BiATNovo: A Self-Attention based Bidirectional Peptide Sequencing Method

Siyu Wu[1], Zhongzhi Luan[1,*], Zhenxin Fu[4], Qunying Wang[1] and Tiannan Guo[2,3,*]

[1] Sino-German Joint Software Institute (JSI), Beihang University, Beijing 100191, China, [2] Zhejiang Provincial Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China, [3] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China, [4] Computer Center, Peking University, Beijing 100871, China

*To whom correspondence should be addressed.

Contact: luan.zhongzhi@buaa.edu.cn; guotiannan@westlake.edu.cn

Abstract

Traditional database-based peptide sequencing methods have shortcomings in discoverability and universality, while de novo sequencing is the essential way to analyze unknown proteins and discover new peptides and proteins. Most existing de novo sequencing algorithms have the problem of accumulated deviation and unbalanced output. At the same time, some algorithms could be more suitable for Data-Independent Acquisition Mass Spectrometry (DIA-MS) data. This paper designed a bidirectional peptide sequencing method to alleviate the problems of unbalanced output and deviation accumulation in the sequencing process. The self-attention mechanism was applied to de novo sequencing to increase the interaction within the peptide sequence and the interaction between the MS/MS spectra and the peptide sequence. On the DIA-MS dataset, the peptide prediction accuracy improved by an average of 15.6% compared with the state-of-the-art method. On the DDA-MS dataset, our method achieved the best performance on partial datasets, the amino acid accuracy improved by an average of 3%. At the same time, two new evaluation scores, Position-BLEU and Alignment score, were proposed to evaluate the misalignment between the predicted sequence and the reference sequence, and the partial absence of fragment ions.

1 Introduction

Since the inception of the Human Proteome Project, Mass Spectrometry (MS)-based proteomics has gained significant momentum and has increasingly become a mainstay in life science research. Proteomics reveals the law of life movement based on protein identification techniques, combining various quantitative techniques (Aebersold and Mann, 2016). In particular, Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) has emerged as the recommended clinical and biochemical small molecule markers method. The resulting biological mass spectrometric data encompasses precursor spectra (LC-MS map) and fragment ion spectra (MS/MS spectra).

With the advent of reliable peptide sequencing methods, the inference of protein sequences has significantly improved, enabling the elucidation of protein expression and interaction relationships. Peptide sequencing is a vital tool for converting mass spectrometry (MS) data into peptide sequences, facilitating identifying and analyzing peptides and proteins. Data-Independent Acquisition (DIA) has recently gained popularity for mass spectrometry due to its high reproducibility and information integrity (Gillet, et al., 2012). The method divides the entire scanning range of m/z into several windows. It selects, fragments, and detects all fragment ions in each window at high speed cyclically.

Traditional MS-based peptide sequencing methods mainly include database-based identification methods and spectral library-based identification methods (Gessulat, et al., 2019; Meyer, et al., 2017; Pino, et al., 2020; Röst, et al., 2014; Tarn and Zeng, 2021; Ting, et al., 2017; Yang, et al., 2020). These methods offer relatively quick and sensitive identification of peptides. However, there are many defects in relying on existing databases: (1) Poor discoverability: only peptides or proteins in the reference database can be identified. New peptides or proteins cannot be discovered. (2) Poor universality: It is not universally applicable to analyzing protein samples from non-model organisms that lack reliable protein reference sequences. Furthermore, the accuracy of identification is constrained by the quality and completeness of the reference library. (3) Low utilization rate of the reference library: The search space of the reference database is constantly expanding. Consequently, a significant portion of the spectrum remains unutilized by database-based identification methods.

The identification of new peptides and proteins that are not present in sequence databases or spectral libraries is crucial in the field of proteomics. Over the past 20 years, many de novo sequencing tools have been proposed, such as NovoHMM (Fischer, et al., 2005), MSNovo (Mo, et al., 2007), UniNovo (Jeong, et al., 2013), Novor (Ma, 2015), etc. NovoHMM used Hidden Markov Model to simulate the generation process in a total probability, which supported the separation of signal and noise in MS data. MSNovo combined a new probability scoring function with a dynamic programming algorithm based on massive arrays, where the structure of massive arrays explicitly encodes all possible peptides and allows the dynamic programming algorithm to find the best peptides. Novor used a modified scoring function that captures dependencies between different ion types, where these dependencies were automatically learned using a modified offset frequency function. However, computational challenges persist due to the high noise and ambiguity in MS/MS spectra. Recent advances in deep learning have led to the development of novel de novo

sequencing algorithms to overcome these challenges. In 2017, Ngoc Hieu Tran et al. developed DeepNovo (Tran, et al., 2017), which first applied an encoder-decoder model to de novo sequencing of Data-Dependent Acquisition Mass Spectrometry (DDA-MS) data. DeepNovo used CNN to encode spectral features and LSTM to generate peptides. The prediction accuracy of the peptide had been improved to 50%-60%. pNovo3 generated candidate peptides through a traditional dynamic programming algorithm and used pDeep to predict the theoretical spectrum of each candidate peptide based on the deep learning method (Yang, et al., 2019). The learning-to-rank model has been applied to re-rank the candidate peptides depending on the feature from the pDeep model. Karunratanakul et al. developed SMSNet to improve the accuracy of de novo peptide sequencing through deep learning (Karunratanakul, et al., 2019). The key innovation in SMSNet was using a multi-step sequence mask search strategy to address the problem of locally incorrect short sequences. Qiao et al. developed PointNovo (Qiao, et al., 2021), which can robustly process DDA-MS data at any resolution level while keeping the computational complexity unchanged. In 2019, the DIA method's de novo sequencing research work appeared for the first time called DeepNovo-DIA (Tran, et al., 2019). In 2020, Ngoc Hieu Tran et al. utilized the DeepNovo-DIA algorithm for training a personal deep-learning model on the immune peptide panel of a single patient, which was then used to predict the patient's mutated neoantigens (Tran, et al., 2020). This personalized learning model discovers HLA-I and HLA-II neoantigens for personalized tumor immunotherapy, including neoantigen peptide-based vaccines, chimeric antigen receptor CAR-T, and T cell receptor TCR-T, which is of improved T-cell therapy.

In the existing work, the algorithms for de novo sequencing are still a one-way prediction, which has the phenomenon of unbalanced output and accumulation of deviations. The unbalance manifests in the high accuracy of the first several amino acids and the low accuracy of the last several amino acids. At the same time, deviation accumulation exists in the peptide sequencing. The current amino acid prediction error will affect the prediction after the current position. Meanwhile, some existing de novo sequencing methods don't apply to the DIA-MS dataset, limiting their ability to discover novel proteins.

In this study, we proposed a novel method for bidirectional peptide sequencing employing an attention mechanism. Our research yields several notable contributions, which we highlight as follows.

