Deep Learning Methods for Novel Peptide Discovery and Function Prediction

by

Shaokai Wang

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

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

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

External Examiner:	Fangxiang Wu Professor, Depart. of Computer Science University of Saskatchewan
Supervisor(s):	Bin Ma Professor, Depart. of Computer Science University of Waterloo
Internal Member:	Yaoliang Yu Associate Professor, Depart. of Computer Science University of Waterloo
	Yang Lu Assistant Professor, Depart. of Computer Science University of Waterloo
Internal-External Member:	Andrew Doxey Associate Professor, Depart. of Biology University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Statement of Contribution

I would like to thank all my co-authors who made this thesis possible. And here, I details the contributions of each co-author for all work presented in my thesis.

- Chapter 3. **Shaokai Wang**, Ming Zhu, Bin Ma. **NeoMS: Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry.** *International Symposium on Bioinformatics Research and Applications(ISBRA). Singapore: Springer Nature Singapore, 2023: 280-291.*

  This work has been accepted by ISBRA 2023, and an extension of this research has been invited for submission to the IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). I am the first author of this work and was responsible for conducting all experiments and writing the manuscript. The conceptual framework for this research was developed through weekly discussions with Bin, who provided the initial idea and helped shape the study. Ming offered valuable suggestions on manuscript writing. All co-authors participated in the review process, providing extensive feedback through multiple rounds of revision.

- Chapter 4. **Shaokai Wang**, Bin Ma. **Deep learning boosted amyloidosis diagnosis.** *IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023: 57-62.*

  This work has been accepted by BIBM 2023. I am the first author and have contributed by proposing the main idea, conducting all experiments, and writing the paper. The majority of this research was completed during my internship at Rapid Novor Inc., where the project was initiated by Anupa and other colleagues. Bin provided valuable advice on experimental design and manuscript preparation. All co-authors have thoroughly reviewed the manuscript and provided extensive feedback over several iterations.

- Chapter 5. **Shaokai Wang**, Bin Ma. **Anti-cancer peptides identification and activity type classification with protein sequence pre-training.** *IEEE Journal of Biomedical and Health Informatics(J-BHI), 2024*

  This work has been accepted by J-BHI in 2024. I am the lead author, having proposed the main concept, carried out all experiments, and undertaken the writing of the manuscript. Bin provided invaluable guidance on the design of the experiments and the drafting of the paper. All co-authors have rigorously reviewed the manuscript and provided comprehensive feedback through multiple iterations.

- Chapter 6. **Shaokai Wang**, Bin Ma. **Novel fine-tuning strategy on pre-trained protein model enhances ACP functional type classfication**.

  This manuscript is ready to submit to Computers in Biology and Medicine (CBM). I hold the position of first author, having initiated the main idea, carried out all related experiments, and undertaken the entirety of the manuscript writing. Bin provided essential guidance on the experimental design and the composition of the manuscript. All co-authors have meticulously reviewed and contributed feedback to the manuscript through several rounds of revisions.

Abstract

This thesis explores deep learning methods for protein identification and property prediction, encompassing two primary areas: mass spectrometry-based protein sequence identification and protein property prediction. We introduce a method that enhances the identification rate of MHC-I peptides and facilitates the discovery of novel mutated MHC-I peptides. In the domain of property prediction, we present three novel approaches for the early diagnosis of amyloidosis, the discovery of anticancer peptides and the classification of anticancer peptide functional type.

NeoMS: Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry [142]**.** The study of immunopeptidomics requires the identification of both regular and mutated MHC-I peptides from mass spectrometry data. For the efficient identification of MHC-I peptides with either one or no mutation from a sequence database, we propose a novel workflow: NeoMS. It employs three main modules: generating an expanded sequence database with a tagging algorithm, a machine learning-based scoring function to maximize the search sensitivity, and a careful target-decoy implementation to control the false discovery rates (FDR) of both the regular and mutated peptides. Experimental results demonstrate that NeoMS both improved the identification rate of the regular peptides over other database search methods and identified hundreds of mutated peptides that have not been identified by any current methods. Further study shows the validity of these new novel peptides.

Deep learning boosted amyloidosis diagnosis [140]**.** Amyloid light chain (AL) amyloidosis is a disorder characterized by the deposition of antibody light chains in organs. The importance of early and accurate diagnosis in AL amyloidosis cannot be overstated, as it enables timely implementation of appropriate treatment strategies and improves patient outcomes. Therefore, developing a highly accurate method using antibody sequencing and computational techniques is crucial to address this urgent need. While several computational methods have been developed to predict AL amyloidosis, they heavily depend on manually extracted features, and their performance falls short of satisfactory levels. We present DeepAL, a deep learning-based approach to predict AL amyloidosis with high precision. DeepAL utilizes a pre-trained model to extract light chain features and then trained with AL amyloidosis knowledge. In evaluations conducted on two benchmark datasets, DeepAL surpasses the performance of previous approaches. Additional experiments demonstrate that features extracted from the pre-trained model have significantly enhanced overall performance.

Anti-cancer peptides identification and activity type classification with protein sequence pre-training [141]**.** Cancer remains a significant global health challenge,

responsible for millions of deaths annually. Addressing this issue necessitates the discovery of novel anti-cancer drugs. Anti-cancer peptides (ACPs), with their unique ability to selectively target cancer cells, offer new hope in discovering low side-effect anti-cancer drugs. We introduce DUO-ACP, a model serving dual roles in ACP prediction: identification and functional type classification. DUO-ACP employs two embedding modules to acquire knowledge about global protein features and local ACP characteristics, complemented by a prediction module. When assessed on two publicly available datasets for each task, DUO-ACP surpasses all existing methods, achieving outstanding results. We further interpret the contribution of each part of our model, including the two types of embeddings as well as ensemble learning. On a new curated dataset, the prediction results of DUO-ACP closely match existing literature, highlighting DUO-ACP's generalization capabilities on previously unseen data and displaying the potential capability of discovering novel ACP.

Novel fine-tuning strategy on pre-trained protein model enhances ACP functional type classification. Cancer remains one of the most formidable health challenges globally. ACPs have recently emerged as a promising new therapeutic strategy, recognized for their targeted and efficient anti-cancer properties. To fully leverage the potential of ACPs, computational methods that can accurately discover and predict their functional types are indispensable. We present ACP-FT, a deep learning model that is fine-tuned from a pre-trained protein model specifically for predicting the functional types of ACPs. Employing a novel fine-tuning approach alongside an adversarial model training technique, our model surpasses existing methods in classification performance on two public datasets. Additionally, we provide a thorough analysis of our training strategy's effectiveness. The experimental results demonstrate that our two-step fine-tuning approach effectively prevents catastrophic forgetting in the pre-trained model, while adversarial training enhances the model's robustness. Together, these techniques significantly increase the accuracy of ACP functional type predictions.

Acknowledgements

I would like to express my deepest gratitude to Professor Bin Ma for his invaluable guidance, patience, and support throughout the course of my PhD journey. His wisdom, knowledge, and commitment to academic excellence have been a constant source of inspiration and have significantly contributed to my personal and professional growth. Professor Ma's mentorship has been instrumental in shaping my research direction and in helping me to navigate the challenges of academic research.

I would also like to extend my thanks to the committee members: Fangxiang Wu, Lu Yang, Yaoliang Yu, and Andrew Doxey. Their insightful feedback, constructive criticisms, and expert advice have been vital in refining my research and enriching my learning experience. I am truly grateful for their time, effort, and dedication in guiding me.

I am immensely grateful to my lab mates: Jiarong Wu, Xiangyuan Zeng, Anupa Murali, Johra Moosa, Shengheng Guan, Soroosh Gholamizoj, and Zhenbo Li. Each of them has contributed to my experience in unique and meaningful ways, providing both academic collaboration and personal camaraderie. Their support, encouragement, and shared insights have greatly enriched my time in the laboratory and have made my research journey both enjoyable and rewarding.

My research would not have been possible without the financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada, the MITACS Accelerate PhD Fellowship program, and Rapid Novor Inc. I am profoundly thankful for their generous funding and for believing in the potential of my research. Their support has been crucial in providing the resources and opportunities necessary for my academic endeavors.

Lastly, I extend my heartfelt thanks to my parents and friends, who have provided me with unwavering support, encouragement, and love throughout my PhD journey. Their belief in me and their sacrifices have been my constant source of strength and motivation. I am forever grateful for their presence in my life and for all the ways they have helped me to achieve this milestone.

Dedication

This is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Proteins are large, complex molecules that play an indispensable role in biological processes, composed of smaller units known as amino acids. These amino acids are linked by peptide bonds into long chains, forming the protein's primary structure [30]. In nature, there primarily exist 20 different amino acids that can combine in nearly limitless sequences to create a vast diversity of proteins. Once synthesized, these amino acid chains fold, coil, and arrange themselves into specific three-dimensional configurations. The unique shapes and compositions of proteins enable them to perform a variety of biological functions. For instance, enzymes, a specialized class of proteins, significantly accelerate biochemical reactions while remaining unchanged themselves, a critical process in metabolism and other biological activities. Additionally, hemoglobin plays a crucial role in oxygen transport within the bloodstream. Antibodies, another type of protein, are essential for immune responses, acting both as defensive agents and signal receptors.

Proteins are essential for biological processes in cells and tissues. It is crucial to deepen our understanding of their roles through research. This is especially important as many diseases are result from protein malfunctions, including mutations, misfoldings, or aberrant expressions, which are central to uncovering the molecular mechanisms of various conditions [53]. Understanding these disease mechanisms provides a critical pathway to developing targeted treatments. Research that focuses on modulating protein functions through drugs aims to discover more effective medications with fewer side effects. Additionally, protein research is fundamental in understanding the functioning of antibodies

within the immune system. This knowledge is crucial for vaccine development and for understanding autoimmune diseases, allergies, and immune deficiencies [75].

The emergence of bioinformatics has fundamentally transformed protein research, arming scientists with advanced tools and methods to decode the mysteries of amino acid sequences, and their structures and functions. This involves comparing amino acid sequences across different proteins, understanding their evolutionary relationships, pinpointing conserved sequences, and predicting proteins' functions and active sites. The development of computational methods has marked significant breakthroughs in accurately determining protein functions and structures [73]. Moreover, the study of protein interactions and networks shedding light on how proteins collectively contribute to various biological processes within cells. This includes identifying interaction partners, constructing interaction networks, and analyzing the dynamics of these networks. In addition to facilitating the study of protein dynamics, analytical tools have streamlined the process of annotating protein functions and led to the creation of large protein databases with annotations, for example, protein structure [7], homology [25], protein-drug interaction [147].

Aside from understanding the fundamental biological process, bioinformatics is also pivotal for transforming protein research into practical clinical uses. Through the analysis of genetic and proteomic information, it is possible to identify biomarkers that assist in diagnosing diseases and prognosing patient responses to specific therapies [92]. Additionally, in the area of pharmaceutical research, computational techniques aids in screening for drug candidates and understanding the molecular basis of drug resistance [101].

The advanced machine learning and deep learning technologies has markedly improved the accuracy of protein studies. These highly accurate and robust computational methods make them feasible in supporting medical applications including peptide identification, disease diagnose and drug discovery. The extensive amount of annotated experimental data facilitates the development of machine learning models. For example, by combining the computational methods for homology modeling, fold identification and *de novo* prediction methods, AlphaFold [61] has marked significant breakthroughs in accurately determining protein structures.

Even though much achievements have been made in bioinformatics area, there are still lots of challenges. Initially, contemporary mass spectrometry-based peptide identification techniques are heavily dependent on protein database searches. This approach encounters significant challenges when attempting to identify peptides with mutations, such as those found in human leukocyte antigens (HLA), which often cannot be matched in the protein database. Additionally, mutations within peptides can alter their structures and functions. Such alterations, especially when occurring in human antibodies, can impact the immune

system's functionality, potentially leading to disease. Given the variability in antibodies, predicting the functional outcomes of these changes is complex. Lastly, with the advancement of biotechnological methods, protein drugs have emerged as promising candidates for cancer treatment, primarily due to their minimal side effects. However, the current screening processes for these protein drugs are both time-intensive and labor-intensive, presenting significant hurdles to their rapid development and deployment.

In this thesis, we address significant challenges in peptide identification, disease diagnosis, and drug discovery by providing computational solutions. Acknowledging the urgent need for enhanced methods in these fields, we delve into a comprehensive exploration of advanced machine learning and deep learning techniques within protein research. Our primary goal is to develop practical, highly accurate methods suitable for biological and medical applications. Through extensive experimentation across diverse domains, we demonstrate the effectiveness and practicality of our approaches. Our research is aimed at advancing the fields of protein sequence identification and function prediction, thereby offering valuable insights and tools for improved disease diagnosis and drug discovery strategies.

1.2 Thesis overview

This thesis explores the deep learning methods for protein identification and function prediction with various approaches to enhance performance and robustness.

In Chapter 2, we provide a comprehensive introduction to the fundamental concepts of protein identification and function. We first introduce the protein sequence identification tool LC-MS/MS and discuss computational methods for interpreting mass spectra, including peptide database search, *de novo* peptide sequencing and sequence based spectrum prediction. Additionally, we cover peptide function prediction techniques, ranging from traditional handcrafted methods to more advanced machine learning and deep learning approaches. We also explore the development of protein language models (pLMs) and their application in downstream tasks. This foundation is crucial for understanding the advanced topics discussed in subsequent chapters.

In Chapter 3, our research concentrates on improving LC-MS/MS-based peptide identification, particularly in the field of immunopeptidomics, which involves identifying both standard and mutated MHC-I peptides. We observed that current database search methods struggle with non-tryptic peptides. By leveraging peptide sequence-based spectrum prediction and retention time prediction, we enhance to distinguish between real peptide

spectrum matches and decoy matches. We propose utilizing the disparity between predicted and experimental spectra as key features for a machine learning classifier. Additionally, recognizing the lack of methods for statistically validating mutated peptides' identification, we explored *de novo* sequencing to create personalized peptide sequences, facilitating the identification of mutated peptides through a personalized database search. This comprehensive approach enables an end-to-end workflow for identifying regular MHC-I peptides and those with mutations.

In Chapter 4, we investigate methods for diagnosing diseases based on protein sequences, specifically focusing on accurately diagnosing AL amyloidosis from antibody light chains. We address the challenge posed by the limited size of datasets and acknowledge that existing sequence-based and structural-based methods for diagnosis show limited promise, with a high risk of overfitting. To overcome these issues, we leverage pre-trained protein models to generate sequence features, significantly reducing the overfitting problem. Furthermore, we explore the use of ensemble learning to fully utilize the training data and introduce a novel loss function to mitigate the imbalance between positive and negative labels. Through these approaches, we propose a highly accurate method for identifying AL amyloidosis from light chain sequences.

In Chapter 5, we expand our exploration into anticancer drug discovery, focusing on Anticancer Peptide (ACP) identification and functional type classification. We critique current methods for their lack of accuracy and depth in addressing ACP's functional classification. Utilizing pre-trained pLMs has shown promise in ACP identification, yet there is potential for enhancement. We introduce a novel approach that captures both global protein features and specific local features relevant to anticancer peptides. Additionally, we tackle the issue of integrating pre-trained and randomly initiated modules within a single system, proposing a two-stage training strategy to harmonize these components. This method enables our model to not only accurately identify ACPs but also classify their functional cancer types effectively.

In Chapter 6, we build upon our anticancer type classification work by enhancing the utilization of pLMs. We discovered that fine-tuning the pLM, rather than keeping it static, significantly improves predictions. Yet, aligning the pre-trained model with a newly initiated projection head poses challenges, often leading to underdeveloped projection heads and distorted features. Meanwhile, with limited sequence data for ACP prediction, we find introducing slight noise to the input enhances model robustness. We propose a novel approach that includes direct fine-tuning of the pre-trained model with adversarial training to boost robustness and a two-step training strategy to address inconsistencies between the pre-trained model and the projection head.

In Chapter 7, we conclude this thesis by summarizing our main contributions and achievements in advancing peptide identification, disease diagnosis, and drug discovery through machine learning and deep learning techniques. We reflect on the successful application of these methods to address complex biological problems, highlighting our innovative approaches in protein sequence analysis and anticancer drug discovery. Looking forward, we discuss potential research avenues that could further explore the integration of advanced computational models with biotechnology, aiming to unlock new insights and methodologies in the field of bioinformatics and personalized medicine.

1.3 Contributions

This thesis makes several important contributions in peptide identification, disease diagnosis, and drug discovery.

Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry

This section contributes an end-to-end peptide identification workflow that can identify both regular MHC-I peptide and MHC-I peptide with one amino acid mutation.

- The proposed workflow, NeoMS, outperforms current database search and post-processing methods in identifying a greater number of regular peptides.
- NeoMS demonstrates the ability to identify MHC-I peptides with single amino acid mutations, validated statistically.
- Experimental results indicate that these mutated peptides have a high likelihood of binding with MHC-I and exhibit a distribution similar to that of regular MHC-I peptides.

Deep Learning Boosted Amyloidosis Diagnosis

This section contributes a deep learning based model for early diagnosing AL amloidosis with human antibody light chain.

- This work introduces DeepAL, the first initiative to leverage a light chain pre-trained model for improving the accuracy of AL amyloidosis identification.
- The proposed DeepAL method surpasses other existing sequence-based identification approaches in performance.

- By integrating a structure-based method, DeepAL demonstrates superior performance, particularly in high scoring regions.

Anti-Cancer Peptides Identification and Activity Type Classification with Protein Sequence Pre-training

This section contributes an protein pre-trained model-based method that can both identify anticancer peptides and classify its functional cancer type.

- The proposed method, DUO-ACP, excels beyond previous methods in both ACP identification and ACP type classification.
- A novel two-step training strategy effectively bridges the gap between the pre-trained module and the randomly initiated module.
- Experimental results indicate that the ensemble learning method effectively mitigates issues arising from insufficient training data.

Novel Fine-tuning Strategy on Pre-trained Protein Model Enhances ACP functional Type Classfication

This section contributes a novel fine-tuning strategy that increases the accuracy of ACP functional type classification.

- We demonstrate that fine-tuning the pre-trained model results in superior performance compared to linear probing.
- Our innovative two-step fine-tuning approach further enhances the model's performance.
- Adversarial training, which introduces perturbations into the feature embedding, increases model robustness.

Chapter 2

Background

In this chapter, we introduce the basic background of this thesis. There are four sections: LC-MS/MS for protein sequencing; Protein sequence based function and property prediction; Protein sequence-based spectrum prediction and pre-trained protein language model.

- In the first section, we lay out the fundamentals of Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) and explore some traditional computational techniques employed to decode spectra into peptide sequences. This exploration not only underscores the importance of LC-MS/MS in proteomics but also highlights the computational challenges and solutions in interpreting spectral data.
- In the second section, we explore computational strategies that derive insights from protein sequences, aiming to predict their respective functions and properties. These foundational approaches set the stage for understanding the complex interplay between sequence characteristics and biological roles.
- In the third section, our focus shifts to deep learning-based methods that have significantly enhanced the accuracy and efficiency of sequence identification. By examining these modern approaches, we illustrate the transformative impact of deep learning on proteomic analysis, offering insights into how these methods outperform traditional computational strategies.
- Finally, the fourth section is dedicated to the construction and application of pre-trained protein language models. We discuss the intricacies of building these models and how they can be effectively utilized. This discussion aims to illuminate the

promising intersection of artificial intelligence and proteomics, showcasing the potential of pre-trained models in advancing our understanding of proteins.

2.1 LC-MS/MS for protein sequencing

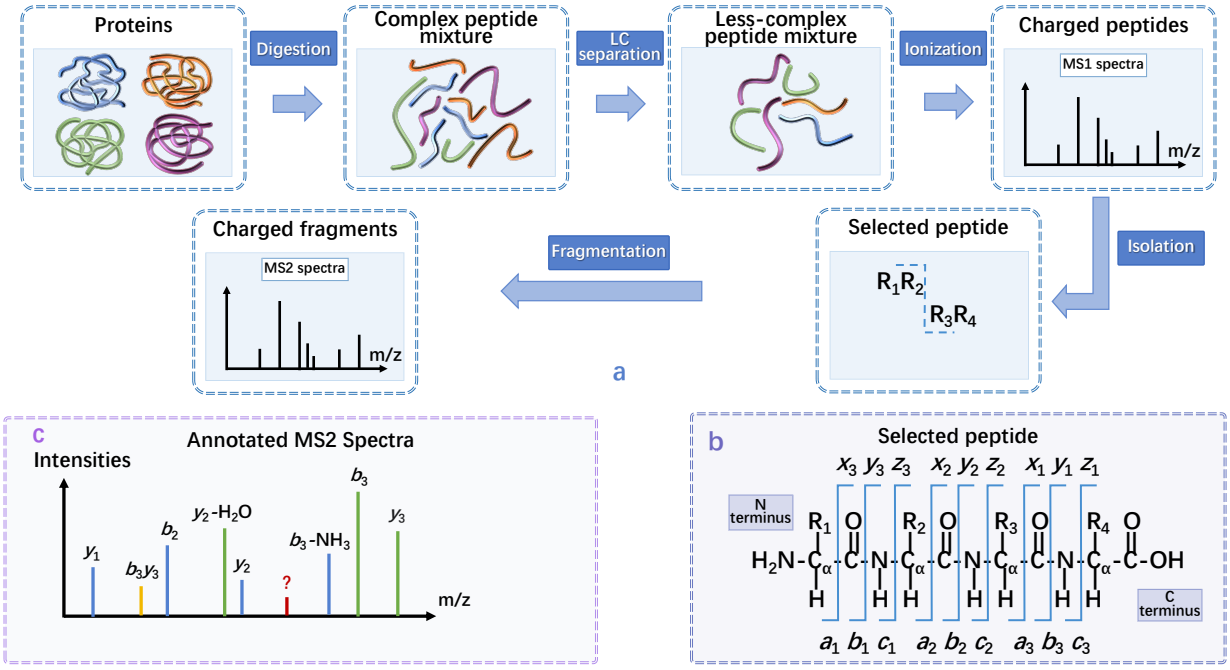

Figure 2.1: (a) The workflow of shotgun proteomics workflow illustrating how to identify protein sequence using LC/MS-MS. (b) An example of a peptide comprising four amino acids, labeled from R1 to R4. Each of which can represent any of the 20 standard amino acid side chains. The blue symbols denote the products resulting from the breakages of main chain bonds. From the N-terminus, a-, b-, c-, three types of fragmentated ions can be produced; From the C-terminus, x-, y-, z-, three types of fragmentad ions can be produces. (c) An example shows a spectrum whose peaks are annotated by the fragmentation of a peptide.

In the field of bottom-up proteomics, commonly referred to as shotgun proteomics, the technique of liquid chromatography tandem mass spectrometry (LC-MS/MS, also known as LC-MS2) stands as the principal method for both identifying and quantifying peptides

and proteins. This approach combines the capabilities of high-performance liquid chromatography (HPLC) with tandem mass spectrometry (MS/MS), facilitating the detection of peptides within intricate mixtures of proteins.

The preparatory stage of LC-MS/MS involves the enzymatic digestion using a protease, usually trypsin, of protein mixtures into smaller peptide fragments. Following this, the LC-MS/MS process in three primary stages. Initially, during the liquid chromatography (LC) phase, peptides are separated within a liquid mobile phase against a solid stationary phase. This separation is based on their distinct retention times (RTs) as they move through the LC column. Subsequently, the peptide samples are ionized, accelerated, and subjected to a first round of mass spectrometric analysis (MS1). In data-dependent acquisition (DDA), peptide precursors are selected in narrow isolation windows in order to select single molecular species. Isolated ions from the MS1 spectrum are then fragmented and further analyzed in a secondary mass spectrometry stage (MS2), producing a detailed spectrum of the ion fragments. The workflow is illustrated in Figure 2.1(a).

The choice of fragmentation techniques, such as collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), or electron transfer dissociation (ETD), plays a pivotal role in MS2 analysis. The fragmentation of ionized peptides can theoretically generate three pairs of ions: a- and x- ions; b- and y- ions; c- and z- ions, as depicted in Figure 2.1(b). These fragmentation techniques produce these ions with varying frequencies [58].

With the knowledge of the physical method of fragmentation, the masses of the peptide fragment can be easily computed from the sequence. An illustrative spectrum with annotation is depicted in Figure 2.1(c). Subsequently, computational methods can analyze the signal intensity of ions across different mass-to-charge (m/z) ratios in the spectrum to deduce the corresponding peptide sequence. The identification of peptides can be seen as two main types: database search, which find the best matching peptided from a given the database, and *de novo* sequencing, where the peptide is directly derived from the spectra without using an existing database. The differences of these two methods are illustrated in Figure 2.2.

2.1.1 Database Searching

Database searching in proteomics typically involves finding the best-matched peptides for a given spectrum. This process encompasses several key steps:

- **Protein Database Creation**: The first step is establishing a protein database,

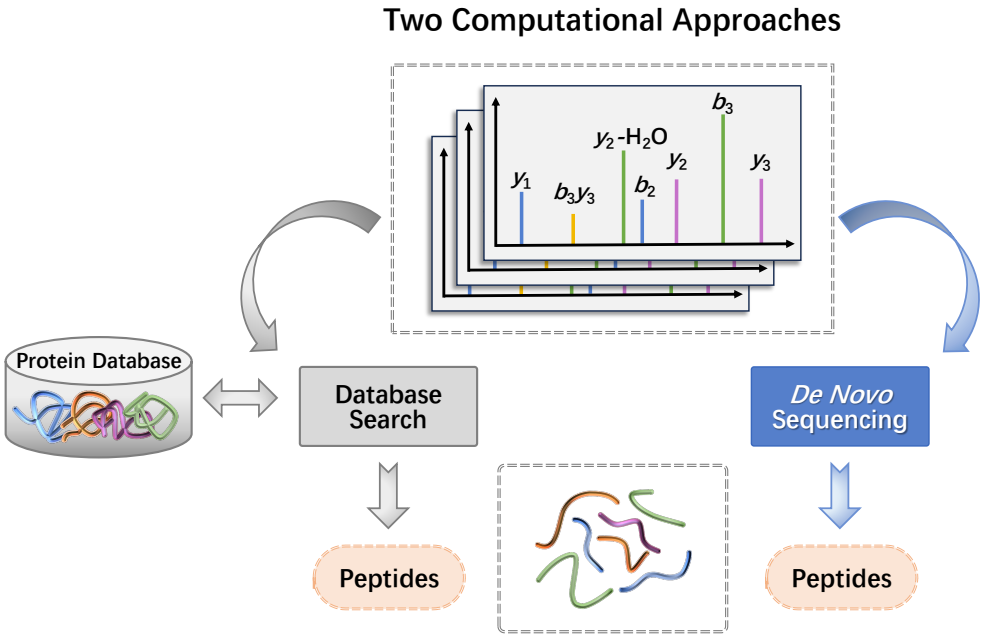

Figure 2.2: The differences between database search and *de novo* sequencing.

which provides a defined search space. This database contains known protein sequences against which the experimental spectra can be compared.

- **Quality Filtering**: The second step involves filtering to remove low-quality peptide candidates while retaining high-quality ones. This selective process is crucial for identifying potential Peptide-Spectrum Matches (PSMs), where candidate peptides are paired with the corresponding spectrum.

- **Scoring Function**: The third step requires an effective scoring function or a scorer. This function assigns scores to the PSMs, effectively ranking them based on how well the peptide candidates match the experimental spectrum. The scoring can be based on various criteria, including the intensity of the spectral peaks, the completeness of the peptide fragmentation, and the accuracy of the mass measurements.

- **Ranking and Output**: The final step is to rank all the PSMs according to their scores. The PSMs that surpass a predefined threshold are then outputted. This threshold is often determined based on the desired level of confidence or the acceptable false discovery rate. High-ranking PSMs are considered to be the most accurate

matches between the spectra and the peptides in the database, thereby providing the most likely identification of the peptides present in the sample.

By following these steps, database search tools in proteomics facilitate the identification of peptides from complex mixtures, contributing significantly to the understanding of protein function and interaction in biological research.

