การถอดรหัสเปปไทด์ตามตำแหน่งด้วยเครือข่ายตัวเข้ารหัสและตัวถอดรหัสโดยรู้ว่าเมื่อใดไม่ควรตอบ

นายกรวีร์ การุณรัตนกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

KNOWING WHEN NOT TO ANSWER: POSITIONAL PEPTIDE SEQUENCING WITH ENCODER-DECODER NETWORKS

Mr. Korrawe Karunratanakul

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2018

Thesis Title	KNOWING WHEN NOT TO ANSWER: POSITIONAL PEPTIDE SEQUENCING WITH ENCODER-DECODER NETWORKS
By	Mr. Korrawe Karunratanakul
Field of Study	Computer Engineering
Thesis Advisor	Ekapol Chuangsuwanich, Ph.D.
Thesis Co-advisor	Sira Sriswasdi, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

Dean of the Faculty of Engineering

.....

(Prof. Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

..... Chairman

(Duangdao Wichadakul, Ph.D.)

..... Thesis Advisor

(Ekapol Chuangsuwanich, Ph.D.)

..... Thesis Co-advisor

(Sira Sriswasdi, Ph.D.)

..... External Member

(Sissades Tongsima, Ph.D.)

กรวีร์ การุณรัตนกุล: การถอดรหัสเปปไทด์ตามตำแหน่งด้วยเครือข่ายตัวเข้ารหัสและตัวถอดรหัสโดยรู้ว่าเมื่อใดไม่ควรตอบ. (KNOWING WHEN NOT TO ANSWER: POSITIONAL PEPTIDE SEQUENCING WITH ENCODER-DECODER NETWORKS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : อ. ดร.เอกพล ช่วงสุวนิช, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : อ. ดร.สิระ ศรีสวัสดิ์ 56 หน้า.

การถอดรหัสเปปไทด์นั้นเป็นองค์ประกอบสำคัญสำหรับการศึกษาโปรตีน โดยทั่วไปแล้วการวิเคราะห์ข้อมูล mass spectrum นั้นจะศึกษาเพียงสายของกรดอะมิโนที่ปรากฏอยู่ในฐานข้อมูลเท่านั้น ทำให้การค้นหาสายเปปไทด์แบบใหม่ที่อาจเกิดจากการกลายพันธุ์นั้นทำได้ยาก วิธีการถอดรหัสด้วยวิธีดีโนโวแก้ไขข้อจำกัดนี้ด้วยการถอดรหัสสายเปปไทด์โดยตรงจากข้อมูล mass spectrum โดยใช้ความรู้เกี่ยวกับกระบวนการแตกตัวของไอออน ทำให้ไม่จำเป็นต้องใช้ฐานข้อมูลโปรตีนช่วย อย่างไรก็ดี วิธีการดังกล่าวยังมีข้อจำกัดด้านความแม่นยำและต้องการการตรวจทานโดยผู้เชี่ยวชาญ วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการถอดรหัสเปปไทด์ด้วยวิธีการดีโนโวแบบใหม่ชื่อ SMSNet โดยใช้โมเดล deep learning เข้าช่วย โดยยังสามารถทำนายกรดอะมิโนได้อย่างครอบคลุมในระดับความแม่นยำของกรดอะมิโนที่ 95% งานฉบับนี้เสนอขั้นตอน ถอดรหัส ตัดออก และสืบค้น เพื่อตัดผลทำนายในตำแหน่งที่มีความกำกวมออกและใช้ข้อมูลจากฐานข้อมูลโปรตีนช่วยเพื่อให้ทำนายสายเปปไทด์ได้ถูกต้องทั้งเส้น นอกจากนี้ งานนี้ได้นำเสนอการใช้ rescorer ในการแก้ไขคะแนนความมั่นใจสำหรับผลทำนายในแต่ละตำแหน่ง ซึ่งส่งผลให้สามารถแยกกลุ่มคะแนนความมั่นใจสำหรับคำตอบที่ถูกต้องและคำตอบที่ผิดได้ดียิ่งขึ้น เมื่อประกอบทุกขั้นตอนวิธีในงานวิจัยฉบับนี้เข้าด้วยกัน พบว่า SMSNet สามารถทำนายสายเปปไทด์ได้ในประสิทธิภาพที่ใกล้เคียงกับการทำนายด้วยฐานข้อมูลในการทดลองจริง

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่ออ.ที่ปรึกษาหลัก .....
ปีการศึกษา	2561	ลายมือชื่อ.ที่ปรึกษาร่วม .....

6170109021: MAJOR COMPUTER ENGINEERING

KEYWORDS: DE NOVO PEPTIDE SEQUENCING / DEEP LEARNING

KORRAWE KARUNRATANAKUL : KNOWING WHEN NOT TO ANSWER: POSITIONAL PEPTIDE SEQUENCING WITH ENCODER-DECODER NETWORKS. ADVISOR : EKAPOL CHUANGSUWANICH, Ph.D., THESIS COADVISOR : SIRA SRISWASDI, Ph.D., 56 pp.

Peptide sequencing is an important component for understanding the characterization of proteins. Typical analyses of mass spectrometry data only identify amino acid sequences that exist in reference databases. This restricts the possibility of discovering new peptides such as those that contain uncharacterized mutations or originate from unexpected proteins. *De novo* peptide sequencing approaches address this limitation by directly deriving peptides from MS/MS spectra using the knowledge of the ion fragmentation process but often suffer from low accuracy and require extensive validation by experts. In this thesis, we develop SMSNet, a deep learning-based hybrid *de novo* peptide sequencing model that achieves >95% amino acid accuracy while retaining good identification coverage. We propose a sequence-mask-search framework which allows the model to recover full-sequence peptide predictions from known database in case the predictions contain ambiguous amino acid positions. Additionally, because the confidence scores of each amino acid are often affected by the predictions in the previous positions, we propose the use of external rescorer for adjusting the scores, which leads to better separation between correct and incorrect amino acids. Using techniques described and proposed in this thesis, we are able to recover a large number of peptides which are in accordance with predictions using database searching techniques, suggesting the potential of SMSNet on other real-life proteomics studies.

Department	: Computer Engineering	Student's Signature	.....
Field of Study	: Computer Engineering	Advisor's Signature	.....
Academic Year	: 2018	Co-advisor's Signature	.....

Acknowledgements

This work is supported by Chula Computer Engineering Graduate Scholarship for CP Alumni from the department of computer engineering, Faculty of Engineering, Chulalongkorn University. We gratefully acknowledge the contribution of mass spectrometry dataset from the Proteomics and Metabolomics Core Facility at The Wistar Institute, the support of the Chulalongkorn Academic Advancement into Its 2nd Century Project, and the donation of TITAN Xp graphic card used in this research by the NVIDIA Corporation. We especially thank Mark A. Knepper (the Epithelial Systems Biology Laboratory, National Heart, Lung, and Blood Institute, National Institute of Health, USA) and Trairak Pisitkul (Systems Biology Center, Chulalongkorn University, Thailand) for facilitating access to computing resources at the National Institute of Health, USA, and for providing critical advice on the manuscript. This work utilized high performance computing resources of the Biowulf cluster, National Institute of Health, USA (http://hpc.nih.gov) and the Center of Excellence for Medical Genomics, Faculty of Medicine, Chulalongkorn University, Thailand.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

Proteins, rather than DNA or RNA, are fundamental molecules that perform critical, basic functions within each cell. Studying proteins therefore provide more direct information on the biological states of the sample of interest. Nonetheless, because proteins consist of amino acids that could not be directly sequenced residue-by-residue, an alternative approach is needed. Mass spectrometry (MS) is an experimental method that allows one to identify amino acid sequences of proteins that are present in a biological sample. During an MS data acquisition stage, proteins are digested into smaller segments, called peptides, which are further broken into fragment ions. For each peptide, the profile of masses and abundances of all fragment ions derived from it is called an MS/MS spectrum. The identity of the peptides, and subsequently the proteins, can then be deduced from these mass spectra.

De novo peptide sequencing from MS/MS spectra is an important building block for characterizing novel protein sequence. Typically, peptide sequencing techniques could be categorized into two main groups: database searching and *de novo* peptide sequencing. Most database search algorithms share the same core principle, they attempt to match the new MS/MS spectrum to the known peptide sequence in the database, and return the sequence with highest similarity score, if any, as an output. This method yields accurate results for species with complete database but performs poorly for novel peptide sequence, such as in less-studied species or peptides with mutation. On the other hand, *de novo* peptide sequencing technique focuses on determining peptide sequence directly from the pattern in MS/MS spectra, thereby being better at handling novel protein sequences.

In the past years, *de novo* peptide sequencing was typically treated as a complex global optimization problem, searching the sequence with best score according to a set of predefined structures. This method usually suffers from the large search space of possible sequences. On the contrary, neural networks are known for it ability to learn complex patterns from the data and have already showed great performance in sequence generating problems. This type of problems share many key features with the way peptide sequence is derived from the spectrum.

Despite many readily available tools for de novo peptide sequencing (Tran et al., 2017; Frank

and Pevzner, 2005; Ma et al., 2003; Ma, 2015), reliably predicting peptides from routine tandem mass spectrometry (MS/MS) spectra is still challenging. Recently, DeepNovo (Tran et al., 2017) has shown that deep learning approach could be effectively applied to de novo peptide sequencing problem and could outperform other standard tools such as PepNovo (Frank and Pevzner, 2005), PEAKS (Ma et al., 2003), and Novor (Ma, 2015). However, performances of these existing tools might not be accurate enough and still require careful validation by experts. Coincidentally, parts of the limitations of these de novo peptide sequencing tools might be due to the nature of MS/MS spectrum. We observed that the problems with sequencing from MS/MS spectrum could be categorized into three main problems: missing data, noise, and ambiguity. In typical MS/MS experiments, evidence for some amino acid residues in the spectrum might be missing entirely or the resulting mass spectrum might contain a lot of noises that are unrelated to the peptide being characterized. In addition, having missing data and noise introduces more ambiguity to the process as noisy spectrum results in more 'paths' for possible peptides. Because peptide sequencing is based on interpreting observed masses as the total weights of some possible amino acid sequences, it remains unclear whether forcibly making predictions even at the position that are subjected to these problems would yield accurate results.

In this thesis, we introduce SMSNet, the Sequence-Mask-Search framework for recovering peptides from MS/MS spectra that addresses these limitations, along with our neural networks-based de novo sequencing tool. First, SMSNet leverages domain-specific deep learning architecture to predict accurate peptide sequences and their positional scores indicating how confidence the model is at those positions. Second, to avoid misinterpreting incomplete MS/MS spectra, SMSNet is allowed to mask low-confidence amino acids from its prediction, producing a partial sequence that excludes low confidence regions for each spectrum. Third, by designing a new searching scheme that matches the partial predictions to the database of known species, we could recover the full peptides sequence. Combining these three steps, we show that our model could enable accurate discovering of peptide in a wide range of proteomes.

1.1 Main Contributions

The main contributions of this thesis can be summarized as follows:

- Developed a neural networks model for *de novo* peptide sequencing. This model achieved state-of-the-art results on various dataset.

- Proposed a rescorer to adjust the positional score of each amino acids prediction, which leads to better separation of score distribution between correct and incorrect predictions.
- Proposed the sequence-mask-search framework, SMSNet, for recovering peptides from MS/MS spectra. SMSNet combined both *de novo* sequencing technique and database searching to resolve ambiguous position in the predictions.
- Investigated the use of SMSNet in real-life preoteomics studies.

This thesis propose a complete and robust deep learning pipeline for *de novo* peptide sequencing, including data preprocessing, deep learning model, and positional filtering model.

