# Alpha-Frag: a deep neural network for fragment presence prediction improves peptide identification by data independent acquisition mass spectrometry

Jian Song[1,2,3], Fangfei Zhang[4,5,6] and Changbin Yu[2,3*]

[1]Zhejiang University, Hangzhou, Zhejiang Province, China

[2]School of Engineering, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

[3]Institute of Advanced Technology, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

[4]Fudan University, Shanghai, China

[5]Zhejiang Provincial Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

[6] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

*Corresponding author, email: yuchangbin@westlake.edu.cn

# ABSTRACT

**Motivation**: Identification of peptides in data-independent acquisition (DIA) mass spectrometry (MS) typically relies on the scoring for the peak groups upon extracted chromatograms of fragment ions. Expanding fragment scoring features closer to the genuine experimental spectra can improve DIA identification. Deep learning is able to predict fragment presence without understanding the fragmentation mechanism that can enrich the scoring features in DIA identification.

**Results**: In this work, we developed a deep neural network-based model, Alpha-Frag, to predict the fragment ions that should be present for a given peptide by reporting their probabilities of existence. The prediction performance was evaluated in terms of intersection over union (IoU), and Alpha-Frag achieved an average of >0.7 and outperformed substantially the benchmarks across the validation datasets. Furthermore, qualitative scores based on Alpha-Frag were designed and incorporated into the peptide statistical validation tools as auxiliary scores. Our preliminary experiments show that the qualitative scores by Alpha-Frag are profitable for DIA identification, especially in the case of short gradient, and yielded an increase of 10.1%-29.3% improvements for the test dataset compared to the same scoring strategy but using Prosit.

**Availability and Implementation**: Source code and the trained model are available at www.github.com/YuAirLab/Alpha-Frag.



**Keywords:** data independent acquisition; fragment presence prediction; peptide identification

# 1 Introduction

Tandem mass spectrometry (MS/MS)-based technologies have been widely used in bottom-up proteomics for peptide identification and quantification (Aebersold and Mann, 2016). The MS/MS spectra record two-dimensional information of fragment ions derived from precursors, mass-to-charge ratio (*m/z*) and their corresponding intensities. Data-independent acquisition (DIA), as a MS/MS strategy, has becoming more and more popular due to its unbiased acquisition fashion by cycling through consecutive precursor isolation windows (Gillet et al., 2012). The most common peptide identification of DIA is the library based targeted approach (Röst et al., 2014; Zhang et al., 2020). In this approach, the engine for DIA identification extracts ion chromatograms (XIC) from MS/MS signals and scores the resulting peak group. Since MS/MS spectra measure the presence and intensity of fragment ions simultaneously, the scoring should take both fragment presence consistency and intensity similarity into consideration (Similar to the term fragment intensity, which describes the intensities of fragment ions, the term fragment presence means the set of present fragment ions irrespective of their intensities). If more genuine set of fragment ions rather than all theoretically possible fragment ions is provided to calculate the fragment presence consistency, the scoring for DIA identification will be more accurate and comprehensive.

The presence and absence of fragment ions observed in an MS/MS spectrum depend on factors including peptide primary sequence, charge state of the peptide, and how the collision energy is introduced (Perkins et al., 1999). To achieve fragment presence prediction, enumerating all the possible fragmentation patterns of a peptide with equal probability is possible theoretically but leads to produce a lot of redundant fragment ions. Alternatively, selecting non-zero intensity fragment ions from the result of fragment intensity prediction is another implementation of fragment presence prediction. Currently, with the advances in deep learning based modeling (LeCun et al., 2015), fragment intensity predictors modeled by neutral network, such as pDeep (Zhou et al., 2017), DeepMass (Tiwary et al., 2019) and Prosit (Gessulat et al., 2019), are making increasingly accurate

fragment intensity prediction. Among these, pDeep and DeepMass followed the framework of bidirectional LSTM (Hochreiter and Schmidhuber, 1997) combined with fully connected layers to predict intensities of fragment ions. Innovatively, Prosit took advantage of the recent advances in neural machine translation (Kingma et al., 2014; Xu et al., 2015) with an encoder to code the amino acids sequence and a decoder to output the fragment ion intensity step by step and had become the state-of-art of intensity predictor. For fragment presence prediction based on intensity predictors, the above-mentioned intensity prediction models aim to spectrum similarities (Toprak et al., 2014) which dominated by high-intensity ions, potentially limiting existence prediction of low-intensity ions. Some low-intensity fragment ions that non-existing but predicting present or present but predicting missing may lead to inaccurate calculation of presence similarity and reduce subsequent statistical power for peptide identification of DIA.