In the field of proteomics, particularly in peptide database searches, the absence of ground truth labels complicates the direct evaluation of search results' quality. To increase confidence in large-scale protein identifications, the target-decoy search strategy is commonly utilized [39]. In this approach, decoy proteins—generated by reversing [96] or shuffling [69] sequences from the original proteins—are incorporated into the database at the initial search step. As the decoy database is intentionally unrelated to the actual protein database, the identification of a decoy PSM represents a scenario where the null hypothesis is validated. Following the search against a combined target-decoy database, calculating the p-value becomes a straightforward method for estimating significance. p-value for a given score s is defined as the proportion of decoy PSMs achieving score s or higher.

However, employing a p-value as a solitary threshold is insufficient due to the extensive number of statistical tests conducted [64]. A correction for multiple testing, such as the False Discovery Rate (FDR), becomes essential. As illustrated in Figure 2.3, given a score threshold s, the decoy samples with score above s is considered as false positive. Due to the null hypothesis, there are the same amount of false positive samples in the identified target samples. The FDR computes the ratio of the false samples in the identified samples. The simplest method for calculating FDR is analogous to determining p-values: for a selected threshold, d denotes the number of decoy PSMs surpassing this threshold, and t denotes the count of target PSMs above the same threshold. The FDR at this threshold is then calculated as follows:

$$FDR = \frac{d}{t} \tag{2.1}$$

Similarly, when calculating the FDR for peptides, PSMs are deduplicated based on their spectrum. Since the FDR curve has the issue that two different scores can lead to the same FDR, a correction of FDR known as q-value is introduced to address this issue. The q-value is defined as the minimum FDR threshold at which a given PSM is accepted. After this correction, the score threshold is determined: target PSMs with FDR below this limit are considered valid. For instance, at a FDR threshold of 1%, if 500 PSMs are accepted, we expect that approximately five of these matches may be incorrect. Under the same threshold, PSMs that accepted by any database search tool are considered to be

of same quality. Consequently, the FDR becomes essential for assessing the precision of peptide identification tools: With a consistent FDR threshold—commonly set at 0.01 or 0.05—the tool that identifies a greater number of PSMs is regarded as more effective in distinguishing between target and decoy PSMs. This approach offers a reliable metric for evaluating identification accuracy in the absence of a ground truth.

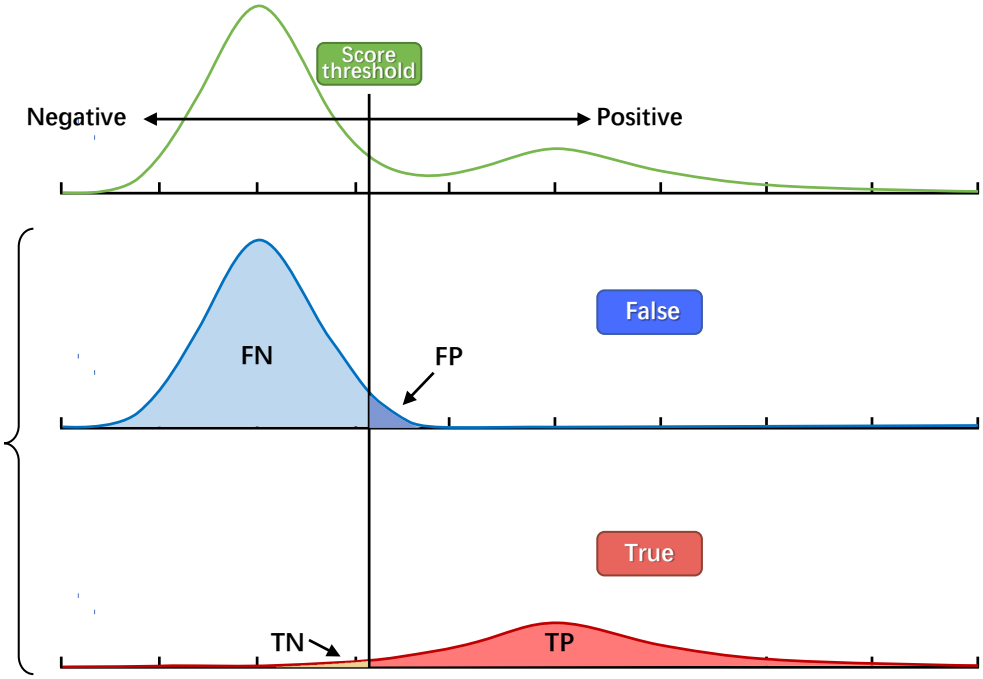

Figure 2.3: The score distribution curve. The green curve is the overall score distribution. The blue curve is the score distribution of decoy sequences and the red curve is it of target sequences.

The development of database search tools for proteomics has significantly evolved since the introduction of Sequest software [42] in 1994, which uses a correlation-based scoring system to match experimental mass spectra against theoretical spectra generated from known protein sequences. Mascot [105] uses a probabilistic scoring model compares experimental data to known protein sequences in a database. It can handle data from various types of mass spectrometers and supports a wide range of data formats. MaxQuant [28] is a computational platform specifically designed for the analysis of large-scale mass-spectrometric data. MaxQuant incorporates advanced algorithms Andromeda [29] for can handling mass spectrum data with arbitrarily high fragment mass accuracy. PeaksDB [159] improves the identification rate by incorporating the *de novo* sequencing results into the database

search. Comet [41] is valued for its speed and efficiency in processing mass spectrometry data. MSGF+ [68] has become notable for its accuracy and statistical models in identifying peptides with unusual modifications or in unexpected charge states. Together, these tools have transformed our ability to identify and quantify proteins, propelling advancements in biological and medical research.

2.1.2 De novo Sequencing

De novo sequencing direcly interpret peptide sequence from the given spectrum. It is comparatively a more difficult problem since it does not have a limited searching range. The aim of *de novo* sequencing is to generate same sequences as it searched from database. Since 1997, many computational methods have developed for this task: Lutefisk [126] that use a graph theory approach was the pioneering software to generate peptides through tandem mass spectrometry using *de novo* methods. PEAKS [87] uses a new model for noise filtering and peak centering and a new dynamic programming algorithm to efficiently compute the best peptide sequences whose fragment ions can best interpret the peaks in the MS/MS spectrum. PepNovo [47] uses a probabilistic network modeling method that obtains higher accuracy. It reflects the chemical and physical rules that govern the peptide fragmentation. With the development of fragmentation technology in mass spectrometer, HCD produces high mass accuracy MS2 spectra without the low-mass cutoff associated with CID in ion trap instruments, providing the avenue for more accurate *de novo* sequencing method. pNovo [23] designs a method that have favorable features to help overcome the obstacles in high resolution *de novo* peptide sequencing. pNovo+ [21] notice the missing fragmentation information in one spectrum may be found in the other, so the antisymmetry restriction is removed and an efficient algorithm pDAG to find the k longest paths is proposed that significantly improves the speed. More mass spectrum data are produced and available on website like ProteomeXchange [34]. This opens a avenue for machine learning based methods that learns from large amount of data. Novor [86] employs probabilistic machine learning models to deduce amino acids and their corresponding quality scores. pNovo3 [150] implement a learning-to-rank framework to distinguish similar peptide candidates for each spectrum. Deepnovo [130] integrates CNN and LSTM network: a spectrum-CNN to learn features of tandem mass spectra, an ion-CNN to learn fragment ions, and a LSTM for learning sequence patterns and predicting peptides. PointNovo [109] is a instrument resolution independent *de novo* peptide sequencing method. It uses a different spectrum representation method with T-Net for feature learning. Casanovo employs a transformer encoder-decoder model that directly translates the spectrum into a peptide sequence [154]. In this model, the spectrum serves as the input for the encoder, while the

precursor and prefix sequence are fed into the decoder. ContraNovo, on the other hand, shares a similar model architecture with Casanovo but includes an additional peptide encoder [59]. Its training involves multi-task learning aimed at minimizing both the sequence difference between the predicted sequence and the actual sequence and the contrastive loss of the sequence representation and spectrum representation.

2.2 Protein sequence-based function and property prediction

Advancements in analytical instruments have rendered protein sequencing more accessible and cost-effective. Concurrently, extensive research efforts have been dedicated to unraveling the structure, properties, and functions of these sequenced proteins. This has led to the development of comprehensive protein databases enriched with specific annotations, facilitating a deeper understanding of the relationship between a protein and its characteristics. Uniprot [25], for instance, is recognized as one of the largest gene-protein databases, notable for its expert-reviewed content. Large language models can leverage such databases, like Uniprot, to construct specialized protein language models by building a corpus based on this data. The Protein Data Bank (PDB) [7] is renowned for housing the largest collection of experimentally-determined 3D structures of proteins. Advanced computational methods, such as Alphafold [61], which was previously considered a state-of-the-art method for protein secondary structure prediction, have been trained on sequence-structure data from the PDB.

Despite the diverse types of proteins (like enzymes and antibodies) and the range of annotations (such as secondary structure, solubility, and binding affinities), the task in bioinformatics can be generally formulated as $y = f(x)$. where x represents the input sequences and y denotes the specific annotations. The primary goal in bioinformatics is to develop models f that accurately describe the relationship between x and y. These models serve multiple purposes, including gaining insights into the intrinsic mechanisms of proteins and predicting annotations for unseen data.

Significant progress has been made in developing computational approaches for predicting protein functions. These methods can be roughly seen as two parts: protein sequence representation and label prediction. The early methods usually use handcrafted features for sequence representation and machine learning methods for label prediction, while the more recent methods use deep learning for both representation learning and label prediction.

2.2.1 Handcrafted feature-based methods

Handcrafted features in protein sequence analysis are specific attributes or characteristics extracted from protein sequences to aid in various bioinformatics tasks like regression, classification, and clustering. These features are designed to capture important biological properties of proteins. Some of the most famous and widely used handcrafted features in protein sequence analysis include:

- Sequence alignment: Protein sequence alignment is a fundamental tool in molecular biology, used to predict the function of proteins. This process involves comparing the amino acid sequences of proteins to identify regions of similarity that may indicate functional, structural, or evolutionary relationships between them. Well known alignment tools includes BLAST (Basic Local Alignment Search Tool) [3] and Clustal Omega [122]. By aligning protein sequences, scientists can identify these conserved domains, providing clues about the protein's function.

- Sequence Motifs and Patterns: Specific short sequences known as motifs, which are often associated with particular functions or structural properties, can be used as features. These motifs are identified through sequence pattern analysis.

- Biochemical and Biophysical Properties: Features based on biochemical properties (e.g., isoelectric point, aromaticity) and biophysical properties (e.g., solvent accessibility) of the protein.

- Position-Specific Scoring Matrix (PSSM): PSSMs [60], generated from multiple sequence alignments, provide a score for each amino acid at each position in the sequence, reflecting its conservation and likelihood of substitution. This is particularly useful for capturing evolutionary information.

- BLOSUM [57] and PAM Matrices: These matrices are used to score substitutions between different amino acids and can be used to generate features that reflect the evolutionary changes in a protein sequence.

With handcrafted features, sequences are transformed into numerical values. These numerical representations allow machine learning models to be trained with corresponding labels for various tasks.

For tasks involving classification and regression, several common machine learning methods are utilized. For classification tasks that involves predicting the category or class of an object, popular algorithms include logistic regression, K-Nearest Neighbors (K-NN),

decision trees, and Support Vector Machines (SVM). In the context of regression tasks that predicts a continuous quantity, commonly employed methods are linear regression, Support Vector Regression (SVR).Certain algorithms can be applied to both classification and regression tasks like Decision trees and random forests. For clustering tasks, which aim to group data points based on similarity measures without prior labeling, common methods include K-means and Hierarchical Clustering. These algorithms are instrumental in uncovering the inherent structure within data sets.

2.2.2 Neural network-based methods

With the advent and evolution of neural networks, deep learning-based methods have demonstrated superior performance in interpreting the representation of sequences. These methods leverage the power of neural networks to automatically learn complex features from protein sequences, bypassing the need for manual feature extraction. As a result, deep learning approaches have increasingly become the state-of-the-art in many protein prediction tasks, outperforming traditional machine learning techniques. Neural network-based sequence models have become increasingly popular and powerful tools in bioinformatics. These models are designed to capture complex patterns and dependencies from protein sequences.

- Recurrent Neural Networks (RNNs): RNNs are designed to process sequences by having a loop within them, allowing information to persist. In the context of sequences, RNNs can use their internal state (memory) to process variable length sequences of inputs. However, standard RNNs often face challenges with long sequences due to problems like vanishing or exploding gradients. Meanwhile, there are special kinds RNN include Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs).

- Convolutional Neural Networks (CNNs) for Sequences: While CNNs are predominantly known for image processing, they can also be applied to sequence data. The convolutional layers can identify and learn patterns or motifs within sequences, making them useful for tasks like sequence classification or feature extraction.

- Transformer Models: Originally developed for natural language processing tasks, transformers have revolutionized the field due to their effectiveness in handling long-range dependencies. They rely on self-attention mechanisms to weigh the influence of different parts of the input data. The self-attention module in transformer is:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

where Q, K, and V represent queries, keys, and values. d_k is the first dimension of matrix K. The multi-head self-attention is computed as

$$MultiHead\left(Q, K, V\right) = Concat\left(h_1, \ldots, h_h\right) W^O \tag{2.3}$$

where $h_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$.

2.3 Protein sequence-based spectrum prediction

In developing database search methods, the accuracy largely hinges on the design of the scoring function. This function assesses the compatibility of a peptide with an experimental spectrum. A proficient scoring function, capable of effectively distinguishing between target and decoy peptides, is essential for identifying peptides with high confidence in PSMs. Traditional database search strategies primarily concentrate on the analysis of predetermined ion series types, such as y- and b-type ions, the masses of which can be directly deduced from the peptide sequence. The development of the scoring function involves calculating the discrepancies between the theoretical and experimental spectra.

Recent advancements in machine learning and deep learning have introduced methods that enhance the accuracy of theoretical spectrum predictions. These advanced techniques go beyond mere mass prediction of ions; they are capable of forecasting the intensity of spectral peaks. Furthermore, some approaches are designed to predict retention times in liquid chromatography (LC), thereby refining the precision of the scoring function. This evolution in methodology not only improves the accuracy of peptide identification but also enriches the overall performance of database search processes.

2.3.1 Peak intensities prediction

The prediction of peak intensities takes the peptide sequence as the input. Aside from the sequence, there are some metadata including charge, Normalized Collision Energy (NCE) and instrument. Peak intensities in mass spectrometry can be predicted through two primary approaches: focusing on pre-defined ion series types or predicting the entire spectrum without relying on ion series annotations.

For the prediction of annotated ions, conventional machine learning methods have been developed, with MS2PIP [32] being a notable example. MS2PIP employs a random forest regression approach to predict the intensities of specific ion types. On the other hand,

advancements in deep learning have led to the development of several innovative methods for this task, including wiNNer [128], pDeep [163] [157], Prosit [51], DeepMass:Prism [128], AlphaPeptDeep [156]. These methods differ in their underlying neural network architectures: wiNNer, pDeep, and DeepMass:Prism utilize LSTM networks; Prosit is built upon a GRU network; and AlphaPeptDeep incorporates Transformer layers, highlighting the diversity in approaches to tackle the challenge of intensity prediction.

When it comes to predicting the full spectrum, PredFull [82] stands out as a unique method employing a Convolutional Neural Network (CNN)-based architecture. This approach models the entire mass spectrum by segmenting the m/z range up to 2,000 Da into bins of 0.1 Da, resulting in a 20,000-dimensional vector to represent the target intensities.

2.3.2 Retention time prediction

The process of predicting retention times (RT) in liquid chromatography utilizes the same input data as intensity prediction, but the output is a single numerical value representing the RT. Due to this similarity in input data, the architectures developed for intensity predictions can be adapted for RT prediction. For instance, Prosit [51], originally designed for intensity prediction, can also be employed for RT prediction, illustrating the versatility of these models.

Additionally, several methods have been specifically devised for the prediction of retention times. DeepLC [13] employs a CNN as its core architecture, leveraging the spatial pattern recognition capabilities of CNNs to predict RT. Conversely, DeepDIA [152] and autoRT [145] utilize a hybrid approach that combines LSTM networks and CNNs. This combination harnesses LSTM's ability to process sequential data and CNN's proficiency in handling spatial features, making these models particularly effective for RT prediction.

2.3.3 PSM rescoring

Database search methods, tailored for large-scale data analysis, aim to strike a balance between speed and precision. To enhance the reliability of peptide identifications, post-processing techniques have been developed to re-rank PSMs. A pioneering tool in this domain is Percolator [63], a semi-supervised learning method that significantly improves the accuracy of PSM identification. Percolator analyzes the output from database search results, computing 20 distinct features for each PSM. It employs half of these PSMs to train a Support Vector Machine (SVM) model. This model is designed to distinguish between

target PSMs and decoy PSMs, subsequently applying the learned distinctions to re-rank all PSMs in the dataset.

Building upon the foundation laid by Percolator, recent years have seen the introduction of several advanced methods, including DeepRescore [79], MS2Rescore [31], AlphaPept-Deep [156], and MSBooster[151]. These innovative approaches extend the capabilities of traditional post-processing by incorporating additional features derived from trained models for intensity prediction and retention time prediction. By integrating these new dimensions of data, these methods further refine the re-ranking process, enhancing the overall precision of peptide identification in proteomic research.

2.4 Pre-train and fine-tune on protein language model

2.4.1 Pre-trained protein language model

In recent years, the field of Natural Language Processing (NLP) has experienced significant advancements, primarily driven by the development of large pre-trained models. These models, known as pre-trained language models (LMs), are built upon extensive datasets comprising long sequences of text data. Central to the concept of pre-training are two main paradigms: causal language modeling (CLM) and masked language modeling (MLM), exemplified by the Generative Pre-trained Transformer (GPT) [110] and Bidirectional Encoder Representations from Transformers (BERT) [67] models, respectively. A crucial distinction between CLM and MLM lies in their predictive focus: MLM is designed to predict only the masked words within a sequence, whereas CLM aims to predict the next word based on the preceding context. This difference necessitates a masking schema for MLM, where tokens are randomly selected and masked from the input sequence, typically replaced with a special token (e.g., [MASK] in BERT). In the case of BERT's pre-training, the sequence length is set at 512 tokens with a masking rate of 15%, meaning that 15% of the tokens are randomly masked for the model to predict, given the unmasked parts of the sequence.

The evolution of CLM from RNNs to Transformer-based architectures marks a significant transition in the field of NLP. Initially, RNNs were the go-to architecture for modeling the probability of a word given the previous words in a sequence. However, the introduction of the Transformer architecture has shifted the paradigm. In the domain of CLM, the Transformer decoder architecture has become the preferred choice, with GPT standing out as the pioneering model for predicting the next token in a sequence. This shift

is attributed to the Transformer's superior handling of long-range dependencies and its efficiency in training over large datasets. Conversely, for MLM, the Transformer encoder architecture has taken precedence. BERT emerged as the trailblazer in employing a bidirectional Transformer encoder for pre-training an MLM. In this framework, the entire input sequence is encoded, and the prediction of masked tokens is performed by the encoder's final layer. This bidirectional approach allows BERT to understand the context from both directions, enhancing its ability to predict the masked words accurately. Following BERT, several variants have expanded on the concept of MLM by introducing additional complexities. For example, T5 [112] employs both a bidirectional Transformer encoder and a unidirectional Transformer decoder, using a span masking schema for a more sophisticated prediction of masked tokens. Similarly, BART [78] utilizes an encoder-decoder model but is distinct in its training with a denoising objective, aiming to reconstruct the original text from a corrupted version. It's important to note that these LMs does not represent a standard NLP task with set datasets and evaluation metrics; instead, it's an unsupervised pre-training task utilizing vast text corpora. An example of such a corpus is WikiText-103 [94], which contains 28,475 articles and over 103 million tokens, serving as a foundational dataset for training models like BERT.

Transformative advancements in natural language processing (NLP) have revolutionized the field. By learning from vast collections of text data, large pre-trained models have significantly enhanced the ability of machines to understand and generate human language. Inspired by these achievements in NLP, the concept of large pre-trained models has been extended to the bioinformatics. By replacing human language with protein sequences, these models, known as protein language models (pLMs), provide novel insights and capabilities in the comprehension and generation of proteins. Trained on extensive datasets of protein sequences, pLMs learn comprehensive representations of protein structure, functions, and interactions. When utilized for feature generation, these models can be fine-tuned for a wide range of specific bioinformatics applications.

Generally, pLMs can be categorized into three types based on their architectural design: encoder-only, decoder-only, and encoder-decoder models. Each type is uniquely suited to different applications in protein research.

Encoder only model

Encoder-only pLMs encode protein sequences and structures into fixed-length vector embeddings. Among the forefront of these advancements are several prominent pre-trained protein sequence encoders, including ESM-1b [117], ESM-1v [93], and ESM-2 [81], ProteinBert [14], ProtTrans [40]. These models leverage the power of Transformer encoder

architectures, akin to those found in BERT [67] and RoBERTa [83], to predict protein structure and function by harnessing the vast sequence information available in protein databases, eliminating the need for manual sequence annotations.

The ESM series focus on utilizing the Transformer's encoder to analyze extensive protein sequence data, aiming to uncover intricate patterns related to protein structure and functionality. ProteinBert [14] introduces an innovative approach to the BERT architecture by incorporating a novel pretraining task specifically designed for protein functionality prediction. This model distinguishes between local (character-level) and global (sequence-level) representations, facilitating multitask learning in a structured manner. ProtTrans [40] takes a different approach by training several auto-encoder models on a vast corpus of sequence data, highlighting the diversity of methods aimed at improving encoder architectures.

The role of Multiple Sequence Alignment (MSA) extends beyond traditional sequence analysis, serving as a computational method to expose common features and variation patterns among sequences. By aligning multiple sequences, MSAs can reveal shared evolutionary relationships, aiding in the identification of functional regions and structural domains. This technique has seen widespread application in protein modeling, exemplified by the MSA Transformer [115]. This model adapts the self-attention mechanism for MSAs, interleaving attention across rows and columns to capture dependencies both within amino acid sequences and across different sequences. Notably, the MSA Transformer has been integrated into the groundbreaking AlphaFold2 [61], underscoring its significance.

For some specific types of protein, for example, antibody protein sequences, its correponding language model is developed to tackle domain-specific downstream challenges. Several studies have trained models either fine-tuned from protein language models (including AntiBERTa[77], AntiBERTy[118]) or directly train on antibody datasets (AbLang[100]). These models have been designed to capitalize on the distinct characteristics and features exhibited by antibody sequences.

Decoder only model

Inspired by the GPT architecture, decoder-based pLMs have emerged as pivotal tools in the novel proteins generation, protein engineering and drug design. These models, through their capacity for autoregressive sequence generation, have opened new avenues for the controlled synthesis of protein sequences, offering promising prospects for protein design and therapeutic development.

A notable example of this approach is ProGen [88], which leverages the GPT architecture for controllable protein generation. Trained on an extensive dataset of 280 million protein sequences, ProGen incorporates conditioning tags to incorporate a wide range of annotations, including taxonomic, functional, and locational information. This allows for the generation of protein sequences that are not only novel but also tailored to specific biological contexts and functions. Building on this foundation, ProGen2 [97] represents a significant advancement, expanding the model to 6.4 billion parameters. It benefits from an even more diverse training dataset, extracted from over one billion proteins across genomic, metagenomic, and immune repertoire databases. This extensive training enables ProGen2 to produce highly varied and functionally relevant protein sequences. ProtGPT2 [46], another GPT-based model, further demonstrates the potential of decoder-based pLMs in protein sequence generation. It is specifically designed to generate protein sequences with amino acid compositions and disorder propensities that mirror those found in natural proteins, showcasing the model's ability to replicate complex biological characteristics.

The application of decoder-based protein language models extends to targeted protein design. For example, PoET [131] focuses on the distribution over protein families, enabling the generation of sets of related proteins with specified characteristics. IgLM [120] employs autoregressive sequence generation techniques for the specific purpose of antibody design, illustrating the model's utility in creating highly specialized protein sequences.

Encoder-decoder model

The encoder-decoder architecture was originally designed for translation tasks, excelling in scenarios that necessitate the transformation between two distinct data types. This makes it particularly well-suited for bioinformatics applications that require translating between different types of biological sequences or structures. This architecture's versatility underscores its utility in bridging various biological data domains. Inspired by the T5 architecture [112], ProtT5 [40] is a protein language model specifically developed to bridge the gap between protein sequences and their properties. In this model, the encoder processes input protein sequences, translating them into a high-dimensional representation that encapsulates the sequences' inherent properties and patterns. The decoder, in turn, utilizes this encoded data to make predictions regarding the protein's structure, function, or other pertinent features. This approach enables ProtT5 to not only learn sequence motifs effectively through the encoder but also to extrapolate protein functions via the decoder, thus providing deeper insights into the its protein representations.

2.4.2 Fine-tune techniques on pre-trained model

Fine-tuning is a crucial technique in leveraging pre-trained models, particularly for developing domain-specific models. There are three main methods of fine-tuning for task prediction: linear probing, full fine-tuning, and parameter-efficient fine-tuning (PEFT).

Linear probing involves using the output of a pre-trained model as a representation at the amino acid or protein level. This is done by freezing the bakcbone model and replacing the last layer with a new output model. During training, the parameters of the pre-trained model remain unchanged. This approach is comparatively faster as features for each sequence only need to be computed once before training. Additionally, since the pre-trained part is frozen, there's less risk of the model overfitting specific tasks and experiencing catastrophic forgetting.

Full fine-tuning involves replacing the last layer with a task-specific model architecture and applying transfer learning on the whole model. Its advantage lies in the adaptability of the representation during training, which can lead to better performance. However this approach is generally slower and more memory-intensive, as it involves training the large model.

PEFT emerges as a solution to the limitations of the first two methods: the challenge in achieving good performance with feature embedding and the high computational resources and risk of catastrophic forgetting with full fine-tuning. PEFT is a more balanced approach that aims to fine-tune models efficiently. It involves updating only a small subset parameters within the model or a small amount of extra parameters, thereby reducing computational requirements while still allowing for model adaptation and learning. This makes PEFT a viable option for tasks requiring the flexibility of fine-tuning without the extensive resource demands of full model training.

Chapter 3

Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry

3.1 Introduction

In adaptive immunity, the major histocompatibility complex (MHC) presents a class of short peptides (also known as MHC peptides or HLA peptides) on the cell surface for T-cell surveillance. The systematic study of the MHC peptides is also referred to as immunopeptidomics. Two major classes of MHC exist: MHC-I and MHC-II. MHC-I molecules are expressed on all nucleated cells and MHC-II molecules are expressed on antigen-presenting cells. Normally, peptides presented by MHC-I are derived from endogenous proteins which are neglected by the cytotoxic T lymphocytes. However, the MHC-I of infected or tumor cells may present exogenous or mutated peptides (neoantigens) derived from either the viral proteome or cancer-related mutations, leading to the activation of specific cytotoxic T lymphocytes to eliminate the neoantigen-presenting cells. These abnormal MHC-I peptides also serve as excellent targets for immunotherapy, such as TCR-T [55] and cancer vaccines [104]. For these reasons, a method that can systematically determine all the MHC-I peptides becomes extremely useful for studying infectious diseases, developing novel cancer immunotherapy, and choosing the right immunotherapy for individual patients. Currently there are two main approaches to identifying abnormal MHC-I peptides: genomics and proteogenomics. In genomics, the identification of somatic mutations on neoepitopes is often performed using whole exome sequencing (WES) or transcriptome

sequencing data [65]. On the other hand, the proteogenomics approach involves analyzing tissue samples using liquid chromatography tandem mass spectrometry (LC-MS/MS) and searching the obtained spectra against a personalized protein database constructed from exome sequencing or RNA sequencing data [5]. However, DNA- or RNA-based methods often require predicting which mutations will generate neoantigens. Protein sequencing allows direct identification of actual peptides produced by mutations, reducing uncertainty based on predictions. Besides, since there's no evidence or statistical validation, the false positives of both methods can be high.