1.2 Thesis Overview

The remainder of this thesis is organized as follows:

- Chapter 2 reviews the works related to mass spectrometry and peptide sequencing, including database searching and *de novo* sequencing.
- Chapter 3 describes the background knowledge necessary for building a neural network for interpreting MS/MS spectra.
- Chapter 4 describes the methods and details of every experiments in this thesis.
- Chapter 5 presents the results and discuss the impact of SMSNet.
- Chapter 6 summarizes the key concepts of the thesis and provides directions for future works.

Chapter II

LITERATURE REVIEW

Peptide sequencing from MS/MS spectrum has been an active research topics for many years as it is a vital part for studying the characterization of proteins. Many tools were proposed and consistent improvement were made over the past few years. Most of the tools, however, focus on identifying peptide sequences by database searching; thus, they are highly dependent on the availability of database. On the other hand, *de novo* peptide sequencing deviates the problem of acquiring database by deducing peptide sequence using only the pattern in the spectrum. As a result, it is more likely to be affected by noise and ambiguity. With the recent advancement in deep learning, there were also attempts to do *de novo* peptide sequencing using neural networks. In this chapter, we explore the process by which MS/MS spectra are obtained, traditional peptide sequencing methods, as well as the recent advancement in deep learning and its application associated with *de novo* peptide sequencing.

2.1 Mass Spectrometry

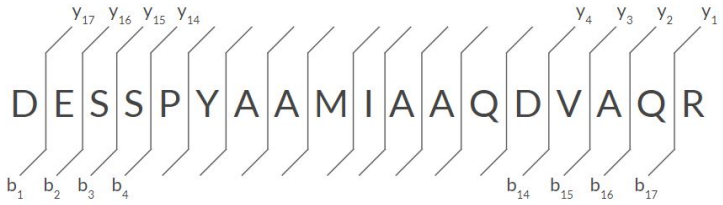
Mass spectrometry is a technique for measuring the masses within the sample. During the process, the sample is ionized and the ions are sorted according to their mass-to-charge ratio. A mass spectrum is obtained by plotting the ion signal intensity as a function of such mass-to-charge ratio. In general, a mass spectrum could be used to determine the masses of molecules within the sample, to identify isotopic signature of the sample, or to annotate the chemical structures of chemical compounds.

Mass spectrometry analysis of proteolytic peptides also plays an important role in protein characterization. By digesting the whole protein into smaller peptide fragments, the sample can be easily prepared, and each peptide fragment can be identified separately. To this end, a method for measuring the fragmented spectra, known as tandem mass spectrometry (MS/MS), is used in order to identify the fragment masses.

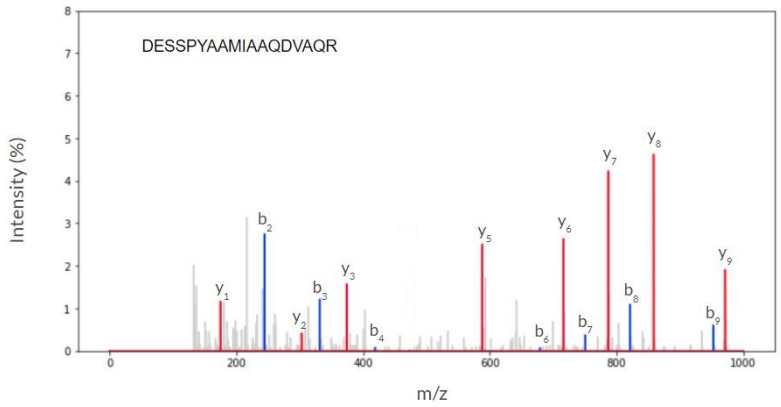
In order to induce fragmentation, many techniques were proposed, each with their own bias and limitation. Higher-energy collisional dissociation (HCD) is one of the possible techniques which induce fragmentation of molecular ions to generate tandem mass spectrometry (Medzihradszky and Chalkley, 2015). In this process, the peptide is bombarded with electrons, causing it to break into charged fragments.

Typically, each peptide breaks only once and produces two ion fragments, b-ion and y-ion, which are measured into tandem mass spectrum. Therefore, by analyzing MS/MS spectrum, which is a collection of b-ions and y-ions from the sample (Figure 2.1), the peptide sequence associated with the given sample could be identified.

In addition, as it could be seen in the example in Figure 2.1(b), MS/MS spectrum usually contains ambiguous information. First, it typically has a lot of noise with high intensity. Second, some b-ion and y-ion evidences might be missing from the spectrum (for example, b_1, b_5, y_4 are not present in the example). Both of these problems make peptide sequencing from MS/MS spectrum challenging.



(a) Possible b- and y- ions fragments



(b) Tandem mass spectrum

Figure 2.1: An example of tandem mass spectrum. The observable peaks of b-ions and y-ions are plotted in blue and red, respectively. The distance between two ion peaks could be used to identify the corresponding amino acid. For example, the difference between y_3 and y_4 equals to the mass of Valine.

2.2 Peptide Sequencing

Given MS/MS spectra, there are generally two approaches that could be used to determine the amino acid sequence in the sample: database search and *de novo* peptide sequencing.

Database Search

Database search is a straight forward approach to identify peptide sequence from MS/MS spectra. The unknown spectra are run through the database to find the best match among the known peptide sequences. This method allows the user to select appropriate threshold by setting the False Discovery Rate. In this approach, the similarity function to compute the best match varies depending on which characteristic of peptide each tool focuses on. For instance, MaxQuant (Cox and Mann, 2008) uses correlation analysis and graph theory to detect peaks and identify amino acid labels, while PEAKS DB (Zhang et al., 2012) integrates *de novo* peptide sequencing into database search to improve accuracy and sensitivity. Nevertheless, every database search techniques rely heavily on the availability of database of interested species and could not recognize novel peptides not presented in the database.

De Novo Peptide Sequencing

De novo peptide sequencing is another sequencing method in which the peptide sequence is determined from MS/MS spectra without any assisting database. This approach focuses more on assigning amino acid labels by considering peaks in the spectrum which are associated with theoretical fragment ions. Because of its ability to discover peptide sequence without database, *de novo* peptide sequencing could be used to study peptide from less-studied species and peptide with mutation.

Because of the nature of MS/MS spectrum which contains high noise and ambiguity, *de novo* peptide sequencing is usually formulated as a complex global optimization problem. Many forms of dynamic programming were developed over the past few years. For example, PepNovo (Frank and Pevzner, 2005) viewed the problem as a probabilistic model and calculated a score for each peak, then found the best path that follows pre-specified rules. PEAKS (Ma et al., 2003), instead of transforming the problem to a graph-based optimization, optimized for a sequence that covers as many high abundance peaks as possible. Finally, Novor (Ma, 2015) used decision trees to compute matching score between MS/MS spectra and peptide fragments, then found the sequence with highest score.

Recently, Tran et al. (2017) proposed a neural networks approach combined with local dynamic programming to solve the problem, and showed improvement over all previous methods. The proposed model, DeepNovo, consumes tandem mass spectrum as input to produce amino acid labels in an auto-regressive manner, one amino acid at a time, while using knapsack algorithm to discard hypothesis which does not match the sample mass. Deepnovo was claimed to perform at 97.2–99.5% accuracy when used together with an assembler to reconstruct antibody light and heavy chains of mouse, while, for peptide sequencing, achieving 38.2-66.1% amino acid recall and 14.6-39.4% peptide recall on various species.

In the neural network aspect of DeepNovo, the model employs Convolution Neural Network together with an encoder-decoder LSTM (Section 3.2), encoding the input MS/MS spectrum into a representation vector. Then, conditioning on given representation, the decoder consumes the previously predicted label and relevant positions based on prefix mass to produce next step prediction. During the decoding process, the model uses beam search algorithm, a techniques that continuously explores and keeps the k most possible hypotheses while searching, to select the sequence with maximum score as output. The knapsack algorithm is also used to help the model filters hypothesis that could not match the total sample mass.

The results of DeepNovo, however, still indicates the gap that could be improved in *de novo* peptide sequencing, particularly, in the aspect of precision of the prediction which could affect the overall usability of the output peptides in different tasks. DeepNovo also has limitation regarding the ability to express uncertainty in the predictions with ambiguous amino acids.

2.3 Review Summary

The review of literature presented in this chapter mostly focuses on the development of peptide sequencing tools and the application of neural networks on sequence transduction tasks. Many works framed *de novo* peptide sequencing as global optimization task, and some attempted to apply machine learning techniques to the problem. It was showed that it is possible to approach *de novo* peptide sequencing task from the neural networks standpoint. We believe there is still a gap in performance of *de novo* peptide sequencing that could be improved.

Chapter III

BACKGROUND KNOWLEDGE

In this chapter, we outlined the knowledge necessary for effectively constructing a deep learning model for *de novo* peptide sequencing.

3.1 Peptide sequencing by database searching

The database search approach for peptide sequencing from MS/MS spectra was used by many peptide sequencing tools such as MaxQuant (Cox and Mann, 2008) and PEAKS DB (Zhang et al., 2012). This approach ensures that the predicted peptides are consistent with the proteins in the known database but lack the ability to handle novel peptides. While each database searching tool differs on how the scores are assigned and how the spectrum is preprocessed, they shared the following key concepts. First, the tool is given a database related to the experiments to search from. Then, every candidate sequence in the database are compared with the input spectrum and are given a score. These scores indicate how likely the sequence could match the input spectrum. The sequence with the best score that passes a certain threshold is then assigned as a prediction for that spectrum. Figure 3.1 illustrates the difference between database searching approach and *de novo* sequencing approach.

3.2 Neural Networks for Transduction Task

Neural Networks are robust machine learning models which have achieved state of the art performance in various fields (LeCun et al., 2015). In sequence modeling and transduction, a form of recurrent neural networks, usually long short-term memory (Hochreiter and Schmidhuber, 1997) or gated recurrent unit (Chung et al., 2014), constantly produced highly competitive results in a variety of tasks, such as machine translation (Sutskever et al., 2014; Wu et al., 2016; Bahdanau et al., 2014), image captioning (Xu et al., 2015), and language modeling (Devlin et al., 2014). In this section, we review the use of neural networks in sequence transduction tasks along with the relevant components for building an effective model.

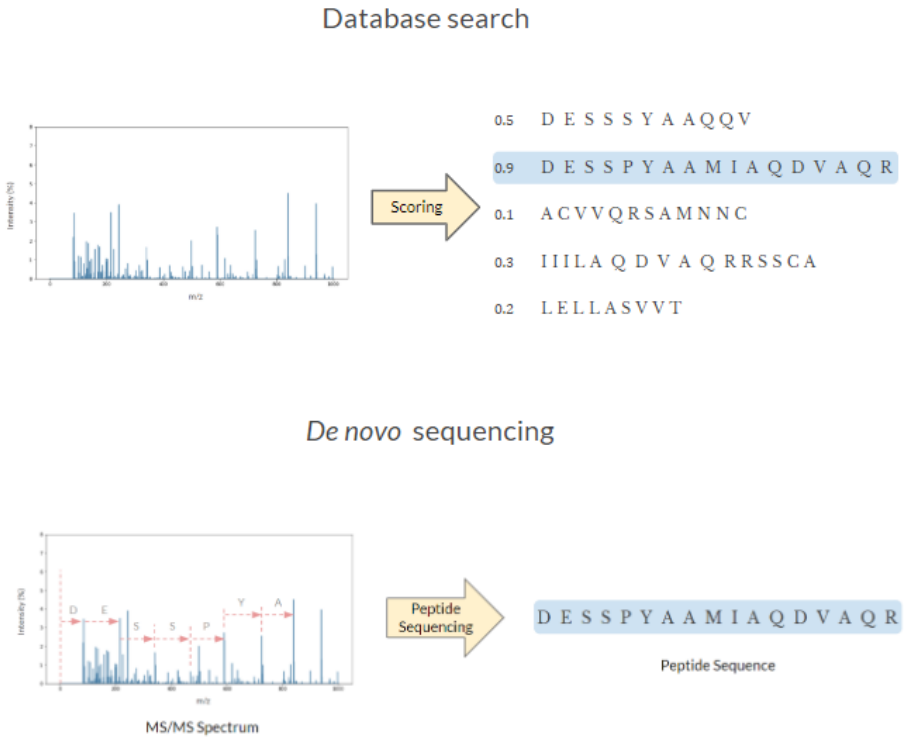


Figure 3.1: A difference between database searching approach and *de novo* sequencing approach. For database search, all peptides in the database are compared to the input spectrum to compute the similarity scores. The peptide with the highest similarity score is then chosen as the prediction for the input spectrum. For *de novo* sequencing, the prediction is directly derived from the input spectrum using the knowledge of the ion fragmentation process.

