A Comprehensive and Systematic Review for Deep Learning-Based De Novo Peptide Sequencing

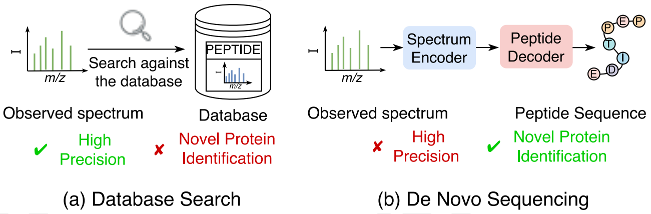
Jun Xia[1], **Jingbo Zhou**[2], **Shaorong Chen**[2], **Tianze Ling**[3] and **Stan Z. Li**[2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Westlake University
[3]Tsinghua University
junxia@hkust-gz.edu.cn, {zhoujingbo, chenshaorong, stan.zq.li}@westlake.edu.cn,
ltz20@mails.tsinghua.edu.cn

Abstract

Tandem mass spectrometry (MS/MS) has revolutionized the field of proteomics, enabling the high-throughput identification of proteins. However, one of the central challenges in mass spectrometry-based proteomics remains peptide identification, especially in the absence of a comprehensive peptide database. While traditional database search methods compare observed mass spectra to pre-existing protein databases, they are limited by the availability and completeness of these databases. *De novo* peptide sequencing, which derives peptide sequences directly from mass spectra, has emerged as a crucial approach in such cases. In recent years, deep learning has made significant strides in this domain. These methods train deep neural networks for translating mass spectra into peptide sequences without relying on any pre-constructed databases. Despite significant progress, this field still lacks a comprehensive and systematic review. In this paper, we provide the first review of deep learning-based *de novo* peptide sequencing techniques from the perspectives of data types, model architectures, decoding strategies, applications and evaluation metrics. We also identify key challenges and highlight promising avenues for future research, providing a valuable resource for the AI and scientific communities.

1 Introduction

Peptide identification through tandem mass spectrometry is a cornerstone of modern proteomics research [Aebersold and Mann, 2003]. The analysis of peptide fragmentation patterns allows for the determination of peptide sequences, which in turn facilitates protein characterization and quantification. As shown in Fig. 1(a), traditional peptide identification has been performed through database search methods, which rely on comparing observed mass spectra to pre-existing databases [Yates III, 1998]. While database search can achieve high precision in many cases, these methods are inherently limited by the completeness and relevance of the available databases. The absence of a suitable database or the presence of novel or uncharacterized peptides requires alternative approaches to peptide identification [VanDuijn *et al.*, 2017; Mayer and Impens, 2021], specifically, *de novo* peptide sequencing shown in Fig. 1(b).



(a) Database Search

(b) De Novo Sequencing

Figure 1: Schematic diagram and comparison of database search and *de novo* peptide sequencing (adapted from our previous work [Xia *et al.*, 2025]).

De novo peptide sequencing directly infers the peptide sequence from mass spectrometry data without relying on a reference database, akin to machine translation in Natural Language Processing (NLP) research [Stahlberg, 2020], where the source language is directly translated into the target language. This approach has become increasingly important in the analysis of complex samples, where unknown peptides may be present, or when studying species with incomplete or unannotated genomes [Nesvizhskii, 2007]. In recent years, deep learning techniques have brought remarkable advancements in the domain of *de novo* peptide sequencing. These methods hold great promise for enhancing sequence accuracy and throughput. The DeepNovo algorithm [Tran *et al.*, 2017], introduced in 2017, was among the pioneering deep learning approaches that significantly improved the performance of *de novo* sequencing. Subsequently, PointNovo [Qiao *et al.*, 2021] innovatively treats mass spectrum data as point clouds and utilizes an order-invariant neural network for peptide sequencing from high-resolution mass spectrometry data. More recently, inspired by the resounding success of the transformer [Vaswani *et al.*, 2017] in natural language processing and computer vision, Casanovo [Yilmaz *et al.*, 2022] was the first to apply a transformer encoder-decoder architecture to predict peptide sequences from observed mass spectra. Following the lead of Casanovo's transformer-based architecture, recent research efforts have been increasingly focused on devising more effective training strategies. For example, ContraNovo [Jin *et al.*, 2024] adopts contrastive learning to extract the subtle correla-

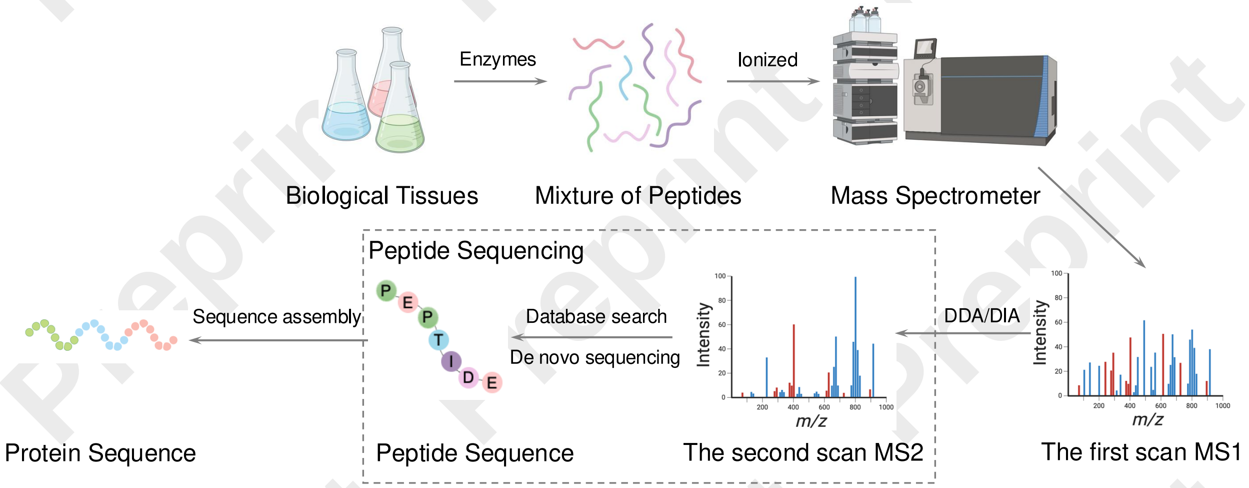


Figure 2: The standard workflow for protein identification in proteomics (adapted from our previous work [Xia *et al.*, 2025]).

tions between spectra and peptides and integrates mass information into the peptide decoding process. AdaNovo [Xia *et al.*, 2024] puts forward conditional mutual information-based re-weighting methods, which are instrumental in identifying amin acids with Post Translational Modifications (PTMs) [Ramazi and Zahiri, 2021]. Furthermore, SearchNovo [Xia *et al.*, 2025] and ReNovo [Chen *et al.*, 2025] leverage database search to enhance *de novo* peptide sequencing, thus enjoying the advantages of both paradigms.

