions, including a^+, a^{2+}, b^+, b^{2+}, y^+, and y^{2+}. We operate on the N-terminal residual masses, denoted as the graph node mass, following Equation 3.1, during which C-terminal ions are converted to their N-terminal residual masses by subtracting their C-terminal residual mass from the precursor mass. Note that although we did not specifically include other common ion types like $b - H_2O$ or $y - NH_3$, their peaks are converted into graph node as well. They do not belong to the target sequence, but their graph node can still contribute to our de novo objective by self attention. We provide an example of spectrum graph in Figure 3.2.

$$m_{\text{nterm}} = \begin{cases} (m/z - m_{\text{proton}}) \times c - \sigma, & \text{if ion} \in \{a, b, c\text{-ions}\} \\ m_{\text{precursor}} - (m/z - m_{\text{proton}}) \times c + \sigma, & \text{if ion} \in \{x, y, z\text{-ions}\} \end{cases} \tag{3.1}$$

where offset σ depends on the ion type, computed as Table 3.1; c is the charge of ion; m_{proton} is the mass of proton and $m_{precursor}$ is the precursor mass [82].

Ion Type	Neutral Offset
a	[N] - CHO
b	[H] - H
y	[C] + H

Table 3.1: Ion offsets for ion types we used. [N] is the molecular mass of the neutral N-terminal group; [C] is the molecular mass of the neutral C-terminal group. C,H,O are the mass of the carbon atom, hydrogen atom and oxygen atom individually. [82]

3.1.2 Time-Series Encoder and Spectrum Encoder

In our approach, we employ a dilated convolutional neural network (CNN) to effectively encode the time-series data derived from the chromatogram and the additional spectral features.

Dilated CNNs [72] are a variant of traditional convolutional networks. The primary distinction of dilated convolutions lies in their ability to expand the receptive field without losing resolution or coverage, achieved by inserting gaps (known as dilation) between each element in the convolution kernel. This allows the network to encode time-dependent information in chromatograms while keeping the number of parameters and memory/computational complexity relatively low.

For the chromatogram and its associated spectral features, dilated CNNs are ideally suited. Our model processes inputs structured as $[5, 1]$ or $[5, 8]$ tensors, where '5' represents the number of neighboring spectra considered, '1' denotes the chromatogram, and '8' reflects the additional computed features. By applying dilated convolutions, the model can integrate both local and broader contextual information across these inputs. The outputs of the dilated CNN are encoded into tensors of size $[hidden\_size]$, where '$hidden\_size$' indicates the dimension of the feature space. This encoding captures information about the temporal dynamics of a specific m/z value.

The time-series embeddings are input directly into the Transformer spectrum encoder, where the self-attention mechanism learns to identify similarities in the time-series data. Ideally, ions from the same peptide will exhibit similar chromatograms, whereas ions from coeluting peptides will display different patterns. By explicitly modeling these similarities, the model's ability to differentiate between ions from coeluting peptides is improved.

In addition, we adopt the FlashAttention 2 [25] implementation of the attention function, allowing us to process very large DIA spectra, with high computation and memory efficiency.

3.1.3 RoPE Integration

GraphNovo employs the Graphformer[139] graph neural network to encode the spectrum graph, capturing comprehensive information about the spectrum. However, applying this approach to DIA data is impractical due to the large size of DIA spectra, which results from high levels of coelution. To address this challenge, we replace the memory-intensive node and path encoders used in Graphformer with Rotary Position Embedding (RoPE). This step is also neccessary for the adoption of Flash Attention 2, since it does not allow us to operate directly on the attention matrix, which is required by Graphformer.

RoPE is a position encoding technique that introduces rotational invariance by representing positional information in a circular format[116]. In our approach, RoPE encodes the mass differences between graph nodes, allowing the model to learn meaningful mass relationships automatically. This effectively transforms the spectrum graph into a fully-connected graph where edges represent mass differences, enabling the model to identify the most relevant mass differences for the task without the computational overhead of traditional node and path encoders. The adoption of RoPE is advantageous, as it captures the critical information conveyed by the mass difference between pairs of graph nodes, effectively representing the cumulative mass of the amino acids that lie between them. It also suggests that adopting RoPE to encode the edges in the graph structure is a viable approach, achieving better memory and computational efficiency. It is important to note that RoPE is a fixed positional embedding method, similar to sinusoidal embeddings. Its key advantage lies in its ability to directly encode mass differences within the spectrum graph. To gain deeper insight into why RoPE yields stronger performance, it would be valuable to conduct vector similarity analysis or examine the key-value attention patterns.

3.1.4 Two Stage Decoder

We adopt a two-stage decoding process, similar to GraphNovo[82] to alleviate the sequence memorization issue during training, where the decoder simply remembers seen training sequences, and fail to generalize on unseen sequences. This approach includes:

- Stage One: Optimal Path Prediction

  In the first stage, the model predicts the optimal path through the spectrum graph. This involves identifying the most probable sequence of nodes (representing graph node mass values) that form a potential peptide sequence. The graph is constructed such that each node represents a possible peptide fragment, and edges denote feasible transitions based on mass differences.

- Stage Two: Sequence Filling

  In the second stage, the optimal path is refined to generate the final peptide sequence. This stage involves filling in the mass tags along the predicted path with corresponding amino acids, ensuring the sequence adheres to known peptide fragmentation patterns.

Reader can refer to Figure 1a of Mao et al. (2023)[82] for a detailed visualization of the two-stage decoding process.

3.1.5 Coelution-aware Pretraining

A key difference between DDA and DIA data is that, in DDA, since precurosr selection is performed, there is typically a small number of peptides in the MS2 spectrum. Meanwhile in DIA, since we uniformly select a precursor range to perform fragmentation, there is a significant amount of coelution, i.e., one spectrum consisting of fragment ions of multiple coeluting peptides. In our study, coeluing peptides are identified by searching through the database search result, and include those whose RT (retention time) overlaps with the target de novo peptide, and their precursor m/z falls in the same isolation window as the target peptide.

In a normal de novo sequencing algorithm, we typically only consider the fragment ions of the de novo target peptide. For instance, in the optimal path task of our program, we predict only the graph nodes (n-terminal masses of peaks) belonging to the de novo target peptide, and the next step we replace these graph nodes with amino acids, resulting in predicted peptide sequence.

However, if we incorporate the information about the fragment ions of other coeluting peptides, we might be able to provide the model broader information about the spectrum, improving de novo performance. More specifically, in traditional de novo algorithms, the fragment ions of other coeluting peptides are treated as noises, but we can give them labels for the model to learn. With such information, the model can make less mistakes distinguishing noise peaks from signal peaks (of the target de novo peptide), since the model knows some noise peak is probably a fragment ion of other coeluting peptide, thus less likely predict it as a signal peak.

To achieve this, we introduce coelution-aware pretraining to our algorithm. We adopt the same Transformer encoder in Section 3.1.1, without the spectrum graph conversion, i.e. we keep all the original m/z values as the input to the Transformer. After layers of self-attention, we obtain the embedding for each m/z value in the spectrum. These embeddings

Ion Type
1a
1b
2a
2b
1a-NH_3
1a-H_2O
1b-NH_3
1b-H_2O
1y
2y
1y-NH_3
1y-H_2O

Table 3.2: Types of Ions Labeled during coelution-aware pretraining. For precursor charge ≤ 2, ions with charge 1 are considered. For precursor charge > 2, ions with charge 1 or 2 are considered.

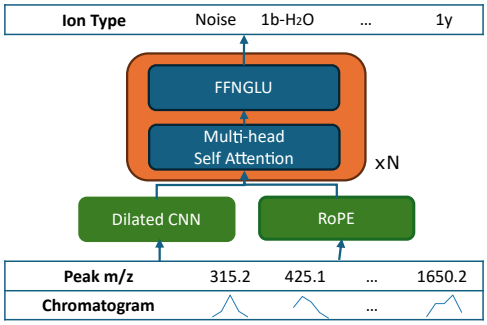


Figure 3.3: Pretrain model architecture.

are trained under the ion type loss, a cross entropy loss, representing the type of fragment ions. 12 types of ions are considered (see Table 3.2), plus noise. These labels corresponds to the fragment ion types of all coeluting peptides, not only the target peptide.

The workflow for the pretraining model is illustrated in Figure 3.3. After pretraining, the trained embeddings are fed to downstream optimal path and sequence generation models as features for the Transformer encoder.

3.1.6 Precursor Feature Detection

For the detection of precursor features from liquid chromatography-mass spectrometry (LC-MS) maps, we utilized the same standard set of precursor information as employed by DeepNovo-DIA [123]. In our experiments, we implemented the detection results from DIA-NN [26] during the main experiment, and DIAUmpire [125] during the comparison with Cascadia; however, these can be substituted with outputs from other existing peak detection algorithms, such as those referenced in Zhang et al. (2012) [145], Taynova et al. (2016) [127] or Tsou et al. (2015) [125]. The outcome of this detection step is a list of precursor features, each comprising the following essential information: feature ID, precursor mass-to-charge ratio (m/z), charge state, retention-time center, and scans across the retention-time range.

Furthermore, given the m/z and retention-time range of a feature, we collected all tandem mass spectrometry (MS/MS) spectra that fell within the feature's retention-time range and whose DIA m/z windows encompassed the feature's m/z value. More specifically, our precursor information includes the following data:

- Feature ID: an unique identifier given to each precursor.
- Precursor m/z: the mass-to-charge ratio of the precursor ion.
- Precursor Charge: the charge state of the precursor ion.
- RT_Mean: the mean of the retention-time range.
- Sequence: this column remains empty during de novo sequencing; in training mode, it contains the peptide sequences identified by the in-house database search for training purposes.
- Scans: a list of all MS/MS spectra associated with the feature as described above.

3.1.7 Model Implementation Details

Our model is trained on a Lion optimizer [19] with learning rate 2×10^{-6}, over 3 epochs. The model includes 4 layers both on encoder and decoder side, with 1,024 hidden size and 8 attention heads.