Recurrent Neural Networks

Recurrent Neural Networks (RNN) (Goodfellow et al., 2016) are a family of Neural Networks designed to handle sequential computation. Typically, recurrent neural networks is a type of neural networks that factors the step position t when processing inputs and outputs. By consuming the input at position t, the model calculates a hidden state h_t as a function of previous state h_{t-1}. Concretely, the model can be described as follows:

$$h_t = f(h^{t-1}, x^t; \theta)$$

where θ are the parameters of the function f, and the hidden state at current time step h^t is a function f of the previous hidden state h^{t-1} and the current input x^t. The output at each step t is then generated conditioning on h^t. This allows RNN to learn the concept of time for the input and summarize the past input sequence into a hidden state h^t.

Long Short-Term Memory

Long short-term memory (Hochreiter and Schmidhuber, 1997) is a variation of recurrent neural networks with an additional gate mechanism to adjust the weights of paremeters within the memory units. In order to avoid the problem with long-term dependencies (Bengio et al., 1994), a memory cell c is added to the model. The formula of long short-term memory networks (LSTM) can be expressed as follows:

$$f_i = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$g_t = tanh(W_g h_{t-1} + U_g x_t + b_g)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot tanh(c_t)$$

where f_t, i_t, o_t represent forget gate, input gate, and output gate respectively, c is the memory cell, h_t is the hidden state at time step t, and $\odot$ is an element-wise dot product operation.

Intuitively, the LSTM decides how much weight each gated elements get based on the previous hidden state h_{t-1} and the current input x_t. Forget gate f decides how much the model should discount the previous memory. Output gate o_t scales the output of the current step. Both input gates i_t and g_t determine how much the current input x_t should be added to the memory cell c.

Given a dataset of input and output pairs (x, y), a neural network is trained to minimize the difference between the predicted value $\hat{y}$ and the true label y according to a loss function. For classification task, the most common loss function is cross entropy loss:

$$loss(y, \hat{y}) = \sum_i y_i log(\hat{y}_i)$$

where y is a one-hot vector with size equals to a number of possible classes and $\hat{y}$ is a vector of predicted probabilities of each class.

In the usual neural networks the gradient of the loss function with respect to the network parameters is calculated by backpropagation algorithm (Rumelhart et al., 1986). For recurrent neural networks, the modified algorithm known as Backpropagation Through Time (Werbos, 1990) is used to propagate the

loss through the unfolded networks. Then, the network parameters are updated according to the optimizer. The most common optimizers are stochastic gradient decent (SGD) (Bottou, 2010) and Adam (Kingma and Ba, 2014).

3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) (LeCun et al., 1998) is another variant of neural networks which aims to capture the locality of the features in the input. The network was first designed to exploit the two properties most images data obeys: highly correlated neighbors and translation invariant. In the networks, many small filters, usually 3x3 or 5x5 pixels, are used to perform 2-dimensional convolution on the input image across vertical and horizontal dimensions. For every positions, the same filters are applied to ensure the translation invariant property. This operation results in a feature vector for each pixel in the image. Because the inputs and outputs of each CNN layer could be very large, max pooling or average pooling is commonly used to downsample the feature vectors. Besides, it is important to note that CNN can be easily extended to handle 1-dimensional data by discarding one dimension from the filters to match with the 1-dimentional input.

3.4 Encoder-Decoder Architecture

Encoder-decoder structure has become a part of many established models for sequence transduction. The idea showed great potential following the successes in many fields such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014) and image captioning (Vinyals et al., 2015). In this architecture, the encoder transforms the input feature $x_1, ..., x_n$ into a vector representation z. Given z, the decoder then computes an output sequence $y_1, ..., y_m$ in an auto-regressive manner. In other words, the model conditions on the previously generated symbol y_{t-1} to produce the next output symbol y_t.

3.5 Embeddings

In sequence transduction models, it is common to have learned embeddings as part of the networks. An embedding layer converts the discrete input tokens into a higher-dimensional, real-valued vectors. The aim is for these vector representations to provide more information about the inputs to the model than their original values, and make the model easier to learn. There are several methods for learning embeddings

such as Continuous Bag-of-Words and Skip-Gram (Mikolov et al., 2013). It is also possible to train an embedding layer together with the model via backpropagation.

3.6 Summary

In this chapter, we described the background knowledge related to the database search approach for peptide sequencing and also each component for building an effective neural network model.

Chapter IV

METHODS

This chapter describes the datasets, experiment settings, and our proposed Sequence-Mask-Search pipeline (SMSNet) for *de novo* peptide sequencing. We will also discuss the evaluation metrics for comparison with the previous model, DeepNovo (Tran et al., 2017), and for validating the usefulness of our prediction in the context of real experiments where some spectra do not contain any peptide information. This chapter was taken from our manuscript which has been submitted for publication.

4.1 Data acquisition

A combined dataset consisting of more than 27 million peptide-spectrum matches (PSM) was obtained from the Proteomics and Metabolomics Core Facility at the Wistar Institute (Philadelphia, PA, USA). All MS/MS spectra were acquired on Q Exactive HF or Q Exactive Plus mass spectrometers (Thermo Fisher Scientific, Bremen, Germany) and processed using MaxQuant (Cox and Mann, 2008) by scientists at the Core Facility. Peptide level false discovery rate was set at 5%. Multiple sets of variable modifications and multiple protein databases were used depending on the goals and scopes of individual mass spectrometry experiments. Importantly, the metadata have been removed to safeguard the identity of principal investigators and the details of their research projects.

From 27 million PSMs, we constructed three individual training datasets: (i) WCU-MS-M, which consists of 25,174,942 MS/MS spectra that correspond to unmodified peptides and peptides containing oxidized Methionine, (ii) WCU-MS-P, which consists of 26,943,975 MS/MS spectra that correspond to unmodified peptides, peptides containing oxidized Methionine, and peptides containing phosphorylated Serine, Threonine, or Tyrosine, and (iii) WCU-MS-BEST, which consists of 1,239,045 MS/MS spectra that were assigned the highest quality scores by MaxQuant (the ”Score” column in evidence output file) for each unique unmodified peptide and charge state. In other words, the WCU-MS-BEST dataset contains the highest quality MS/MS spectrum for each unmodified peptide.

We also acquired two external datasets to evaluate SMSNet’s performance on Q Exactive MS/MS data from diverse species and laboratories. For direct comparison with DeepNovo (Tran et al., 2017), we combined 1,422,793 PSMs from 9 studies of distinct species (PRIDE accessions PXD005025,

PXD004948, PXD004325, PXD004565, PXD004536, PXD004947, PXD003868, PXD004467, and PXD004424) that were previously curated by DeepNovo's developers. Finally, high-quality MS/MS spectra of synthetic peptides were acquired from the ProteomeTools HCD Spectral Library (Zolg et al., 2017). It should be noted that this dataset was acquired on Orbitrap Fusion Lumos mass spectrometer. We named both datasets DEEPNOVO-M and ProteomeTool, respectively.

4.2 Training, validation, and test sets partitioning

To ensure that training, validation, and testing sets do not share a common peptide, we first partitioned unique peptides into three sets, then constructed training, validation, and testing sets from mass spectrum data associated with these peptides. Accounting for the fact that some peptides appear in the datasets much more often than the others, we kept only one random data entry per peptide in validation and testing sets. The validation set was used for choosing the model architecture and determining the number of training steps. For WCU-MS datasets, we used validation and test sets of size 50,000 but used smaller subset in other datasets to maintain the number of unique peptides in each dataset. During training, peptides with more than 30 amino acids were ignored to account for the fact that longer peptides are likely to have more noise in the spectrum. The details of each dataset after partitioning are presented in table 4.1.

Dataset	Unique Peptide		
	Train	Validation	Test
DEEPNOVO-M	216,200	20,000	20,000
WCU-MS-BEST	769,208	50,000	49,998
WCU-MS-M	864,990	50,000	50,000
WCU-MS-P	994,786	50,000	50,000
ProteomeTools	162,648	20,000	20,000
Dataset	Total Spectra		
	Train	Validation	Test
DEEPNOVO-M	1,198,433	111,365	112,995
WCU-MS-BEST	1,095,941	50,000	49,998
WCU-MS-M	22,607,416	50,000	50,000
WCU-MS-P	24,499,353	50,000	50,000
ProteomeTools	212,454	26,154	26,228

Table 4.1: Datasets description. DEEPNOVO-M, WCU-MS-BEST, and ProteomeTools were used for comparison with DeepNovo.

4.3 Data preprocessing

MS/MS spectra in the WCU-MS training sets were extracted from raw files and centroided using Thermo Fisher Scientific's MSFileReader version 3.0. MS/MS spectra in the HLA peptidome and phosphoproteome datasets were extracted from raw files into mgf format using ProteoWizard version 3.0.11133 (Chambers et al., 2012) with the following filter parameters: Peak Picking = Vendor for MS1 and MS2, Zero Samples = Remove for MS2, MS Level = 2-2, and the default Title Maker. Charge state deconvolution was not performed.

The data of a tandem mass spectrum in mgf format is stored as a list of mass and intensity tuples, where each tuple contains the intensity of each mass value. In all experiments, any mass intensity above 5,000 Da was discarded as we found it is often noisy and uninformative for the model. For each spectrum, we used two resolutions of 0.1 Da and 0.01 Da for discretizing it to vector representations of length 50,000 and 500,000. The lower resolution vector provides an overview of the spectrum for the encoder while the higher resolution vector is used by the candidate ion stack. The details of each component are described in the next section.

4.4 Neural networks model architecture

Inspired by DeepNovo (Tran et al., 2017), we developed our deep learning model focusing on integrating domain knowledge to create a specialized model for *de novo* peptide sequencing, which we called SMSNet. By viewing a peptide sequence as a list of amino acids, we can view the peptide sequencing problem as a problem of predicting a series of amino acid for each position. Let X be an input mass spectrum data, the model can be written as:

$$P(Peptide|X) = \prod_{i=1}^{N} P(y_i|y_0, y_1, y_2, ..., y_{i-1}; X)$$

where y_i is the predicted amino acid at position i, y_0 is a special start token, and N is the peptide length.

Our model consists of three main components: an encoder, a decoder, and an ion stack. In general, the encoder tries to capture an overview of input mass spectrum and use it to initialize the decoder. Then, conditioning on the predicted prefix, the ion stack focuses on the relevant part of the spectrum and uses it to compute features for predicting the next amino acid. Finally, the decoder calculates probabilities for the next amino acid using its previous prediction and features from the ion stack. The model architecture

is illustrated in Figure 4.1 and Figure 4.2. Every layers in the networks used rectified linear unit (ReLU) as the activation function unless specified otherwise.

