Peptide Sequencing with Deep Learning

by

Rui Qiao

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

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

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

External Examiner:	Ting Chen Professor, Tsinghua University
Supervisor(s):	Ali Ghodsi Professor, University of Waterloo
Internal Member:	Mu Zhu Professor, University of Waterloo
Internal Member:	Leilei Zeng Associate Professor, University of Waterloo
Internal-External Member:	Ming Li University Professor, University of Waterloo

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, software development, and implementation of DeepNovoDIA. I conveived the research idea of PointNovo, developed the PointNovo model and software, and performed data analysis. I contributed to the design of the personalized neoantigen identification workflow, implemented the software for and performed data analysis.

Abstract

In shotgun proteomics, *de novo* peptide sequencing from tandem mass spectrometry data is the key technology for finding new peptide or protein sequences. It has successful applications in assembling monoclonal antibody sequences and great potentials for identifying neoantigens for personalized cancer vaccines. In this thesis, I propose a novel deep neural network-based *de novo* peptide sequencing model: PointNovo. The proposed PointNovo model not only outperforms the previous state-of-the-art model by a significant margin but also solves the long-standing accuracy–speed/memory trade-off problem that exists in previous *de novo* peptide sequencing tools. Further, our experiment results show that even though PointNovo is not trained to distinguish between true and false peptide spectrum matching, its resulting log probability score can be used as a scoring function to perform database searching. On several different datasets, we show that PointNovo, when used as a database search engine, can achieve an identification rate that is at least comparable to existing popular database search softwares.

We also extend and adapt an existing model to process Data Independent Acquisition (DIA) data and propose the first *de novo* peptide sequencing algorithm for DIA tandem mass spectra.

Finally, we develop a workflow that can identify tumor-specific antigens directly and purely from mass spectrometry data of tumor tissues and test it on a published dataset of tumor samples from melanoma patients. Our workflow applies *de novo* peptide sequencing to detect mutated endogenous peptides, in contrast to the prevalent indirect approach of combining exome sequencing, somatic mutation calling, and epitope prediction in existing methods. More importantly, we develop machine learning models that are tailored to each patient based on their own MS data. Such a personalized approach enables accurate identification of neoantigens for the development of personalized cancer vaccines. We applied the workflow to datasets of five melanoma patients and expanded their immunopeptidomes by 5% to 15%. Subsequently, we discovered 17 neoantigens of both HLA–I and HLA–II, including those with validated T cell responses and those novel neoantigens that had not been reported in previous studies.

Chapter 3 of this thesis is based on the author's published paper "Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry," *Nature methods* 16.1 (2019): 63–66. Chapter 4 and Chapter 5 are based on the author's preprint paper "DeepNovoV2"[1] (conditionally accepted by *Nature Machine Intelligence*)

[1] https://arxiv.org/abs/1904.08514

and "Identifying neoantigens for personalized cancer vaccines by personalized *de novo* peptide sequencing."[2] (conditionally accepted by *Nature Machine Intelligence*)

[2]https://www.biorxiv.org/content/biorxiv/early/2019/04/26/620468.full.pdf

Acknowledgments

First and foremost, I want to express my gratitude to my supervisor Prof. Ali Ghodsi, who has always encouraged me to explore the problems I am interested in. Your immense knowledge in different research fields has inspired me to keep expanding my horizons and developing new methods. I truly appreciated all your time, patience, ideas, discussions, and advises.

I would like to thank Prof. Ming Li for introducing me to the area of bioinformatics. Your generous advice and guidance have greatly helped me in my research and in the writing of this thesis. My sincere thanks must also go to my collaborators from the Bioinformatics Solution Incorporation: Dr. Ngoc Hieu Tran, Dr. Lei Xin, Dr. Xin Chen, and Dr. Baozhen Shan. Thank you for patiently teaching me concepts in mass spectrometry and lending your expertise to my research.

I thank Dr. Kun Xiong and Dr. Zefeng Zhang for giving me the opportunity to intern at RSVP technologies incorporation. The invaluable hands-on experience I gained there in developing software and natural language processing models is essential to the completion of this thesis.

Knox Waterloo Church greatly enriched my time at Waterloo. I am grateful to have met such a wonderful group of people and appreciate the spiritual support I received from them when I was at a low point in my life.

Finally, I want to thank my loving, encouraging, and supporting wife, Chuyi Liu. This thesis would not have been possible without your presence and support.

Dedication

This thesis is dedicated to my grandmother, Xiuzhen Wen (1943–2019). Your life shall not be forgotten.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction of tandem mass spectrometry and peptide sequencing

1.1 Tandem Mass Spectrometry

Mass spectrometry (MS) is a popular and powerful tool for chemical analysis. It has contributed to different areas of research, including, but not limited to, chemistry, physics, and biochemistry[79]. In MS, samples, typically in liquid or gas form, are loaded into a mass spectrometer which comprises an ion source, a mass analyzer, and an ion detector. The process is shown in Figure 1.1. The ion source produces gas phase ions from the sample being studied, the mass analyzer separates those ions according to their mass-to-charge ratio (m/z), and the ion detector detects the ions and record their relative abundance.

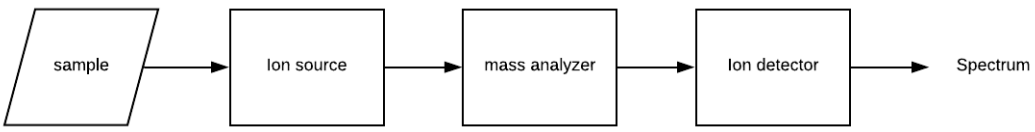


Figure 1.1: Components of a mass spectrometer

Tandem mass spectrometry (MS/MS) is a technique of utilizing two or more different types of mass analyzers to enhance analysis through fragmentation of the input molecules[55]. In this approach, the sample first goes through an ion source, a mass analyzer, and an ion detector, as in MS; the output is called MS1 spectrum. Distinct ions

of interest are then selected and are further fragmented by several different dissociation methods, e.g., collision-induced dissociation (CID) and higher energy collision dissociation (HCD). These fragments are then processed by the second mass spectrometer, and the output is MS2 spectrum.

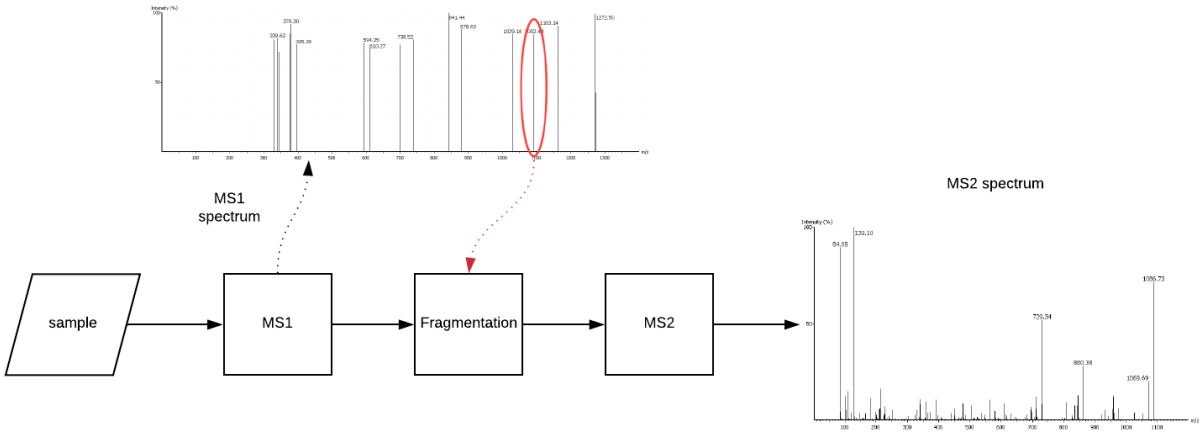


Figure 1.2: Diagram of tandem mass spectrometry

In MS/MS shotgun proteomics (also known as "bottom-up" proteomics), the biological sample (e.g., tumor tissues, plasma, and urine) often contains multiple proteins. First, the sample is pre-processed with certain proteases (e.g., trypsin and Endoproteinase Lys-C), such that the proteins are cleaved into shorter peptides. These peptides are then fed into a spectrometer, and the output is denoted as MS1 spectra. Each signal in an MS1 spectrum is called a precursor ion, which typically represents a certain kind of peptide. Next, the mass spectrometer selects some precursor ions to perform fragmentation based on certain strategies. The two commonly used strategies are Data Dependent Acquisition (DDA) and Data Independent Acquisition (DIA). Their main difference is that only a fixed number of precursor ions are picked in DDA, while in DIA all ions within a certain m/z range are fragmented and analyzed. Next, the selected precursor ions are further processed by the second round of MS and are fragmented into smaller pieces called fragment ions. The final outputs are denoted as MS2 spectra. From the precursor mass (i.e., the mass of the peptide) and the fragment ion signals information contained in the MS2 spectrum, it is possible to recover the exact amino acid sequence of the original peptide. This solution is called peptide sequencing. In this thesis, I focus on the algorithm part of the peptide sequencing problem. I develop novel algorithms for MS/MS data generated through different strategies

and demonstrate real-world applications for identifying tumor-specific antigens with our proposed models.

1.2 Peptide Sequencing

1.2.1 Definitions and Notations

Denote $\mathcal{A} = \{a_1, a_2, \cdots, a_v\}$ as the set of amino acid residues with molecular masses $m(a), a \in \mathcal{A}$. A length n peptide $P = (p_1, \cdots, p_n),\ p_i \in \mathcal{A}$ is a sequence of n amino acids. The mass of a peptide P can be computed using the following formula:

$$m(P) = \sum_{i=1}^{n} m(p_i) + m_{\mathrm{H_2O}} \tag{1.1}$$

where $m_{\mathrm{H_2O}} \approx 18.0106$ Da. A partial peptide $P' \subset P$ is a substring $(p_i, \cdots, p_j)$ of P, where either $i = 1$ or $j = n$, with mass $m(P') = \sum_{i \le l \le j} m(p_l)$. We define the complement of P' as the remaining amino acids of the peptide, denoted by P'_{c}. For example, if $P' = (p_1, \cdots, p_j)$, then $P'_{\mathsf{c}} = (p_j, \cdots, p_n)$ and $m(P'_{\mathsf{c}}) = \sum_{j \le l \le n} m(p_l)$.

The set of possible fragment ion types is denoted as $\Delta = \{\delta_1, \cdots, \delta_k\}$, where each δ_i is a $\mathbb{R} \longmapsto \mathbb{R}$ mass transformation (from the mass of a partial peptide to the theoretical mass over charge value of the fragment ion of this partial peptide) for a specific type of ion. For example, if δ_1 represents the map for b-ion, then $\delta_1(m(P')) = m(P') + m_H$, $m_H = 1.0078$ Da is the mass of a hydrogen atom. A spectrum $\mathbf{S} = \{(m/z_1, I_1), \cdots, (m/z_s, I_s)\}$ is a set of peaks, where each peak is a tuple of mass over charge ratio (m/z) and intensity (i). A partial peptide P' is said to match a spectrum $\mathbf{S}$ if:

$$\exists\ j, l \quad s.t. \quad |\delta_j(m(P')) - m/z_l| < t$$

where t is the fragment ion mass difference tolerance, a parameter related to the resolution of the mass spectrometer. For MS2 spectra from modern mass spectrometers like Orbitrap Fusion, the parameter t is usually set to be between 0.02 Da and 0.05 Da. Given the above notations, the peptide sequencing problem can be defined as: Given the peptide mass m_p, spectrum $\mathbf{S}$, and the set of ion types Δ, finding the peptide of mass m_p with maximal partial peptide matches to spectrum $\mathbf{S}$[19].

In general, existing peptide sequencing algorithms can be classified into three categories: (1) database searching, (2) spectral library searching, and (3) *de novo* peptide sequencing.

1.2.2 Database Searching

In database searching peptide identification, MS2 spectrums are searched against a known database of proteins. Typically, the database of proteins is built from genome information. For each MS2 spectrum, the mass of the precursor ion is retrieved from the corresponding MS1 spectrum. Using the precursor mass, together with the cleavage rule defined by the protease, we can then search the protein database to find all possible peptides whose mass difference with the precursor mass is smaller than a certain threshold, e.g., < 15 parts per million (ppm). Next, the searching algorithm will rank all of these peptide candidates according to a scoring function for peptide-spectrum matches (PSM). The peptide with the highest PSM score is returned as the search result.

An important advantage of database searching peptide sequencing algorithms is that it is easy to control the false discovery rate (FDR) with the target-decoy search strategy[23]. In this strategy, we add some incorrect "decoy" sequences into the search space (i.e., protein database). By counting the number of decoy peptides identified by the searching algorithm, we can then estimate the FDR on the MS/MS dataset. Further, we can filter the identified peptides based on their PSM scores such that only the high confidence ones (e.g., FDR $< 1\%$) are kept. In the past 20 years, many tools and software for database searching have been proposed, including SEQUEST, X!Tandem, Mascot, Comet, MaxQuant, and PEAKS.

1.2.3 Spectral Library Searching

Spectral library searching is another peptide sequencing method that is gaining more and more popularity. This approach requires a peptide spectral library, an annotated collection of MS/MS peptide spectra. For each unidentified MS/MS spectrum, the spectral library searching algorithm will compare its similarity with all spectra in the spectral library. Then, the annotated peptide of the most similar spectrum in the library will be reported as the identification. Compared to the database searching method, spectral library searching algorithms usually run much faster because of the greatly reduced search space. In addition, these algorithms can better discriminate between true and false matches by taking full advantage of spectral features like relative fragment intensities[45]. In general, however, it is trickier to apply the target-decoy quality control strategy for the spectral library searching method. A commonly used method is to create the decoy spectral library by randomly permuting peptides and their corresponding fragment ions in the true spectral library. For some biological samples, especially data acquired by DIA, the spectral library searching method has already shown a higher identification rate than the database searching

method. Popular tools for spectral library searching include SpectraST[45], Spectronaut[9] and OpenSWATH[64].

1.2.4 *De novo* Peptide Sequencing

For both of the aforementioned methods, the FDR of the identification result can be estimated and controlled with the target-decoy search strategy[23]. This makes them popular choices for many applications. On the other hand, since both methods require some level of prior knowledge (e.g., a protein database or a spectral library), they could not search for unknown peptide sequences. An alternative method is *de novo* peptide sequencing, which predicts the peptide sequence directly, and solely from the MS/MS spectrum. It has shown successful results in applications that require finding novel peptides and proteins, such as monoclonal antibody(mAb) assembling[76] and identifying tumor-specific antigens[46].

The basic idea for *de novo* sequencing is simple and straightforward. Suppose we start from the left-hand side of the peptide (N-terminal). The first amino acid is among $\mathcal{A} = \{a_1, a_2, \cdots, a_v\}$. Then, for each a_i we check if there exists $\delta_j \in \Delta$ such that $\delta_j m(a_i)$ matches a peak in the spectrum. If only a_1 has a fragment ion match and all other a_i do not, then we are confident that the first amino acid should be a_1. We can repeat this process to keep predicting the next amino acid residue until the mass of the amino acid sequence matches the precursor mass.

Unfortunately, real-world MS2 spectra are notoriously known for being noisy and incomplete. The incompleteness of spectra means that often we could not detect fragment ions for all partial peptides. On the other hand, MS2 spectrum typically contains many noisy peaks, which means that when predicting the next amino acid, we may find multiple candidates that all match the spectrum, thus making it difficult to decide which is correct. In the past 20 years, different algorithms and technologies have been applied to solve this problems. These include, but are not limited to, the spectrum graph method[19], dynamic programming[14, 51], probabilistic network[28] and hidden markov model[27]. More recently, Tran et al. first introduced deep learning to *de novo* peptide sequencing and proposed DeepNovo, a neural network-based *de novo* peptide sequencing model for DDA MS/MS data that outperforms the previous state-of-the-art models by a large margin[77].

1.3 Contribution

The contribution of this thesis lies in the following aspects: In Chapter 3 we propose the DeepNovo-DIA, the first *de novo* peptide sequencing model for DIA data, and show that our model could efficiently detect more peptides of low abundance. In Chapter 4, we propose a novel *de novo* sequencing model, PointNovo, that outperforms DeepNovo by at least 15%. More importantly, our novel method of spectrum representation not only improves the final peptide accuracy but also solves the accuracy-speed/memory trade-off problem that has long existed in this area. Unlike DeepNovo or previous spectrum graph-based and dynamic programming-based tools, PointNovo can directly benefit from the higher resolution data generated by next-generation mass spectrometers without any increase in computational complexity. Finally, in Chapter 5, we demonstrate an application of finding neoantigens with a personalized *de novo* peptide sequencing model.

Chapter 2

Background

2.1 DeepNovo

Inspired by the success of the image captioning models[88], DeepNovo integrated two fundamental types of neural networks—convolutional neural networks (CNNs) and long short-term memory networks (LSTM)[34]—in order to extract features from both the spectrum and the "language model of peptides." In DeepNovo, each spectrum is represented as a long intensity vector, and CNNs are applied on segments of this vector to extract features and make predictions of the next amino acid. CNNs have been proven effective tools for pattern recognition in different applications, including image classification, object detection, and sentiment analysis[70, 61, 43]. By applying CNNs to the intensity vector, DeepNovo can learn from the noisy spectrum. It is reported that DeepNovo outperformed the past decade's long-standing, state-of-the-art records of *de novo* sequencing algorithms by a large margin of 38.1–64.0% at the peptide level[77]. The structure of DeepNovo is shown in Figure 2.1

2.1.1 Spectrum Representation

DeepNovo discretizes an MS2 spectrum into an intensity vector, in which masses correspond to indices and intensities are the values[77]. In the original code published by the authors[1], the default maximum mass is 3,000 Da, and the default spectrum resolution is 10. This means the intensity vector will be of size 30,000, and every peak within a 0.1 Da bin

[1] https://github.com/nh2tran/DeepNovo

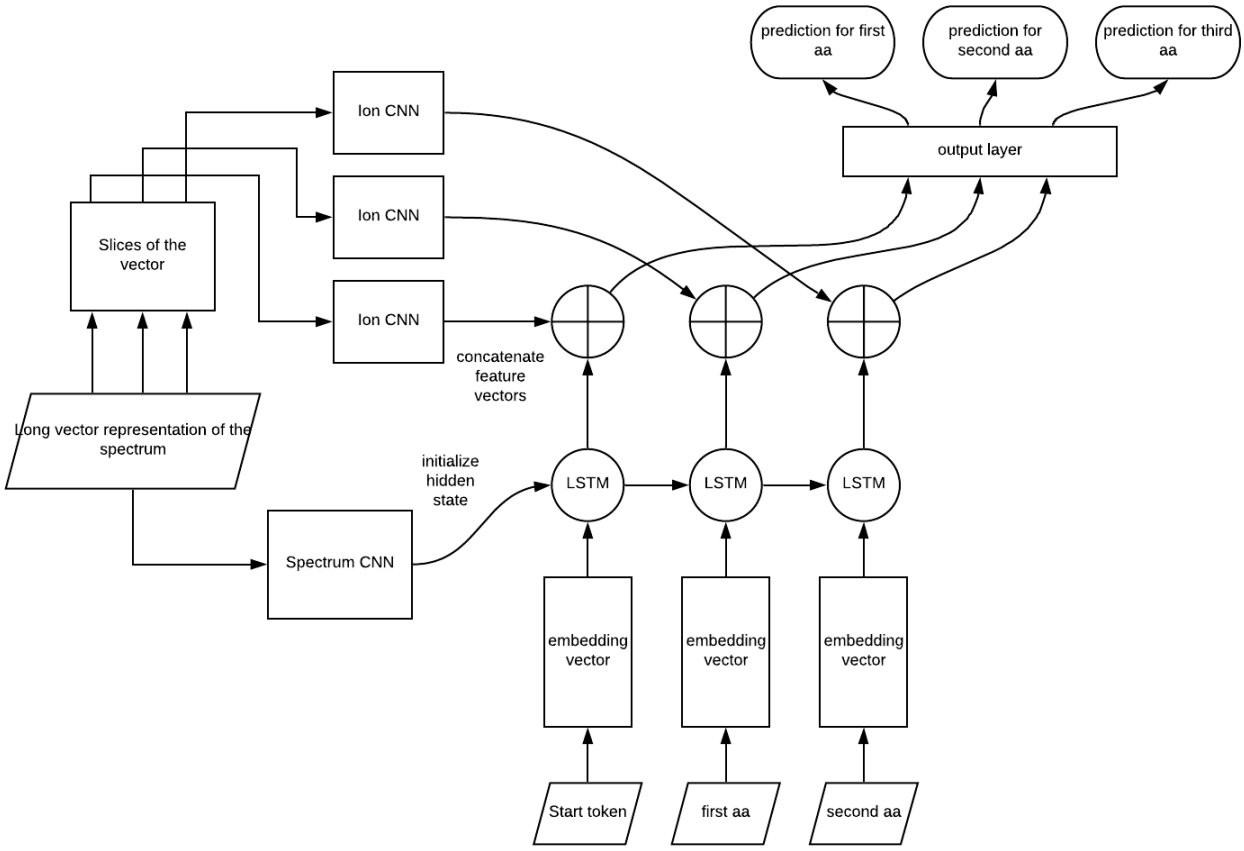


Figure 2.1: DeepNovo

will be merged together and represented as an element of the intensity vector. Figure 2.2 demonstrates the discretization method used by DeepNovo.

This discretization method creates fixed-dimensional vector representations that can be directly processed by CNNs. It is commonly adopted by other neural networks models, such as DeepMatch[66]. However, when merging multiple peaks into a single value, we lose valuable information about the exact mass value of each peak. Also, an MS2 spectrum usually contains only 300 to 1,000 peaks, which means a spectrum could be stored on the disk as 2,000 float numbers. But the intensity vector method needs 30,000 float numbers to represent a spectrum. When experiment scientists build a more accurate mass spectrometer in the future, the DeepNovo model needs to increase the spectrum resolution to take

advantage of the improved accuracy. This will result in a significantly longer intensity vector and will require more memory and time to train the model.

2.1.2 Ion CNN

By default, DeepNovo has a vocabulary of size $v = 26$, which consists of 20 amino acid residues, 3 post-translational modification (PTM) residues, and 3 special tokens: "start", "end" and "padding". When predicting sequences from MS2 spectra, DeepNovo starts with a "start" token, then predicts one token at a time, step by step, until an "end" token is encountered. At each step of prediction, DeepNovo uses an Ion CNN module to extract features from spectrum's vector representation. By default, DeepNovo includes eight fragment ions types: b, y, b(2+), y(2+), b-H_2O, y-H_2O, b-NH_3, and y-NH_3. I denote these eight ions as $\Delta = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8\}$. The transform of each δ is defined by the following formula:

$$
\begin{aligned}
\delta_1(m(P')) &= m(P') + m_H \\
\delta_2(m(P'_c)) &= m(P'_c) + m_H + m_{\mathrm{H_2O}} \\
\delta_3(m(P')) &= \frac{m(P') + 2 \times m_H}{2} \\
\delta_4(m(P'_c)) &= \frac{m(P'_c) + 2 \times m_H + m_{\mathrm{H_2O}}}{2} \\
\delta_5(m(P')) &= \delta_1(m(P')) - m_{\mathrm{H_2O}} \\
\delta_6(m(P'_c)) &= \delta_2(m(P'_c)) - m_{\mathrm{H_2O}} \\
\delta_7(m(P')) &= \delta_1(m(P')) - m_{\mathrm{NH_3}} \\
\delta_8(m(P'_c)) &= \delta_2(m(P'_c)) - m_{\mathrm{NH_3}}
\end{aligned}
\tag{2.1}
$$

Suppose the model already predicted an amino acid sequence $(p_1, \cdots, p_j)$. The next amino acid could be one of the 26 tokens. This then leads to 26 potential partial peptide $\{P'_i\}$, $i \in \{1, 2, \cdots, 26\}$, where $P'_i = (p_1, \cdots, p_j, a_i)$. For each candidate partial peptide P'_i, we can then calculate the m/z values of its eight ion types by Equation 2.1. In total there are 26×8 m/z values of interest.

The inputs of Ion CNN are thus slices of the intensity vector around these locations of interest. To be precise, for each m/z value of interest, DeepNovo identifies its corresponding location index (e.g., as shown in Figure 2.2, the location index of 205.0 Da is 2051), then

extracts a short sub-vector of length 10 around it. Thus, the input of Ion CNN is a 3-dimensional array (denoted by X) of shape $26 \times 8 \times 10$.