For older-generation data, training and validation sets are Pain, PXD019777, and PXD003179, with 680,947/254,467/1,206,052 PSMs respectively. For Astral data, they

Dataset	Link
Hela [120]	Chorus Project, number 1105
Pain [89]	ftp://PASS00706:YP9554a@ftp.peptideatlas.org/
PXD003179 [126]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD003179
PXD026600 [44]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD026600
PXD019777 [58]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD019777
PXD046386 [46]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD046386
PXD046453 [46]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD046453
PXD046444 [46]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD046444
PXD046283 [46]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD046283
PXD046471 [46]	https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD046471

Table 3.3: Links to each dataset

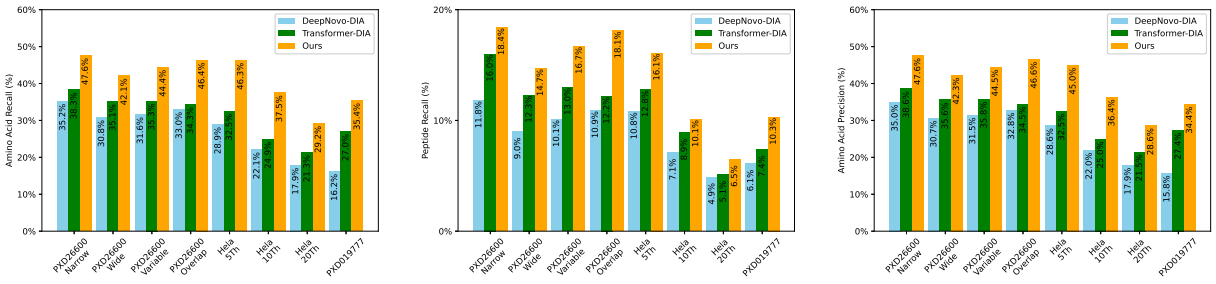
are PXD046386, PXD046453, and PXD046444, with 1,896,662, 1,086,577, and 2,136,215 PSMs respectively, these PSMs are chosen from the DIA-NN search result with 1% FDR. For the test set, we ensured that no testing sequences appeared in the training set.

The links to all datasets utilized in this project is listed in Table 3.3

3.2 Results

Our experiments comprehensively evaluated the performance and robustness of our de novo peptide sequencing algorithm across various datasets and conditions. We found that older-generation mass spectrometers exhibit relatively lower peptide recall, while Astral data showed better results due to improved fragment ion coverage. Our model consistently outperformed the baselines DeepNovo-DIA and Transformer-DIA model across multiple datasets, highlighting its superior capability in complex DIA settings. We also include a comparison with a recent state-of-the-art model, Cascadia, demonstrating the superior performance of our approach.

Furthermore, in the comparison of DDA and DIA on the Hela [120] (older-generation) and PXD046453 [46] Hek293T (Astral) dataset, DIA demonstrated advantage in peptide detection when isolation window is small, with our de novo sequencing algorithm identifying additional peptides that DDA missed. These findings collectively underscore the robustness, versatility, and superior performance of DIA in diverse and challenging proteomics scenarios. These findings provide valuable insights into the proper scenarios for applying DIA de novo sequencing and highlight DIA's capability to extend identification beyond DDA. Finally, we provide an ablation study justifying our model design choices.



(a) Amino acid recall (b) Peptide recall (c) Amino acid precision

Figure 3.4: Amino acid recall (a), peptide recall (b), and amino acid precision (c) of our method vs DeepNovo-DIA and Transformer-DIA, training sequences excluded from test set, on various older-generation datasets, measured over 10,000 randomly selected peptide precursors per dataset.

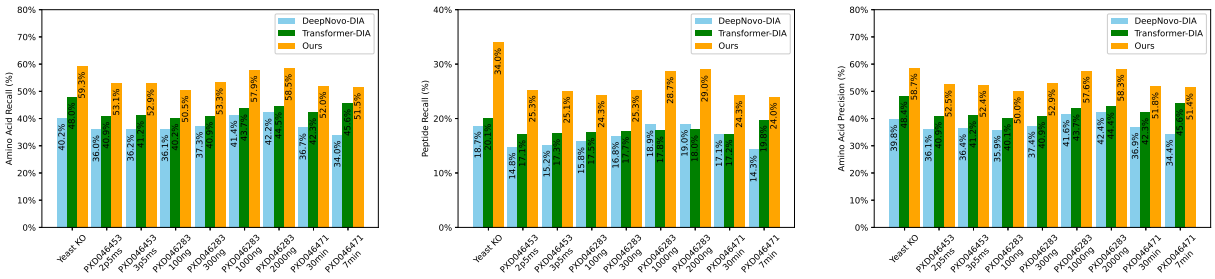
3.2.1 De Novo Performance on Older-Generation Data

We compared our model's performance with the baselines across several datasets [58] [44] [120] in Figure 3.4, where peptide recall is defined as the proportion of peptides whose entire sequences exactly match the database search results, without any mismatches or errors. On most datasets, our model demonstrated significantly better performance than the baselines, however the results are relatively lower compared to DDA levels across the board, indicating that there is a significant gap between database and de novo identification performance in older-generation mass spectrometers. Our amino acid recall is on average 60% higher than DeepNovo-DIA, while peptide recall being 53% higher. Comparing to Transformer DIA, our amino acid recall and peptide recall is 32% and 24% higher respectively.

Although our method shows greater ability beyond the baselines, the results indicate that further refinements are needed to achieve satisfactory accuracy in de novo peptide sequencing in older-generation instruments.

3.2.2 Performance on Orbitrap Astral Data

Our experiments on Astral data demonstrate superior de novo performance compared to Orbitrap data. Our findings indicate that Astral data outperforms older-generation data in de novo peptide sequencing. This superior performance can be attributed to better fragment ion coverage in the Astral data, resulting in fewer missing fragments during the



(a) Amino acid recall (b) Peptide recall (c) Amino acid precision

Figure 3.5: Amino acid recall (a), peptide recall (b), and amino acid precision (c) of our method vs baselines on various Astral datasets.

sequencing process. The performance differences are visually represented in the Figure 3.5, comparing our model's performance against the baseline across the different datasets.

The performance of our model on the Astral datasets shows a marked improvement over the baselines. On average we can expect a 43%/60% increase of amino acid / peptide recall with our methods compared to DeepNovo-DIA, or a 27%/47% improvement compared to Transformer-DIA. Furthermore, Astral data delivers a significant boost to de novo performance, affirming the effectiveness of DIA de novo sequencing in such system.

3.2.3 Sensitivity to Coelution Number

In this section, we investigate how our algorithm responds to varying levels of coelution. The coelution number is defined as the number of peptides that coelute with a target peptide, specifically when their precursor m/z values fall within the same precursor isolation window and their retention times overlap.

To visualize the relationship between coelution number and de novo sequencing performance, we generated a plot (shown in Supporting Figure 3.6) that demonstrates our algorithm's performance across different levels of coelution. The plot illustrates that our algorithm maintains consistent performance irrespective of the coelution number, indicating its robustness in handling complex spectral data.

The ability of our method to maintain performance in the presence of numerous coeluting peptides underscores its effectiveness in dealing with the inherent complexity of DIA data. This resilience is critical for practical applications in proteomics, where accurate peptide detection amidst complex mixtures is essential.

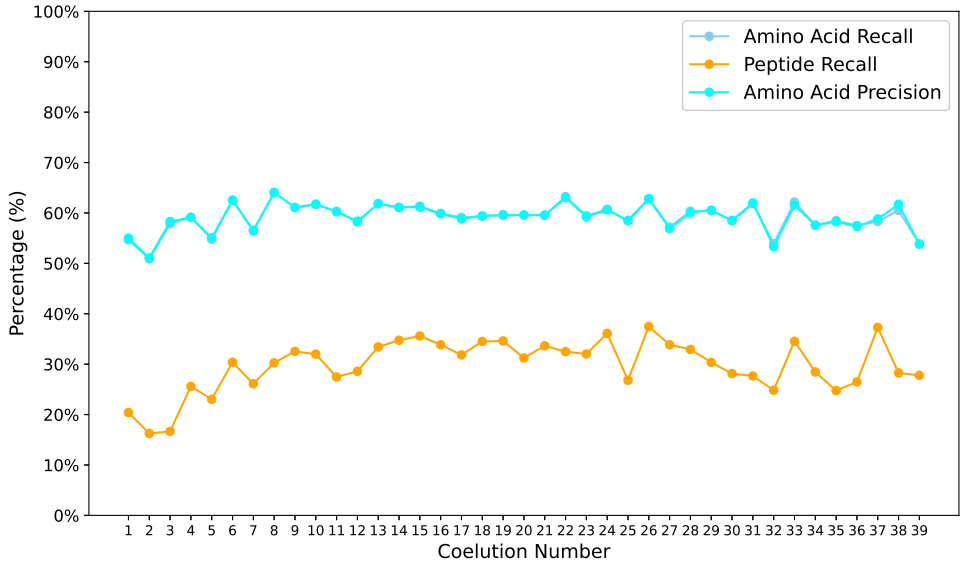


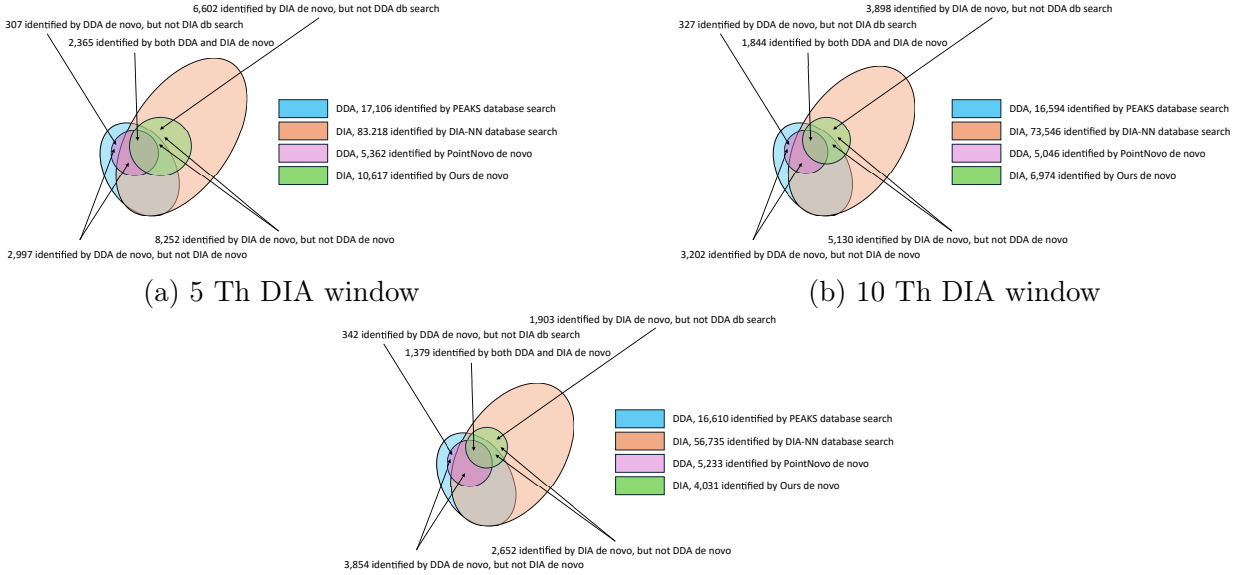
Figure 3.6: Performance of our method vs baselines, on the Yeast KO Dataset, on peptides with varying coelution number.

3.2.4 Comparison of Peptide Detection by DDA and DIA

In this experiment, we conducted a detailed comparison of the number of peptides detected by de novo sequencing using both Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA) modes, utilizing both older-generation mass spectrometers and the newer Orbitrap Astral model. The goal was to understand how these acquisition strategies perform across different isolation windows, particularly in the context of de novo peptide identification.