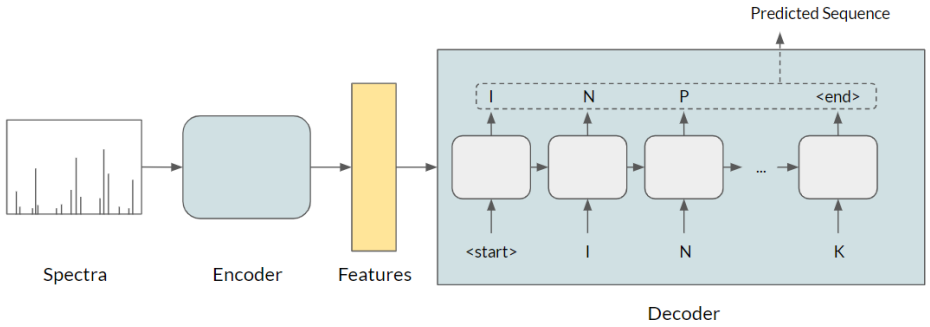


Figure 4.1: A simplified architecture of SMSNet. The encoder captures peak pattern in the input spectrum and encodes it into a feature vector. The decoder then iteratively predicts amino acid based on the feature vector and the previous prediction.

Encoder

The encoder was designed to encode an overview of the input spectrum vector into a feature vector of size 1024 which will be used to initialize the hidden state and cell state of the decoder. To integrate the knowledge from the peptide fragmentation process into the model, we restructured the input to make it more likely for the encoder to capture the relationship between positions that could be used to determine amino acid presences. Firstly, the input vector of length 50,000 was duplicated A times, where A is the number of possible amino acids, into a tensor of shape (50000,A). (A is 21 when training on datasets with 20 amino acids plus oxidized Methionines and 24 when training on datasets with 20 amino acids plus oxidized Methionines and phosphorylated Serines, Threonines, and Tyrosines). Each copy of the original input vector is shifted to the left according to each amino acid mass, then padded with zeros. For example, with the resolution of 0.1 Da, the vector representing Alanine is shifted to the left by $floor(71.037 \times 10) = 710$. The first 710 values in the vector are discarded, and 710 zeros are padded to the right. This process resulted in a tensor of shape $(50000, A)$. Secondly, we created another vector of values from 0 to 49,999 to indicate the index of each positions on the spectrum, then normalized it to have zero mean and unit variance. The index vector was then concatenated to the input to provide the information regarding the position, resulting in a tensor of shape $(50000, A + 1)$.

The restructured input was then passed to the encoder neural networks consisting of three 1x1 convolution layers, followed by three fully connected layers. Each of the 1x1 convolution layers applied the

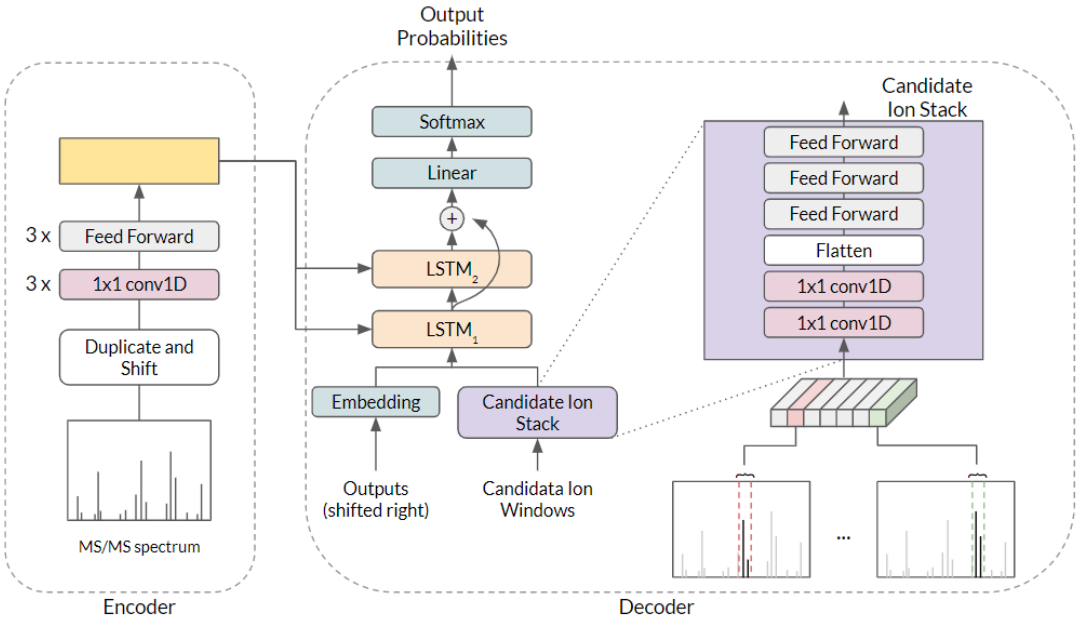


Figure 4.2: The core neural networks of SMSNet. The SMSNet model comprises of three main parts: the encoder, the decoder, and the candidate ion stack, each part focusing on finding patterns in one aspect of MS/MS spectra. In the encoder, input spectrum is duplicated and shifted left according to amino acid masses to highlight relationship between peaks with mass difference equal to mass of an amino acids. The decoder uses feature representation computed by the encoder to initialize the Long short-term memory (LSTM) layers, then iteratively predicts a series of amino acids by conditioning on the previous output and current-step features calculated by the candidate ion stack.

same transformation to every input position separately and compute features along the second dimension of the input tensor. This forces the encoder to learn about the structure at each location. The three kernels had shape $(1, 32)$, $(1, 64)$, and $(1, 2)$ that would produce a tensor of shape $(50000, 32)$, $(50000, 64)$, and $(50000, 2)$ respectively after each layer. After that, the feature vector was flatten and passed through three fully connected layers with dimension 512, 512, and 1024, finally resulting in a vector of size 1,024. For regularization, a dropout layer with dropout rate of 0.4 was used between the first and second fully connected layer.

Decoder

The decoder is a type of recurrent neural network that receives the feature vector from the encoder and uses it to generated a sequence of amino acids by outputting amino acids one by one. This is similar to the technique used in training neural networks for image captioning (Vinyals et al., 2015) or machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014) where the input information

(an image or a sentence in one language) is encoded into a vector representation, then passed to a decoder to generate the intended output (a caption or a sentence in a different language). Normally, the decoder for image captioning takes only the previously outputted word as input. In SMSNet, the decoder also takes as input a feature vector calculated by the candidate ion stack based on previous predictions for each step. This additional features were designed to provide more context about the next amino acid to the model.

In the decoder, we used two layers of long short-term memory (LSTM) of size 512 with layer normalization (Ba et al., 2016) on top of each layer and a residual connection (He et al., 2016) around the second layer. The same encoded vector of length 1,024 was partitioned into two halves and used as initial values for the hidden state and memory in both layers. At each step, the LSTMs takes as input a vector of length 544, a concatenated vector between a feature vector of length 512 from the candidate ion stack and an embedding vector of size 32 of the previous amino acid. Then, the output from LSTMs is passed through a fully-connected layer with a softmax activation function to produce probabilities for each amino acid. The shape of the last output depends on the number of possible amino acids (20, 21, or 24 depending on the number of modified amino acids considered)

Candidate ion stack

Given the total mass of the previously predicted amino acids, the candidate ion stack retrieved relevant sections of the mass spectrum to compute a feature vector for the decoder. Specifically, it looks for evidence supporting the next prediction by focusing on regions that can be the next amino acid. For each possible amino acid, 8 ion types were considered: b, b(2+), b-H2O, b-NH3, y, y(2+), y-H2O, and y-NH3. Supposed that there are 21 different amino acids, for each of the 8 ions, we sliced a small window of size 0.2 Da (20 elements at 0.01 resolution) from the original input vector of size 500,000, resulting in 168 20-element vectors. These vectors were stacked together to form an input of shape $(168, 20)$.

The candidate ion stack consisted of two 1x1 convolution layers followed by two fully-connected layers. The idea is to force the model to first learn the peak patterns of each ion, then learn the relationship between ions based on the calculated features. The two 1x1 convolution layers had 32 and 64 filters respectively, while both fully-connected had 512 dimensions. The output feature tensor was then used as input for the decoder. Figure 4.3 illustrates the process of the candidate ion stack at each time step.

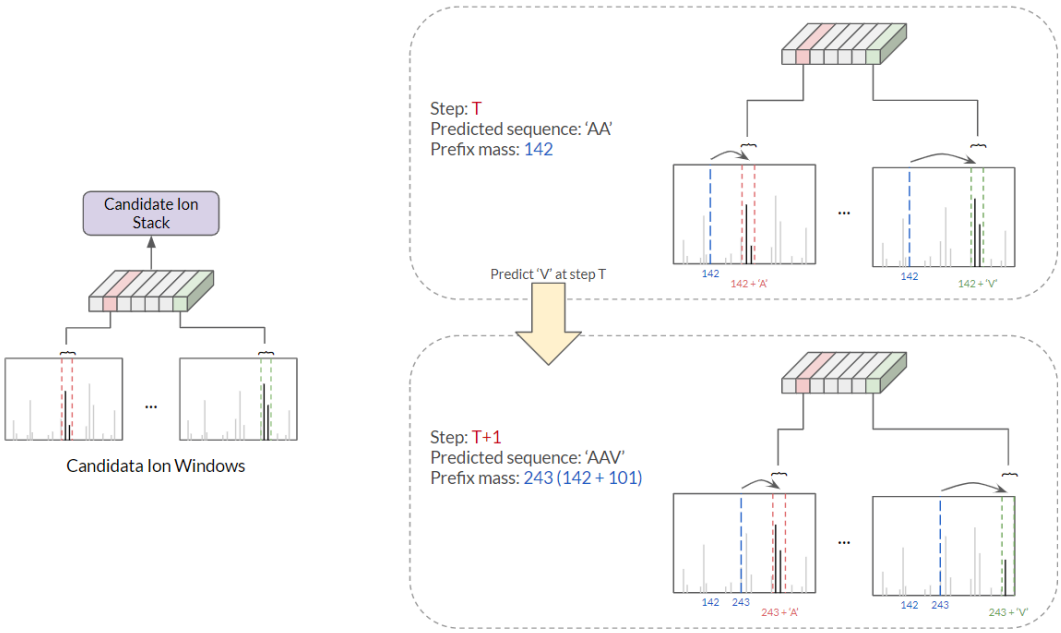


Figure 4.3: The candidate ion stack at each time step. The candidate ion stack used the total mass of the predicted sequence as a starting point, then cut small windows that might contain the information of the next amino acid from the input spectrum to be as input.

4.5 Inference

During inference, we used beam search with beam size of 20 to explore and find the most likely sequence of amino acids. At each step, every remaining hypotheses are ranked by the following formulas, which is a modified version from Wu et al. (2016):

$$score(Y, X) = log(P(Y|X))/length\_penalty(Y)$$

$$length\_penalty(Y) = \frac{5 + |Y|}{6}$$

where $P(Y|X)$ is a product of the previously predicted amino acid probabilities. The length penalty term is used to compensate longer sequences which usually have lower product value than the shorter ones. Additionally, during each step, we filtered out hypotheses that the difference between its current mass and the precursor mass did not match any possible amino acid combinations using the knapsack search algorithm.

The beam search decoding would continue until a special ending token is produced or a maximum length of 50 is reached for every remaining beam. After the decoding process ended, the amino acid sequence with the best score according to the provided formulas was selected as the final output.