The Ion CNN module consists of two 2d-convolutional layers and one fully connected layer. Rectified linear unit (ReLU) is used as the activation function[54]. In the convolution operations, the first dimension of X is treated as the dimension of channels. The convolutional kernel sizes for the second and third dimensions of X are 1 and 3, respectively. Eventually, the Ion CNN module transforms an input X into a feature vector F_{ion} of length 512.

2.1.3 LSTM and Spectrum CNN

DeepNovo integrates an LSTM module to learn the sequence patterns of peptides. At each step of prediction, the previous amino acid will be embedded into a 512-dimensional vector. The LSTM module then outputs a feature vector F_{lstm} conditioned on its hidden state and the embedded vector. To achieve a meaningful prediction, the LSTM module should be initialized with information from the original spectrum. DeepNovo uses a spectrum CNN, which consists of a max pooling layer followed by two convolutional layers and a fully connected layer, to extract features from intensity vectors and then uses the extracted features as the initial states of the LSTM module.

As for the final prediction of the next amino acid, DeepNovo concatenates F_{ion} and F_{lstm} into a feature vector F of length 1024 and applies a fully connected layer with 26 hidden neurons and softmax activation on F. The output can be then viewed as a probability distribution over the 26 tokens.

2.1.4 Training and Searching

DeepNovo uses cross-entropy (CE) loss as the loss function. During training, a forward model (predicting from the left-hand side, or N-terminal, of the peptide) and a backward model (predicting from the right-hand side, or C-terminal, of the peptide) are trained together. Both models are trained with the Adam optimization algorithm[44]. After each 500 training steps, the CE loss on validation is calculated and the weight matrix that has the smallest validation loss is selected for testing.

During prediction, the knapsack dynamic programming algorithm is applied to reduce the search space. The beam search algorithm is applied to search for the best amino acid sequence within a reasonable time. By default, the beam size is set to be 5. The

forward model and backward models each give a predicted peptide, and the one with the highest score, defined as length normalized log probability, will be reported as the identified sequence.

2.1.5 Result

DeepNovo is the first deep neural network-based *de novo* sequencing model. It is reported to achieve a 7.7–22.9% higher accuracy at the amino acid level and 38.1–64.0% higher accuracy at the peptide level. Additionally, in the application of mAb assembling, DeepNovo is shown to be capable of reconstructing the complete sequences of light and heavy chains of a mouse antibody without assisting databases[77].

2.2 Therapeutic Cancer Vaccines

Typically, vaccines are made from weakened or harmless versions of the disease-causing microorganism. After being injected into the human body, vaccines stimulate the body's immune system and provide the recipient an active acquired immunity to the particular pathogen. Various preventive vaccines (e.g., flu shots, HPV vaccine) have been used to provide the infected population immunities against contagious diseases. In the meantime, vaccines could also be adopted to treat existing disease. These are referred to as therapeutic vaccines. Therapeutic cancer vaccines are a type of cancer immunotherapy that aims to treat existing tumors.

In the human immune system, the major histocompatibility complex (MHC) brings short peptides to the surface of cells. Often referred to as MHC peptides (or HLA peptides, where HLA is the gene complex that encodes MHC), these peptides are produced from digested proteins that are broken down in the proteasomes. When a cell is infected by a virus or grows malignant, non-self MHC peptides (from the virus or mutated proteins) will be presented on the surface of the cell so that T cells can recognize and subsequently kill the cell[7]. The MHC peptides on cancer cells (also referred to as antigens) are usually the main component of cancer vaccines. Choosing effective antigens is the single most important step in designing a cancer vaccine. Ideally, the antigens should be expressed specifically by cancer cells (not in normal cells), presented on all cancer cells, and elicit strong immune response[35].

Many cancer vaccines have taken aim at tumor-associated antigens (TAAs), which are abnormally expressed self-proteins. For some types of cancers, the same TAAs are observed

among different patients. With shared TAAs, therefore, it is possible to develop "off-the-shelf" vaccines that are ready-to-use for any eligible cancer patients[73]. However, several challenges remain for developing vaccines against TAAs. Since TAAs are self-antigens, the immune system may develop immune tolerance against the lymphocytes that strongly recognize these peptides. Thus, a cancer vaccine must break the tolerance by using adjuvants, co-stimulators, or repeated vaccination. The expression of TAAs in normal cells may also lead to collateral damage. For example, a recent clinical trial of receptor-engineered T cell therapy (CAR-T) targeting a shared TAA (the colorectal carcinoembryonic antigen) caused severe colitis in a high percentage of patients, as this antigen is also expressed in normal intestinal tissue[57].

Tumor-specific mutated antigens (often referred to as neoantigens) represent another type of target for therapeutic cancer vaccines. Unlike TAAs, neoantigens are mutated non-self peptides that are not found in normal tissues. Thus, vaccines against neoantigens have better specificity and are less likely to elicit collateral damage. Multiple independent clinical trials have confirmed the potential of personalized neoantigen vaccines for patients with melanoma[56, 65, 13]. More recent research shows that neoantigen vaccination is a feasible therapeutic strategy, even for immunological cold tumors with a relatively low mutational burden (e.g., glioblastoma)[42]. The vaccination was found to generate prominent T cell responses against immunizing neoantigens. In addition to developing cancer vaccines, neoantigens help to improve the prediction of response to immune checkpoint inhibitor therapies. Indeed, numerous studies have shown that the response to immune-mediated therapies often correlates with high numbers of identified neoantigens[35].

Since neoantigens carry mutations unique to each patient, identifying them requires a personalized approach. The current prevalent approach is a proteogenomics method that typically involves some of the following steps:

- Sequencing the genome for both tumor and normal tissues of the patient (by whole exome sequencing and RNA sequencing)
- Identifying mutations with somatic variant calling
- Predicting the likelihood of a mutated peptide to be a neoantigen with pMHC binding affinity prediction tools(e.g. NetMHCpan[38], Edge[11])
- Validating with MS

This method has been proven efficient and feasible in multiple studies[6, 52, 42]. It is capable of finding neoantigens that derived from non-synonymous single-nucleotide variants

(SNVs). However, recent research suggests that the majority of neoantigens might be peptides translated from the non-coding region[46] or resulting from alternative splicing[71, 26]. In both cases, the neoantigens are hard to detect by the aforementioned proteogenomics method. A workflow that can identify a neoantigen, regardless of whether it is from SNV, a non-coding region, or alternative splicing, is desirable.

In Chapter 5 we propose a novel pipeline to detect neoantigens by personalized *de novo* sequencing models. Since our method focuses on the MS data, neoantigens from all sources can be identified. Also, by training separate *de novo* sequencing models for each patient, our method could take advantage of the T cell epitope recognition patterns with respect to the patient's specific HLA alleles.

2.3 Order Invariant Networks

Deep neural networks (DNNs, often referred to as deep learning) have gained tremendous attention in recent years. DNNs have set new records on multiple supervised learning tasks and have become the tool of choice in different areas, including, but not limited to, image classification[32, 37], object detection[61, 47], machine translation[5, 81], and speech recognition[33]. Aside from doing end-end training with the gradient-based method on labeled data, DNNs can also serve as a powerful feature detector and be trained with reinforcement learning (RL) algorithms. By combining deep learning with RL, researchers are now able to propose solutions to some difficult and important research problems that could not be solved efficiently before, such as Go playing[68], drug discovery[91], and protein folding[25].

2.3.1 Carefully Designed Model Leads to Better Performance

The success of DNNs is often explained by the representational power that comes with their immense number of parameters. It is widely known that a simple feed-forward network with a single hidden layer containing a finite number of neurons could approximate any continuous functions on compact subsets of $\mathbb{R}^n$ under mild assumptions on the activation function[18]. However, we currently lack theoretical results about the learnability of DNNs with different structures. In practice, researchers often find that given the same number of parameters, a fully connected feed-forward network tends to perform worse than a carefully hand-crafted network structure. Indeed, many successful applications of DNNs rely heavily on some specific types of neural networks. For example, CNNs are crucial

components in almost all DNNs models that process image data. CNNs are inspired by the biological processes on the animal visual cortex[29] and have the desired characteristics of local connectivity and translation invariance. These properties make them a perfect choice for image processing models. Another example would be the attention mechanism in neural machine translation. Based on the observation that an encoder-decoder model deteriorates rapidly as the length of an input sentence increases, Bahdanau et al. proposed to store a vector representation for each input word and let the model jointly learn to align and translate. This method is often referred to as (soft) attention mechanism. It has been shown that by applying the attention modules, models based DNNs could perform significantly better with longer sentences[5, 48]. Thus, the attention mechanism becomes a standard component for different models in natural language processing[86, 20, 60].

2.3.2 T Net

A point cloud is a set of points in a metric space. Point clouds could be generated by 3D scanners and are used for different purposes including 3D computer-aided design modeling and quality control for 3D printing. Figure 2.3 shows a sample point cloud for a 3D ball.

Due to the irregular format of point clouds, most previous research transformed such data to regular 3D voxel grids and applied 3D CNNs to process them. In 2017, Qi et al. proposed PointNet, the first order invariant neural network structure for set data such as point clouds[59]. PointNet was reported to be highly efficient and effective. In the task of point cloud classification, PointNet outperforms the previous state-of-the-art methods by a significant margin.

The building block of PointNet is a structure called T Net. In essence, PointNet is a stack of T Net modules and matrix multiplication operations. The structure of T Net is shown in Figure 2.4.

Here, N denotes the number of data points in a point cloud, $D = H_0$ is the dimension of input data, and H_i represents the number of hidden neurons in a layer. The shared MLP denotes a simple matrix multiplication followed by a non-linear activation function. Suppose we denote the input N by H_i matrix as X, then a shared MLP layer is:

$$f(X) = \sigma(XW) + b$$

where W is a H_i by H_{i+1} matrix of trainable parameters. b is scalar, also a trainable parameter, and σ represents a non-linear activation function. The fully connected layer operates similarly, except that it expects a length H_i row vector $\mathbf{v}$ as input:

$$f(\mathbf{v}) = \sigma(\mathbf{v}W) + \mathbf{b}$$

Here, W is a H_i by H_{i+1} matrix and $\mathbf{b}$ is a vector of length H_{i+1}. The global max pooling operation takes in a matrix and returns the maximum along its first axis. The inclusion of a global max pooling operation guarantees that exchanging the orders of input data points would not affect the output. In popular deep learning frameworks like Tensorflow[1] and PyTorch[58], the shared MLP module can be implemented as a 1D convolution layer and the global max pooling operation can be implemented by the maximum function. Therefore we can build the T Net structure efficiently with built-in functions provided by these frameworks.

Qi et al. reported that PointNet could obtain on par or better results than the previous state-of-the-art methods on a number of 3D recognition tasks, including object classification, part segmentation, and semantic segmentation. The success of PointNet once again demonstrates that it is meaningful to tailor a model structure for a specific type of data. By integrating domain-specific expertise into the neural networks structure, researchers can expect to see a potential improvement in performance.

Here I note that an MS2 spectrum is similar to a point cloud. Indeed, a spectrum is a set of data points where each data point is a peak with two attributes: m/z value and intensity. Therefore, I believe order invariant network structures like T Net have great potential for analyzing mass spectra.

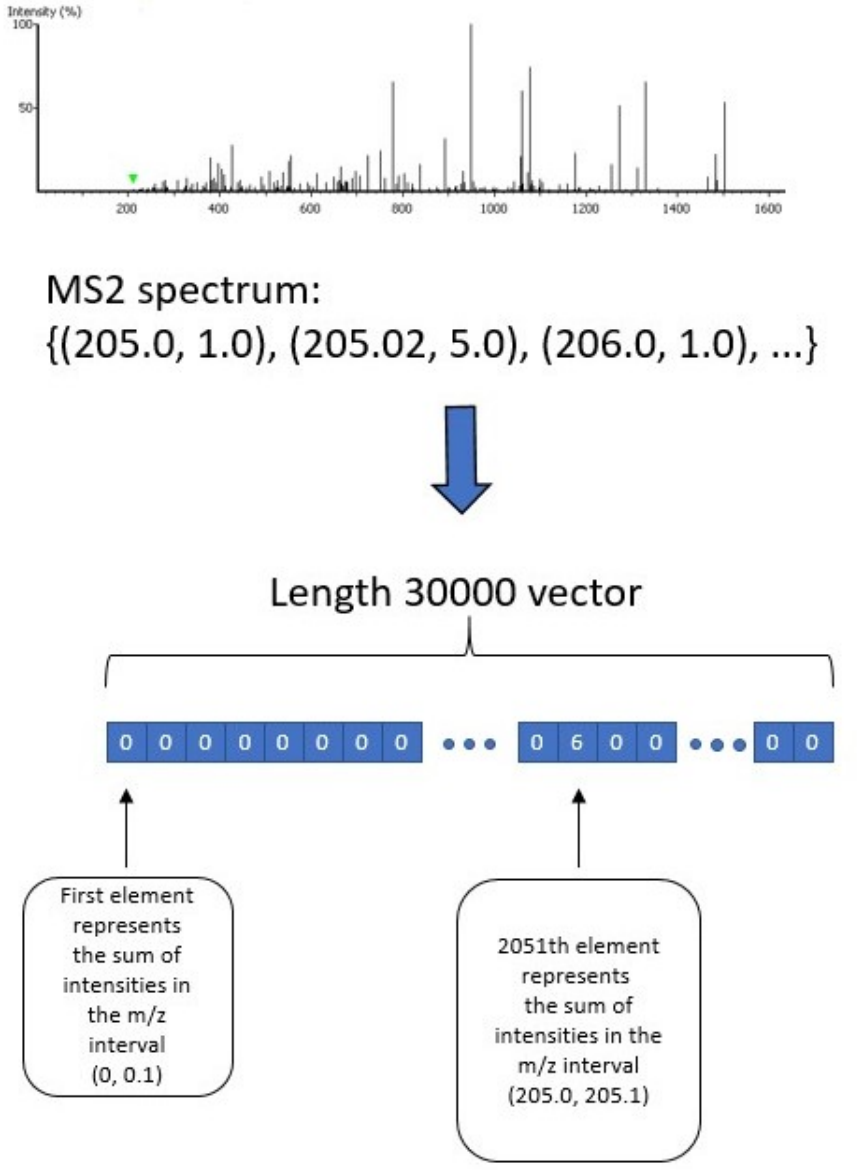


Figure 2.2: Spectrum representation in DeepNovo

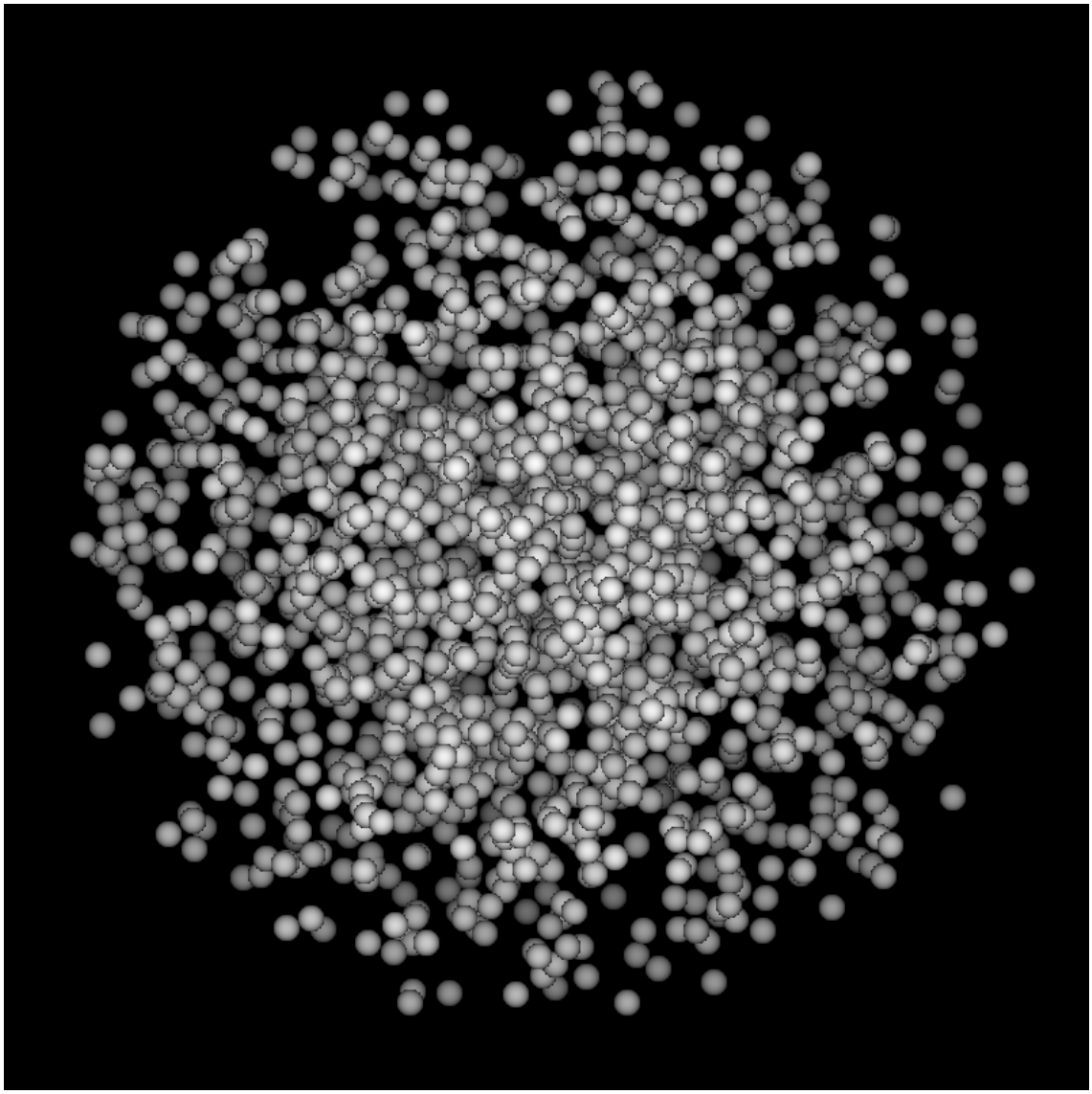


Figure 2.3: Point clouds for a 3D ball

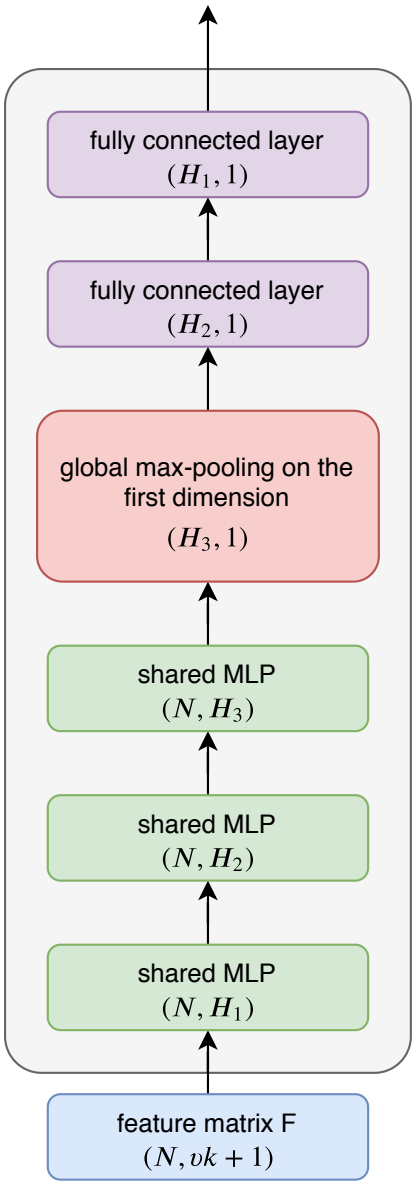


Figure 2.4: Structure of T Net. The output shape is annotated below each block.

Chapter 3

DeepNovo-DIA

As described in Section 1.1, there are two general strategies for selecting features to do second round MS: DDA and DIA. Comparing to DIA, DDA usually generates MS2 spectrums with higher signal-to-noise ratios, i.e. fewer noisy peaks. In DDA, however, only a proportion of the precursor ions appeared in MS1 spectrum will be selected to do further fragmentation. This limits the number of peptides that could be identified in the biological samples. Recently, advances in DIA strategy allow the fragmentation of all precursor ions within a certain range of m/z and retention time in an unbiased and untargeted fashion[82, 64], which means, DIA experiments could produce a complete record of all peptides that are present in a sample, including those with low abundance. This is an important property for applications like personalized immunotherapy[56, 65, 30]. Because in those applications the peptides of interest like TSAs are often of low-abundance.

A remaining question is how to decode these data to extract meaningful information. MS/MS spectra from DIA are notoriously hard to interpret because they are highly multiplexed. Each spectrum contains fragment ions from multiple precursor ions, and the link between a precursor ion and its fragment ions is unknown. This challenge prevents many DIA database search engines from achieving identification power comparable to that of their DDA counterparts[64, 22, 78, 74]. The problem is even more acute for the *de novo* sequencing approach, and to the best of our knowledge, no method has been proposed to address it before.

Recently Tran et al. proposed DeepNovo, a deep-learning-based model for *de novo* sequencing using DDA data. We have observed that, in contrast to many complicated optimization algorithms, the iterative sequencing framework of DeepNovo makes it possible to extend to DIA without any increase in complexity. More importantly, to address the

problem of highly multiplexed spectra, we restructure the neural networks to utilize the extra dimensionality of DIA data (m/z and retention time) to identify coeluting patterns of a precursor ion and its fragment ions, as well as fragment ions across multiple neighbor spectra. This evidence allows DeepNovo-DIA to pick up the correct signal for *de novo* sequencing amid a large amount of noise in a DIA spectrum. Taking all these considerations into account, we designed DeepNovo-DIA to enable *de novo* sequencing using DIA data.

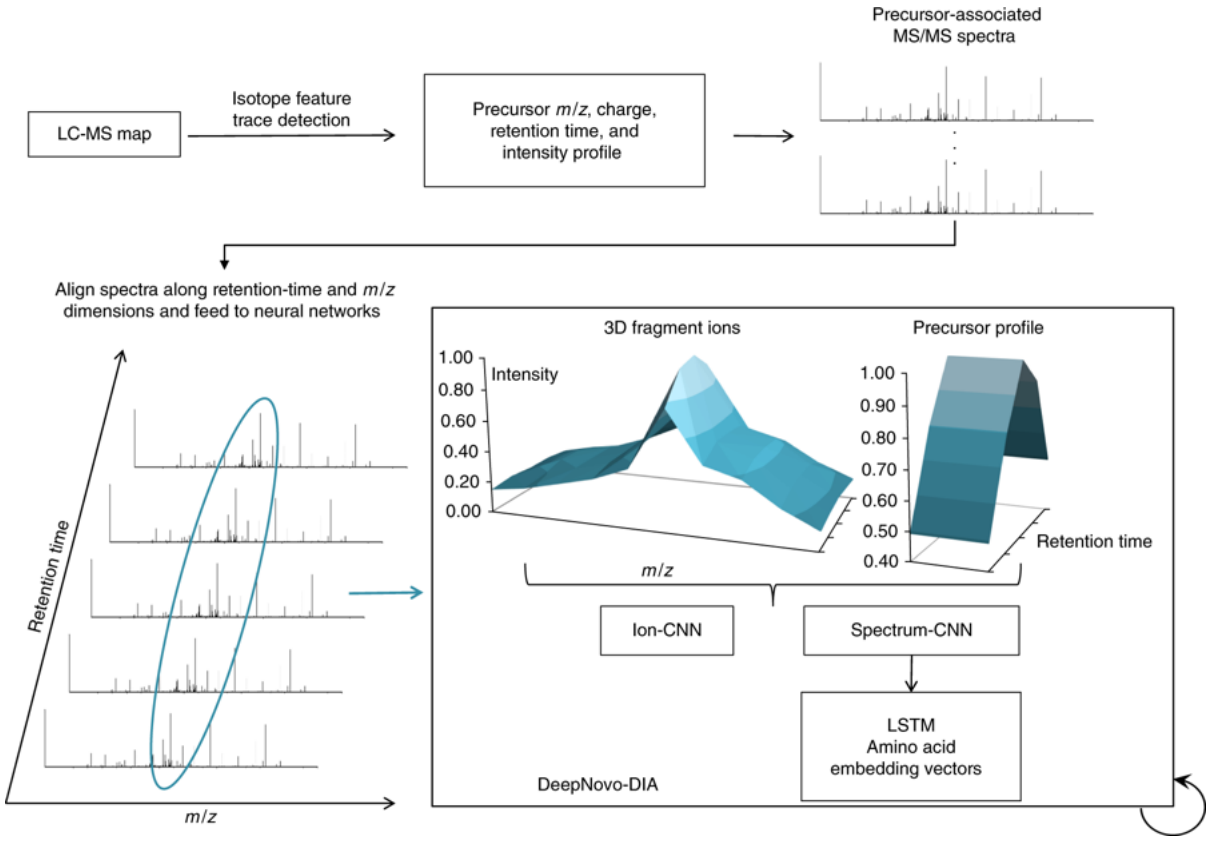


Figure 3.1: The workflow of DeepNovo-DIA for *de novo* sequencing of DIA data.