Although deep learning-based *de novo* peptide sequencing methods have achieved overwhelming success in protein identification, this rapidly expanding field still lacks a systematic review. Also, we focus solely on deep learning methods, as previous reviews have adequately covered earlier work based on traditional methods in this field [Vitorino *et al.*, 2020; Ng *et al.*, 2023]. In this paper, we present the first review to assist audiences of diverse backgrounds in understanding, using, and developing *de novo* peptide sequencing tools or methods for various practical tasks.

The contributions of this work can be summarized from the following four aspects.
(1) *A structured taxonomy.* A broad overview of the field is presented with a structured taxonomy that categorizes existing works from 5 perspectives (Fig. 3): data type, model architectures, decoding strategies, applications, and evaluation metrics.
(2) *Thorough review of the current progress.* Based on the taxonomy, the current research progress of deep learning-based *de novo* peptide sequencing is systematically delineated.
(3) *Abundant additional resources.* Abundant resources are collected and can be found at https://github.com/jingbo02/Awesome-Denovo-Peptide-Sequencing. These resources will be continuously updated on a regular basis.
(4) *Discussion of future directions.* The limitations of existing works are discussed and several promising research directions are highlighted.

2 Background

To help the AI community better understand mass spectrometry data and the task of *de novo* peptide sequencing, we first provide a brief overview of the workflow of mass spectrometry-based protein identification. As shown in Fig. 2, a standard protein identification workflow in shotgun proteomics [Zhang *et al.*, 2013] begins with enzymatic digestion of proteins, producing a mixture of peptides. These peptides are then separated using liquid chromatography before being introduced into a mass spectrometer. The first scan (MS1) records the mass-to-charge (*m/z*) ratios of intact peptides. Subsequently, peptides undergo fragmentation in the mass spectrometer based on different precursor ion selection strategies, generating second scan (MS2) spectra, which consist of multiple peaks. In Data-Dependent Acquisition (DDA) [Bateman *et al.*, 2014], the instrument selects the most intense precursor ions from the MS1 scan for fragmentation, resulting in high-quality MS2 spectra but potentially missing low-abundance peptides. In contrast, Data-Independent Acquisition (DIA) [Doerr, 2015] fragments all precursor ions within a predefined *m/z* range, ensuring comprehensive peptide coverage at the cost of increased spectral complexity. Each peak in an MS2 spectrum is represented as a tuple containing an *m/z* value and an associated intensity. MS2 spectral data can thus be categorized into two types based on the acquisition strategy: DDA and DIA. A detailed discussion of data types is provided in Section 3. The core of this workflow is peptide sequencing, where we aim to predict the peptide sequence using the observed MS2 spectrum and the corresponding precursor information (mass and charge of the intact peptide). However, accurate sequencing is complicated by challenges such as incomplete fragmentation, noisy spectra, and the presence of post-translational modifications (PTMs). To overcome these issues, computational methods leverage database search strategies or deep learning-based *de novo* sequencing approaches to improve sequence prediction accuracy. Finally, the entire protein sequence can be inferred using assembly tools [Liu *et al.*, 2015].