(1) We have designed and executed a bidirectional de novo peptide sequencing model. Candidate sequences were obtained by *Bidirectional Independent peptide sequencing* and *Bidirectional Synchronous peptide sequencing*, and then the peptide with the highest score was selected by score re-ranking. Our model achieves state-of-the-art results in the DIA-MS datasets.

(2) Our peptide sequencing method is suitable for both DDA-MS and DIA-MS datasets, and only a few parameters need to be changed during the training and testing of the model.

(3) We proposed two novel evaluation scores, *Position-BLEU* and *Alignment score*, to evaluate the misalignment between the predicted sequence and the reference sequence or the partial absence of fragment ions in the MS/MS spectra. We applied the new scores to evaluate the existing de novo sequencing algorithms.

2 Materials and Methods

Aiming at the low accuracy of de novo peptide sequencing, we designed a bidirectional prediction peptide sequencing algorithm, which solved the problem of accumulation of amino acid mass deviations and unbalanced output during one-way prediction. It will be described specifically in *Bidirectional Independent peptide sequencing*. Our approach takes extracted peptide features and MS/MS spectra as input and generates the peptide sequence as output. Fig. 1 shows the overview of our approach.

"Bidirectional prediction" mainly includes three parts: bidirectional independent prediction to generate *Ide candidate sequences*, bidirectional interactive prediction to generate *Syn candidate sequences* and re-ranking the score of all candidate sequences. The model comprises an encoder and a decoder. Encoder Global-CNN is used to extract the Global features of the MS/MS spectra. The self-attention module in the decoder will calculate the interior of the peptide sequence and the spectrum and sequence attention. The local CNN module in the decoder will calculate the candidate intensity features predicted in each step.

Transforming from raw data to the method's input requires peptide feature extraction and MS/MS spectra association. Peptide feature extraction can be performed using PEAKS (Ma, et al., 2003) or MaxQuant (Cox and Mann 2008), which provide information such as mass charge ratio (m/z), charge quantity, and retention time. Next, peptide features are associated with MS/MS spectra by filtering out the spectra whose RT field between RT Begin and RT end of the peptide feature and *PepMass* can cover the m/z of the peptide.

In the following subsections, we will introduce the fundamental model and present the central innovation of this research, namely "bidirectional", which encompasses *Bidirectional Independent peptide sequencing* and *Bidirectional Synchronous peptide sequencing*. Lastly, we will elucidate the candidate sequences scoring and re-ranking.

2.1 Attention-based peptide sequencing

The attention mechanism is widely applied in various deep-learning tasks, such as natural language processing, image classification, and speech recognition. The attention mechanism discussed in this section reflects in two parts, namely the Peptide Sequence self-attention and Spectrum-Peptide Attention.

Global CNN, as an encoder, encodes the global features of MS/MS spectra. In the Attention decoder, the Positional Encoding module encodes the peptide sequence. It is followed by Peptide Sequence Self-Attention and Spectrum-Peptide Attention. The output of the Attention decoder is then combined with the candidate intensity extracted by Local-CNN and is finally mapped to the probability of different amino acids.

Global CNN. Global CNN encodes MS/MS spectra. Like DeepNovo (Tran, et al., 2017), the global representation of the MS/MS spectra is a vector. In this work, the global representation is processed through a max pooling layer and two layers of 3D convolution kernels, and then a max pooling and fully connected layer to form a global representation of MS/MS spectra.

Peptide Sequence Self-Attention. This paper utilizes the self-attention mechanism (Vaswani, et al., 2017) to mine the relationship within the peptide sequence. Among them, the critical module is *Scaled Dot-Product Attention*. *Scaled Dot-Product Attention* maps a query and a set of key-value pairs to an output. Since the peptide sequencing task is similar to the text generation task in natural language processing, the amino acid at the current position can only be predicted using the already predicted amino acid. Q_j is the Query vector at the *jth* position, $K_{\leq j}$ is the Key vector before the *jth* position, and $V_{\leq j}$ is the Value vector before the *jth* position. Therefore, when calculating the attention mechanism, Q_j can only calculate the attention weight matrix with $K_{\leq j}$, and then multiply the attention weight matrix with $V_{\leq j}$ to obtain the amino acid vector representation of the *jth* position. The calculation formula is as (1), where d_k is the dimension of the Key vector to prevent the obtained score from being too large.

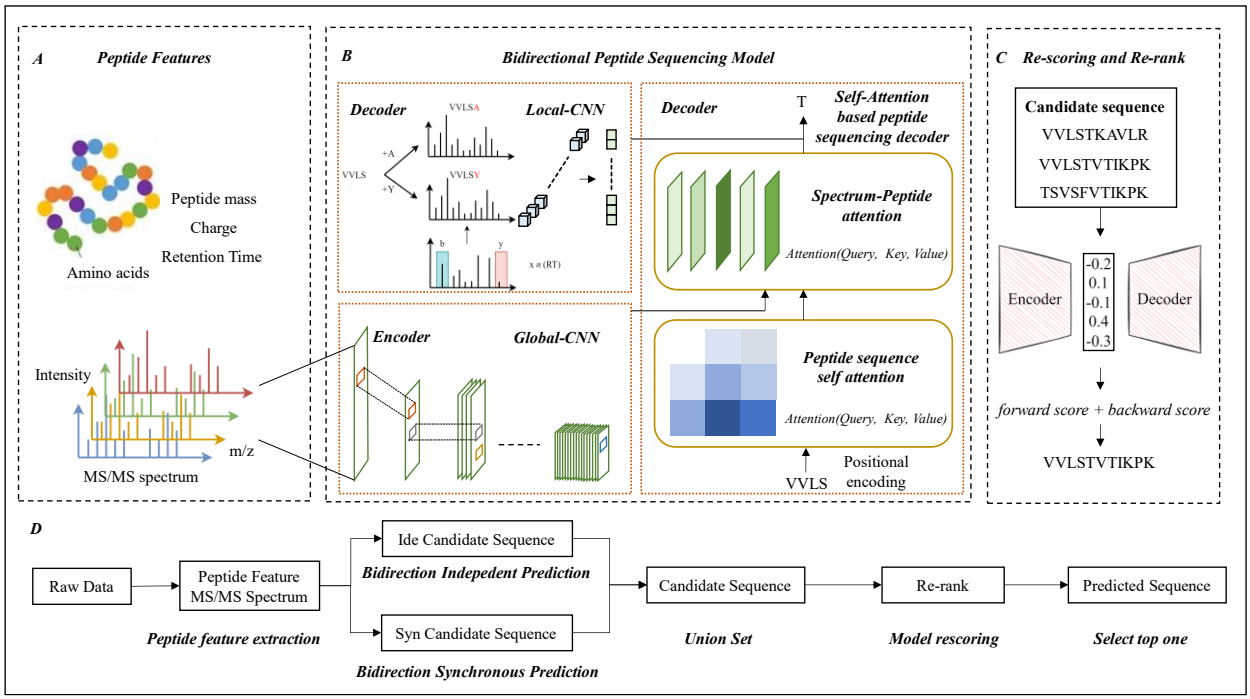

Fig. 1. The overall process of our method. Subfigure A shows the input of the method, subfigure B shows the model framework of the method, subfigure C represents the Re-scoring, and Re-ranking of the method and subfigure D is the overall flow chart of the bidirectional peptide sequencing method.