For older-generation instruments, we adopted the Hela dataset [120], analyzing peptides identified from the same biological sample in the DDA and DIA modes with varying DIA isolation windows (5 Th, 10 Th, and 20 Th). In the DDA experiments, we performed a database search using the PEAKS[145] DB search engine and employed PointNovo [101] for de novo peptide sequencing. For DIA, we utilized the DIA-NN[26] search engine for database search and applied our de novo sequencing methods.

The results (shown in Figure 3.7) reveal a clear relationship between the isolation window size and the efficacy of peptide detection. When the isolation window is narrow (5 Th), DIA de novo sequencing significantly outperforms DDA, identifying almost twice as many peptides. This suggests that in older-generation mass spectrometers, smaller

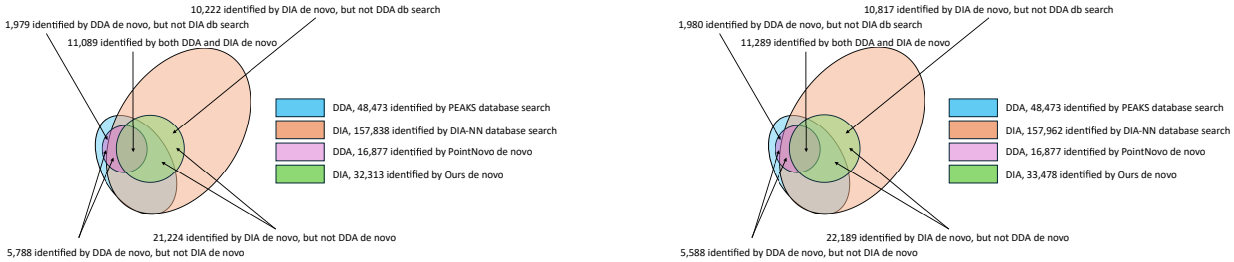


(a) 5 Th DIA window

(b) 10 Th DIA window

(c) 20 Th DIA window

Figure 3.7: Venn diagram, comparison of peptide identification under DDA or DIA mode, with Orbitrap Q Exactive (older-generation), where blue and orange circles refer to number of peptides identified in database search, while pink and green circles refer to number of peptides identified in de novo mode, under DDA and DIA respectively.



(a) 2.5ms DIA Data

(b) 3.5ms DIA Data

Figure 3.8: Venn diagram, comparison of peptide identification Under DDA or DIA mode, with Orbitrap Astral.

isolation windows in DIA mode allow for more precise detection, enhancing the depth of peptide identification. However, as the isolation window increases to 10 Th and 20 Th, the performance of DIA begins to diminish in both de novo sequencing and database search, losing its advantage over DDA. At a 20 Th isolation window, while DIA database search still

detects more peptides than DDA, de novo sequencing lags behind. The additional peptides identified in database search mode cannot be accurately recovered by de novo sequencing. This decline illustrates that larger isolation windows introduce higher-level of coelution, leading to unwanted noise and overlapping ion signals, reducing the accuracy and efficiency of peptide detection. Therefore, increasing the isolation window in older-generation mass spectrometers appears to be a suboptimal approach for DIA experiments.

In contrast, the performance of the Orbitrap Astral mass spectrometer, using the PXD046453 [46] HEK293T dataset, presents a more consistent scenario (shown in Figure 3.8) due to the enabled narrow window mode. This dataset provides both DDA data and DIA data acquired at different cycle times (2.5 ms and 3.5 ms), both using a narrow 2 Th isolation window. The Astral data demonstrates that DIA consistently outperforms DDA, regardless of the cycle time. The DDA de novo sequencing identifies approximately 17,000 peptides, whereas the DIA de novo sequencing detects nearly 32,000 peptides, showing a considerable increase over DDA. The additional 15,000 peptides identified by DIA de novo are likely missed by DDA due to biased sampling. This highlights the key advantage of DIA over DDA—its comprehensive and unbiased sampling of peptides, which is further enhanced by the speed and sensitivity of the Orbitrap Astral. In this case, the smaller isolation window and improved performance of Astral's DIA mode enable the identification of a broader range of peptides, overcoming the limitations observed with DDA.

Overall, the results of both datasets underscore the advantages of using DIA de novo sequencing with narrow isolation windows, particularly when coupled with the advanced capabilities of next-generation instruments like the Orbitrap Astral. This mode not only boosts peptide detection rates but also minimizes the sampling bias inherent in DDA, providing a more comprehensive view of the proteome.

3.2.5 Comparison with Cascadia [108]

To evaluate the performance of our approach in a practical setting, we conduct a comparative analysis against Cascadia, a recently introduced state-of-the-art model. For this purpose, we selected a representative raw file from the PXD046386 Yeast Knockout (KO) dataset. This dataset provides a relevant benchmark for assessing model effectiveness in proteomics data analysis. Both our method and Cascadia were applied to this identical data source under equivalent preprocessing and parameter settings to ensure a fair comparison. The resulting performance differences are summarized and visualized in Figure 3.9.

Our model demonstrates a 21% increase in peptide identifications compared to Casca-

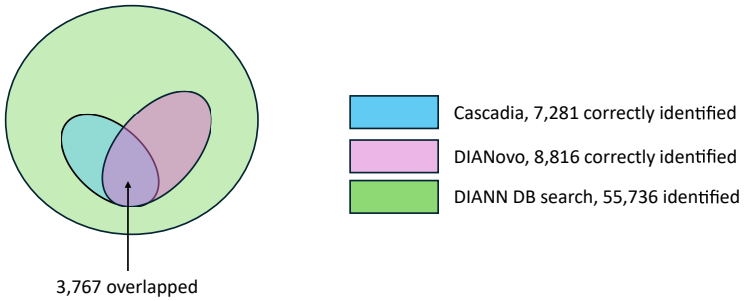


Figure 3.9: Venn diagram for Cascadia comparison.

dia. Notably, the overlap in identified peptides between the two models is relatively limited, indicating that each model captures distinct subsets of the data. This suggests that our approach exhibits different identification characteristics, potentially offering complementary insights to those provided by Cascadia.

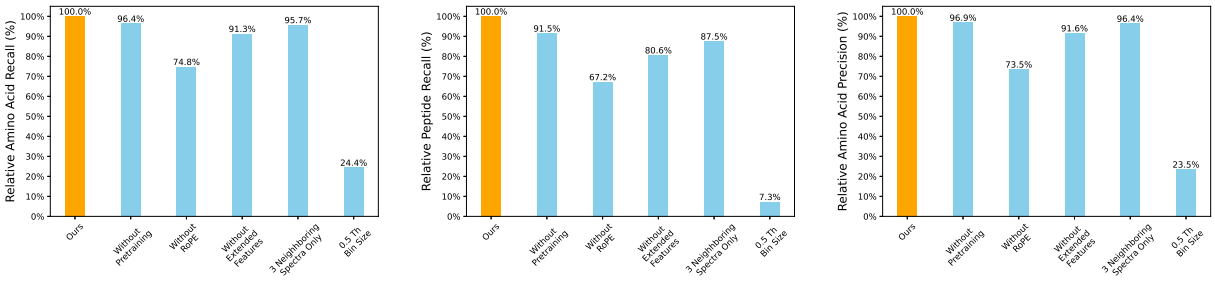
3.2.6 Ablation Study

In this ablation study, illustrated in Figure 3.10, we evaluate the impact of different configurations on our model using the Yeast KO dataset. The results highlight that coelution-aware pretraining plays a helpful role in model performance, as its removal reduces peptide recall by 8.5% compared to DIANovo. More notably, the removal of Rotary Positional Encoding (RoPE) leads to a severe performance drop, with peptide recall plunging to 67.2% of the original. This decline suggests that absolute positional embeddings alone are insufficient for encoding graph node masses effectively. Additionally, excluding extended peak features or limiting the input to only three neighboring spectra results in moderate performance losses, demonstrating the importance of both extended features and broader spectral context. Finally, binning spectra into 0.5 Th intervals causes a significant drop in performance due to reduced spectral resolution, which increases ambiguity in peptide identification.

3.2.7 Visualization of Learned Embeddings

In this section, we present a visualization of the learned peak embeddings to better understand the features captured by our pretrained model.

We generate a t-SNE [130] plot from the learned peak embeddings, and observe that the model implicitly captures the relationship between coeluting precursors and their cor-



(a) Relative amino acid recall (b) Relative peptide recall (c) Relative aa precision

Figure 3.10: Relative amino acid recall (a), peptide recall (b), and amino acid precision (c) of our method vs different configurations, compared to our method, on the Yeast KO dataset.

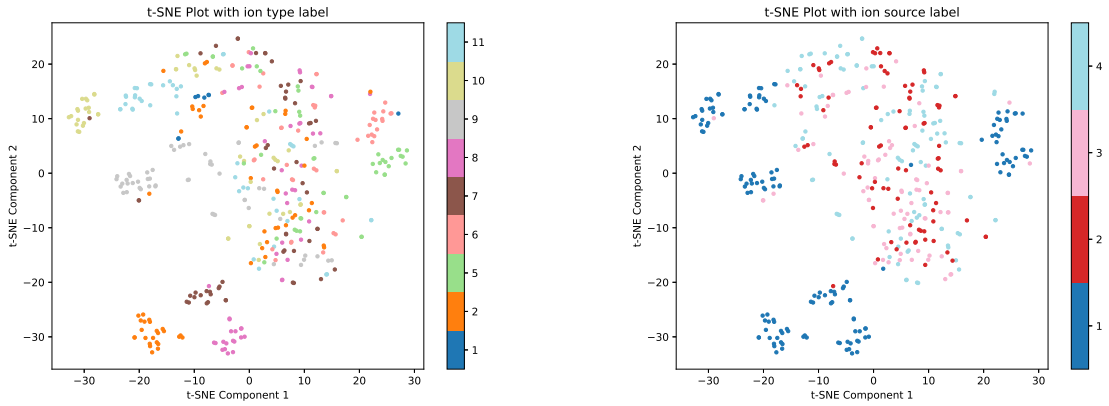
responding fragment ions, even when trained solely with the ion type loss. The t-SNE plots (shown in Figure 3.11, 3.12, and 3.13) further reveal that peak embeddings are organized not only by fragment ion type (ion type label) but also by their source peptide (ion source label). For instance, in Figure 3.11, a dark blue cluster on the right side of the graph represents peaks originating from peptide 1 in the ion source plot. Within this cluster, the peaks are further subdivided into distinct colors (pink and green) in the ion type plot, corresponding to different fragment ion types. We provide the t-SNE plots from three different peptides to illustrate this effect.

These results suggest that the model inherently learns to associate each fragment ion with its coeluting peptide, highlighting its ability to extract meaningful structural relationships from the data.