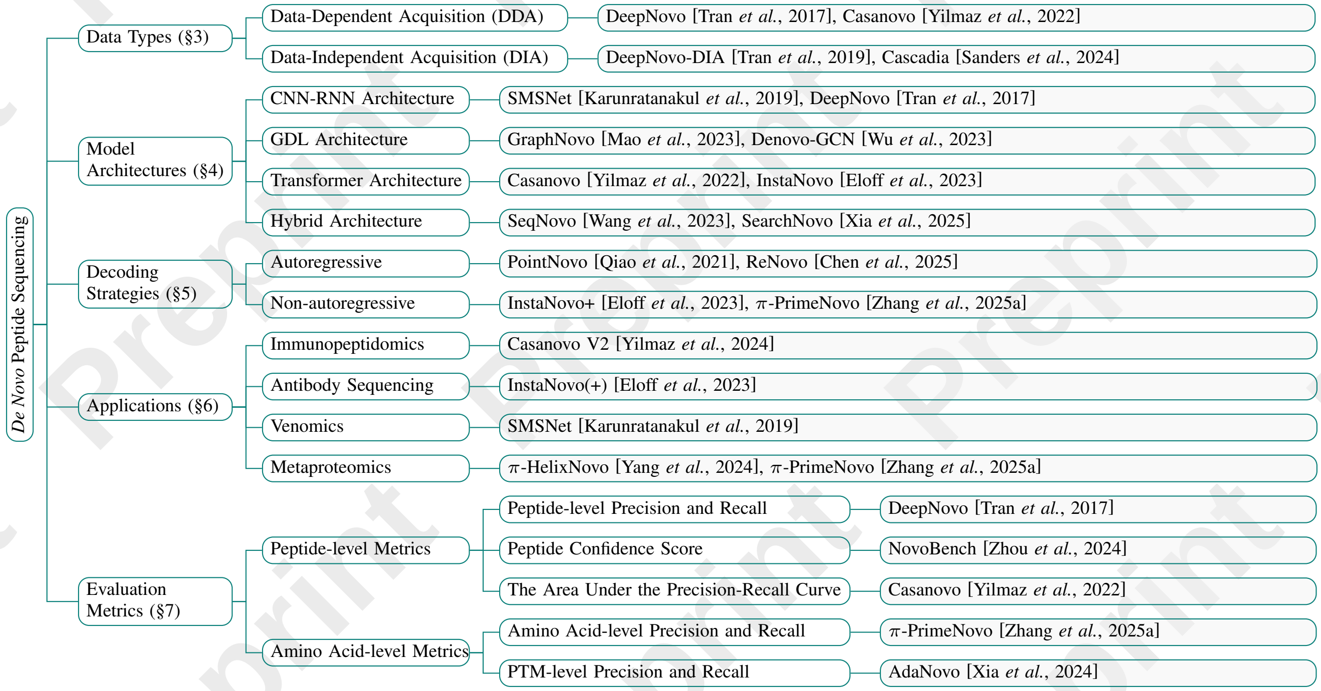

De Novo Peptide Sequencing
Data Types (§3)
Data-Dependent Acquisition (DDA)
DeepNovo [Tran *et al.*, 2017], Casanovo [Yilmaz *et al.*, 2022]
Data-Independent Acquisition (DIA)
DeepNovo-DIA [Tran *et al.*, 2019], Cascadia [Sanders *et al.*, 2024]
Model Architectures (§4)
CNN-RNN Architecture
SMSNet [Karunratanakul *et al.*, 2019], DeepNovo [Tran *et al.*, 2017]
GDL Architecture
GraphNovo [Mao *et al.*, 2023], Denovo-GCN [Wu *et al.*, 2023]
Transformer Architecture
Casanovo [Yilmaz *et al.*, 2022], InstaNovo [Eloff *et al.*, 2023]
Hybrid Architecture
SeqNovo [Wang *et al.*, 2023], SearchNovo [Xia *et al.*, 2025]
Decoding Strategies (§5)
Autoregressive
PointNovo [Qiao *et al.*, 2021], ReNovo [Chen *et al.*, 2025]
Non-autoregressive
InstaNovo+ [Eloff *et al.*, 2023], π-PrimeNovo [Zhang *et al.*, 2025a]
Applications (§6)
Immunopeptidomics
Casanovo V2 [Yilmaz *et al.*, 2024]
Antibody Sequencing
InstaNovo(+) [Eloff *et al.*, 2023]
Venomics
SMSNet [Karunratanakul *et al.*, 2019]
Metaproteomics
π-HelixNovo [Yang *et al.*, 2024], π-PrimeNovo [Zhang *et al.*, 2025a]
Evaluation Metrics (§7)
Peptide-level Metrics
Peptide-level Precision and Recall
DeepNovo [Tran *et al.*, 2017]
Peptide Confidence Score
NovoBench [Zhou *et al.*, 2024]
The Area Under the Precision-Recall Curve
Casanovo [Yilmaz *et al.*, 2022]
Amino Acid-level Metrics
Amino Acid-level Precision and Recall
π-PrimeNovo [Zhang *et al.*, 2025a]
PTM-level Precision and Recall
AdaNovo [Xia *et al.*, 2024]

Figure 3: A taxonomy of *De novo* peptide sequencing with representative examples.

3 Mass Spectrometry Data Type

In mass spectrometry-based proteomics, two main data acquisition strategies are employed: Data-Dependent Acquisition (DDA) [Bateman *et al.*, 2014] and Data-Independent Acquisition (DIA) [Doerr, 2015]. These techniques determine how the mass spectrometer collects and processes ion fragmentation data, which directly influences the peptide identification and quantification process.

3.1 Data-Dependent Acquisition (DDA)

Data-Dependent Acquisition (DDA) is a traditional approach in mass spectrometry where the instrument first performs a full survey scan to detect the total ion spectrum (precursor ion spectrum) across a wide *m/z* range. Based on the intensity of the detected peaks in MS1, the most abundant ions are selected for fragmentation in subsequent scans (MS2). The selection of precursor ions for fragmentation is dynamic, meaning that only the strongest ions are chosen for analysis. This process is repeated multiple times, with different precursor ions being targeted in each cycle. DDA is highly effective for identifying peptides that are abundant in the sample, making it well-suited for discovery-based proteomics. However, because it focuses on the most abundant ions, DDA may miss low-abundance peptides and thus offer incomplete proteome coverage. As shown in Table 1, most *de novo* sequencing methods currently focus on DDA data, as it is more widely available and easier to process.

3.2 Data-Independent Acquisition (DIA)

Data-Independent Acquisition (DIA) is a more advanced and systematic approach that differs from DDA by fragmenting all precursor ions within predefined *m/z* windows, regardless of their intensity. Rather than dynamically selecting precursor ions based on their intensity, DIA fragments ions across the entire *m/z* range in a non-discriminatory manner. This ensures that even low-abundance peptides, which might be overlooked in DDA, are included in the analysis. DIA provides a more comprehensive and reproducible dataset, making it particularly useful for quantitative proteomics and large-scale studies. While it offers better proteome coverage and is less biased toward high-abundance peptides, DIA fragments all precursor ions within a given mass range, resulting in highly complex and overlapping spectra. This makes it harder for deep learning models to correctly associate fragment ions with their corresponding precursor peptides compared to DDA. Additionally, since all ions are fragmented simultaneously, the sensitivity for individual peptides may be slightly reduced compared to DDA, but the method's overall coverage and consistency make it ideal for more in-depth analyses of complex biological samples. As shown in Table 1, *de novo* peptide sequencing methods for DIA data are relatively fewer compared to DDA, primarily due to the lack of large and well-annotated DIA datasets for model training.

4 Model Architectures

Numerous powerful model architectures have been adopted in the field of *de novo* peptide sequencing. Specifically, the model architectures of current methods fall into four categories: CNN-RNN architecture, Transformer architecture [Vaswani *et al.*, 2017], Geometric Deep Learning (GDL) architecture [Cao *et al.*, 2022], and Hybrid architecture.

Model	Data Type	Model Architecture	Decoding Strategy	Code Link
DeepNovo [Tran *et al.*, 2017]	DDA	CNN-RNN	AR	Link
DeepNovo-DIA [Tran *et al.*, 2019]	DIA	CNN-RNN	AR	Link
SMSNet [Karunratanakul *et al.*, 2019]	DDA	CNN-RNN	AR	Link
RANovo [Liu and Zhao, 2020]	DDA	CNN-RNN	AR	Unavailable
PointNovo [Qiao *et al.*, 2021]	DDA	GDL	AR	Link
Casanovo [Yilmaz *et al.*, 2022]	DDA	Transformer	AR	Link
DPST [Yang *et al.*, 2022]	DDA	Transformer	AR	Link
DEPS [Ge *et al.*, 2022]	DDA	CNN-RNN	AR	Unavailable
PepNet [Liu *et al.*, 2023]	DDA/DIA	CNN-RNN	NAR	Link
BiATNovo [Yang *et al.*, 2023]	DDA/DIA	CNN-RNN	AR	Link
GraphNovo [Mao *et al.*, 2023]	DDA	GDL	NAR	Link
PGPointNovo [Xu *et al.*, 2023]	DDA	GDL	AR	Link
Denovo-GCN [Wu *et al.*, 2023]	DDA	GDL	AR	Unavailable
SeqNovo [Wang *et al.*, 2023]	DDA	Hybrid	AR	Unavailable
InstaNovo [Eloff *et al.*, 2023]	DDA	Transformer	AR	Link
InstaNovo+ [Eloff *et al.*, 2023]	DDA	Transformer	NAR	Link
π-HelixNovo [Yang *et al.*, 2024]	DDA	Transformer	AR	Link
ContraNovo [Jin *et al.*, 2024]	DDA	Transformer	NAR	Link
NovoB [Lee and Kim, 2024]	DDA	Transformer	AR	Link
AdaNovo [Xia *et al.*, 2024]	DDA	Transformer	AR	Link
Transformer-DIA [Ebrahimi and Guo, 2024]	DIA	Transformer	AR	Link
Cascadia [Sanders *et al.*, 2024]	DIA	Transformer	AR	Link
Spectralis [Klaproth-Andrade *et al.*, 2024]	DDA	CNN-RNN	AR	Link
PowerNovo [Petrovskiy *et al.*, 2024]	DDA	Hybrid	AR	Link
CrossNovo [Zhang *et al.*, 2025b]	DDA	Transformer	AR	Link
SearchNovo [Xia *et al.*, 2025]	DDA	Hybrid	AR	Link
RankNovo [Qiu *et al.*, 2025]	DDA	Transformer	AR	Link
ReNovo [Chen *et al.*, 2025]	DDA	Hybrid	AR	Link
π-PrimeNovo [Zhang *et al.*, 2025a]	DDA	Transformer	NAR	Link