$$Attention(\vec{Q}_j, \vec{K}_{\leq j}, \vec{V}_{\leq j}) = softmax(\frac{\vec{Q}_j \vec{K}_{\leq j}}{\sqrt{d_k}})\vec{V}_{\leq j} \quad (1)$$

Spectrum-Peptide Attention. The global representation of the MS/MS spectra contains all peptide information. We aim to focus on different MS/MS spectra regions when predicting peptide sequences in different steps, giving different weights to the MS/MS spectra regions. We treat the MS/MS spectra as a picture, and the task is similar to the image caption. Specifically, the Query vector is derived from the output of the self-attention layer, whereas the Key and Value vectors come from the global features of the MS/MS spectra (i.e., the encoder's output). As illustrated in Fig. 2, the attention calculation involves matching the query with the key and value vectors to obtain the output. After that, the output is concatenated with the residual and passed through a normalization layer and a feedforward network.

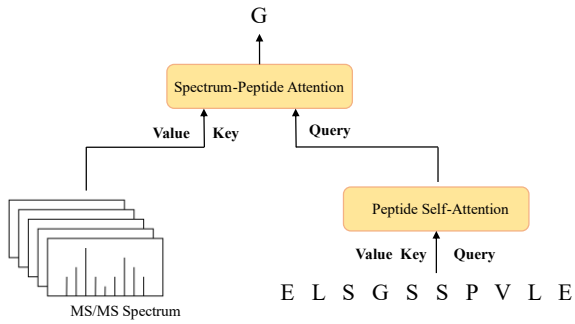

Fig. 2. Query, Key, Value representation of Spectrum-Peptide Attention.

Local CNN. This module aims to capture the correlation between the theoretically possible intensity features and the MS/MS spectra at each step. We followed the approach used in previous works such as DeepNovo (Tran, et al., 2017) and DeepNovo-DIA (Tran, et al., 2019) to achieve this. Specifically, in each step, the candidate intensity features are [26, 8, 5, 10] in DIA mode and [26, 8, 1, 10] in DDA mode. There are 26 candidate ions, consisting of 23 candidate amino acids and three identifiers (PAD: gap identifier; BOS: start identifier; EOS: end identifier). The second dimension represents the theoretically generated fragment ion species, the third represents the retention time, and the last represents the error tolerance window. Local CNN is then used to convolve the candidate intensity values, where three 3D convolution kernels, max pooling layer, and fully connected layer are used for dimension mapping.

2.2 Bidirectional independent peptide sequencing

Typically, peptide sequencing methods employ a single-direction decoding approach in the decoding stage, which involves the prediction of peptide sequences either from left to right (forward) or right to left (backward). Single-direction decoding conforms to the logic of text generation, like written text, from front to back or from back to front. When predicting the amino acid at the *t-th* position, the model can only use the amino acid information generated before the *t-th* position. Finally, the one with the highest score among the predicted peptides is selected as the final output.

However, an issue with single-directional decoding is that it can result in unbalanced output, as it cannot fully utilize the information of the future output from reverse decoding. We analyzed the single-directional decoding results of DeepNovo-DIA on the OC, UTI, and Plasma test datasets. Specifically, we calculated the accuracy of the first four amino acids and the last four amino acids both in the forward and backward prediction. The amino acids accuracy are presented in Table 1.

Table 1. Prediction imbalance appearance.

Dataset	Forward first_4	Forward_ Last_4	Backward Last 4	Backward First_4
OC	72.87	62.56	57.10	72.13
UTI	71.09	61.27	55.79	69.48
Plasma	64.26	55.34	49.70	76.90

For both directions, the accuracy of the first four acids is high. The accuracy of generated acids becomes low with the prediction going on. Generally speaking, we hope that the first part of the sequence uses the result of forward prediction and the last part uses the result of backward prediction, that is, bidirectional generation.

At the same time, during the peptide sequence prediction, due to the quality constraints of the peptide sequence, the current amino acid prediction error can accumulate and affect subsequent predictions, which is the accumulation of deviations. Based on the above problem, we designed peptide sequencing models for bidirectional prediction to alleviate the problems of deviation accumulation and unbalanced output.

The bidirectional independent prediction approach involves the model generating predictions independently from both the forward and backward directions and using all intermediate positions as junctions to form candidate prediction sequences. To ensure the accuracy of the candidate sequences, the sum of amino acid masses must be equal to the mass of the precursor ions. Any candidate sequence that does not satisfy this criterion will be filtered out. A visual representation of the bidirectional independent prediction process can be found in Fig. 3.

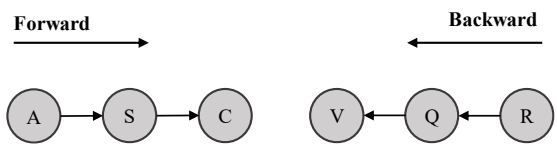

Fig. 3. Bidirectional independent prediction diagram.

Assuming that the peptide sequence for forward prediction is $forward_{peptide} = \{a_1, a_2, a_3, \dots, a_m\}$ and the peptide sequence for backward prediction is $backward_{peptide} = \{b_1, b_2, b_3, \dots, b_n\}$. Formula (2) illustrates the candidates generated by the bidirectional independent prediction approach.

In order to ensure the validity of the formed bidirectional independent prediction candidate set, specific constraints must be satisfied. To meet these constraints, it is necessary to ensure that the candidate sequence falls within the error range of the peptide sequence mass, which is expressed in formula (3). $Bi^k_{candidate}$ is the *kth* candidate sequence. $mass(precursor)$ is the mass of the peptide sequence and Δ indicates the absolute value of the change in mass. And in this candidate, both the confidence of the first few amino acids and the confidence of the last amino acids are higher, which also increases the accuracy of the predicted peptide sequence.

$$Bi_{candidate} = \{a_1, \dots, a_i, b_j, \dots, b_1\} \ i, j \epsilon (1, \max(m, n)) \tag{2}$$

$$\Delta |mass(Bi^k_{candidate}) - mass(precursor)| < 0.1 \tag{3}$$

2.3 Bidirectional synchronous peptide sequencing

The fundamental difference between Bidirectional Synchronous Prediction and Bidirectional Independent Prediction is explored in this study, inspired by the bidirectional synchronous generation of natural language (Zhang, et al., 2020; Zhou, et al., 2019). In Bidirectional Independent Prediction, both forward and backward predictions generate the amino acid based solely on the amino acids that have been predicted (i.e., historical information). In particular, synchronous prediction incorporates future information, allowing the forward and backward prediction sequences to converge better at a specific position in the middle while dynamically expanding the candidate sequence. Furthermore, Bidirectional Independent Prediction may reject certain combinations of candidate peptides that do not meet the mass requirements, while Bidirectional Synchronous Prediction has a reverse guiding effect in prediction, resulting in more qualified peptides. A schematic diagram of bidirectional synchronous prediction is shown in Fig. 4. When predicting amino acid C, forward-predicted A and S and the backward-predicted Q and R information will be used, enabling more global information to be considered in the prediction process.