3.3 Discussion

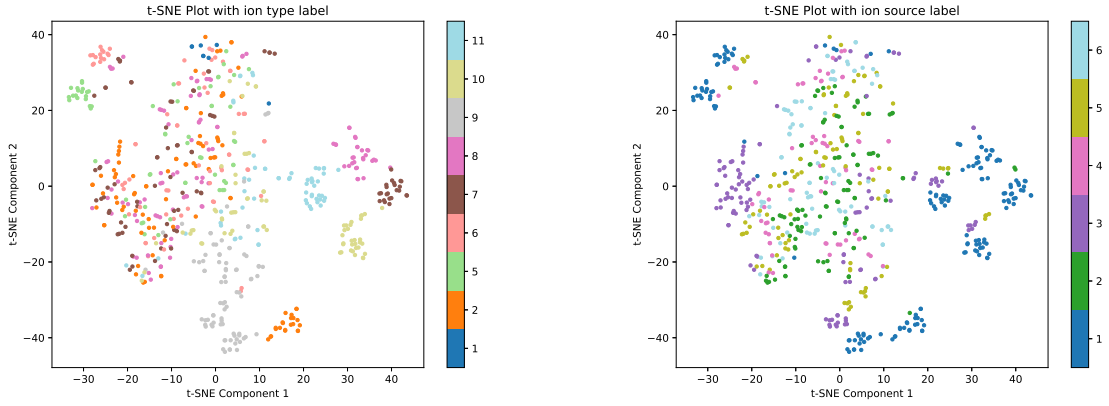
This study introduces a robust and highly accurate method for de novo peptide sequencing within the DIA setting, offering substantial improvements over existing approaches. By incorporating dilated convolutional neural networks, FlashTtention-2, RoPE, as well as coelution-aware pretraining, our model is adept at handling the complex, highly-multiplexed and noisy spectral data inherent to DIA experiments. The use of these advanced techniques allows our approach to effectively capture intricate spectral patterns, enabling reliable peptide sequence predictions even in challenging conditions.

Our comprehensive experiments demonstrate the superior performance of the proposed



(a) Ion Type Label (b) Ion Source Label

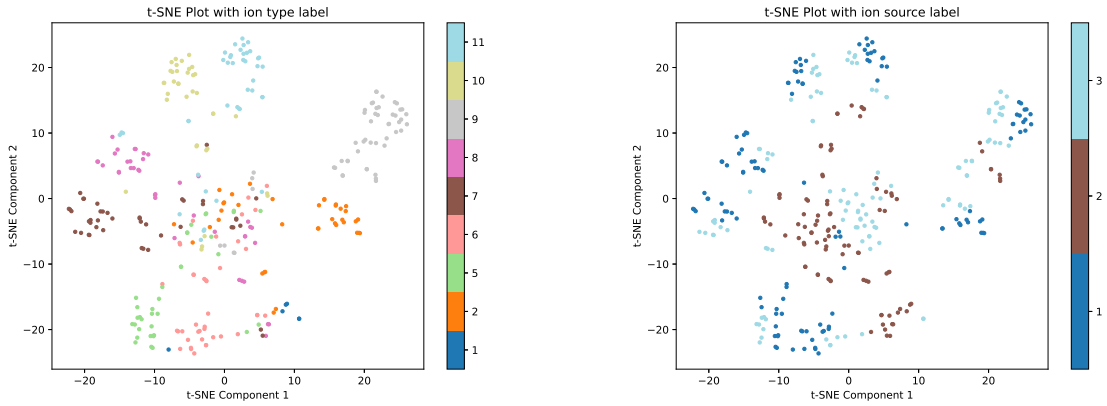
Figure 3.11: tSNE plot for ion type label (a) and ion source label (b) of peptide DHGEGGIIVGSALENK2. The color for ion type label refers to the fragment ion type of each peak, while the color for ion source label refers to which coeluting peptide a peak comes from. Noise peaks (which do not belong to any coeluting peptide) are excluded.



(a) Ion Type Label (b) Ion Source Label

Figure 3.12: tSNE plot for ion type label (a) and ion source label (b) of peptide GYWGTNLGQPHSLATK2.

method, particularly when applied to data from the Orbitrap Astral mass spectrometer. The Astral's enhanced fragment ion coverage and consistent fragmentation patterns pro-



(a) Ion Type Label

(b) Ion Source Label

Figure 3.13: tSNE plot for ion type label (a) and ion source label (b) of peptide EYLPEMAASYSHPK2.

vide a solid foundation for our de novo sequencing approach, leading to a marked increase in peptide identification accuracy, even in case of high coelution level. In contrast, older-generation mass spectrometers, such as those used in older DDA and DIA experiments, reveal the limitations of existing de novo sequencing methodologies, especially when larger isolation windows are used. Our DDA vs. DIA comparison study highlights these challenges, showing that while DIA outperforms DDA with narrower isolation windows, it begins to lose this advantage as the window size increases. This issue is particularly evident with older-generation instruments, where DIA performance in both database search and de novo sequencing diminishes at larger window sizes. However, the Orbitrap Astral consistently demonstrates DIA's superiority over DDA, facilitated by the narrow isolation windows enabled by this instrument, underscoring the pivotal role that modern mass spectrometers play in advancing peptide sequencing capabilities.

Furthermore, the sensitivity analysis reinforces the robustness of our method, showing stable performance even in scenarios with high numbers of coeluting peptides—common in proteomics workflows. Maintaining high peptide recall under these challenging conditions is essential for practical applications in the field, where complex mixtures and high levels of coelution are routine. Our method's ability to handle these complexities underscores its versatility and reliability across a range of proteomic datasets. In addition, our ablation study verifies the validity and neccessity of our design.

Looking forward, one particularly promising application of our model lies in the field of

immunopeptidomics. Immunopeptides—short peptides presented by major histocompatibility complex (MHC) molecules—are central to immune recognition and are key targets in cancer immunotherapy, infectious disease surveillance, and autoimmune disease research [9, 20]. Their identification poses unique challenges due to their high sequence diversity, non-enzymatic cleavage patterns, and typically low abundance. Conventional database search methods often miss these peptides, particularly when they are not represented in reference proteomes. De novo sequencing combined with data-independent acquisition (DIA) offers an attractive solution, enabling unbiased discovery of MHC-bound peptides without reliance on predefined databases. Our model, designed to operate robustly on coeluted and noisy DIA spectra, is ideally suited for this task. Its capacity to generalize across diverse sequence contexts and maintain high recall in complex spectra makes it a promising tool for advancing immunopeptidomic profiling, particularly when applied to high-resolution DIA datasets generated from instruments like the Orbitrap Astral.

In conclusion, our study represents an advancement in de novo peptide sequencing within the DIA framework. By integrating advanced computational techniques to tackle the coeluted nature of DIA spectra, and leveraging the advanced capabilities of next-generation mass spectrometers like the Orbitrap Astral, we offer a robust and accurate solution for analyzing complex proteomic data. The findings of this study not only demonstrate the limitations of older-generation instruments in de novo peptide sequencing but also highlight the potential of DIA to provide more comprehensive and precise peptide identification. This work paves the way for future innovations in de novo sequencing, offering new opportunities for discovering novel peptides and advancing our understanding of proteomic diversity.

3.4 Conclusion

In this project, we present a novel and highly effective method for de novo peptide sequencing in the DIA setting, offering significant improvements over traditional approaches. A key focus of our study is the comparison between DDA and DIA, where we demonstrate that while DIA outperforms DDA with narrow isolation windows, its advantage diminishes with wider windows in older-generation instruments. However, with the Orbitrap Astral, DIA consistently surpasses DDA, highlighting the importance of next-generation mass spectrometers in improving peptide identification.

Our method, specifically designed to tackle the highly mutliplexed DIA data, along with leveraging advanced instrumentation, provides robust performance in challenging proteomic environments. These findings emphasize the value of using DIA with mod-

ern technologies for comprehensive peptide detection, paving the way for more reliable and effective de novo sequencing methods in proteomics.

Chapter 4

Theoretical Analysis on Peptide Identification Performance

To better understand the challenges in de novo peptide sequencing and peptide database search, we developed a theoretical framework that explains how the balance between signal enhancement and noise accumulation in different acquisition methods affects de novo peptide sequencing performance. Specifically, we analyze the signal and noise characteristics of DDA, older-generation DIA, and Astral DIA, and their impact on the efficacy of peptide matching algorithms like XCorr.[34] [55] Our model reveals that older-generation DIA introduces substantial noise that outweighs the marginal gain in signal peaks—diminishing de novo sequencing performance—while Astral DIA provides a disproportionate increase in signal peaks despite higher noise levels. This increase effectively enhances peptide identification by improving the confidence of peptide-spectrum matches. This theoretical insight underscores the critical importance of optimizing signal-to-noise ratios in mass spectrometry data to improve de novo sequencing or database searching outcomes.

4.1 Methods

We implemented simulations to generate synthetic mass spectra reflecting the characteristic signal and noise levels associated with DDA, older-generation DIA, and Astral DIA methods.

- Addition of Signal and Noise Peaks

Dataset	Coeluting Number	# Signal Peaks	# Noise Peaks	Median Peptide Length	Median Noise Intensity	Isolation Window	Mass Spectrometer
Hela 5 Th	18.50	72	1,965	11	0.44	5	Q Exactive
Hela 10 Th	28.03	80	1,965	12	0.52	10	Q Exactive
Hela 20 Th	36.20	91	2,438	13	0.56	20	Q Exactive
PXD026600 Narrow	5.73	59	967	13	0.64	8	Fusion
PXD026600 Wide	7.24	65	1,485	14	0.69	15	Fusion
PXD026600 Overlap	5.79	53	1,130	13	0.65	8	Fusion
PXD026600 Variable	5.95	60	1,275	13	0.67	8-15	Fusion
PXD019777	12.34	91	3,057	13	0.73	24.25	Q Exactive
Yeast KO	19.49	152	9,404	12	0.34	3	Astral
PXD046453 2p5ms	10.15	152	12,284	12	0.21	2	Astral
PXD046453 3p5ms	11.99	155	14,069	12	0.18	2	Astral
PXD046283 100ng	4.66	82	2,790	12	0.54	2	Astral
PXD046283 300ng	5.01	105	4,007	12	0.49	2	Astral
PXD046283 1000ng	5.41	131	6,114	12	0.40	2	Astral
PXD046283 2000ng	6.04	137	6,960	12	0.37	2	Astral
PXD046471 30min	6.97	136	6,526	12	0.34	2	Astral
PXD046471 7min	14.04	130	6,359	12	0.51	4	Astral

Table 4.1: Experimental Parameters for test datasets. Coeluting number refers to how many coeluting peptides one peptide has on average, number of signal or noise peaks refers to the median number of peaks of the neighboring five spectrums which are fragment ions or the target peptide or not, median noise intensity refers to the ratio between median intensity for noise peaks and signal peaks, and isolation window has unit Th.

Signal peaks were uniformly sampled from the theoretical peaks of the target peptide to represent peptide evidence. To simulate experimental conditions, random noise peaks were added to each spectrum, with the quantity adjusted to reflect typical noise levels for each acquisition method, introduced by different level of coelution. Specifically, for DDA, we included 50 signal peaks and 500 noise peaks; for older-generation DIA, 60 signal peaks and 1,750 noise peaks were added; and for Astral DIA, 150 signal peaks and 9,000 noise peaks were included. Noise peaks were assigned random m/z values with uniform intensities to mimic realistic spectral conditions.

