Retrieve and Revise: Improving Peptide Identification with Similar Mass Spectra

Zhengcong Fei[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
feizhengcong@ict.ac.cn

Abstract

Tandem mass spectrometry is an indispensable technology for identification of proteins from complex mixtures. Accurate and sensitive analysis of large amounts of mass spectra data is a principal challenge in proteomics. Conventional deep learning-based peptide identification models usually adopt an encoder-decoder framework and generate target sequence from left to right without fully exploiting the global information. A few recent approaches seek to employ two-pass decoding, yet have limitations when facing the spectra filled with noise. In this paper, we propose a new paradigm for improved peptide identification, which first retrieves a similar mass spectrum from the database as a reference and then revise the matched sequence according to the difference information between the referenced spectrum and current context. The inspiration of design comes that the retrieved peptide-spectrum pair provides a good start point and indirect access to both past and future information, such that each revised amino acid can be produced with better noise perception and global understanding. Moreover, a disturb-based optimization process is introduced to sharpen the attention for difference vector with reinforcement learning before fed to decoder. Experimental results on several public datasets demonstrate that prominent performance boost is obtained with the proposed method. Remarkably, we achieve new state-of-the-art identification results on these datasets.

1 Introduction

Proteins are key actors in all cellular processes and pathways (Bantscheff et al. 2007). Moreover, almost all diseases are linked to perturbations of proteins. Consequently, the fundamental study of proteins is underlying most biological questions (Yates III 1998; Mann, Hendrickson, and Pandey 2001; Venable et al. 2004). Mass spectrometry (MS) is applied to obtain sequential information of proteins, such as amino acid sequence, composition, and modifications (Yates et al. 1995). Among these, determining the protein subsequence, named peptide, to each produced mass spectrum is the workhorses in the entire identification pipelines and has attracted much research attention recently (Shao and Lam 2017). The advances in deep neural networks (DNNs) have demonstrated promising performances in peptide identification tasks (Tran et al. 2017; Qiao et al. 2019; Tran et al. 2019; Fei 2020c,b). Most existing deep learning-based identification methods learn a neural network in a supervised learning manner based on elaborately designed peptide-sequence matches. The essential practice of such models follows encoder-decoder frameworks. In between, convolutional neural network (CNN) is utilized to encode an input mass spectrum, and recurrent neural network (RNN) is adopted as a decoder to generate target peptide sequences (Tran et al. 2017). During inference, the peptide identification models usually generate the sequence through one-pass decoding from left to right, that is, only conditions on the extracted spectrum features and previously generated subsequence to determine the next target amino acid.

One major limitation of such a one-pass process is that the next amino acid generation only uses partial information of the previously generated incomplete peptide sequence rather than considering the global information carried by a complete target candidate. Meanwhile, the peptide fragments are more likely to occur in the middle position in a practical biological experiment, which results in a lower abundance of useful signal peaks on the sides of spectra and easy to be concealed by noise peaks (Fei 2020b). Future information of the target domain is intuitively beneficial for the current sequence generation since each amino acid in the target peptide has to be consistent with its surrounding subsequences, *i.e.*, both before and after it. Different approaches have been proposed to leverage the global information, *e.g.*, (Qiao et al. 2019) designs an order invariant network structure to encode entire spectrum effectively. (Fei 2020b) introduces a hierarchical multi-stage framework, which starts the inference with a selected high-confidence guiding tag and provides the complete sequence based on this tag. (Fei 2020c) train a value network to estimate all possible sequence extensions of reward to assist current decision. These approaches are of relatively good quality, but inefficient to generate and train.

Inspired from (Zhang et al. 2018; Gu et al. 2018), we propose a new paradigm for peptide identification, which consists of a similar spectrum retriever and a peptide sequence reviser. In specific, given a mass spectrum, we first retrieve a similar spectrum and its associated peptide sequence from database, *i.e.*, the original training data. Then, we concatenate both spectrum vectors weighted by attention module to a difference vector. Finally, we revise the reference peptide sequence condition on the difference vector. The proposed approach does not only benifits from the informativeness advantages of the retrieval pair but also poses the flexibility of target sequence generation. Techniqully, it can alleviate the noisy issue of one-pass generative models by leveraging existing information and is more efficient than multi-stage models. In addition, we introduce a disturb-based optimization to force the model to attend the useful spectrum features and explore more accurate peptide sequences with reinforcement learning. To evaluate the effectiveness of the proposed revised approach, we perform comprehensive experiments with different baselines, including traditional and deep learning-based identification tools, and demonstrate strong superiority on different metrics.

In summary, our contributions are listed as follows:

- We propose a novel framework, retrieve and revise, for mass spectrum data analysis, which combines the input spectrum as well as retrieval reference. To the best of our knowledge, it is the first work to combine retrieval and generation for improved peptide sequence identification.
- We introduce a disturb-based optimization reinforcement learning to help the peptide reviser focus on the effective difference features and discriminate against noise.
- Extensive experiments are conducted on various mass spectra datasets. The results show that our proposed model outperforms all the baseline in a large margin and achieves a new state-of-the-art performance.