Proteomic approaches that rely exclusively on MS spectra for peptide identification provide direct experimental validation, which helps reduce false positives and enables statistical verification of the False Discovery Rate (FDR). Traditional database search methods, such as Peaks [159], Comet [41], and MaxQuant [28], are prevalently employed for identifying peptides, particularly those generated from tryptic digestion. Tryptic digestion involves the cleavage of proteins at the C-terminal ends of lysine (K) and arginine (R) residues. However, these methods often fall short in analyzing complex MHC peptides, which undergo non-tryptic digestion. Non-tryptic digestion results in protein cleavage at various sites, not limited to K and R residues, significantly expanding the search space and complicating the identification process.

To enhance the identification rates for non-tryptic peptides, current efforts focus on rescoring the database search results. Percolator [127] is a widely used tool that employs an SVM-based semi-supervised machine learning approach to rescore Peptide-Spectrum Matches (PSMs), thereby enhancing sensitivity in peptide identification. MHCquant [9] integrates Percolator into an immunopeptidomics data analysis workflow for MHC peptide identification, benefiting from improved sensitivity and accuracy. Recent advances in peptide property prediction, such as retention time (RT), MS/MS spectrum, and collisional cross sections, have enabled innovative workflows like DeepRescore [79], Prosit [146], MS2Rescore [31], and AlphaPeptDeep [156]. These workflows utilize prediction tools to generate PSM features and leverage Percolator for re-rank the identified PSMs.

Much experiments have shown these methods help improve the identification rate, but a significant limitation of these tools is their inability to identify mutated peptides not present in the sequence database, which are the key of finding targets for immunotherapy. This hinders their applicability in detecting novel peptides with mutations. Current sequencing methods that can identify mutated peptides include *de novo* sequencing and open-search sequencing. *De novo* sequencing methods, for example Deepnovo [129] [109] can identify novel peptides as they do not rely on a reference database. However, since it completely has no constraint from the reference, it suffer from a higher error rate and the lack of a universally accepted result validation method, which undermines confidence in

the identified sequences. Open-search methods such as MSFragger [70], Open-pFind [22], TagGraph [35], and PROMISE [62] provide tag-based techniques for identifying peptides with unexpected post-translational modifications (PTMs), semi- and nonspecific digestion, in-source fragmentation, and cofragmentation of coeluting peptides. Still, they do not support to find peptide with mutations.

To confidently identify MHC peptides, we have developed a novel workflow called NeoMS that combines *de novo* sequencing and PSM rescoring techniques. NeoMS generates a candidate neoepitope database using *de novo* sequencing and k-mer tagging. To enhance peptide identification, we employ a trained lightGBM model instead of using semi-supervised learning to avoid overfitting. To ensure the accuracy of identified mutated peptides, NeoMS implements rigorous FDR control measures. The performance of NeoMS is evaluated using publicly available mass spectrometry data for MHC peptides. In comparison to other tested methods, NeoMS outperforms them by identifying a greater number of peptides. Importantly, NeoMS confidently identifies hundreds of mutated MHC-I peptides, providing high-confidence results. By combining *de novo* sequencing, PSM rescoring, and stringent FDR control, NeoMS offers a powerful and accurate approach for MHC peptide identification, including the detection of mutated peptides. This workflow has the potential to advance research in immunopeptidomics and facilitate the development of personalized cancer immunotherapy.

3.2 Methods

The input of the workflow is the peptide MS/MS spectra and a reference protein sequence database. The raw MS file is converted to the mgf format file using msConvert [1] before the analysis. There are four main steps of the analysis:

1. Expanded database generation: Generate an expanded sequence database that consists of both the original sequences and possibly mutated sequences. To generate the mutated sequences, *de novo* sequence is used to locate the possible point mutations.

2. Database search: Conduct database search analysis by using the input spectra and the expanded sequence database.

3. PSM Rescoring: Rescore the PSMs found by the database search with newly computed scoring features and a machine-learning scoring function.

4. Result analysis: Control the FDR of the identified regular and mutated peptides.

These steps are further elaborated on Figure 3.1. In the following sections, we describe the details of each step. In the implementation of NeoMS, each module is dockerized, and the whole workflow is compiled in Nextflow [36]. The code is available on GitHub (https://github.com/waterlooms/NeoMS).

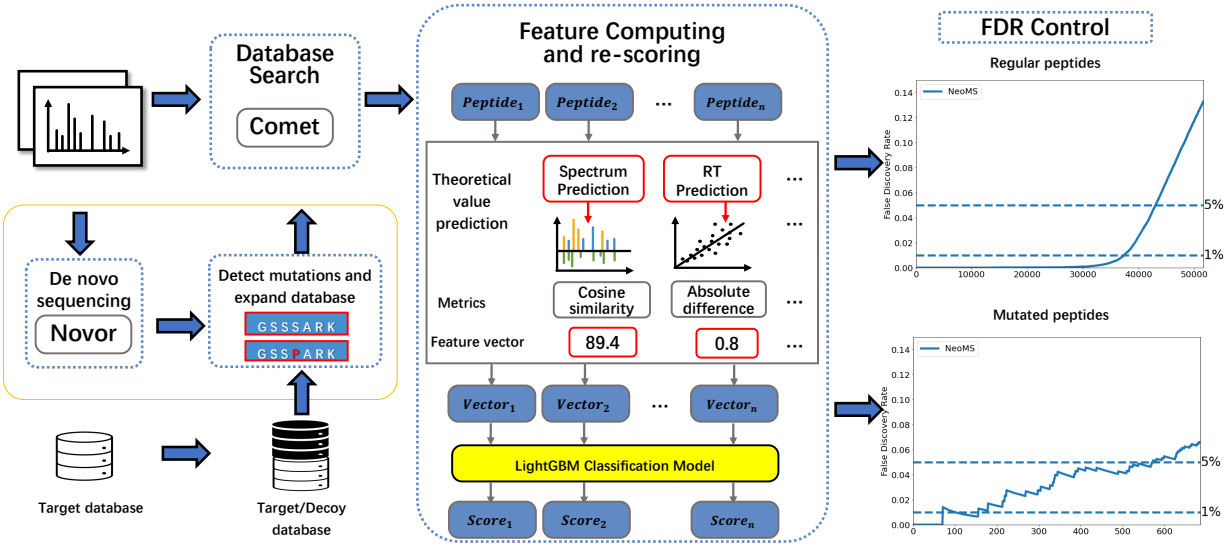

Figure 3.1: The overall workflow of NeoMS.

3.2.1 Generation of expanded database

The MS/MS spectra are *de novo* sequenced by Novor [86]. Novor is machine learning based *de novo* sequencing method. For each spectrum, Novor computes a peptide sequence and a positional confidence score for each amino acid of the peptide.

Within the *de novo* sequences, a confident sequence tag is defined as a length-k substring (continuous subsequence) where each amino acid has a confidence score above a threshold t. By default, NeoMS sets $k = 7$ and $t = 60$, which were selected empirically. These tags are searched in a target/decoy database to find approximate matches. This target sequence database is human protein database downloaded from uniprot [26]. Considering the non-tryptic manner of searching, the decoy is generated by random shuffling target protein sequences. Hence, the target/decoy database is the concatenation of these two database.

A hit to a tag is a length-k substring in the database that approximately matches the tag with exactly one amino acid mutation. For each hit, a mutated sequence is constructed by concatenating the at-most n amino acids immediately before the hit, the mutated amino acids, and the at-most n amino acids immediately after the hit. By default, NeoMS sets $n = 12$. Since MHC-I peptides have lengths up to 13, this choice of n allows the inclusion of every mutated MHC-I peptide that has a confident *de novo* tag covering the point mutation.

The newly generated mutated sequences are appended to the target/decoy database to form an expanded database. Note that a hit can be either from a target or a decoy sequence. The target and decoy hits are treated equally throughout the analysis until the FDR control step. In the end, the resulting expanded database contains the target sequences and the decoy sequences, as well as the mutated sequences generated from the target and decoy sequences. The process of generating personalized database is shown in Figure 3.2.

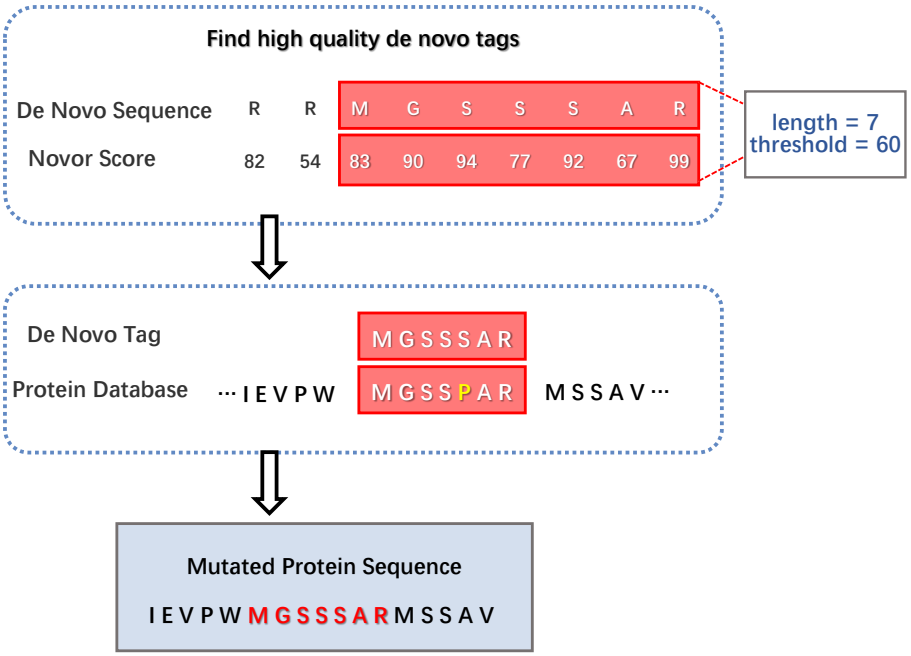

Figure 3.2: An example of personalized database generation. In this example, a high quality tag of length 7 *MGSSSAR* is generated. This *de novo* tag is matched to a subsequence *MGSSPAR* in the protein database with one amino acid mutation. This protein is mutated an extended to a new protein sequence and appended to the original database.

3.2.2 Database search in the expanded database

Comet [41] is used for the database search analysis due to its ability to search multiple candidate peptides for each spectrum. By using the input MS/MS spectra and the expanded sequence database, it finds the best k peptides for each spectrum. In our workflow, We used the default Comet high-high parameter and made three adjustments:

1. Set *num_output_lines* to 10. Up to 10 candidate peptides are computed for each spectrum. These candidate peptides are further evaluated in the downstream rescoring analysis to choose the optimal one for each spectrum. For a certain spectrum, some true peptides with lower Comet scores could have higher rankings after rescoring. Meanwhile, the statistical relations of peptides for one spectrum produce important features for rescoring. It is noticed that in practice, any other database search tool that allows the output of multiple candidates per spectrum can be used in lieu of Comet as the base engine of NeoMS.

2. Set enzyme to *cut_everywhere*. By default, Comet searches peptides in tryptic digestion that only cut cleaves the C-terminal to lysine (K) and arginine (R). We set it as *cut_everywhere* to search in a non-tryptic manner.

3. Set *mass_tolerance* to 0.02 Dalton. To narrow down the search space and filters out the incorrect PSMs.

3.2.3 Rescoring

For each spectrum S, the top 10 peptide candidates $P_1, P_2, \ldots, P_n$. Five comet computed values: *xcorr*, *delta_cn*, *sp_score*, *mass_error*, and *e_value* are taken as features. Here the *e_value* score is converted to $log(e\_value)$ before using it in machine learning. Besides, the following set of peptide features is computed for each P_i:

1. The absolute difference between predicted RT for $P_i(RT_{predicted})$ and the experimental RT of the spectrum $S(RT_experimental)$ can be denoted as $RT-ABS$. In our NeoMS, AutoRT [145] was used to make the prediction. It can be formulated as:

$$\text{RT\_ABS}(RT_{predicted}, RT_{experimental}) = |RT_{predicted} - RT_{experimental}| \tag{3.1}$$

2. The similarity between S and the MS/MS spectrum predicted for P by pDeep2 [157]. The feature is computed by the Pearson correlation coefficient between the predicted

b and y-ion intensities and their experimental intensities, denoted as PCC. The b- and y-ions of spectrum P and S are denoted as P_i and S_i, respectively. PCC can be formulated as:

$$\text{PCC}(P, S) = \frac{\sum(P_i - \bar{P})(S_i - \bar{S})}{\sqrt{\sum(P_i - \bar{P})^2 \sum(S_i - \bar{S})^2}} \tag{3.2}$$

3. The similarity between S and the MS/MS spectrum predicted for P by PredFull [82]. PredFull predicts a sparse vector of length 20,000, where each dimension represents the maximum peak intensity in an m/z bin of width 0.1 mass units. The experimental spectrum is also converted to such a sparse vector. The intensity of each bin in P and S are denoted as P_i and S_i. The feature is the Cosine Similarity of two vectors, denoted as CS, can be formulated as :

$$\text{CS}(P, S) = \frac{P \cdot S}{\|P\|\|S\|} = \frac{\sum P_i S_i}{\sqrt{\sum P_i^2}\sqrt{\sum S_i^2}} \tag{3.3}$$

This list comprises eight peptide features. In addition, for the $log(e\_value)$ and the 3 peptide features above, three spectrum features are computed: maximum, mean, and variance of the feature's values on the top 10 peptides for the spectrum. In total, there are $4 \times 3 = 12$ features, consisting of 8 peptide features and 12 spectrum features. A machine learning model based on LightGBM is used to calculate a numeric score based on the 20 features.

3.2.4 FDR control

The target-decoy approach [39] is adapted for controlling the FDR. The target and decoy databases are combined and analyzed together. After the NeoMS search, the regular and mutated peptides are separated, and their FDRs are also controlled separately. The score thresholds of the regular and mutated peptides are usually different because of their different distributions. In addition, a user can choose to use different FDR thresholds for the regular and mutated peptides, respectively. In our experiment, due to the higher complexity of identifying mutated peptides compared to regular peptides, we established distinct thresholds for each category: 1% for regular peptides and 5% for mutated peptides.

3.2.5 Training of the scoring function

The training of the scoring function is conducted as a distinct phase, separate from the scoring process itself. Unlike Percolator's semi-supervised learning method, our approach

involves supervised learning, which eliminates the risk of overfitting by avoiding training on test data. The training is carried out only once using a designated set of training data and is kept constant for subsequent analyses. We utilize the LightGBM [66] package for the training process, with specific parameters set: *max_depth* to 9 and *num_leaves* to 51.

The training proceeds iteratively across multiple iterations to enhance the model's performance optimally. Before initiating our training, we lacked precise labels for positive and negative PSMs. In the initial iteration, target peptides identified by Comet's search results at a 1% FDR formed the preliminary positive set, while an equivalent quantity of top-ranked decoy peptides established the initial negative set. LightGBM was employed to construct the initial Gradient Boosting Machine (GBM) model. In each subsequent iteration, the GBM model from the previous round was applied to perform the search. Subsequently, target peptides identified at 1% FDR were added to the existing positive set after removing duplicates, and a similar quantity of top-ranked decoy peptides were added to the ongoing negative set, also with deduplication. The model was then updated using the expanded sets of positive and negative instances. This iterative process was repeated several times until there was no further improvement in performance.

3.3 Results

3.3.1 Datasets

Two LC-MS/MS datasets of human HLA I peptides: PXD000394 and PXD004894 were downloaded from the proteomeXchange [34] repository. The first dataset is used for training our lightGBM model and the second dataset is used for testing. The two datasets are derived from separate experiments, resulting in no overlap between them.

Pride PXD000394 was acquired from a Thermo Q-Exactive instrument and contained 41 MS raw files [6]. It is a collection of six cell lines: JY, SupB15WT, HCC1143, HCC1937, Fib, and HCT116. All 41 MS raw files (12.6 million MS/MS spectra) were used for training our machine learning model. The detail of this dataset is in Table 3.1.

Pride PXD004894 was acquired from a Thermo Q-Exactive HF instrument [5]. This is a survey conducted on tissue samples associated with melanoma-associated tumors. A total of 25 melanoma patients were included in the study, and we specifically focused on three patients: Mel5, Mel8, and Mel15. To test our software, we utilized 24 MS raw files, which consisted of 1.12 million MS/MS spectra associated with these three patients. Among these files, 16 were associated with Mel15, containing 720,557 spectra. There were 4 raw

Cell line	Tissue origin	Number of raw files	Total number of spectra
JY	B-cells EBV transformed	5	164,844
SupB15	B-cell leukemia	8	257,775
HCC1143	Basal like breast cancer	4	116,148
HCC1937	Basal like breast cancer	5	161,211
HCT116	Colon carcinoma	6	203,290
Fibroblast	Primary fibroblast cells	13	546,519

Table 3.1: Information about the six cell lines for dataset PXD000394.

files associated with Mel5, comprising 118,003 spectra, and another 4 raw files associated with Mel8, comprising 132,297 spectra. The detail of this dataset is in Table 3.2.

Patient	Number of raw files	Total number of spectra
Mel-15	16	720,557
Mel-5	4	118,003
Mel-8	4	132,297

Table 3.2: Information about the three patients' datasets selected for this study from PXD004894.

3.3.2 Regular peptide identification

The performance of NeoMS was compared against four other database search methods: Comet [41], MaxQuant [28], PeaksX [159], and DeepRescore [79]. The datasets from three patients (Mel 5, Mel 8, and Mel 15) within PXD004894 and the human sequence database (UniProt UPID: UP000005640) were utilized to evaluate performance. For the Comet and DeepRescore analyses, we executed the software using our test dataset. Specifically, Comet was run with the default parameter settings, albeit with three modifications previously mentioned. DeepRescore was configured to utilize Comet's default parameters. For PeaksX and MaxQuant, we leveraged results from prior studies: PeaksX's database search results provided from the database searching part of an individual immunopeptidomes [129] on the same datasets were used. MaxQuant's results provided by dataset paper [5] were used. As the other software does not search for mutated peptides, the mutation finding function was turned off in NeoMS here for a fair comparison. The number of PSMs and the number of peptides identified at 1% FDR by each software are plotted in Figure 3.3 and Figure 3.4, respectively.

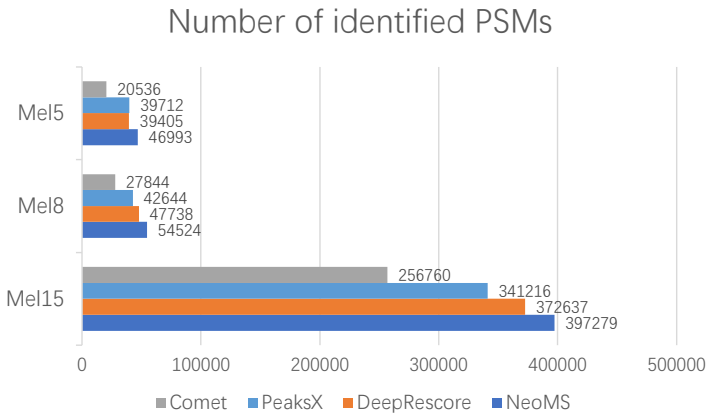

Figure 3.3: Comparing the number of identified MHC-I PSMs without mutations of three patients' samples (Mel5, Mel8, Mel15). X-axis indicates the number of PSM identifications at 1% FDR. NeoMS is compared with three methods: Comet, PeaksX, and DeepRescore.

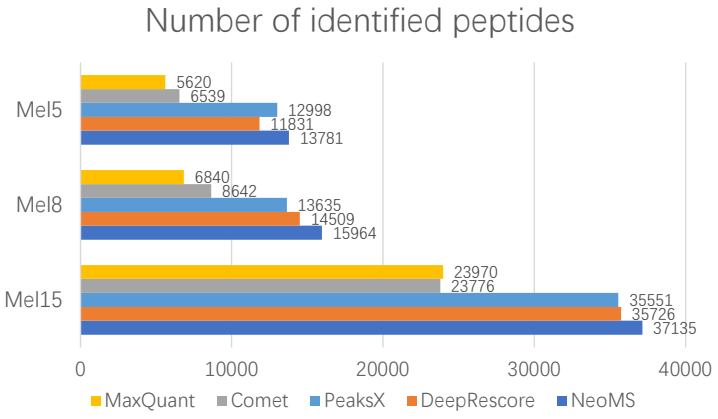

Figure 3.4: Comparing the number of identified MHC-I peptides without mutations of three patients' samples (Mel5, Mel8, Mel15). X-axis indicates the number of unique peptide identifications at 1% FDR. NeoMS is compared with four methods: Comet, MaxQuant, PeaksX, and DeepRescore.

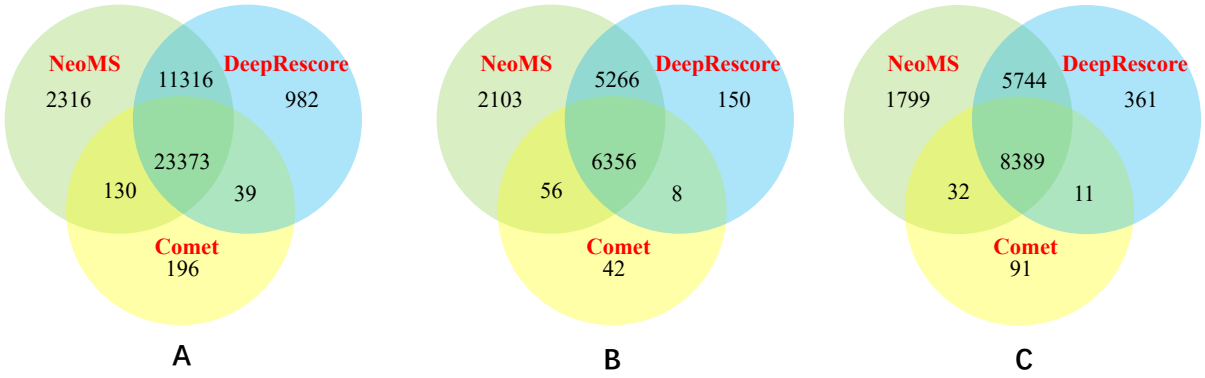

Figure 3.5: Venn diagram of the unique peptides identified at 1% FDR on Mel 15(a), Mel 5(b), and Mel 8(c) by the three search engines: NeoMS, DeepRescore, and Comet, respectively. The number in each area indicates the number of identified peptides.

From the results, it is evident that the methods designed for identifying MHC-I peptides—namely DeepRescore, PeaksX, and NeoMS—significantly enhance the identification rate compared to the baseline method. Specifically, for the three samples, our NeoMS method has improved the identification count by two to three times compared to our base method, Comet. Regarding the number of peptides, PeaksX and DeepRescore exhibit comparable performances: PeaksX identifies more peptides in the Mel 5 samples, while DeepRescore is more effective in identifying peptides in the Mel 8 and Mel 15 samples. Nevertheless, across these three samples, NeoMS surpasses all other methods in terms of both peptide and PSM numbers. When compared to other rescoring methods, NeoMS achieves a 5% to 10% higher identification rate than the second-best methods.

We compare the peptides identified by NeoMS, DeepRescore, and Comet across three patient datasets. The Venn diagram illustrating the overlap of identified peptides by the three methods is presented in Figure 3.5. Across all datasets, NeoMS identifies the highest number of peptides and uniquely discovers novel peptides not detected by the other methods. Specifically, NeoMS identifies 2,316, 2,103, and 1,799 novel peptides in the three datasets, respectively.

We next look into the detailed identification results for each raw file, as depicted in Figure 3.6. Due to the absence of individual raw file results in previous studies by PeaksX and MaxQuant, our analysis specifically contrasts our NeoMS method with Comet and DeepRescore. NeoMS consistently outperforms the other methods in identifying a greater number of peptides across all individual raw files: 16 for Mel 15, and four each for Mel 5 and Mel 8. Specifically, in comparison to DeepRescore, NeoMS increases the peptide

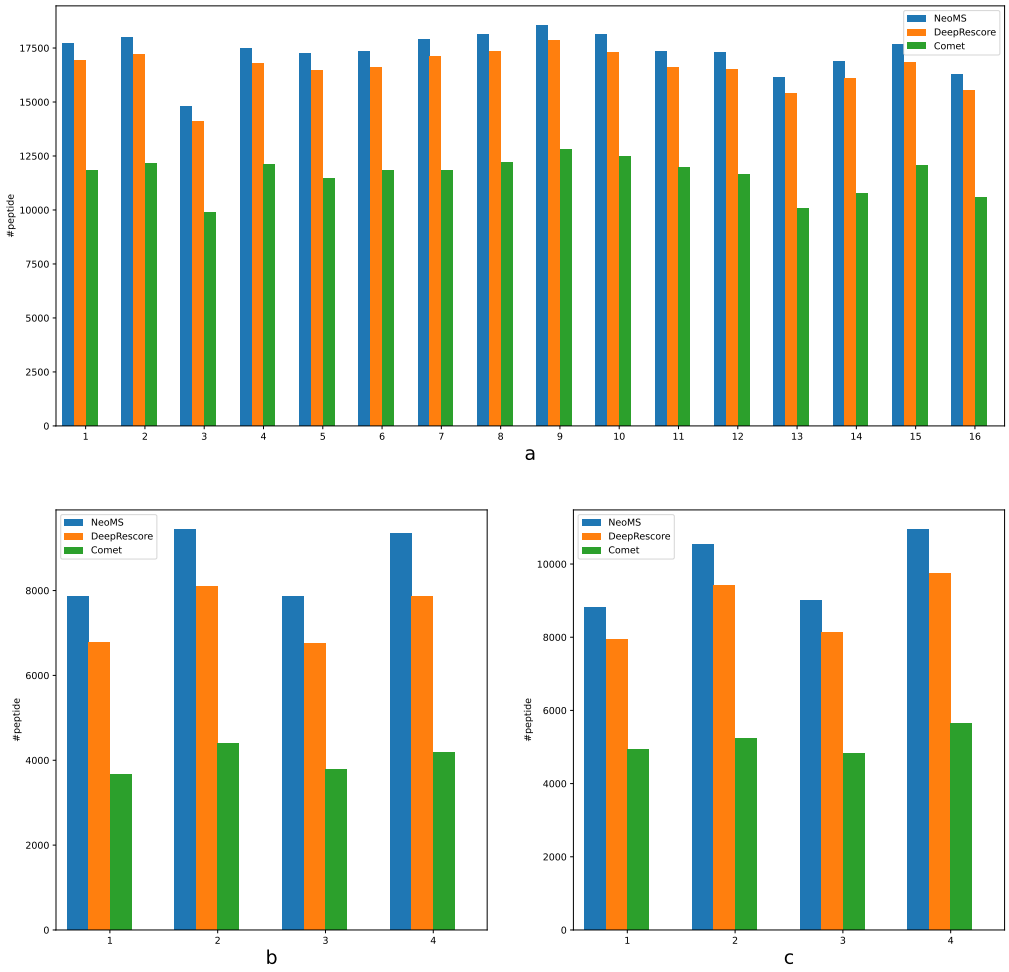

Figure 3.6: The number of identified peptides from NeoMS on each raw files. (a) The number of peptides identified by NeoMS for the 16 raw files in Mel 15. (b) The number of peptides identified by NeoMS for the 4 raw files in Mel 5. (c) The number of peptides identified by NeoMS for the 4 raw files in Mel 8.

identification numbers by 4.7% for the Mel 15 dataset, and significantly more for Mel 5 and Mel 8 datasets, by 16.8% and 11.4%, respectively. This improvement can be attributed to the underlying mechanisms of DeepRescore, which is based on Percolator, a semi-supervised machine learning method that requires half of the input data for training. The Mel 5 and Mel 8 datasets, with an average of 29,500 and 33,074 spectra per raw file respectively, provide fewer spectra than the Mel 15 dataset, which has 45,035 spectra. This discrepancy results in a less effectively trained machine learning model due to the insufficient number of spectra. In contrast, NeoMS utilizes a supervised learning approach, ensuring robust performance across different input files.