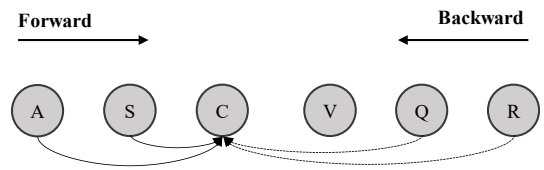

Fig. 4. Bidirectional synchronous prediction diagram.

To realize bidirectional synchronous prediction, we have transformed the Peptide Sequence Self-Attention module in subsection A into a bidirectional synchronized self-attention module. The model employs a decoder to decode left-to-right and right-to-left predictions simultaneously interactively. The necessary modification involves using bidirectional matrices for Query, Key, and Value, which are integral to the attention calculation. Specifically, attention calculation involves Query, Key, and Value vectors in the same direction and opposite directions. Finally, the hidden layer vectors obtained by forward and backward calculation are fed into a linear layer. Moreover, we concatenated the outputs from both directions to form the input of the subsequent layer. The computation of the bidirectional synchronized self-attention module is illustrated in Fig. 5. Among them, $\vec{Q}_{\leq j}$ is the Query vector in the forward prediction, $\vec{V}_{\leq j}$ and $\vec{K}_{\leq j}$ are the Value and Key vectors of the forward prediction, respectively. $\overleftarrow{Q}_{\leq j}$ is the Query vector for backward prediction, and $\overleftarrow{V}_{\leq j}$ and $\overleftarrow{K}_{\leq j}$ are the Value and Key vectors for backward prediction, respectively. $\overleftarrow{H}_j$ is the hidden layer output during forward prediction, and $\overleftarrow{H}_j$ is the hidden layer output during backward prediction. $\vec{H}^f_j$ is the hidden layer output obtained by using only historical prediction information in forward prediction, and $\vec{H}^b_j$ is the hidden layer output obtained by using only future information during forward prediction. $\overleftarrow{H}^f_j$ is the hidden layer output obtained by using only historical prediction information during backward prediction. $\overleftarrow{H}^b_j$ is the hidden layer output obtained by using only future information during backward prediction. The formula is shown in (4) to (7), The calculation formula of attention is consistent with formula (1).

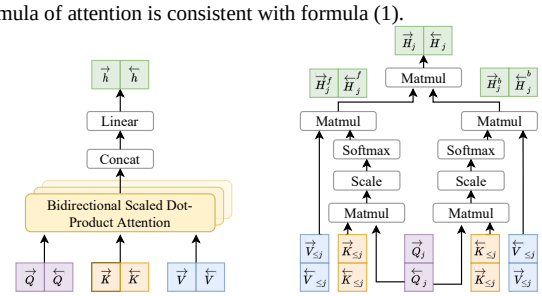

Fig. 5. Bidirectional synchronous self-attention calculation.

$$\vec{H}^f_j = Attention(\vec{Q}_j, \vec{V}_{\leq j}, \vec{K}_{\leq j}) \tag{4}$$

$$\vec{H}^b_j = Attention(\vec{Q}_j, \overleftarrow{V}_{\leq j}, \overleftarrow{K}_{\leq j}) \tag{5}$$

$$\overleftarrow{H}_j^f = Attention(\overleftarrow{Q}_j, \overleftarrow{V}_{\leq j}, \overleftarrow{K}_{\leq j}) \quad (6)$$

$$\overleftarrow{H}_j^b = Attention(\overleftarrow{Q}_j, \overrightarrow{V}_{\leq j}, \overrightarrow{K}_{\leq j}) \quad (7)$$

The hidden layer output $\overrightarrow{H}_j$ during forward prediction combines $\overrightarrow{H}_j^f$ and $\overrightarrow{H}_j^b$, and $\overrightarrow{H}_j^b$ is linearly combined with $\overrightarrow{H}_j^f$ after the ReLU activation function. $\overleftarrow{H}_j$ is calculated in the same way. The formulas are shown in (8) to (10).

$$\overrightarrow{H}_j = \overrightarrow{H}_j^f + 0.1 * ReLU(\overrightarrow{H}_j^b) \quad (8)$$

$$\overleftarrow{H}_j = \overleftarrow{H}_j^f + 0.1 * ReLU(\overleftarrow{H}_j^b) \quad (9)$$

$$ReLU(x) = \max(0, x) \quad (10)$$

The calculation of bidirectional synchronous self-attention can also be split into multiple subspaces for multi-head bidirectional synchronous self-attention calculation.

$\overrightarrow{Q}_i$, $\overrightarrow{K}_i$, and $\overrightarrow{V}_i$ are the Query, Key, and Value vectors of the *ith* subspace of forward prediction. $\overleftarrow{Q}_i$, $\overleftarrow{K}_i$, and $\overleftarrow{V}_i$ are the Query, Key, and Value vectors of the *ith* subspace for backward prediction. $\langle \overrightarrow{head}_i, \overleftarrow{head}_i \rangle$ is the hidden layer output of the *ith* subspace. The calculation of $\overrightarrow{H}_i$ and $\overleftarrow{H}_i$ is consistent with the calculation methods of formula (8) and formula (9).

During the model prediction stage, dynamic beam search is employed, , which requires dynamic adjustments in attention calculation. Unlike single-directional prediction, where the beam search size is k, bidirectional interactive prediction simultaneously decodes forward and backward predictions, necessitating a beam search size of $\left\lfloor \frac{k}{2} \right\rfloor$ for each direction. It is crucial to distinguish between the forward and backward predictions during beam search, for which we use the first identifier for Begin Of Sequence (BOS) and End Of Sequence (EOS), respectively.

The calculation of bidirectional synchronous attention involves the correspondence between the candidate sequence of forward prediction and the candidate sequence of backward prediction. In this paper, we adopt the dynamic calculation of bidirectional synchronous self-attention.

2.4 Re-rank

The bidirectional prediction peptide sequencing model described in this paper synthesizes bidirectional independent prediction and bidirectional synchronous prediction. The candidate sequence generated by the model is the union of the *Ide Candidate Sequence* and *Syn candidate Sequence*.