2 Background

Mass spectrometry has become a powerful tool in life science research and is nowadays an integral analytical method in proteomics (Aebersold and Mann 2003; Patterson and Aebersold 2003). For example, MS-based peptide identification is the key to increase our understanding of cell biology and human disease, which has numerous applications such as peptide drugs designing for 2019-nCoV (Zhang et al. 2020) In a standard approach in MS-based proteomics (Domon and Aebersold 2006), the biological sample is first lysed, and proteins are extracted followed by proteolytic cleavage into peptides. The peptides are often further fractionated to reduce sample complexity or to enrich for certain subsets of peptides. The first MS stage measures the accurate mass of the entire peptides. In the second stage, referred to as tandem mass spectrometry (MS/MS), peptides are fragmented, and the masses and intensity of the resulting fragment ions are detected. MS spectra usually contain charge-to-mass ratio and relative intensity information, and the identity of the peptides can be deduced by matching the MS/MS spectra against a sequence candidate (Taylor and Johnson 1997; Aebersold and Mann 2003).

In this work, our goal is to model the amino acid sequence of a peptide according to the given spectrum. The major obstacle lies that: (1) The exponential peptide candidates need to be considered when matching with a given spectrum. Most sequence models employ a common decoder mechanism using a greedy or beam search to release this issue (Fei 2020c), while such a mechanism can miss correct peptides at early steps. (2) There are multiple types of ions that have quite different intensity values when the peptide is digested by the enzyme. Meanwhile, the fragmentation rule is critical but remains understudied (Tran et al. 2017). (3) A large number of noise peaks blending with the real useful ions that limit the determination of correct sequences.

There is already a substantial literature on computational proteomics. These algorithms assign a peptide sequence to each tandem mass spectrum (MS/MS) by integrating graph algorithms and dynamic programming to reduce the task complexity (Elias et al. 2004; Zhang 2004; Dasari et al. 2010). Although traditional mechanism learning plus human design features work well, there remains substantial room for improvement. Specially, these methods set strong assumptions, and there are many limitations in practical application. Encouragingly, deep learning technology was introduced to peptide sequencing (Tran et al. 2017; Zhou et al. 2017; Tran et al. 2019; Fei, Wang, and Chi 2020). These methods exploit the encoder-decoder paradigm that firstly utilizes CNN to encode spectrum and then adopt an RNN-based decoder to generate the output sequence, leading to promising results for this task. Nevertheless, such a one-pass framework has encountered a performance bottleneck when facing a seriously noisy filled spectrum. In contrast, our retrieve-and-revise framework incorporates indirect access to global information to compensate such conditions as well as avoid a decoding time punishment.

3 Approach

3.1 Model Overview

The entire workflow of our approach is displayed in Figure 1, consisting of a similar mass spectra retriever and a peptide sequence reviser. In specific, to analyze an input mass spectrum S, we first utilize the similar spectrum retriever to search a reference peptide-spectrum match $(S_i, P_i) \in \mathcal{D}$, where a peptide-spectrum matched database is denoted as $\mathcal{D} = \{(S_i, P_i)\}_{i=1}^{N}$, comprise a sequence of mass spectrum S_i and its corresponding peptide sequence $P_i = \{a_1, \ldots, a_L\}$, where a_l represents the l-th amino acid letter of peptide sequence P_i. Then, the peptide sequence reviser builds a difference vector $v_{diff} = f(S_i, S)$ to encode the information about the difference context between reference spectrum S_i and target spectrum S. Finally, we generate a target peptide sequence according to the conditional probability of $p(P|v_{diff}, P_i)$. In the following, we will introduce how to design the similar mass spectrum retriever and the peptide sequence reviser in detail.

3.2 Similar Spectrum Retriever

Following the common practice (Zhou et al. 2017), we combine Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SPC) jointly to compute the mass spectrum similarity. Firstly, the spectrum is represented as intensity vectors where each index of the vector represents a small mass-to-charge $\frac{m}{z}$ bin, and the value represents the sum of intensities of all peaks which fall into that bin. In concrete, we utilize a spectrum resolution of 10, which means every peak within a 0.1 Da $\frac{m}{z}$ bin will be merged together and represented as an element of the intensity. Then, both