To assess the validity of these identified peptides as MHC-bound, MHCflurry [102], a deep learning-based tool for predicting peptide-MHC-I binding affinity is employed. MHCflurry evaluates the binding affinity for a given peptide sequence and allele, assigning an affinity score and ranking the peptide within the top percentile of binding strength. Following established benchmarks [79], peptides ranked in the top 2% for any allele in the sample are deemed capable of binding to MHC-I. In the Mel 15 dataset, out of 37,135 peptides identified by NeoMS, 33,810 (91.05%) were predicted to bind to MHC-I. In comparison, DeepRescore identified only 32,600 peptides as MHC-bound. Notably, while there is a significant overlap between the peptides identified by NeoMS and DeepRescore, each method also pinpoints a substantial number of unique peptides. Among the 2,316 NeoMS-exclusive peptides, 1,955 (84.41%) were predicted to be MHC-bound, underscoring the efficacy of NeoMS in accurately identifying MHC-associated peptides. In other two dataset, it has similar result: for all the NeoMS identified peptides, 80.28% and 92.03% are bound to MHC, respectively; for the NeoMS-exclusive peptide, 78.32% and 84.41% are bound to MHC, respectively.

3.3.3 Mutated MHC-I peptides identification

NeoMS demonstrates the capability to identify both MHC-I peptides with and without mutations in a unified search. In Figure 3.7 and Figure 3.8, the number of mutated peptides and PSMs identified by NeoMS on the Mel15 dataset (from PXD004894) is presented, respectively. The search was conducted on 16 MS files, comprising a total of 164,844 spectra, using the Uniprot human sequence database. The identification of mutated and unmutated peptides was performed independently, with separate FDR controls. NeoMS identified 37,042 peptides (393,871 PSMs) without any mutations at 1% FDR and 544 mutated peptides (4,028 PSMs) with exactly one amino acid mutation at a 5% FDR. The other two samples, Mel5 and Mel8, yielded the identification of 191 and 86 mutated peptides, respectively, at a 5% FDR. These results highlight NeoMS's ability to efficiently identify

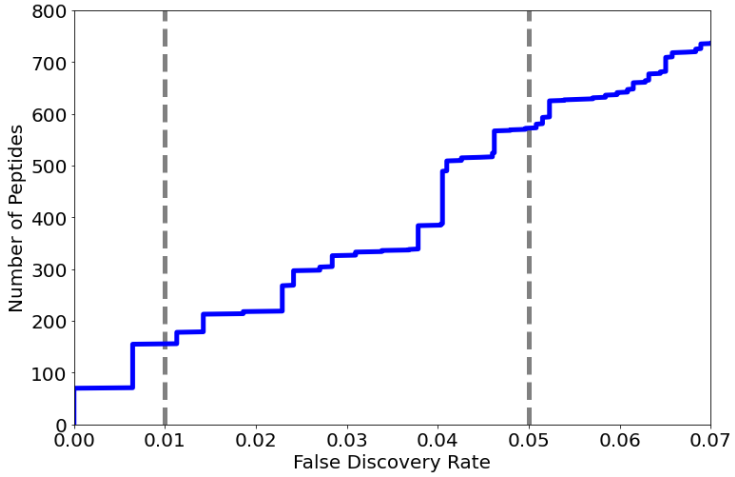

Figure 3.7: The number of mutated peptides identified by NeoMS from the Mel15 sample in dataset PXD004894. X-axis is the FDR threshold and y-axis is the number of peptides.

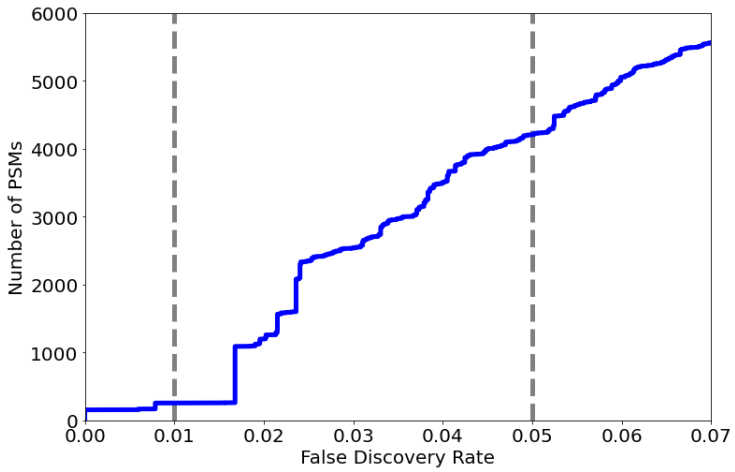

Figure 3.8: The number of mutated PSMs identified by NeoMS from the Mel15 sample in dataset PXD004894. X-axis is the FDR threshold and y-axis is the number of PSMs.

MHC-I peptides, both with and without mutations, in large-scale datasets. The accurate and comprehensive identification of mutated peptides opens up new avenues for investigating their associations with tumor cells and their potential as targets for T lymphocytes in cancer immunology research.

For each mutated peptide identified with a 5% FDR threshold by NeoMS, we retrieved its corresponding original peptide in the sequence database. The pair of peptides differ by only one amino acid. Their MHC-I binding affinities were predicted by MHCflurry, using the six alleles provided together with the dataset in the paper [5]. The maximum affinity of a peptide from the six alleles was used as the affinity for the peptide. Figure 3.9 shows the scatter plot of the predicted affinity scores of the identified peptide pairs. Among these, there were 81 mutated target peptides with binding affinity scores increased by at least 0.1. In contrast, only 8 have binding affinity scores decreased by at least 0.1. These clearly show that the single amino acid mutation in these mutated peptides generally improved the binding affinity.

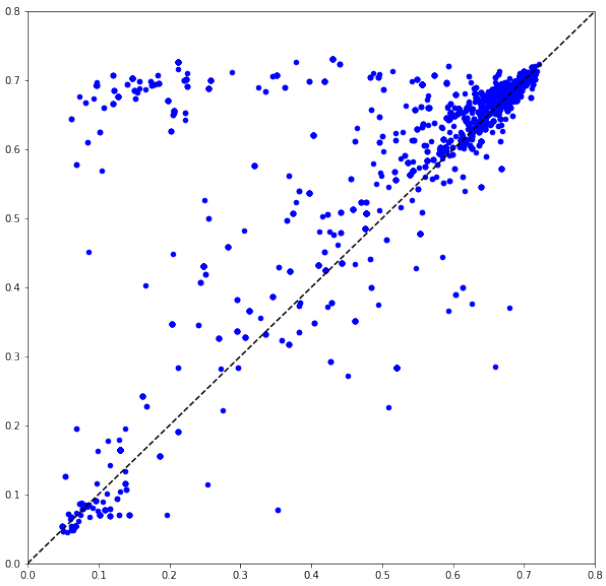

Figure 3.9: This analysis involves comparing the affinities between mutated peptides and their corresponding original peptides found in the sequence database. Each data point on the plot represents a pair consisting of a mutated peptide and its original counterpart. Y-axis and x-axis display their respective predicted affinity scores. The peptides located in the upper left quadrant are of particular interest, as they exhibit both high affinity scores and a significant increase in score due to the single amino acid mutation.

3.3.4 Case study

In the dataset PXD004894, a proteogenomics approach was utilized to identify mutated peptides [5]. Exome sequencing was conducted on the DNA from tumor samples of three patients, followed by somatic nucleotide variant (SNV) calling. This process led to the creation of a customized personalized reference database incorporating all protein isoforms affected by amino acid sequence alterations due to detected SNVs. The MaxQuant search tool was then applied to this personalized database, resulting in the identification of 11 mutated peptides.

In our study, we also build a personalized database, but not utilizing genomic data. We conduct the same FDR control as theirs and the comparative results between the proteogenomics approach and NeoMS are presented in Table 3.3. NeoMS successfully identified 6 out of the 11 peptides. Of the remaining 5 peptides, 4 were not within the 1% FDR threshold in the original study, suggesting weak spectrum signals. The weak signals make our *de novo* sequencing method fail to generate correct mutated tags, leading to their absence in our identification result.

Sequence	Mutation	FDR(proteogenomics)	FDR(NeoMS)	Patient
GRIAFFLKY	S-F	0.01	0.01	Mel 15
RLFKGYEGSLIK	P-L	0.01	0.01	Mel 15
LPIQYEPVL	P-L	0.01	N/A	Mel 15
RIKQTARK	T-I	0.05	N/A	Mel 15
KLILWRGLK	P-L	0.01	0.05	Mel 15
KLKLPIIMK	M-I	0.01	0.01	Mel 15
ASWVVPIDIK	E-K	0.05	N/A	Mel 15
GRTGAGKSFL	S-F	0.05	N/A	Mel 15
SPGPVKLEL	P-L	0.05	0.01	Mel 8
ETSKQVTRW	E-K	0.05	0.01	Mel 5
YIDERFERY	Q-R	0.05	N/A	Mel 5

Table 3.3: The mutated peptide identified by proteogenomics method and its FDR in the search result of NeoMS.

3.3.5 Ablation study

In this section, we conduct two ablation study for studying our NeoMS model. We first evaluate the capability of rescue low-ranked peptides. Then we perform how the iterative

training helps to construct a large and high quality labeled dataset and trains a robust machine learning model.

NeoMS identifies low-ranked peptide

In conventional rescoring methods, database search tools are configured to identify a single peptide for each spectrum. However, when identifying MHC peptides, these methods may not always detect the most suitable peptides for a given spectrum. In our experiments, we contend that certain peptides, despite not being highly ranked by the database search tool, exhibit a strong match to the spectrum.

In the NeoMS workflow, we configure Comet to identify the top k candidate peptides for each spectrum. We conducted an experiment to investigate the peptides identified by our method and assess their original rankings within Comet. We deply a 10-fold validation on the training set. Applying a 1% FDR cutoff, there are totally 542,682 PSMs are identified in the 10 rounds. The numbers of identified peptides of each ranking were shown in Figure 3.10. It is observed that a majority of the peptides (93.5%) are ranked at the top by the database search tool. However, if only these top-ranked peptides were considered for rescoring, the remaining 6.5% would not be rescued. While considering more candidate PSMs for feature computing does increase the runtime, it notably improves the overall performance.

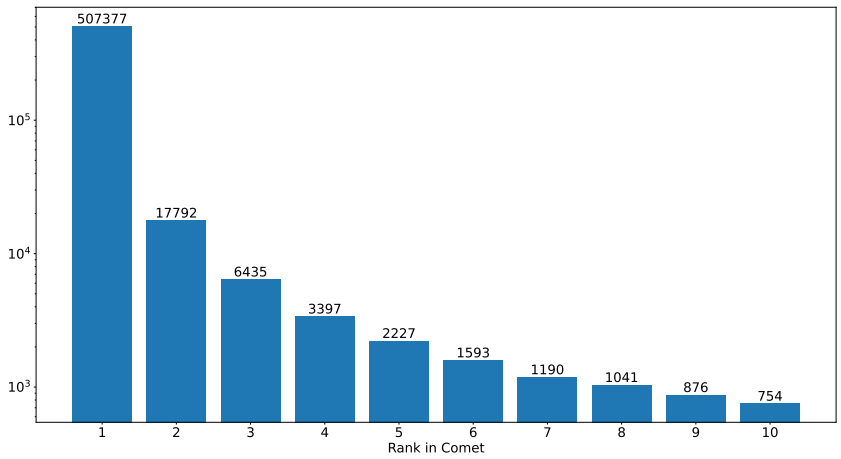

Figure 3.10: The original ranking of identified peptides. X-axis is the ranking and y-axis is the logarithm of number of peptides to the base 10 for better comprehension.

Iterative training

The training of machine learning models necessitates labels. However, due to the absence of ground truth labels in our training set, it is necessary to select high-confidence samples and assign labels before initiating the training process. One straightforward approach is to use Comet to search the dataset, considering PSMs that fall within the top 1% FDR as true labels, and selecting an equal number of decoy PSMs as negative labels. However, this initial labeled set, derived solely from Comet's differentiation, is suboptimal due to insufficient label quantity and compromised data quality.

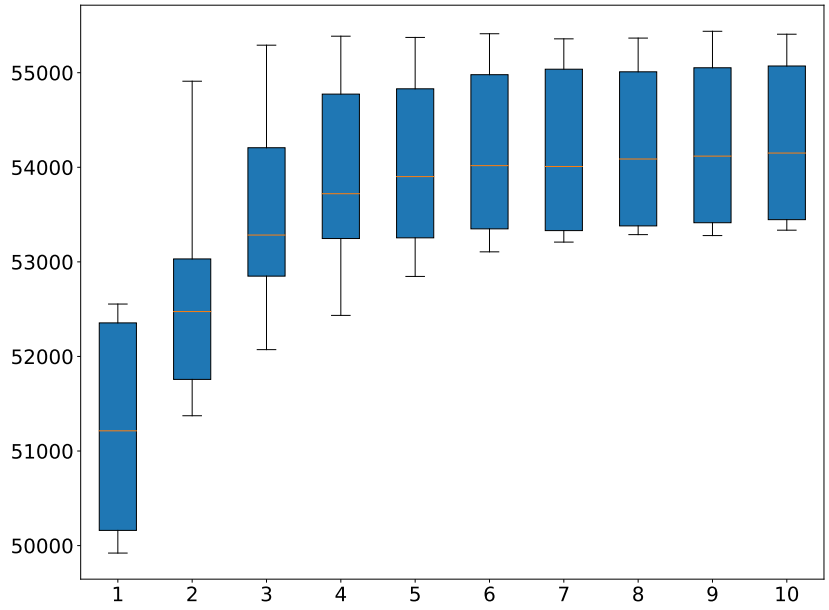

Figure 3.11: The graph depicts the number of peptides identified following iterative learning processes on cross validation. On x-axis, we display the number of training rounds completed. Y-axis shows the number of peptides identified by the model after completing x rounds of training.

To address these issues, we propose an iterative training approach, initially leveraging Comet's results for the first round of training and subsequently using the outcomes of our model for further refinement and training. To demonstrate the effectiveness of iterative training, we documented the model's performance after ten rounds of training. Similar to previous experiment, we conduct a 10 fold validation on the training set. The count of identified peptides (applying a 1% FDR cutoff) on the validation set for each round is

depicted in Figure 3.11. We noted progressive enhancements in the model's performance with each training round. Significantly, the most pronounced improvements were observed during the initial rounds, highlighting the importance of our refined dataset construction in boosting model accuracy. A comparative analysis between the outcomes after one round of training and after ten rounds reveals that the median number of identified peptides increased by approximately 5.96%, underscoring the effectiveness of our iterative training approach.

3.4 Discussion

Immunopeptidomics studies require the identification of MHC-I peptides containing amino acid mutations with high confidence from a sequence database and MS data. The lack of protease specificity with the consideration of mutations together increased the search space as well as the spectrum complexity. Both a better scoring function and rigorous result validation are required. In this work, we proposed a novel computational workflow, NeoMS to meet the needs for MHC-I peptide identification. Based on a *de novo*-based approach to detect mutations and to expand the sequence database, NeoMS could identify both the regular peptides that do not contain mutations and the mutated peptides containing exactly one amino acid mutation in a unified and efficient search workflow. For the identification of regular peptides, NeoMS outperformed all other search engines: under the same FDR constraints, NeoMS identifies most PSMs and peptides. These peptides are further examined by MHC-Flurry and more than 90% of our peptides are bound to MHC. A distinct advantage of NeoMS is that it can identify MHC-I peptides with a single amino acid mutation. In our melanoma dataset, NeoMS identifies 544 mutated peptides. We examine the binding affinities of these peptides before mutation and after mutation. Most of the peptides binding affinity are increased by the mutation.

Chapter 4

Deep learning boosted amyloidosis diagnosis

4.1 Introduction

Amyloidosis is a rare and progressive systemic disorder characterized by the accumulation of abnormal protein fibrils in various body tissues. It includes several types, such as AL, AA, ALECT2, and ATTR amyloidosis [106], among which AL is the most clinically significant form. AL amyloidosis is specifically associated with the production of monoclonal antibody-free light chains (LCs) by an abnormal clone of plasma cells. These free LCs aggregate into insoluble fibrils, which then deposit in different organs, causing dysfunction and ultimately leading to organ failure. AL amyloidosis can impact multiple organ systems, including the heart, kidneys, liver, and nervous system, resulting in a broad spectrum of clinical symptoms that severely affect the quality of life of patients [119]. Specifically, cardiac involvement in AL amyloidosis often leads to the most severe outcomes, including patient mortality.

Diagnosing AL amyloidosis can be challenging due to its nonspecific symptoms and only occurs after irreversible organ damage. Pre-symptomatic early diagnosis is crucial for implementing appropriate management strategies and improving patient outcomes [99] [54]. AL symptoms are always preceded by the circulation of the monoclonal LC in patient blood [103], and the amino acid sequence of the LC is a dominating factor in determining whether and where the LC deposit [108]. AL is primarily influenced by the sequence alterations on the V-domain of light chain (V_L) because it leads to abnormal folding and aggregation of the LC, resulting in the formation of amyloid deposits [10]. Mutations or

abnormalities in the V_L region sequence can contribute to the occurrence and progression of AL. As such, V_L peptide sequencing combined with computational prediction gives a potential solution for early diagnosis. In specific, there are two main types of LCs: lambda(λ) LC and kappa(κ) LC. In normal human body, the lambda/kappa ratio in AL patients is approximately 3:1, whereas in healthy individuals the ratio is around 1:2 [50].

LCs in immunoglobulins are highly diverse due to the unique sequences produced by plasma cells and the occurrence of somatic mutations during clonal expansion. This diversity contributes to the complexity of diagnosing, prognosis, and treating light chain amyloidosis, as it results in various clinical outcomes and treatment responses. There has been extensive work for predicting protein aggregation, particularly in the context of amyloidosis. Early efforts involved the development of algorithms like Tango [45], which is a statistical mechanics algorithm for predicting protein aggregation. Researchers later discovered that assessing the aggregation propensity of individual amino acids can enhance the accuracy of aggregation prediction. Several algorithms, including Aggrescan [24], Waltz [91], FoldAmyloid [49], PASTA 2.0 [136], have been developed to identify amyloidogenic regions in protein sequences and predict aggregation. The integration of machine learning methods further improved prediction accuracy. For example, APPNN [43] utilized a recursive feature selection method and employed a shallow neural network to predict aggregation. AmyloGram [16] and RFAmy [98] both employ a random forest classifier for the classification task with different feature extraction methods. iAMY-SCM [17] adopted a scoring card-based method for prediction. Subsequently, methods were developed to predict the amyloidogenic nature of the entire protein using sequence or structural information. VLAmY-Pred [116] incorporated 70 single amino acid features from the AAIndex database and literature, employing a decision tree algorithm called PART, for classification. LICTOR [50] extracted three types of structural features and utilized various machine learning classifiers for amyloidosis prediction. AB-Amy [164] employed the MRMD2.0 feature selection algorithm to select the most relevant features and utilized an SVM-based model for predicting amyloidogenic risk.

The progress in predicting AL amyloidosis has opened pathways for a better understanding of the disease. However, despite these advancements, there are still challenges to address, particularly in refining the capture of features from the complex nature of protein folding and aggregation, as well as improving the performance of current methods. To address this issue, an intuitive solution is to leverage deep learning techniques that directly learn from the sequences, eliminating the need for manual feature extraction.

In this study, we have developed DeepAL, a deep learning-based approach that accurately predicts AL from all types of light chain sequences, including both lambda and kappa. Instead of manually selecting features, DeepAL learns directly from the raw se-

quences. Given the limited availability of AL sequences for training deep learning models, starting from scratch to build a model poses challenges in acquiring comprehensive knowledge and presents a higher risk of overfitting the training data. To address this, the concept of few-shot learning (FSL) has emerged as a solution. FSL leverages prior knowledge to quickly adapt to new tasks with only a small number of samples and supervised information [143]. In DeepAL, we harness a pre-trained model that has learned general LC knowledge from an extensive collection of LC sequences. With the knowledge extracted from the pre-trained model, the rest of our model consists of a transformer module and a linear module that is trained to predict AL amyloidosis. To further enhance our training performance, we integrate an ensemble technique. Through meticulous experimentation, we compare DeepAL to previous methods using two distinct datasets from prior studies. On the first dataset, DeepAL achieves 90.72% the area under the ROC curve (AUROC), 87.97% accuracy, and 65.52 F_1score. On the second dataset, DeepAL achieves 89.19% AUROC, 81.95% accuracy and 77.55% F_1score. On both datasets, DeepAL surpasses all previous approaches in terms of various metrics. The high accuracy prediction of DeepAL make it potentially for medical use.

4.2 Methods

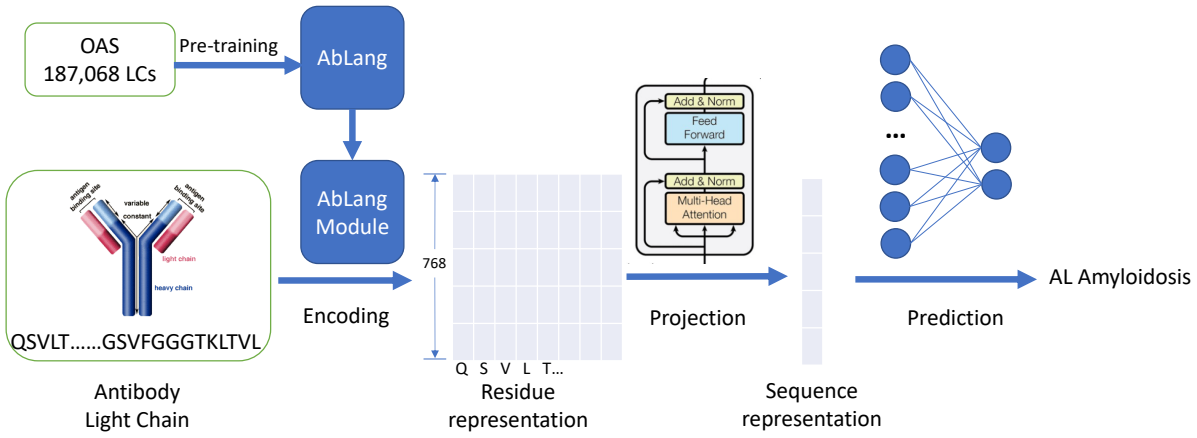

Figure 4.1: The architecture of DeepAL Model.

The architecture of DeepAL is depicted in Figure 4.1 and consists of three main modules: encoding, projection, and prediction. In the DeepAL framework, the encoding module utilizes AbLang [100], for sequence encoding, which converts a given LC sequence into a

feature space where each amino acid is represented by a 768-dimensional vector. Following this, the projection module is employed to reduce the 2-dimensional residue representation to a 1-dimensional sequence representation. In contrast to the untrainable averaging approach of AbLang, we integrate a one-layer transformer encoder module [134] within this module to enhance the learning of sequence features effectively. Subsequently, the prediction module is applied to generate the final prediction from the output of the projection module. This module comprises a one-layer linear model designed to identify and leverage critical sequence-level features crucial for accurate predictions. To improve training efficacy, we adopt an ensemble learning strategy known as Bootstrap Aggregating (Bagging). Throughout the training phase, k models are developed using unique train/validation splits to enhance model robustness and generalizability. During the prediction phase, the final output is derived by aggregating the predictions from these k models, thus ensuring a more reliable and stable performance.

4.2.1 Encoding module

AbLang follows the architectural principles of RoBERTa [83], and stands as the foremost antibody pre-training model. There are three modules in AbLang: Embedding module, Positioning module Transformer module. The Embedding module is designed to process an amino acid sequence, converting it into a vector through tokenization. In the case of RoBERTa, an English word vocabulary with a default size of 50,265, is employed for tokenization. In contrast, AbLang adopts 20 amino acids as tokens. Since in the transformer module, the position information cannot be learned by the orders, a module to inject position information is needed. Unlike RoBERTa which uses sine and cosine to represent positions, in AbLang's position module, considering the length of light chains are relatively small, it uses a learned positional embedding layer with a max length of 160. Transformer module within AbLang consists of 12 transformer blocks. Each of these blocks integrates 12 attention heads with an inner hidden size of 3,072, and an overall hidden size of 768. Much like RoBERTa, AbLang undergoes training via masked language modeling(MLM). Throughout the training process, between 1% and 25% of the residues within each sequence are selected. Within this subset, 80% undergo masking, 10% are randomly substituted with different residues, and 10% remain unchanged. The training dataset encompasses 187,068 light chains, sourced from the observed antibody space (OAS) [72] database. The model is trained for 40 epochs with a batch size of 4096. In DeepAL, we utilize AbLang to encode each sample's antibody LC raw sequences to a $768 \times L$ feature map for residue representation, where L is the length of the LC sequence.

Antibodies are Y-shaped proteins composed of two different types of polypeptide chains:

heavy chains and light chains. Within the OAS, there are approximately 14 million heavy chains and merely 0.19 million light chains [139]. Models trained on all antibody sequences run the risk of being dominated by heavy chains, resulting in a less understanding of light chain characteristics. Among the array of pre-trained models, AbLang is the only one that trained two models separately for heavy chains and light chains. The model that trained with light chain sequences is called AbLang-L. This model is a particularly fitting choice for our approach.

4.2.2 Projection module

In pre-trained language models like RoBERTa, the first token is [CLS], which represents the feature of the whole sequence. However, AbLang directly starts with the first amino acid and does not have a [CLS] token to represent the whole sequence. It provides sequence encoding by averaging the feature map on each position. Nevertheless, this averaging operation is non-trainable, resulting in a loss of information. Instead, we introduce an additional layer of transformer architecture into our model. This layer operates on the residue feature map, allowing the model to capture contextual information and interactions between amino acids in the sequence. By leveraging the transformer's self-attention mechanism, the model can effectively capture long-range dependencies and learn meaningful sequence representations.

4.2.3 Prediction module

After AbLang's encoding, the feature maps are padded to a uniform length of 128, with masking applied to the padding region to prevent interference during training. The output of the transformer layer retains the input size of 768×128. We extract the 768-dimension feature of the first position to represent the entire sequence. With the 768-dimension feature map, we apply a linear layer to predict. The input size is 768 and the output size is 2. The two output represents the probability of positive and negative, respectively.

4.2.4 Ensemble learning

In traditional deep learning training, the best model is typically selected based on its performance on the validation set, without learning from the validation set itself. This approach works well for larger datasets since the train/validation datasets are randomly split and their distributions remain consistent. However, in this case, where the dataset

is relatively small, the train/validation split may result in different distributions. Furthermore, not leveraging the learning potential from the sequences in the validation set could potentially decrease the model's performance. To mitigate these concerns, we employ a Bagging approach.

- Randomly select 80% samples (with replacement) from the training set.
- Train a separate instance of the base model using the sampled data.
- Repeat this process k times, where k is the number of base models.
- Use each trained base model to make predictions on the validation or test set. Combine the k predictions from all base models using an aggregation mechanism.

The final output is by averaging the k values, which is also known as soft voting. Based on our ablation study, DeepAL achieves optimal performance when the value of k is set to 5. This ensemble technique enhances the overall performance of models by leveraging the strengths of multiple models and reducing their collective weaknesses, resulting in a more robust and accurate ensemble prediction.

4.2.5 Model training

During our training, the AbLang module is frozen and the best part is trained. In all our experiments, the learning rate is set as 1e-5, and the batch size is set to 32. For each round, the training dataset is randomly split into training data and validation data with a ratio of 8:2. The maximum training epoch is 300, and early stopping regularization is set to 20 epochs. The checkpoint model that achieves the best performance in validation data is selected as the trained model.

In traditional binary classification tasks, the widely used loss function is cross entropy. However, cross-entropy treats incorrect positive samples and incorrect negative samples equally, which can pose challenges when dealing with imbalanced datasets like ours. This equality in punishment may result in the model being biased towards the majority class. To mitigate this problem, we have opted to use focal loss as our chosen loss function. Focal loss, initially introduced for object detection tasks, has shown effectiveness in addressing label imbalance issues. Unlike cross-entropy, focal loss assigns different weights to samples based on their difficulty in classification. By focusing more on challenging samples, focal loss helps the model better handle imbalanced datasets. The formula of focal loss is given in E.q. 4.1

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma log(p_t) \tag{4.1}$$

Here p_t is the model's estimated probability for the class with the label $y = 1$. There are two parameters in this formula: α and γ. α balances the importance of positive/negative examples and γ is called the **focusing** parameter that smoothly adjusts the rate at which easy examples are down-weighted. Through grid search experiment, we have set the parameters of the focal loss to their default values: $\alpha = 0.25$ and $\gamma = 2$.