Upon acquiring candidate sequences for bidirectional prediction, it is imperative to identify the sequence with the highest score as the ultimate result, a process referred to as re-scoring and re-ranking. In this regard, the candidate sequence is re-entered into the model for scoring, where the scoring formula is represented by fomula (11). The final score is obtained by computing the sum of both forward and backward prediction scores, and subsequently, the peptide sequence with the highest score is selected as the output. The forward and backward prediction scores are calculated by adding the prediction scores at each position within the peptide sequence. The prediction score for each position is obtained using the *logsoftmax* probability of the prediction result at that position. The calculation formulas are shown in (12) and (13), where x_i is the amino acid at the *ith* position of the $Bi^k_{candidate}$peptide sequence. Self-Attention based bidirectional peptide sequencing method is shown as **Algorithm 1**.

$$Score(Bi^k_{candidate}) = forw(Bi^k_{candidate}) + backw(Bi^k_{candidate}) \quad (11)$$

$$forw(Bi^k_{candidate}) = \sum_i^n logsoftmax(x_i) \quad (12)$$

$$logsoftmax(x_i) = \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right) \quad (13)$$

Algorithm 1: Bidirectional Peptide Sequencing

Input: *Precursor_mass*: Peptide mass
Peptide_Spectrum: MS/MS Spectrum
mass_Aaid {AAid: mass}: Amino acid mass set
beam size: k

Output: *Predict_sequence*

```
Candidate_sequence_independent = Bidirection_Independent_Prediction (Precursor_mass, Peptide_Spectrum, mass_AAid, k)
Candidate_sequence_synchronous = Bidirection_Synchronous_Prediction(Precursor_mass, Peptide_Spectrum, mass_AAid, k)
Candidate_sequences = Set(Candidate_sequence_independent ∪ Candiate_sequence_synchronous)
for sequence in Candidate_sequences:
    tmp_score = LogSoftmax(ModelInference(sequence))
    Candidate_sequences[sequence] = tmp_score
sort(Candidate_sequences, key = score)
Predict_sequence = Candidate_sequences[0]
return Predict_sequence
```

3 Results

3.1 Datasets and experimental settings

The DIA dataset was derived from urine samples collected from voluntary patients with abdominal pain in the emergency room of Children's Hospital Boston, Massachusetts, USA (Muntel, et al., 2015). The pain sample control group was stratified into two distinct subsets, namely Pain_train and Pain_valid. The samples with ovarian cysts, urinary tract infections and plasma dataset disclosed in DeepNovo-DIA (Tran, et al., 2019) were taken as the test dataset.

The DDA-MS datasets collected in this experiment correspond to those utilized in DeepNovo (Tran, et al., 2017), comprising of nine high-resolution HCD mass spectrum datasets obtained using the Thermo Scientific Q-Exactive instrument. The original high-resolution datasets can be conveniently accessed and downloaded from the ProteomeXchange database (Vizcaíno, et al., 2014). In this experiment, we adopted the leave-one cross-validation method. The remaining eight species datasets, except one species dataset, were used to train and verify the de novo sequencing model. The training and verification sets were divided in 9:1. The test set was conducted on the set aside one species dataset.

During the training process, the focal loss function (Lin, et al., 2017) was employed, an advanced version of the cross-entropy loss function that addresses sample imbalance concerns. This function effectively reduces the emphasis on easily classifiable samples and accentuates the focus on complex samples, ultimately leading to improved training outcomes. The formula for calculating Focal Loss is shown in formula (14), where q_t refers to the probability of predicting the correct category and γ denotes the scaling factor, which defaults to 2.

$$Focal\ loss = -(1 - q_t)^{\gamma} log q_t \quad (14)$$

When solving the minimum value of the loss function, the iterative step-by-step solution can be attained using the gradient descent method. This study employed optimization techniques, including the Adam adaptive optimizer (Kingma and Ba 2014), learning rate decay, and warm-up strategies to refine the gradient descent algorithm. Meanwhile, we adopted the

Early Stopping mechanism to prevent over-fitting during training. As the number of global steps increases, the learning rate increases rapidly to pre-heat the network and then decreases slowly to find the optimal advantage of gradient descent. The accuracy rate of amino acids in the validation set steadily increased until it reached a certain value, after which no significant improvement was observed, indicating optimal network training and resulting in the cessation of model training. The changes in the learning rate of the model in the training stage and the changes in amino acid accuracy in the training and validation are clearly demonstrated in Fig. 6.

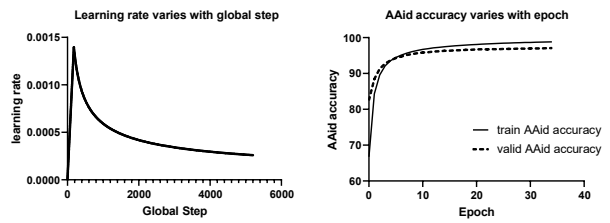

Fig. 6. Learning rate and amino acid accuracy during training.

3.2 Evaluation score design

The evaluation scores in the existing method calculate the proportion of amino acid prediction accuracy among the total prediction results. However, these scores will lead to two problems: (1) the misalignment between the prediction and the reference sequences cannot be reasonably measured. (2) Due to the partial missing of fragment ions in the MS/MS spectra, the order of some amino acids would be reversed. The reversal of amino acid sequence cannot be reasonably evaluated.

Therefore, we have proposed two novel evaluation scores to address the abovementioned challenges in this section. One is to add distance weight, named Position-BLEU, referring to the idea of BLEU (Papineni, et al., 2002), a typical evaluation score in neural machine translation. The other is to introduce a pairwise alignment algorithm named Alignment Score to calculate the similarity between the predicted sequence and the reference sequence.

Position-BLEU. We proposed improvements to the BLEU-1 gram concept by incorporating a distance weight in the matching of each position to address the issue of misalignment between the predicted and reference sequences. The calculation formula for this new approach is provided in formula (15) through (18).

$$Position\ BLEU = \frac{\sum_0^N match\ score(pre_sequence_i, trg_sequence_i)}{\sum_0^N length(pre_sequence_i)} \quad (15)$$

$$match\ score(pre_i, trg_i) = \sum_0^n \frac{1}{1+\min\left(dist_forw(pre_i^j)+dis_backw(pre_i^j)\right)} \quad (16)$$

$$dist_forw(aaid) = |trg_{idx}(aaid) - pre_idx(aaid)|\ (0, \dots, n) \quad (17)$$

$$dist_backw(aaid) = |trg_{idx}(aaid) - pre_{idx}(aaid)|(n, \dots, 0) \quad (18)$$

pre_i^j is the amino acid in position j of the *ith* prediction sequence, $dist_forw(pre_i^j)$ is the absolute distance difference between the pre_i^j in the reference sequence and the predicted sequence calculated from left to right, denoted as the forward distance of the amino acid matching. $dist_backw(pre_i^j)$ is the distance calculated from right to left, denoted as the backward distance of the amino acid matching. After calculating the sum of the forward and backward distances, the minimum value was taken to consider only the distance between pre_i^j and the nearest identical amino acid in the reference sequence. This approach considers the closest amino acid match and enables better alignment between the predicted and reference sequences.