During beam search, we prune the hypotheses using its remaining mass, $total\_mass - prefix\_mass(Y)$, to keep only hypotheses that are possible to match the total mass of the spectra. The hypothesis is discarded if there is no amino acid combination that has a mass equal to the remaining mass. Knowing every possible amino acids, we can construct a look-up table using dynamic programming described as follows:

$$DP[i] = \begin{cases} 1, & \text{if any } DP[i - aa_j] = 1. \\ 0, & \text{otherwise.} \end{cases}$$

where $DP[0] = 1$ and aa_j is a mass of possible amino acids. The table is built with 0.0005 Da resolution and allows 0.01 Da tolerance when searching. Concretely, a hypothesis is deemed possible if $\exists_i DP[i] = 1$ where $i \in [suffix\_mass(Y) - 0.01, suffix\_mass(Y) + 0.01]$.

4.6 Model training

We modeled the peptide sequencing task as a series of amino acid predictions where each prediction is a multi-class classification problem. We chose the focal loss (Lin et al., 2017), which is a dynamically scaled cross-entropy loss, as a loss function for our model. For binary classification tasks, the focal loss is defined as:

$$Focal\ Loss = -\alpha(1 - p_t)^\gamma$$

where $p_t = p$ for the class with label $y = 1$ and $p_t = 1 - p$ otherwise, p is the model's estimated probability for the class with label $y = 1$, α and γ are hyperparameters for balancing the importance of positive/negative examples and easy/hard examples, respectively. We set α to 0.25 and γ to 1.0 as it performed best on the validation set. The focal loss is chosen instead of normal cross-entropy loss because we suspected that there is an imbalance between easy examples with complete mass spectrum data evidence and hard examples with missing peaks.

To extend the focal loss to multi-class classification, we can view a multi-class classification problem as many binary classification problems. Concretely, we can pass the output of the last layer of the model through multiple sigmoid functions to obtain binary probabilities of being each class, then use the provided formula to calculate the focal loss. For inference, the sigmoid function was substituted with a softmax function to compute probability scores which can be summed to 1.

We initialized all parameters by drawing from a uniform distribution between -0.1 and 0.1, and

trained the model using stochastic gradient descent with learning rate decay. An initial learning rate of 0.01 was used until two-thirds of the maximum training step. Afterwards, as a form of learning rate decay for the model to better fit the data when loss is small, the learning rate was halved every one-twelfth of the maximum training steps. The gradient of the loss was normalized so that its L2-norm was less than or equal to 5. With a batch size of 32, the models were trained for 4,000,000 steps on WCU-MS-M and WCU-MS-P, which took roughly one month on Nvidia GeForce GTX 1080 Ti.

4.7 Ablation study

To evaluate the impact of each component to the performance of SMSNet, we performed ablation studies by making some modifications to the model, then measuring the performance degradation caused by those modifications. The following modifications were tested:

- Removing the encoder entirely and initializing the decoder with a vector of zeros.
- Removing the shift mechanism in the encoder. In this variation, we removed the 1x1 convolution layers and fed the low-resolution input vector of size 50,000 directly to the fully-connected layer.
- Using normal cross-entropy loss instead of the focal loss.
- Not using layer normalization after LSTM layers in the decoder.
- Not considering b-H2O, b-NH3, y-H2O, and y-NH3 ions in the candidate ion stack.

Every modified models were trained for 20 epochs on WCU-MS-BEST.

4.8 Rescorer

Once the entire amino acid sequence has been predicted, SMSNet adjusts the confidence score for each position in the prediction through another model called the rescorer. We designed the rescorer to be a shallow neural network consisting of two fully connected layers of size 64. To train the model, we used binary cross-entropy loss and the Adam optimizer with default parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The rescoring validation set was used for early stopping.

4.9 Data preprocessing for the rescorer

Unlike the main model, the rescorer operates solely on the level of amino acids. For each hypothesized amino acid, it predicts the confidence level of the prediction. The following features were used:

peptide length, numbers of amino acids with probability more than 0.7, 0.8, and 0.9, a geometric mean of amino acid probabilities in the peptide, the position of amino acid normalized by the peptide length, probabilities of amino acids at index $t-1$ to $t+2$ for current index t. We chose these features on the basis that they are not visible to the main model during the decoding process and they gave the lowest loss on the validation set. The label for each data point is 1 if the given *de novo* amino acid matches the true label and 0 otherwise.

As the rescoring model is designed to evaluate amino acids labels predicted by the main model, we could not used the original training set that the main model was trained on. Therefore, the amino acids in the original validation set was partitioned into rescoring training and validation sets with ratio of 90:10. The test set were still the same for both tasks.

4.10 Comparison with DeepNovo

To compare the performance of our model with DeepNovo (Tran et al., 2017), we trained both DeepNovo and SMSNet on two datasets, one from nine species used in DeepNovo (Tran et al., 2017) and one from our new dataset, which were used for comparison only. Both models used the same training, validation, and test sets. For DeepNovo, we used the code provided together with their publication.

The first dataset is constructed by combining together all high-resolution datasets in DeepNovo publication. As we only focused on amino acid with Methionine-oxidation, any peptide that contains amino acid with Asparagine- or Glutamine-deamidation in the original dataset was discarded. The remaining data consisted of 1,422,793 mass spectra from 256,200 unique peptides. Due to its lower number of unique peptides, instead of using 50,000 unique peptides as validation and testing sets as in other experiments, we sampled only 20,000 unique peptides from the dataset and used all of their associated spectra, resulting in validation and test sets of size 111,365 and 112,995, respectively.

The second dataset, called WCU-MS-BEST, is a subset of WCU-MS-M dataset that contained only peptide with no amino acid modification. We selected only spectrum with best quality score according to MaxQuant (Cox and Mann, 2008) for each unique peptide and charge state to form an easy but diverse dataset. In total, there are 1,239,045 spectrum of 869,206 unique peptide. The validation and test set each contains 50,000 unique peptide spectra (two spectra were later removed from the test set due to mismatches between their precursor masses and the labels, resulting in the test set of size 49,998). For

peptides with many charge states, we randomly chose one charge state and discarded the rest.

The third dataset is the high-quality MS/MS spectra of synthetic peptides from the ProteomeTools HCD Spectral Library (Zolg et al., 2017). To preserve the number of unique peptides in each of train, validation, and test set, only 20,000 unique peptides out of 242,648 were randomly chosen for validation and test set. In total, there were 212,454, 26,154, 26,228 spectra in train, validation, and test set, respectively. As the dataset was acquired on Orbitrap Fusion Lumos mass spectrometer, the noise patterns in the spectrum were different from those in WCU-MS datasets. As a result, using the models trained on WCU-MS-BEST for predicting peptides from ProteomeTools spectra yielded near-zero recall. Thus, by using this dataset, we could evaluate the generalizability of SMSNet and DeepNovo model by retraining the models on the MS/MS spectra from different mass spectrometer.

The amino acid vocabulary were set according to the dataset for both models, with 20 possible amino acids for the first dataset and 21 for the second dataset. Apart from the amino acid vocabulary, our model settings were the same as in other experiments. For DeepNovo, we set the spectrum resolution to 0.02 Da and kept other default parameters. At inference time, both models used beam search with beam size 20 to find the most probable peptide for each input.

4.11 Evaluation metrics

For evaluation, we considered the performance on both amino acid level and peptide level. A predicted amino acid is considered matched to the ground truth only if their masses differs less than 0.0001 Da and their prefix masses differs less than 0.03 Da, and a peptide sequence is considered matched only if all of its amino acids match the true labels. As the models provided confidence scores that reflect the quality of each amino acid predictions, we could set a threshold below which the predictions are discarded. Then, by varying the threshold, we can plot precision-recall curves to summarize the performance of the models. In addition, if a peptide has less than four amino acids left, we will also discard all of the remaining amino acids regardless of their scores. The precision and recall were measured by the number of matched predictions divided by the number of total predictions and the number of ground truth labels, respectively.

The formulas for calculating precision and recall can be described as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

4.12 Mass tag and database search

After the scores of each amino acid prediction were adjusted by the rescorer in the ”sequence” phase, we set a threshold below which a prediction is considered ambiguous. Then, during the ”mask” phase, all of the continuous amino acids positions whose confidence scores lie below a user-specified threshold were grouped together and replaced by a mass tag that reflect their combined masses. Finally, during the ”search” phase, SMSNet attempts to recover the exact amino acid sequences from masked positions by searching all predictions against a reference amino acid sequence database. Combining the three steps together, the sequence-mask-search framework can be illustrated as in Figure 4.4

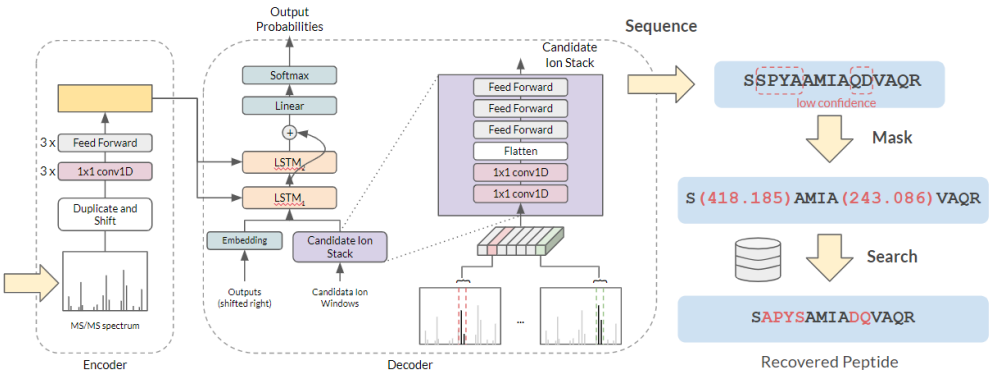


Figure 4.4: Overview of the Sequence-Mask-Search framework. SMSNet encodes the input MS/MS spectrum and passes the information to the decoder module which outputs amino acid sequentially. During the sequencing process, relevant m/z regions from the input MS/MS spectrum are extracted and fed to the decoder. Post-processing steps involve the adjustment of positional confidence scores, the replacement of low confidence positions by mass tags, and the recovery of exact amino acid sequences in masked segments through database search.

To identify the exact amino acid sequences for predictions that contain mass tags or ambiguous Leucine/Isoleucine positions, we search for possible matches within a given protein sequence database. For analyzing HLA peptidome and human phosphoproteome dataset, the Uniprot (Consortium, 2018) reference human proteome was used. For evaluating whether each of SMSNet's predictions could be matched to a unique possibility, the Uniprot reference proteome for either human, mouse (*Mus muscu-*

lus), budding yeast (*Saccharomyces cerevisiae*), *Escherichia coli* strain K12, or a combination of all four species was used. All databases include isoforms and predictions proteins. An amino acid sequence within the database is considered a match to an ambiguous prediction if (i) all non-Isoleucine positions in both sequences match, (ii) all Isoleucines in the prediction match to either Leucine or Isoleucine in the database sequence, and (iii) all mass tags in the prediction match to amino acid substrings in the database sequence whose weights differ less than 20 ppm from the corresponding mass tags. For evaluating the fraction of SMSNet's prediction that can be uniquely mapped to a peptide in a database, we considered only the spectra whose labels were also present in that database.

4.13 Definition of amino acid's evidence

Given a mass spectrum, we determined that an amino acid has supporting evidence if it follows our defined criteria. Firstly, for an amino acid with mass M_{aa} and prefix mass M_{prefix}, there must be ions with mass M_{prefix} and $M_{prefix} + M_{aa}$ present in the spectrum. Secondly, a fragmented ion is said to be present in the spectrum if there is at least a peak with any intensity within 0.1 Da of its theoretical b-, b(2+)-, y-, or y(2+)-ion. The first and last amino acid in a peptide only require one ion mass presence.