4.3 Results

4.3.1 Datasets

Three public AL amyloidosis datasets have been used in this study.

- The first dataset (dataset I) is obtained from VLAmY-Pred [116], consisting of 348 amyloidogenic and 1,480 non-amyloidogenic amino acid sequences of antibody light chains. This dataset originates from AL-Base [11]. In AL-Base, LC sequences are classified into amyloid plasma cell disorder (AL-PCD), other plasma cell disorder (other-PCD), and non-plasma cell disorder (non-PCD). In this dataset, the AL-PCD sequences are considered amyloidogenic (positive) and the others are non-amyloidogenic (negative). Notice that the sequences with missing or unmatched Framework Regions (FRs) and Complementarity-determining regions (CDRs) are excluded. This dataset is downloaded from the VLAmY-Pred website.

- The second dataset (dataset II) is sourced from LICTOR [50], comprising 428 amyloidogenic and 590 non-amyloidogenic sequences from AL-Base [11], adding with 57 non-amyloidogenic sequences collected at the Institute for Research in Biomedicine (IRB-DB).

- The third dataset (dataset III) is also sourced from LICTOR [50], including 7 sequences associated with AL with cardiac involvement and 5 from multiple myeloma (MM) patients. These 12 sequences are with known clinical phenotypes but not present in the previous datasets. Here we use this dataset for the case study.

The details of the three datasets are shown in Table 4.1. Note that both dataset I and 2 were extracted from AL-Base for different ways of selecting positive and negative samples. We include both datasets here for comparison purpose with VLAmY-Pred and LICTOR, respectively.

Name	Source	Type	Positive	Negative
dataset I	VLAmY-Pred	Lambda and Kappa	348	1480
dataset II	LICTOR	Lambda	428	647
dataset III	LICTOR	Lambda	7	5

Table 4.1: The details of three benchmark datasets

4.3.2 Metrics

Following the previous study [116] [50], we utilized area under receiver operating characteristic curve (AUROC) as the evaluation metric for our models. Considering the imbalanced nature of the datasets, we also measure the area under precision recall curve (AUPRC) as an evaluation metric. Additionally, we estimated our models using the following performance metrics. Here we use TP, TN, FP, FN to represent the number of True Positive, True Negative, False Positive, and False Negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

$$F_1\text{score} = \frac{2 \times precision \times recall}{precision + recall} \tag{4.7}$$

4.3.3 Identification performance

In this section, we compared three methods specifically developed for determining AL: VLAmY-Pred [116], LICTOR [50] and AB-Amy [164]. Additionally, we included iAMY-SCM [17] in our comparison, which is the state-of-the-art method for predicting amyloid

proteins. It is important to note that iAMY-SCM is designed for general proteins and not specifically tailored for antibody light chains. For methods VLAmY-Pred, AB-Amy, iAMY-SCM, we upload our fasta format dataset files to their corresponding web server to get results. For LICTOR, we use their open source code for training and testing.

We first utilized dataset I, sourced from VLAmY-Pred's paper [116]. For our DeepAL method, we performed a 10-fold cross-validation on this dataset, ensuring rigorous evaluation. However, due to the unavailability of source code for the three other methods (VLAmY-Pred, AB-Amy, and iAMY-SCM), we were unable to conduct the same training and testing process as with DeepAL. Instead, we test the entire dataset and obtain the results on their respective web servers. Although this approach differed from our DeepAL evaluation, it allowed us to include these methods in our comparison.

The ROC curves for each method are plotted in Figure 4.2 and PR curves are plotted in Figure 4.3. In our experiment, DeepAL achieved an AUROC of 90.72% and an AUPRC of 73.14%, surpassing all other methods. On the other hand, iAMY-SCM only achieved an AUROC of 56.42%, slightly better than random prediction. This suggests that methods designed specifically for predicting protein amyloidosis cannot be directly transferred to predicting AL amyloidosis. Both VLAmY-Pred and AB-Amy achieved similar results in terms of AUROC: 86.95% and 86.94% for each method respectively. However, in terms of AUPRC, VLAmY-Pred achieves 66.23%, much higher than AB-Amy's 55.99%.

It is important to consider the potential overlaps between their training datasets and the dataset used in our experiment. Specifically, VLAmY-Pred is trained on the same dataset, and the training dataset of AB-Amy contains sequences from AL-Base, which is the same data source as our dataset. However, in our 10-fold validation, there were no overlaps between the training and testing data. Since the model will not see the testing data during training, a fair evaluation is ensured. Despite these potential overlaps, DeepAL still demonstrated superior performance compared to VLAmY-Pred and AB-Amy.

We also utilized other metrics for a comprehensive evaluation. The classifier's threshold significantly impacts the metrics, including accuracy, sensitivity, and specificity. In VLAmY-Pred the result, score of each protein is from -1 to 1. In iAMY-SCM is a float number. Other methods are ranging from 0 to 1. We take VLAmY-Pred using 0 as the threshold. For iAMY-SCM, we use its default threshold, 288.5625. For other methods, 0.5 is set as the threshold. From Table 4.2 we can see that DeepAL outperforms all other methods: in terms of accuracy, DeepAL achieve 87.97%, significantly higher than VLAmY_Pred's 79.39% and AB-Amy's 69.89%. In terms of other evaluation metrics including sensitivity, specificity and F_1 score, DeepAL still have reached the best performance overall.

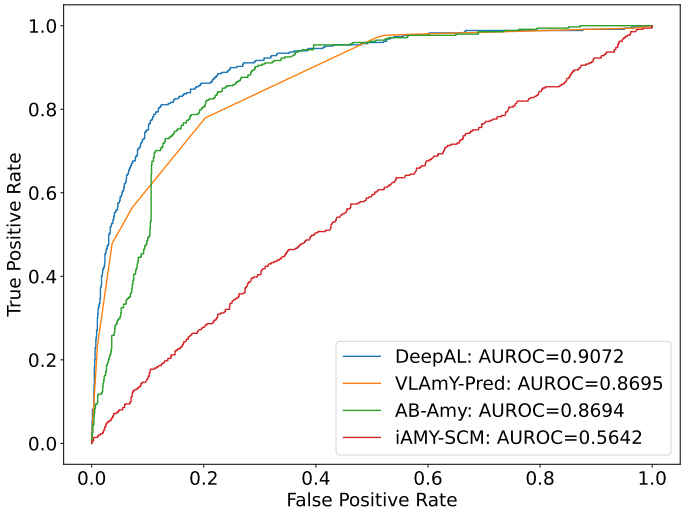

Figure 4.2: The ROC curves of DeepAL, VLAmY-Pred, AB-amy and iAMY-SCM on dataset I. X-axis is false positive rate and y-axis is true positive rate.

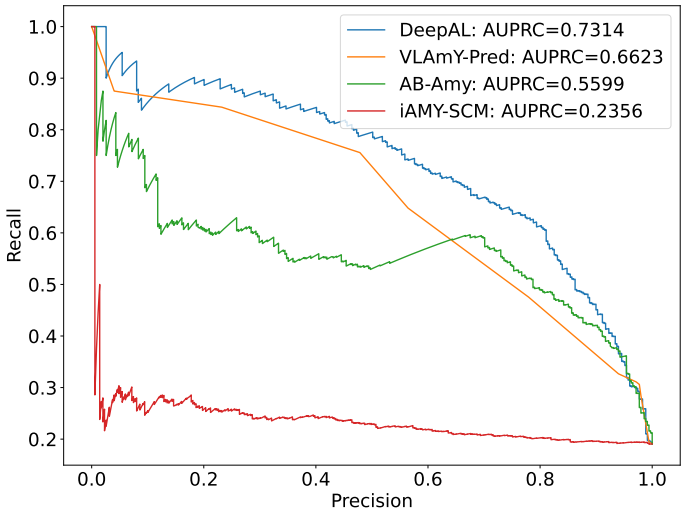

Figure 4.3: The PR curves of DeepAL, VLAmY-Pred, AB-amy and iAMY-SCM on dataset I. X-axis is precision and y-axis is recall.

Method	Accuracy	Sensitivity	Specificity	F_1score
DeepAL	**87.97%**	**72.32%**	**94.59%**	**65.52%**
VLAmY-Pred	79.39%	47.55%	79.73%	59.07%
AB-Amy	69.86%	38.02%	64.53%	53.89%
iAMY-SCM	21.92%	19.47%	3.85%	32.51%

Table 4.2: Comparison of the prediction results of DeepAL and three state-of-the-art computational methods (VLAmY-Pred, AB-Amy, iAMY-SCM) at VLAmY-Pred's benchmark dataset I.

In our next experiment, we utilized dataset II sourced from LICTOR's paper [50]. Considering the lambda/kappa ratio in AL patients is approximately 3:1, whereas in healthy individuals the ratio is around 1:2, LICTOR only included lambda sequences in the dataset. Since the source code for LICTOR is openly available, we utilze LICTOR's source code to perform the same train/validation split for both DeepAL and LICTOR. In this experiment, sequences along with their aligned germline sequences are provided. We conducted a 10-fold cross-validation for LICTOR and DeepAL, and for other methods, we test the entire dataset and obtain the results from their respective web server.

The ROC curve and PR curve of the 5 methods on dataset II are depicted in Figure 4.4 and Figure 4.5. DeepAL has an 89.19% AUROC and 82.66% AUPRC, which evidently shows that DeepAL outperforms other methods. Among the other comparing methods, LICTOR has the performance closest to DeepAL: an 86.18% AUROC and 81.08% AUPRC. We also compute the other metrics, shown in Table 4.3. DeepAL has the highest accuracy, sensitivity, and F_1 score, while LICTOR has specificity 86.09%, slightly higher than DeepAL's 84.39%

Method	Accuracy	Sensitivity	Specificity	F_1score
DeepAL	**81.95%**	**76.83%**	84.39%	**77.55%**
VLAmY-Pred	66.05%	55.37%	59.51%	64.04%
AB-Amy	54.51%	46.63%	25.35%	63.32%
iAMY-SCM	39.91%	39.72%	1.24%	56.59%
LICTOR	77.3%	75.27%	**86.09%**	69.19%

Table 4.3: Comparison of the prediction results of DeepAL and four state-of-the-art computational methods (VLAmY-Pred, AB-Amy, iAMY-SCM, LICTOR) at LICTOR's benchmark dataset II.

Although DeepAL exhibits superior AUROC compared to other methods on Dataset II, we observe that in the high-score range, LICTOR demonstrates better performance.

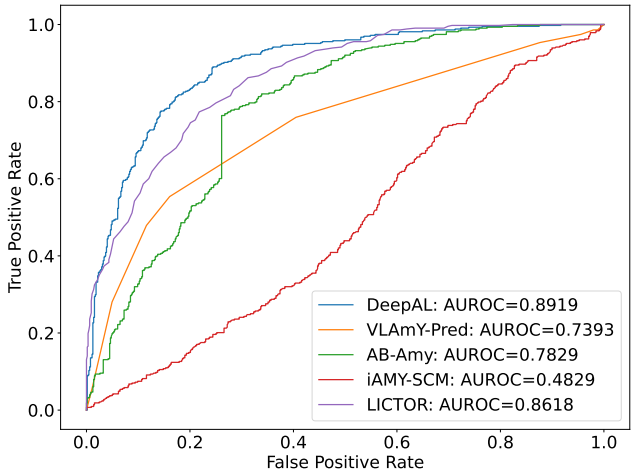

Figure 4.4: The ROC curves of DeepAL, VLAmY-Pred, AB-amy, iAMY-SCM and LICTOR on dataset II. X-axis is false positive rate and y-axis is true positive rate.

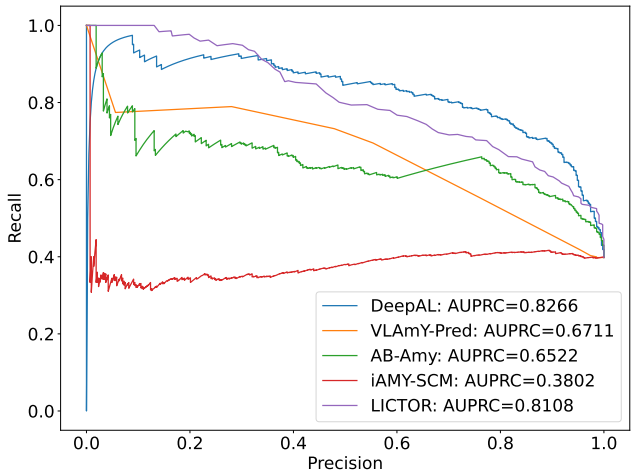

Figure 4.5: The PR curves of DeepAL, VLAmY-Pred, AB-amy, iAMY-SCM and LICTOR on dataset II. X-axis is precision and y-axis is recall.

LICTOR is a handcrafted feature-based method. They hypothesis that the toxity of AL is result from the somatic mutations on LCs. They totally detect three features:

- AMP (Amino acid at each mutated position). For a sequence, a list of variable pairs (p, aa) describes if amino acid aa is mutated on posision p.
- MAP (Monomeric amino acid pairs). They hypothesis all LCs share a conserved 3D structure(PDB ID: 2OLD). For a sequence, a list of variable pairs (p_1, aa_1, p_2, aa_2) describes if amino acid aa_1 and aa_2 is mutated on posision p_1 and p_2, and having a distance between the respective $C\beta$ atoms less than 7.5 Å in the X-ray structure.
- DAP (Dimeric amino acid pairs). Similar to MAP, DAP is also a list of variable pairs (p_1, aa_1, p_2, aa_2) but the two amino acids are located in different chains.

LICTOR employs a sequence alignment method to match LCs with their corresponding germline sequences and identify somatic mutations. Based on these mutations, three features (AMP, MAP, DAP) are extracted to represent each AL sequence. A random forest model is then utilized to learn from these features. Conversely, our DeepAL model directly learns from sequences without relying on handcrafted features. This absence of structural information may contribute to LICTOR's superior performance in the high-score region compared to DeepAL. To overcome this limitation, we propose a straightforward yet effective approach: aggregating the LICTOR model with our DeepAL model by adding their prediction results together. This integrated approach is termed DeepAL_LICTOR. It is important to note that, although these two methods are trained separately, they are evaluated using the same training and testing splits in the 10-fold validation.

We evaluated these three methods on Dataset II. The results, showcased in Figures 4.6 and 4.7, reveal that the combined DeepAL_LICTOR method not only maintains DeepAL's performance across most regions but also addresses the underperformance issue in the high-score region. This innovative approach results in a significant improvement, achieving an AUROC of 90.38% and an AUPRC of 85.57%, markedly enhancing the performance of DeepAL.

4.3.4 Ablation study

In this section, we evaluate the contribution of each module in our model: encoding module, projection module, and ensemble learning. In this experiment, we use dataset II as our benchmark dataset for the following experiments.

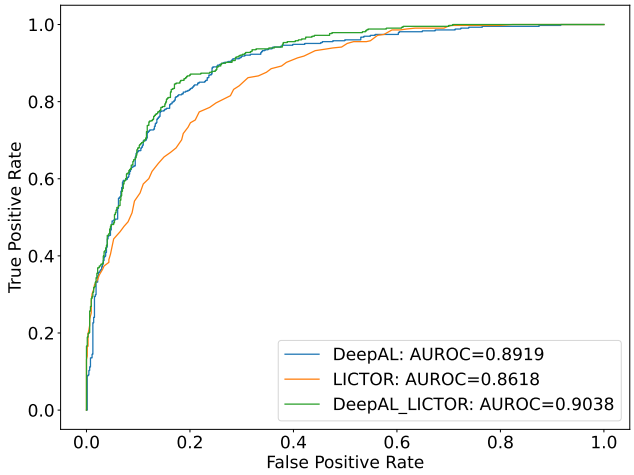

Figure 4.6: The ROC curves of DeepAL, LICTOR and DeepAL_LICTOR on dataset II. X-axis is false positive rate and y-axis is true positive rate.

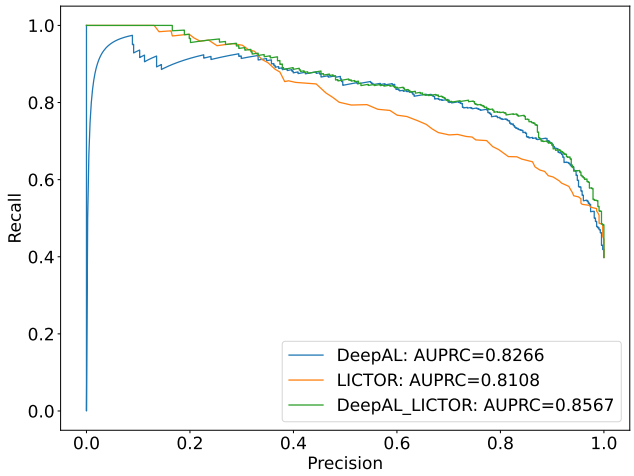

Figure 4.7: The PR curves of DeepAL, LICTOR and DeepAL_LICTOR on dataset II. X-axis is precision and y-axis is recall.

In DeepAL encoding layer, we used a pre-trained model AbLang to extract LC features. To see how much it improves the performance, we replace this module with a general embedding layer. We call this new model DeepAL_no_AbLang. In our projection module, a one-layer transformer encoder is used to extract sequence features. To find out how much this module learns sequence features, we replace this with an averaging pooling layer. We call this new model DeepAL_no_transformer. Similar to the previous experiment, we implement a 10-fold cross-validation on dataset II. The results of these three models are shown in Figure 4.9.

The AUROC of DeepAL_no_AbLang is recorded at 77.69%, which is significantly lower than the 89.19% achieved by DeepAL. This discrepancy underscores the critical role of the pre-trained model's capabilities. Through transfer learning from pre-established knowledge, DeepAL demonstrates enhanced performance. Additionally, the importance of the transformer module is highlighted, as DeepAL equipped with this module achieves an AUROC that is 5.38% higher than its absence. This indicates that a single linear layer is insufficient for accurately predicting AL.

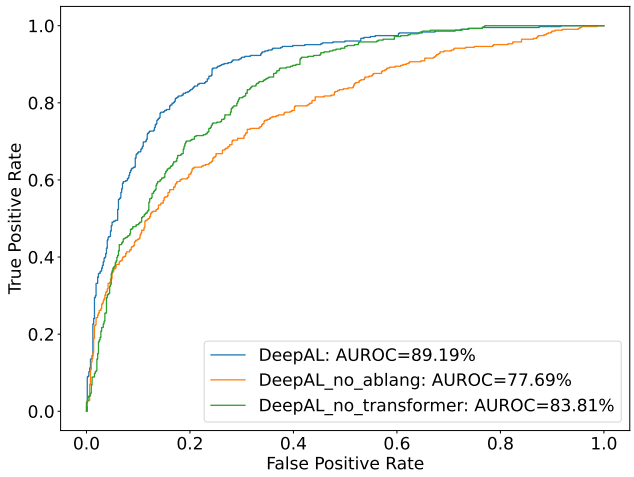

Figure 4.8: The ROC curves of DeepAL, DeepAL_no_AbLang and DeepAL_no_transformer. X-axis is false positive rate and y-axis is true positive rate.

4.3.5 Ensemble learning

In DeepAL, we implemented an ensemble learning technique known as Bagging. This approach was deemed necessary due to the limited sample size available for training. In this experiment, five models were trained by randomly splitting the training and validation datasets. X-axis represents the number of models (k) employed in the Bagging algorithm, while y-axis represents AUROC.

When k is set to 1, we have $\binom{5}{1}$ model results. The average AUROC across the five models is 87.33%. When k increases to 2, there are $\binom{5}{2} = 10$ combinations. The average AUROC improves to 88.42%. The results depicted in Figure 4.9 indicate the AUROC exhibits an upward trend with increasing k. When k is set to 5, it has the best result of 89.19%, 1.86% improvement compared without using Bagging, showing that multi-times of training make the model learn more knowledge from the training set.

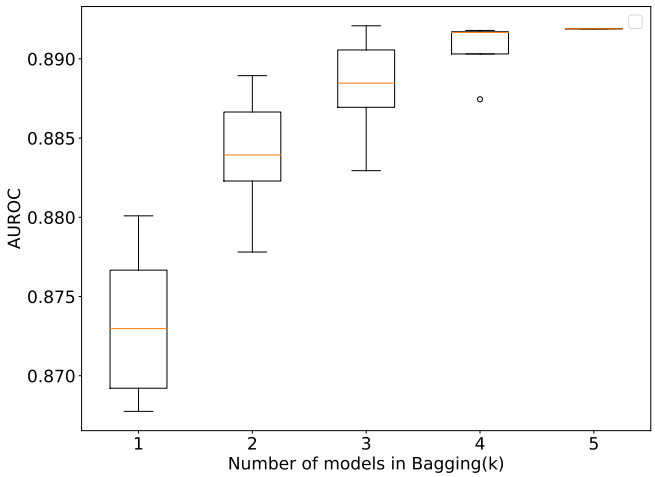

Figure 4.9: Boxplot of AUROC of using different numbers of models. X-axis is the number of models and y-axis its AUROC.

Not only the average AUROC is increasing, but the minimum AUROC is also increasing, which shows the utilization of the Bagging method not only enhances the overall performance of the model but also increases its robustness. By mitigating the risk of significantly poor performance on data with distinct distributions from the training set, the Bagging technique provides greater resilience. Although the performance gains achieved

through Bagging may be marginal, the improved robustness makes it less likely to encounter situations where the model performs inadequately.

4.3.6 Case study

To further investigate the performance of DeepAL, we utilized an additional dataset from LICTOR's paper that included new clinical phenotypes that are not present in the AL-Base. This dataset consisted of 12 sequences, with 7 positive samples and 5 negative samples.

In this experiment, we trained our DeepAL model using the LICTOR dataset and evaluated its performance on this LICTOR clinical dataset. DeepAL achieved an AUROC of 0.86 on this dataset. When using a threshold of 0.5, our model successfully predicts 6 positive sequences and 4 negative sequences, shown in Table 4.4. It demonstrated an accuracy of 83.3% and an F_1score of 85.71%, which aligns with the performance observed during the 10-fold validation in the previous experiment.

ID CODE	CLINICAL PHENOTYPE	DeepAL Prediction
H3	Toxic	**Toxic**
H6	Toxic	**Toxic**
H7	Toxic	Non-toxic
H9	Toxic	**Toxic**
H15	Toxic	**Toxic**
H16	Toxic	**Toxic**
H18	Toxic	**Toxic**
M2	Non-toxic	**Non-Toxic**
M7	Non-toxic	**Non-Toxic**
M8	Non-toxic	**Non-Toxic**
M9	Non-toxic	**Non-Toxic**
M10	Non-toxic	Toxic

Table 4.4: The label and predicting results in dataset III.

4.4 Discussion

Predicting AL amyloidosis from antibody sequences has presented itself as a significant yet formidable challenge. This paper introduces DeepAL, a deep learning-based prediction

approach tailored specifically for AL amyloidosis prediction. A remarkable feature setting DeepAL apart is its ability to achieve precise predictions without relying on manually feature extraction. Rigorous experimentation has unequivocally demonstrated that DeepAL outperforms previous methodologies. Moreover, our results underscore the distinctive advantages of employing pre-training, particularly in the context of training deep learning models for rare diseases characterized by limited available data. We believe that DeepAL will hold immense value in the diagnosis and treatment of AL amyloidosis. Furthermore, we aspire for our work to act as a driving force for additional research and innovation in this promising realm, fostering progress that not only aids patients but also contributes to the broader landscape of amyloidosis research.

Chapter 5

Anti-Cancer Peptides Identification and Activity Type Classification with Protein Sequence Pre-training

5.1 Introduction

Cancer stands as one of the most devastating and relentless human diseases, contributing to millions of deaths worldwide each year [121]. In the current landscape, radiotherapy and chemotherapy serve as the primary strategies for cancer treatment. However, radiotherapy approaches are inherently destructive, as they not only target cancer cells but also harm healthy cells. Chemotherapy involves the introduction of chemicals into the body to attack cancer cells, which increases the likelihood of drug resistance and recurrence [160] [84]. In the pursuit of anti-cancer drugs that can effectively target tumor cells while minimizing side effects, a class of short peptides, known as anti-cancer peptides (ACPs), has emerged as a potential breakthrough in cancer treatment. ACPs possess a unique ability to selectively target cancer cells while sparing healthy ones, making them promising candidates for novel therapeutic interventions [148] [4].

ACPs are a class of biologically active peptides that exhibit anti-tumor activity. Their unique sequences and structures enable them to interact with cancer cells and inhibit or destroy malignant cells. This makes ACPs a promising approach in combatting cancer. ACPs employ a variety of mechanisms to selectively eliminate cancer cells: Firstly, ACPs create pores in cell membranes, causing the release of cellular contents and resulting in necrosis. Moreover, they impede the formation of new blood vessels that supply nutrients

to tumors. Additionally, ACPs activate caspase cascades and mitochondrial pathways, initiating programmed cell apoptosis [71] [90]. Through these cytotoxic effects and precisely targeted actions, ACPs demonstrate potential in the development of effective and safe cancer drugs.

The development of ACP drugs involves two primary tasks: determining the peptide's capability to activate anti-cancer properties and categorizing the specific tissue it targets. However, the heavy reliance on wet laboratory experiments for ACP identification or classification presents significant challenges, including time-consuming, labor-intensive, and costly. Consequently, there is an urgent need for more efficient and cost-effective approaches to identify and classify ACPs. In this context, computational methods emerge as a promising solution to address this challenge. By harnessing the power of machine learning and statistical techniques, these methods expedite the ACP discovery process and accelerate the development of effective therapies for cancer patients.

To develop an accurate ACP identification tool, several methods have been proposed. AntiCP [132] is one of the first computational methods, which relies on three handcrafted features tailored to ACPs: Amino Acid Composition (AAC), Dipeptide Composition (DPC), and Binary Profile (BP). These features are harnessed in conjunction with a Support Vector Machine (SVM) model. Building on this foundation, AntiCP2.0 [2], ACPred-FL [144], and ACPred-Fuse [113] have emerged as methods dedicated to ACP identification, each with improvements in predictive performance. These tools incorporate more powerful handcrafted features and employ feature selection techniques to reduce dimensionality. Furthermore, they utilize a variety of machine-learning models to enhance their prediction processes.

The emergence and widespread adoption of deep learning have catalyzed the development of several deep learning-based methods for prediction. ACP-DL [153] employs a Long Short-Term Memory (LSTM)-based deep learning model that leverages k-mer sparse matrices and binary features. Another notable model, ACPred-LAF [56], utilizes a transformer encoder-based approach with multi-sense and multi-scale embedding algorithms to capture the context and sequential characteristics of ACPs. iACP-DRLF [85] is a model that takes advantage of both LightGBM and deep representation learning techniques for feature embedding. ME-ACP [44] employs ensemble learning as a preprocessing step for embedding, followed by inputting the processed data into a hybrid neural network that comprises residual modules and Bidirectional LSTM (Bi-LSTM) layers. ACP-ODE [155] utilizes a unique peptide sequence encoding method to represent ACPs. This model combines techniques from both deep learning, such as Bi-LSTM and Convolutional Neural Networks (CNN), and machine learning, particularly LightGBM, within its prediction framework.

One of the major challenges in developing highly accurate ACP prediction models is the limited size of ACP datasets, which typically contain fewer than 1,000 ACPs. This scarcity of ACP data can lead to overfitting in machine learning models. To address this issue, several strategies commonly used in deep learning have been applied. One such strategy is data augmentation, which is employed by methods like ACP-DA [20], ACP-DAD [8], and ACP-ASSF [125]. These techniques use data augmentation to artificially expand the dataset, helping the model generalize better and make more accurate predictions. Another approach involves leveraging pre-trained models, as demonstrated in Unidl4biopep [38]. This method utilizes pre-trained representations to extract informative features and predicts using CNNs. Pre-trained models capture valuable information from larger datasets and transfer that knowledge to enhance ACP prediction performance on smaller datasets.