For real world test dataset, we substitute the experimental parameters with real data statistics, and take into consideration the median peptide length, as well as median signal and noise intensities. Table 4.1 provides the key parameters for each test dataset; these parameters are utilized in our theoretical analysis of peptide recall.

- Calculation of XCorr Scores

  We calculated the XCorr scores for the simulated spectra by cross-correlating the ex-

perimental spectra with the theoretical spectra of the target peptides. This involved binning the spectra, normalizing the intensities, and computing the dot product between the experimental and theoretical spectra after shifting the theoretical spectrum over a range of lags to find the maximum correlation.

Computation of p-values: We employed the dynamic programming approach described by Noble, et al. [55] to compute the p-values associated with the XCorr scores. This method estimates the distribution of XCorr scores under the null hypothesis and determines the statistical significance of the observed XCorr scores.

- Adjustment for Multiple Hypotheses

  Theoretically, within de novo peptide sequencing, the p-value is inversely related to peptide recall. This inverse relationship arises because, when computing the Sidak-corrected p-value [112], we must adjust for a significantly larger number of peptide hypotheses in de novo sequencing than in database searches. In database searches, only the peptides existing in the database are considered, limiting the number of comparisons. In contrast, de novo sequencing must account for all possible peptide sequences, dramatically increasing the number of hypotheses and necessitating a stricter correction for multiple testing. Consequently, the p-value threshold becomes more stringent in de novo sequencing, potentially reducing peptide recall. The relationship between peptide recall and p-value is given by the following equation:

$$Peptide\ Recall = \mathbb{E}\left[\frac{\#Denovo\ Success}{\#Database\ Success}\right] \approx \frac{(1-p)^{\#\ De\ Novo\ Peptides}}{(1-p)^{\#\ Database\ Peptides}} \tag{4.1}$$

  where $\#\ De\ Novo\ Success$ and $\#\ Database\ Success$ are the number of successfully identified peptides by de novo mode/ database search mode, p is the p value, and $\#\ De\ Novo\ Peptides$ and $\#\ Database\ Peptides$ refer to the number of peptides with the same precursor mass to compare for de novo or database search, with $\#\ De\ Novo\ Peptides \gg \#\ Database\ Peptides$

  Additionally, in peptide database search, the true positive is also inversely related to the p-value, as evidenced by the probability of successful database identification computed as

$$DB\ Identification\ Prob = \mathbb{E}\left[\frac{\#Database\ Success}{\#Total}\right] \approx (1-p)^{\#\ Database\ Peptides} \tag{4.2}$$

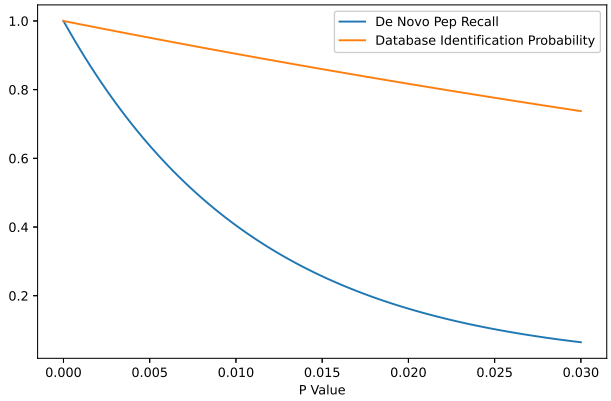


Figure 4.1: De novo peptide recall, and database identification probability, vs different p-values.

Although both de novo peptide recall and database-based identification probability exhibit an inverse relationship with the corresponding p-values, the impact of increasing p-value is more pronounced in the de novo approach. This trend, as illustrated in Figure 4.1, suggests that de novo identification methods are more sensitive to statistical confidence levels compared to database search strategies, due to their reliance on sequence inference without prior knowledge.

This observation helps to explain why de novo peptide sequencing experiences a notable decline in performance as p-values increase—often a consequence of degraded signal-to-noise profiles, such as those observed in older-generation DIA data. In contrast, database search methods demonstrate greater robustness under such conditions, maintaining relatively stable identification performance despite the diminished data quality.

4.2 Results

We present a theoretical framework to explain the performance variations observed in de novo peptide sequencing across different mass spectrometry acquisition methods: Data-Dependent Acquisition (DDA), older-generation Data-Independent Acquisition (DIA), and Astral DIA. Our analysis focuses on how the balance between signal and noise in the generated spectra affects the efficacy of peptide matching algorithms like XCorr in both

database searches and de novo sequencing.

4.2.1 Signal and Noise Characteristics in Different Acquisition Methods

We define signal peaks as the fragment ions originating from the target peptide. All other peaks are considered noise, which typically includes fragment ions from coeluting peptides, immonium ions, and instrumental noise. In DDA spectra, we typically observe approximately 50 signal peaks corresponding to fragment ions of the peptide of interest, along with about 500 noise peaks. Transitioning from DDA to older-generation DIA results in a slight increase in signal peaks to around 60 but introduces a substantial increase in noise peaks to approximately 1,750, due to spectra being highly-multiplexed. We argue that this significant escalation in noise outweighs the marginal gain in signal peaks, leading to diminished de novo sequencing performance in older-generation DIA compared to DDA.

When moving from older-generation DIA to Astral DIA, one might expect the scenario to revert to that of DDA due to the employment of narrower isolation windows. Contrary to this expectation, Astral DIA produces a considerable increase in signal peaks to approximately 150 and an even larger surge in noise peaks to around 9,000, while demonstrating high level of coelution in spite of the narrow isolation window. Additionally, Astral DIA data exhibits lower noise peak intensity relative to signal peaks compared to older-generation equipment. Despite the higher noise levels, Astral DIA demonstrates significantly better de novo sequencing performance than older-generation DIA.

4.2.2 Theoretical Model Explaining the Observed Performance

To elucidate this phenomenon, we propose a theoretical model based on the difficulty of matching peptides against noise using the widely adopted peptide matching algorithm, XCorr. Our central argument is that a higher XCorr value for a de novo peptide indicates a more straightforward and confident match of the peptide sequence in both database searches and de novo sequencing algorithms.

In database searches, the XCorr score [34] quantifies the similarity between the experimental spectrum and theoretical spectra generated from candidate peptides in the database. A higher XCorr score reflects a better match, leading to increased confidence in peptide identification. This is because the probability of a high-scoring match occurring by chance decreases exponentially with increasing score, enhancing the specificity of the identification.

For de novo sequencing, the advantage of a higher XCorr score stems from the intrinsic similarities between database search algorithms and de novo methods. De novo algorithms are heuristic approaches designed to reconstruct peptide sequences directly from spectra without relying on a predefined database. Although de novo methods may employ advanced scoring functions tailored for sequence reconstruction, the XCorr score serves as a valuable lower-bound approximation of their performance. A higher XCorr score implies that the spectrum contains clear and informative fragment ion peaks that can be effectively utilized by de novo algorithms to deduce the peptide sequence.

4.2.3 Simulation of Experimental Scenarios

To validate our theory that signal-noise profile plays a crucial rule in de novo performance, we conducted simulations under the three experimental conditions: older-generation DDA, older-generation DIA, and Astral DIA. We generated synthetic spectra that reflect the characteristic signal and noise levels associated with each of the three methods. Additionally, we created synthetic spectra that replicate the signal and noise profiles of each test dataset, accounting for peptide length and noise intensity, thereby simulating real experimental conditions. Utilizing the approach proposed by Noble et al. (2012)[55], we computed the p-value for a single peptide-spectrum match based on dynamic programming and XCorr scoring. This method allows for the estimation of the statistical significance of the observed XCorr scores, providing insight into the confidence of peptide identifications under varying signal-to-noise conditions.

4.2.4 Relationship Between p-value and Peptide Recall

In theory, the p-value should exhibit an inverse correlation with peptide recall, as discussed in detail in Section 4.1. This theoretical relationship is supported by our simulation results using real test datasets, as shown in Figure 4.2. The figure clearly demonstrates an inverse relationship between peptide recall and p-value, indicating that lower p-values are associated with higher recall rates. Furthermore, the results reveal a clear separation between older-generation DIA and Astral DIA data. Older-generation DIA data predominantly occupies the upper-left quadrant, characterized by lower peptide recall and higher p-values. In contrast, Astral DIA data is concentrated in the lower-right quadrant, reflecting its higher peptide recall and lower p-values. This distinction highlights the superior performance of Astral DIA, which can be attributed to its improved signal-to-noise profile and lower noise intensity, resulting in enhanced data quality and spectral characteristics.

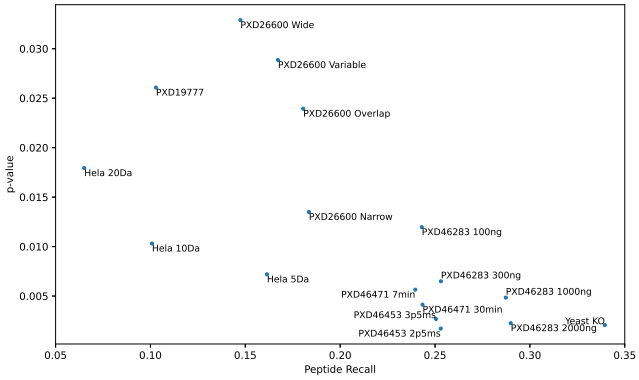


Figure 4.2: Simulated p-values vs peptide recall for test datasets, with a Pearson correlation coefficient of -0.68.

This finding highlights the utility of the simulated p-value as a reliable indicator of de novo sequencing performance. By effectively capturing the underlying relationship between statistical significance and identification accuracy, the p-value provides valuable insights into the quality of peptide identifications. These results further validate the robustness of our simulations and their alignment with real experimental conditions, reinforcing the practical relevance of this theoretical framework.

Our simulation results are presented as a heatmap in Figure 4.3, illustrating the impact of the signal-to-noise profile on de novo sequencing performance. The three experimental scenarios are labeled within the figure for reference. The simulations reveal that the computed p-values align with our theoretical expectations. Specifically, both older-generation DDA and Astral DIA exhibit similar p-values, which are significantly lower than those observed for older-generation DIA. This finding suggests that peptide-spectrum matching is intrinsically more favorable in the contexts of Astral DIA and DDA compared to older-generation DIA. The improved performance of Astral DIA compared to older-generation methods demonstrates that, despite its higher noise levels, it also produces significantly more signal peaks. In this scenario, the increased number of signal peaks effectively compensates for the additional noise. This suggests that the absolute signal peak count plays a more critical role than the signal-to-noise ratio, as Astral DIA yields a lower signal-to-noise ratio than older-generation methods yet still achieves superior results.