Our DIA *de novo* sequencing workflow is shown in Figure 3.1. First, precursor features are detected together with their m/z, charge, retention time, and intensity profile. Next, for each precursor, we collect all MS2 spectra so they are within the precursor's retention-time range and ensure that their DIA m/z windows cover the precursor's m/z. Because the number of spectra collected for a precursor may vary, we select a fixed number of spectra that are closest to the center of the precursor's retention time. The closer a spectrum is to the center, the stronger its fragment ion signals are for *de novo* sequencing. The correlation

between the precursor's intensity profile and its fragment ions is also a good indicator for *de novo* sequencing. Thus, we feed the precursor and its associated MS/MS spectra into DeepNovo-DIA neural networks to learn (1) the 3D shapes of fragment ions along m/z and retention-time dimensions, (2) the correlation between the precursor and its fragment ions, and (3) the peptide sequence patterns. Similar to DeepNovo, Our *de novo* sequencing framework operates in a recurrent and beam-search fashion: at each iteration, the model predicts the next amino acid by conditioning on the output of previous steps and keeps track of only a constant number of top candidate sequences. As a result, its complexity does not increase with the number of peptides or with the number of ions in the spectrum.

3.1 Method

Because a DIA spectrum is highly multiplexed, it is important to use high resolution to distinguish fragment ions from different precursors that happen to have similar masses. In DeepNovo-DIA, we used 50 bins to represent 1.0 Da, that is, a spectrum resolution of 50. We also defined a maximum mass value of 3,000 Da. Thus, each spectrum was represented by a vector of length 150,000, in which the mass of an ion corresponded to an index and the ion intensity was the vector value at that index. For the retention-time dimension, we fixed this number and selected those spectra closest to the feature's retention-time mean. If there were not enough spectra, we appended zeros. In this study, we used five spectra (the use of ten spectra led to minor improvements). We stacked the spectra along the retention-time dimension so that the middle one was the closest to the feature's retention-time mean. The five selected MS/MS spectra of a feature were stored in a matrix of size $5 \times 150{,}000$. To normalize the intensities, we divided the matrix element-wise by its maximum. We also extracted the MS1 intensity profile of a given feature at the respective retention times of those five MS/MS spectra. The resulting normalized $5 \times 150{,}000$ matrix, together with the length 5 MS1 intensity profile vector were then fed to the DeepNovo-DIA model for *de novo* sequencing.

In general, the *de novo* sequencing framework is the same for DDA and DIA data, except that extra preprocessing is needed to add the retention-time dimension of DIA data. The DeepNovo-DIA model structure is illustrated in Figure 3.2. At each step, the input to Ion CNN module is now a four dimensional array X of shape $26 \times 8 \times 5 \times 10$. Thus we change the 2D convolution operation in DeepNovo to 3D convolution. The first dimension of X is still viewed as the "channel" dimension. To make use of the MS1 intensity profile information, we also compute its correlation with fragment ions. The output is a vector of length 26 and is concatenated with the 512 dimensional feature vector returned

by the convolutional operations. The structure of Ion CNN in DeepNovo-DIA is shown in Figure 3.3. As for the Spectrum CNN module, we simply change the 2D convolution to 3D.

Previously, DeepNovo used cross-entropy loss as the loss function. For DIA, the presence of multiple peptides in the same spectrum inspired us to view *de novo* sequencing as a multi-label classification problem with dense signals, and hence to apply focal loss[47] as the suitable objective function for DIA. Our experiment shows that the switch to focal loss improved DeepNovo-DIA's performance considerably. Lin et al. proposed focal loss to solve the class-imbalance issue in object detection[47]. The focal loss down-weights the contribution of easy predictions and puts more focus on hard predictions, and therefore could help to address the problems of noisy targets and class imbalance. In object-detection problems, the neural networks need to classify whether a patch of an image is an object or background. Because of the nature of this problem, most patches neural networks can see are background, and this causes problems for end-to-end training with cross-entropy loss. To deal with this problem, Lin et al. proposed a dynamically scaled cross-entropy loss that they named focal loss. For a binary classification problem, we denote $y \in \{0, 1\}$ as the ground-truth class for a data point, and p as the model's predicted probability for class 1. Then the focal loss is defined by the following formula:

$$\text{Focal Loss} = -(1 - p_t)^{\gamma} \log(p_t) \tag{3.1}$$

where $p_t = p$ if y is class 1 and $p_t = 1 - p$ if y is class 0. And γ is a hyperparameter greater than 1.

From the definition, we can see that, compared with cross-entropy loss, focal loss scales down the loss by a factor of $(1 - p_t)^{\gamma}$. This means that focal loss down-weights the contribution of easy examples (where $1 - p_t$ is small), and the model is likely to focus more on hard examples.

In our case, we found that the DeepNovo-DIA model also had a class-imbalance problem, as the frequency for amino acids varies a lot. Therefore, we suspected that focal loss could help us to better train the DeepNovo-DIA model. During training, we changed the activation function of the last layer from a softmax function to a sigmoid function, which led the model to give a probability between 0 and 1 for each of the 26 classes (note that here the sum of these 26 probabilities might not amount to 1). Then, for each class we computed the focal loss using the formula above, and used the average of those 26 losses as the final loss. At inference time, we switched the activation function back to softmax because we found that this led to better performance. Overall, our experiments show that the focal loss improved the amino acid accuracy by 20% on the plasma dataset.

3.2 Results

We trained DeepNovo-DIA on a previously obtained dataset of urine samples from 64 subjects[53]. We evaluated DeepNovo-DIA on two other datasets from different subjects who had been diagnosed with ovarian cyst (OC; six subjects) or urinary tract infection (UTI; six subjects). We also tested DeepNovo-DIA on a previously obtained dataset of plasma samples[74]. The test datasets were not used during model development.

We built an in-house database search tool to generate training data. In particular, we followed the approach of DIA-Umpire[78] to generate a pseudo-spectrum from each precursor feature and its associated spectra. Then we used a conventional DDA database search tool, PEAKS DB[90], to search the pseudo-spectra against the Swiss-Prot human database. The peptides identified at 1% FDR were assigned to the corresponding precursors and were used as ground-truth labels for training. Our training set included 2,177,667 spectra, 202,114 labeled precursor features, and 14,400 unique peptides. For evaluation, we compared DeepNovo-DIA to DIA database search tools including PECAN[74], Spectronaut[10], and OpenSWATH[64]. Such comparisons illustrate (1) the accuracy of *de novo* sequencing (based on overlapping identifications) and (2) DeepNovo-DIA's identification of new peptides not found in the database.

We first calculated the accuracy of DeepNovo-DIA using labeled features from the in-house database search. For each labeled feature, we compared the *de novo* peptide predicted by DeepNovo-DIA with the ground-truth sequence on the basis of the alignment of their mass fragments. We measured the sequencing accuracy at the amino acid level (i.e., the ratio of the total number of matched amino acids to the total length of predicted peptides) and at the peptide level (i.e., the fraction of fully matched peptides). As shown in Figure 3.4a, DeepNovo-DIA accurately predicted 63.8–68.1% of amino acids and 37.4–52.4% of peptides of the labeled features. Moreover, DeepNovo-DIA provides a confidence score for each predicted amino acid. Figure 3.4b shows the distribution of sequencing accuracy with respect to confidence score that allows one to select high-confidence *de novo* peptides with a certain expected accuracy.

We then applied DeepNovo-DIA to all features, labeled and unlabeled, and used the confidence-score distribution in Figure 3.4b to select high-confidence predicted peptides with an expected sequencing accuracy of 90%. Figure 3.4c shows the substantial overlap of precursor features with peptide identifications by the database search and DeepNovo-DIA. The amino acid accuracy of overlapping features was close to 90%, as expected (Figure 3.4d), thus demonstrating the reliability of the DeepNovo-DIA confidence score for quality control. More important, DeepNovo-DIA identified peptides for 33.0–72.6%

of extra features (e.g., plasma dataset 33% = 2529/(4207 + 3466)). We also observed that DeepNovo-DIA's performance was better for the UTI and OC datasets than for the plasma dataset; we suggest this is because the UTI and OC datasets were more similar to the training data.

Next, we compared DeepNovo-DIA to PECAN and Spectronaut, using the plasma dataset[74]. DeepNovo-DIA correctly predicted the full sequences of 1,023 database peptides that were reported by PECAN or Spectronaut (Figure 3.4e). Among 2,091 peptides reported by both PECAN and Spectronaut, which can be considered as high-quality database search results, DeepNovo-DIA identified 778 (37.2%). This is comparable to the performance of *de novo* sequencing tools for DDA data (25-40% at the peptide level[77]). Among peptides reported only by DeepNovo-DIA, 587 could be found in the database and 2,011 were *de novo*. To ensure that the *de novo* peptides were supported by significant peptide–spectrum matches, we augmented the database FASTA file with the *de novo* peptides and re-ran the in-house database search. We found that 1,143 *de novo* peptides passed 1% FDR after the search was re-run. Thus, 1,730 peptides were identified only by DeepNovo-DIA.

Finally, we show an example of DeepNovo-DIA's application to a DIA spectrum from the plasma dataset that contained mixed fragment ions from three different peptides (3.4g-i). DeepNovo-DIA was able to identify all of them. The last two peptides were predicted by both DeepNovo-DIA and the database search; however, the first one did not exist in the database. Thus, the combination of DIA and *de novo* sequencing has the potential to help scientists discover novel peptides and enable more complete profiling of biological samples.

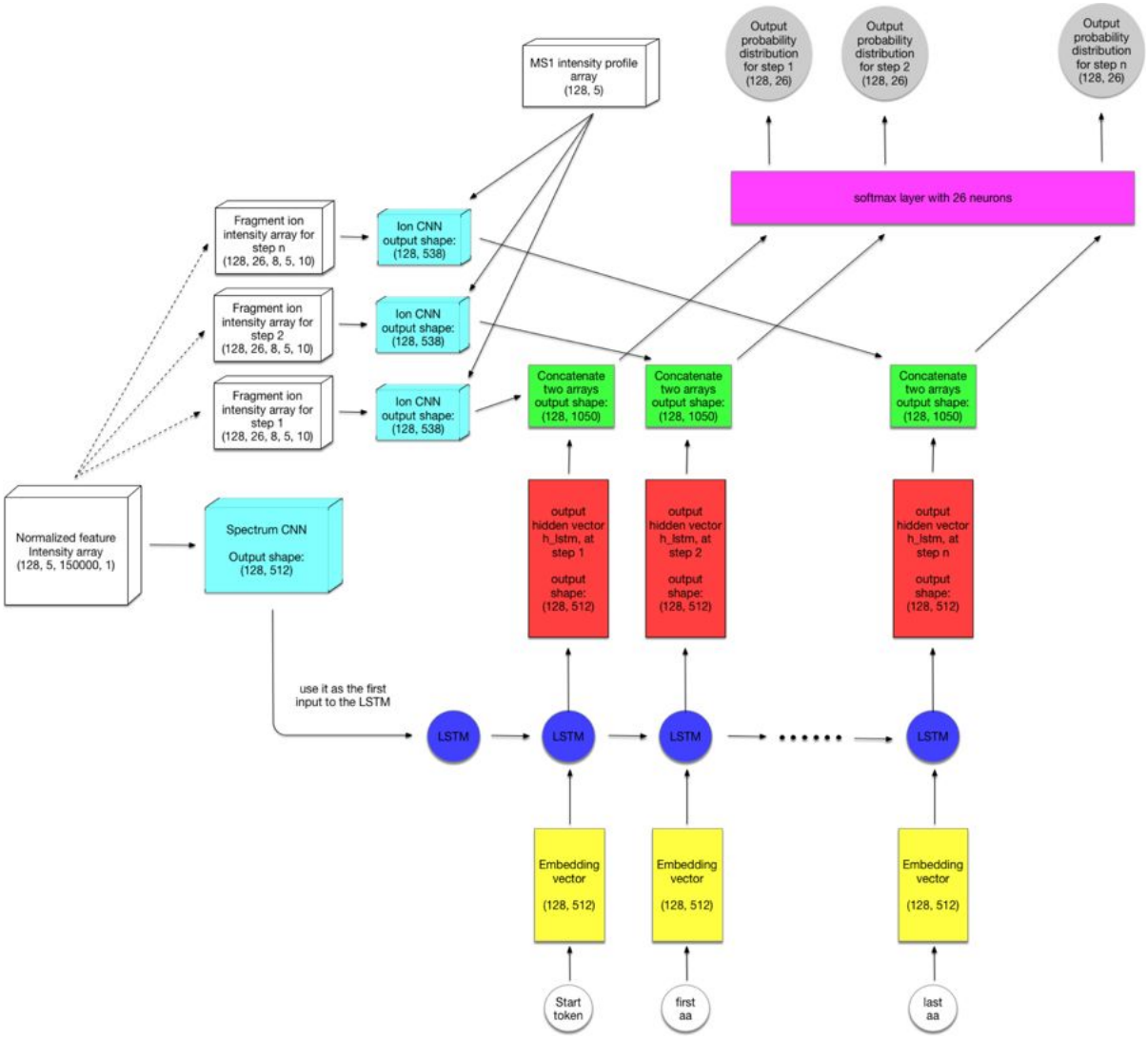


Figure 3.2: DeepNovo-DIA

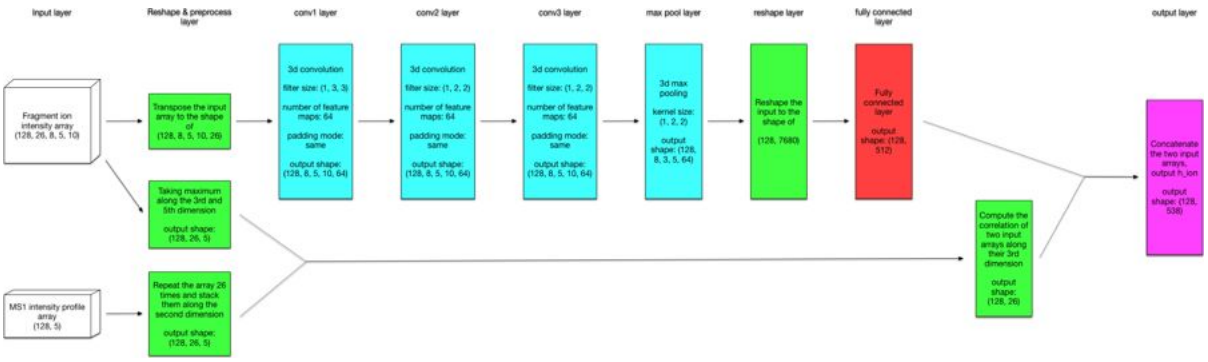


Figure 3.3: DeepNovo-DIA Ion CNN module

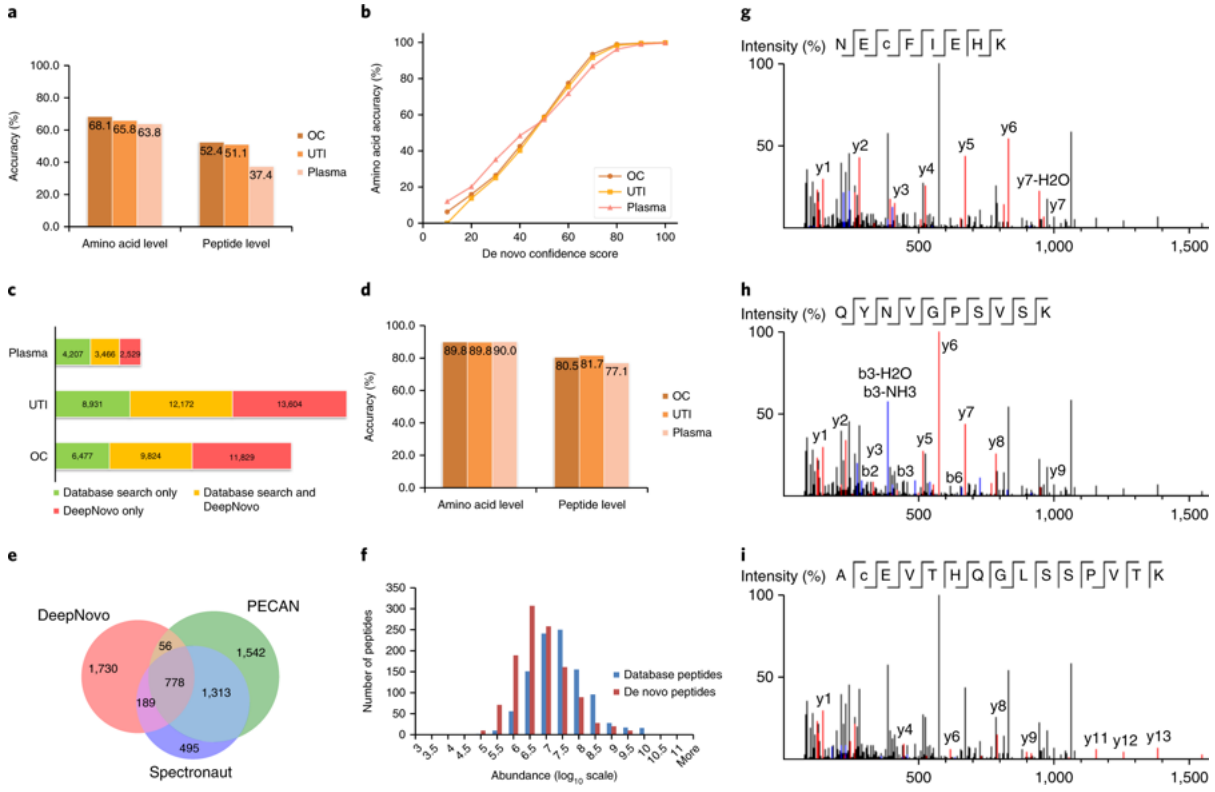


Figure 3.4: DeepNovo-DIA evaluation. (**a**) Accuracy of DeepNovo-DIA on labeled features. (**b**) Distribution of DeepNovo-DIA accuracy and confidence scores. (**c**) Precursor features with peptide identifications by in-house database search or DeepNovo-DIA. (**d**) DeepNovo-DIA accuracy on overlapping features in (**c**). (**e**), Comparison of unique peptides identified by DeepNovo-DIA, PECAN, and Spectronaut from the plasma dataset. (**f**), Abundance distributions of 1,143 *de novo* peptides identified by DeepNovo-DIA and 1,023 database peptides identified by DeepNovo-DIA and PECAN or Spectronaut. (**g**–**i**), Examples of a DIA spectrum that contains three different peptides, all of which were predicted by DeepNovo-DIA. In each panel, the fragment ions supporting the corresponding peptide are highlighted (red, y ion; blue, b ion).

Chapter 4

PointNovo

De novo peptide sequencing is the problem of reconstructing the peptide sequence directly from a tandem mass spectrum and the peptide mass. In the past 20 years different *de novo* peptide sequencing tools have been proposed and successful applications have been shown in assembling monoclonal antibody sequences and identifying tumor specific antigens (TSA), especially those resulting from noncoding region or alternative splicing. However, it still remains challenging for a *de novo* peptide sequencing tool to discriminate between amino acids pairs that have similar masses, e.g. glutamine (Q) and lysine (K), methionine sulfoxide (M(Oxi)) and phenylalanine (F). For instance, when evaluating the accuracy of *de novo* peptide sequencing, some previous studies [49, 77] considered a predicted amino acid matching a real amino acid if their mass difference is smaller than 0.1 Da and if the prefix masses before them differ by less than 0.5 Da. This means, for example, if a *de novo* sequencing tool reports a Q for a ground truth K, it will still be labeled as "correct" by the evaluation criteria since the mass difference between Q and K is smaller than 0.05 Da. However, for applications of antibody sequencing or TSA finding, it is important for the *de novo* sequencing tool to be able to reconstruct the exact sequence of a peptide. Otherwise an amino acid difference could result in an ineffective drug or vaccine. With recent advances in mass spectrometers, the mass accuracy could be improved to around 1 ppm. For a fragment ion of mass 1000 Da, this means the measurement error is smaller than 0.001 Da. Such high-resolution data allows accurate de novo peptide sequencing.

On the other hand, most existing de novo sequencing tools were developed back in the days when the mass error was greater than 100 ppm. It is not trivial for those tools to take full advantage of the higher precision provided by the latest generation of mass spectrometers. For spectrum graph-based methods [19, 14, 28], a higher precision means less nodes shall be merged and the generated spectrum graph would have more vertices.

This leads directly to a higher computational complexity. Similarly, the complexity of dynamic programing based methods such as PEAKS [50] and Novor [49] are sensitive with respect to the spectrum resolution. For instance, the computational complexity of the dynamic programing proposed by [51] is inversely proportional to the cube of the finest calibration of the mass spectrometer. In addition, the current existing neural network based *de novo* sequencing models, e.g. DeepNovo [77] and SMSNet [41], need to first discretize a spectrum to an intensity vector. For example, DeepNovo uses a length 150,000 vector to represent a spectrum when the spectrum resolution parameter is set to 50. The creation and process of the long intensity vectors require significant memory and CPU time. In fact, in the original implementation of DeepNovo, GPU is often not fully utilized because the program needs to wait for the CPU to build and process those long vectors. In order to take advantage of the improved precision offered by spectra of higher resolution, both DeepNovo and SMSNet need to discretize spectra with a higher spectrum resolution parameter R. For these models, the computation and memory demands grow linearly with respect to R (i.e. complexity of $O(R)$).

To fully benefit from the high precision that the latest mass spectrometers offer, we present PointNovo, a neural network based de novo peptide sequencing tool that does not vectorize the mass spectrum. PointNovo is ready to be applied on higher resolution data that may be generated in the future, without any added complexity. Moreover, our experiment results show that PointNovo also significantly outperforms previous state of the art methods. PointNovo achieves this by directly representing a spectrum as a set of m/z value and intensity pairs, and through the use of an order invariant network structure [59] to learn from data of such structure. Our extensive experiment results show PointNovo outperforms existing *de novo* peptide sequencing tools by capitalizing on the ultra-high resolution of the latest mass spectrometers.

Further, we demonstrate that the PointNovo model could also be used for database searching, even though it is not trained to distinguish between true and false peptide spectrum matching (PSM). Our experiments on several different datasets show that by using the fragment ion scores predicted by PointNovo, together with other common PSM features, we can achieve an identification rate at least comparable with other popular database search tools such as MaxQuant[17].

4.1 Method

4.1.1 Spectrum Representation

In DeepNovo and SMSNet, spectra are represented as intensity vectors, where each index of the vectors represents a small m/z bin and the value represents the sum of intensities of all peaks fall into that bin. This representation of spectra naturally has the problem of accuracy and speed/memory trade-off. In PointNovo, we propose to directly represent a spectrum as a set of *(m/z, intensity)* pairs. For each spectrum we select the top N most intense peaks (by default $N = 1000$), and represent the spectrum as $\{(m/z_i, I_i)\}_{i=1}^{N}$. Further, we denote $M_{observed} = (m/z_1, \cdots, m/z_N)$ as the observed m/z vector and $I = (I_1, \cdots, I_N)$.