Furthermore, alongside the accurate identification of ACPs, there is an increasing focus on classifying their functional types. Since an ACP is able to activate in several different tissues, this becomes a multi-label classification problem. For this problem, xDeep-AcPEP [19] utilizes multi-tasking learning within a CNN-based framework to predict the IC50 values of ACPs across six different tissue categories. Another approach, ACP-MLC [33], introduces a two-level prediction model: one for identification and another for classification. The two models share the same architecture but are trained on distinct datasets.

Advanced research has contributed significantly to an enhanced understanding of ACPs. However, it is important to acknowledge certain limitations in existing research efforts. One of the critical aspects of designing an ACP identification model is the representation of peptide sequences. Many methods rely on handcrafted approaches to represent ACPs, which often fall short of capturing the full spectrum of features. On the other hand, some approaches turn to deep learning methods to extract peptide features. However, the limited number of ACP peptides in datasets can lead to overfitting when attempting to learn representations directly from peptides. Furthermore, many existing methods take a hybrid approach, combining both machine learning and deep learning techniques. While this can be effective, it also introduces additional complexity in model development and is prone to overfitting training data.

In our research, we have introduced a novel method called DUO-ACP, which demonstrates dual prediction capability: performing both ACP identification and activity type classification. DUO-ACP leverages two distinct modules for peptide representation: a global feature embedding and a local feature embedding. We conducted comprehensive evaluations using one identification dataset and one classification dataset. For ACP identification task, DUO-ACP achieved an accuracy of 0.828 and an AUROC of 0.895, outperforming all other existing methods in this task. In the multi-label classification task,

DUO-ACP attained a Macro-averaged accuracy of 0.835 and a Macro-averaged AUROC of 0.886, both of which also exceed the current best results. To gain deeper insights into DUO-ACP's performance, we conducted experiments revealing how DUO-ACP integrates knowledge from both modules and how ensemble learning enhances robustness. In addition, we curated a new dataset containing ACPs that were not present in previous datasets. DUO-ACP was applied to both ACP identification and classification on this dataset, successfully identifying 92.6% of ACPs and achieving a Macro-averaged AUROC of 0.946 in classification. These results highlight DUO-ACP's potential for discovering novel ACPs. Overall, DUO-ACP has demonstrated its effectiveness and versatility in ACP identification and classification tasks, offering a valuable tool for advancing research in anti-cancer peptide discovery.

5.2 Methods

Our method addresses two crucial challenges in ACP analysis: determining whether a given protein sequence belongs to the ACP category and classifying it into specific tissue types. Our approach is capable of performing both identification and classification within a single model architecture. For both tasks, the input to our model is a protein sequence. In the first task, our model outputs the probability of the input sequence being identified as an ACP sequence. In the second task, it predicts the probability of activation in each tissue category.

The overall workflow is illustrated in Figure 5.1. Our model comprises two embedding modules: global feature embedding and local feature embedding, along with a prediction module. The embedding modules are responsible for mapping the input protein sequence into a learnable vector representation. The global feature embedding module uses a pre-trained model to capture global protein sequence information, while the local feature embedding module utilizes a randomly initialized vector to directly learn ACP-specific features. The prediction module's objective is to integrate these embeddings and output the probability of each ACP type.

5.2.1 Global feature embedding

In the realm of NLP, pre-trained language models have ushered in significant advancements in representation learning. Similarly, in the field of bioinformatics, the development of effective representation models for biological sequences, including gene and protein sequences,

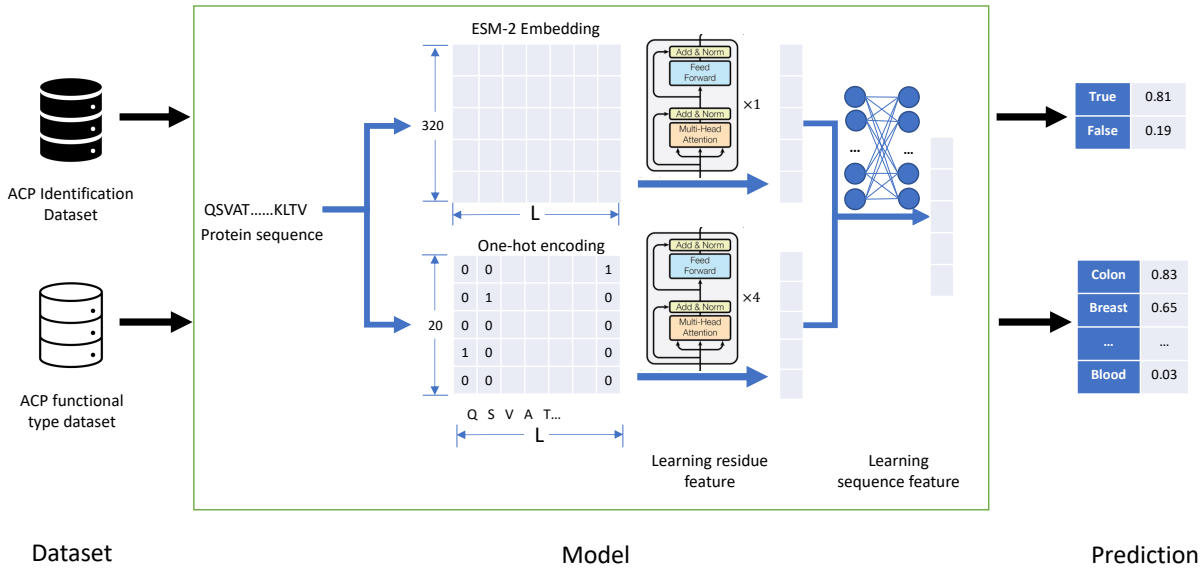

Figure 5.1: The DUO-ACP model's structure. This model processes protein sequences derived either from an ACP identification dataset or an ACP functional type dataset. Each sequence undergoes embedding through two distinct modules: a global feature embedding module and a local feature embedding module. The outputs from these modules are combined into a single vector. For ACP identification predictions, the model outputs the likelihood of the sequence being an ACP. In the case of predicting ACP functional types, it outputs the probability distribution across different types.

is crucial for predictive tasks. ESM-2 [117], a prominent protein sequence representation model, is built upon a transformer-based architecture and trained on an extensive dataset containing up to 250 million sequences from the UniParc database [25], encompassing a staggering 86 billion amino acids. In DUO-ACP, we download and leverage a 6-layer pre-trained model [81] for computing global features. This model generates a $320 \times L$ matrix, where L represents the length of the input sequence.

The $320 \times L$ matrix is then projected to a vector for peptide feature representation. Similar to previous DeepAL method, to represent the sequence, we introduce an additional layer of transformer architecture into our model. This layer operates on the residue feature map, enabling the model to capture contextual information and interdependencies between amino acids in the sequence. By leveraging the transformer's multi-head self-attention mechanism, our model can effectively capture long-range dependencies and learn meaningful sequence representations. This enhanced representation contributes to the overall performance of DUO-ACP.

The typical length of ACP is from 10-50. However, there are a few ACP in our dataset exceeded 50. To train or predict on these long sequences, we set out maximum input as 128. Before this step, the feature maps are padded to a uniform length of 128, with masking applied to the padding region to prevent interference during training. The output of the transformer layer retains the input size of 320×128. We extract the 320-dimension feature of the first position to represent the entire sequence.

5.2.2 Local feature embedding

The pre-trained model excels at learning general protein features from a vast corpus of protein sequences. However, it may not capture specific knowledge related to ACPs. To address this gap, we introduce a local feature embedding sub-module dedicated to learning ACP-specific domain knowledge. In this module, protein sequences are initially converted into a 20-dimensional matrix using one-hot embedding. We then employ a lookup table that projects this matrix into a 256-dimensional space. Similar to the previous sub-module, the input to this sub-module is in the form of a $20 \times L$ matrix, which is padded to create a 20×128 matrix. However, unlike utilizing a pre-trained model where a single layer of transformer encoder can suffice for feature extraction, converting raw input sequences into meaningful representations typically necessitates a deeper neural network. To this end, we employ a four-layer transformer encoder within this sub-module to learn ACP-specific features. This deeper architecture enables the model to capture intricate patterns and domain-specific knowledge related to anti-cancer peptides, enhancing the model's ability to make accurate predictions for ACP identification and classification.

5.2.3 Prediction module

In this module, we concatenate features from the global and local feature embedding whose dimensions are 320 and 256. Moreover, we add one hand-crafted feature: sequence length. To combine these features, in our prediction module, the combined feature vector of dimension 577 is projected for the final outputs. For the binary classification problem, the output size is two with a SoftMax activation function. The two output values represent the probability of true or false. For multi-tasking problems, the output size is the number of types. In dataset III, there are 7 types of ACP (Colon, Breast, Cervix, Lung, Skin, Prostate, Blood), so the output size is 7, with a Sigmoid activation function.

5.2.4 Model training

Our model's training is divided into two distinct steps. The first step focuses on training the two embedding modules, while the second step is dedicated to training the prediction module. We adopt this approach due to the different initial states of the two modules: the global embedding module is pre-trained, whereas the local feature embedding module begins with a random initialization. Directly training the entire model can result in uneven updates, often leading to an overreliance on the pre-trained global module and insufficient learning of local features. To mitigate this, we train each module separately and then combine them for the prediction phase.

In the first step of our training, the global feature embedding module is linked to a 320-dimensional linear layer, and the local feature embedding module is connected to a 256-dimensional linear layer. Each of these models is trained independently with a learning rate 1e-5. After this step is complete, we extract the embedding modules from both models to initialize the respective modules in our comprehensive model. Then in the second step, we train our comprehensive model with a learning rate 1e-6.

In both training steps, we set the batch size to 32 and limit the training to a maximum of 300 epochs. We also implement early stopping regularization, halting training if there's no improvement after 20 epochs. The choice of loss function is tailored to the specific training task: for binary classification tasks, considering the dataset is balance for positive and negative samples, we use the Cross-Entropy (CE) loss function, while for multi-tasking problems, the Binary Cross-Entropy (BCE) loss function is applied. It's important to note that during training, the model is not permitted to use test data for the purpose of early stopping. The chosen model checkpoint is the one that records the lowest loss on the validation set.

5.3 Results

5.3.1 Datasets

Dataset name	#Positive Entries	#Negative Entries
dataset I	242	242
dataset II	61	61

Table 5.1: The number of peptides for each type in ACP identification dataset.

In ACP identification, the ACP-Mixed-80 dataset originally sourced from the ACPred-LAF paper [56] is utilized. The ground truth sets were sourced from two databases: CancerPPD [133] and MLACP [89]. Five datasets were merged and validated against these ground truth sets, forming a new dataset that comprises 1,054 ACPs and 4,895 non-ACPs. The non-ACPs were derived from regular peptides and antimicrobial peptides that lack anticancer properties.

To ensure data quality, a series of preprocessing steps were performed. These included label verification, correction, removal of duplicates, and separation of ambiguous samples. Following these preprocessing steps, 736 unique ACPs were selected, along with an equal number of unique non-ACPs, chosen randomly from the remaining 4,577 samples. To create balanced training and testing datasets, the data was split into an 80:20 ratio, resulting in a training dataset with 558 ACPs and 558 non-ACPs and an independent test set comprising 148 ACPs and 148 non-ACPs. To reduce sequence similarity within the datasets, peptides with more than 80% sequence identity were removed using the CD-HIT [48] program. As a result, a training dataset (referred to as dataset I) consisting of 242 positive and 242 negative samples and a testing dataset (referred to as dataset II) containing 61 positive and 61 negative samples are obtained.

A multi-label dataset was obtained from ACP-MLC [33]. It is originated from CancerPPD [133], where ACPs are categorized into 21 functional groups based on the types of tissues they target. To ensure statistical significance, functional types with fewer than 40 unique entries were excluded from the dataset. Further preprocessing steps involved removing duplicate sequences, non-linear peptides, excessively long peptides (those with lengths exceeding 100 amino acids), and peptides containing non-standard amino acids. Additionally, peptides sharing more than 90% pairwise sequence identity with any other sequences were eliminated using the CD-HIT program [48]. This curated dataset, referred to as dataset III, comprises 211 ACPs that target various tissue types, including colon, breast, cervix, skin, lung, prostate, and blood.

Type	Number of Entries
Colon	137
Breast	124
Cervix	105
Skin	106
Lung	94
Prostate	89
Blood	63
Total	211

Table 5.2: The number of peptides for each type in ACP type classification dataset (dataset III).

5.3.2 Evaluation metrics

To evaluate the performance of ACP identification, we employ several widely used metrics commonly utilized in previous studies. These metrics include accuracy (ACC), sensitivity (SE, also called recall), specificity (SP), precision, Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic ROC curve (AUROC). Here, we use TP, TN, FP, and FN to represent True Positive, True Negative, False Positive, and False Negative, respectively. MCC is defined as follows:

$$MCC = \frac{TP \times TN - FP \times \ FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{5.1}$$

In a multi-label classification task, the prediction for each label can be treated as a binary classification task. Therefore, the metrics mentioned earlier for evaluating identification can also be applied here for each label. When it comes to integrating the evaluation metrics across all labels, there are two primary approaches: Macro-averaging and Micro-averaging. Macro-averaging calculates each metric independently for each label and then computes the average to obtain the final metric. In contrast, Micro-averaging involves first aggregating TP, TN, FP, and FN values across all labels and then calculating the evaluation metrics. The definitions for Macro-averaging and Micro-averaging are as follows:

$$\text{Macro Metric} = \frac{1}{L}\sum_{l=1}^{L} evalMetric\,(TP_l, TN_l, FP_l, FN_l) \tag{5.2}$$

$$\text{Micro Metric} = evalMetric\left(\sum_{l=1}^{L} TP_l, \sum_{l=1}^{L} TN_l, \sum_{l=1}^{L} FP_l, \sum_{l=1}^{L} FN_l\right) \tag{5.3}$$

In addition to the previously mentioned metrics, another important evaluation metric for our multi-label classification task is Hamming Loss. Hamming Loss is a measure that quantifies the fraction of labels that are incorrectly predicted for a given sample. It takes into account both false positives and false negatives for each label, providing a comprehensive view of the classification performance across multiple labels simultaneously. Hamming Loss is defined as follows:

$$\text{Hamming Loss} = \frac{1}{D}\sum_{i=1}^{D}\frac{r_i \Delta Z_i}{L} \tag{5.4}$$

Here D is the number of instances in the dataset. r_i is the set of true labels for instance i. Z_i is the set of the predicted labels for instance i. $r_i \Delta Z_i$ calculates the symmetric difference between the true and predicted labels for instance i. L is the total number of labels.

5.3.3 ACP identification results

In this section, we conducted an evaluation of DUO-ACP by comparing its ACP identification capabilities against four existing state-of-the-art methods: ACPred-LAF [56], ACP-MLC [33], ACP-ODE [155], and UniDL4BioPep [38]. This evaluation was carried out by training the models on dataset I and testing them on dataset II. For the first two methods, ACPred-LAF and ACP-MLC, we obtained the results directly from their published papers respectively. ACPred-LAF developed four models using different embedding methods: ACPred-LAF Basic, ACPred-LAF MSE, ACPred-LAF MSC, and ACPred-LAF MSMC. Among these, the ACPred-LAF MSMC model achieved the highest AUROC, so we utilized it as the representative of ACPred-LAF. In the case of ACP-MLC, the results of seven models employing different machine-learning algorithms were provided. Similarly, we selected the random forest model since it yielded the highest AUROC to represent ACP-MLC. Concerning the other two methods, ACP-ODE and UniDL4BioPep, we employed their source code to train and test. In order to ensure the trained models are optimum, for each method we have tried several hyper parameter settings and the model with the highest AUROC in test set is used in comparison. During training our approach keeps the test

dataset unseen, and we choose the best model based on its performance on the validation dataset. This rigorous evaluation methodology ensures a fair and reliable comparison of our model against existing methods in the context of ACP identification. The comparison results are presented in Table 5.3.

Model	ACC	SE	SP	MCC	AUROC
ACPred-LAF	0.812	0.721	**0.902**	0.633	0.827
ACP-MLC	0.787	0.803	0.770	0.574	0.888
ACP-ODE	0.820	0.770	0.869	0.642	0.862
UniDL4BioPep	0.787	0.797	0.778	0.558	0.853
DUO-ACP	**0.828**	**0.885**	**0.902**	**0.663**	**0.895**

Table 5.3: The testing results on dataset II of ACP identification dataset

Among the five methods compared, it is evident that our DUO-ACP model excels by achieving the best performance across multiple key metrics. Given that the dataset maintains a balanced distribution between true and false labels, the AUROC is considered a crucial metric. In this regard, DUO-ACP outperforms all other methods, achieving an impressive AUROC value of 0.895. Regarding other metrics, it's important to note that the choice of threshold settings can impact the results. In this comparison, we selected a threshold of 0.8. While each of the other four methods achieved the best performance in a specific metric (ACP-ODE for accuracy, UniDL4BioPep for sensitivity, ACPred-LAF for specificity, and ACP-MLC for AUROC), our DUO-ACP consistently outperforms the best model across all metrics. This outstanding performance underscores the effectiveness and superiority of our DUO-ACP model in the realm of ACP identification.

5.3.4 ACP activity type classification results

Aside from identification, we also evaluate our method's capability in activity type classification. In this experiment, DUO-ACP is compared with the state-of-the-art method ACP-MLC [33] using their provided dataset III. In this dataset, only training data is provided, and there is a lack of an independent testing dataset. To ensure a rigorous assessment of our model and to facilitate a fair comparison, we adopted the same 10-fold cross-validation strategy that ACP-MLC employed on this dataset. Similar to previous experiments, ACP-MLC's results are obtained from their paper. This approach allows us to maintain consistency in evaluation methodologies and provides a robust basis for assessing the performance of our model in the absence of an independent test dataset. The comparison results are presented in Table 5.4.

Method	AUROC	ACC	SE	SP	F1-score	MCC
ACP-MLC	0.868	0.773	0.608	**0.854**	0.676	0.509
DUO-ACP	**0.886**	**0.835**	**0.812**	0.819	**0.819**	**0.647**

Table 5.4: The testing results on dataset III of ACP type classification dataset.

The predictive results of DUO-ACP and ACP-MLC for each targeted tissue type are illustrated in Figure 5.2, showcasing a comparison of accuracy, MCC, AUROC, and F1-score for each tissue type. In this experiment, a threshold of 0.5 was set. For the six tissue types (breast, cervix, lung, skin, prostate, blood), DUO-ACP consistently achieves higher MCC than ACP-MLC, and it also demonstrates better performance across other metrics. The only exception is the tissue type colon, for which ACP-MLC achieves better results across all metrics.

For the micro-averaged results, DUO-ACP achieves a sensitivity of 0.824, a recall of 0.840, and an F1-score of 0.832. The Hamming loss between the prediction results and labels is 0.165. To provide a comprehensive comparison between DUO-ACP and ACP-MLC, the macro-averaged results of the two methods are presented in Table 5.4. Across all six metrics considered, DUO-ACP outperforms ACP-MLC in five of them, with the only exception being specificity. Overall, our model attains impressive macro-averaged results, boasting a Macro-AUROC of 0.886 and a Macro-ACC of 0.835. These results surpass the respective values of 0.868 and 0.509 obtained by ACP-MLC. These superior performance metrics underscore the effectiveness and advantages of DUO-ACP in the activation classification task.

5.3.5 Ablation study

In DUO-ACP, there are two embedding modules: global feature embedding and local feature embedding, along with a prediction module that integrates them. We refer to these three models as Global-ACP, Local-ACP, and DUO-ACP. To understand how these modules contribute to performance improvement, we evaluated the performance of each of them. In this study, we focused on binary classification tasks and evaluated dataset I and dataset II. To ensure statistical robustness and reduce randomness, we created a new dataset by combining dataset I and dataset II into a single dataset, containing 606 peptides (303 positive and 303 negative). Then, we performed 5-fold cross-validation on this new dataset for all three methods.

In the first experiment, we extracted the embeddings from the last layer before the output for all three models: DUO-ACP, Global-ACP, and Local-ACP. We then applied

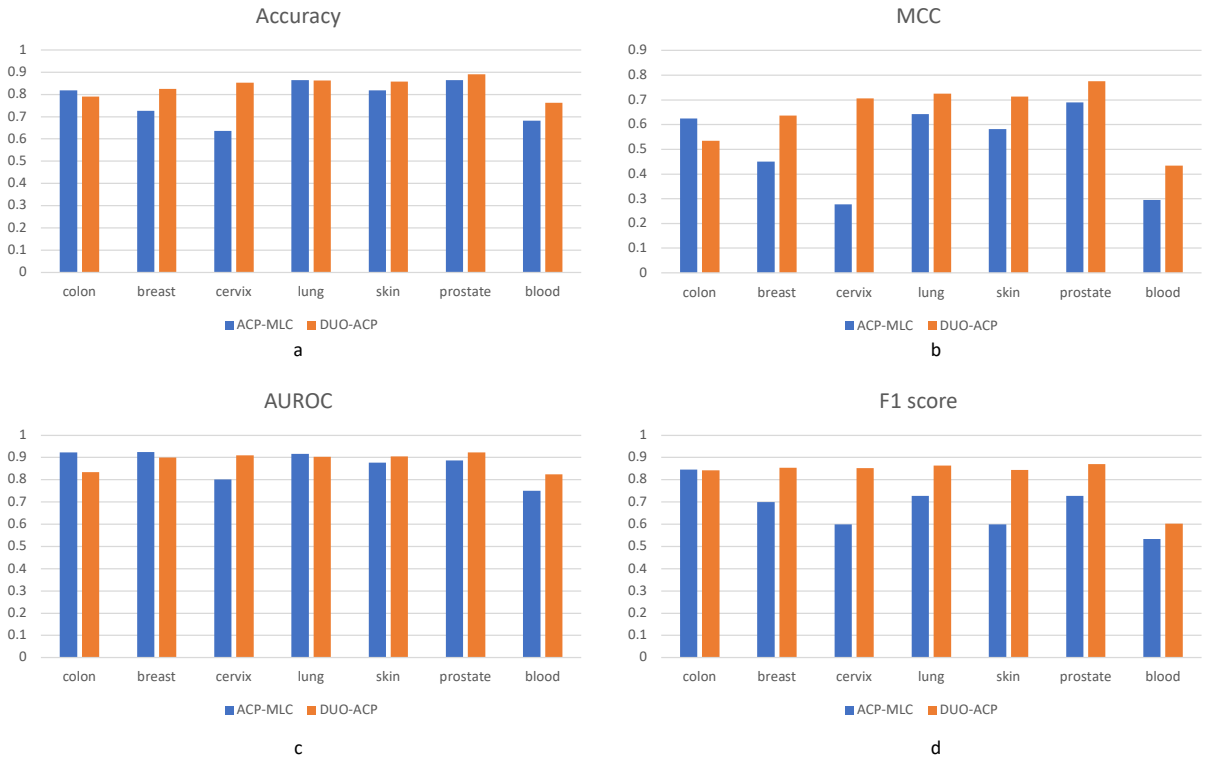

Figure 5.2: The performances comparison between ACP-MLC (represented in blue) and DUO-ACP (represented in orange) on the multi-label classification task are depicted in panels (a) through (d). These panels illustrate the accuracy, MCC, AUROC, and F1-score, respectively, for each tissue type being analyzed.

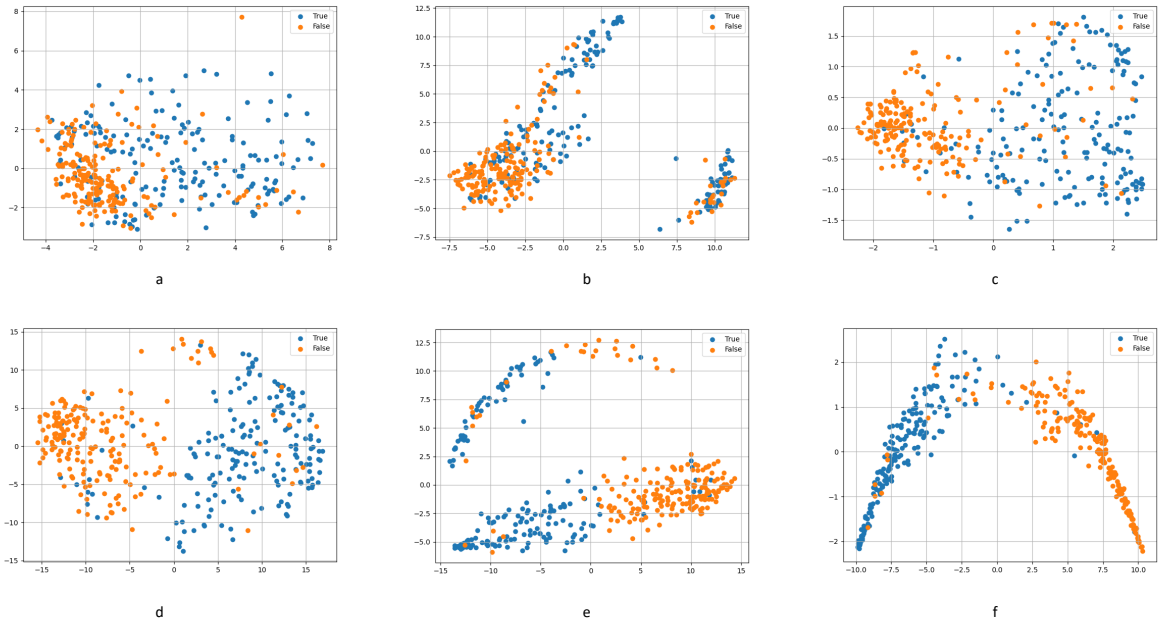

Figure 5.3: The visualization of PCA dimensional reduction of three embeddings. (a)(b)(c): The embedding of training for 10 epochs of Global-ACP, Local-ACP, and DUO-ACP. (d)(e)(f): The final trained embedding of Global-ACP, Local-ACP, and DUO-ACP.

PCA dimensional reduction to these embeddings to visualize how they can separate positive and negative samples. Figure 5.3 depicts the embeddings of the three methods after training for 10 epochs and at the end of training. When comparing Figure 5.3 (a)(b)(c), it's evident that the DUO-ACP model converges faster than Global-ACP and Local-ACP due to the components that have already been trained. For the final trained models, both Global-ACP and Local-ACP achieve good results in separating positive and negative samples (Figure 5.3 (d)(e)). The distributions of these two embeddings are different, indicating that they potentially distinguish samples from different perspectives. DUO-ACP integrates the advantages of both modules and exhibits the best performance (Figure 5.3 (f)). This suggests that DUO-ACP effectively leverages the different features learned by the global and local embedding modules to improve overall performance.

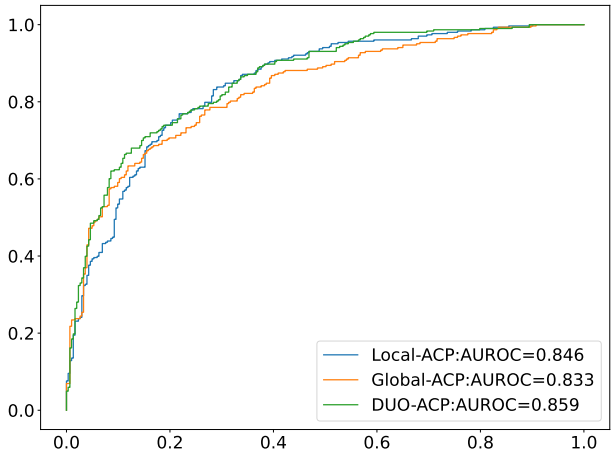

Figure 5.4: ROC curves of Local-ACP, Global-ACP, and DUO-ACP.

Figure 5.4 displays the ROC curves of the three models in the ablation study. It's evident that Global-ACP or Local-ACP experiences a decrease in AUROC compared to the original model. During this 5-fold cross-validation, DUO-ACP achieves an AUROC of 0.859. When the local embedding module is removed, there is a 1.5% drop in AUROC, and when the global embedding module is removed, there is a 3% drop in AUROC. These results indicate that both local and global feature embedding modules contribute significantly to the overall performance of DUO-ACP.