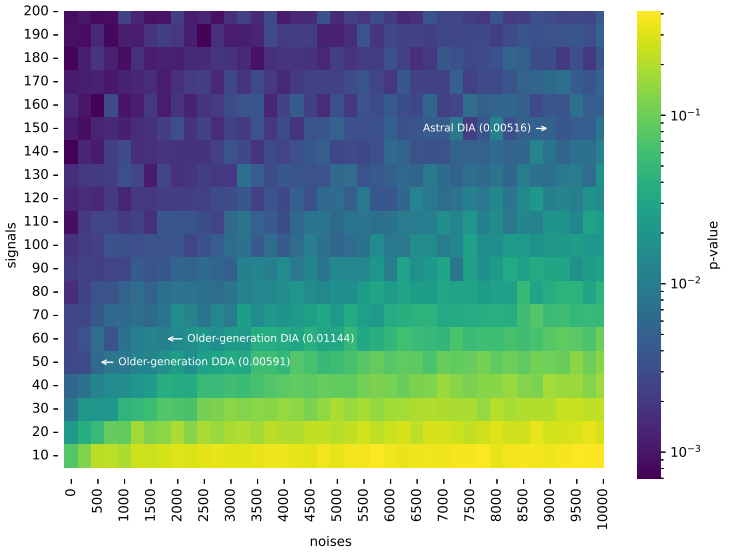


Figure 4.3: Simulated p-values for different signal and noise values. In the figure, older-generation DDA points to 50 signal peaks and 500 noise peaks with p-value of 0.00591, older-generation DIA points to 60 signal peaks and 1750 noise peaks with p-value 0.01144, while Astral-DIA points to 90 signal peaks and 9000 noise peaks with p-value 0.00516.

4.3 Conclusion

During our theoretical analysis of de novo and database search performance, we show that the balance between signal enhancement and noise accumulation is critical. In the case of Astral DIA, the substantial increase in informative signal peaks provides a richer dataset for de novo algorithms to process, improving the accuracy and confidence of peptide sequence reconstruction. This outcome underscores the importance of optimizing acquisition parameters to maximize signal quality while managing noise levels.

Our findings suggest that acquisition methods like Astral DIA, which can substantially increase signal peaks while effectively managing noise, are more conducive to de novo peptide sequencing. This has important implications for the development of mass spectrometry techniques and the selection of acquisition parameters in proteomics research. Conversely, older-generation DIA does not offer the same advantage because the modest gain in signal peaks is insufficient to counterbalance the significant escalation in noise peaks. This results in lower confidence in peptide identifications, as reflected by higher p-values.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis addresses several of the most pressing challenges in modern proteomics, focusing on enhancing protein inference and de novo peptide sequencing in the context of tandem mass spectrometry (MS/MS). Traditional approaches in protein inference struggle with ambiguous peptide mappings, and a lack of labeled training data, while traditional approaches in DIA de novo sequencing suffers from noise, multiplexed DIA data. To overcome these limitations, we introduced novel deep learning–based frameworks—GraphPI and DIANovo—as well as a theoretical analysis to systematically examine the factors influencing peptide identification performance.

In Chapter 2, we presented GraphPI, a graph neural network (GNN)–based framework that recasts protein inference as a node classification problem on a tripartite graph comprising proteins, peptides, and peptide-spectrum matches (PSMs). The architecture, inspired by GraphSAGE, is extended to handle heterogeneous nodes and edge types with specialized aggregation functions. Importantly, GraphPI uses a semi-supervised learning paradigm, leveraging pseudo-labels derived from existing inference tools and refining them through iterative self-training. Experimental results across multiple benchmark datasets demonstrate that GraphPI achieves superior performance in both accuracy and computational efficiency. It significantly reduces reliance on large manually labeled datasets while providing generalizable protein scores, highlighting the benefits of integrating graph-based representations into proteomics pipelines.

In Chapter 3, we introduced DIANovo, a Transformer-based de novo peptide sequencing model specifically designed to address the challenges posed by DIA data. DIA's mul-

tiplexed nature leads to overlapping signals and coeluting peptides, complicating peptide identification. DIANovo addresses this by constructing a fully connected spectrum graph and encoding edge features such as mass differences using Rotary Positional Embeddings. It incorporates coelution-aware pretraining, enabling the model to learn robust spectral representations from mixed signals. A two-stage decoding strategy is employed to trace a confident path through the spectrum graph and refine it into a complete peptide sequence. Evaluation on both older-generation and next-generation DIA datasets—particularly those from the Orbitrap Astral—demonstrates significant improvements in recall rates compared to existing methods. Our results also provide a nuanced understanding of when DIA can outperform DDA.

In Chapter 4, we complemented our empirical studies with a theoretical framework that models how signal and noise accumulation impacts peptide identification in different acquisition modes. By simulating varying signal-to-noise conditions for DDA, older-generation DIA, and Astral DIA, we showed that Astral DIA offers a favorable balance that improves identification accuracy, while older-generation DIA suffers from excessive noise. This analysis explains observed trends in empirical performance and offers predictive guidance for the design of future mass spectrometry experiments.

Together, these contributions represent a significant step forward in proteomic analysis. GraphPI improves the speed and accuracy of protein inference using graph-based modeling and semi-supervised learning. DIANovo pushes the frontier of de novo peptide sequencing in DIA by integrating advanced neural architectures with spectrum-specific inductive biases. The theoretical analysis provides a principled lens through which acquisition strategies can be evaluated and optimized.

5.2 Future Work

This thesis also opens several promising directions for future research:

- Generalization to Unseen Organisms and PTMs: Both GraphPI and DIANovo are trained on curated datasets that may not capture the full diversity of the proteome. Future work could explore domain adaptation and transfer learning techniques to enhance performance on unseen species or under-represented post-translational modifications (PTMs).
- Integration with Protein-Protein Interaction Networks: For GraphPI, incorporating biological priors such as protein-protein interactions or functional annotations may

further improve inference accuracy, especially in cases where peptide evidence alone is ambiguous.

- Real-Time and On-Instrument Deployment: As mass spectrometry moves toward real-time analysis and clinical applications, deploying lightweight versions of GraphPI and DIANovo on instrument-connected systems could make immediate peptide and protein identification feasible during acquisition. This would require us to further improve the computational efficiency of our methods.

- Identification of immunopeptides: As future work, we aim to apply our model to immunopeptidomics datasets, where identifying non-tryptic, low-abundance MHC-bound peptides remains a major challenge. De novo sequencing within the DIA framework offers a promising avenue for unbiased immunopeptide discovery. Evaluating our model in this context will help assess its potential for applications such as immune profiling and neoantigen identification.

- Improved Theoretical Bounds and Simulation Frameworks: The theoretical model in Chapter 4 offers a first step toward understanding performance limits. Future work could expand this framework to incorporate factors like noise profiles or retention time to better simulate experimental variability. It would also be worthwhile to develop a more rigorous analytical bound for peptide identification performance, providing theoretical guarantees and deeper insights into the limitations and capabilities of current sequencing algorithms.

References

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

[3] S. E Ahmed. Bayesian networks and decision graphs. *Technometrics*, 50, 2008.

[4] Erik Ahrné, Lars Molzahn, Timo Glatter, and Alexander Schmidt. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*, 13:2567–2578, 2013.

[5] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, et al. *Molecular biology of the cell*, volume 3. Garland New York, 1994.

[6] N. Leigh Anderson and Norman G. Anderson. The human plasma proteome: history, character, and diagnostic prospects., 2002.

[7] Anonymous. The problem with neoantigen prediction. *Nature biotechnology*, 35:97, 2017.

[8] Christian Bartels. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical & environmental mass spectrometry*, 19(6):363–368, 1990.

[9] Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature communications*, 7(1):13404, 2016.

[10] Michal Bassani-Sternberg and David Gfeller. Unsupervised hla peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-hla interactions. *Journal of Immunology*, 197(6):2492–2499, 2016.

[11] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1):26–31, November 2009.

[12] Mamatha Bhat, Sergi Clotet-Freixas, Cristina Baciu, Elisa Pasini, Ahmed Hammad, Tommy Ivanics, Shelby Reid, Amirhossein Azhie, Marc Angeli, Anand Ghanekar, Sandra Fischer, Gonzalo Sapisochin, and Ana Konvalinka. Combined proteomic/transcriptomic signature of recurrence post-liver transplantation for hepatocellular carcinoma beyond milan. *Clinical Proteomics*, 18:1–16, 2021.

[13] Stephen J Blanksby and Todd W Mitchell. Advances in mass spectrometry for lipidomics. *Annual Review of Analytical Chemistry*, 3(1):433–465, 2010.

[14] Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Saša M Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & Cellular Proteomics*, 14(5):1400–1410, 2015.

[15] Tomas Cajka and Oliver Fiehn. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *TrAC Trends in Analytical Chemistry*, 61:192–206, 2014.

[16] Tomas Cajka and Oliver Fiehn. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical chemistry*, 88(1):524–545, 2016.

[17] Ilaria Cela, Maria Concetta Cufaro, Maurine Fucito, Damiana Pieragostino, Paola Lanuti, Michele Sallese, Piero Del Boccio, Adele Di Matteo, Nerino Allocati, Vincenzo De Laurenzi, and Luca Federici. Proteomic investigation of the role of nucleostemin in nucleophosmin-mutated oci-aml 3 cell line. *International Journal of Molecular Sciences*, 23:7655, 2022.

[18] Matthew C. Chambers, Brendan MacLean, Robert Burke, Dario Amodei, Daniel L. Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jar-

rett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A. Baker, Mi Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L. Seymour, Lydia M. Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W. Deutsch, Robert L. Moritz, Jonathan E. Katz, David B. Agus, Michael MacCoss, David L. Tabb, and Parag Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30:918–920, 2012.

[19] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

[20] Chloe Chong, Fabio Marino, HuiSong Pak, Julien Racle, Roy T Daniel, Markus Müller, David Gfeller, George Coukos, and Michal Bassani-Sternberg. High-throughput and sensitive immunopeptidomics platform reveals profound interferonγ-mediated remodeling of the human leukocyte antigen (hla) ligandome. *Molecular & Cellular Proteomics*, 17(3):533–548, 2018.

[21] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[22] Jürgen Cox and Matthias Mann. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an orbitrap. *Journal of the American Society for Mass Spectrometry*, 20(8):1477–1485, 2009.

[23] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[24] Benjamin F Cravatt, Gabriel M Simon, and John R Yates Iii. The biological impact of mass-spectrometry-based proteomics. *Nature*, 450(7172):991–1000, 2007.

[25] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[26] Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17:41–44, 2020.

[27] Eleftherios P. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations, 2004.

[28] Bruno Domon and Ruedi Aebersold. Mass spectrometry and protein analysis. *science*, 312(5771):212–217, 2006.

[29] Shiva Ebrahimi and Xuan Guo. Transformer-based de novo peptide sequencing for data-independent acquisition mass spectrometry, 2024.

[30] Jarrett D Egertson, Brendan MacLean, Richard Johnson, Yue Xuan, and Michael J MacCoss. Multiplexed peptide analysis using data-independent acquisition and skyline. *Nature protocols*, 10(6):887–903, 2015.