Alignment Score. Dynamic programming is adopted based on the Smith-Waterman algorithm (Pearson 1991) of pairwise local alignment to find local similarity regions. The basic idea of the algorithm is to calculate the similarity scores of two sequences using the iterative method and store them in the score matrix. Then according to the score matrix, the optimal alignment sequence is found backtracking by the dynamic programming method. The score value of each unit in the score matrix will be calculated by the formula (19).

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j) & x_i\ aligned\ y_j \\ F(i-1,j) + d & x_i\ aligned\ gap \\ F(i,j-1) + d & y_j\ aligned\ gap \end{cases} \quad (19)$$

In this study, the mass substitution matrix of amino acids, named "mass_matrix", was generated to replace $s(x_i, y_j)$ in formula (19) for comparison purposes. The amino acid quality was arranged in ascending order of mass to construct this matrix. The difference between the maximum and minimum mass was divided into ten equal parts to determine the amino acid quality intervals corresponding to score values, as specified in the formula (20). When the amino acid at *ith* in the predicted sequence matched the amino acid at *jth* in the reference sequence, a full score of 10 was given. However, when the masses differed, a negative score was assigned based on the magnitude of the mass difference, as calculated in the formula (21). An open penalty of -1 was set for the matching score calculation.

$$per\ mass = (\max(\text{mass_AAid}) - \min(mass_AAid)) / 10 \quad (20)$$

$$mass_matrix[i,j] = -\left\lfloor \frac{(\text{mass_AAid[i]} - \text{mass_AAid[j]})}{per\ mass} \right\rfloor \quad (21)$$

3.3 Comparisons with baselines

Given the same test set, we compared the performance of DeepNovo-DIA (Tran, et al., 2019) and BiATNovo. Table 2 shows the comparison of the results, which indicate that BiATNovo outperforms the baseline on the same test set. Specifically, the accuracy of amino acid prediction increased by 6.97%, and peptide sequence prediction increased by 15.6% on average. Moreover, based on the performance of Position-BLEU and Alignment score, it is evident that BiATNovo has certain advantages in terms of prediction effectiveness. The Precision_AAid calculated the accurate proportion of amino acid prediction in all prediction results, the Recall_AAid calculated the accurate proportion of amino acid prediction in all reference sequences, and the Precision_Peptide calculated the accurate proportion of peptide sequence in all prediction results. Peptide count is the number of correctly predicted peptides.

In order to assess the sensitivity of the algorithm and the accuracy of peptide prediction, we conducted a comparative analysis of BiATNovo and the baseline approach across various score ranking percentages. This result enabled us to gauge the relative performance of the two methods under different conditions. The specific calculation is as follows: the peptide sequence and the scores in the prediction result are formed into <Key, Value> list, and then sorted based on descending order of the scores. Higher scores indicate greater likelihood of accurate peptide sequences. Then, the peptide sequences were successively selected from the Top N predictive sequences after ordering scores. Precision AAid, Precision_Pep, Position-BLEU and Alignment score indexes were changed, respectively, as shown in Fig. 7. As the percentage of score ranking increases in UTI and Plasma datasets, the difference between BiATNovo and Baseline widens gradually. In contrast, the performance of the Baseline in the OC dataset is unsatisfactory. When the score ranking percentage is low, indicating high confidence, the accuracy of BiATNovo significantly outperforms the four evaluation scores.

Table 2. Results of BiATNovo and Baseline.

Evaluation	OC		UTI		Plasma	
	baseline	BiATNovo	baseline	BiATNovo	baseline	BiATNovo
Precision_AAid(%)	68.12	74.57	65.81	70.78	63.79	66.28
Recall_AAid(%)	68.22	75.13	65.98	71.51	64.30	67.07
Peptide count	8160	9574	10359	11686	2833	3305
Precision_Pep(%)	52.40	61.47	51.06	57.60	37.40	43.64
Position-BLEU	0.6749	0.7456	0.6604	0.7117	0.6270	0.6458
Alignment score	7.5050	7.9811	7.3053	7.6583	7.1192	7.2317

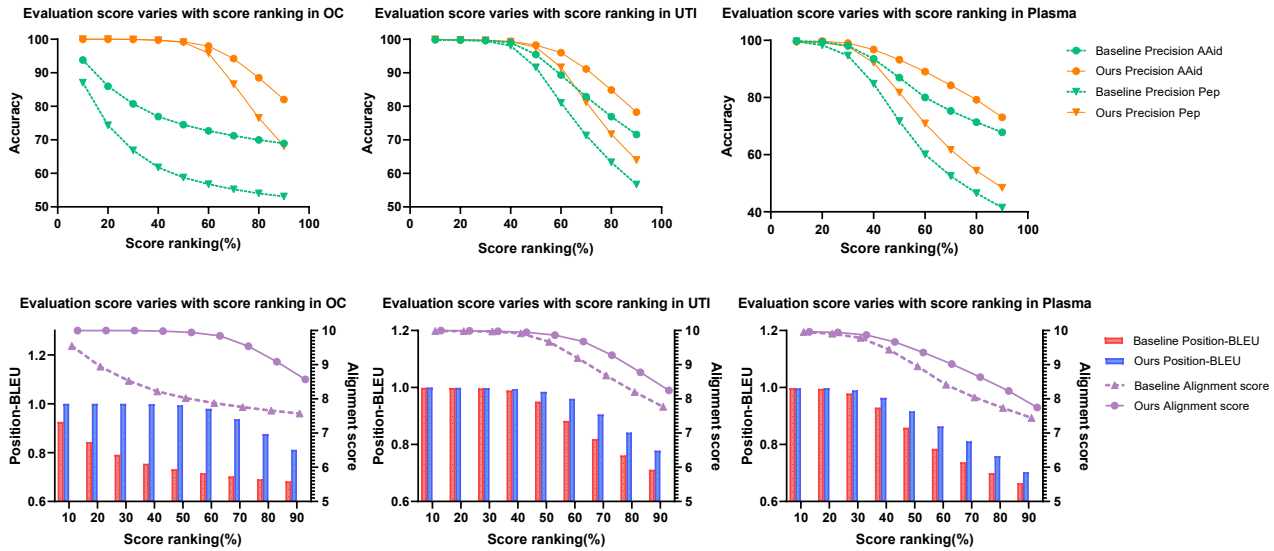

Fig. 7. The change of evaluation with score ranking. The top half of the picture shows Precision_AAid and Precision_pep vary with score ranking in three datasets, and the lower part of the picture shows Position-BLEU and Alignment score varies with score ranking in datasets.