4.1.2 Feature Extraction

Aside from the 20 amino acid residues and their PTMs, we include three special tokens—"start", "end", and "padding"—in our model's vocabulary set. Following the notations defined in Section 1.2.1, we denote the number of tokens as v and number of ion types as k. PointNovo use the 12 types of ion ($k = 12$): b, y, a, b(2+), y(2+), a(2+), b-H2O, y-H2O, a-H2O, b-NH3, y-NH3, and a-NH3. Their mass transforms are defined by the following formula:

$$
\begin{aligned}
\delta_{\mathrm{b}}(m(P')) &= m(P') + m_H \\
\delta_{\mathrm{y}}(m(P'_{\mathrm{c}})) &= m(P'_{\mathrm{c}}) + m_H + m_{\mathrm{H_2O}} \\
\delta_{\mathrm{a}}(m(P')) &= m(P') + m_H - m_{\mathrm{CO}} \\
\delta_{\mathrm{b(2+)}}(m(P')) &= \frac{m(P') + 2 \times m_H}{2} \\
\delta_{\mathrm{y(2+)}}(m(P'_{\mathrm{c}})) &= \frac{m(P'_{\mathrm{c}}) + 2 \times m_H + m_{\mathrm{H_2O}}}{2} \\
\delta_{\mathrm{a(2+)}}(m(P')) &= \frac{m(P') - m_{\mathrm{CO}} + 2 \times m_H}{2} \\
\delta_{\mathrm{b\text{-}H2O}}(m(P')) &= \delta_{\mathrm{b}}(m(P')) - m_{\mathrm{H_2O}} \\
\delta_{\mathrm{y\text{-}H2O}}(m(P'_{\mathrm{c}})) &= \delta_{\mathrm{y}}(m(P'_{\mathrm{c}})) - m_{\mathrm{H_2O}} \\
\delta_{\mathrm{a\text{-}H2O}}(m(P')) &= \delta_{\mathrm{a}}(m(P')) - m_{\mathrm{H_2O}} \\
\delta_{\mathrm{b\text{-}NH3}}(m(P')) &= \delta_{\mathrm{b}}(m(P')) - m_{\mathrm{NH_3}} \\
\delta_{\mathrm{y\text{-}NH3}}(m(P'_{\mathrm{c}})) &= \delta_{\mathrm{y}}(m(P'_{\mathrm{c}})) - m_{\mathrm{NH_3}} \\
\delta_{\mathrm{a\text{-}NH3}}(m(P')) &= \delta_{\mathrm{a}}(m(P')) - m_{\mathrm{NH_3}}
\end{aligned}
\tag{4.1}
$$

where $m_{\mathrm{CO}} \approx 27.9949$ Da. At each prediction step, we compute the theoretical m/z values for each token and ion type pair. The result is a matrix of shape (v, k) and is denoted as $M_{theoretical}$. Next we expand the dimension of $M_{observed}$ to make it a 3-dimensional tensor of shape $(N, 1, 1)$, and then repeat $M_{observed}$ on second dimension for v times and on third dimension for k times. The result is denoted as $M'_{observed}$ and it is a tensor of shape (N, v, k). Similarly, we expand $M_{theoretical}$ to the shape of $(1, v, k)$, repeat on first dimension for N times and denote the result as $M'_{theoretical}$. We can then compute the m/z difference tensor (denoted as D) in which each element represents the difference between the m/z value for an observed peak and the theoretical m/z for a token and ion type pair.

$$
D = M'_{observed} - M'_{theoretical} \tag{4.2}
$$

It is worth noting that Equation 4.2 can be computed efficiently by using the "broadcast" behavior in popular frameworks like Tensorflow[1] and PyTorch[58].

$$
\sigma(D) = exp\{-|D| * c\} \tag{4.3}
$$

Based on expert knowledge of *de novo* peptide sequencing, we design an activation function σ, shown in Equation 4.3. Here, the exponential and absolute operations are

all element-wise operations. The intuition for σ is that an observed peak could only be considered matching a theoretical m/z location if the absolute m/z difference is small. For example, if we set $c = 100$, then an observed peak that is 0.02 Da away from a theoretical location would generate a signal of $e^{-2} \approx 0.135$, which is only one-seventh of the signal of a perfect match. In our experiments, we tried setting c to be a trainable parameter and updating it through backpropagation. It shows similar performance with setting $c = 100$. For better model interpretability, we set $c = 100$ in all experiments reported in this manuscript. However, setting c to a learnable parameter would require less prior knowledge about the resolution of training spectra and might be preferable in certain cases.

$$F = \sigma(D)' \oplus I' \tag{4.4}$$

Next, we reshape the N by v by k tensor $\sigma(D)$ to a matrix $\sigma(D)'$ of shape N by vk, reshape I to a N by 1 vector I'. Finally, the feature matrix F used for predicting the next amino acid is simply the concatenation of $\sigma(D)'$ and I', as shown in Equation 4.4. Here $\oplus$ represents concatenation along the second dimension. The output F is matrix of shape N by $vk + 1$

A spectrum is set of $(m/z, intensity)$ pairs, which means the order of peaks should be irrelevant. Therefore, the prediction network should have order invariant property with respect to the first dimension of F. To the best of our knowledge, T Net (introduced in Section 2.3) is the first model designed for this kind of order invariant data. It showed state-of-the-art performance on the point cloud classification task. Therefore, we apply T net to learn from the feature matrix F. The global max pooling operation in T Net guarantees that the output would not change for any row permutations of F.

4.1.3 The Initial state for LSTM

The LSTM module is an optional component in PointNovo. In some applications, e.g. training an allele aware de novo sequencing model for HLA peptides, it might be desirable for the model to remember some peptide sequence patterns. In such cases, we can include an LSTM module in PointNovo. The full model structure of PointNovo (both with and without an LSTM module) is shown in Figure 4.1.

We need to initialize the hidden states of LSTM with information from the original spectrum. Inspired by the success of positional embedding introduced by Vaswani et al. [81], we choose to embed each peak into a vector. In more detail, the input spectrum is

first discretized at 0.1 Da resolution. When applied to the case of without LSTM, the discretization step is not needed.

Next, we create a sinusoidal m/z positional embedding matrix E in the way suggested by [81].

$$
\begin{aligned}
E_{(loc,2j-1)} &= sin(loc/10000^{\frac{2j-2}{512}}) \\
E_{(loc,2j)} &= cos(loc/10000^{\frac{2j-2}{512}}) \\
\forall j &\in \{1, 2, \cdots, 256\}
\end{aligned}
$$

Here loc represents the m/z index after discretization. We use E_l to denote the lth row vector of E. The sinusoidal embedding has a desired property that for any distance d, E_{loc+d} could be represented as a linear function of E_{loc}. This property is important because in mass spectra the m/z difference between observed peaks contains useful information that indicates which amino acids possibly exist. For an input spectrum: $\{(m/z_i, I_i)\}_{i=1}^{N}$ we denote loc_i to represent the index of m/z_i after discretization and we use $I_i E_{loc_i}$ as the vector representation of the ith peak. A spectrum representation vector S can then be generated by taking the summation of the vector representations of all peaks:

$$
S = \sum_{i=1}^{N} I_i E_{loc_i} \tag{4.5}
$$

We multiplied the intensities with the embedded peak vectors because we think the effect of a single peak, in the representation of a spectrum, should be proportional to its intensity. Finally, the hidden states of the LSTM module are initialized to S.

4.1.4 Training and Searching

As suggested by Tran et al.[75], we used focal loss[47] instead cross-entropy loss when training the model. We train PointNovo with Adam algorithm[44] with an initial learning rate of 10^{-3}. After every 300 training steps, the loss on the validation set is computed. If the validation loss has not achieved a new low in the recent ten evaluations, then the learning rate would be dropped by half. As for the searching part, we applied the beam search algorithm used by DeepNovo. Similar to DeepNovo, PointNovo also uses knapsack algorithm to reduce the search space.

4.1.5 Speed of PointNovo

On an RTX 2080 TI GPU, a training step (batch size 16) takes around 0.4 seconds. And for inference (i.e. *de novo* peptide sequencing), PointNovo (with LSTM) could process around 20 spectra per second. In the without LSTM model, both training and inference would be faster.

4.1.6 Database Search

For database search experiments, we use PointNovo model **without** the LSTM component. This is because that in the training dataset, all labeled sequences are target peptides. Thus if we train a model with a recurrent neural network component, the model will learn to remember the sequence pattern of target peptides. Previous research of deep neural network-based database searching models used a RNN component and did cross-species training[66]. Here we argue that different species still share some common protein sequences and peptides, thus using RNN with cross-species training would still break the assumptions of target-decoy strategy and make evaluation results biased.

Following the notations defined in section 1.2.1, the set of amino acids is denoted as $\mathcal{A} = \{a_1, a_2, \cdots, a_v\}$ and $v = |\mathcal{A}|$. We use a one-to-one mapping function e to encode each amino acid into a unique integer between 1 and v. Given a spectrum $\mathbf{S}$ and a prefix mass m_{prefix}, the PointNovo model is trained to predict a probability distribution for the next amino acid.

We denote the probability distribution as

$$\mathbf{ps}(m_{\text{prefix}}) = (p^1_{\mathbf{S}}(m_{\text{prefix}}),\ p^2_{\mathbf{S}}(m_{\text{prefix}}),\ \cdots,\ p^v_{\mathbf{S}}(m_{\text{prefix}}))$$

where $p^i_{\mathbf{S}}(m_{\text{prefix}})$ represents the probability of next amino acid being $e^{-1}(i)$. Then for any length n peptide $p_1p_2\cdots p_n$, $p_i \in \mathcal{A}$, we could define a DeepNovo PSM score using the following formula:

$$f(\mathbf{S},\ p_1p_2\cdots p_n) = \sum_{i=0}^{n-1} \log p^{e(p_{i+1})}_{\mathbf{S}}(m_i) \tag{4.6}$$

where $m_0 = 0$ and

$$m_i = \sum_{j=1}^{i} mass(p_j),\quad \forall i \in \{1, 2, \cdots, n-1\}$$

To compare PSM scores among different candidate peptides, we need to normalize these scores by the peptides' length. We choose to include both length normalized scores and log-length normalized scores.

$$f_{\text{length}}(\mathbf{S},\ p_1p_2\cdots p_n) = \frac{f(\mathbf{S},\ p_1p_2\cdots p_n)}{n}$$
$$f_{\text{loglength}}(\mathbf{S},\ p_1p_2\cdots p_n) = \frac{f(\mathbf{S},\ p_1p_2\cdots p_n)}{\log n}$$

Also, to make these PSM scores comparable across different spectra, we follow the normalization procedure proposed by PEAKS DB[90]:

$$\bar{f}_{\text{length}}(\mathbf{S}, p_1p_2\cdots p_n) = \frac{f_{\text{length}}(\mathbf{S}, p_1p_2\cdots p_n) - \mu(\mathbf{S})}{\sigma(\mathbf{S})}$$

where $\mu(\mathbf{S})$ represents the mean f_{length} scores for top 10 candidate peptides of spectrum $\mathbf{S}$, and $\sigma(\mathbf{S})$ represents the standard deviation of top n_{std} candidate peptides' f_{length} scores.

When sequencing peptide against a target protein database, we first create a decoy protein database by reversing all protein sequences in the target database. We then combine the target and decoy database. Next, for each spectrum $\mathbf{S}$, we retrieve all candidate peptides (with a mass close to the precursor mass) from the combined database. Then for each candidate peptides, we compute the PointNovo scores f_{length}, $f_{\text{loglength}}$, $\bar{f}_{\text{length}}$ and $\bar{f}_{\text{loglength}}$ together with other features such as peptide length, charge state, mass difference and number of variable modifications. The top 10 candidate peptides with the highest $\bar{f}_{\text{length}}$ scores will be saved for re-ranking by percolator[39]. Finally, percolator will report the list of identified peptides at 1% FDR.

4.2 Results

4.2.1 *De novo* Sequencing Results

We downloaded the nine species data used by the original publication of DeepNovo (MSV000081382) and applied our model to this data. We implement the same leave-one-out cross-validation scheme as described in [77], i.e. all except one of the nine datasets were used to train PointNovo and the trained model is tested on the remaining dataset. When calculating

the amino acid precision, amino acid recall, and peptide recall, we used the same evaluation metric adopted by DeepNovo and Novor, i.e., a predicted amino acid matching a real amino acid if their mass difference is smaller than 0.1 Da and if the prefix masses before them are different by less than 0.5 Da. To make a fair comparison, we used PointNovo with a long short-term memory (LSTM) module [34] in this experiment, because by default DeepNovo includes an LSTM module. The test results and comparison with DeepNovo are shown in Figure 4.2. PointNovo outperforms DeepNovo consistently on peptide level by a large margin of 13.01%–23.95%. We note out here that in the cross-species training for humans, DeepNovo reports a slightly lower amino acid recall but a higher peptide recall rate than PEAKS. Similar results are also observed for PointNovo. We suggest this is because in the cross-species training scheme, some peptides in the test set also appear in the training set. The LSTM modules in PointNovo and DeepNovo will be trained to predict the sequences that existed in the training set. It might be a desired property in some applications (e.g. training an allele aware *de novo* sequencing model for HLA peptides), for evaluating machine learning models it is not the best practice. To better compare our proposed model with DeepNovo, SMSNet [41] and pNovo3 [89], we collected three high-resolution MS/MS spectra datasets provided by different labs (Hela samples from ABRF3, PXD008844 [92] and PXD010559 [67]). On each of the three datasets, we first ran a database search using PEAKS X. The post translational modifications (PTMs) settings are included in the online Method. The identified PSMs at 1% FDR, on each dataset, are split into train, validation and test set in the ratio of 8:1:1. During the split, we made sure that no common peptide sequences are shared among the train, validation and test sets. Then for each of the three high resolution MS/MS spectra datasets, two PointNovo models (with and without LSTM) and two DeepNovo models (with and without LSTM) are trained from scratch on the train set. The weights that show the best validation loss during training are saved as the trained model weights. Finally, trained models are evaluated on the test set. The mean value and standard deviation of metrics from 5 independent runs on the three test datasets are shown in Table 4.1–4.3 and in Figure 4.3. In the case of including an LSTM module, PointNovo improves on peptide level recall by 15.05%–23.32%. And in the case of not including an LSTM module, PointNovo outperformed DeepNovo by 25.61%–31.94%.

In a procedure similar to the above experiments, we also compared PointNovo with SMSNet. The results are shown in Figure 1d. In this comparison, we applied SMSNet without re-scoring [41]. Because PointNovo does not contain any post-processing. Due to a limitation of SMSNet, all PSMs that contain PTMs other than carbamidomethylation of C or oxidation of M are removed from our datasets. As a result, the train, validation and test sets are slightly different from previous experiments and that is the reason why accuracies of PointNovo reported in Figure 1d are different from those reported in

	DeepNovo without lstm	DeepNovo with lstm	PointNovo without lstm	PointNovo with lstm
AA recall	66.44%(±0.18%)	67.81%(±0.21%)	71.59%(±0.18%)	72.46%(±0.15%)
AA precision	66.42%(±0.22%)	67.67%(±0.25%)	71.65%(±0.24%)	72.25%(±0.20%)
peptide recall	30.33%(±0.15%)	32.96%(±0.31%)	38.10%(±0.06%)	39.24%(±0.29%)

Table 4.1: ABRF DDA dataset. Carbamidomethylation of C was set as a fixed modification. Oxidation of M and deamidation of N or Q were set as a variable modification.

	DeepNovo without lstm	DeepNovo with lstm	PointNovo without lstm	PointNovo with lstm
AA recall	66.98%(±0.17%)	67.31%(±0.33%)	73.22%(±0.14%)	73.45%(±0.45%)
AA precision	66.51%(±0.15%)	66.65%(±0.45%)	73.03%(±0.14%)	72.66%(±0.25%)
peptide recall	40.37%(±0.31%)	42.03%(±0.48%)	52.07%(±0.31%)	51.83%(±0.09%)

Table 4.2: PXD008844 dataset. Carbamidomethylation of C was set as a fixed modification. Oxidation of M was set as a variable modification.

Figure 4.4. We notice that in the case of without re-scoring, SMSNet sometimes predicts exceptionally long sequences for spectra of poor quality. The existence of such long sequences undermines the amino acid level accuracy. Therefore, the peptide level recall metric shows a better comparison of the performance of the two models. Nevertheless, PointNovo outperforms SMSNet (without re-scoring) on peptide level by over 17%. We want to point out here that the contribution of sequence-mask-search made by SMSNet is orthogonal to the improvement made by PointNovo. A similar post-processing could be applied to the output of PointNovo. In Figure 4.5 we show the comparison results between PointNovo and pNovo3. Because pNovo3 is distributed as pretrained software, we cannot adopt the same training procedure as the previous experiments since that would give PointNovo an unfair advantage. To make a fair comparison, we collected four other high-resolution MS/MS spectra datasets: PXD008808 [72], PXD011246 [8], PXD012645 [69] and PXD012979 [31]. We trained a PointNovo without an LSTM model on the identified PSMs of these four datasets and applied the trained model on the test sets of ABRF, PXD008844 and PXD010559. In this experiment, we again exclude all PSMs that contain PTMs other than carbamidomethylation of C or oxidation of M from the train and test sets because we need to apply the same trained model on all three test sets. Figure 4.5 shows that our trained PointNovo without LSTM model outperforms pNovo3 by more than 25.5% on peptide level. More interestingly, the performance gap of PointNovo between Figure 4.4 and Figure 4.5 gives us an estimate of the generalizability of our proposed

	DeepNovo without lstm	DeepNovo with lstm	PointNovo without lstm	PointNovo with lstm
AA recall	69.03%(±0.11%)	73.94%(±0.12%)	74.62%(±0.17%)	79.00%(±0.14%)
AA precision	68.71%(±0.17%)	73.63%(±0.13%)	74.60%(±0.18%)	78.92%(±0.10%)
peptide recall	38.26%(±0.24%)	52.48%(±0.40%)	50.48%(±0.15%)	60.73%(±0.27%)

Table 4.3: PXD010559 dataset. Carbamidomethylation of C was set as a fixed modification. Oxidation of M, deamidation of N or Q and phosphorylation of S, T, or Y were set as variable modifications.

model. The metrics reported in Figure 4.4 represent the performance in the best-case scenario, where the training spectra are acquired in the same experiment setting as the test spectra (e.g. different fractions of the same sample). As well, Figure 4.5 results represent the performance in the normal scenario, where training spectra are collected from multiple experiments conducted by different labs. Above all, our results as shown in Figure 4.2–4.5 demonstrate that PointNovo consistently outperforms DeepNovo, SMSNet (without rescoring) and pNovo3 on all three different test sets. Here we want to explain again that, even though the results shown in Figure 4.3–4.5 are from the same test datasets, these results cannot be merged because the three experiments are conducted in different settings (i.e. different PTMs included, different training datasets) for the purpose of making a fair comparison.

To further demonstrate that our proposed PointNovo model could take full advantage of the high-resolution data and better discriminate between amino acids pairs that have similar masses, we calculate the precision and recall for amino acid pairs F and M(Oxi) (the mass difference is smaller than 0.035 Da), Q and K. In this analysis, a predicted amino acid is considered as matching the ground truth amino acid in the target sequence if and only if the amino acids are exactly the same and the prefix masses before them are different by less than 0.5 Da. Both DeepNovo and PointNovo are trained without the LSTM modules, since we want to compare their ability of learning from spectra, not their ability to remember the sequence patterns. The precision-recall curves for two datasets are shown in Figure 4.6 and 4.7. PointNovo improves the Average Precision (AP) for all four amino acids, which are widely known for being hard to discriminate. Specifically, for amino acid Q and M(Oxi), we observe a significant improvement of more than 15%. Figure 4.8 shows Venn diagrams of the peptide sets identified by PEAKS X (database search), predicted by PointNovo, and predicted by DeepNovo on ABRF, PXD008844 and PXD010559 datasets. Following the practice introduced by [75], we filtered the *de novo* peptides based on their peptide scores given by the models. Peptide score cutoffs are selected so that the amino acid accuracy is

90%. The intersection between two sets represents peptides of the exact same amino acid sequence. As can be seen from the Venn diagrams, PointNovo's prediction always covers more peptides identified by PEAKS X as compared to DeepNovo's prediction.

Finally, to show that PointNovo can potentially benefit from the improved precision of higher-resolution spectra generated in the future, we simulate low-resolution spectra of ABRF, PXD008844 and PXD010559 datasets and report PointNovo's performance on these spectra in Figure 4.9. The low-resolution spectra are generated by adding random parts per million (ppm) errors $\epsilon \sim U(-10, 10)$ to the m/z value of every peak in original spectra datasets. PointNovo is then trained and tested on the jittered train and test spectra. The comparison results in Figure 4.9 demonstrate that, with PointNovo, we could indeed expect better performance on spectra of higher resolution.

The above results demonstrate that PointNovo outperforms previous state-of-the-art *de novo* peptide sequencing tools by a significant margin and could better discriminate between similar amino acids pairs. Also, unlike previous neural network based de novo peptide sequencing tools, PointNovo does not include any spectrum vectorization, thus is ready to be applied on the more precise MS data generated in the future.

4.2.2 Database Search Result

For database search experiments, we collected the training spectra from PXD008844(Mouse), PXD012979(Mouse) and PXD008808(Dolphin) and train a PointNovo model without the LSTM component. After training, we test our model on three human dataset: PXD008999, PXD007890, and PXD009021. Due to computational limitations, we select one fraction (one raw file) per each dataset for testing. In all three experiments, the Swiss-Prot human protein database[16] is used as the reference database and precursor tolerance is set to be 20 ppm. Under this setting, there are often more than 1000 candidate peptides for a given precursor mass. It is too expensive to compute the PointNovo scores for all of these candidates. To speed up the experiment, we use the number of matching partial peptides as a quick scorer to filter candidate peptides. Only the top 150 candidate peptides are fed into the PointNovo model and have their scores computed. After applying the quick filtering step, our proposed database search by PointNovo method could process more than 6000 spectra per hour on two GTX 1070 GPUs.

We compare PointNovo's performance with PEAKS DB[90], MaxQuant[17] and Comet[24]. The decoy database is generated by each software's default method. And for PointNovo, we generate the decoy database by reversing all protein sequences in the reference database.

As for the evaluation criteria, we compare the number of identified PSMs and unique peptides under 1% FDR.

	Identified PSMs	Identified peptides
PEAKS	2462	1296
Comet + Percolator	1248	676
MaxQuant	1601	917
PointNovo + Percolator	2190	1104

Table 4.4: Database search on Sclera_IG_20.raw, PXD008899. Carbamidomethylation of C was set as a fixed modification. Oxidation of M was set as variable modifications.

	Identified PSMs	Identified peptides
PEAKS	39897	23752
Comet + Percolator	30442	17624
MaxQuant	39524	23625
PointNovo + Percolator	40102	22889

Table 4.5: Database search on B02_06.raw, PXD007890. Carbamidomethylation of C was set as a fixed modification. Oxidation of M was set as variable modifications.

	Identified PSMs	Identified peptides
PEAKS	7977	5832
Comet + Percolator	7478	4841
MaxQuant	4834	3518
PointNovo + Percolator	7062	4904

Table 4.6: Database search on liver_20.raw, PXD009021. Carbamidomethylation of C was set as a fixed modification. Oxidation of M, deamidation of N and Q were set as variable modifications.

Our results in Table 4.4–4.6 demonstrate that even though PointNovo was not designed and trained to discriminate between true and false PSM, it can achieve an identification rate that is at least comparable to popular database search tools such as MaxQuant.