[31] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4:207–214, 2007.

[32] Jimmy K. Eng, Michael R. Hoopmann, Tahmina A. Jahan, Jarrett D. Egertson, William S. Noble, and Michael J. MacCoss. A deeper look into comet - implementation and features. *Journal of the American Society for Mass Spectrometry*, 26:1865–1874, 2015.

[33] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: an open-source ms/ms sequence database search tool. *Proteomics*, 13(1):22–24, 2013.

[34] Jimmy K Eng, Alan L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[35] Vanessa C Evans, Gary Barker, Kate J Heesom, Jun Fan, Conrad Bessant, and David A Matthews. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature methods*, 9(12):1207–1211, 2012.

[36] Romanos Fasoulis, Georgios Paliouras, and Lydia E Kavraki. Graph representation learning for structural proteomics. *Emerging Topics in Life Sciences*, 5(6):789–802, 2021.

[37] Jorge Fernandez-de Cossio, Javier Gonzalez, Lazaro Betancourt, Vladimir Besada, Gabriel Padron, Yasutsugu Shimonishi, and Toshifumi Takao. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'seqms', a software aid for de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 12(23):1867–1878, 1998.

[38] Veronique Fischer, Vincent Hisler, Elisabeth Scheer, Elisabeth Lata, Bastien Morlet, Damien Plassard, Dominique Helmlinger, Didier Devys, László Tora, and Stephane D. Vincent. Supt3h-less saga coactivator can assemble and function without significantly perturbing rna polymerase ii transcription in mammalian cells. *Nucleic Acids Research*, 50:7972–7990, 2022.

[39] Ari Frank and Pavel Pevzner. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005.

[40] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, Nuno Bandeira, Ronald J Moore, Gordon A Anderson, Richard D Smith, and Pavel A Pevzner. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature methods*, 8(7):587–591, 2011.

[41] Philipp E Geyer, Nils A Kulak, Garwin Pichler, Lesca M Holdt, Daniel Teupser, and Matthias Mann. Plasma proteome profiling to assess human health and disease. *Cell systems*, 2(3):185–195, 2016.

[42] Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6), 2012.

[43] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[44] Clarisse Gotti, Florence Roux-Dalvai, Charles Joly-Beauparlant, Loïc Mangnier, Mickaël Leclercq, and Arnaud Droit. Extensive and accurate benchmarking of dia acquisition methods and software tools using a complex proteomic standard. *Journal of proteome research*, 20(10):4801–4814, 2021.

[45] Shenheng Guan, Michael F. Moran, and Bin Ma. Prediction of lc-ms/ms properties of peptides from sequence by deep learning. *Molecular and Cellular Proteomics*, 18:2099–2107, 2019.

[46] Ulises H Guzman, Ana Martinez-Val, Zilu Ye, Eugen Damoc, Tabiwang N Arrey, Anna Pashkova, Santosh Renuse, Eduard Denisov, Johannes Petzoldt, Amelia C Pe-

terson, et al. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nature Biotechnology*, pages 1–12, 2024.

[47] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 1025–1035. Curran Associates, Inc., 2017.

[48] Samir M Hanash, Sharon J Pitteri, and Vitor M Faca. Mining the plasma proteome for cancer biomarkers. *Nature*, 452(7187):571–579, 2008.

[49] Kimberly D. Herman, Carl G. Wright, Helen M. Marriott, Sam C. McCaughran, Kieran A. Bowden, Mark O. Collins, Stephen A. Renshaw, and Lynne R. Prince. The egfr/erbb inhibitor neratinib modifies the neutrophil phosphoproteome and promotes apoptosis and clearance by airway macrophages. *Frontiers in Immunology*, 13:956991, 2022.

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[51] Dawn E. Holmes. *Bayesian Networks: Theory and Philosophy*, volume 24, pages 103–119. Springer, 2022.

[52] Andrew L. Hopkins and Colin R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1:727–730, 2002.

[53] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[54] Takashi Hotta, Thomas S. McAlear, Yang Yue, Takumi Higaki, Sarah E. Haynes, Alexey I. Nesvizhskii, David Sept, Kristen J. Verhey, Susanne Bechstedt, and Ryoma Ohi. Eml2-s constitutes a new class of proteins that recognizes and regulates the dynamics of tyrosinated microtubules. *Current Biology*, 32:3898–3910, 2022.

[55] J Jeffry Howbert and William Stafford Noble. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular & Cellular Proteomics*, 13(9):2467–2479, 2014.

[56] En Chi Hsu, Meghan A. Rice, Abel Bermudez, Fernando Jose Garcia Marques, Merve Aslan, Shiqin Liu, Ali Ghoochani, Chiyuan Amy Zhang, Yun Sheng Chen,

Aimen Zlitni, Sahil Kumar, Rosalie Nolley, Frezghi Habte, Michelle Shen, Kashyap Koul, Donna M. Peehl, Amina Zoubeidi, Sanjiv S. Gambhir, Christian A. Kunder, Sharon J. Pitteri, James D. Brooks, and Tanya Stoyanova. Trop2 is a driver of metastatic prostate cancer with neuroendocrine phenotype via parp1. *Proceedings of the National Academy of Sciences of the United States of America*, 117:2032–2042, 2020.

[57] Ellis G. Jaffray, Michael H. Tatham, Barbara Mojsa, Magda Liczmanska, Alejandro Rojas-Fernandez, Yili Yin, Graeme Ball, and Ronald T. Hay. The p97/vcp segregase is essential for arsenic-induced degradation of pml and pml-rara. *Journal of Cell Biology*, 222:e202201027, 2023.

[58] Mathias Kalxdorf, Torsten Müller, Oliver Stegle, and Jeroen Krijgsveld. Icer improves proteome coverage and data completeness in global and single-cell proteomics. *Nature communications*, 12(1):4787, 2021.

[59] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Research*, 32:D277–D280, 2004.

[60] Minseung Kim, Ameen Eetemadi, and Ilias Tagkopoulos. Deeppep: Deep proteome inference from peptide profiles. *PLoS Computational Biology*, 13:e1005661, 2017.

[61] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[62] John Klimek, James S. Eddes, Laura Hohmann, Jennifer Jackson, Amelia Peterson, Simon Letarte, Philip R. Gafken, Jonathan E. Katz, Parag Mallick, Hookeun Lee, Alexander Schmidt, Reto Ossola, Jimmy K. Eng, Ruedi Aebersold, and Daniel B. Martin. The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools. *Journal of Proteome Research*, 7:96–103, 2008.

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

[64] Dominique Kreutz, Andrea Bileck, Kerstin Plessl, Denise Wolrab, Michael Groessl, Bernhard K. Keppler, Samuel M. Meier, and Christopher Gerner. Response profiling

using shotgun proteomics enables global metallodrug mechanisms of action to be established. *Chemistry - A European Journal*, 23:1881–1890, 2017.

[65] Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–925, 2007.

[66] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, Nichole King, Stephen E Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–667, 2007.

[67] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[69] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[70] Jiapeng Li, Logan S Smith, and Hao-Jie Zhu. Data-independent acquisition (dia): an emerging proteomics technology for analysis of drug-metabolizing enzymes and transporters. *Drug Discovery Today: Technologies*, 39:49–56, 2021.

[71] Yong Fuga Li and Predrag Radivojac. Computational approaches to protein inference in shotgun proteomics, 2012.

[72] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[73] Kaiyuan Liu, Yuzhen Ye, Sujun Li, and Haixu Tang. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications*, 14(1):7974, 2023.

[74] Ping Liu, Xi Wang, Yiwei Sun, Hongyu Zhao, Fang Cheng, Jifeng Wang, Fuquan Yang, Junjie Hu, Hong Zhang, Chih-chen Wang, et al. Sars-cov-2 orf8 reshapes the er through forming mixed disulfides with er oxidoreductases. *Redox Biology*, 54:102388, 2022.

[75] Wen Ting Lo, Hassane Belabed, Murat Kücükdisli, Juliane Metag, Yvette Roske, Polina Prokofeva, Yohei Ohashi, André Horatscheck, Davide Cirillo, Michael Krauss, Christopher Schmied, Martin Neuenschwander, Jens Peter von Kries, Guillaume Médard, Bernhard Kuster, Olga Perisic, Roger L. Williams, Oliver Daumke, Bernard Payrastre, Sonia Severin, Marc Nazaré, and Volker Haucke. Development of selective inhibitors of phosphatidylinositol 3-kinase c2α. *Nature Chemical Biology*, 19:18–27, 2023.

[76] Bingwen Lu and Ting Chen. Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discovery Today: BioSilico*, 2(2):85–90, 2004.

[77] Christina Ludwig, Ludovic Gillet, George Rosenberger, Sabine Amon, Ben C Collins, and Ruedi Aebersold. Data-independent acquisition-based swath-ms for quantitative proteomics: a tutorial. *Molecular systems biology*, 14(8):e8126, 2018.

[78] Christopher J. Lupton, Charles Bayly-Jones, Laura D'Andrea, Cheng Huang, Ralf B. Schittenhelm, Hari Venugopal, James C. Whisstock, Michelle L. Halls, and Andrew M. Ellisdon. The cryo-em structure of the human neurofibromin dimer reveals the molecular basis for neurofibromatosis type 1. *Nature Structural and Molecular Biology*, 28:982–988, 2021.

[79] Bin Ma. Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, 2015.

[80] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

[81] Zheng Ma, Jiazhen Chen, Lei Xin, and Ali Ghodsi. Graphpi: Efficient protein inference with graph neural networks. *Journal of Proteome Research*, 23(11):4821–4834, 2024.

[82] Zeping Mao, Ruixue Zhang, Lei Xin, and Ming Li. Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model. *Nature Machine Intelligence*, 5(11):1250–1260, 2023.

[83] Sean McIlwain, Kaipo Tamura, Attila Kertesz-Farkas, Charles E. Grant, Benjamin Diament, Barbara Frewen, J. Jeffry Howbert, Michael R. Hoopmann, Lukas Käll, Jimmy K. Eng, Michael J. MacCoss, and William Stafford Noble. Crux: Rapid open

source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, 13:4488–4491, 2014.

[84] Samuel M. Meier, Dominique Kreutz, Lilli Winter, Matthias H.M. Klose, Klaudia Cseh, Tamara Weiss, Andrea Bileck, Beatrix Alte, Johanna C. Mader, Samir Jana, Annesha Chatterjee, Arindam Bhattacharyya, Michaela Hejl, Michael A. Jakupec, Petra Heffeter, Walter Berger, Christian G. Hartinger, Bernhard K. Keppler, Gerhard Wiche, and Christopher Gerner. An organoruthenium anticancer agent shows unexpected target selectivity for plectin. *Angewandte Chemie - International Edition*, 56:8267–8271, 2017.

[85] Annette Michalski, Juergen Cox, and Matthias Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc- ms/ms. *Journal of proteome research*, 10(4):1785–1793, 2011.