Table 1: A summary of representative *de novo* peptide sequencing methods in literature.

4.1 CNN-RNN Architecture

The application of CNN-RNN architectures in de novo peptide sequencing methods has revolutionized the field of proteomics by enhancing the accuracy and efficiency of peptide identification from tandem mass spectra. DeepNovo [Tran *et al.*, 2017] utilizes a hybrid architecture that combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to learn complex features from mass spectra, allowing it to predict peptide sequences with significant improvements in accuracy compared to traditional methods. This model iteratively predicts amino acids and integrates local dynamic programming to optimize the sequencing process, achieving high coverage and accuracy for antibody sequences without reliance on existing databases. DeepNovo-DIA [Tran *et al.*, 2019] extends this concept to data-independent acquisition (DIA) mass spectrometry, employing neural networks to capture relationships across multiple dimensions of spectral data, thus addressing challenges posed by multiplexed spectra. DePS [Ge *et al.*, 2022] first processes the input mass spectrometry data through a CNN in the feature extraction module to extract important features. Then, the LSTM captures the sequential dependencies in the peptide sequence, effectively modeling the relationships between amino acids. This dual approach allows DePS to maintain good performance even under challenging conditions, such as missing signal peaks and excessive noise. SMSNet [Karunratanakul *et al.*, 2019] adopts a similar architecture but emphasizes the identification of novel peptides, achieving over 95% amino acid accuracy while maintaining good coverage. It employs an excitation mechanism to discern important pairwise relationships among input features, enabling it to discover previously uncharacterized peptides effectively. PepNet [Liu *et al.*, 2023], on the other hand, is a fully convolutional network that processes high-dimensional input spectra through a series of residual dilated convolution blocks and a residual Transformer block. This design captures both local and global sequence information, significantly outperforming existing algorithms in peptide-level accuracy and processing speed. Lastly, BiATNovo [Yang *et al.*, 2023] introduces an attention-based bidirectional framework that enhances prediction accuracy for longer peptides by effectively capturing relationships between mass spectra and fragment ions through a two-phase training strategy. Its post-processing module further refines predictions by mitigating biases commonly observed in sequence prediction. Above methods have small parameter sizes and fast running speeds; however, their model expressiveness is limited.

4.2 Geometric Deep Learning Architecture

The application of geometric deep learning architectures in *de novo* peptide sequencing has gained significant attention due to their ability to effectively model complex relationships between mass spectrometry peaks, which is important for peptide sequence generation. Among the prominent methods, GraphNovo [Mao *et al.*, 2023] employs a two-stage graph-based approach using graph neural networks (GNNs) [Wu *et al.*, 2020], where the first stage identifies optimal paths in spectrum graphs through a Graphormer [Ying *et al.*, 2021] encoder, while the second stage resolves unknown mass tags using transformer decoders to address missing fragmentation issues. Denovo-GCN [Wu *et al.*, 2023] combines graph convolutional networks (GCN) [Kipf and Welling, 2017] with convolutional neural networks, constructing undirected spectrum graphs where nodes represent spectral peaks and edges encode mass relationships, enabling robust feature extraction through hybrid architectures. PointNovo [Qiao *et al.*, 2021] utilizes an order-invariant neural network that directly processes raw peak sets through a novel T-Net structure, achieving instrument-resolution independence by avoiding spectral discretization while maintaining constant computational complexity. PGPointNovo [Xu *et al.*, 2023] extends PointNovo's architecture through PyTorch-based data parallelization, implementing gradient synchronization across multiple GPUs and advanced optimization techniques like Rectified Adam to enable large-scale processing without sacrificing precision-recall performance.

4.3 Transformer Architecture

Transformer [Vaswani *et al.*, 2017] architectures have revolutionized *de novo* peptide sequencing by enabling end-to-end learning from mass spectrometry (MS) data while handling variable-length input spectra and output peptide sequences. These models typically employ encoder-decoder frameworks with attention mechanisms to map spectral peaks to amino acid sequences, often incorporating specialized components for spectral processing, precursor mass integration, and iterative refinement. Casanovo [Yilmaz *et al.*, 2022] pioneered the transformer-based approach with a vanilla encoder-decoder architecture that processes raw MS/MS spectra without *m/z* binning, using sinusoidal embeddings for peak features and precursor information. Its encoder contextualizes spectral peaks through self-attention, while the decoder autoregressively predicts amino acids using cross-attention to encoded spectra. InstaNovo [Eloff *et al.*, 2023] enhanced this paradigm with multi-scale sinusoidal embeddings for peak resolution adaptation and introduced InstaNovo+ [Eloff *et al.*, 2023], a diffusion model that iteratively refines predictions through multinomial denoising. DPST [Yang *et al.*, 2022] introduced amino-acid-aware attention through a confidence value aggregation encoder that prioritizes spectral peaks based on local amino acid connectivity, coupled with a global-local fusion decoder integrating both contextualized spectrum representations and amino acid priors. π-HelixNovo [Yang *et al.*, 2024] processes complementary synthetic spectra alongside experimental data through dual encoders to address missing ion challenges, while π-PrimeNovo [Zhang *et al.*, 2025a] employs non-autoregressive decoding with parallel amino acid prediction and mass constraint verification for 69x faster inference. NovoB [Lee and Kim, 2024] introduced bidirectional decoding via twin decoders that predict sequences from N- to C-terminus and vice versa, leveraging complementary ion series information. For data-independent acquisition (DIA) spectra, Transformer-DIA [Ebrahimi and Guo, 2024] extends Casanovo with hybrid encoders integrating MS1/MS2/precursor features, and Cascadia [Sanders *et al.*, 2024] implements transformer-based multiplexed spectrum interpretation specifically optimized for DIA workflows, demonstrating improved variant peptide detection through learned attention patterns across co-fragmented precursors.