5.3.6 Case study

To demonstrate the potential capabilities of DUO-ACP in the discovery of novel ACPs, we conducted case studies involving samples that were not previously included in any existing datasets. Specifically, we gathered all samples from the CancerPPD database and subjected them to a series of filtering and organization steps as outlined below:

1. We retain only linear peptides composed of the 20 natural L-form amino acids, excluding cyclic peptides, D-form peptides, and mixed peptides.
2. Peptides that exhibited activity against at least one of the seven tissues in dataset II are retained.
3. Entries that included EC50, LC50, IC50, or LD50 values are retained, while those with missing values are removed.
4. Peptides already present in dataset I, II or III are excluded.
5. To minimize redundancy, we utilized CD-HIT to eliminate peptides sharing more than 80% sequence identity.

Following these procedures, we obtained a dataset comprising 109 ACPs along with their corresponding activity types. This new dataset was then employed to assess the efficacy of DUO-ACP's identification model, which had been trained on dataset I and II, in predicting whether these peptides qualified as ACPs. Remarkably, DUO-ACP identified 92.7% of the peptides in this dataset as ACPs. Subsequently, DUO-ACP's classification model, trained on dataset III, was used to predict the probability of each activity type for the ACPs in this dataset. The macro-averaged result across the seven tissues yielded a remarkable 94.6% Macro-averaged AUROC score, underscoring the exceptional generalization capabilities of our model.

To further investigate the analysis, we selected the top 10 candidate peptides based on their probability scores. The functional type of these peptides from CancerPPD is in Table 5.5 and our DUO-ACP's prediction results on each type is in Table 5.6. Notably, all these peptides exhibited probabilities exceeding 98.5% for being classified as ACPs. In the context of the classification task, DUO-ACP's results closely align with those in the CancerPPD database. However, it's worth noting a single peptide, GFKMALKL-LKKVL, which was predicted to be active in blood tissue by DUO-ACP but lacked such a label in CancerPPD. Our research did not uncover concrete evidence supporting its role in interacting with blood cancer cells. Nevertheless, it has exhibited toxicity in human erythrocytes [123], suggesting its potential function in blood cancer cells.

Sequence	Label
FALALKLAKKL	Breast, Cervix, Colon, Lung, Prostate, Skin
FAKKLLAKALKL	Breast, Cervix, Colon, Lung, Prostate, Skin
GFKMALKLLKKVL	Cervix,Colon
KWFKKIPKFLHLAKKF	Blood,Breast
KWKLFKKIPLAKKF	Blood,Breast
KAKLAKKALAKLL	Breast, Cervix, Colon, Lung, Prostate, Skin
FAKALAKLAKKLL	Breast, Cervix, Colon, Lung, Prostate, Skin
GKWKKILGHLIR	Cervix,Colon
GKWMSLLKHIWK	Cervix,Colon
FAKKLAKLAKKALAL	Breast, Cervix, Colon, Lung, Prostate, Skin

Table 5.5: The functional types of the top 10 peptides.

Sequence	Colon	Breast	Cervix	Lung	Skin	Prostate	Blood
FALALKLAKKL	0.981	0.974	0.984	0.958	0.977	0.940	0.002
FAKKLLAKALKL	0.989	0.986	0.980	0.982	0.990	0.973	0.002
GFKMALKLLKKVL	0.982	0.020	0.995	0.039	0.024	0.030	0.902
KWFKKIPKFLHLAKKF	0.012	0.964	0.013	0.088	0.005	0.015	0.989
KWKLFKKIPLAKKF	0.072	0.958	0.043	0.792	0.087	0.052	0.564
KAKLAKKALAKLL	0.995	0.978	0.971	0.957	0.993	0.925	0.005
FAKALAKLAKKLL	0.986	0.985	0.954	0.994	0.993	0.992	0.002
GKWKKILGHLIR	0.977	0.009	0.979	0.006	0.006	0.006	0.294
GKWMSLLKHIWK	0.993	0.009	0.996	0.010	0.017	0.010	0.029
FAKKLAKLAKKALAL	0.986	0.989	0.925	0.993	0.994	0.992	0.002

Table 5.6: The prediction results of the top 10 peptides.

5.4 Discussion

In the field of anti-cancer peptide research, there is a growing demand for computational methods capable of efficiently identifying ACPs and categorizing them into specific functional types. In this section, we introduce DUO-ACP, a deep learning-based method with a dual capability for ACP identification and activity type classification. DUO-ACP distinguishes itself with its unique approach to learning protein sequence representation. It employs protein sequence pre-training to acquire global features and employs an embedding module for learning local features from ACP sequences. Furthermore, we introduce a two-step training process to balance the initial state of the two representation modules. Additionally, we apply an ensemble learning approach to enhance its performance and robustness. Through rigorous experimentation, we have demonstrated its superior performance in both ACP identification and classification tasks. DUO-ACP achieves an ACP identification accuracy of 89.5% and a Macro-averaged AUROC of 88.6% in ACP functional type classification, surpassing all existing methods. Finally, a case study is conducted on a newly curated dataset. A peptide predicted to possess a new activity type that not shown in database, indicating the potential of DUO-ACP for discovering novel ACPs or new functional types.

Chapter 6

Novel Fine-tuning Strategy on Pre-trained Protein Model Enhances ACP functional Type Classfication

6.1 Introduction

Cancer has long been a major health challenge worldwide [121]. There are various established medical approaches to treating cancer, such as chemotherapy, radiation therapy, and surgical interventions. Despite their effectiveness, these methods often come with significant drawbacks, including side effects and varying degrees of success depending on the cancer stage and type [149]. Recently, the development of ACPs has emerged as a promising new strategy in both cancer diagnosis and treatment. ACPs offer several benefits compared to traditional chemotherapy methods, including higher specificity, reduced side effects, and potential effectiveness in drug-resistant cancer forms. Research indicates that numerous ACPs have been developed and are currently being utilized in clinical settings. For instance, some specific ACPs has been effectively used in treating melanoma [137] and breast cancer [138]. This progress in ACP research and application has significantly contributed to the field, laying the groundwork for further advancements and potentially more effective cancer treatments in the future.

Considering the significant advantages of ACPs, it is crucial to develop computational methods for accurately identifying new ACPs. These techniques predict whether a specific peptide possesses anti-cancer properties by analyzing its amino acid sequence. AntiCP [132], as the first computational method for identifying ACPs, employs three hand-

crafted features in a Support Vector Machine (SVM) model. The advent of deep learning techniques has significantly enhanced prediction accuracy and efficiency. For instance, ACP-DL [153] employs a Long Short-Term Memory (LSTM) model to extract features from peptide sequences. ACPred-LAF [56] developed a transformer encoder-based deep learning method with learnable and adaptive embedding features. These methods improve the model's ability to learn embedding features from ACPs as well as to predict ACP properties. Currently, ACP-ODE [155], which integrates deep learning with machine learning to learn ACP features, and Unidl4biopep [38], which leverages a pre-trained model to learn peptide features, are considered state-of-the-art due to their exceptional accuracy on ACP identification.

ACP research goes further than just confirming if a peptide has anti-cancer qualities. As highlighted in prior studies [133], an individual ACP can act against multiple cancer cell lines. Therefore, an equally critical aspect is to predict which cancer types ACPs can target effectively. These predictions are crucial for the tailored use of ACPs in cancer treatment, allowing therapies to be specifically designed according to the unique features of the cancer.

Several computational methods have been developed for this task: xDeep-AcPEP [19], for instance, uses multi-task learning in a CNN framework to predict the IC50 values of ACPs across six different tissue categories. ACP-MLC [33], employs a two-level prediction model that includes one layer for identification and another for type classification. Each model are trained on distinct datasets yet sharing the same architecture. Our previous work, DUO-ACP [141], marks a significant advancement as utilizing a pre-trained protein model for generating amino acid features. Leveraging the knowledge from the pre-trained model, DUO-ACP has achieved current best results on this multi-label classification problem. ACPScanner [162] presents a two-level prediction framework that leverages a pre-trained protein model, protein secondary structure prediction, and physicochemical properties for feature representation. The LightGBM and GAT [135] algorithms are employed to learn from these features, with their outcomes subsequently combined to enhance prediction accuracy.

The notable achievements in this field are largely credited to the adoption of pre-trained protein models. Nevertheless, there's considerable potential for enhancement in optimizing the utilization of these pre-trained models. The current approach predominantly involves employing pre-trained models to extract features from amino acids or peptide sequences. In contrast, practices such as fine-tuning or applying transfer learning from pre-trained models to downstream tasks—strategies that have become standard in the NLP domain [37]—offer promising avenues for boosting model performance. An additional pathway for enhancing outcomes involves refining the training methodology. It has been

observed that models integrating pre-trained and randomly initialized components tend to disproportionately rely on the pre-trained aspect, leading to the underdevelopment of other modules and distortion of the pre-trained features [74]. While DUO-ACP introduces a dual-phase training approach to address this inconsistency, it stops short of fine-tuning the core of the pre-trained model, opting instead to keep it unchanged.

In this work, we propose a novel model that directly fine-tunes the pre-trained model, following a two-step training process. Our method begins with training the projection head, followed by fine-tuning the entire model. We evaluated our model against the current best methods using 10-fold validation on a dataset and found that it outperforms others in several key metrics. The ablation study further demonstrates the advantage of initializing the projection head over direct fine-tuning of the pre-trained model. Our model continues to show superior performance on this dataset, underscoring its effectiveness and potential in advancing the field of ACP type classification.

6.2 Methods

6.2.1 Model architecture

In our ACP-2FT, we use ESM-2 [80] as our backbone model. In our model, it connects to a linear module that consists of two Full Connect(FC) linear layer. Assume the dimension of ESM-2 is d. The input and output dimension of first FC layer are d. A TanH activation function is added after the first layer. For the second layer, the dimension of input is d and output is the number of labels.

6.2.2 Two step fine-tuning strategy

When transferring a pre-trained model, there are currently two main methods: fine-tuning, where all parameters in the pre-trained model can be trained, and linear probing, where only the last linear layers are trainable. In identification or classification tasks, fine-tuning usually achieves higher accuracy than linear probing [158], as the backbone is fine-tuned to provide better sequence representation. However, fine-tuning has its issues. Since the linear layers are not well-initialized while the backbone is pre-trained, there's a tendency for the model to focus training on the backbone rather than the linear layers. This not only results in undertrained linear layers but also causes the pre-trained model to overfit the task and distort the pre-trained sequence features, leading to catastrophic forgetting [74].

Considering the problem stems from poorly initialized linear layers, an intuitive solution is a two-step fine-tuning strategy that combines the concepts of both transfer learning methods: in the first step, apply linear probing, freezing the backbone and training only the linear layers. In the second step, implement fine-tuning, showing in Figure 6.1. Since the entire model is well-initialized, the linear layer guides the backbone to find the local minimum.

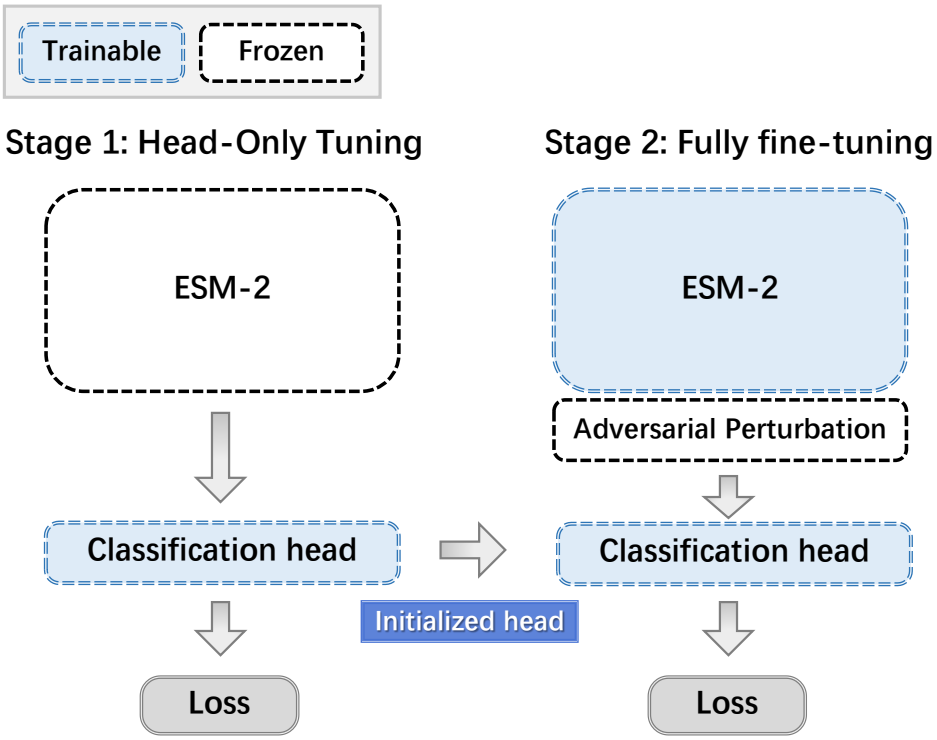

Figure 6.1: The workflow illustrating two fine-tuning strategies. In the first stage, pre-trained backbone is frozen while the classification head is trainable. In the second stage, the whole model is trainable.

6.2.3 Adversial training

When the training set is not sufficiently large, a key issue in supervised learning is the model's propensity to overfit. To address this problem, one approach is to augment the training data with additional pseudo training data. For instance, in the context of generating protein sequences, the challenge lies in accurately labeling these sequences. Unlike

in natural language processing (NLP), where minor modifications to an input sentence usually do not alter its meaning, mutations or deletions in a protein sequence can lead to fundamental changes in function. Another strategy involves incorporating adversarial examples during training. These examples, created by applying small perturbations to the input data, are intended to markedly increase the loss of these samples, thereby enhancing the model's robustness [52]. In our model, we an adversial an adversarial training strategy called Fast Gradient Method (FGM) [95]. FGM adds an adversarial perturbation to the embeddings of amino acids according to the updated gradients. The adversarial perturbation radv can be defined as Eq. 6.1.

$$r_{adv} = -\frac{\epsilon g}{\|g\|_2} \tag{6.1}$$

The gradient is defined in Eq. 6.2 as follows:

$$g = \nabla_e \log p(y|e;\hat{\theta}) \tag{6.2}$$

Let e denote the embedding of peptide sequences, and let y represent the true functional labels. The term $p(y|e;\theta)$ denotes the conditional probability of y, given e, parameterized by θ. Here, $\hat{\theta}$ is set as a constant, corresponding to the current parameters of the model. Additionally, ϵ represents the shared norm constraint, which is a critical component of the adversarial loss. Through grid search, we select ϵ as 1 in training our model.

With the adversarial training, the aim became a multi-task training. The first task is to minize the original loss without perturbation, which is a binary cross-entropy loss function, defined in Eq. 6.3.

$$L_{\text{init}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log p(y_i|x_i;\theta) + (1-y_i)\cdot \log(1-p(y_i|x_i;\theta)) \tag{6.3}$$

where N is the number of batch size, x_i,y_i are the input peptide sequence, label on the i-th sample in a batch and θ denotes the model's parameters.

The second task is to minize the adversarial loss with the adversarial perturbation, defined in Eq. 6.4

$$L_{\text{adv}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p(y_i|x_i + r_{\text{adv},i};\theta) + (1-y_i)\cdot \log(1-p(y_i|x_i + r_{\text{adv},i};\theta)) \tag{6.4}$$

where $r_{\text{adv},i}$ is the perturbation on the i-th sample. The multi-task learning is aim to minimize the sum of $L_{\text{init}}(\theta)$ and $L_{\text{adv}}(\theta)$.

6.2.4 Experiment settings

In the first step of our training, the backbone is frozen. The model is trained with a learning rate 1e-3. Then in the second step, we train our comprehensive model with a learning rate 1e-5. In both training steps, we set the batch size to 16 and limit the training to a maximum of 500 epochs. We also implement early stopping regularization, halting training if there's no improvement after 40 epochs. Binary Cross-Entropy (BCE) loss is applied for our loss function. The chosen model checkpoint is the one that records the lowest loss on the validation set. The dropout rate for backbone module is set to 0.1 and for linear layers is set to 0.5.

6.3 Results

6.3.1 Datasets

The first dataset is obtained from ACP-MLC [33]. The data is originally sourced from CancerPPD [133]. It contains totally 211 ACPs that have 7 tissue types, including colon, breast, cervix, skin, lung, prostate and blood. The detail of this dataset is listed in Table 5.2.

The second dataset is obtained from ACPScanner. Unlike Dataset I that every sequences are assigned to at least one tissue type, non-ACP sequences are included in Dataset II. The sequences with anti-cancer types are originally sourced from Database of Antimicrobial Activity and Structure of Peptides (DBAASP) [107], while the non-ACPs are obtained from negative samples from dataset built by AntiCP [2]. Dataset II obtains 701 sequences and split into training set containing 563 sequences and testing set containing 138 sequences. There are 10 anti-cancer categories include blood, breast, cervical, colon, liver, histiocyte, lung, myeloma, prostate, and ACPs not in these type. The details of training and testing set are displayed in Table 6.1.

6.3.2 Classification performance

In this section, We conducted an evaluation of our model on two datasets and compared with three methods: ACP-MLC [33], DUO-ACP [141] and ACPScanner [162].

For the the experiment on dataset I, following the approach of previous studies, we adopted a 10-fold validation. In this dataset, we compare our method with ACP-MLC

Tissue	#Entries in training set	#Entries in testing set
Blood	23	16
Breast	129	25
Cervix	104	16
Colon	47	17
Liver	66	16
Histiocyte	19	16
Lung	119	19
Myeloma	16	16
Prostate	24	16
Other-ACPs	60	14
Non-ACPs	120	30
Total	563	138

Table 6.1: The number of each labels in training set and testing set.

and DUO-ACP, whose results are obtained from their respective papers. The overall comparison results are shown in Table 6.2, with the metrics being macro-averaged. Across all metrics, ACP-2FT achieves the best current results. Compared with DUO-ACP, our AUROC has increased from 0.886 to 0.912.

We also evaluted the performances on each label. The results of the label-wise comparison are shown in Figure 6.2. For all seven labels, ACP-2FT demonstrates superior performance compared to DUO-ACP. Specifically, in predicting the 'Colon and Blood labels, ACP-2FT shows significant improvements. These two labels represent the most unbalanced categories among the seven labels. For both DUO-ACP and ACP-2FT, these labels had the lowest AUROC scores. In this experiment, ACP-2FT's better predictive results in these more challenging tasks highlight its ability to learn features from a limited number of samples.

Method	AUROC	ACC	SE	SP	F1-score	MCC
ACP-MLC	0.868	0.773	0.608	**0.854**	0.676	0.509
DUO-ACP	0.886	0.835	0.812	0.819	0.819	0.647
ACP-2FT	**0.910**	**0.855**	**0.834**	**0.854**	**0.837**	**0.690**

Table 6.2: The testing results of ACP-MLC, DUO-ACP and ACP-2FT on 10-fold validation of dataset I.

In dataset II, we compare our method with ACPScanner. In ACPScanner, two machine

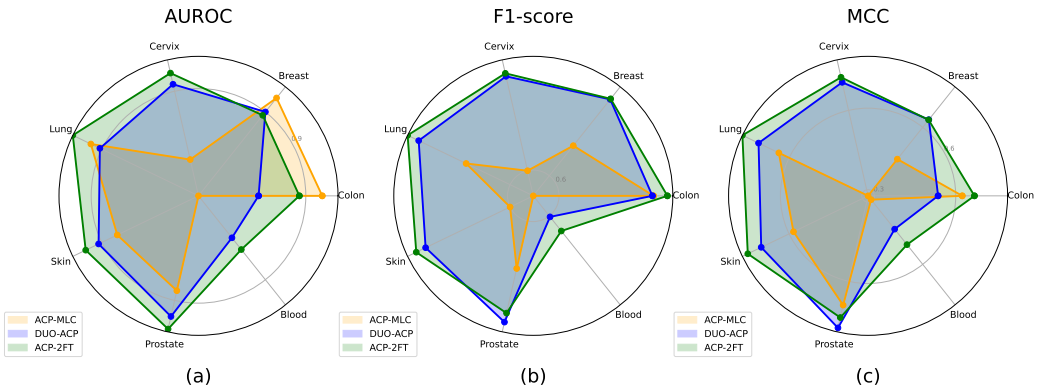

Figure 6.2: The rader chart showing the performances of ACP-MLC, DUO-ACP and ACP-2FT on 7 tissue types. (a) The AUROC of three methods on 7 tissue types. (b) The F1-score of three methods on 7 tissue types. (c) The MCC of three methods on 7 tissue types.

learninng models are trained: one is based on LightGBM model and another is based on GAT model. The model is trained on the training set and is evaluated on the independent testing set. We adopt the same training/testing for ACP-2FT and ACPScanner. Considering the training set is a more balanced dataset while in testing set negative samples got higher ratio, we set our threshold as 0.3 to balance the difference. The results of each label are macro averaged and shown in Table 6.3. ACP-2FT achives better results than the two models of ACPScanner on all metrics except sensitivity. The main reason for this is due to the threshold setting. In AUROC (Area under ROC curve) and AUPRC(Area under PR curve), ACP-2FT got significant improvement. Considering that ACPScanner also uses ESM model for feature embedding, this increase is be attributed to our specific training strategies.

The label-wise results are shown in Figure 6.3. In the comparison result for the 9 cancer tissue types, ACP-2FT has the best AUROC and F1-score for 7 cancer types. In terms of MCC, ACP-2FT got the best performance over all types. We investigated the reason.

6.3.3 ACP-2FT for ACP identification

In Dataset II, which includes samples without anti-cancer properties, ACPScanner employs a distinct binary classification model to ascertain whether a peptide exhibits anti-cancer properties. We evaluate the efficacy of ACP-2FT model that trained type classification

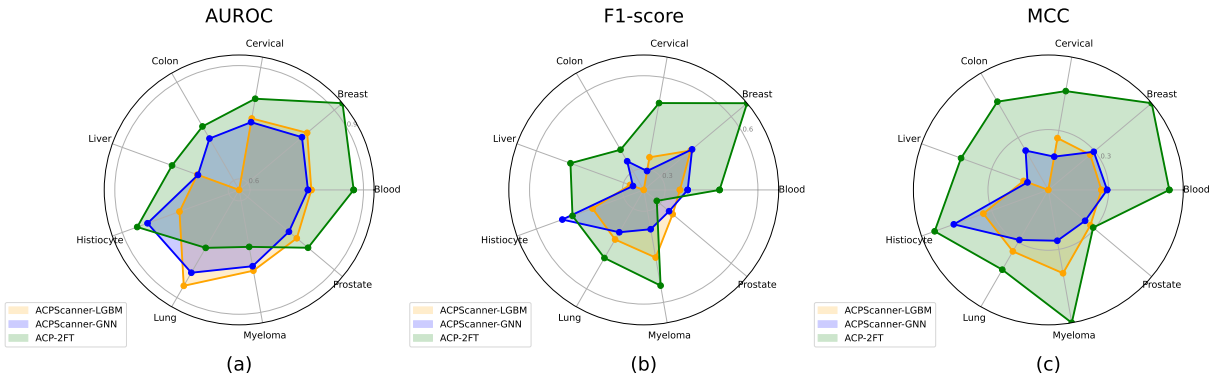

Figure 6.3: The rader chart showing the performances of two sub model of ACPScanner(ACPScanner LGBM and ACPScanner GAT) and ACP-2FT on 9 tissue types. (a) The AUROC of three methods on 9 tissue types. (b) The F1-score of three methods on 9 tissue types. (c) The MCC of three methods on 9 tissue types.

	AUROC	AUPRC	ACC	SE	SP	F1-score	MCC
ACPScanner(LGBM)	0.75	0.342	0.669	**0.725**	0.661	0.361	0.267
ACPScanner(GAT)	0.765	0.376	0.681	0.708	0.676	0.367	0.272
ACP-2FT	**0.809**	**0.567**	**0.898**	0.393	**0.968**	**0.479**	**0.478**

Table 6.3: The testing results of ACPScanner-LGBM, ACPScanner-GAT and ACP-2FT on 10-fold validation of dataset II.

labels in identifying ACPs using this dataset. For each sample, we determine the likelihood of it being an ACP by selecting the highest probability across all labels. The results of our model and binary classification model of ACPScanner, as detailed in Table 6.4. The comparison reveals that ACPScanner's LGBM model generally outperforms our model. Nonetheless, even without being explicitly trained on a binary model, ACP-2FT demonstrates comparable effectiveness to ACPScanner. This underscores the feasibility of employing a single model for both the identification and classification of ACPs, highlighting its versatile potential.

	AUROC	ACC	F1-score	MCC
LGBM	**0.983**	0.913	0.941	**0.797**
GAT	0.979	0.906	0.937	0.772
ACP-2FT	0.981	**0.92**	**0.947**	0.791

Table 6.4: The performances comparison of ACP identification of three models: two sub model of ACPScanner(ACPScanner-LGBM, ACPScanner-GAT) and ACP-2FT.

6.3.4 Strategies that enhance performance

In this section, we evaluate the impact of two-step training on enhancing the model's performance. Three training strategies: linear probing, fine-tuning, and two-step training are tested on Dataset I. The results for AUROC and AUPRC of each tissue type are presented in Figure 6.4. We observed that linear probing yields the least favorable performance, as it does not alter the backbone model. While fine-tuning and two-step training demonstrate comparable performance across several tissue types, the two-step training strategy exhibits distinct advantages, particularly for tissue types such as skin and colon. This leads to a superior macro-averaged AUPRC of 90.3%, surpassing the 88.8% achieved through fine-tuning.

In our model, another crucial training strategy employed is adversarial training. The fast gradient method introduces perturbations into the feature embeddings alongside a multi-task loss to enhance the model's robustness. We assess the impact of adversarial training by comparing our two-step training model, with and without FGM, on Dataset II. The results, presented in Table 6.5, clearly indicate that FGM enhances the model's robustness, leading to significant improvements across all performance metrics.

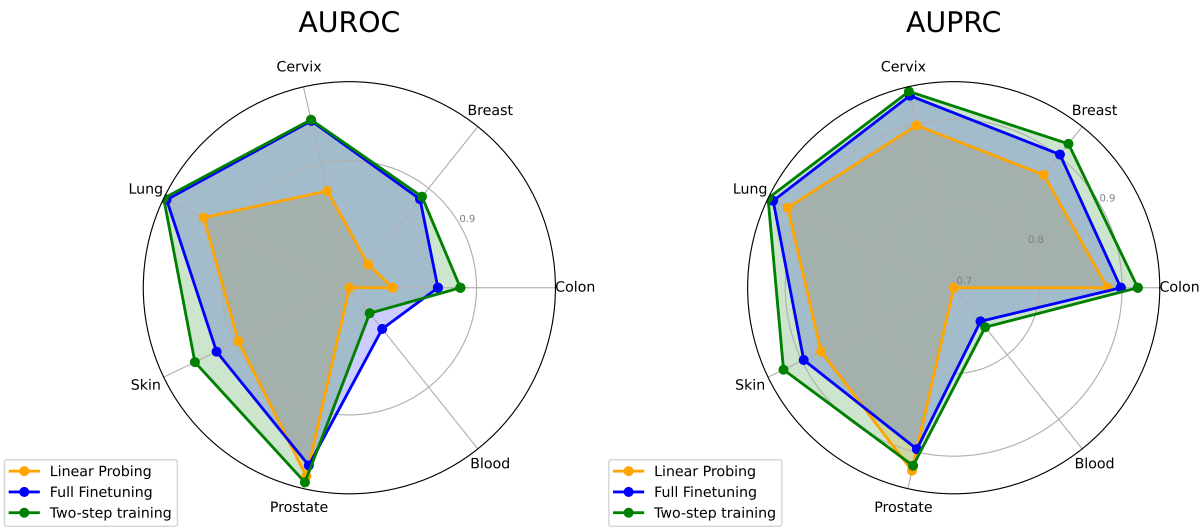

Figure 6.4: The rader chart showing the performances of the three training strategy: linear probing, fine-tuning and two-step training. (a) The AUROC of three methods on 7 tissue types. (b) The AUPRC of three methods on 7 tissue types.