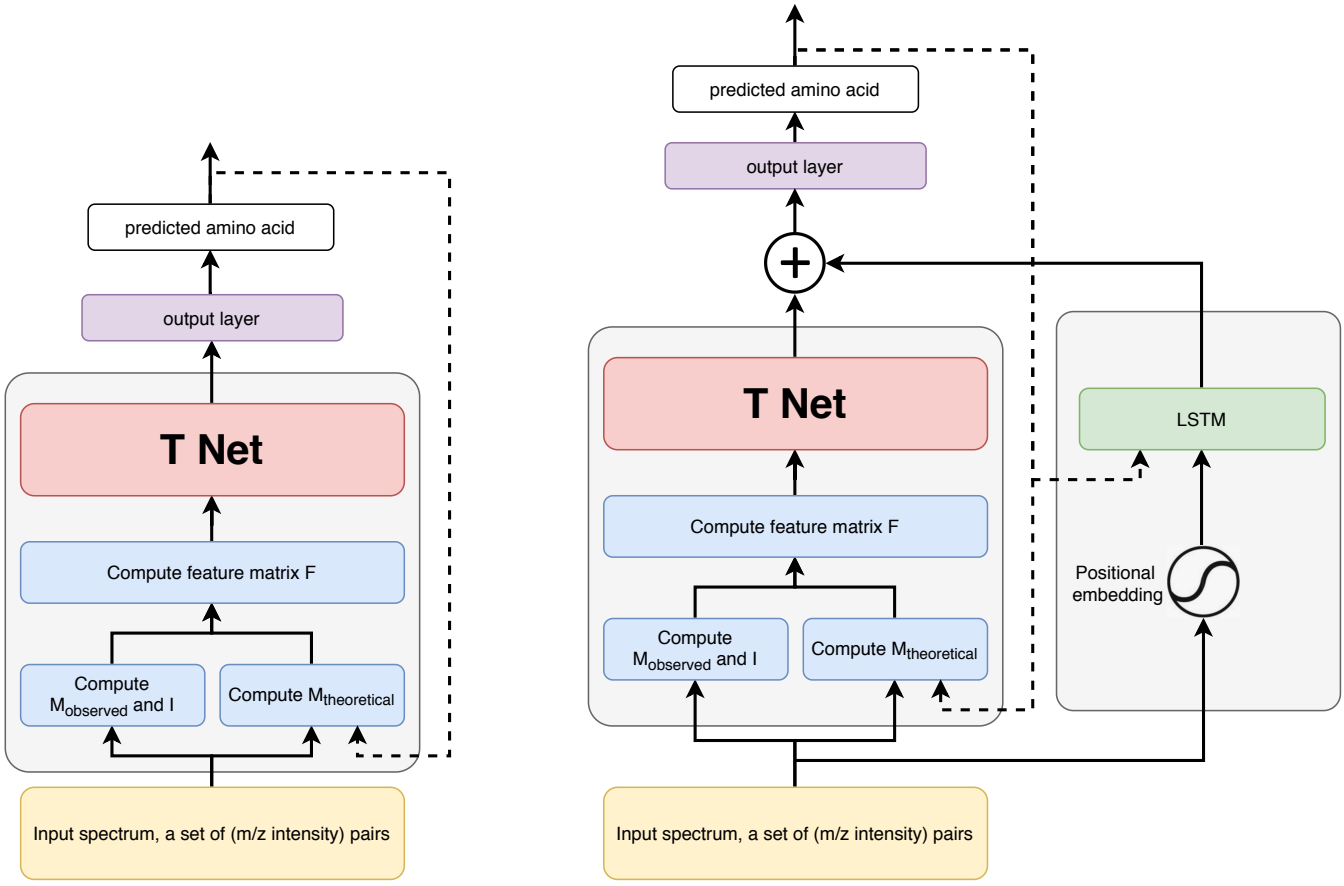


Figure 4.1: Structure of PointNovo

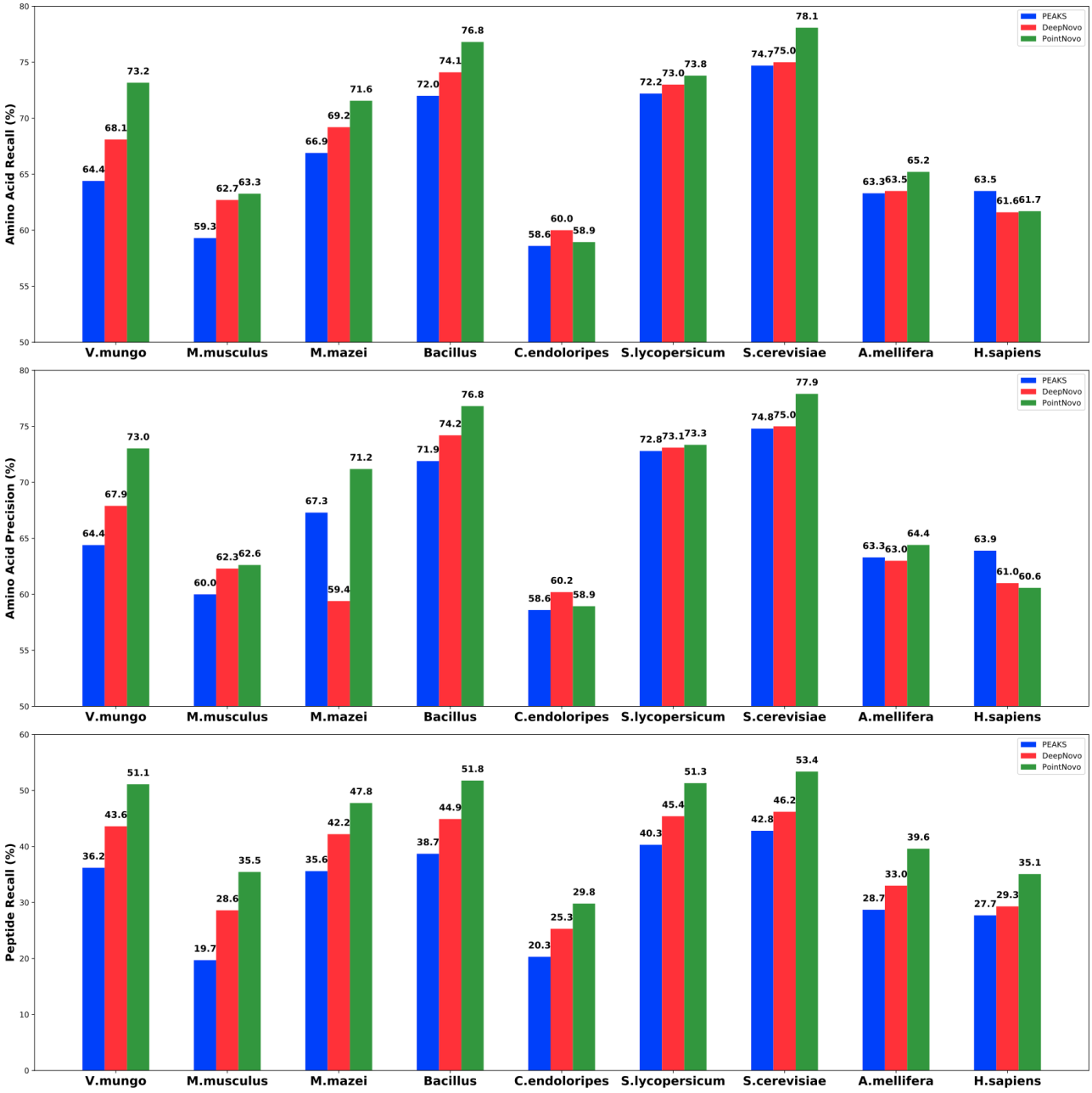


Figure 4.2: Amino acid recall, amino acid precision and peptide recall of DeepNovo and PointNovo

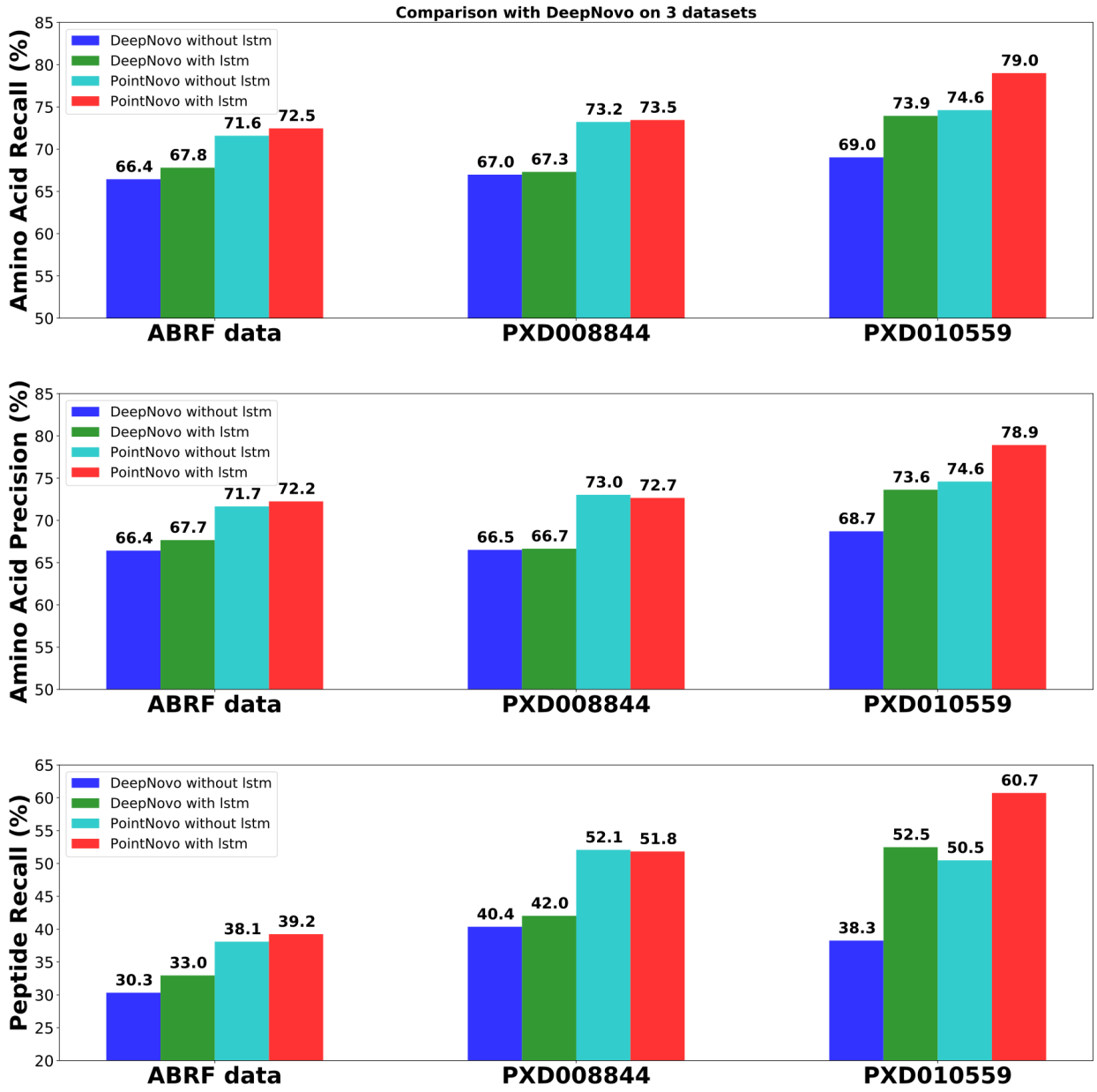


Figure 4.3: Amino acid recall, amino acid precision, and peptide recall of DeepNovo and PointNovo on three test datasets

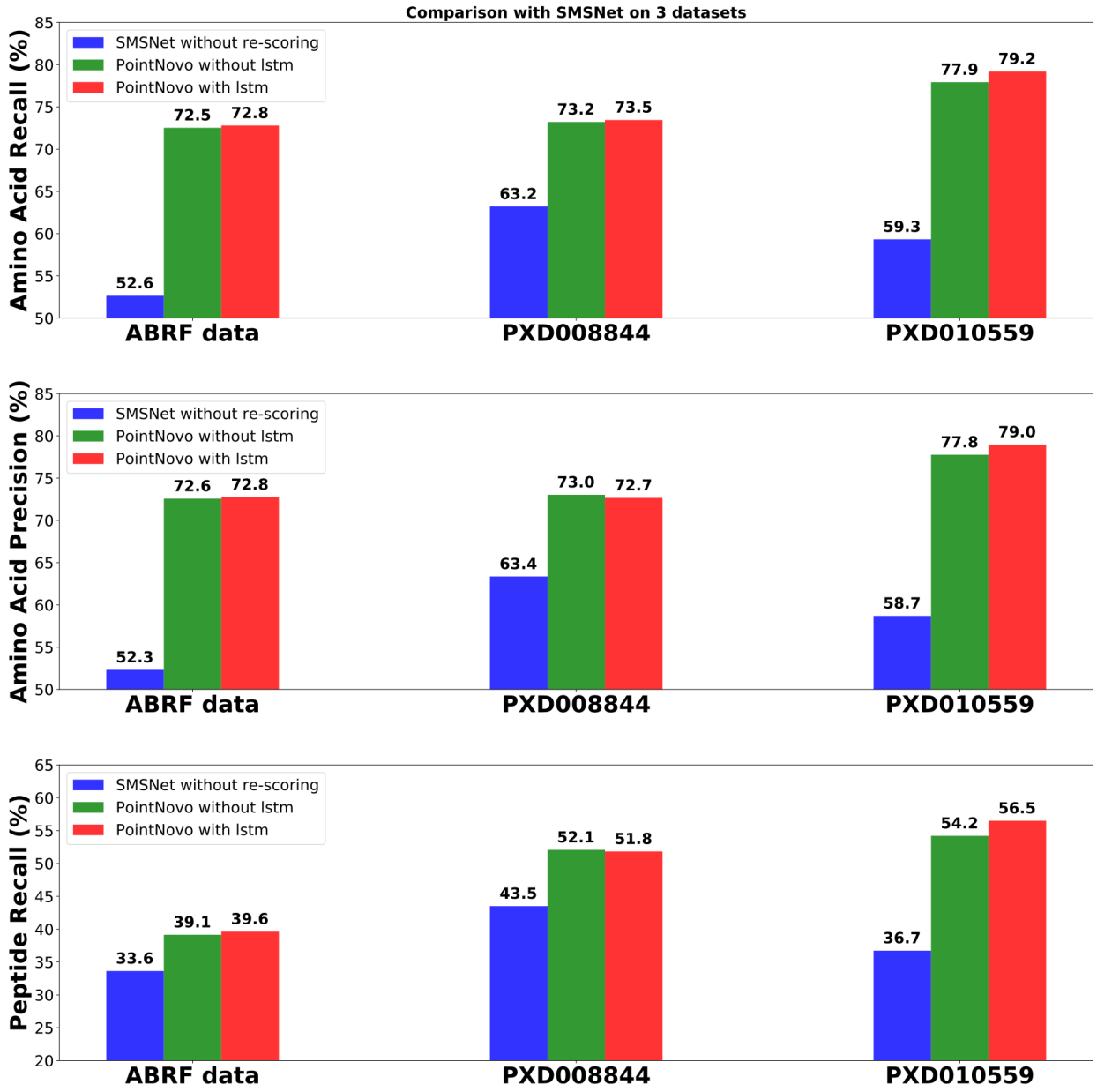


Figure 4.4: Amino acid recall, amino acid precision, and peptide recall of SMSNet and PointNovo on three test datasets

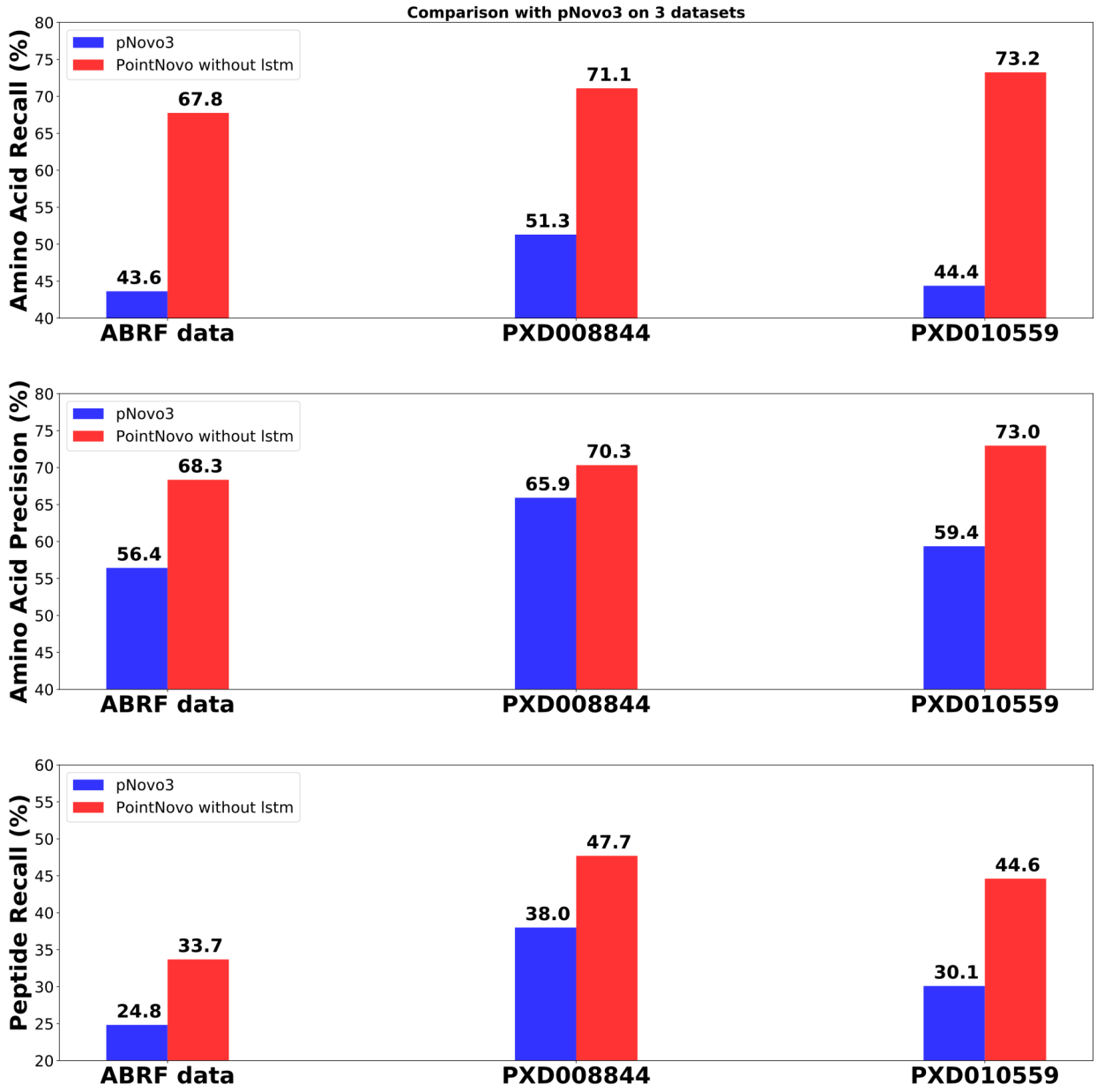


Figure 4.5: Amino acid recall, amino acid precision, and peptide recall of SMSNet and PointNovo on three test datasets