4.4 Hybrid Architecture

Recent advancements in *de novo* peptide sequencing have introduced hybrid architectures that integrate diverse computational strategies to address longstanding challenges like post-translational modification identification, spectral noise, and missing peaks. These methods combine machine learning paradigms, retrieval mechanisms, and mass spectrometry data fusion to enhance accuracy and robustness. AdaNovo [Xia *et al.*, 2024] employs conditional mutual information (CMI) to adaptively weigh spectral-peptide relationships during training, prioritizing informative amino acids and PTMs while down-weighting noisy data. Its architecture uses CMI to dynamically adjust loss functions, improving PTM detection in low-frequency training scenarios. ContraNovo [Jin *et al.*, 2024] leverages contrastive learning to model pairwise spectra-peptide interactions and uniquely incorporates prefix/suffix mass data during decoding. By embedding mass compatibility checks into its transformer-based framework, it refines amino acid predictions at each step. ReNovo [Chen *et al.*, 2025] introduces a retrieval-augmented approach, building a datastore of training-derived spectral-peptide pairs to guide inference. This hybridizes database search principles with de novo flexibility, enabling novel peptide identification while leveraging retrieved contextual patterns. PowerNovo [Petrovskiy *et al.*, 2024] combines Transformer-based sequence-to-sequence learning with a BERT-inspired evaluator, forming an ensemble that corrects sequencing errors and assesses detectability. Finally, SeqNovo [Wang *et al.*, 2023] integrates multilayer perceptrons (MLPs) with attention mechanisms to emphasize critical spectral features.

5 Decoding Strategies

In the field of machine learning, two primary methodologies for sequence generation are autoregressive (AR) models and non-autoregressive (NAR) models. Autoregressive models are particularly effective in scenarios that demand high accuracy and the modeling of dependencies, whereas non-autoregressive models are favored for their efficiency and rapid performance in real-time applications. This distinction also applies to *de novo* peptide sequencing, where models can be categorized into AR and NAR types based on their sequence generation patterns.

AR models are a class of generative models that rely on previously generated peptide sequences to iteratively predict the next amino acid identity iteratively. The fundamental concept of AR model is that the generation of the next amino acid identity is contingent upon the peptide sequence that have been previously predicted. Specifically, the AR *de novo* peptide sequencing models are designed to predict the peptide $\mathbf{y} = \{y_i\}_{i=1}^{N} = (y_1, y_2, \ldots, y_N)$ given MS2 data $\mathbf{s}$, precursor $\mathbf{p}$, and model parameter θ:

$$P(\mathbf{y} \mid \mathbf{s}, \mathbf{p}; \theta) = \prod_{t=1}^{N} p(y_t \mid y_{1:t-1}, \mathbf{s}, \mathbf{p}; \theta) \tag{1}$$

Non-autoregressive (NAR) models are designed to enhance the efficiency of peptide sequence generation by generating the entire amino acid sequence in parallel and reduce reliance on previous outputs. Although NAR models offer superior efficiency in sequence generation compared to AR models, they often fall short in their ability to capture the dependencies within amino acid sequences.

6 Applications

De novo peptide sequencing has been widely applied in various fields where reference databases are incomplete or unavailable. Its ability to directly infer peptide sequences from mass spectrometry data makes it particularly valuable in immunology, antibody research, venomics, and metaproteomics studies. Below, we highlight some of its key applications.

6.1 Immunopeptidomics

One of the most common applications of *de novo* peptide sequencing is the identification of neoantigens and noncanonical antigens, which play crucial roles in cancer immunotherapy and autoimmune disease research. Neoantigens are tumor-specific peptides arising from somatic mutations, making them promising targets for personalized cancer vaccines. Noncanonical antigens, including those derived from alternative splicing, post-translational modifications, or cryptic translation, expand the repertoire of potential immunogenic peptides. Previous methods including DeepNovo and pNovo 3 have been employed to discover novel peptides without relying on a reference database [Tran *et al.*, 2020; Li *et al.*, 2023], making them particularly valuable for immunopeptidomics studies.

6.2 Antibody Sequencing

Antibody sequencing is another key area where *de novo* peptide sequencing is widely used. Unlike DNA-based sequencing, which requires prior knowledge of antibody genes, *de novo* sequencing directly reconstructs the amino acid sequence from mass spectrometry data. This approach is particularly useful for characterizing monoclonal antibodies [Singh *et al.*, 2018], studying immune repertoire diversity, and guiding therapeutic antibody development. By overcoming limitations posed by somatic hypermutation and sequence variability, *de novo* sequencing ensures accurate and high-throughput analysis of antibody sequences. Many *de novo* sequencing tools, such as DeepNovo, Casanovo, InstaNovo and PointNovo, have been widely used in antibody protein sequencing [Beslic *et al.*, 2023].

6.3 Venomics

The study of venom proteins and peptides, known as venomics, benefits significantly from *de novo* peptide sequencing, as many venomous species lack well-annotated genomes. Venom peptides exhibit diverse bioactive properties, including antimicrobial, neurotoxic, and anticoagulant effects, making them valuable for drug discovery and biomedical applications. *De novo* sequencing methods allows researchers to identify and characterize novel venom peptides from various species, facilitating evolutionary studies and the development of venom-derived therapeutics [Saethang *et al.*, 2022].

6.4 Metaproteomics

In metaproteomics, peptides are extracted from, e.g., environmental samples or a gut microbiome, constructing a relevant peptide database is challenging. *De novo* peptide sequencing is thus essential for identifying peptides in the absence of complete reference genomes. This approach is particularly useful in microbiome research, enabling the discovery of novel functional peptides and proteins in environmental, gut, and clinical microbiomes. By bypassing the need for pre-existing protein databases, *de novo* sequencing enhances the ability to study microbial diversity, host-microbe interactions, and ecosystem dynamics at the proteomic level. Many previous works leveage powerful *de novo* peptide sequencing tools such as π-HelixNovo, Casanovo, and SMSNet to conduct sequencing in metaproteomics [Kleikamp *et al.*, 2021] or detect giant genes in bacteria from metaproteomics data [West-Roberts *et al.*, 2023].