Method	AUROC	AUPRC	Accuracy	F1score	MCC	Hamming Loss
ACP-2FT	0.815	0.578	0.901	0.427	0.438	0.099
ACP-2FT W/O FGM	0.791	0.527	0.888	0.322	0.340	0.111

Table 6.5: The performances comparison of ACP-2FT with and without FGM.

6.4 Discussion

In the discovery of novel anticancer peptides, it is crucial not only to determine whether a peptide possesses anticancer properties but also to identify the specific tissue types it targets. In this section, we introduce ACP-2FT, a model designed to enhance the classification of functional types. ACP-2FT undergoes direct fine-tuning from a pre-trained protein language model, employing a two-step strategy along with a fast gradient method for adversarial training. Across two distinct datasets, our method outperforms existing approaches. We further validate our model through ablation studies, which reveal that our training strategies significantly contribute to improving the model's effectiveness and robustness.

Chapter 7

Conclusion

7.1 Summary

This thesis focuses on the development of deep learning-based methods for protein sequence identification and function prediction. Despite the rapid advancements in deep learning technologies, we argue that modeling methods for bioinformatics tasks remains a complex challenge.

In this work, we initially apply deep learning techniques to the problem of mass spectrometry-based peptide identification. A novel workflow designed to enhance performance is presented in Chapter 3. We then explore the application of pre-trained models in predicting properties of antibodies, proposing a deep learning framework that leverages features from pre-trained models to improve the diagnosis of AL amyloidosis, as detailed in Chapter 4. Additionally, we introduce a new model capable of identifying ACP and classifying their functional types, which is elaborated in Chapter 5. Lastly, we enhance our classification performance by implementing a novel training strategy, discussed in Chapter 6.

7.1.1 MHC-I peptide identification

In immunology, MHC-I peptides play a crucial role by presenting cellular information, crucial for immune surveillance. Identifying these peptides' sequences remains challenging due to their non-tryptic nature and potential somatic mutations. Chapter 3 introduces

a computational workflow specifically designed for MHC-I peptide identification. Our experiments demonstrate that our method, NeoMS, outperforms existing methodologies in identifying conventional peptides. Notably, this is the inaugural study focused on identifying MHC-I peptides with single amino acid mutations. The mutated peptides we identified exhibit distribution patterns akin to regular peptides and demonstrate high binding affinities to MHC-I, underscoring the authenticity of these peptides.

7.1.2 AL amyloidosis diagnosing

Antibodies are crucial for protecting the human body from external diseases due to their ability to bind antigens. However, structural changes due to mutations can lead to AL amyloidosis, a serious condition. Chapter 4 introduces a novel approach for the early diagnosis of AL amyloidosis, focusing on the antibody light chain. This method addresses the challenge of the limited size of AL peptide datasets by leveraging a pre-trained antibody sequence model. Utilizing features extracted from this pre-trained model, our DeepAL framework sets a new standard in accurately identifying light chains that could potentially lead to AL amyloidosis.

7.1.3 ACP identification

Cancer continues to be one of the most devastating diseases globally. ACPs represent a novel therapeutic approach, offering benefits such as effective penetration and reduced drug resistance. In Chapter 5, we introduce an innovative method for ACP discovery, capable of identifying ACPs from a pool of candidate peptides. We propose a novel deep learning architecture that integrates protein knowledge with specific insights into ACP characteristics. Our experimental results demonstrate that the DUO-ACP model exhibits outstanding performance, underscoring its potential applicability in the field of drug discovery.

7.1.4 ACP type classification

Current research on ACPs primarily focuses on determining their anti-cancer properties. However, predicting the functional type of ACPs—specifically, identifying the types of cancer cells they can target—is equally critical. In Chapter 6, we leverage a pre-trained protein model combined with an innovative two-step training strategy and an adversial training. Compared to previous methodologies, including the approach discussed in Chapter 5, our proposed ACP-2FT model demonstrates superior performance across a majority

of tissue types. This advancement represents a significant step forward in the personalized and targeted application of ACPs in cancer treatment.

7.2 Future directions

In this thesis, we have introduced various methodologies for protein sequence identification and protein function prediction, contributing significantly to the field's advancement. Despite the progress made, our research is not without limitations. These constraints pave the way for several future research directions, outlined as follows:

7.2.1 End-to-end peptide identification model

The NeoMS workflow integrates a database search tool with various deep learning methodologies, demonstrating proficiency in identifying regular peptides from mass spectrometry data. However, the training of these models is conducted independently, suggesting for enhanced performance through the development of a unified end-to-end model. Such a model would cohesively learn from both peptide sequences and mass spectrometry data, potentially unveiling underlying relationships between them. This integrative approach could lead to significant advancements in the accuracy and efficiency of peptide identification.

7.2.2 Complete identification of MHC peptide

Even though NeoMS is the first method that can identify peptides with a single mutation under strict statistic restriction, there is still research value in identifying peptides with multiple mutations, as they can provide valuable insights into tumor heterogeneity and immune response. Meanwhile, MHC-II molecules primarily interact with immune cells, and their peptide length ranges from 13 to 25 amino acids. The longer length of MHC-II peptides adds complexity to the identification of mutated peptides. Future work should focus on improving NeoMS to enable the identification of mutated peptides in the context of MHC-II. The *de novo* sequencing method has shown its capability of generating personalized database. The future work can apply *de novo* sequencing database for peptides with multiple mutations and MHC-II peptides.

7.2.3 Exploring structural information for AL prediction

While DeepAL's exceptional performance is evident among existing methods, there exist promising avenues for future exploration and refinement: DeepAL model exclusively relies on sequence-based information. However, the occurrence of protein amyloidosis predominantly arises from misfolding processes intricately tied to protein structure. Notably, LICTOR [50] has emphasized the importance of incorporating structural information to address this issue. Exploring the integration of both sequence and structural information within our framework holds the potential for augmenting DeepAL's predictive capabilities.

7.2.4 ACP sequence design

Both our model and current methods are discriminative models reliant on pre-specified sequences. The existing pipeline for identifying new ACPs is marked by a relatively high cost associated with the screening of candidate peptides. Developing a computational strategy to autonomously generate high-confidence ACP candidates presents a significant challenge. However, recent advancements in decoder-based pLMs leveraging the GPT architecture, such as ProGen [88] and ProtGPT2 [46], have shown potential in synthesizing proteins with designated characteristics. Future research endeavors could explore fine-tuning these models to facilitate the generation of novel ACP candidates, potentially streamlining the discovery process and reducing associated costs.

References

[1] Ravali Adusumilli and Parag Mallick. Data conversion with proteowizard msconvert. *Proteomics: methods and protocols*, pages 339–368, 2017.

[2] Piyush Agrawal, Dhruv Bhagat, Manish Mahalwal, Neelam Sharma, and Gajendra PS Raghava. Anticp 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics*, 22(3):bbaa153, 2021.

[3] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[4] Fatemeh Araste, Khalil Abnous, Maryam Hashemi, Seyed Mohammad Taghdisi, Mohammad Ramezani, and Mona Alibolandi. Peptide-based targeted therapeutics: Focus on cancer treatment. *Journal of controlled release*, 292:141–162, 2018.

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

[6] Michal Bassani-Sternberg, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann. Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & Cellular Proteomics*, 14(3):658–673, 2015.

[7] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[8] Sadik Bhattarai, Kyu-Sik Kim, Hilal Tayara, and Kil To Chong. Acp-ada: A boosting method with data augmentation for improved prediction of anticancer peptides. *International Journal of Molecular Sciences*, 23(20):12194, 2022.

[9] Leon Bichmann, Annika Nelde, Michael Ghosh, Lukas Heumos, Christopher Mohr, Alexander Peltzer, Leon Kuchenbecker, Timo Sachsenberg, Juliane S Walz, Stefan Stevanovic, et al. Mhcquant: automated and reproducible data analysis for immunopeptidomics. *Journal of proteome research*, 18(11):3876–3884, 2019.

[10] Luis M Blancas-Mejía and Marina Ramirez-Alvarado. Systemic amyloidoses. *Annual review of biochemistry*, 82:745–774, 2013.

[11] Kip Bodi, Tatiana Prokaeva, Brian Spencer, Maurya Eberhard, Lawreen H Connors, and David C Seldin. Al-base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. *Amyloid*, 16(1):1–8, 2009.

[12] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[13] Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, Lennart Martens, and Sven Degroeve. Deeplc can predict retention times for peptides that carry as-yet unseen modifications. *Nature methods*, 18(11):1363–1369, 2021.

[14] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[16] Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz, and Małgorzata Kotulska. Amyloidogenic motifs revealed by n-gram analysis. *Scientific reports*, 7(1):12961, 2017.

[17] Phasit Charoenkwan, Sakawrat Kanthawong, Chanin Nantasenamat, Md Mehedi Hasan, and Watshara Shoombuatong. iamy-scm: Improved prediction and analysis

of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics*, 113(1):689–698, 2021.

[18] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

[19] Jiarui Chen, Hong Hin Cheong, and Shirley WI Siu. xdeep-acpep: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *Journal of chemical information and modeling*, 61(8):3789–3803, 2021.

[20] Xiangan Chen, Wen Zhang, Xiaofei Yang, Chenhong Li, and Hengling Chen. Acp-da: improving the prediction of anticancer peptides using data augmentation. *Frontiers in Genetics*, 12:698477, 2021.

[21] Hao Chi, Haifeng Chen, Kun He, Long Wu, Bing Yang, Rui-Xiang Sun, Jianyun Liu, Wen-Feng Zeng, Chun-Qing Song, Si-Min He, et al. pnovo+: de novo peptide sequencing using complementary hcd and etd tandem mass spectra. *Journal of proteome research*, 12(2):615–625, 2013.

[22] Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, Rui-Min Wang, Xiu-Nan Niu, Yue-He Ding, Yao Zhang, et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature biotechnology*, 36(11):1059–1061, 2018.

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

[24] Oscar Conchillo-Solé, Natalia S de Groot, Francesc X Avilés, Josep Vendrell, Xavier Daura, and Salvador Ventura. Aggrescan: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides. *BMC bioinformatics*, 8:1–17, 2007.

[25] UniProt Consortium. The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl_1):D190–D195, 2007.

[26] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

[27] Jürgen Cox. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology*, 41(1):33–43, 2023.

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

[29] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of proteome research*, 10(4):1794–1805, 2011.

[30] Srinivasan Damodaran. Amino acids, peptides and proteins. *Fennema's food chemistry*, 4:425–439, 2008.

[31] Arthur Declercq, Robbin Bouwmeester, Aurélie Hirschler, Christine Carapito, Sven Degroeve, Lennart Martens, and Ralf Gabriels. Ms2rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. *Molecular & Cellular Proteomics*, 21(8), 2022.

[32] Sven Degroeve and Lennart Martens. Ms2pip: a tool for ms/ms peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, 2013.

[33] Hua Deng, Meng Ding, Yimeng Wang, Weihua Li, Guixia Liu, and Yun Tang. Acpmlc: A two-level prediction engine for identification of anticancer peptides and multilabel classification of their functional types. *Computers in Biology and Medicine*, 158:106844, 2023.

[34] Eric W Deutsch, Nuno Bandeira, Yasset Perez-Riverol, Vagisha Sharma, Jeremy J Carver, Luis Mendoza, Deepti J Kundu, Shengbo Wang, Chakradhar Bandla, Selvakumar Kamatchinathan, et al. The proteomexchange consortium at 10 years: 2023 update. *Nucleic acids research*, 51(D1):D1539–D1548, 2023.

[35] Arun Devabhaktuni, Sarah Lin, Lichao Zhang, Kavya Swaminathan, Carlos G Gonzalez, Niclas Olsson, Samuel M Pearlman, Keith Rawson, and Joshua E Elias. Taggraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nature biotechnology*, 37(4):469–479, 2019.

[36] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[37] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[38] Zhenjiao Du, Xingjian Ding, Yixiang Xu, and Yonghui Li. Unidl4biopep: a universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 24(3):bbad135, 2023.

[39] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207–214, 2007.

[40] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

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

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

[43] Carlos Família, Sarah R Dennison, Alexandre Quintas, and David A Phoenix. Prediction of peptide and protein propensity for amyloid formation. *PloS one*, 10(8):e0134679, 2015.

[44] Guanwen Feng, Hang Yao, Chaoneng Li, Ruyi Liu, Rungen Huang, Xiaopeng Fan, Ruiquan Ge, and Qiguang Miao. Me-acp: Multi-view neural networks with ensemble model for identification of anticancer peptides. *Computers in Biology and Medicine*, 145:105459, 2022.

[45] Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–1306, 2004.

[46] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

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

[48] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[49] Sergiy O Garbuzynskiy, Michail Yu Lobanov, and Oxana V Galzitskaya. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332, 2010.

[50] Maura Garofalo, Luca Piccoli, Margherita Romeo, Maria Monica Barzago, Sara Ravasio, Mathilde Foglierini, Milos Matkovic, Jacopo Sgrignani, Raoul De Gasparo, Marco Prunotto, et al. Machine learning analyses of antibody somatic mutations predict immunoglobulin light chain toxicity. *Nature Communications*, 12(1):3532, 2021.

[51] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 16(6):509–518, 2019.

[52] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[53] Niels Gregersen, Peter Bross, Søren Vang, and Jane H Christensen. Protein misfolding and human disease. *Annu. Rev. Genomics Hum. Genet.*, 7:103–124, 2006.

[54] Martha Grogan, Angela Dispenzieri, and Morie A Gertz. Light-chain cardiac amyloidosis: strategies to promote early diagnosis and cardiac response. *Heart*, 103(14):1065–1072, 2017.

[55] Qinghua He, Xianhan Jiang, Xinke Zhou, and Jinsheng Weng. Targeting cancers through tcr-peptide/mhc interactions. *Journal of hematology & oncology*, 12(1):1–17, 2019.

[56] Wenjia He, Yu Wang, Lizhen Cui, Ran Su, and Leyi Wei. Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. *Bioinformatics*, 37(24):4684–4693, 2021.

[57] Steven Henikoff and Jorja G Henikoff. Performance evaluation of amino acid substitution matrices. *Proteins: Structure, Function, and Bioinformatics*, 17(1):49–61, 1993.

[58] Mark P Jedrychowski, Edward L Huttlin, Wilhelm Haas, Mathew E Sowa, Ramin Rad, and Steven P Gygi. Evaluation of hcd-and cid-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Molecular & Cellular Proteomics*, 10(12), 2011.

[59] Zhi Jin, Sheng Xu, Xiang Zhang, Tianze Ling, Nanqing Dong, Wanli Ouyang, Zhiqiang Gao, Cheng Chang, and Siqi Sun. Contranovo: A contrastive learning approach to enhance de novo peptide sequencing. *arXiv preprint arXiv:2312.11584*, 2023.

[60] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

[61] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[62] Assaf Kacen, Aaron Javitt, Matthias P Kramer, David Morgenstern, Tomer Tsaban, Merav D Shmueli, Guo Ci Teo, Felipe da Veiga Leprevost, Eilon Barnea, Fengchao Yu, et al. Post-translational modifications reshape the antigenic landscape of the mhc i immunopeptidome in tumors. *Nature biotechnology*, 41(2):239–251, 2023.

[63] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923–925, 2007.

[64] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(01):29–34, 2008.

[65] Takahiro Karasaki, Kazuhiro Nagayama, Hideki Kuwano, Jun-ichi Nitadori, Masaaki Sato, Masaki Anraku, Akihiro Hosoi, Hirokazu Matsushita, Masaki Takazawa, Osamu Ohara, et al. Prediction and prioritization of neoantigens: integration of rna sequencing data with whole-exome sequencing. *Cancer science*, 108(2):170–177, 2017.

[66] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[67] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[68] Sangtae Kim and Pavel A Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5(1):1–10, 2014.

[69] Aaron A Klammer and Michael J MacCoss. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *Journal of proteome research*, 5(3):695–700, 2006.

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

[71] Masoumeh Kordi, Zeynab Borzouyi, Saideh Chitsaz, Mohammad hadi Asmaei, Robab Salami, and Maryam Tabarzad. Antimicrobial peptides with anticancer activity: Today status, trends and their computational design. *Archives of Biochemistry and Biophysics*, page 109484, 2022.

[72] Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M Deane, and Konrad Krawczyk. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

[73] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.

[74] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[75] Himanshu Kumar, Taro Kawai, and Shizuo Akira. Pathogen recognition in the innate immune response. *Biochemical Journal*, 420(1):1–16, 2009.

[76] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[77] Jinwoo Leem, Laura S Mitchell, James HR Farmery, Justin Barton, and Jacob D Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), 2022.

[78] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[79] Kai Li, Antrix Jain, Anna Malovannaya, Bo Wen, and Bing Zhang. Deeprescore: leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics*, 20(21-22):1900334, 2020.

[80] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[81] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[82] Kaiyuan Liu, Sujun Li, Lei Wang, Yuzhen Ye, and Haixu Tang. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical chemistry*, 92(6):4275–4283, 2020.

[83] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[84] Yong-Qiang Liu, Xiao-Lu Wang, Dan-Hua He, and Yong-Xian Cheng. Protection against chemotherapy-and radiotherapy-induced side effects: A review based on the mechanisms and therapeutic opportunities of phytochemicals. *Phytomedicine*, 80:153402, 2021.

[85] Zhibin Lv, Feifei Cui, Quan Zou, Lichao Zhang, and Lei Xu. Anticancer peptides prediction with deep representation learning features. *Briefings in bioinformatics*, 22(5):bbab008, 2021.

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

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

[88] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[89] Balachandran Manavalan, Shaherin Basith, Tae Hwan Shin, Sun Choi, Myeong Ok Kim, and Gwang Lee. Mlacp: machine-learning-based prediction of anticancer peptides. *Oncotarget*, 8(44):77121, 2017.

[90] Susan Marqus, Elena Pirogova, and Terrence J Piva. Evaluation of the use of therapeutic peptides for cancer treatment. *Journal of biomedical science*, 24(1):1–15, 2017.

[91] Sebastian Maurer-Stroh, Maja Debulpaep, Nico Kuemmerer, Manuela Lopez De La Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods*, 7(3):237–242, 2010.

[92] Richard Mayeux. Biomarkers: potential uses and limitations. *NeuroRx*, 1:182–188, 2004.

[93] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[94] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[95] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[96] Roger E Moore, Mary K Young, and Terry D Lee. Qscore: an algorithm for evaluating sequest database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[97] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.

[98] Mengting Niu, Yanjuan Li, Chunyu Wang, and Ke Han. Rfamyloid: a web server for predicting amyloid proteins. *International journal of molecular sciences*, 19(7):2071, 2018.

[99] MIFJ Oerlemans, KHG Rutten, MC Minnema, RAP Raymakers, FW Asselbergs, and Nicolaas de Jonge. Cardiac amyloidosis: the need for early diagnosis. *Netherlands Heart Journal*, 27:525–536, 2019.

[100] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[101] Si-sheng Ou-Yang, Jun-yan Lu, Xiang-qian Kong, Zhong-jie Liang, Cheng Luo, and Hualiang Jiang. Computational drug discovery. *Acta Pharmacologica Sinica*, 33(9):1131–1140, 2012.

[102] Timothy J O'Donnell, Alex Rubinsteyn, and Uri Laserson. Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell systems*, 11(1):42–48, 2020.

[103] Giovanni Palladini, Paolo Milani, and Giampaolo Merlini. Management of al amyloidosis in 2020. *Blood, The Journal of the American Society of Hematology*, 136(23):2620–2627, 2020.

[104] Miao Peng, Yongzhen Mo, Yian Wang, Pan Wu, Yijie Zhang, Fang Xiong, Can Guo, Xu Wu, Yong Li, Xiaoling Li, et al. Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular cancer*, 18(1):1–14, 2019.

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

[106] Maria M Picken. The pathology of amyloidosis in classification: a review. *Acta haematologica*, 143(4):322–334, 2020.

[107] Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research*, 49(D1):D288–D297, 2021.

[108] Tanya L Poshusta, Laura A Sikkink, Nelson Leung, Raynell J Clark, Angela Dispenzieri, and Marina Ramirez-Alvarado. Mutations in specific structural regions of immunoglobulin light chains are associated with free light chain levels in patients with al amyloidosis. *PloS one*, 4(4):e5169, 2009.

[109] Rui Qiao, Ngoc Hieu Tran, Lei Xin, Xin Chen, Ming Li, Baozhen Shan, and Ali Ghodsi. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3(5):420–425, 2021.

[110] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[111] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[112] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[113] Bing Rao, Chen Zhou, Guoying Zhang, Ran Su, and Leyi Wei. Acpred-fuse: fusing multi-view information improves the prediction of anticancer peptides. *Briefings in bioinformatics*, 21(5):1846–1855, 2020.

[114] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[115] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[116] Puneet Rawat, R Prabakaran, Sandeep Kumar, and M Michael Gromiha. Exploring the sequence features determining amyloidosis in human antibody light chains. *Scientific Reports*, 11(1):13785, 2021.

[117] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[118] Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

[119] Susan C Shin and Jessica Robinson-Papp. Amyloid neuropathies. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 79(6):733–748, 2012.

[120] Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Generative language modeling for antibody design. *bioRxiv*, pages 2021–12, 2021.

[121] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023.

[122] Fabian Sievers and Desmond G Higgins. Clustal omega, accurate alignment of very large numbers of sequences. *Multiple sequence alignment methods*, pages 105–116, 2014.

[123] Jiřina Slaninová, Veronika Mlsová, Hilda Kroupová, Lukáš Alán, Tereza Tmová, Lenka Monincová, Lenka Borovičková, Vladimír Fučík, and Václav Čeřovskỳ. Toxicity study of antimicrobial peptides from wild bee venom and their analogs toward mammalian normal and cancer cells. *Peptides*, 33(1):18–26, 2012.

[124] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[125] Huawei Tao, Shuai Shan, Hongliang Fu, Chunhua Zhu, and Boye Liu. An augmented sample selection framework for prediction of anticancer peptides. *Molecules*, 28(18):6680, 2023.

[126] J Alex Taylor and Richard S Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 11(9):1067–1075, 1997.

[127] Matthew The, Michael J MacCoss, William S Noble, and Lukas Käll. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of the American Society for Mass Spectrometry*, 27:1719–1727, 2016.

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

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

[130] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017.

[131] Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36, 2024.

[132] Atul Tyagi, Pallavi Kapoor, Rahul Kumar, Kumardeep Chaudhary, Ankur Gautam, and GPS Raghava. In silico models for designing and discovering novel anticancer peptides. *Scientific reports*, 3(1):2984, 2013.

[133] Atul Tyagi, Abhishek Tuknait, Priya Anand, Sudheer Gupta, Minakshi Sharma, Deepika Mathur, Anshika Joshi, Sandeep Singh, Ankur Gautam, and Gajendra PS Raghava. Cancerppd: a database of anticancer peptides and proteins. *Nucleic acids research*, 43(D1):D837–D843, 2015.

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

[135] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[136] Ian Walsh, Flavio Seno, Silvio CE Tosatto, and Antonio Trovato. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, 42(W1):W301–W307, 2014.

[137] Che Wang, Yin-Wang Chen, Liang Zhang, Xian-Ge Gong, Yang Zhou, and De-Jing Shang. Melanoma cell surface-expressed phosphatidylserine as a therapeutic target for cationic anticancer peptide, temporin-1cea. *Journal of Drug Targeting*, 24(6):548–556, 2016.

[138] Che Wang, Yang Zhou, Song Li, HuiBing Li, LiLi Tian, He Wang, and DeJing Shang. Anticancer mechanisms of temporin-1cea, an amphipathic α-helical antimicrobial peptide, in bcap-37 human breast cancer cells. *Life sciences*, 92(20-21):1004–1014, 2013.

[139] Danqing Wang, YE Fei, and Hao Zhou. On pre-training language model for antibody. In *The Eleventh International Conference on Learning Representations*, 2022.

[140] Shaokai Wang and Bin Ma. Deep learning boosted amyloidosis diagnosis. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 57–62. IEEE, 2023.

[141] Shaokai Wang and Bin Ma. Anti-cancer peptides identification and activity type classification with protein sequence pre-training. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[142] Shaokai Wang, Ming Zhu, and Bin Ma. Neoms: Identification of novel mhc-i peptides with tandem mass spectrometry. In *International Symposium on Bioinformatics Research and Applications*, pages 280–291. Springer, 2023.

[143] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[144] Leyi Wei, Chen Zhou, Huangrong Chen, Jiangning Song, and Ran Su. Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, 34(23):4007–4016, 2018.

[145] Bo Wen, Kai Li, Yun Zhang, and Bing Zhang. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature communications*, 11(1):1759, 2020.

[146] Mathias Wilhelm, Daniel P Zolg, Michael Graber, Siegfried Gessulat, Tobias Schmidt, Karsten Schnatbaum, Celina Schwencke-Westphal, Philipp Seifert, Niklas de Andrade Krätzig, Johannes Zerweck, et al. Deep learning boosts sensitivity of

mass spectrometry-based immunopeptidomics. *Nature communications*, 12(1):3346, 2021.

[147] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[148] Dongdong Wu, Yanfeng Gao, Yuanming Qi, Lixiang Chen, Yuanfang Ma, and Yanzhang Li. Peptide-based cancer therapy: opportunity and challenge. *Cancer letters*, 351(1):13–22, 2014.

[149] Mingfeng Xie, Dijia Liu, and Yufeng Yang. Anti-cancer peptides: Classification, mechanism of action, reconstruction and modification. *Open biology*, 10(7):200004, 2020.

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

[151] Kevin L Yang, Fengchao Yu, Guo Ci Teo, Kai Li, Vadim Demichev, Markus Ralser, and Alexey I Nesvizhskii. Msbooster: improving peptide identification rates using deep learning-based features. *Nature Communications*, 14(1):4539, 2023.

[152] Yi Yang, Xiaohui Liu, Chengpin Shen, Yu Lin, Pengyuan Yang, and Liang Qiao. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications*, 11(1):146, 2020.

[153] Hai-Cheng Yi, Zhu-Hong You, Xi Zhou, Li Cheng, Xiao Li, Tong-Hai Jiang, and Zhan-Heng Chen. Acp-dl: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Molecular Therapy-Nucleic Acids*, 17:1–9, 2019.

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

[155] Qitong Yuan, Keyi Chen, Yimin Yu, Nguyen Quoc Khanh Le, and Matthew Chin Heng Chua. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. *Briefings in Bioinformatics*, 24(1):bbac630, 2023.

[156] Wen-Feng Zeng, Xie-Xuan Zhou, Sander Willems, Constantin Ammar, Maria Wahle, Isabell Bludau, Eugenia Voytik, Maximillian T Strauss, and Matthias Mann. Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics. *Nature Communications*, 13(1):7238, 2022.

[157] Wen-Feng Zeng, Xie-Xuan Zhou, Wen-Jing Zhou, Hao Chi, Jianfeng Zhan, and Si-Min He. Ms/ms spectrum prediction for modified peptides using pdeep2 trained by transfer learning. *Analytical chemistry*, 91(15):9724–9731, 2019.

[158] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

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

[160] Qing-Yu Zhang, Fei-Xuan Wang, Ke-Ke Jia, and Ling-Dong Kong. Natural product interventions for chemotherapy and radiotherapy-induced side effects. *Frontiers in pharmacology*, 9:1253, 2018.

[161] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

[162] Guolun Zhong and Lei Deng. Acpscanner: Prediction of anticancer peptides by integrated machine learning methodologies. *Journal of Chemical Information and Modeling*, 2024.

[163] Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, and Zhifei Zhang. pdeep: predicting ms/ms spectra of peptides with deep learning. *Analytical chemistry*, 89(23):12690–12697, 2017.

[164] Yuwei Zhou, Ziru Huang, Yushu Gou, Siqi Liu, Wei Yang, Hongyu Zhang, Anthony Mackitz Dzisoo, and Jian Huang. Ab-amy: machine learning aided amyloidogenic risk prediction of therapeutic antibody light chains. *Antibody Therapeutics*, page tbad007, 2023.