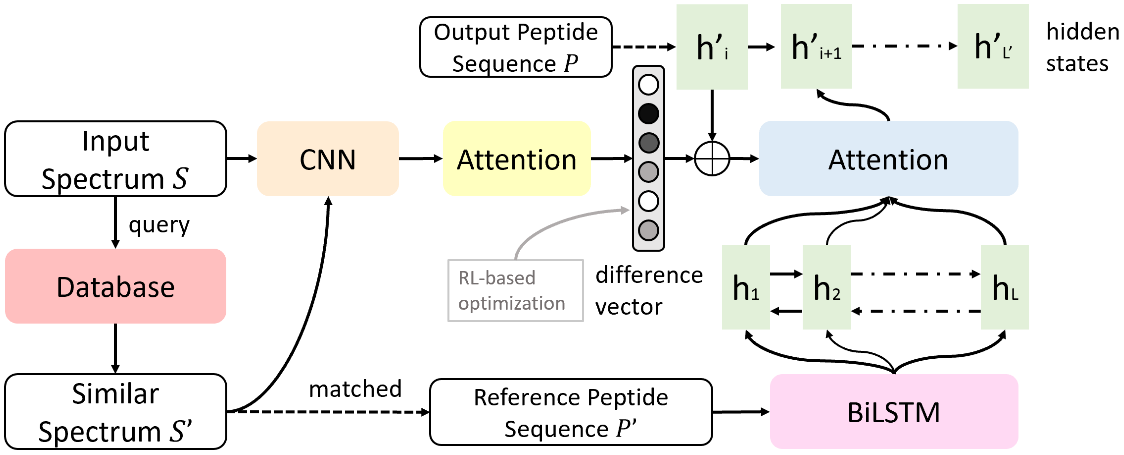

Figure 1: An overview of our proposed peptide identification model, which includes a similar spectrum retriever to search a reference from the database and a peptide sequence reviser to generate the output peptide sequence conditioned on both the difference vector and retrieved sequence.

PCC and SPC are linked linearly to score the similarity between two formalized spectrum vector as follows:

$$PCC = \frac{\sum_i (x_i - \overline{X})(x'_i - \overline{X'})}{\sqrt{\sum_i (x_i - \overline{X})^2}\sqrt{\sum_i (x'_i - \overline{X'})^2}} \quad (1)$$

$$SPC = 1 - \frac{6\sum_i d_i^2}{L_v(L_v^2 - 1)} \quad (2)$$

$$sim = \lambda PCC + (1-\lambda)SPC \quad (3)$$

where X and X' represent the input spectrum vector and reference spectrum vector from the database, d_i denotes the difference between ranked vector pairs, L_v is the vector length, and λ is the balancing factor.

For any (S_i, P_i), our goal is to maximize the conditional probability of P_i by retrieving a reference similar pair $(S'_i, P'_i) \in \mathcal{D}$. In practice, we retrieve twenty similar mass spectra and corresponding peptide sequence candidates $\{(S'_{i,j}, P'_{i,j})\}_{j=1}^{20}$ based on similarity score, and then select the retrieved peptide-spectrum matches whose similarity score are in the range of $[0.3, 0.7]$. Here, each sample in the training dataset is processed with the above procedure, so we can obtain enormous tuples $\{(S_i, P_i, S'_{i,j}, P'_{i,j})_{j=0}^{M}\}_{i=0}^{N}$ after this step. The motivation behind filtering out instances with a similarity score < 0.3 is that a neural reviser model performs well only if reference PSM is similar to its ground-truth and can provide sufficient useful global information. Besides, we also hope the peptide reviser does not only copy the reference peptide sequence, that is, hold the strong revision capability, so we discard instances where the reference sequence and target sequence are nearly identical.

3.3 Peptide Sequence Reviser

Peptide sequence reviser aims to modify a reference peptide sequence to adapt to the current input mass spectrum. Formally, given a tuple (S, P, S', P'), a peptide sequence reviser first forms a difference vector v_{diff} using spectra S and S', and then updates parameters of sequence decoding model by maximizing conditional probability of $p(P|v_{diff}, P')$. Next, we will describe how to obtain the difference vector as well as the peptide generation.

Difference Information Encoding Here, we first encode the difference context between input spectrum S and reference spectrum S' before input to the decoder for revising the reference peptide sequence P'. Formally, reference peptide sequence P' is firstly transformed to hidden vectors through a biLSTM to consider the two-side context as:

$$h_k = \overleftarrow{h}_k \oplus \overrightarrow{h}_k \quad (4)$$

$$\overleftarrow{h_k} = \text{LSTM}(\overleftarrow{h}_{k-1}, a'_j); \quad \overrightarrow{h_k} = \text{LSTM}(\overrightarrow{h}_{k+1}, a'_j) \quad (5)$$

where $\oplus$ denotes concatenation operation and a'_j is the j-th amino acid of reference peptide P'. Then we compute a difference vector v_{diff} with a dual attention mechanism following (Park, Darrell, and Rohrbach 2019), which act as a difference information localizer between S and S'. Assume V and V' are spectrum features of S and S' respectively, encoded by a pre-trained CNN (Qiao et al. 2019; Fei 2020a; Park, Darrell, and Rohrbach 2019). Spectrum features are then used to generate two separate spatial attention maps A and A'. We utilize element sigmoid for computing our attention maps as:

$$A = \sigma(\text{Conv}_2(\text{ReLu}(\text{Conv}_1(V)))) \quad (6)$$

$$A' = \sigma(\text{Conv}_2(\text{ReLu}(\text{Conv}_1(V')))) \quad (7)$$

$$l = \sum A \odot V; \quad l' = \sum A' \odot V' \quad (8)$$

$$c_{diff} = l \oplus l' \quad (9)$$

where Conv, σ and $\odot$ indicate convolution layer, element sigmoid and element wise multiplication. c_{diff} explicitly encodes difference context information between S and S'. Then we compute the difference vector v_{diff} by following non-linear transformation:

$$v_{diff} = \tanh(W_d \cdot c_{diff} + b_d) \tag{10}$$

where W_d and b_d are two learnable parameters. Equation 10 can be regareded as a feature space mapping from spectrum difference to peptide sequence difference.