4.14 Application on real-world studies

In order to assess the performance of SMSNet as part of the real-world studies, we run the trained models on the spectra from two external datasets: human leukocyte antigen (HLA) peptidome and human phosphoproteome. In both studies, some spectra might not correspond to any peptide as a result of noise in the detection process or chemical contamination. We set the cutoff of SMSNet at 95% amino acid-level precision on the training dataset of the model. The predicted sequences which contain less than four remaining amino acids were discarded. For evaluating SMSNet's ability to discover new peptides, we downloaded 83 raw files consisting of more than 3.5 million MS/MS spectra from an HLA peptidome study of mono-allelic cell lines (Abelin et al., 2017) (MassIVE accession MSV000080527). Finally, for testing SMSNet-P model's ability to identify phosphorylated peptides, we downloaded 12 raw files consisting of more than 676,000 MS/MS spectra from a comprehensive phosphoproteome study of control and epidermal growth factor-treated glioblastoma cells (Humphrey et al., 2018) (PRIDE accession PXD009227). The details of both studies are beyond the scope of this thesis. We refer the reader to Abelin et al. (2017) and Humphrey et al. (2018) for more information.

4.15 Summary

In this chapter, we described the datasets used in this thesis, the SMSNet model, and the evaluation on various aspect. We started by providing a step by step detail of how the data were curated and preprocessed. Then, we explained the motivation and construction of each part of SMSNet, including the encoder, the decoder, and the candidate ion stack. Finally, we outlined the methods for comparing our results to the state-of-the-art *de novo* sequencing tool DeepNovo and for evaluating SMSNet's performance in the real-world setting. The results of each evaluation will be presented in the next chapter.

Chapter V

RESULTS

In this chapter, we present the results from the experiments described in the previous chapter. We tested SMSNet on both controlled datasets and real-world datasets where some spectra might not have a label. We will also discuss the results, especially in term of model architecture comparison between SMSNet and DeepNovo (Tran et al., 2017), and the impact of each component of SMSNet to the overall performance.

5.1 Dataset compositions and SMSNet model

To train our neural networks model, we gathered a large collections of mass spectrum data from experiments at Proteomics Core Facility at Wistar Institute to be used as training data. Our dataset contains a diverse range of species, including peptides that can be uniquely matched to the protein database of Homo sapiens(17%), Mus musculus(9%), Saccharomyces cerevisiae(6%), and Escherichia coli(3%). The complete compositions of species-specific peptides in WCU-MS dataset are shown in table 5.1. First, by combining 1,543,394 spectra of peptides with Methionine-oxidation and peptides with no modification, we constructed a dataset of size 25,174,942, which we called WCU-MS-M dataset. Then, an addition of 1,769,033 spectra with Serine-, Threonine-, or Tyrosine-phosphorylation were added to WCU-MS-M to form the second dataset called WCU-MS-P, consisting of 26,943,975 spectra. For comparison with the current state-of-the-art de novo sequencing tools DeepNovo (Tran et al., 2017), we prepared another two sets of spectra. The first set was a collection of high-resolution spectra from nine species used in DeepNovo (Tran et al., 2017) with a total size of 1,422,793. The second set was a high-quality subset of WCU-MS-M of size 1,239,045 (see Methods for details). We chose these two sets on the basis that DeepNovo settings were configured for the dataset of this size. We named these two sets DEEPNOVO-M and WCU-MS-BEST, respectively.

Using these mass spectrometry data, we trained four neural network models for de novo sequencing, one for each dataset. We named our neural network model SMSNet, reflecting the Sequence-Mask-Search mechanism used in our sequencing pipeline. To reduce the complexity of the output space, we reframed de novo sequencing task from outputting the whole peptide chain all at once into a task of

Species	Number of Unique Peptides
Homo sapiens	165,783
Mus muculus	91,482
Saccharomyces cerevisiae	55,466
Escherichia coli	31,585
Rattus norvegicus	17,587
Arabidopsis thaliana	13,865
Xenopus laevis	11,441
Bos taurus	6,515

Table 5.1: Compositions of species-specific peptides in WCU-MS dataset. A peptide is counted only if it could be mapped to exactly one of the databases considered. All databases were downloaded from Uniprot. Isoforms and predicted proteins are included.

predicting a series of amino acids to match the encoder-decoder framework (Sutskever et al., 2014; Venugopalan et al., 2015; Vinyals et al., 2015; Cho et al., 2014), where the inputs are first encoded into a fixed-length vector representation that is then used for generating outputs. More specifically, the model takes a MS/MS spectrum as the input then predicts one amino acid at each step by conditioning on the spectrum and previously predicted amino acid. SMSNet comprises of three main components: the encoder, the candidate ion stack, and the decoder (Fig 4.4). Each component focuses on one aspect of the mass spectrum data: (i) the encoder captures an overview of mass spectrum in low resolution, (ii) the candidate ion stack considers only a few locations necessary for predicting the next amino acid in high resolution, and (iii) the decoder learns the peptide sequence pattern based on the other two features. During inference, we used beam search to keep only a constant number of top candidate sequences at each decoding step. More details can be found in Methods.

5.2 Model evaluation on held-out mass spectrometry data

We evaluated performance of SMSNet on the held-out test sets of each datasets by using peptide labels from a database searching tool MaxQuant (Cox and Mann, 2008) at <5% false discovery rate (FDR) as ground truth and comparing them with the predicted sequences.

First, we compared SMSNet results to the current state-of-the-art de novo sequencing model DeepNovo (Tran et al., 2017) on two datasets, DEEPNOVO-M and WCU-MS-BEST, with both model trained and tested on the same datasets. The recalls of SMSNets trained on DEEPNOVO-M were 47.11% on peptide-level and 71.24% on amino acid-level, outperforming DeepNovo which obtained 44.41% on amino acid-level and 65.57% on peptide-level by 2.7% and 5.7%, respectively. The precision-recall curve

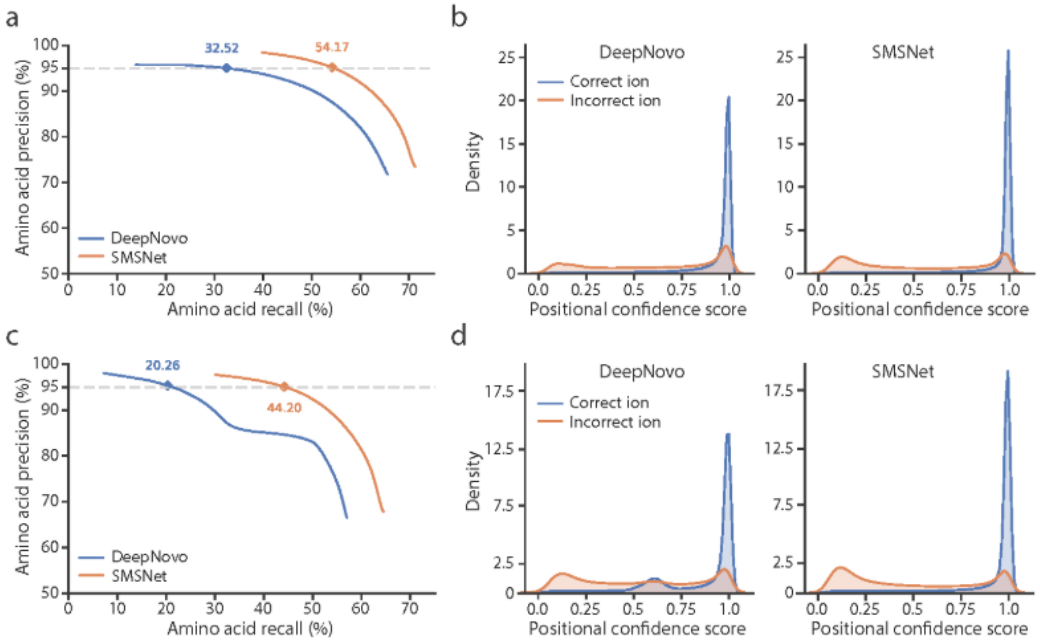


Figure 5.1: Comparison between SMSNet and DeepNovo on DEEPNOVO-M and WCU-MS-BEST. **(a)** Amino acid-level precision-recall curves for SMSNet and DeepNovo when evaluated on the dataset curated by DeepNovo's authors. The corresponding recalls at 5% amino acid false discovery rate are indicated. **(b)** Histograms showing the distributions of positional confidence scores produced by SMSNet and DeepNovo. **(c-d)** Similar plots showing performances of SMSNet and DeepNovo when evaluated on our WCU-MS-BEST dataset.

(Fig 5.1(a-b)), produced by varying the threshold for each amino acid predictions, showed that SMSNet provided better recalls at every precision levels. At 95% precision (5% FDR), SMSNet achieved as large as 21.7% improvement over DeepNovo. On WCU-MS-BEST (Fig 5.1(c-d)), the result follows the same behaviour where SMSNet attained 44.73%/64.45% peptide/amino acid recall, surpassing DeepNovo which acheived 37.57% and 57.02% recall by 7.16% and 7.43%, while showing improvements at every precision threshold. Overall, our model outperforms the current state-of-the-art tool DeepNovo on both datasets.

Second, we evaluated the performance of both SMSNet and DeepNovo on the high-quality MS/MS of synthetic peptides. Because the peptide labels of the MS/MS spectra in WCU-MS dataset and DEEPNOVO-M dataset were acquired by databases searching, the labels might have some bias towards peptides which are known to databases, or the data might contain some wrong labeling that would affect the model when used for training. Thus, we used an addition synthesized dataset from the ProteomeTools HCD Spectral Library (Zolg et al., 2017) (described in Methods) which could restrict the effect of database search to analyze the models. Furthermore, this evaluation also functioned as a generalizability

test to show that SMSNet architecture does not overfit to only MS/MS spectra from Q Exactive HF or Q Exactive Plus mass spectrometers. Both model were trained on a portion of ProteomeTools dataset. The results showed that, considering all predicted peptides, SMSNet achieved 53.51% peptide recall and 75.06% amino acid recall while DeepNovo achieved 50.69% peptide recall and 72.49% amino acid recall, respectively. The precision-recall curves of both models are shown in Figure 5.2.

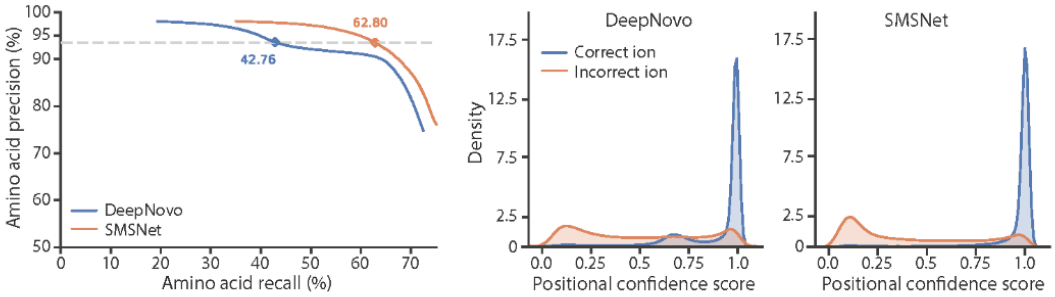


Figure 5.2: (left) Precision-recall curve illusting the performance of SMSNet and DeepNovo on high-quality MS/MS of synthetic peptides. (right) Density plots of amino acid confidence scores of the predictions from both models.

We then evaluated the other two SMSNet models trained on the entirety of our available data, WCU-MS-M and WCU-MS-P datasets. We referred to both models as SMSNet-M and SMSNet-P from here on. The models were tested against their own held-out test sets. SMSNet-M and SMSNet-P acheived 40.24% and 38.23% peptide recall, and 60.4% and 59.39% amino acid recall, respectively.

5.3 Ablation studies

The results of our ablation studies are shown in table 5.2. The peptide recall and amino acid recall degradation range between -0.63 to -4.67% and -0.56 to -2.64%, respectively. The modification which impacted the model the most are the removal of the encoder at -4.67% peptide recall, followed closely by the removal of shift layer at -2.64%.