[86] Annette Michalski, Eugen Damoc, Olaf Lange, and et al. Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, 10(9):M111–011015, 2011.

[87] Annette Michalski, Eugen Damoc, Olaf Lange, and et al. Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down lc ms/ms and versatile peptide fragmentation modes. *Molecular & Cellular Proteomics*, 10(9):M111–011015, 2011.

[88] Madiha Mumtaz, Irene V. Bijnsdorp, Franziska Böttger, Sander R. Piersma, Thang V. Pham, Samiullah Mumtaz, Ruud H. Brakenhoff, M. Waheed Akhtar, and Connie R. Jimenez. Secreted protein markers in oral squamous cell carcinoma (oscc). *Clinical Proteomics*, 19:1–15, 2022.

[89] Jan Muntel, Yue Xuan, Sebastian T Berger, Lukas Reiter, Richard Bachur, Alex Kentsis, and Hanno Steen. Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer. *Journal of proteome research*, 14(11):4752–4762, 2015.

[90] Thilo Muth and Bernhard Y Renard. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*, 19(5):954–970, 2018.

[91] Pedro Navarro, Jörg Kuharev, Ludovic C Gillet, Oliver M Bernhardt, Brendan MacLean, Hannes L Röst, Stephen A Tate, Chih-Chiang Tsou, Lukas Reiter, Ute

Distler, et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nature biotechnology*, 34(11):1130–1136, 2016.

[92] Richard E. Neapolitan. *Probabilistic methods for bioinformatics: With an introduction to bayesian networks*. Morgan Kaufmann, 2009.

[93] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11(11):1114–1125, 2014.

[94] Alexey I Nesvizhskii and Ruedi Aebersold. Interpretation of shotgun proteomic data. *Molecular & cellular proteomics*, 4(10):1419–1440, 2005.

[95] Alexey I. Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75:4646–4658, 2003.

[96] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nature methods*, 4(9):709–712, 2007.

[97] Gilbert S Omenn, Lydie Lane, Christopher M Overall, Ileana M Cristea, Fernando J Corrales, Cecilia Lindskog, Young-Ki Paik, Jennifer E Van Eyk, Siqi Liu, Stephen R Pennington, et al. Research on the human proteome reaches a major milestone:$>$ 90% of predicted human proteins now credibly detected, according to the hupo human proteome project. *Journal of proteome research*, 19(12):4735–4746, 2020.

[98] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12173–12182, 2020.

[99] David N Perkins, Darryl JC Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, 20(18):3551–3567, 1999.

[100] Julianus Pfeuffer, Timo Sachsenberg, Tjeerd M.H. Dijkstra, Oliver Serang, Knut Reinert, and Oliver Kohlbacher. Epifany: A method for efficient high-confidence protein inference. *Journal of Proteome Research*, 19:1060–1072, 2020.

[101] Rui Qiao, Ngoc Hieu Tran, Lei Xin, Xin Chen, Ming Li, Baozhen Shan, and Ali Ghodsi. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3:420–425, 2021.

[102] S Ramakrishnan and C Vogel. Gold standard of protein expression in yeast, 2009. Available from: http://www.marcottelab.org/MSdata/gold_yeast.html.

[103] Mathias Rask-Andersen, Markus Sällman Almén, and Helgi B. Schiöth. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery*, 10:579–590, 2011.

[104] Andreas Römpp and Bernhard Spengler. Mass spectrometry imaging with high resolution in mass and space. *Histochemistry and cell biology*, 139:759–783, 2013.

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

[106] Hannes L. Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E. Wolski, Oliver Schilling, Jyoti S. Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. Openms: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13:741–748, 2016.

[107] Alexander B. Saltzman, Mei Leng, Bhoomi Bhatt, Purba Singh, Doug W. Chan, Lacey Dobrolecki, Hamssika Chandrasekaran, Jong M. Choi, Antrix Jain, Sung Y. Jung, Michael T. Lewis, Matthew J. Ellis, and Anna Malovannaya. Gpgrouper: A peptide grouping algorithm for gene-centric inference and quantitation of bottom-up proteomics data. *Molecular and Cellular Proteomics*, 17:2270–2283, 2018.

[108] Justin Sanders, Bo Wen, Paul Rudnick, Rich Johnson, Christine C Wu, Sewoong Oh, Michael J MacCoss, and William Stafford Noble. A transformer model for de novo sequencing of data-independent acquisition mass spectrometry data. *bioRxiv*, pages 2024–06, 2024.

[109] Oliver Serang, Michael J. MacCoss, and William Stafford Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of Proteome Research*, 9:5346–5357, 2010.

[110] Oliver Serang and William Noble. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface*, 5(1):3, 2012.

[111] Gloria M Sheynkman, Michael R Shortreed, Brian L Frey, and Lloyd M Smith. Discovery and mass spectrometric analysis of novel splice-junction peptides using rna-seq. *Molecular & Cellular Proteomics*, 12(8):2341–2353, 2013.

[112] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

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

[114] Marina Spivak, Jason Weston, Daniela Tomazela, Michael J. MacCoss, and William Stafford Noble. Direct maximization of protein identifications from tandem mass spectra. *Molecular and Cellular Proteomics*, 11:M111.012161, 2012.

[115] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nature reviews Molecular cell biology*, 5(9):699–711, 2004.

[116] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[117] Yanting Su, Yuanyuan Guo, Jieyu Guo, Ting Zeng, Ting Wang, and Wu Liu. Study of foxo1-interacting proteins using turboid-based proximity labeling technology. *BMC Genomics*, 24:1–13, 2023.

[118] John EP Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences*, 101(26):9528–9533, 2004.

[119] Matthew The, Fredrik Edfors, Yasset Perez-Riverol, Samuel H. Payne, Michael R. Hoopmann, Magnus Palmblad, Björn Forsström, and Lukas Käll. A protein standard that emulates homology for the characterization of protein inference algorithms. *Journal of Proteome Research*, 17:1879–1886, 2018.

[120] Ying S Ting, Jarrett D Egertson, James G Bollinger, Brian C Searle, Samuel H Payne, William Stafford Noble, and Michael J MacCoss. Pecan: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature methods*, 14(9):903–908, 2017.

[121] Shivani Tiwary, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature methods*, 16(6):519–525, 2019.

[122] Ngoc Hieu Tran, Rui Qiao, Zeping Mao, Shengying Pan, Qing Zhang, Wenting Li, Lei Xin, Ming Li, and Baozhen Shan. Novoboard: a comprehensive framework for evaluating the false discovery rate and accuracy of de novo peptide sequencing. *Molecular & Cellular Proteomics*, 23(11), 2024.

[123] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, 16:63–66, 2019.

[124] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114:8247–8252, 2017.

[125] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I Nesvizhskii. Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods*, 12(3):258–264, 2015.

[126] Chih-Chiang Tsou, Chia-Feng Tsai, Guo Ci Teo, Yu-Ju Chen, and Alexey I Nesvizhskii. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using orbitrap mass spectrometers. *Proteomics*, 16(15-16):2257–2271, 2016.

[127] Stefka Tyanova, Tikira Temu, and Juergen Cox. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, 11(12):2301–2319, 2016.

[128] Julian Uszkoreit, Alexandra Maerkens, Yasset Perez-Riverol, Helmut E. Meyer, Katrin Marcus, Christian Stephan, Oliver Kohlbacher, and Martin Eisenacher. Pia:

An intuitive protein inference engine with a web-based user interface. *Journal of Proteome Research*, 14:2988–2997, 2015.

[129] Aaron van den Oord, Sander Dieleman, Heiga Zen, and et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[130] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[131] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[132] John D Venable, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, and John R Yates III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods*, 1(1):39–45, 2004.

[133] Anandalakshmi Venkatraman, Bamaprasad Dutta, Elavazhagan Murugan, Hao Piliang, Rajamani Lakshminaryanan, Anita Chan Sook Yee, Konstantin V. Pervushin, Siu Kwan Sze, and Jodhbir S. Mehta. Proteomic analysis of amyloid corneal aggregates from tgfbi-h626r lattice corneal dystrophy patient implicates serine-protease htra1 in mutation-specific pathogenesis of tgfbip. *Journal of Proteome Research*, 16:2899–2913, 2017.

[134] Antonella Vitiello and Maurizio Zanetti. Neoantigen prediction and the need for validation. *Nature biotechnology*, 35(9):815–817, 2017.

[135] Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S Pullman, Seong Won Cha, and Nuno Bandeira. Assembling the community-scale discoverable human proteome. *Cell systems*, 7(4):412–421, 2018.

[136] Craig D Wenger and Joshua J Coon. A prospective peptide sequencing-based pipeline for rapid and comprehensive proteome characterization. *Nature Methods*, 10(4):333–337, 2011.

[137] Dirk A Wolters, Michael P Washburn, and John R Yates. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry*, 73(23):5683–5690, 2001.

[138] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[139] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810, 2021.

[140] John R Yates, Jimmy K Eng, Ashley L McCormack, and David Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, 67(8):1426–1436, 1995.

[141] John R Yates III. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, 33(1):1–19, 1998.

[142] Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William S Noble. De novo mass spectrometry peptide sequencing with a transformer model. In *International Conference on Machine Learning*, pages 25514–25522. PMLR, 2022.

[143] Yonghan Yu and Ming Li. Towards highly sensitive deep learning-based end-to-end database search for tandem mass spectrometry. *Nature Machine Intelligence*, pages 1–11, 2025.

[144] Amr El Zawily, Frederick S. Vizeacoumar, Renuka Dahiya, Sara L. Banerjee, Kalpana K. Bhanumathy, Hussain Elhasasna, Glinton Hanover, Jessica C. Sharpe, Malkon G. Sanchez, Paul Greidanus, R. Greg Stacey, Kyung Mee Moon, Ilya Alexandrov, Juha P. Himanen, Dimitar B. Nikolov, Humphrey Fonge, Aaron P. White, Leonard J. Foster, Bingcheng Wang, Behzad M. Toosi, Nicolas Bisson, Tajib A. Mirzabekov, Franco J. Vizeacoumar, and Andrew Freywald. A multipronged unbiased strategy guides the development of an anti-egfr/epha2–bispecific antibody for combination cancer therapy. *Clinical Cancer Research*, 29:OF1–OF16, 2023.

[145] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A Lajoie, and Bin Ma. Peaks db: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics*, 11(4), 2012.

[146] Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343–2394, 2013.

[147] Ziqiang Zhang, Shiaw-Lin Wu, and David L Stenoien. De novo peptide sequencing using complementary tandem mass spectrometry. *Journal of Proteome Research*, 11(11):5609–5616, 2012.

[148] Yingming Zhao and Ole N Jensen. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics*, 9(20):4632–4641, 2009.

[149] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[150] Fatema Tuz Zohora, M. Ziaur Rahman, Ngoc Hieu Tran, Lei Xin, Baozhen Shan, and Ming Li. Deepiso: A deep learning model for peptide feature detection from lc-ms map. *Scientific Reports*, 9:1–13, 2019.