7 Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of the *de novo* peptide sequencing models. These metrics help quantify various aspects of the model's effectiveness, reliability, and efficiency. The following are key metrics that are typically used in this evaluation.

7.1 Amino Acid-level Metrics

Amino Acid-level Precision and Recall. The number of matched amino acid predictions, N^{aa}_{match}, is usually defined as the predicted amino acids that exhibit a mass difference of less than 0.1 Da from the ground truth amino acids. Additionally, these predictions must have either a prefix or a suffix with a mass difference of no more than 0.5 Da from the corresponding ground truth amino acid sequence in the ground truth peptide. Amino acid-level precision is then defined as:

$$\text{Amino Acid-level Precision} = \frac{N^{aa}_{\text{match}}}{N^{aa}_{\text{pred}}}, \tag{2}$$

where N^{aa}_{pred} represents the number of predicted amino acids in the predicted peptide sequences. Similarly, amino acid-level recall is defined as:

$$\text{Amino Acid-level Recall} = \frac{N^{aa}_{\text{match}}}{N^{aa}_{\text{truth}}}, \tag{3}$$

where N^{aa}_{truth} represents the number of amino acids in the ground truth peptide sequences.

PTM Precision and Recall. Amino acids with PTMs are specialized amino acids that play a crucial role in biology as these modifications can significantly impact protein structure, activity, and interactions. Accurately identifying PTMs is essential for drug development and biomarker discovery. Similar to amino acid-level metrics, post-translational modifications (PTMs) identification precision and recall can be defined as:

$$\text{PTM Precision} = \frac{N^{ptm}_{\text{match}}}{N^{ptm}_{\text{pred}}}, \quad \text{PTM Recall} = \frac{N^{ptm}_{\text{match}}}{N^{ptm}_{\text{truth}}}, \tag{4}$$

where N^{ptm}_{match} denotes the number of matched PTMs, N^{ptm}_{pred} represents the number of predicted amino acids with PTMs, and N_{truth}^{ptm} refers to the number of PTMs in the ground truth peptide sequence. These metrics provide a detailed evaluation of model performance at the individual amino acid level.

7.2 Peptide-level Metrics

Since the fundamental objective of *de novo* peptide sequencing model is to assign a complete peptide sequence to each spectrum, peptide-level performance serve as the primary quantifier for evaluating the effectiveness of the *de novo* peptide sequencing model. The peptide-level metrics are summarized as follows.

Peptide-level Precision and Recall. A predicted peptide is considered a correct match only if all of its amino acids are matched based on the criteria mentioned in the previous paragraph. In a collection of $N_{\text{truth}}^{\text{peptide}}$ spectra, if a model makes predictions for $N_{\text{pred}}^{\text{peptide}}$ of these spectra and accurately predicts $N_{\text{match}}^{\text{peptide}}$ peptides, the peptide-level precision and recall are:

$$\text{Peptide-level Precision} = \frac{N_{\text{match}}^{\text{peptide}}}{N_{\text{pred}}^{\text{peptide}}}, \tag{5}$$

$$\text{Peptide-level Recall} = \frac{N_{\text{match}}^{\text{peptide}}}{N_{\text{truth}}^{\text{peptide}}}. \tag{6}$$

Peptide Confidence Score. The confidence score is a metric used to evaluate the reliability of predicted peptide sequences when the ground-truth sequence is unavailable. It is computed as the average softmax probability of each predicted amino acid type in the sequence, representing the model's overall confidence in its predictions [Zhou *et al.*, 2024].

Peptide AUC-PR. Given the peptide-level recall, precision, and confidence scores, one effective way to evaluate *de novo* sequencing accuracy is by plotting precision-recall curves and calculating the area under the curve (AUC-PR). This is done by first ranking the predictions from each model based on their confidence scores, from highest to lowest. Starting with the most confident prediction, we accumulate the model's recall and precision values. These accumulated values are then used to plot the precision-recall curve, where precision is represented on the y-axis and recall on the x-axis. The AUC-PR of this curve provides a thorough evaluation of the model's performance across various confidence levels.

8 Conclusions and Future Outlooks

In conclusion, this paper provides a comprehensive overview of *de novo* peptide sequencing methods. We start by reviewing the mass spectral data types, then present or compare the representative models from the perspectives of decoding strategies and model architectures. We also showcase various successful applications of *de novo* peptide sequencing tools in biology. Despite the fruitful progress, there are several areas of improvement and emerging trends that hold promise for the next generation methods. In this section, we discuss potential future directions for research and development in the field.

8.1 Improved Handling of Low-Quality Data

Mass spectrometry data can often be noisy, incomplete, or of low resolution, particularly when analyzing samples with low abundance or complex matrices. Current deep learning models may struggle with such data, leading to inaccurate or incomplete peptide identifications. Future models should incorporate more robust preprocessing and noise-filtering techniques, or perhaps even develop models that are explicitly designed to handle low-quality or noisy spectra. Approaches like data augmentation or self-supervised pre-training could help improve model robustness in such challenging conditions.

8.2 Integration with Other Omics Data

De novo peptide sequencing can benefit from integration with other types of omics data. For example, combining *de novo* peptide sequencing results with transcriptomics data could provide additional context for interpreting peptide sequences, particularly in the case of novel or poorly characterized proteins. Similarly, integrating with metabolic profiling could help identify post-translational modifications (PTMs) or peptide variants that might be difficult to detect from mass spectrometry data alone. Future research should focus on developing multimodal learning frameworks that integrate these various data types to provide more holistic insights into proteomics.

8.3 Real-Time Peptide Sequencing

Currently, deep learning-based *de novo* peptide sequencing typically requires batch processing, which means the peptide identification process happens after the mass spectrometry experiment is complete. For applications in real-time analysis, such as in clinical settings or during live experiments, there is a need for faster, more efficient models capable of delivering peptide sequences in real time. Developing models that can handle streaming data and provide rapid feedback would have significant implications for the pace of scientific discovery and clinical decision-making.