5.4 Post-processing

We first evaluated the completeness of information contained in the MS/MS spectra from our datasets and found that correctly recovering the whole peptide from MS/MS spectrum could be challenging. By comparing MS/MS spectra in the test sets with their amino acid labels (more details in Methods), the results show that less than 30% of MS/MS spectra in WCU-MS-P and WCU-MS-M dataset con-

Modification	Peptide recall (%)	Difference (%)	Amino acid recall (%)	Difference (%)
SMSNet	44.73	-	64.45	-
No layer normalization	44.10	-0.63	63.89	-0.56
Using cross-entropy loss	43.41	-1.32	63.79	-0.66
No neutral loss	43.01	-1.72	63.00	-1.45
No shift layer	40.23	-4.50	61.90	-2.55
No encoder	40.06	-4.67	61.81	-2.64

Table 5.2: Ablation study results. Removing the shift layer resulted in the most degradation in performance of the model, almost as large as removing the encoder entirely.

tain complete information that would permit correct de novo sequencing of all amino acids. On amino acid-level, among all datasets, at least 24% of all amino acids do not have any mass in the corresponding location in the spectrum that could be used to recover those labels. Only 65.53% of amino acids in WCU-MS-P have corresponding masses, 66.24% in WCU-MS-M, 70.13% in WCU-MS-BEST, and 75.68% in DEEPNOVO-M. As such, we proposed that the model should have a mechanism to filter out some positions from the predicted peptides. Our proposed method works as follows: first, amino acids that have their positional scores below a certain threshold are considered to be ”uncertain predictions” and are masked, then the sequence of masked amino acids is replace with a mass tag equaling it sum of masses to be used for database search. We called this method ”masking.”

Based on the hypothesis and positional score for each amino acid from SMSNet deep learning model, we evaluated three methods which could be used for masking ”uncertain” amino acid: normal thresholding, extended masking, and rescoring with another neural network. In the first method, amino acids with scores lower than a threshold were masked independently then grouped together if it has neighboring masked amino acids. For the second method, we also masked the amino acid immediately following the amino acids masked by the first method. By design, SMSNet is a kind of auto-regressive model commonly used in machine translation, meaning that it assumes the previous output as being correct when making the next prediction. Thus, the model tends to overestimate the confidence of the next prediction even when the previous one has low confidence. This second method (extended masking) was used as a baseline for further comparisons with the rescorer as it outperformed the normal thresholding procedure for every precision level (Figure 5.4).

Nevertheless, because of the auto-regressive property in SMSNet, amino acids predictions and their scores are made based solely on the previous predictions and without the knowledge of any following

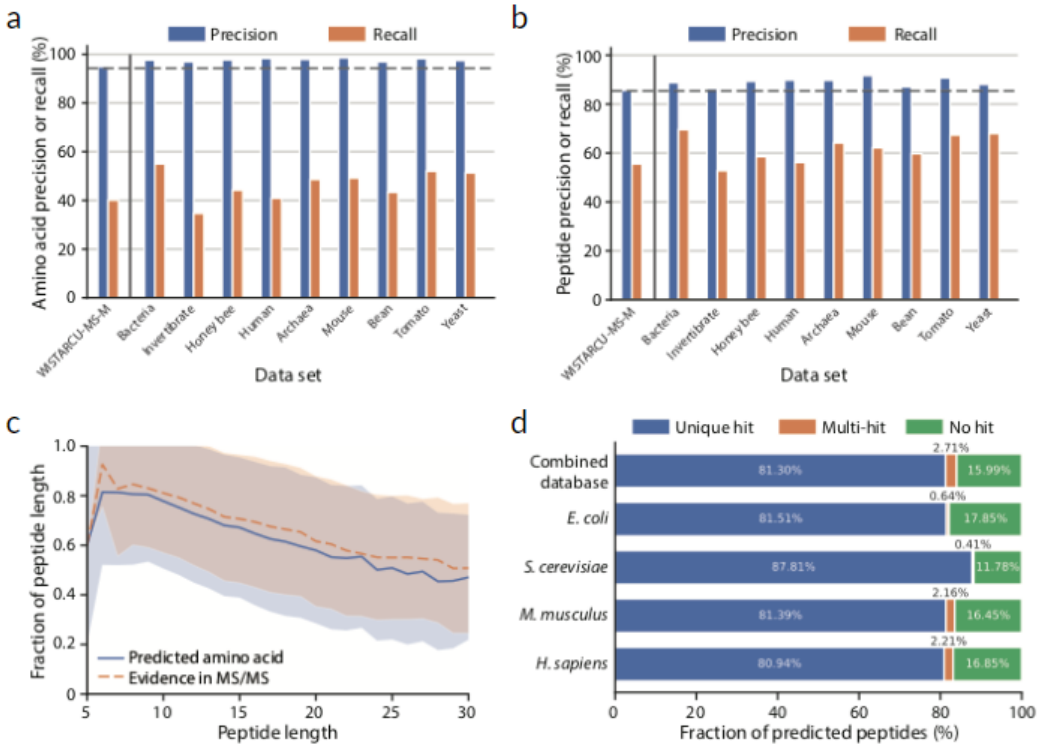


Figure 5.3: Evaulation of SMSNet-M model after rescoring. **(a)** Bar plots showing amino acid-level precisions and recalls of SMSNet on a test set derived from WCU-MS-M and on MS/MS spectra of nine species that comprise the dataset curated by DeepNovo's authors. The threshold on positional confidence score was selected so that 5% amino acid false discovery rate was achieved on the WCU-MS-M test set (the leftmost bars). **(b)** Similar bar plots showing the results at peptide-level. **(c)** Line plots comparing the fraction of predicted amino acid positions that pass the same score threshold used in **a-b** in peptides of various lengths (blue line) to the fraction of amino acid positions that can be definitely determined based on observed ions in the MS/MS spectra (orange dashed line). Shaded area indicate the ± 1 standard deviation ranges. **(d)** Stacked bar plots showing the fraction of predicted peptides that could be matched to various protein sequence databases. Amino acid sequence database for each species was downloaded from Uniprot Consortium (2018) (see Methods). Combined database integrates amino acid sequences from all four species considered. In each bar, only predictions whose ground truths exist within the corresponding database were counted. "Unique hit" means that there the predicted sequence matches to exactly one possibility in the database. "Multi-hit" means that the predicted sequence matches to multiple possibilities. "No hit" means the predicted sequence does not match to anything in the database.

amino acids. The nature of auto-regressive prediction, while ensuring the consistency among amino acids, does not provide any mean to re-calculate previous scores if the next output drastically changes the context of the sequence. Thus, we developed another neural networks-based post-processing model capable of using the whole context of the sequence to adjust the scores of each amino acid. This post-processing model was designed to focus only on rescoring the positional scores of each amino acids without changing

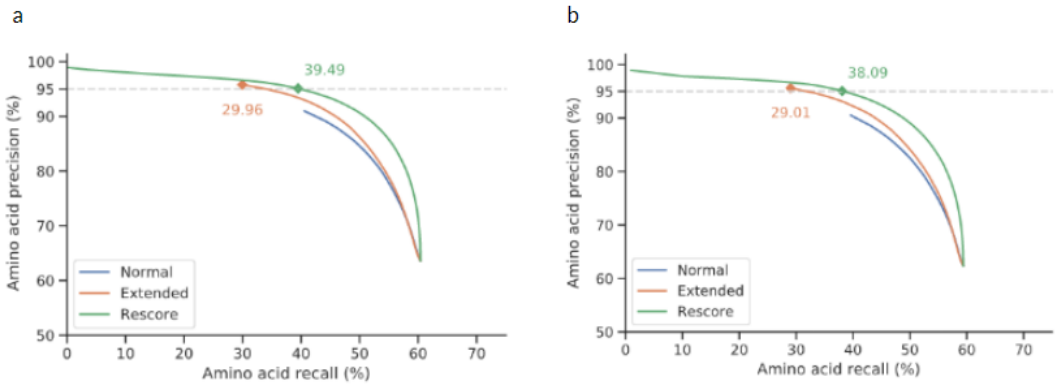


Figure 5.4: **(a)** Precision-recall curves of SMSNets trained on WCU-MS-M. The three lines indicate three masking methods: normal thresholding, extended masking, and rescoring with another neural network. **(b)** Similar plot on WCU-MS-P.

the predicted peptide. We trained this model on the original validation sets using the correctness of the de novo model as labels. The objective of this model is simply to maximize scores of amino acids that match the labels while minimizing ones that do not. We compared this rescoring method with the other two on the same test sets of WCU-MS-M and WCU-MS-P. The rescoring method improved amino acid recalls at 95% precision by 9.53% and 9.08% from the baseline, pushing amino acid recalls to 39.49% and 38.09%, respectively (Fig 5.4). Additionally, at 95% precision, the average number of remaining amino acids of the rescored predictions almost reached the number of amino acids with evidence on the spectrum (Fig 5.3(c)). Furthermore, using the rescoring method, we evaluated the ability of SMSNet to generalize to different species and to datasets that the model was not trained on by using SMSNet-M model to do de novo sequencing on test set of DEEPNOVO-M. Figure 5.3(a,b) shows SMSNet achieves comparable performance across all nine species with 53%-70% peptide recall and 35%-55% amino acid recall at above 95% amino acid precision. Based on these results, the rescoring model was integrated into SMSNet and used by default for any subsequent evaluations.

Lastly, to evaluate the ability to uniquely identify proteins in the database by the partially predicted peptide, we compared the results of database search using the ground truths and our predictions on the top-hit species of WCU-MS-M datasets. For every ground truth peptides in the test set that could be found in the Uniprot database (Consortium, 2018), we counted the number of peptide matched by SMSNet prediction in those species. Because our partial predictions could not be used for string matching, we devised a new searching scheme as described in Methods. The results in Figure 5.3d showed that the partial predictions could still maintain a high number of peptide matches in the database of all species.

In summary, with our novel partially predicting method and database searching scheme, SMSNet could provided de novo sequenced peptides with high amino acids precision while still being useful for searching in a database of known species.

5.5 Applications of SMSNet

To measure the performance of SMSNet on real-world proteomics experiments, we run SMSNet on two datasets from HLA peptidome and human phosphoproteome studies. Both studies shared the same characteristics that some spectra might not have a corresponding peptide and each spectrum might not have a verified true label. To verify the new peptides recovered by SMSNet would require complex external validations which are beyond the scope of this works. Thus, we used our trained models with the cutoff at 5% false discovery rate as in the previous section on these two tasks, then report only the number of discovered peptides instead.

First, we used the SMSNet-M model trained on WCU-MS-M dataset to analyze a large-scale HLA class I peptidome dataset of mono-allelic human B lymphoblastoid cell lines (Abelin et al., 2017) which consists of more than 35 million MS/MS spectra. SMSNet made 95,062 full-sequence predictions and 68,159 partial predictions. From the 20,780 unique fully-predicted sequences, 7,217 peptides were new according to the Immune Epitope Database (Vita et al., 2015). Additionally, most of the SMSNet's identified peptides were also of the right length between 8 and 12 amino acids (88,611 peptides out of 95,062 full-sequence predictions). Note that we did not compare the predicted results in a peptide-to-peptide basis as Abelin et al. (2017) provided only the best spectrum-peptide pairs for each peptide and the database searching software is commercialized.