Peptide Sequence Generation We build our peptide sequence generation model upon a conventional attention-based encoder-decoder framework, integrating the difference vector and reference sequence into the decoder. The decoder takes $\{h_k\}$ as an input and generates a target peptide sequence S by a LSTM model with weighted attention. The hidden state of the decoder can be acquired by:

$$h'_j = \text{LSTM}(h'_j, a_{j-1} \oplus v_{diff}) \tag{11}$$

where the input of j-th time step is the last step hidden state h'_{j-1} as well as the concatenation of the (j-1)-th amino acid a_{j-1} embedding and the difference vector v_{diff} obtain in Equation 11. Then we comput a context vector c_i, which is a weight sum of $\{h_k\}$.

$$c_i = \sum_j \alpha_{i,j} h_j \tag{12}$$

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum(e_{i,j})} \tag{13}$$

$$e_{i,j} = W_o \tanh(W_i[h_j \oplus h_i]) \tag{14}$$

where weight $\alpha_{i,j}$ is linked to relation between hidden states and Ws are trainable weights. The final generative probability distribution is given by,

$$p = \text{softmax}(W_p[a_{i-1} \oplus h'_i \oplus c_i] + b_p) \tag{15}$$

Note that Equation 8 and 12 are the attention mechanism following (Chorowski et al. 2015), which can mitigate the long-term dependency issue of the original Seq2Seq model. We append the difference vector to every input embedding of the decoder in Equation 11, so the difference information can be utilized conveniously in the entire generation process.

3.4 Reinforcement Learning-based Optimization

Similar to previous works, we can train our peptide sequence generation model by minimizing the conventional negative log likelihood loss, which aims to maximize the probability of the ground truth peptide sequence provided with the difference vector and current context as:

$$\mathcal{L}_{XE} = -\sum_{j=1}^{L} \log p(a_j | v_{diff}, P', a_{<j}) \tag{16}$$

However, it is proved that only training with cross-entropy loss is insufficient, and the resulting model usually suffers from the exposure bias problem (He et al. 2019). What's more, the peptide identification model should not only learn the difference between two spectra but also know which is useful regions while others are noisy or invalid. Therefore, we introduce a reinforcement learning-based strategy that incorporating the attention perturbation to further optimize our cross-modality attention module.

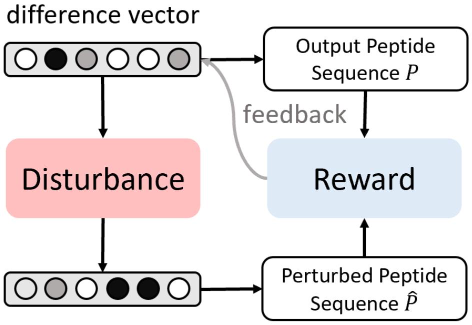
difference vector
Output Peptide Sequence P
feedback
Disturbance
Reward
Perturbed Peptide Sequence $\hat{P}$

Figure 2: A schematic diagram of reinforcement learning-based optimization to force model to attend the effective difference features with feedback reward.

During training, our model generates two peptide sequence candidates for the same input spectrum. The first peptide sequence $P = \{a_1, \ldots, a_L\}$ is generated in the usual way with original difference vector, while the second peptide sequence $\hat{P} = \{\hat{a}_1, \ldots, \hat{a}_L\}$ is generated in a slightly different way according to a sampled probability map $\hat{v}_{diff}$, as shown in Figure 2. Here, $\hat{v}_{diff}$ is a modified map of difference vector v_{diff} by integrating some small disturbation, which will lead to the generation of a new peptide sequence. By comparing the estiamtion score of two sequences P and $\hat{P}$, and rewarding the better one with reinforcement learning technique, the peptide identification model can virtually explore for a better match.

Specially, to generate a disturbed $\hat{v}_{diff}$, we first randomly select some elements from the original feature map v_{diff} to perturb with a sampling of γ. For the selected element $v^{i,j}_{diff}$, we generate its new perturbed version by sampling from the following Gaussian distribution:

$$\hat{v}^{i,j}_{diff} \sim \mathcal{N}(v^{i,j}_{diff}, \sigma) \tag{17}$$

where the standard deviation σ corresponds to a small preset hyperparameter. For the residual elements, we maintain the original form, $\hat{v}^{i,j}_{diff} = v^{i,j}_{diff}$. With $\hat{v}$, we can then generate the peptide sequence $\hat{P}$ with peptide reviser accordingly. After that, we calculate the accuracy metric for the two peptide sequences and treat them as rewards, *i.e.*, $r(P) = \text{AAP}(P, P')$ and $r(\hat{P}) = \text{AAP}(\hat{P}, P')$. If $r(P)$ is higher than $r(\hat{P})$, it means that $\hat{P}$ shows better quality and matchness of peptide sequences than P. Consequently, the attended difference vector v_{diff} should be more close to the sampled vector $\hat{v}_{diff}$. Along with this intuition, we can define our reinforcement loss as follows:

$$\mathcal{L}_{RL} = \max(r(\hat{P}) - r(P), 0) ||\hat{v}_{diff} - v_{diff}||^2_2 \tag{18}$$

Further more, the overall training loss can be rewritten as:

$$\mathcal{L} = \mu \mathcal{L}_{RL} + \mathcal{L}_{XE} \tag{19}$$

where μ denotes a balancing factor to control the relative weights of two losses.