8.4 Exploring Post-Translational Modifications

Post-translational modifications (PTMs) are a critical aspect of proteomics, as they influence protein function, interactions, and localization. *De novo* peptide sequencing, when coupled with deep learning methods, offers the potential to identify and map PTMs directly from mass spectrometry data. However, the complexity of PTM identification remains a significant challenge, as modifications can occur at multiple sites and vary in their fragmentation patterns. Future research will likely focus on developing specialized models that can detect and interpret PTMs alongside peptide sequences, potentially leading to a more comprehensive understanding of protein regulation and function.

Contribution Statement

Jun Xia, Jingbo Zhou, Shaorong Chen, and Tianze Ling contribute equally to this work. Stan Z. Li is the corresponding author.

Acknowledgements

This work was supported by National Natural Science Foundation of China Project No. 623B2086 and No. U21A20427, the Science & Technology Innovation 2030 Major Program Project No. 2021ZD0150100.

References

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

[Bateman *et al.*, 2014] Nicholas W Bateman, Scott P Goulding, Nicholas J Shulman, Avinash K Gadok, Karen K Szumlinski, Michael J MacCoss, and Christine C Wu. Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (dda). *Molecular & Cellular Proteomics*, 13(1):329–338, 2014.

[Beslic *et al.*, 2023] Denis Beslic, Georg Tscheuschner, Bernhard Y Renard, Michael G Weller, and Thilo Muth. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Briefings in Bioinformatics*, 24(1):bbac542, 2023.

[Cao *et al.*, 2022] Wenming Cao, Canta Zheng, Zhiyue Yan, and Weixin Xie. Geometric deep learning: progress, applications and challenges. *Science China. Information Sciences*, 65(2):126101, 2022.

[Chen *et al.*, 2025] Shaorong Chen, Jun Xia, Jingbo Zhou, Lecheng Zhang, Zhangyang Gao, Bozhen Hu, Cheng Tan, Wenjie Du, and Stan Z. Li. Renovo: Retrieval-based \emph{De Novo} mass spectrometry peptide sequencing. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Doerr, 2015] Allison Doerr. Dia mass spectrometry. *Nature methods*, 12(1):35–35, 2015.

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

[Eloff *et al.*, 2023] Kevin Eloff, Konstantinos Kalogeropoulos, Oliver Morell, Amandla Mabona, Jakob Berg Jespersen, Wesley Williams, Sam van Beljouw, Marcin Skwark, Andreas Hougaard Laustsen, Stan J. J. Brouns, Anne Ljungars, Erwin M. Schoof, Jeroen Van Goey, Ulrich auf dem Keller, Karim Beguir, Nicolas Lopez Carranza, and Timothy P. Jenkins. De novo peptide sequencing with instanovo: Accurate, database-free peptide identification for large scale proteomics experiments. *bioRxiv*, 2023.

[Ge *et al.*, 2022] Cheng Ge, Yi Lu, Jia Qu, Liangxu Xie, Feng Wang, Hong Zhang, Ren Kong, and Shan Chang. Deps: an improved deep learning model for de novo peptide sequencing. *arXiv preprint arXiv:2203.08820*, 2022.

[Jin *et al.*, 2024] Zhi Jin, Sheng Xu, Xiang Zhang, Tianze Ling, Nanqing Dong, Wanli Ouyang, Zhiqiang Gao, Cheng Chang, and Siqi Sun. Contranovo: A contrastive learning approach to enhance de novo peptide sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 144–152, 2024.

[Karunratanakul *et al.*, 2019] Korrawe Karunratanakul, Hsin-Yao Tang, David W Speicher, Ekapol Chuangsuwanich, and Sira Sriswasdi. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular & Cellular Proteomics*, 18(12):2478–2491, 2019.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[Klaproth-Andrade *et al.*, 2024] Daniela Klaproth-Andrade, Johannes Hingerl, Yanik Bruns, Nicholas H Smith, Jakob Träuble, Mathias Wilhelm, and Julien Gagneur. Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. *Nature Communications*, 15(1):151, 2024.

[Kleikamp *et al.*, 2021] Hugo BC Kleikamp, Mario Pronk, Claudia Tugui, Leonor Guedes da Silva, Ben Abbas, Yue Mei Lin, Mark CM van Loosdrecht, and Martin Pabst. Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems*, 12(5):375–383, 2021.

[Lee and Kim, 2024] Sangjeong Lee and Hyunwoo Kim. Bidirectional de novo peptide sequencing using a transformer model. *PLOS Computational Biology*, 20(2):e1011892, 2024.

[Li *et al.*, 2023] Ming Li, Ngoc Hieu Tran, Chao Peng, Qingyang Lei, Lei Xin, Jingxiang Lang, Qing Zhang, Wenting Li, Rui Qiao, Haiming Qin, et al. A complete mass spectrometry-based immunopeptidomics pipeline for neoantigen identification and validation. 2023.

[Liu and Zhao, 2020] Zihang Liu and Chunhui Zhao. A residual network for de novo peptide sequencing with attention mechanism. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1165–1170. IEEE, 2020.

[Liu *et al.*, 2015] Fan Liu, Dirk TS Rijkers, Harm Post, and Albert JR Heck. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nature methods*, 12(12):1179–1184, 2015.

[Liu *et al.*, 2023] 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.

[Mao *et al.*, 2023] 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.

[Mayer and Impens, 2021] Rupert L Mayer and Francis Impens. Immunopeptidomics for next-generation bacterial vaccine development. *Trends in microbiology*, 29(11):1034–1045, 2021.

[Nesvizhskii, 2007] Alexey I Nesvizhskii. Protein identification by tandem mass spectrometry and sequence database searching. *Mass Spectrometry Data Analysis in Proteomics*, pages 87–119, 2007.

[Ng *et al.*, 2023] Cheuk Chi A Ng, Yin Zhou, and Zhong-Ping Yao. Algorithms for de-novo sequencing of peptides by tandem mass spectrometry: a review. *Analytica Chimica Acta*, 1268:341330, 2023.

[Petrovskiy *et al.*, 2024] Denis V Petrovskiy, Kirill S Nikolsky, Liudmila I Kulikova, Vladimir R Rudnev, Tatiana V Butkova, Kristina A Malsagova, Arthur T Kopylov, and Anna L Kaysheva. Powernovo: de novo peptide sequencing via tandem mass spectrometry using an ensemble of transformer and bert models. *Scientific Reports*, 14(1):15000, 2024.

[Qiao *et al.*, 2021] 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(5):420–425, 2021.

[Qiu *et al.*, 2025] Zijie Qiu, Jiaqi Wei, Xiang Zhang, Sheng Xu, Kai Zou, Zhi Jin, ZhiQiang Gao, Nanqing Dong, and Siqi Sun. Ranknovo: A universal reranking approach for robust de novo peptide sequencing, 2025.