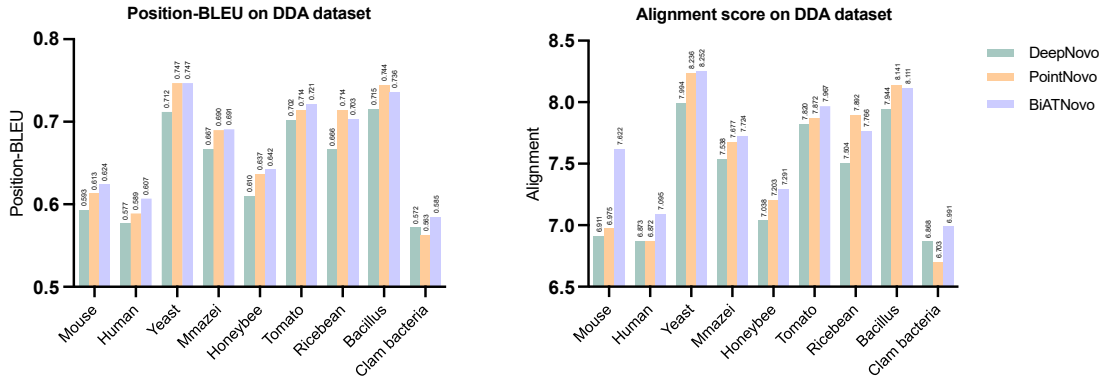

Fig. 8. Position-BLEU and Alignment score on DDA-MS dataset compared with DeepNovo and PointNovo.

BiATNovo can maintain high confidence and high accuracy in the prediction sequence with the highest score ranking in the three datasets. This finding indicates that BiATNovo outperforms the other methods, making it a promising approach for peptide sequence prediction.

For the DDA-MS dataset, we compared BiATNovo with two recent deep learning-based peptide sequencing models in the Position-BLEU and Alignment score (as illustrated in Fig. 8). The BiATNovo showed state-of-the-art results on most datasets except *Ricebean* and *Bacillus*. In terms of amino acid accuracy, BiATNovo improved by 3% on average. In BiATNovo, only the b-ions and y- ions in the MS/MS spectra were calculated, while PointNovo considered the existence of a-ions and explored multiple possibilities in the spectrum resolution at each step. More types of fragment ions can also be added to the feature to improve the prediction accuracy of peptides in DDA-MS data.

Meanwhile, BiATNovo can be applied to DDA-MS and DIA-MS data, whereas PointNovo is limited to DDA-MS data. BiATNovo can achieve

the current optimal prediction effect in DIA-MS data, and the prediction accuracy of DDA-MS data will not decrease significantly. Therefore, BiATNovo is considered a highly compatible and accurate peptide sequencing model.

3.4 Ablation studies

In this section, we investigated why BiATNovo can achieve inspiring prediction by ablation studies (Table 3). We made four variants of BiATNovo, of which the first removes the Bidirectional synchronous peptide sequencing (denoted as BiATNovo-Ide), the second one lacks the Bidirectional peptide sequencing (denoted as AttnNovo), the third (denoted as Local-CNN) alters the model removing Self-Attention mechanism and bidirectional prediction and the fourth (denoted as Attention-decoder) only uses Self-Attention mechanism.

BiATNovo significantly outperforms Local-CNN over all the evaluation scores. In detail, compared with Local-CNN, AttnNovo is more than twice as accurate as Local-CNN in peptides. This shows that the self-attention mechanism captures the rule of peptide sequence in the peptide sequencing process. Moreover, the comparison with AttnNovo and BiATNovo shows a similar improvement in overall evaluation scores. In detail, BiATNovo improves the Precision_AAid by 15.5%, the Precision_Peptide by 24.1%, and the Position-BLEU by 16.9%. These results show that the bidirectional peptide sequencing method can make the forward and backward predicted sequences better converge in the middle position and dynamically expand the candidate sequences. Similar results are found in UTI and Plasma datasets. Precision_AAid increased by 14.9% and 22.1%, respectively. Precision_Peptide increased by 21.6% and 28.6%, respectively, and Position-BLEU increased by 16.0% and 22.3%, respectively.

Table 3. Ablation results of BiATNovo on OC dataset.

Model	Precision_AAid (%)	Recall_AAid (%)	Precision_Pep (%)	Position-BLEU	Alignment score
Local-CNN	50.25	50.64	20.61	0.455	5.750
Attention-decoder	45.51	46.99	34.43	0.491	5.775
AttnNovo	64.59	65.34	49.52	0.638	7.197
BiATNovo-Ide	69.47	70.15	55.59	0.690	7.575
BiATNovo	74.57	75.13	61.47	0.746	7.981

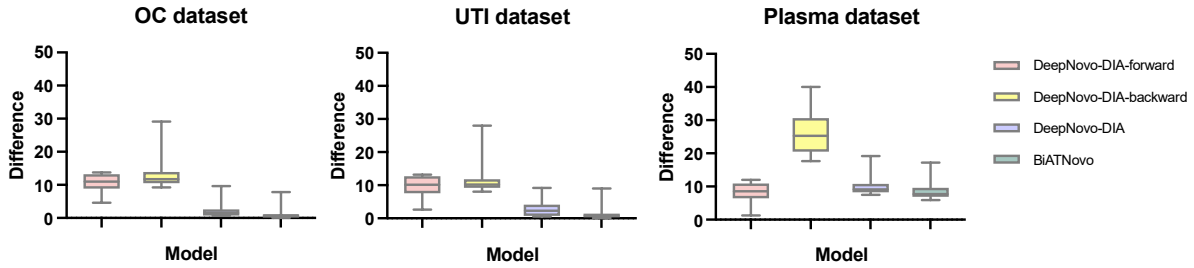

Fig. 9. Results of *Difference-i* in the forward and backward one-way prediction of DeepNovo-DIA, DeepNovo-DIA and BiATNovo.

3.5 Analysis of the imbalance phenomenon

This paper uses bidirectional peptide sequencing to mitigate the problem of output imbalance generated in one-way prediction, the motivation for which is described in the method. In this section, we analyze the effect of the method on the output imbalance problem.

We defined an evaluation index of output imbalance, *Difference-i*, to quantify the output imbalance between the forward and backward predictions in one-way peptide sequencing models. *Difference-i* represents the absolute difference between the accuracy of the *ith* amino acid in the forward and backward predictions. A higher value of Difference-i indicates a larger output imbalance between the two predictions. We compare the results of *Difference-i* in the forward and backward one-way prediction of DeepNov-DIA (denoted DeepNovo-DIA-forward and DeepNovo-DIA-backward), DeepNov-DIA, and BiATNovo (Fig. 9).