Data set	Lab	Instrument	Resolution	Species	#Spectra	Publication
Mann-Human-QE	Mann	Q Exactive	high	Human	27,570	(Michalski et al. 2011)
Mann-Mouse-QEHF	Mann	Q Exactive HF	high	Mouse	172,000	(Sharma et al. 2015)
Gygi-Human-QE	Gygi	Q Exactive	high	Human	176,000	(Chick et al. 2015)
Dong-Ecoli-QE	Dong	Q Exactive	high	*E. coli*	15,000	(Liu et al. 2014)
Xu-Yeast-QEHF	Xu	Q Exactive HF	high	Yeast	243,000	(Chi et al. 2018)

Table 1: Basic mass spectrum dataset information integrated in the experiments.

4 Experiments

We empirically verify the merit of our retrieve-and-revise model by conducting experiments on different mass spectrum datasets for peptide identification task.

4.1 Experimental Settings

Datasets Diverse datasets are collected from the previous work (Tran et al. 2017; Fei 2020b), which are all publicly available. Dataset information is listed in Table 1. Following the common practice in (Fei 2020b), we utilize Open-pFind (Chi et al. 2018) to process these raw data sets and search against the reviewed peptide sequence database of human, mouse, *E.coli*, and yeast, respectively, which can be downloaded from Uniprot and their versions are consistent with (Chi et al. 2018). We also set the precursor ion tolerance and the fragment ion tolerance as ±20 ppm. The FDR is controlled at 1% at the spectrum level. At last, $\sim$ 920,000 high-quality PSMs were obtained in the experiments. Importantly, the peptide sequences identified from Open-pFind were assigned to the corresponding MS/MS spectra and regarded as ground truth for estimating peptide identification results' correctness.

Evaluation Metrics In this paper, the generated amino acid can be considered as correct when the mass difference between the predicted amino acid and a ground-truth amino acid is less than 0.1 Da, and the prefix mass before them as well as the suffix mass behind them are different by less than 0.5 Da. Following (Qiao et al. 2019), we adopt three types of metrics: *precision*, *recall* and *area under curve* (AUC) to evaluate the performance of peptide sequencing. In between, *amino acid level precision* denotes the ratio of the total number of matched amino acids over the total number of amino acids in the generated peptide sequences while *peptide level precision* the fraction of correct peptide sequences in total predicted peptide sequences. Similar definitions can be applied to recall and AUC as well.

4.2 Compared Methods

(1) PEAKS (Ma et al. 2003) adopts a sophisticated dynamic programming algorithm to make peptide correspond to as many high abundance peaks as possible. (2) Novor (Ma 2015) employs a decision tree based scoring function with two-stage refinement to select the peptide candidates. (3) pNovo3 (Yang et al. 2019) uses a learning-to-rank framework to distinguish similar peptide candidates for each spectrum. (4) DeepNovo (Tran et al. 2017) is the first work to incorporate deep learning technology with peptide identification, which is under the conventional CNN plus RNN framework. (5) DeepNovoV2 (Qiao et al. 2019) combines an order invariant network and RNN to predict peptide sequence patterns. (6) DRL (Fei 2020c) consider peptide identification as a multi-step decision making process and optimize with a reward function. (7) DeepTag (Fei 2020b) starts the inference with a selected high-confidence guiding tag and produce the complete sequence based on this guiding tag further. Please note that the first three methods are constructed with a traditional search strategy, while the last two methods are deep learning-based.

4.3 Implementation Details

The Mann-Human-QE dataset is served as our retrieval database in the whole experiment. We utilize pre-trained T-Net (Tran et al. 2017) as spectrum feature extractor. Meanwhile, two-layers biLSTM is adopted to encode reference peptide sequence, whose hidden size is set to 256, and dropout is set to 0.2. The same structure of biLSTM is used to peptide sequence decoding, and both are trained independently. The λ is set as 0.5 in the similar mass spectrum retrieval. Due to the low loss value, we set the hyperparameter μ in Equation 19 as 200 to put more weight on the reinforcement learning loss. In the reinforcement learning optimization, we set the random sampling rate of γ as 0.025. As for the gaussian distribution, we set σ as 0.1. During training, we use a minibatch size of 64, the Adam optimizer (Kingma and Ba 2014) with learning rate 3e-4.

We implement our experiments in PyTorch. After each epoch, we evaluate the model performance on the validation set and choose the identification model with the best performance. Following setting in (Fei 2020b), our model was first trained on the Mann-Human-QE data set and then tested on Mann-Mouse-QEHF for cross-species validation and Gygi-Human-QE data set for cross-lab validation. The rest of the data sets were used to test the robustness of the proposed model. Please note that the training dataset and testing dataset come from different species (Zhou et al. 2017). The cross-validation is used to guarantee unbiased training and testing and does not give our model any extra advantage.