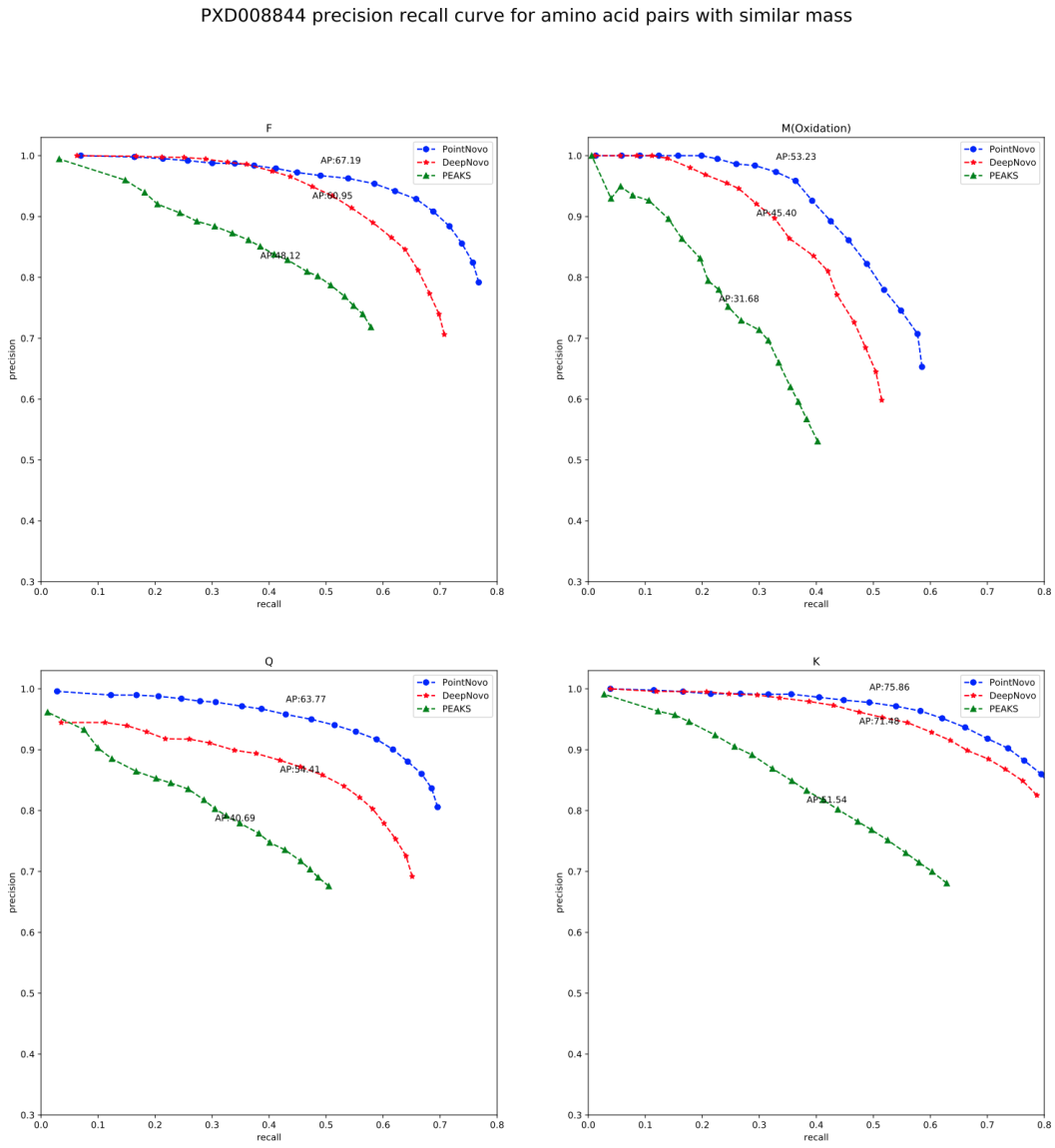


Figure 4.6: Precision recall curve for certain amino acid on PXD008844

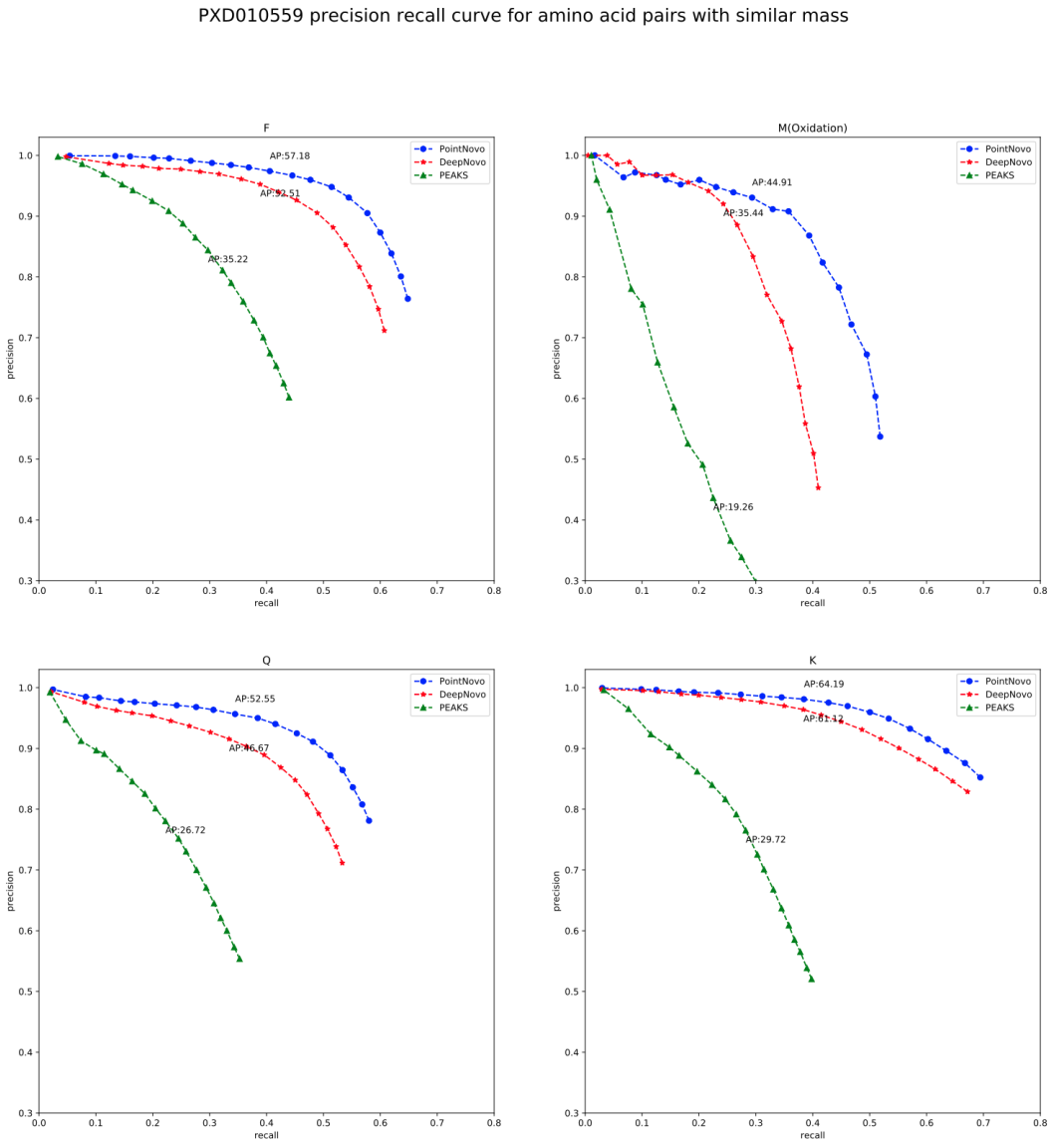


Figure 4.7: Precision recall curve for certain amino acid on PXD010559

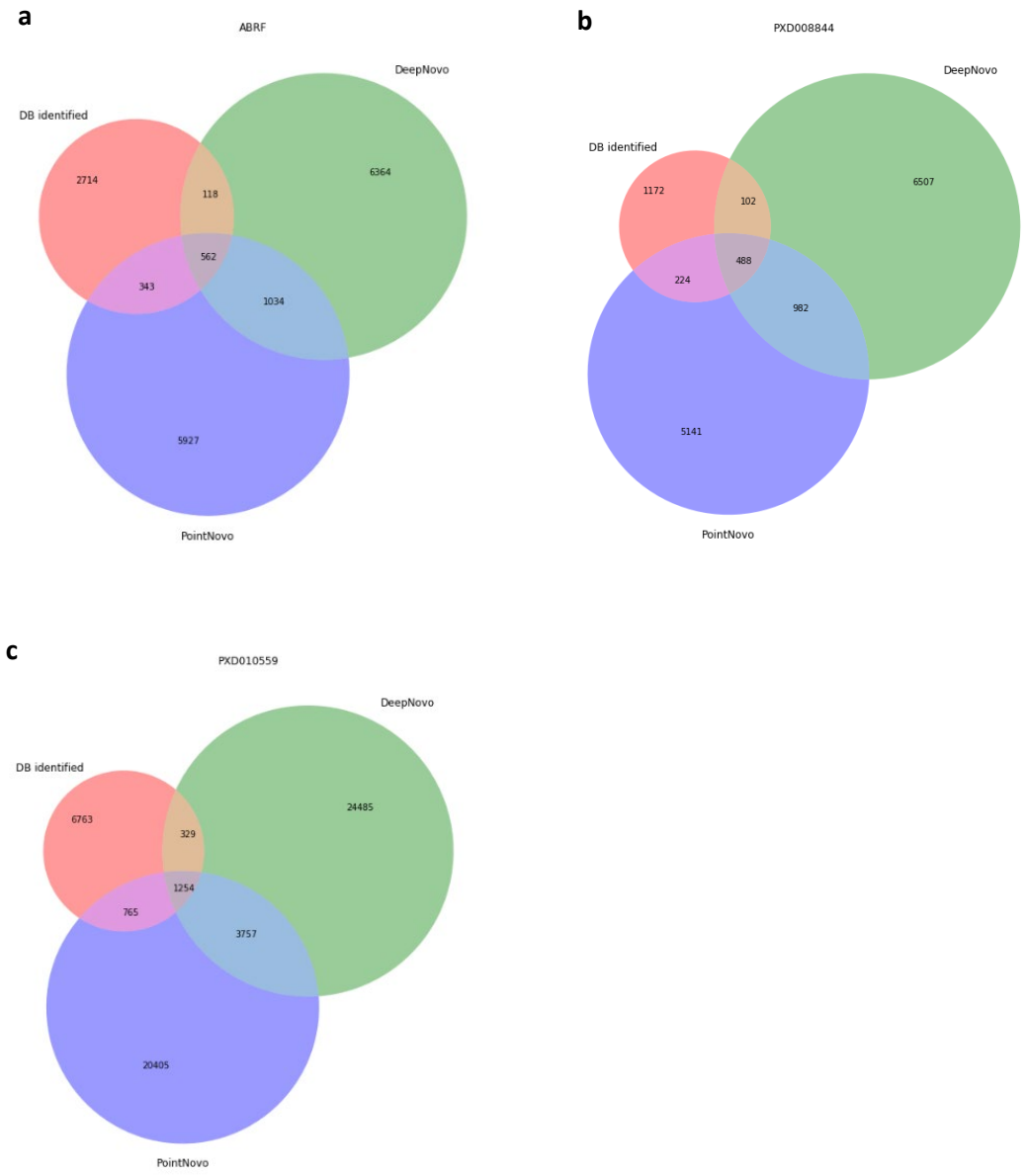


Figure 4.8: Set of peptides predicted by PointNovo and DeepNovo, comparing with the set of peptides identified by PEAKS DB. Both DeepNovo and PointNovo are trained without the LSTM modules.Peptide score cutoff is applied to the results given by PointNovo and DeepNovo. We select the cutoff scoresso that the amino acid accuracy of the remaining predicted peptides is 90%.Here, the overlap between two sets represents the peptides that are exactly the same (i.e. same amino acid residue sequence).

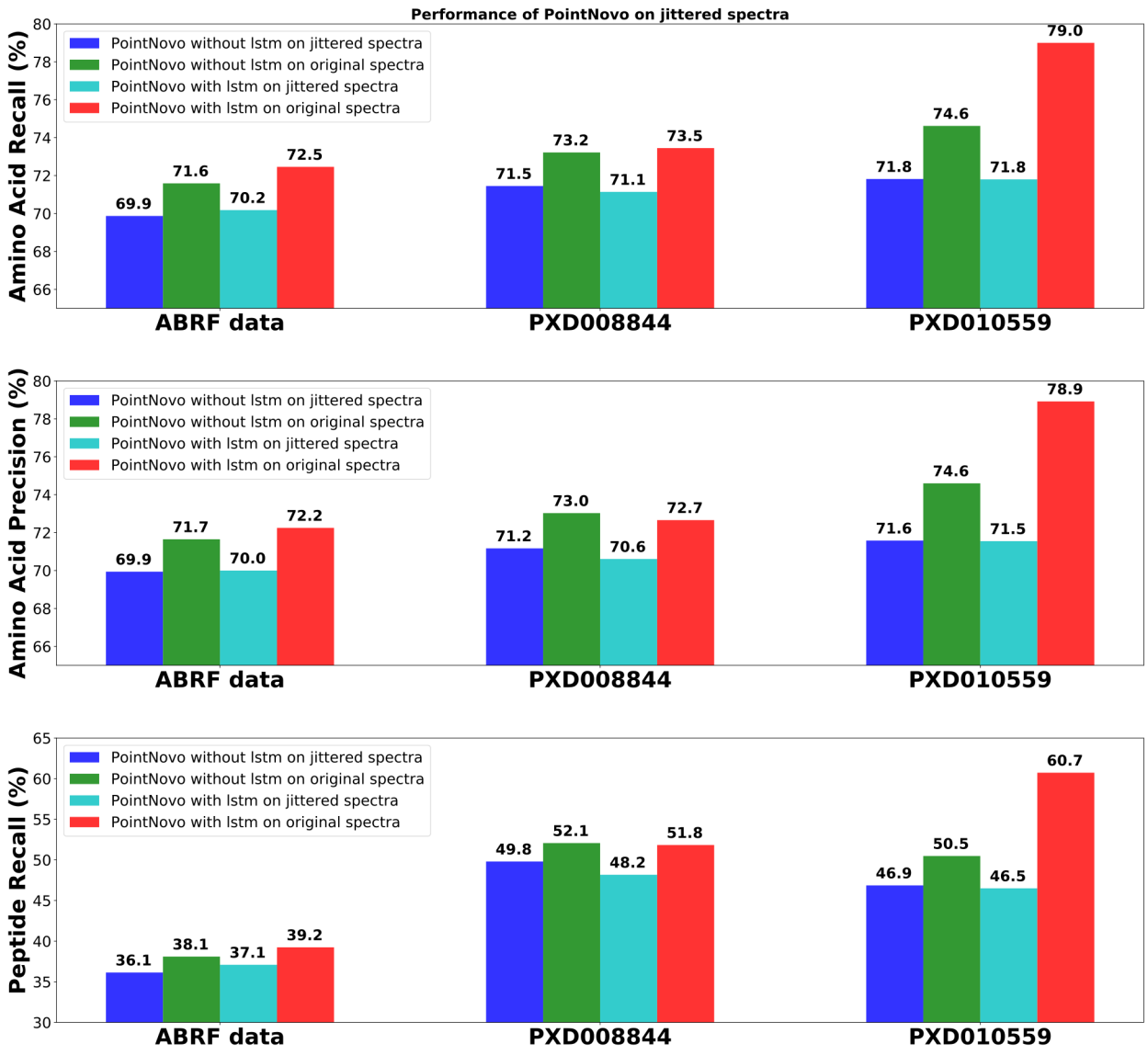


Figure 4.9: Performance of PointNovo on jittered spectra. To jitter the spectra, we add uniformly distributed random ppm errors to the m/z value of every peak in the original datasets. These jittered spectra could be considered as spectra of lower resolution

Chapter 5

Identifying Neoantigens by Personalized *De Novo* Peptide Sequencing

Neoantigens are tumor-specific mutated peptides that are brought to the cancer cell surface by major histocompatibility complex (MHC) proteins for T-cell recognition. As neoantigens carry tumor-specific mutations that are not found in normal tissues, they represent ideal targets for the immune system to distinguish cancer cells from non-cancer ones [36, 84]. The potential of neoantigens for cancer vaccines is supported by multiple evidences, including the correlation between mutation load and response to immune checkpoint inhibitor therapies[80, 63], neoantigen-specific T cell responses detected even before vaccination (naturally occurring)[56, 65, 13]. Indeed, three independent studies have further demonstrated successful clinical trials of personalized neoantigen vaccines for patients with melanoma[56, 65, 13]. The vaccination was found to reinforce pre-existing T cell responses and to induce new T cell populations directed at the neoantigens. In addition to developing cancer vaccines, neoantigens may help to identify targets for adoptive T cell therapies, or to improve the prediction of response to immune checkpoint inhibitor therapies.

The current prevalent approach to identifying candidate neoantigens often includes two major phases: (1) exome sequencing of cancer and normal tissues to find somatic mutations and (2) predicting which mutated peptides are most likely to be presented by MHC proteins for T-cell recognition. The first phase is strongly backed by high-throughput sequencing technologies and bioinformatics pipelines that have been well established through several genome sequencing projects during the past decade. The second phase, however, is still

facing challenges due to our lack of knowledge of the MHC antigen processing pathway: how mutated proteins are processed into peptides; how those peptides are delivered to the endoplasmic reticulum by the transporter associated with antigen processing; and how they bind to MHC proteins. To make it further complicated, human leukocyte antigens (HLA), those genes that encode MHC proteins, are located among the most genetically variable regions and their alleles basically change from one individual to another. The problem is especially more challenging for HLA class II (HLA–II) peptides than HLA class I (HLA–I), because of the more complicated heterodimer-based nature of MHC–II molecules and the limited understanding of peptide binding properties in the cleavage process.

Current *in silico* methods focus on predicting which peptides bind to MHC proteins given the HLA alleles of a patient, e.g., NetMHC[4, 38]. usually very few, less than a dozen from thousands of predicted candidates are confirmed to be presented on the tumor cell surface and even less are found to trigger T cell responses, not to mention that real neoantigens may not be among top predicted candidates[36]. In addition, binding prediction models do not perform equally well on different HLA alleles and the binding affinities of some less common HLA alleles still remain poorly characterized. Several efforts have been made to improve the MHC binding prediction, including using mass spectrometry data in addition to binding affinity data for more accurate prediction of MHC antigen presentation[2, 11]. Recently, proteogenomic approaches have been proposed to combine mass spectrometry and exome sequencing to identify neoantigens directly isolated from MHC proteins, thus overcoming the limitations of MHC binding prediction[6][46]. In those approaches, exome sequencing was performed to build a customized protein database that included all normal and mutated protein sequences. The database was further used by a search engine to identify endogenous peptides, including neoantigens, that were obtained by immunoprecipitation assays and mass spectrometry.

Existing database search engines, however, are not designed for MHC peptides and are biased towards tryptic peptides[17, 90]. They may have sensitivity and specificity issues when dealing with a very large search space created by (i) all mRNA isoforms obtained from exome sequencing and (ii) unknown digestion rules for HLA peptides. Furthermore, recent proteogenomic studies reported a weak correlation between proteome- and genome-level mutations, where the number of identified mutated HLA peptides was three orders of magnitudes less than the number of somatic mutations that were provided to the database search engines[6, 46]. A large number of genome-level mutations were not presented at the proteome level, while at the same time, some mutated peptides might be difficult to detect at the genome level. For instance, Faridi et al. found evidence of up to 30% of HLA–I peptides that were cis- and trans-splicing, which couldn't be detected by exome sequencing nor protein database search[26]. Thus, an independent approach that does not rely heavily

on genome-level information to identify mutated peptides directly from mass spectrometry data is needed, and *de novo* sequencing is the key to address this problem.

In this study, we propose, for the first time, a personalized *de novo* sequencing workflow to identify HLA–I and HLA–II neoantigens directly and solely from mass spectrometry data. *De novo* sequencing is the process of reconstructing the amino acid sequence of a peptide from its tandem mass spectrum and its molecule mass, without an assisting protein database. This technique is invented for the purpose of discovering novel peptides and proteins, genetic variants or mutations. Thus, its application to identify neoantigens is a perfect match. Since both tumor mutations and HLA alleles are specific to each individual patient, a personalized approach is desirable to detect mutated HLA peptides. We bring *de novo* sequencing to the "personalized" level by training a specific machine learning model for each individual patient using her/his own data. In particular, we use the collection of normal HLA peptides, i.e. the immunopeptidome, of a patient to train a model and then use it to predict mutated HLA peptides of that patient. Learning an individual's immunopeptidome is made possible by our recent deep learning model, DeepNovo[77][75], which uses an LSTM to capture sequence patterns in peptides or proteins, in a similar way to natural languages. This personalized learning workflow significantly improves the accuracy of *de novo* sequencing for comprehensive and reliable identification of neoantigens. Furthermore, our *de novo* sequencing approach predicts peptides solely from mass spectrometry data and does not depend on genomic information as existing approaches. We applied the workflow to the datasets of five melanoma patients Mel-5, Mel-8, Mel-12, Mel-15, and Mel-16, which were published recently by Bassani-Sternberg et al. [6]. The datasets were selected because of the availability of mass spectrometry data, RNA-Seq and T-cell assay information together to validate the results of our workflow. HLA peptides were purified from the patients' tumor tissues and their mass spectrometry data were made publicly available by the authors in [6].

5.1 Results

5.1.1 Personalized *De novo* Sequencing of Individual Immunopeptidomes

Figure 5.1 describes five steps of our personalized *de novo* sequencing workflow to predict HLA peptides of an individual patient from mass spectrometry data: (1) build the immunopeptidome of the patient; (2) train personalized machine learning model; (3) personalized *de novo* sequencing; (4) quality control of *de novo* peptides; and (5) neoantigen

Neoantigen discovery workflow	HLA–I	HLA–II
Step 1: Build the immunopeptidome of the patient		
Number of identified peptide-spectrum matches	341,216	67,021
Number of identified database peptides	35,551	9,664
Number of unlabeled spectra	596,915	135,490
Step 2: Train personalized machine learning model		
Number of training PSMs	307,058	60,822
Number of validation PSMs	17,217	2,999
Number of test PSMs	16,941	3,200
Step 3: Personalized *de novo* sequencing		
Number of raw *de novo* peptides	441,274	93,983
Step 4: Quality control		
Number of high-confidence *de novo* peptides	16,226	2,717
Number of *de novo* peptides at 1% FDR	5,320	863
Step 5: Neoantigen selection Binding affinity for patient's HLA alleles; Missense mutation against wild-types; Neoepitope expression level	158	37

Table 5.1: Personalized workflow of neoantigen discovery for patient Mel–15

selection. The step-by-step results on five melanoma patients Mel–5, Mel–8, Mel–12, Mel–15, and Mel–16 from [13] are provided in Table 5.1–5.3

In step 1 of the workflow, to build the immunopeptidome of the patient, we searched the mass spectrometry data against the standard Swiss-Prot human protein database. Normal HLA peptides and their PSMs at 1% FDR were identified. Note that the immunopeptidome included both normal and mutated HLA peptides, and only normal HLA peptides were identified at this step. Mutated HLA peptides were not presented in the protein database, so they were not detected, and their spectra remained unlabeled. We identified from 36,369–341,216 PSMs and from 10,068–35,551 peptides of HLA–I per patient (Table 5.1–5.3). The numbers of PSMs and peptides indicated a wide range of depth between the immunopeptidomes of five patients. We also noticed that, for the same patient, the numbers of PSMs and peptides of HLA–II were 3–5 times lower than those of HLA–I (Mel–15 and Mel–16). Most importantly, for all five patients, the number of unlabeled spectra was much higher than the number of identified PSMs, thus highlighting the need of *de novo* peptide sequencing to improve the identification rate.

In step 2, we used the identified normal HLA peptides and PSMs of each patient as patient-specific training data to train DeepNovo. In addition to capturing fragment ions in

Neoantigen discovery workflow	HLA–I	HLA–II
Step 1: Build the immunopeptidome of the patient		
Number of identified peptide-spectrum matches	207,332	39,630
Number of identified database peptides	25,274	6,171
Number of unlabeled spectra	487,233	102,615
Step 2: Train personalized machine learning model		
Number of training PSMs	185,823	35,315
Number of validation PSMs	10,900	2,364
Number of test PSMs	10,609	1,951
Step 3: Personalized *de novo* sequencing		
Number of raw *de novo* peptides	327,415	77,480
Step 4: Quality control		
Number of high-confidence *de novo* peptides	6,444	2,257
Number of *de novo* peptides at 1% FDR	1,259	722
Step 5: Neoantigen selection Binding affinity for patient's HLA alleles; Missense mutation against wild-types; Neoepitope expression level	80	23

Table 5.2: Personalized workflow of neoantigen discovery for patient Mel–16

tandem mass spectra, DeepNovo learns sequence patterns of peptides by modeling them as a special language with an alphabet of 20 amino acid letters. This unique advantage allowed us to train a personalized model to adapt to a specific immunopeptidome of an individual patient and achieved much better accuracy than a generic model (results are shown in a later section). At the same time, it was essential to apply counter-overfitting techniques so that the model could predict new peptides that it had not seen during training. We partitioned the PSMs into training, validation, and test sets (ratio 90-5-5, respectively) and restricted them not to share common peptide sequences. We stopped the training process if there was no improvement on the validation set and evaluated the model performance on the test set. As a result, our personalized model was able to both achieve very high accuracy on an individual immunopeptidome and detect mutated peptides. This approach is particularly useful for missense mutations (the most common source of neoantigens) because they still preserve most patterns in the peptide sequences.

In step 3, we used the personalized DeepNovo model to perform *de novo* peptide sequencing on both labeled spectra (i.e., the PSMs identified in step 1) and unlabeled spectra. Results from labeled spectra were needed for accuracy evaluation and calibrating prediction confidence scores. Peptides identified from unlabeled spectra and not presented in the

Neoantigen discovery workflow	Mel–8 HLA–I	Mel–12 HLA–I
Step 1: Build the immunopeptidome of the patient		
Number of identified peptide-spectrum matches	42,644	36,369
Number of identified database peptides	13,635	10,068
Number of unlabeled spectra	142,794	221,532
Step 2: Train personalized machine learning model		
Number of training PSMs	38,372	32,620
Number of validation PSMs	2,200	1,802
Number of test PSMs	2,072	1,947
Step 3: Personalized *de novo* sequencing		
Number of raw *de novo* peptides	126,813	142,052
Step 4: Quality control		
Number of high-confidence *de novo* peptides	2,109	2,632
Number of *de novo* peptides at 1% FDR	1,235	1,354
Step 5: Neoantigen selection Binding affinity for patient's HLA alleles; Missense mutation against wild-types; Neoepitope expression level	135	169

Table 5.3: Personalized workflow of neoantigen discovery for patient Mel–8 and Mel–12, HLA–I

protein database were defined as "*de novo* peptides" and would be further analyzed in the next steps to find candidate neoantigens of interest.

In step 4, a quality control procedure was designed to select high-confidence *de novo* peptides and to estimate their FDR. We first calculated the accuracy of *de novo* sequencing on the test set of PSMs by comparing the predicted peptide to the true one for each spectrum. DeepNovo also provides a confidence score for each predicted peptide, which can be used as a filter for better accuracy. Since the test set did not share common peptides with the training set, we expected the distribution of accuracy versus confidence score on the test set to be close to that of *de novo* peptides which the model had not seen during training. Thus, we calculated a score threshold at a precision of 95% on the test set and used it to select high-confidence *de novo* peptides (Figure 5.2b). Finally, to estimate the FDR of high-confidence *de novo* peptides, we performed a second-round PEAKS X search of all spectra against a combined list of those peptides and the database peptides (i.e. normal HLA peptides identified in step 1). Only *de novo* peptides identified at 1% FDR were retained. Table 5.4 shows the number of *de novo* HLA peptides identified at 1% FDR,

Patient ID	HLA-I		HLA-II	
	Database	*De novo*	Database	*De novo*
Mel-5	12,998	1,272	MS data not available	
Mel-8	13,635	1,235	MS data not available	
Mel-12	10,068	1,354	MS data not available	
Mel-15	35,551	5,320	9,664	863
Mel-16	25,274	1,259	6,171	722

(FDR: False Discovery Rate; MS: Mass Spectrometry)

Table 5.4: Number of *de novo* and database HLA peptides identified at 1% FDR.

on top of the corresponding number of database peptides for each of the five patients. Our *de novo* peptide sequencing results expanded the immunopeptidomes by 5%–15% (Mel–16 HLA–I: 5%=1,259/25,274; Mel–15 HLA–I: 15%=5,320/35,551; other cases were within that range).

5.1.2 Advantages of Personalized Model over Generic Model

To demonstrate the advantages of our personalized approach, we compared the personalized model of patient Mel–15's HLA–I to a generic model, which had the same neural network architecture but was trained on a combined HLA–I dataset of 9 other patients from the same study[6]. All datasets were derived from the same experiment and instrument, the only difference is the immunopeptidomes of the patients. The combined dataset has 477,482 PSMs, which is 39.9% larger than the Mel–15 dataset. Figure 5.2a shows the accuracy of the personalized model versus the generic model on the Mel–15 test set. As mentioned earlier, this test set did not share common peptides with the Mel–15 training set, so both models had not seen the test peptides during training. The personalized model achieved 14.3% higher accuracy at the peptide level (0.6939 / 0.6070 = 1.143) and 3.8% higher accuracy at the amino acid level (0.8668 / 0.8349 = 1.038), despite its smaller training set. The superiority of the personalized model over the generic one can also be seen from the accuracy-versus-score distribution in Figure 5.2b. At the same level of amino acid accuracy, e.g., 95%, the personalized model required a lower score cutoff, thus allowing more *de novo* peptides to be identified. Indeed, Figure 5.2c shows that the personalized model identified 87.8% more high-confidence *de novo* peptides (16,226 / 8,642 = 1.878) and 38.9% more *de novo* peptides at 1% FDR (5,320 / 3,829 = 1.389). More importantly,

the personalized model was able to capture 6 of 8 target neoantigens of patient Mel–15, while the generic model only recovered 3 of them. Those results demonstrate that our personalized approach substantially improves the accuracy and identification rate of *de novo* peptides by adapting to a specific immunopeptidome of an individual patient.

5.1.3 Analysis of Immune Characteristics of *De novo* HLA peptides

In this section, we studied common immune features of *de novo* HLA peptides and compared them to normal HLA peptides, i.e. those identified by the database search engine in step 1 of the workflow. We also compared to previously reported human epitopes from the Immune Epitope Database (IEDB)[83].

Figure 5.2d shows the distribution of PEAKS X identification scores of *de novo* PSMs against those of database and decoy PSMs for HLA–I peptides of patient Mel–15. The distributions confirm that the *de novo* peptides have strong supporting PSMs as the database peptides and are clearly distinguishable from the decoy ones.

Next, we compared *de novo* and database HLA–I peptides of patient Mel–15 to 18,022 IEDB epitopes, which were retrieved according to the patient's six alleles (HLA–A03:01, HLA–A68:01, HLA–B27:05, HLA–B35:03, HLA–C02:02, HLA–C04:01). The Venn diagram in Figure 5.2e shows that 56 *de novo* peptides have been reported as epitopes in earlier studies. Note that the *de novo* peptides were specific to an individual patient and were not presented in the protein database, so the chance to find them in IEDB is rare. Even 81.4% (28,943 / 35,551) of the database peptides were not found in IEDB. This is due to the large variation of HLA peptides and further emphasizes the importance of our personalized approach. Figure 5.2f further shows that both *de novo* and database peptides have the same characteristic length distribution as IEDB epitopes. For the other four patients Mel–5, Mel–8, Mel–12, and Mel–16, we also found that the length distributions of their *de novo* HLA–I peptides are very similar to those of database peptides, as shown in Figure 5.3a–d. However, for HLA–II, the *de novo* peptides tend to be longer than the database ones (Figure 5.3e, f). We hypothesize that it might be challenging for the database search engine to identify long HLA–II peptides when the digestion rule is unknown.