The figure reveals that generating forward or backward predictions independently produces a large *Difference-i* value, indicating a significant output imbalance. DeepNovo-DIA adopts the best forward and backward outputs for each sample, resulting in a balanced output with little difference in forward and backward accuracy. Thanks to independent prediction and interactive prediction, BiATNovo can better balance the forward and backward output, and the difference between the forward and backward output accuracy is smaller, indicating that our model has the optimal performance in output imbalance.

3.6 Case study: Analysis of evaluation score

The two evaluation scores proposed in this paper focus on amino acid accuracy in the case of misalignment and reversed position. In this section, we analyze in detail the differences between the predicted and the reference sequences to illustrate the validity of the proposed evaluation scores.

Precision_AAid could not reasonably evaluate the case of amino acid order reversal (as demonstrated in Fig. 10 *A*). In the first example, despite the common subsequence NEQVTR between the reference and predicted sequences in the case of misalignment, Precision_AAid yields a score of 0, thereby indicating incorrect matching. In the second example, the reference and the predicted sequences' first four amino acids are the same, but the order is reversed. The reason may be that the signal intensity of some fragment ions in the MS/MS spectra is missing. During the sequencing process, we know the sum of the mass of the four amino acids of CYTC, but we cannot determine the corresponding order of the four amino acids of CYTC. The Precision_AAid ignores this matter, and we hope to

have a reasonable evaluation of this reversed order rather than simply assuming that the match score of the first four amino acids is zero.

In this study, we counted the partial cases of two, three, and four-reversal. *K-reversal* is defined as the predicted and the reference sequences that differ in only k position amino acids, and the set of these k amino acids is the same (Fig. 10 *B*). To better analyze the experimental data, we excluded sequences with significant differences between the predicted and reference sequences, ensuring that sequences with over 30% completely different amino acid sets are not considered. From the figure, we can find that nearly 10% of sequences are *k-reversal*. Considering that we have only counted a few typical types of reversal, the actual proportion is greater than 10%. The high proportion of *k-reversal* proved the necessity of our proposed evaluation scores. Among these results, our evaluation indexes could be evaluated with more reasonable accuracy of amino acids.

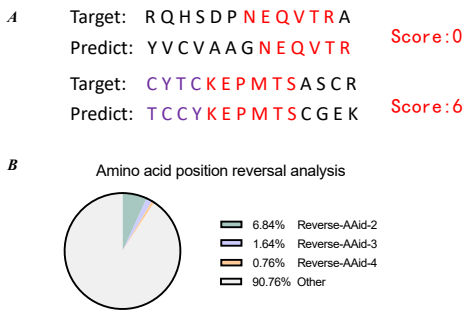

Fig. 10. Evaluation scores analysis. Subfigure *A* shows the defects of the existing evaluation scores, subfigure *B* analyzes the proportion of k-reversal in the dataset.

4 Discussion and conclusion

In the experiment, we used labeled peptide features obtained by searching the database. We noted that the yield of peptides detected by de novo sequencing was proportional to the number of peptide features discovered from the LC-MS map. The correctness of unlabeled peptides (not in the database) can be evaluated by artificial biological experiments or homologous similarity search software.

The accuracy of de novo sequencing falls behind traditional database-based search methods, due to the following reasons: (1) MS/MS spectra often include noise peaks that compromise the accuracy of sequencing. (2) The absence of some fragment ions in MS/MS spectra can cause amino acid sequence reversals, further undermining sequencing accuracy. (3) There is a mass constraint in the process of peptide sequencing, which leads to the accumulation of deviations.

In our contribution, we proposed a de novo sequencing method that facilitates the discovery of new peptide proteins by inferring directly from MS/MS spectra and precursor ions without relying on a database. We designed a bidirectional peptide sequencing and re-scoring method to solve the problems of unbalanced output and deviation accumulation in peptide sequencing. Furthermore, we incorporated a self-attention mechanism in our model to learn the internal rules of peptide sequence and fragment ion patterns, which presents a new solution for the analysis and inference of MS/MS spectra. Meanwhile, this method applies to both DDA-MS and DIA-MS data. Finally, we designed the evaluation score of de novo peptide sequencing, which more comprehensively considers the absence of fragment ions in MS/MS spectra and the misalignment of amino acids in the peptide sequence. In the future, we plan to provide pre-training models for multiple tissue samples, enabling users to fine-tune the models with minimal data for de novo sequencing of different tissue samples.

Acknowledgments

This work was supported by National Natural Science Foundation Program of China “Geometric deep learning methods and application of proteomics cancer pattern analysis” (Grant No. U21A20427) 2022.

References

Aebersold, R. and M. Mann (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**, 347-355.

Cox, J. and M. Mann (2008) MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, **26**, 1367-1372.

Fischer, B., et al. (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. *Analytical Chemistry*, **77**, 7265-7273.

Gessulat, S., et al. (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, **16**, 509-518.

Gillet, L. C., et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, **11**.

Jeong, K., et al. (2013) UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics*, **29**, 1953-1962.

Karunratanakul, K., et al. (2019) Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular & Cellular Proteomics*, **18**, 2478-2491.

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

Lin, T.-Y., et al. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision.

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

Ma, B., et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, **17**, 2337-2342.

Meyer, J. G., et al. (2017) PIQED: automated identification and quantification of protein modifications from DIA-MS data. *Nature methods*, **14**, 646-647.

Mo, L., et al. (2007) MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Analytical Chemistry*, **79**, 4870-4878.

Muntel, J., et al. (2015) Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer. *Journal of proteome research*, **14**, 4752-4762.

Papineni, K., et al. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Pearson, W. R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635-650.

Pino, L. K., et al. (2020) The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass spectrometry reviews*, **39**, 229-244.

Qiao, R., et al. (2021) Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, **3**, 420-425.

Röst, H. L., et al. (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology*, **32**, 219-223.

Tarn, C. and W.-F. Zeng (2021) pDeep3: toward more accurate spectrum prediction with fast few-shot learning. *Analytical Chemistry*, **93**, 5815-5822.

Ting, Y. S., et al. (2017) PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature methods*, **14**, 903-908.

Tran, N. H., et al. (2019) Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature methods*, **16**, 63-66.

Tran, N. H., et al. (2020) Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence*, **2**, 764-771.

Tran, N. H., et al. (2017) De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, **114**, 8247-8252.

Vaswani, A., et al. (2017) Attention is all you need. *Advances in neural information processing systems*, **30**.

Vizcaíno, J. A., et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, **32**, 223-226.

Yang, H., et al. (2019) pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, **35**, i183-i190.

Yang, Y., et al. (2020) In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications*, **11**, 1-11.

Zhang, J., et al. (2020) Synchronous bidirectional inference for neural sequence generation. *Artificial Intelligence*, **281**, 103234.

Zhou, L., et al. (2019) Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, **7**, 91-105.