4.4 Comparison with State-of-the-art

The performances of different peptide identification models on public mass spectrum datasets for the peptide identification task are summarized in Table 2. Overall, the results across all metrics consistently indicate that our revise-based model exhibits better performances than other approaches, including statistical-based methods (PEAKS, PepNovo, and

Dataset	Mann-Mouse-QEHF			Gygi-Human-QE			Dong-Ecoli-QE			Xu-Yeast-QEHF		
Metric	AAR	AAP	PR	AAR	AAP	PR	AAR	AAP	PR	AAR	AAP	PR
Statistical based methods												
PEAKS	0.342	0.483	0.145	0.365	0.421	0.152	0.425	0.462	0.178	0.382	0.453	0.155
Novor	0.371	0.502	0.152	0.383	0.412	0.187	0.461	0.501	0.218	0.401	0.512	0.171
pNovo3	0.402	0.508	0.193	0.416	0.419	0.23	0.494	0.508	0.244	0.423	0.535	0.182
Deep learning based methods												
DeepNovo	0.427	0.512	0.241	0.454	0.428	0.251	0.513	0.521	0.321	0.466	0.561	0.253
DeepNovoV2	0.467	0.532	0.266	0.484	0.448	0.281	0.533	0.538	0.345	0.482	0.583	0.262
DRL	0.480	0.562	0.284	0.495	0.464	0.294	0.554	0.558	0.347	0.501	0.612	0.269
DeepTag	0.492	0.568	0.289	0.515	0.486	0.307	0.581	0.580	0.382	0.512	0.605	0.271
Ours	0.501	0.575	0.293	0.527	0.497	0.315	0.590	0.586	0.390	0.522	0.610	0.275

Table 2: Evaluation results of popular statistical-based, deep learning-based and our retrieve-and-revise model on different data sets.AAR represents amino acid recall, AAP represents amino acid precision, and PR represents peptide recall. The metric values of baselines are from the corresponding paper.

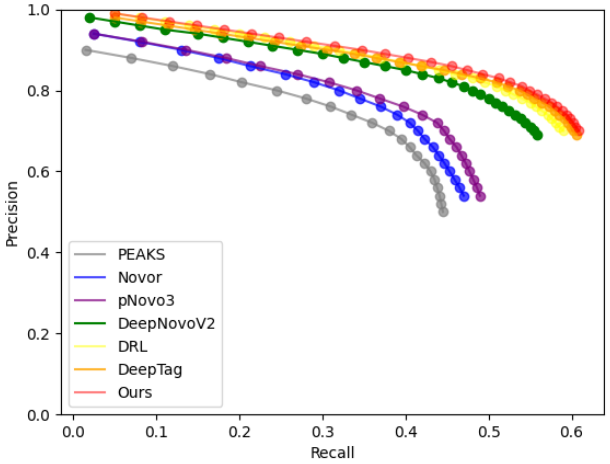

Figure 3: The precision-recall curves of various baselines and our proposed peptide identification model on Mann-Mouse-QEHF dataset.

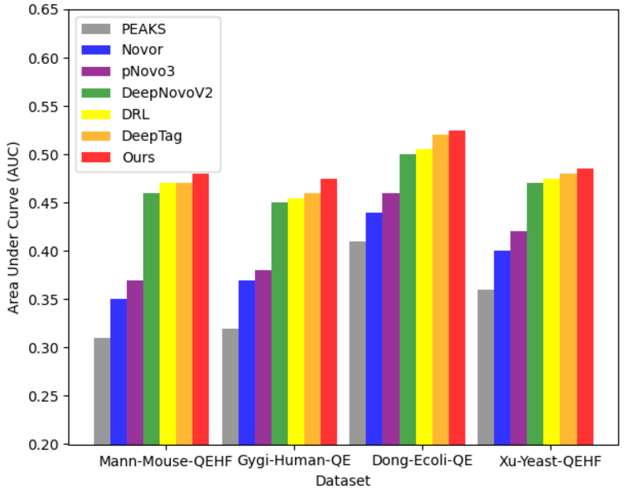

Figure 4: The area under curve of various baselines and our proposed peptide identification model on different species and labs data sets.

Novor) and deep learning-based methods (DeepNovo, DeepNovoV2, and DeepTag). In particular, our model by integrating similar references with global information, makes the absolute improvement over the best competitor DeepTag by 3.1% in terms of PR score, which achieves a new state-of-the-art result. The results generally highlight the key advantage of exploiting the past and after features on the basis of similar spectrum-peptide pair, persuing a through difference information understanding. Similar to the observations (Tran et al. 2019), deep learning-based methods lead to better performances than conventional statistical-based method.

On the other hand, we should also be noted that all peptide identification models report confidence scores for their own predictions, and setting a higher threshold of confidence score will lead to a smaller part of peptides with high precision but will make the rest of the dataset without results (Tran et al. 2017). It is difficult for biologists to accept in practical application. In this end, it is reasonable to depict precision-recall curves and incorporate the area under curve (AUC) as metrics of peptide sequencing quality. Figure 3 and 4 display the precision-recall curves on the Mann-Mouse-QEHF dataset and the AUC of different peptide sequencing methods on different data sets, respectively. We can observe that our retrieve-and-revise model maintains superiority performance against other peptide identification methods, including conventional human-designed and deep learning-based. Encouragingly, extensive quantify experimental results demonstrated that the improvement of our method was efficient and reliable.