One of the most widely used measures to assess HLA peptides is their binding affinity to MHC proteins. We used NetMHCpan[38] to predict the binding affinity of the *de novo*, database, and IEDB peptides for HLA–I alleles of patient Mel–15. Figures 5.2g shows the binding affinity distribution of *de novo* peptides, database peptides and IEDB peptides.

From Mann-Whitney U test (p-value > 0.23), we could not reject the null hypothesis that *de novo* peptides have the same binding affinity distribution as IEDB peptides. Furthermore, the majority of the *de novo* peptides were predicted as good binders by multiple criteria: 79.3% (4,220 / 5,320) weak-binding, 51.8% (2,757 / 5,320) strong-binding, and 74.0% (3,938 / 5,320) with binding affinity less than 500 nM (Figure 5.4). Similar results were observed for *de novo* peptides of different HLA–I alleles of the other four patients.

We also applied GibbsCluster[3], an unsupervised alignment and clustering method to identify binding motifs without the need of HLA allele information. We found that the *de novo* peptides of patient Mel–15 were clustered into four groups, of which motifs corresponded exactly to four alleles of the patient (Figure 5.2h). Note that both *de novo* sequencing and unsupervised clustering methods do not use any prior knowledge such as protein database or HLA allele information, yet their combination still revealed the correct binding motifs of the patient. This suggests that our workflow can be used to identify novel HLA peptides of unknown alleles. Results from the database peptides also yielded the same binding motifs (Figure 5.5).

Finally, we used an IEDB tool[1][12] to predict the immunogenicity of *de novo* HLA–I peptides and then compared to database, IEDB, and human immunogenic peptides that were used in that original study (Figure 5.2i). We found that 38.8% (2,065 / 5,320) of the *de novo* peptides had positive predicted immunogenicity (log-likelihood ratio of immunogenic over non-immunogenic[12]). The *de novo* peptides had lower predicted immunogenicity than the database and IEDB peptides. This was expected because the tool had been developed on a limited set of a few thousand well-studied peptides. The predicted immunogenicity of *de novo* HLA–I peptides of the other four patients are provided in Figure 5.6.

Overall, our analysis results confirmed the correctness—and, more importantly—the essential characteristics of *de novo* HLA peptides for immunotherapy. The remaining question is to select candidate neoantigens from *de novo* HLA peptides based on their characteristics.

5.1.4 Neoantigen Selection and Evaluation

We considered several criteria that had been widely used in previous studies for neoantigen selection[56, 65, 13, 6, 42]. Specifically, we checked whether a *de novo* HLA peptide carried one amino acid substitution by aligning its sequence to the Swiss-Prot human protein

[1]http://tools.iedb.org/immunogenicity/

database, and whether that substitution was caused by one single nucleotide difference in the encoding codon. In this paper, we refer to those substitutions as "missense-like mutations." For each mutation, we recorded whether the wild-type peptide was also detected and whether the mutated amino acid was located at a flanking position. For expression level information of a peptide, we calculated the number of its PSMs, their total identification score, and their total abundance. Finally, we used NetMHCpan and IEDB tools to predict the binding affinity and the immunogenicity of a peptide. As a result, we find 10,440 HLA–I and 1,585 HLA–II *de novo* peptides of five patients.

To select candidate neoantigens, we focused on *de novo* HLA peptides that carried one single missense-like mutation. This criterion reduced the number of peptides considerably, e.g. from 5,320 to 328 HLA–I and from 863 to 154 HLA–II peptides of patient Mel–15. We further filtered out peptides with only one supporting PSM or with mutations at flanking positions because they were more error prone and less stable to be effective neoantigens. On average, we obtained 154 HLA–I and 47 HLA–II candidates per patient. Expression level, binding affinity, and immunogenicity can be further used to prioritize candidates for experimental validation of immune response; we avoided using those information as hard filters.

Tab	Position	Ref Allele	Alt Allele	Genename	Transcript ID	Effect	Aa change	wildtype peptide	de novo peptide
X	100687163	G	A	SYTL4	ENST00000263033	missense_variant	Ser363Phe	GRIAFSLKY	GRIAFFLKY
14	32822259	G	A	AKAP6	ENST00000280979	missense_variant	Met1482Ile	KLKLPMIMK	KLKLPIIMK
10	17229543	G	A	VIM	ENST00000224237	missense_variant	Gly41Ser	SLGSALRPSTSRSLY	SLSSALRPSTSRSLY
7	158680743	G	A	NCAPG2	ENST00000441982	missense_variant	Pro134Leu	KPILWRGLK	KLILWRGLK
8	30474849	C	T	RBPMS	ENST00000517860	missense_variant	Pro46Leu	RPFKGYEGSLIK	RLFKGYEGSLIK

Table 5.5: Identified neoantigens for patient Mel–15.Green rows: MHC class 1; yellow row: MHC class 2; red letters: mutated amino acids.

We cross-checked our *de novo* HLA peptides against the nucleotide mutations and mRNA transcripts in the original publication[6]. We identified seven HLA–I and ten HLA–II candidate neoantigens that matched missense variants detected from exome sequencing (Table 5.5 and Table 5.6). The first seven were among eleven neoantigens reported by the authors using proteogenomic approach that required both exome sequencing and proteomics database search. Two HLA–I neoantigens, "GRIAFFLKY" and "KLILWRGLK", had been experimentally validated to elicit specific T-cell responses. We indeed observed that those two peptides had superior immunogenicity, and especially, expression level of

…SYVTTSTRTYSLGSALRPSTSRSLY…	binding %	num_psm	total_score	total_abundance
YVTTSTRTYSLSSALRPSTS		8	573.18	4.26E+06
VTTSTRTYSLSSALRPSTS		5	290.37	2.25E+06
SLSSALRPSTSRSLY	0.08	8	351.79	1.70E+07
SLSSALRPSTSRSLY.1	0.08	11	622.22	3.44E+07
SYVTTSTRTYSLSSALRPSTS		8	666.41	9.78E+06
VTTSTRTYSLSSALRPS		3	115.28	3.03E+06
TTSTRTYSLSSALRPS		6	340.06	1.10E+07
YVTTSTRTYSLSSALRPST		2	104.34	3.68E+05
TTSTRTYSLSSALRPSTS		3	120.71	7.46E+05
YVTTSTRTYSLSSALRPS		2	121.59	8.45E+05
TSTRTYSLSSALRPS	0.43	12	498.53	1.26E+07

Table 5.6: Alignment of candidate mutated peptides against the reference sequence from the MHC class 2 dataset of patient Mel–15. The mutated site is highlighted in green and yellow colors, for reference and mutated amino acids respectively. The columns provide supporting evidence of binding affinity rank (lower is better), number of PSMs, the total confidence score of PSMs, and the total abundance of PSMs. Two candidate neoantigens "SLSSALRPSTSRSLY" and "TSTRTYSLSSALRPS" are highlighted in red color. "SLSSALRPSTSRSLY.1" shows the identification of this peptide from the MHC class 1 dataset.

up to one order of magnitude higher than the other neoantigens (Table 5.5). This observation confirms the critical role of peptide-level expression for effective immunotherapy, in addition to immunogenicity and binding affinity.

The ten HLA–II candidate neoantigens were novel and had not been reported in [6]. They were clustered around a single missense mutation and were a good example to illustrate the complicated digestion of HLA–II peptides (Table 5.6). Eight of them were predicted as strong binders by NetMHCIIpan (rank $\leq 2\%$), two as weak binders (rank $\leq 10\%$). The peptide located at the center of the cluster, "TSTRTYSLSSALRPS", showed both highest expression level and binding affinity, thus representing a promising target for further experimental validation. Interestingly, another peptide, "SLSSALRPSTSRSLY", showed up in both HLA–I and HLA–II datasets with very high expression level (Tables 5.5 and Table 5.6). Using a consensus method of multiple binding prediction tools from IEDB to double-check, we found that this peptide had a binding affinity rank of 0.08%, instead of

4.5% as predicted by NetMHCIIpan, and exhibited a different binding motif from the rest of the cluster. Thus, given its superior binding affinity and expression level, this peptide would also represent a great candidate for immune response validation.

We also investigated the four HLA–I neoantigens that had been reported in [6] but were not detected by our method. Three of them were not supported by good PSMs, and, in fact, DeepNovo and PEAKS X identified alternative peptides that better matched the corresponding spectra (Figure 5.7–5.9). The remaining neoantigen was missed due to a *de novo* sequencing error. We noticed that all four peptides had been originally identified at 5% FDR in stead of 1%, so their signals were possibly too weak for identification. DeepNovo model, and other *de novo* peptide sequencing tools in general, rely mainly on fragment ions in a spectrum to predict its peptide. Thus, the model may miss potential neoantigens that have low MS signals but are still capable of triggering T cell response. However, DeepNovo also includes a recurrent neural network to learn sequence patterns of the peptides to assist the signals from fragment ions, especially when the signals are weak[77]. This aspect of the model can be improved to address the problem of low sensitivity in MS data.

Since exome sequencing only covers 1% of the human genome, many *de novo* HLA peptides that did not match nucleotide variants reported in [6] could have originated from non-coding regions. For instance, Laumont et al. suggested that non-coding regions were the main source of neoantigens[46].

5.2 Discussion

In this study, we proposed a personalized *de novo* peptide sequencing workflow to identify HLA neoantigens directly and solely from mass spectrometry data. The key advantage of our method is the ability of its deep learning model to adapt to a specific immunopeptidome of an individual patient. This personalized approach greatly improved the performance of *de novo* peptide sequencing and allowed accurate identification of mutated HLA peptides. We applied the workflow to five melanoma patients and expanded their immunopeptidomes by 5%–15%. Our analysis also demonstrated that the *de novo* HLA peptides exhibited the same immune characteristics as previously reported human epitopes, including binding affinity, immunogenicity, and expression level, which are essential for effective immunotherapy. On the Mel–15 dataset, we cross-checked our *de novo* HLA peptides against exome sequencing results and discovered 15 neoantigens of both HLA–I and HLA–II, including ten novel HLA–II neoantigens that had not been reported earlier. This result demonstrated the capability of our *de novo* peptide sequencing approach to overcome the challenges of unknown degradation and binding prediction for HLA-II peptides. Last but not least, our

de novo peptide sequencing workflow directly predicted neoantigens from mass spectrometry data and required neither genome-level information nor the patient's HLA alleles, as in existing approaches.

Our current workflow focuses on neoantigens carrying missense mutations. However, there is emerging interest and evidence of neoantigens that result from other sources such as frameshift mutations, non-coding regions, or cis- and trans-splicing events [46, 26]. Thus, our workflow needs further improvement to address those cases. In the cases of frameshift mutations and non-coding regions, integrating genomic information to the current workflow may be needed since it is difficult to confirm those types of mutations if we only look at the protein-level information. For cis- and trans-spliced peptides, while some *de novo* HLA peptides could be explained by cis- or trans-splicing events, the more important question is to establish the statistical significance of such events.

In conclusion, our personalized *de novo* peptide sequencing workflow to predict mutated HLA peptides from mass spectrometry data presents a simple and direct solution to discover neoantigens for cancer immunotherapy. As Newton said, "nature is pleased with simplicity".

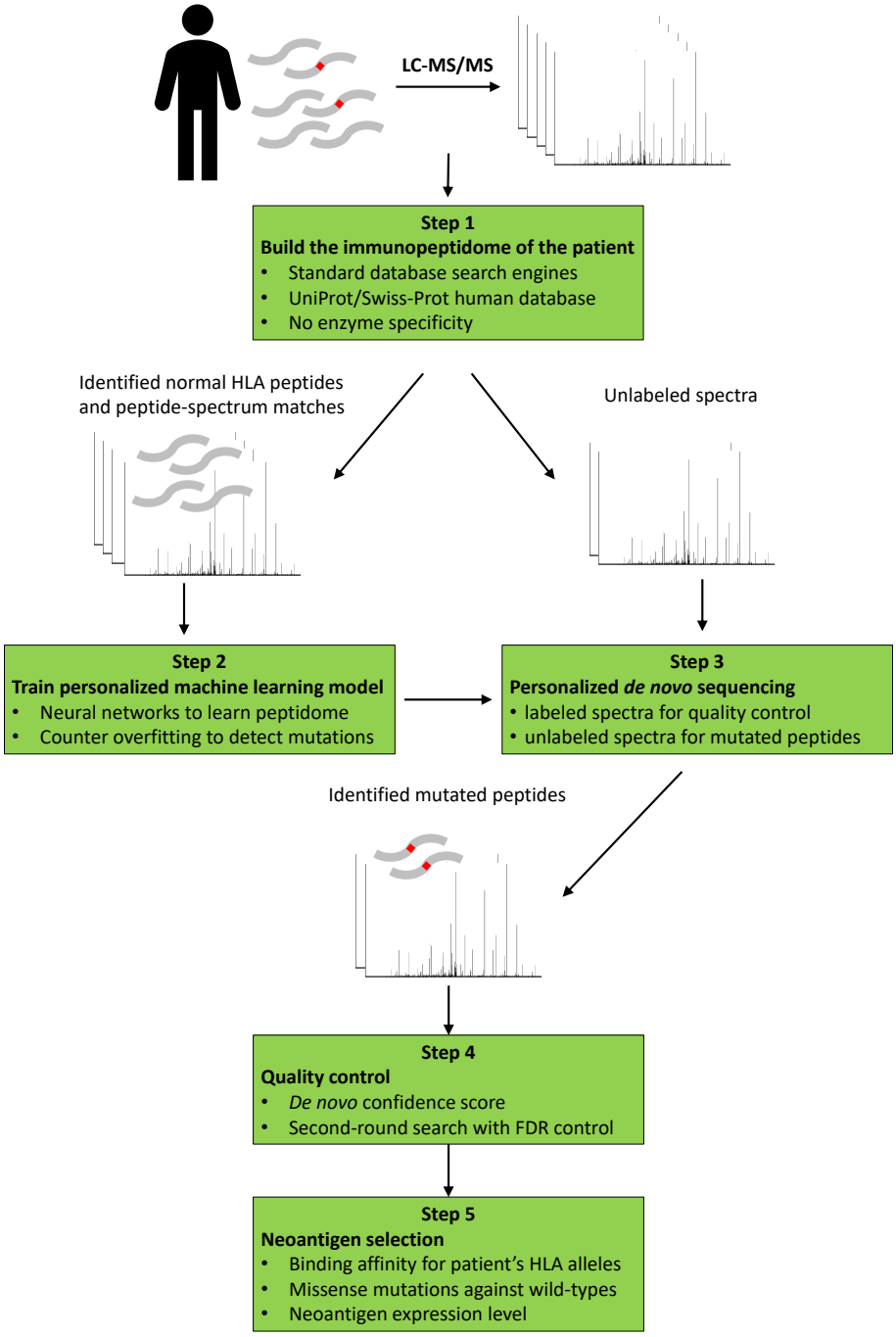


Figure 5.1: Personalized *de novo* sequencing workflow

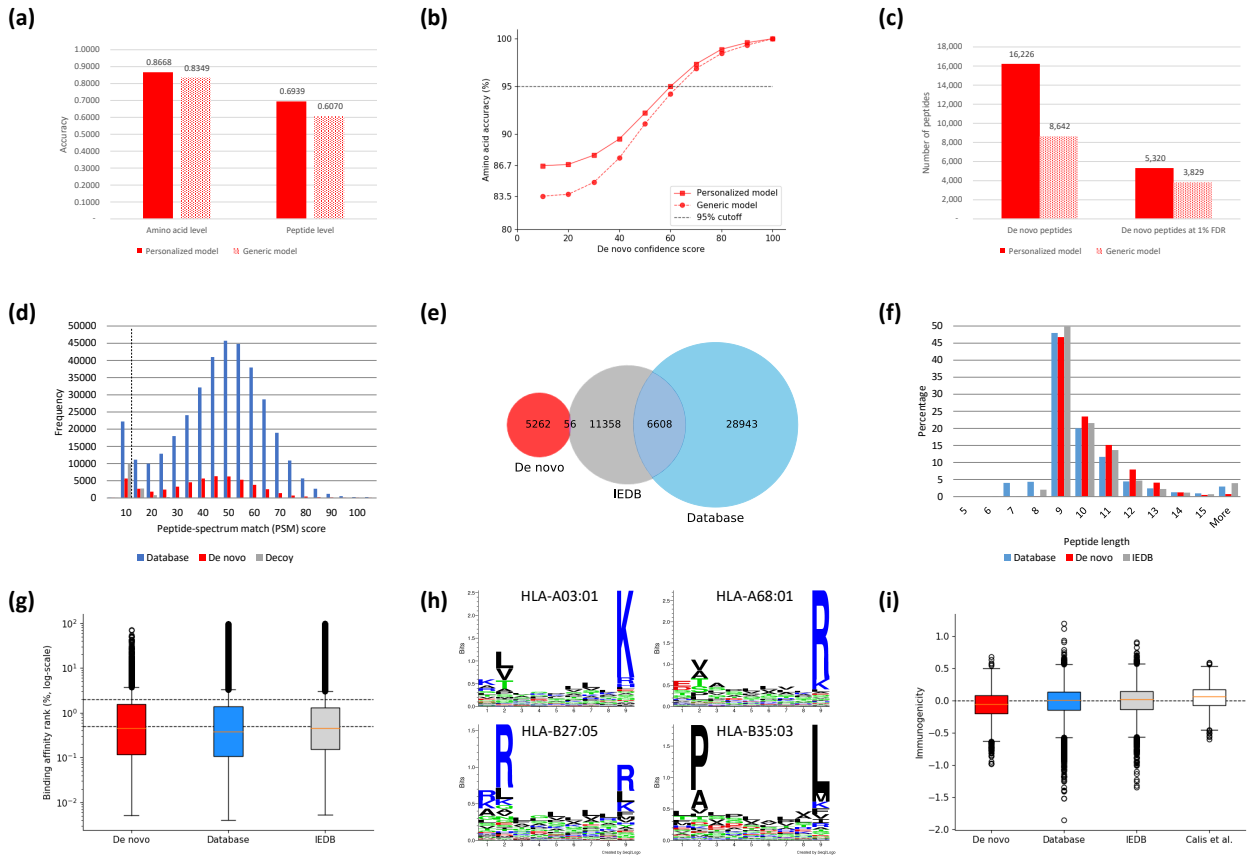


Figure 5.2: Accuracy and immune characteristics of *de novo* HLA–I peptides from patient Mel-15 dataset. (a) Accuracy of *de novo* peptides predicted by personalized and generic models. (b) Distribution of amino acid accuracy versus DeepNovo confidence score for personalized and generic models. (c) Number of *de novo* peptides identified at high-confidence threshold and at 1% FDR by personalized and generic models. (d) Distribution of identification scores of *de novo*, database, and decoy peptide-spectrum matches. The dashed line indicates 1% FDR threshold. (e) Venn diagram of *de novo*, database, and IEDB peptides. (f) Length distribution of *de novo*, database and IEDB peptides. (g) Distribution of binding affinity ranks of *de novo*, database, and IEDB peptides. Lower rank indicates better binding affinity. The two dashed lines correspond to the ranks of 0.5% and 2%, which indicate strong and weak binding, respectively, by NetMHCpan. (h) Binding sequence motifs identified from *de novo* peptides by GibbsCluster. (i) Immunogenicity distribution of *de novo*, database, IEDB, and Calis et al.'s peptides[12].

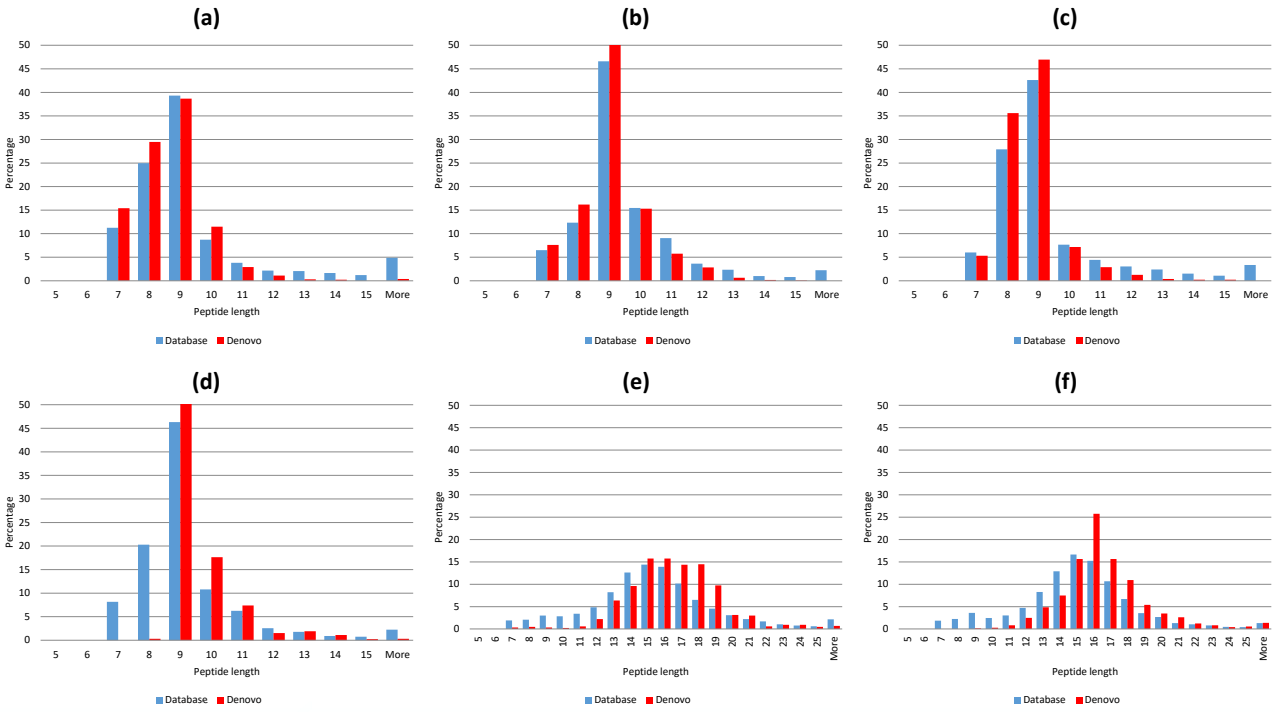


Figure 5.3: Length distributions of HLA *de novo* and database peptides. (a) Mel–5 HLA–I;(b) Mel–8 HLA–I; (c) Mel–12 HLA–I; (d) Mel–16 HLA–I; (e) Mel–15 HLA–II; (f) Mel–16 HLA–II

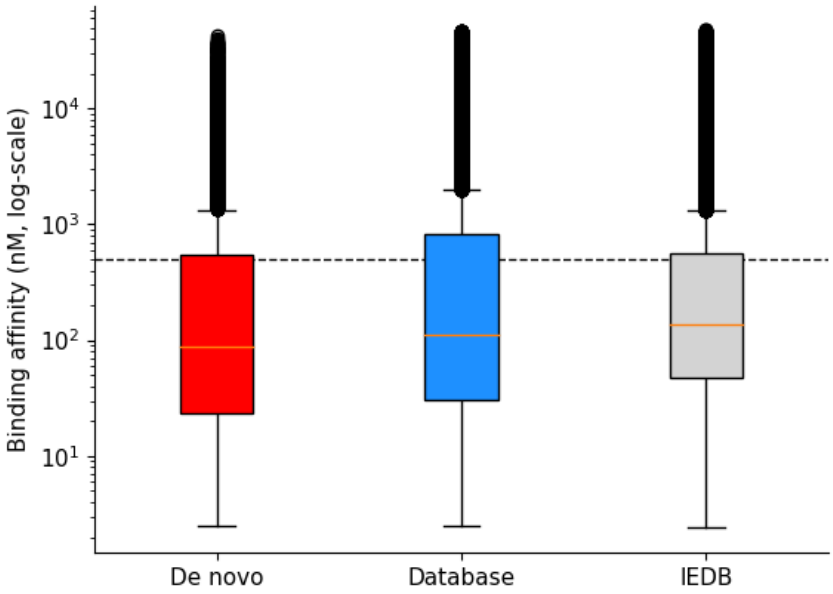


Figure 5.4: Binding affinity distributions of *de novo*, database and IEDB HLA–I peptides of patient Mel–15. The dashed line indicates the value of 500 nM, a common threshold to select good binders.