[Ramazi and Zahiri, 2021] Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021:baab012, 2021.

[Saethang *et al.*, 2022] Thammakorn Saethang, Poorichaya Somparn, Sunchai Payungporn, Sira Sriswasdi, Khin Than Yee, Kenneth Hodge, Mark A Knepper, Lawan Chanhome, Orawan Khow, Narongsak Chaiyabutr, et al. Identification of daboia siamensis venome using integrated multi-omics data. *Scientific reports*, 12(1):13140, 2022.

[Sanders *et al.*, 2024] 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.

[Singh *et al.*, 2018] Surjit Singh, Nitish K Tank, Pradeep Dwiwedi, Jaykaran Charan, Rimplejeet Kaur, Preeti Sidhu, and Vinay K Chugh. Monoclonal antibodies: a review. *Current clinical pharmacology*, 13(2):85–99, 2018.

[Stahlberg, 2020] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

[Tran *et al.*, 2017] 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*, 114(31):8247–8252, 2017.

[Tran *et al.*, 2019] 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(1):63–66, 2019.

[Tran *et al.*, 2020] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Baozhen Shan, and Ming Li. Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence*, 2(12):764–771, 2020.

[VanDuijn *et al.*, 2017] Martijn M VanDuijn, Lennard J Dekker, Wilfred FJ van IJcken, Peter AE Sillevis Smitt, and Theo M Luider. Immune repertoire after immunization as seen by next-generation sequencing and proteomics. *Frontiers in Immunology*, 8:1286, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vitorino *et al.*, 2020] Rui Vitorino, Sofia Guedes, Fabio Trindade, Inês Correia, Gabriela Moura, Paulo Carvalho, Manuel AS Santos, and Francisco Amado. De novo sequencing of proteins by mass spectrometry. *Expert Review of Proteomics*, 17(7-8):595–607, 2020.

[Wang *et al.*, 2023] Ke Wang, Mingjia Zhu, Wadii Boulila, Maha Driss, Thippa Reddy Gadekallu, Chien-Ming Chen, Lei Wang, Saru Kumari, and Siu-Ming Yiu. Seqnovo: De novo peptide sequencing prediction in iomt via seq2seq. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[West-Roberts *et al.*, 2023] Jacob West-Roberts, Luis Valentin-Alvarado, Susan Mullen, Rohan Sachdeva, Justin Smith, Laura A Hug, Daniel S Gregoire, Wentso Liu, Tzu-Yu Lin, Gabriel Husain, et al. Giant genes are rare but implicated in cell wall degradation by predatory bacteria. *BioRxiv*, pages 2023–11, 2023.

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

[Wu *et al.*, 2023] Ruitao Wu, Xiang Zhang, Runtao Wang, and Haipeng Wang. Denovo-gcn: De novo peptide sequencing by graph convolutional neural networks. *Applied Sciences*, 13(7):4604, 2023.

[Xia *et al.*, 2024] Jun Xia, Shaorong Chen, Jingbo Zhou, Xiaojun Shan, Wenjie Du, Zhangyang Gao, Cheng Tan, Bozhen Hu, Jiangbin Zheng, and Stan Z. Li. Adanovo: Towards robust \emph{De Novo} peptide sequencing in proteomics against data biases. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Xia *et al.*, 2025] Jun Xia, Sizhe Liu, Jingbo Zhou, Shaorong Chen, hongxin xiang, Zicheng Liu, Yue Liu, and Stan Z. Li. Bridging the gap between database search and \emph{De Novo} peptide sequencing with searchnovo. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Xu *et al.*, 2023] Xiaofang Xu, Chunde Yang, Qiang He, Kunxian Shu, Yuan Xinpu, Zhiguang Chen, Yunping Zhu, and Tao Chen. Pgpointnovo: an efficient neural network-based tool for parallel de novo peptide sequencing. *Bioinformatics Advances*, 3(1):vbad057, 2023.

[Yang *et al.*, 2022] Yan Yang, Zakir Hossain, Khandaker Asif, Liyuan Pan, Shafin Rahman, and Eric Stone. Dpst: de novo peptide sequencing with amino-acid-aware transformers. *arXiv preprint arXiv:2203.13132*, 2022.

[Yang *et al.*, 2023] Shu Yang, Siyu Wu, Binyang Li, Yuxiaomei Liu, Fangzheng Li, Jiaxing Qi, Qunying Wang, Xiaohui Liang, Tiannan Guo, and Zhongzhi Luan. Biatnovo: An attention-based bidirectional de novo sequencing framework for data-independent-acquisition mass spectrometry. *bioRxiv*, pages 2023–05, 2023.

[Yang *et al.*, 2024] Tingpeng Yang, Tianze Ling, Boyan Sun, Zhendong Liang, Fan Xu, Xiansong Huang, Linhai Xie, Yonghong He, Leyuan Li, Fuchu He, et al. Introducing π-helixnovo for practical large-scale de novo peptide sequencing. *Briefings in Bioinformatics*, 25(2):bbae021, 2024.

[Yates III, 1998] John R Yates III. Database searching using mass spectrometry data. *Electrophoresis*, 19(6):893–900, 1998.

[Yilmaz *et al.*, 2022] 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.

[Yilmaz *et al.*, 2024] Melih Yilmaz, William E Fondrie, Wout Bittremieux, Carlo F Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William Stafford Noble. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature communications*, 15(1):6427, 2024.

[Ying *et al.*, 2021] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

[Zhang *et al.*, 2013] 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.

[Zhang *et al.*, 2025a] Xiang Zhang, Tianze Ling, Zhi Jin, Sheng Xu, Zhiqiang Gao, Boyan Sun, Zijie Qiu, Jiaqi Wei, Nanqing Dong, Guangshuai Wang, et al. π-primenovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nature Communications*, 16(1):267, 2025.

[Zhang *et al.*, 2025b] Xiang Zhang, Jiaqi Wei, Zijie Qiu, Sheng Xu, Zhi Jin, ZhiQiang Gao, Nanqing Dong, and Siqi Sun. Distilling non-autoregressive model knowledge for autoregressive de novo peptide sequencing, 2025.

[Zhou *et al.*, 2024] Jingbo Zhou, Shaorong Chen, Jun Xia, Sizhe Liu, Tianze Ling, Wenjie Du, Yue Liu, Jianwei Yin, and Stan Z. Li. Novobench: Benchmarking deep learning-based \emph{De Novo} sequencing methods in proteomics. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.