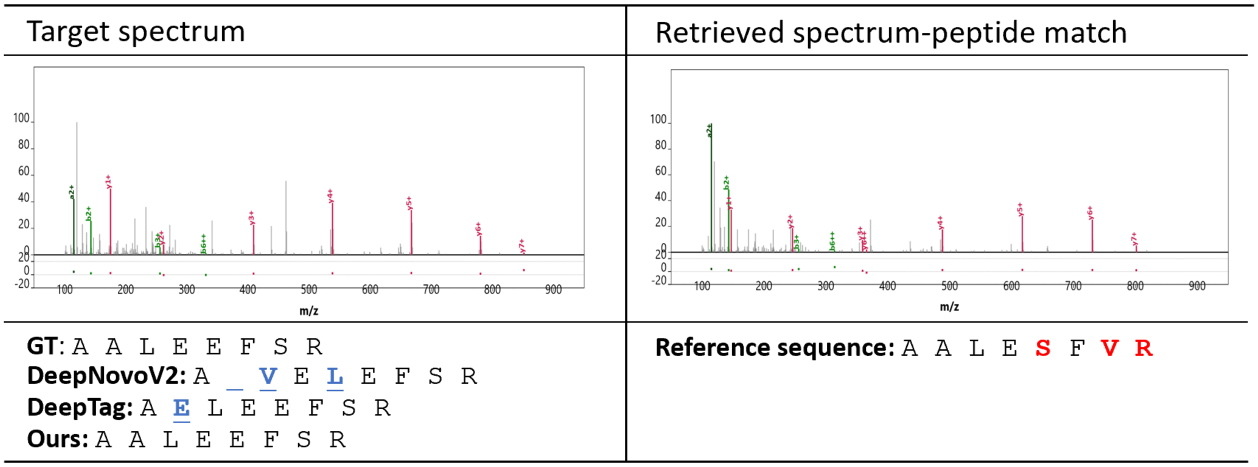

Figure 5: One mass spectrum examples form Mann-Mouse-QEHF dataset with retrieved similar peptide-spectrum match and peptide identification results from the training dataset. The peptide sequences are generated by 1) Ground Truth (GT), 2) DeepNovoV2, 3) DeepTag, and 4) our proposed method (Ours). Peaks colored in red and blue are effective peaks, amino acid letters colored in red denotes the difference between the reference sequence and target sequence, and amino acid letters colored in blue represents the difference between the candidates and ground-truth peptide sequences. Spaces are underlined to indicate missing amino acid letters.

4.5 Qualitative Analysis

Figure 5 showcases a mass spectrum identification result with retrieved spectrum-peptide pair produced by DeepNovoV2, DeepTag, and our proposed revised based model, respectively. As illustrated in the examplar results, the peptide sequences output by our model is most correct to the ground truth. We can observe that there are a lot of noise peaks (color in gray) on both sides of the mass spectrum. For DeepNovoV2, which inference the peptide sequence from left to right in one stage, is prone to make mistakes at early stages. Though DeepTag incorporates a high-confidence tag to release the issue, it still comes to a tight spot sometimes. In contrast, our model effectively solves this problem. We speculate the results are benefited from the integration of the reference template. The results again indicate the advantage of revising the peptide sequence from a good point and considering the past and after information.

4.6 Effect of Reinforcement Learning-based Optimization

To clarify the effect of the RL-based optimization, we illustrate the performance over automatic metrics with different parameters μ in Table 3. We can see that: 1) By incorporating the RL-based optimization, that is, $\mu > 0$, all of the metrics are increased, indicating the effectiveness of perturbing with reforcement learning technique. 2) Along with the column of metric scores listed in Table 3, most performance metrics are generally like the shapes hillside (first increase then decrease) when it varies in a range from 0 to 2000. Correspondingly, we set the parameter μ as 200 in our experiments, which can achieve the best performance.

μ	AAR	AAP	PR	AUC
0	0.494	0.567	0.288	0.47
100	0.498	0.572	0.291	0.48
200	0.501	0.575	0.293	0.48
500	0.500	0.575	0.292	0.48
1000	0.495	0.572	0.290	0.47
2000	0.490	0.565	0.283	0.46

Table 3: Effect of RL-based optimization. Results are evaluated on Mann-Mouse-QEHF. μ denotes the balancing factor to reweight of the perturbed loss.

5 Conclusion

In this paper, we present a new paradigm, retrieve and revise, for improved MS-based peptide identification, enabling a peptide sequence generation model to leverage effective retrieved information. Compared with other conventional peptide identification methods, our proposed approach can effectively exploit global information and accurately discriminate against the changes between spectrum pairs. To further utilize the sequence information to guide the diversity judgment, we have further proposed a reinforcement learning-based optimization to supervise the difference vector with the peptide sequence evaluation rewards. Experiment results on various datasets show that our model outperforms both traditional and deep learning methods on some metrics. More remarkably, we achieve new state-of-the-art performances on these datasets. The results on our model also validate the potential of useful knowledge retrieving.