Then, we used the SMSNet model trained on WCU-MS-P dataset was used to analyze a phosphoproteome dataset of control and epidermal growth factor (EGF)-treated glioblastoma cells (Humphrey et al., 2018), which was previously analyzed by the database searching tool MaxQuant (Cox and Mann, 2008). At 5% amino acid false discovery rate, SMSNet made 181,144 full-sequence predictions and 81,874 predictions with mass tag. As MaxQuant did not output partially predicted peptides, we considered only the full sequence for comparison. Within the full-sequence predicitons, 134,562 peptides are in agreement with the previous study, 1,808 peptides are mismatched, and 45,430 peptides are only answered by SMSNet. This high accordance with MaxQuant predictions illustrates the high accuracy of SMSNet that is on par with state-of-the-art database-search approach.

5.6 Discussion

As SMSNet shares the same underlining concept of neural networks encoder-decoder with DeepNovo, we noted that the key differences between SMSNet and DeepNovo are as follows: (i) SMSNet is explicitly constructed to integrate the nature of mass spectrum and de novo sequencing process into the model, reflected by the shift encoder, and the positioning of the LSTM, (ii) we factorized the bell-shape signals feature processing into two separate steps: detecting signal presences and computing the relevance between signals, and (iii) we incorporated various new techniques from computer vision and machine translation into the model, including focal loss (Lin et al., 2017), residual connection (He et al., 2016), layer normalization (Ba et al., 2016), and new normalization formulas for beam search during inference. Notably, the LSTM layers in SMSNet take both the previous prediction and the output of the candidate ion stack as inputs instead of only the previous prediction as in DeepNovo. Doing so lets the LSTM layers keep track of information not only from the previous prediction but also from the previous input as well. Likewise, the focal loss was used to reflect the nature of MS/MS data where it is relatively easy to determine the presence of an amino acid when all fragmented ions are presented in the spectrum but remarkably harder when some ions are missing or ambiguous. Coincidentally, this feature has also been introduced in a later version of DeepNovo (Tran et al., 2019). Overall, each of these modifications leads to an improvement on the final model.

From the ablation studies, interestingly, the performance degradation by removing the shift layer, which explicitly integrates the knowledge of ion fragmentation into the model, was almost as large as removing the encoder entirely (-4.5%/-2.55% on peptide/amino acid recall compared to -4.67%/-2.64%). Comparing to other modifications which showed lower performance degradation (Table 5.2), this indicates that incorporating domain-specific knowledge is highly critical for handling complex data generated in the field of biotechnology.

For every machine learning system, generalizability is always of great concern. We have shown that SMSNet could acheive high accuracy on MS/MS spectra from a wide range of species and laboratories. We also demonstrated that the SMSNet framework not only can perform well on the data from Q Exactive mass spectrometer with higher-energy collisional dissociation (HCD), but can also adapt to data from different mass spectrometers and peptide fragmentation methods, such as the MS/MS spectra acquired on Orbitrap Fusion Lumos (Zolg et al., 2017).

One limitation in the sequence-mask-search framework of SMSNet is that the ”search” step effectiveness is still highly dependent on the quality of the database provided. Even though multiple databases can be given to SMSNet to search all at once, too much sequence possibilities could also degrade the ability to recover the full sequence from the mass tag. Another possibility is to generate all sequence arrangements that can fit the mass tag, then do a scoring-based search to match the spectrum with the highest score peptide. With the recent advance in using a deep learning-based model for predicting fragment ion intensity (Gessulat et al., 2019), the ”search” step of SMSNet could be improved in the future.

5.7 Summary

In this chapter, we presented the performance of our SMSNet model on all of the WCU-MS datasets and DEEPNOVO-M dataset. We compared SMSNet to the current state-of-the-art *de novo* sequencing tool DeepNovo, and showed that SMSNet could consistently outperformed DeepNovo on all datasets. SMSNet is not limited to any species or the Q Exactive mass spectrometer but could also generalize to other species and other mass spectrometer as well.

Our ablation studies demonstrated the importance of the shift layer in the encoder which engineers the knowledge of the fragmentation process into the model. This suggests that incorporating domain-specific knowledge is highly critical for handling complex data generated in the field of biotechnology.

Chapter VI

CONCLUSION

6.1 Summary

In this thesis, we explored and further improved the task of *de novo* peptide sequencing from MS/MS spectrum. The research is motivated by the limitations of the database-search approaches and the previous *de novo* sequencing tools. The searching technique is highly dependent on the completeness of the database provided, while the *de novo* approach derives peptide directly using the theoretical ion fragmentation process which overcomes the database dependency but is more susceptible to the noise in MS/MS spectrum.

We examined various techniques that are useful for sequencing peptide from MS/MS spectrum, then conceived a new framework, SMSNet, which considerably improve the performance of *de novo* peptide sequencing. SMSNet is a deep learning-based model which incorporates both new deep learning techniques and the specific knowledge of the sequencing process altogether. We proposed the sequence-mask-search frameworks which overcame many problems in prior works. To help boost the accuracy of the raw amino acid prediction from the model, we constructed our neural networks such that each part can focus on only one aspect of the data and became a more specialized model. To account for the fact that the scores of each amino acid prediction often affected by prior outputs, we introduced a rescorer to adjust those score. Finally, to handle the ambiguities normally present in MS/MS spectra, we allowed the model to mask out some amino acid position with low confidence then recovered those positions using the databases of known proteins.

6.2 Future work

Improve the search step

One possibility lies in the ”search” step of the sequence-mask-search frameworks. In this step, the model attempts to resolve the mass tag using known sequence in the database. In the current setting, if a ambiguous peptide can be matched to more than one sequence in the database, the peptide is considered cannot be resolved and discarded. With the recent advancement in database search using deep learning-

assisted prediction of fragment ion intensity (Gessulat et al., 2019), it might be possible to generate all amino acid arrangements that can fit the mass tag, then performs a scoring-based search and picks the most likely sequence.

Incorporate attention mechanism into the model

Another possibility to enhance the performance of SMSNet is in the construction of encoder-decoder framework in the "sequence" model. With many works trying to explore the attention mechanism (Vaswani et al., 2017), it might be possible to incorporate it into SMSnet model. One problem that prevents the use of the attention mechanism in SMSNet is that a MS/MS spectrum has a very sparse structure yet requires very high resolution which makes using the attention in the encoder very computationally expensive. SMSNet's accuracy would be boosted even further if one know how to make it computationally possible to integrate the attention layers into the model.

Input spectrum denoising

There is also a possibility of denoising the input spectrum before using it as input for the model. In this work, we tried to improve the input spectrum quality by removing some peaks according to the following conditions. First, we viewed the set of peaks in the spectrum as a possibly fragmented graph where peaks are nodes and edges only exist between peaks that have mass difference equal to any amino acid mass. Then, we added additional edges that connected two peaks with mass difference equal to the collective mass of any combination of two amino acids. This process essentially created a graph that connected nodes with non-adjacent missing peaks from the input spectrum. Finally, we kept only nodes that connected to N-terminus and discarded the rest. The processed spectrum was supposed to have cleaner data compared to the spectrum before denoising. However, we found that training and testing on the cleaned data resulted in a significant drop of performance to around 10% amino acid recall. The result suggested that only allowing length-one adjacent missing peaks might discard too much information from the input as there might be several longer adjacent missing peaks that are still useful for the model. Furthermore, there might exist other pattern that could be used for detecting the presence of amino acids.

In addition, we also tried training the model on a subset of training spectra with fewer missing amino acid evidences. We defined the existence of each amino acid evidence such that an amino acid is deemed as having support evidences when there is at least one peak present at its b-ion, y-ion, b(2+)-ion,

or y(2+)-ion position, regardless of its intensity. The subset of peptides in WCU-MS-BEST training set which more than 40% of their amino acids have evidences were selected as a clean training set. The results, however, showed the drop of around 3-5% in both amino acid-level and peptide-level recalls in the test set. This indicated that it might be useful patterns for the model to learn on in the peptides with missing evidences that still require further investigation.

REFERENCES

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. 2017. Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity 46.2 (2017): 315–326.

Ba, J. L., Kiros, J. R., and Hinton, G. E. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016):

Bahdanau, D., Cho, K., and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014):

Bengio, Y., Simard, P., and Frasconi, P. 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5.2 (1994): 157–166.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, pp. 177–186. : Springer.

Chambers, M., MacLean, B., Burke, D. R. D. N. S. G. L. F. B., R.and Amode., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S., Nuwaysir, L., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E., Moritz, R., Katz, J., Agus, D., MacCoss, M., Tabb, D., and Mallick, P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 30 (2012): 918–920.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014):

Chung, J., Çaglar Gülçehre, Cho, K., and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014):

Consortium, U. 2018. Uniprot: a worldwide hub of protein knowledge. Nucleic acids research 47.D1 (2018): D506–D515.

Cox, J. and Mann, M. 2008. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature biotechnology 26.12 (2008): 1367.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. 2014. Fast and robust neural network joint models for statistical machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pp. 1370–1380. :

Frank, A. and Pevzner, P. 2005. Pepnovo: de novo peptide sequencing via probabilistic network modeling. Analytical chemistry 77.4 (2005): 964–973.

Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., and Wilhelm, M. 2019. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nature Methods 16.6 (2019): 509–518.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep learning, volume 1. MIT press Cambridge.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. :

Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. Neural computation 9.8 (1997): 1735–1780.

Humphrey, S. J., Karayel, O., James, D. E., and Mann, M. 2018. High-throughput and high-sensitivity phosphoproteomics with the easyphos platform. Nature Protocols 13 (2018): 1897–1916.

Kingma, D. P. and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014):

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86.11 (1998): 2278–2324.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. nature 521.7553 (2015): 436.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pp. 2980–2988. :

Ma, B. 2015. Novor: real-time peptide de novo sequencing software. Journal of the American Society for Mass Spectrometry 26.11 (2015): 1885–1894.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. 2003. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid communications in mass spectrometry 17.20 (2003): 2337–2342.

Medzihradszky, K. F. and Chalkley, R. J. 2015. Lessons in de novo peptide sequencing by tandem mass spectrometry. Mass spectrometry reviews 34.1 (2015): 43–63.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013):

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning representations by back-propagating errors. nature 323.6088 (1986): 533.

Sutskever, I., Vinyals, O., and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pp. 3104–3112. :

Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. 2017. De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences 114.31 (2017): 8247–8252.

Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. 2019. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nature Methods 16 (2019): 63–66.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998–6008. :

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. 2015. Sequence to sequence-video to text. In Proceedings of the IEEE international conference on computer vision, pp. 4534–4542. :

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164. :

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. 2015. The immune epitope database (iedb) 3.0. Nucleic Acids Research 43.D1 (2015): D405–D412.

Werbos, P. J. 1990. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE 78.10 (1990): 1550–1560.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016):

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pp. 2048–2057. :

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. 2012. Peaks db: de novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics 11.4 (2012): M111–010587.

Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H.-C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R. L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., and Kuster, B. 2017. Building proteometools based on a complete synthetic human proteome. Nature Methods 14 (2017): 259–262.

Appendix I

HYPERPARAMETER TUNING

For the main model training, we considered the followings hyperparameter ranges:

- Learning rate: 0.01 to 0.5
- Focal loss α: 0.25 to 2.0
- Focal loss γ: 0.5, 1.0, 1.5, 2.0
- Max gradient norm: 1.0, 5.0
- Embedding size: 5, 32, 64
- Candidate ion stack resolution: 0.01 Da, 0.02 Da, 0.05 Da, 0.1 Da

The hyperparameter that affected the model the most is the learning rate, which varied the final amino acid recall up to 5%. The rest of the hyperparameters contributed up to 1-2% amino acid recall on the test set of WCU-BEST. For the focal loss, any other combination of α and γ beside the one used in our final model resulted in the model not converging at all. We chose the largest layer size for all layers such that the model could still fit in the GPU memory.

Biography

Korrawe Karunratanakul was born in Bangkok on February 21, 1996. He received a bachelor's degree in computer engineering from Chulalongkorn University in 2018.