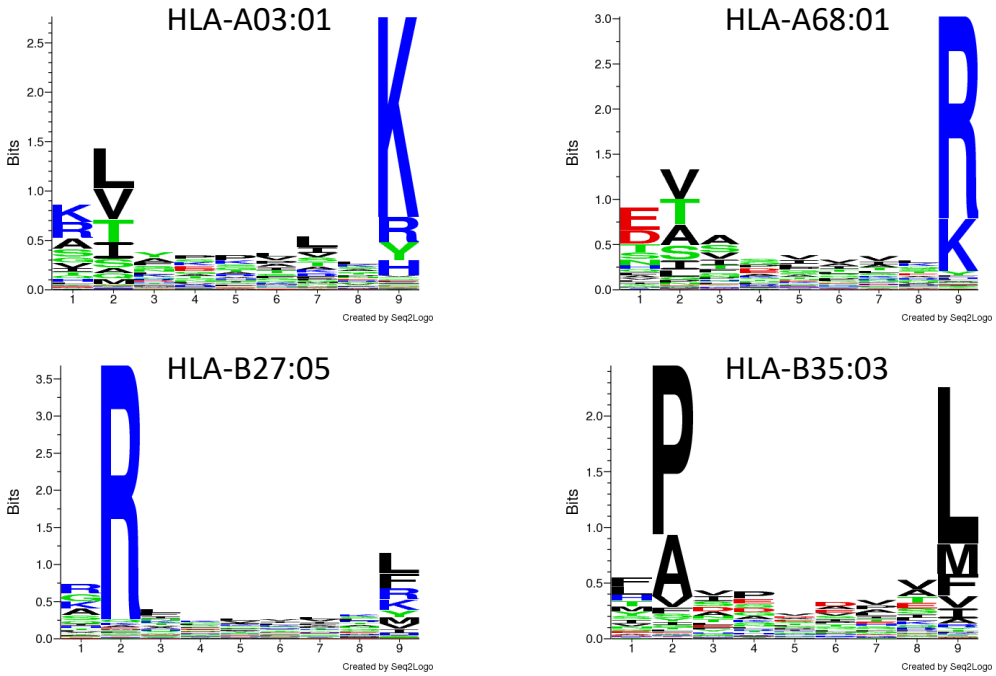


Figure 5.5: Binding motifs of database HLA–I peptides of patient Mel–15

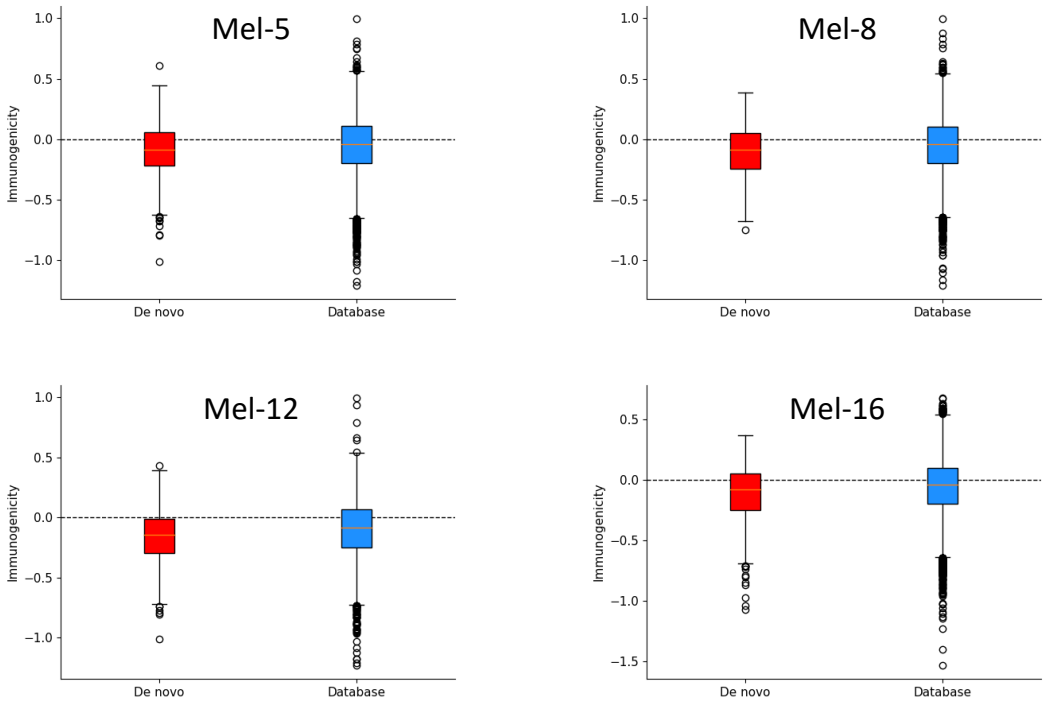
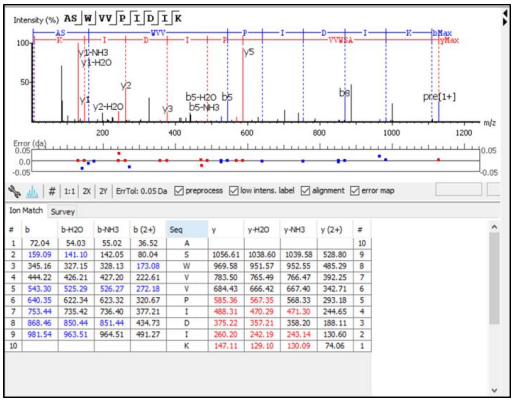


Figure 5.6: Immunogenicity of *de novo* and database HLA–I peptides

Fraction: 20141208_QEp7_MiBa_SA_HLA-I-p_MM15_4_B.raw
Scan ID: 49534
Retention time: 82.974
M/z: 564.327
Charge: 2

MaxQuant



DeepNovo

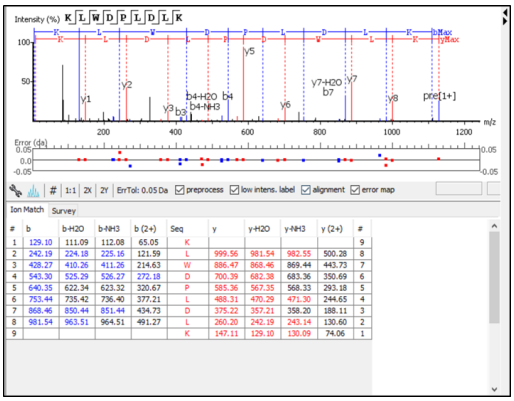
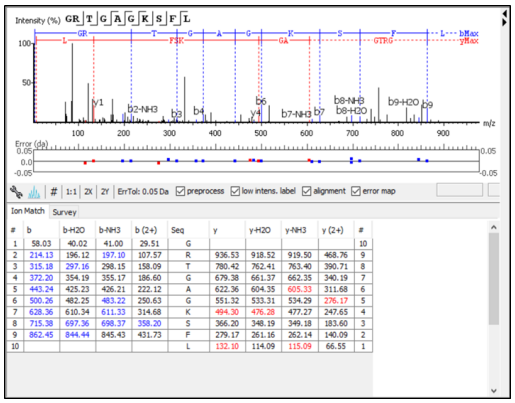


Figure 5.7: MaxQuant and DeepNovo spectrum identification difference 1

Fraction: 20141210_QEp7_MiBa_SA_HLA-I-p_MM15_2_B_1.raw
Scan ID: 21931
Retention time: 37.371
M/z: 331.854
Charge: 3

MaxQuant



DeepNovo & PEAKS DB

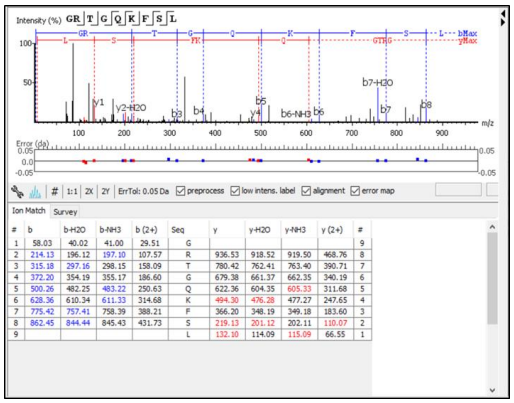
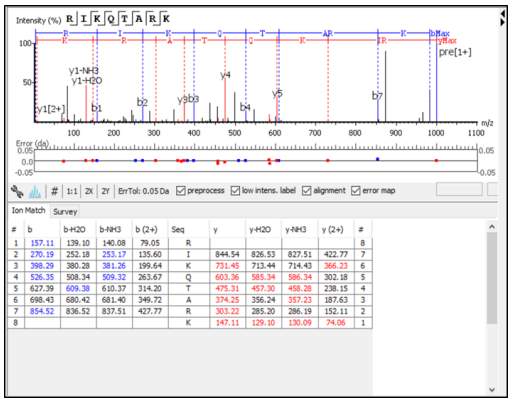


Figure 5.8: MaxQuant and DeepNovo spectrum identification difference 2

Fraction: 20141208_QEp7_MiBa_SA_HLA-I-p_MM15_3_B.raw
Scan ID: 2606
Retention time: 6.317
M/z: 334.217
Charge: 3

MaxQuant



DeepNovo

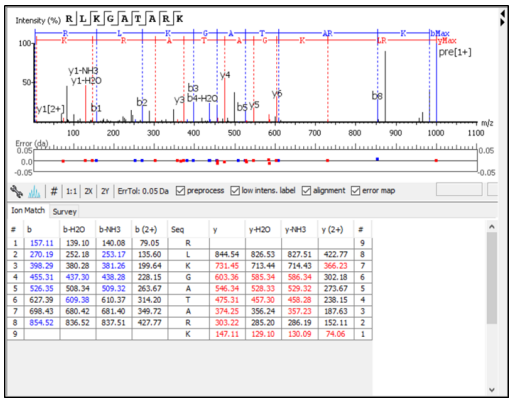


Figure 5.9: MaxQuant and DeepNovo spectrum identification difference 3

Chapter 6

Conclusions and Future Research

6.1 Impact of this Thesis

In Chapter 3, we developed the first *de novo* peptide sequencing tool for DIA data. Our proposed DeepNovo-DIA model could identify peptides for more features, especially those of low abundance. Thus, the combination of DIA and *de novo* sequencing has the potential to help scientists discover novel peptides and enable more complete profiling of biological samples.

In Chapter 4, we developed PointNovo, a *de novo* sequencing model that does not suffer from the accuracy-speed trade-off and outperforms the previous state-of-the-art method, DeepNovo, by a significant margin of at least 15%. Our proposed PointNovo model could directly benefit from the higher resolution data generated by next-generation mass spectrometers, without any increase in computation complexity. We demonstrated that PointNovo could better discriminate between pairs of amino acids that have similar mass (like M(Oxidation) and F). The predicted log probability could be used as PSM scores for doing database searching and the identification rate is at least comparable with existing database search tools. Our novel method of spectrum representation and feature extraction have great potentials for other important problems in MS. The promising results given by PointNovo further confirm that domain-specific expertise is still important and valuable in the era of deep learning. Even though the commonly used DNNs structures are powerful feature extractors, researchers could still expect to observe improvements by proper feature engineering and a carefully designed model structure.

Finally, in Chapter 5, we developed the first personalized sequencing workflow that finds neoantigens directly and solely from mass spectrometry data. Compared to the current

prevalent proteogenomics method, our approach could find more candidate neoantigens, including those resulting from non-coding region or alternative splicing.

6.2 Future Research

Our following research interests are as follows:

- Aside from the mass spectrum, liquid chromatography MS provides another dimension of data: retention time (RT). Previous research on spectral library search and database search showed that including the RT information significantly improved the performance. However, since RT is a peptide property, it is generally hard to use RT in *de novo* sequencing. One approach is to use Monte Carlo tree search to replace the current beam search strategy in PointNovo. This way, we may guide the searching process to favor a *de novo* sequence with the desired RT.

- In Chapter 4, I used a T Net structure to process the order invariant feature matrix. Recent research on point clouds classification showed that including graph structures of the point clouds helps the model to make better predictions. On the other hand, an MS2 spectrum could be represented as a graph (often referred to as the spectrum graph) and *de novo* peptide sequencing problems could be organized as finding the best path on the spectrum graph[19]. So far my experiments of using Graph Neural Networks (GNNs) to process mass spectra have not shown any improvements on the task of *de novo* peptide sequencing. Nevertheless, it is worth putting more effort in investigating the use of GNNs in MS.

- In Chapter 5, our workflow found a list of candidate neoantigens in which only very few are reported by the existing proteogenomic approach. Possible explanation are that some neoantigens might be from a non-coding region of the genome[46], or they might be results of trans or cis splicing[21]. Next, we plan to generate immunopeptidomics MS dataset, together with whole genome and RNA sequencing results to validate our hypothesis. Hopefully, clinical trials will validate the candidate neoantigens identified in our workflow.

Copyright Permissions

Parts of Chapter 3, 4 and 5 are reprinted from the following papers.

- Tran, Ngoc Hieu, 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, no. 1 (2019): 63–66.
- Qiao, Rui, Ngoc Hieu Tran, Lei Xin, Baozhen Shan, Ming Li, and Ali Ghodsi. "Deepnovov2: Better de novo peptide sequencing with deep learning." *arXiv preprint* arXiv:1904.08514 (2019).
- Tran, Ngoc Hieu, Rui Qiao, Lei Xin, Xin Chen, Baozhen Shan, and Ming Li. "Identifying neoantigens for cancer vaccines by personalized deep learning of individual immunopeptidomes." *bioRxiv* (2019): 620468.

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] Jennifer G Abelin, Derin B Keskin, Siranush Sarkizova, Christina R Hartigan, Wandi Zhang, John Sidney, Jonathan Stevens, William Lane, Guang Lan Zhang, Thomas M Eisenhaure, et al. Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*, 46(2):315–326, 2017.

[3] Massimo Andreatta, Bruno Alvarez, and Morten Nielsen. Gibbscluster: unsupervised clustering and alignment of peptide sequences. *Nucleic acids research*, 45(W1):W458–W463, 2017.

[4] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the mhc class i system. *Bioinformatics*, 32(4):511–517, 2015.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] 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*, 7:13404, 2016.

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

[8] Aleyda Benitez-Amaro, Chiara Pallara, Laura Nasarre, Andrea Rivas-Urbina, Sonia Benitez, Angela Vea, Olga Bornachea, David de Gonzalo-Calvo, Gabriel Serra-Mir, Sandra Villegas, et al. Molecular basis for the protective effects of low-density lipoprotein receptor-related protein 1 (lrp1)-derived peptides against ldl aggregation. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1861(7):1302–1316, 2019.

[9] Oliver M Bernhardt, Nathalie Selevsek, Ludovic C Gillet, Oliver Rinner, Paola Picotti, Ruedi Aebersold, and Lukas Reiter. Spectronaut: A fast and efficient algorithm for mrm-like processing of data independent acquisition (swath-ms) data. *Biognosys. ch*, 2012.

[10] 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.

[11] Brendan Bulik-Sullivan, Jennifer Busby, Christine D Palmer, Matthew J Davis, Tyler Murphy, Andrew Clark, Michele Busby, Fujiko Duke, Aaron Yang, Lauren Young, et al. Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification. *Nature biotechnology*, 37(1):55, 2019.

[12] Jorg JA Calis, Matt Maybeno, Jason A Greenbaum, Daniela Weiskopf, Aruna D De Silva, Alessandro Sette, Can Keşmir, and Bjoern Peters. Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS computational biology*, 9(10):e1003266, 2013.

[13] Beatriz M Carreno, Vincent Magrini, Michelle Becker-Hapak, Saghar Kaabinejadian, Jasreet Hundal, Allegra A Petti, Amy Ly, Wen-Rong Lie, William H Hildebrand, Elaine R Mardis, et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific t cells. *Science*, 348(6236):803–808, 2015.

[14] Ting Chen, Ming-Yang Kao, Matthew Tepel, John Rush, and George M Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8(3):325–337, 2001.

[15] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, et al. pnovo: de novo peptide sequencing and identification using hcd spectra. *Journal of proteome research*, 9(5):2713–2724, 2010.

[16] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2018.

[17] 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, 2008.

[18] Balázs Csanád Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[19] Vlado Dančík, Theresa A Addona, Karl R Clauser, James E Vath, and Pavel A Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology*, 6(3-4):327–342, 1999.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Gianfranco Di Segni, Serena Gastaldi, and Glauco P Tocchini-Valentini. Cis-and trans-splicing of mrnas mediated by trna sequences in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 105(19):6864–6869, 2008.

[22] 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, 2015.

[23] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. In *Proteome bioinformatics*, pages 55–71. Springer, 2010.

[24] 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.

[25] R Evans, J Jumper, J Kirkpatrick, L Sifre, TFG Green, C Qin, A Zidek, A Nelson, A Bridgland, H Penedones, et al. De novo structure prediction with deeplearning based scoring. *Annu Rev Biochem*, 77:363–382, 2018.

[26] Pouya Faridi, Chen Li, Sri H Ramarathinam, Julian P Vivian, Patricia T Illing, Nicole A Mifsud, Rochelle Ayala, Jiangning Song, Linden J Gearing, Paul J Hertzog, et al. A subset of hla-i peptides are not genomically templated: Evidence for cis-and trans-spliced peptide ligands. *Science immunology*, 3(28):eaar3947, 2018.

[27] Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem, and Joachim M Buhmann. Novohmm: a hidden markov model for de novo peptide sequencing. *Analytical chemistry*, 77(22):7265–7273, 2005.

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

[29] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[30] Sebastian Giwa, Jedediah K Lewis, Luis Alvarez, Robert Langer, Alvin E Roth, George M Church, James F Markmann, David H Sachs, Anil Chandraker, Jason A Wertheim, et al. The promise of organ and tissue preservation to transform medicine. *Nature biotechnology*, 35(6):530, 2017.

[31] Elizabeth Haythorne, Maria Rohm, Martijn van de Bunt, Melissa F Brereton, Andrei I Tarasov, Thomas S Blacker, Gregor Sachse, Mariana Silva dos Santos, Raul Terron Exposito, Simon Davis, et al. Diabetes causes marked inhibition of mitochondrial metabolism in pancreatic β-cells. *Nature communications*, 10(1):1–17, 2019.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385. IEEE, 2019.

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

[35] Robert E Hollingsworth and Kathrin Jansen. Turning the corner on therapeutic cancer vaccines. *NPJ vaccines*, 4(1):7, 2019.

[36] Zhuting Hu, Patrick A Ott, and Catherine J Wu. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nature Reviews Immunology*, 18(3):168, 2018.

[37] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[38] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. Netmhcpan-4.0: improved peptide–mhc class i interaction predictions integrating eluted ligand and peptide binding affinity data. *The Journal of Immunology*, 199(9):3360–3368, 2017.

[39] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923, 2007.

[40] Eugene Kapp and Frédéric Schütz. Overview of tandem mass spectrometry (ms/ms) database search algorithms. *Current protocols in protein science*, 49(1):25–2, 2007.

[41] 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.

[42] Derin B Keskin, Annabelle J Anandappa, Jing Sun, Itay Tirosh, Nathan D Mathewson, Shuqiang Li, Giacomo Oliveira, Anita Giobbie-Hurder, Kristen Felt, Evisa Gjini, et al. Neoantigen vaccine generates intratumoral t cell responses in phase ib glioblastoma trial. *Nature*, 565(7738):234, 2019.

[43] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] 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.

[46] Céline M Laumont, Krystel Vincent, Leslie Hesnard, Éric Audemard, Éric Bonneil, Jean-Philippe Laverdure, Patrick Gendron, Mathieu Courcelles, Marie-Pierre Hardy, Caroline Côté, et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Science translational medicine*, 10(470):eaau5516, 2018.

[47] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[48] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

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

[50] 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.

[51] Bin Ma, Kaizhong Zhang, and Chengzhi Liang. An effective algorithm for peptide de novo sequencing from ms/ms spectra. *Journal of Computer and System Sciences*, 70(3):418–430, 2005.

[52] Spencer D Martin, Scott D Brown, Darin A Wick, Julie S Nielsen, David R Kroeger, Kwame Twumasi-Boateng, Robert A Holt, and Brad H Nelson. Low mutation burden in ovarian cancer may limit the utility of neoantigen-targeted vaccines. *PloS one*, 11(5):e0155189, 2016.

[53] 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.

[54] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[55] Russell P Newton, A Gareth Brenton, Chris J Smith, and Edward Dudley. Plant proteome analysis by mass spectrometry: principles, problems, pitfalls and recent developments. *Phytochemistry*, 65(11):1449–1485, 2004.

[56] Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217, 2017.

[57] Maria R Parkhurst, James C Yang, Russell C Langan, Mark E Dudley, Debbie-Ann N Nathan, Steven A Feldman, Jeremy L Davis, Richard A Morgan, Maria J Merino, Richard M Sherry, et al. T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Molecular Therapy*, 19(3):620–626, 2011.

[58] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[59] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[61] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[62] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[63] Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, et al. Mutational landscape determines sensitivity to pd-1 blockade in non–small cell lung cancer. *Science*, 348(6230):124–128, 2015.

[64] 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, 2014.

[65] Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D Tadmor, Ulrich Luxemburger, Barbara Schrörs, et al. Personalized rna mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 547(7662):222, 2017.

[66] Samuel S Schoenholz, Sean Hackett, Laura Deming, Eugene Melamud, Navdeep Jaitly, Fiona McAllister, Jonathon O'Brien, George Dahl, Bryson Bennett, Andrew M Dai, et al. Peptide-spectra matching from weak supervision. *arXiv preprint arXiv:1808.06576*, 2018.

[67] Melanie J Shears, Raja Sekhar Nirujogi, Kristian E Swearingen, Santosh Renuse, Satish Mishra, Panga Jaipal Reddy, Robert L Moritz, Akhilesh Pandey, and Photini Sinnis. Proteomic analysis of plasmodium merosomes: the link between liver and blood stages in malaria. *Journal of proteome research*, 18(9):3404–3418, 2019.

[68] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[69] Seo Young Sim, Yu Ri Choi, Jun Hyung Lee, Jae Min Lim, Seung-Eun Lee, Kwang Pyo Kim, Jin Young Kim, Seung Hyeun Lee, and Min-Sik Kim. In-depth proteomic analysis of human bronchoalveolar lavage fluid toward the biomarker discovery for lung cancers. *PROTEOMICS–Clinical Applications*, 13(5):1900028, 2019.

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[71] C. C. Smith, S. R. Selitsky, S. Chai, P. M. Armistead, B. G. Vincent, and J. S. Serody. Alternative tumour-specific antigens. *Nat. Rev. Cancer*, 19(8):465–478, Aug 2019.

[72] Philip Sobolesky, Celeste Parry, Baylye Boxall, Randall Wells, Stephanie Venn-Watson, and Michael G Janech. Proteomic analysis of non-depleted serum proteins from bottlenose dolphins uncovers a high vanin-1 phenotype. *Scientific reports*, 6(1):1–10, 2016.

[73] Maria Tagliamonte, Annacarmen Petrizzo, Maria Lina Tornesello, Franco M Buonaguro, and Luigi Buonaguro. Antigen-specific vaccines for cancer treatment. *Human vaccines & immunotherapeutics*, 10(11):3332–3346, 2014.

[74] 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, 2017.

[75] 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.

[76] Ngoc Hieu Tran, M Ziaur Rahman, Lin He, Lei Xin, Baozhen Shan, and Ming Li. Complete de novo assembly of monoclonal antibody sequences. *Scientific reports*, 6:31730, 2016.

[77] 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.

[78] 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, 2015.

[79] Pawel L Urban. Quantitative mass spectrometry: an overview, 2016.

[80] Eliezer M Van Allen, Diana Miao, Bastian Schilling, Sachet A Shukla, Christian Blank, Lisa Zimmer, Antje Sucker, Uwe Hillen, Marnix H Geukes Foppen, Simone M Goldinger, et al. Genomic correlates of response to ctla-4 blockade in metastatic melanoma. *Science*, 350(6257):207–211, 2015.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[82] 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, 2004.

[83] Randi Vita, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, Alessandro Sette, et al. The immune epitope database (iedb) 3.0. *Nucleic acids research*, 43(D1):D405–D412, 2014.

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

[85] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018.

[86] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[87] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[88] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[89] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. pnovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i183–i190, 2019.

[90] 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):M111–010587, 2012.

[91] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[92] Ying Zhu, Maowei Dou, Paul D Piehowski, Yiran Liang, Fangjun Wang, Rosalie K Chu, William B Chrisler, Jordan N Smith, Kaitlynn C Schwarz, Yufeng Shen, et al. Spatially resolved proteome mapping of laser capture microdissected tissue with automated sample transfer to nanodroplets. *Molecular & Cellular Proteomics*, 17(9):1864–1874, 2018.