References

Aebersold, R.; and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* 422(6928): 198–207.

Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; and Kuster, B. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* 389(4): 1017–1031.

Chi, H.; Liu, C.; Yang, H.; Zeng, W.-F.; Wu, L.; Zhou, W.-J.; Wang, R.-M.; Niu, X.-N.; Ding, Y.-H.; Zhang, Y.; et al. 2018. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature biotechnology* 36(11): 1059–1061.

Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; and Gygi, S. P. 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 33(7): 743–749.

Chorowski, J. K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-based models for speech recognition. In *Proc. NIPS*, 577–585.

Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A.-J. L.; and Tabb, D. L. 2010. TagRecon: high-throughput mutation identification through sequence tagging. *Journal of proteome research* 9(4): 1716–1726.

Domon, B.; and Aebersold, R. 2006. Mass spectrometry and protein analysis. *science* 312(5771): 212–217.

Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; and Gygi, S. P. 2004. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature biotechnology* 22(2): 214–219.

Fei, Z. 2020a. Actor-critic sequence generation for relative difference captioning. In *Proc. ICMR*.

Fei, Z. 2020b. Improving Tandem Mass Spectra Analysis with Hierarchical Learning. In *Proc. IJCAI*, 4345–4351.

Fei, Z. 2020c. Novel Peptide Sequencing With Deep Reinforcement Learning. In *Proc. IEEE ICME*, 1–6.

Fei, Z.; Wang, K.; and Chi, H. 2020. GameTag: A New Sequence Tag Generation Algorithm Based on Cooperative Game Theory. *Proteomics* 2000021.

Gu, J.; Wang, Y.; Cho, K.; and Li, V. O. 2018. Search Engine Guided Neural Machine Translation. In *AAAI*, 5133–5140.

He, T.; Zhang, J.; Zhou, Z.; and Glass, J. 2019. Quantifying exposure bias for neural language generation. *arXiv preprint arXiv:1905.10617* .

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

Liu, C.; Song, C.-Q.; Yuan, Z.-F.; Fu, Y.; Chi, H.; Wang, L.-H.; Fan, S.-B.; Zhang, K.; Zeng, W.-F.; He, S.-M.; et al. 2014. pQuant improves quantitation by keeping out interfering signals and evaluating the accuracy of calculated ratios. *Analytical chemistry* 86(11): 5286–5294.

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

Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; and Lajoie, G. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry* 17(20): 2337–2342.

Mann, M.; Hendrickson, R. C.; and Pandey, A. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annual review of biochemistry* 70(1): 437–473.

Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; and Horning, S. 2011. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & Cellular Proteomics* 10(9): 111–120.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proc. IEEE CVPR*, 4624–4633.

Patterson, S. D.; and Aebersold, R. H. 2003. Proteomics: the first decade and beyond. *Nature genetics* 33(3): 311–323.

Qiao, R.; Tran, N. H.; Xin, L.; Shan, B.; Li, M.; and Ghodsi, A. 2019. DeepNovoV2: Better de novo peptide sequencing with deep learning. *arXiv preprint arXiv:1904.08514* .

Shao, W.; and Lam, H. 2017. Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass spectrometry reviews* 36(5): 634–648.

Sharma, K.; Schmitt, S.; Bergner, C. G.; Tyanova, S.; Kannaiyan, N.; Manrique-Hoyos, N.; Kongi, K.; Cantuti, L.; Hanisch, U.-K.; Philips, M.-A.; et al. 2015. Cell type–and brain region–resolved mouse brain proteome. *Nature neuroscience* 18(12): 1819–1830.

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

Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; and Li, M. 2019. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature methods* 16(1): 63–66.

Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; and Li, M. 2017. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences* 114(31): 8247–8252.

Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; and Yates, J. R. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods* 1(1): 39–45.

Yang, H.; Chi, H.; Zeng, W.-F.; Zhou, W.-J.; and He, S.-M. 2019. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* 35(14): i183–i190.

Yates, J. R.; Eng, J. K.; McCormack, A. L.; and Schieltz, D. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry* 67(8): 1426–1436.

Yates III, J. R. 1998. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* 33(1): 1–19.

Zhang, H.; Saravanan, K. M.; Yang, Y.; Hossain, M. T.; Li, J.; Ren, X.; Pan, Y.; and Wei, Y. 2020. Deep learning based drug screening for novel coronavirus 2019-nCov. *Interdisciplinary Sciences, Computational Life Sciences* 1.

Zhang, J.; Utiyama, M.; Sumita, E.; Neubig, G.; and Nakamura, S. 2018. Guiding Neural Machine Translation with Retrieved Translation Pieces. In *Proc. NAACL*, 1325–1335.

Zhang, Z. 2004. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* 76(14): 3908–3922.

Zhou, X.-X.; Zeng, W.-F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S.-M.; and Zhang, Z. 2017. pdeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical chemistry* 89(23): 12690–12697.