The purpose of this study is to develop a deep neural network-based model, Alpha-Frag, for fragment presence prediction. Our study design, including model training, evaluation and application, is shown in Fig. 1. Peptide sequence and charge are taken as the input for Alpha-Frag to predict the presence of fragments ions. Unlike intensity prediction, which is a regression problem in most cases, Alpha-Frag models fragment presence prediction as a multi-label classification problem. Taking advantage of the ProteomeTools (Zolg et al., 2017) dataset for model training, Alpha-Frag can output the most likely set of fragment ions for a given peptide. The result of prediction was evaluated by intersection over union (IoU) between experimental and predicted fragment ions. To illustrate the possible application of Alpha-Frag, we designed the qualitative scores based on Alpha-Frag as auxiliary scores to improve DIA identification. The preliminary results showed that Alpha-Frag is able to predict the presence of fragment ions accurately and substantially improves DIA identifications, especially in the case of short gradient.

# 2 Material and methods

## 2.1 Data preparation

In this paper, multiple datasets were used for model training & test, evaluation and application as for DIA identification (shown in Fig. 1). Like Prosit, training and test datasets (table in the top panel of Fig. 1) were sourced from ProteomeTools (PXD004732 and PXD010595 in PRIDE) of which the raw data were searched by MaxQuant (v.1.5.3.30; Cox and Mann, 2008) at 1% peptide false discovery rate (FDR). To select a high confident dataset, only PSMs met these criteria were considered: Andromeda score is >50, charge is <5, peptide length is restricted to 7-30 amino acids. Alpha-Frag used MS/MS data with normalized collision energy (NCE) 35 suggested by Zolg et al., 2017. Meanwhile, y/b fragment ions with charges up to 2 were extracted and annotated from the corresponding MS/MS for each identified peptide by MaxQuant, and these y/b-ions were used as the ground truth of present fragment ions for training and test. In total, 452,558 precursors (1,252,829 PSMs) were kept and split into a training set and a test set randomly by a ratio at 4:1.

To ensure the generality of model evaluation, we relied on three diverse data-dependent acquisition datasets. Details were summarized in the table in the middle panel of Fig. 1. Among these, Ammar's dataset (Ammar et al., 2019) was acquired by TripleTOF 6600 with CID fragmentation. After searched by MaxQuant (v.1.5.3.12) at 1% FDR and filtration (Andromeda score is >50, precursor charge is <5, peptide length is restricted to 7-30 amino acids), Ammar's dataset has 18,700 PSMs. Huang's dataset (Huang et al., 2020) was obtained from Q Exactive HF (Thermo Scientific) with HCD fragmentation (NCE=27) and has 270,125 PSMs after the same filtration (MaxQuant v.1.6.0.1). Bekker-Jensen's tryptic dataset (Bekker-Jensen et al., 2017) searched by MaxQuant (v.1.5.3.6) was conducted in Q Exactive HF with HCD fragmentation (NCE=28) and has 315,975 PSMs after merged search results. Each of the PSMs for these three datasets had a corresponding set of present fragment ions determined by MaxQuant. For training & test and evaluation datasets, the search result of MaxQuant were available on their corresponding PRIDE.

For the application of Alpha-frag in DIA identification, Bruderer's dataset (Bruderer et al., 2017) which has been specifically created for testing DIA identification was adopted. It is a HeLa whole-proteome tryptic lysate collected on a Q Exactive HF (NCE 27.5) with 0.5h, 1h, 1.5h, 2h chromatographic gradient lengths (referred as HeLa-0.5h, HeLa-1h, HeLa-1.5h, HeLa-2h). The DIA-MS method of Bruderer's dataset consisted of one MS1 survey scan (120,000 resolution; AGC target, 3e6; maximum fill time, 60ms; covering m/z range of 350-1650) followed by a few sequential MS2 windows of variable width (30,000 resolution; AGC target was 3e6 and auto for injection time; minimum m/z: 200). The number of variable isolation windows were set to 18, 21, 26 and 37, respectively. Software OpenSWATH (Röst et al., 2014) and PyProphet (Teleman et al., 2015) were selected to identify the DIA files as they are convenient to add custom function. Meanwhile, CiRT peptides (Parker et al., 2015) and pan-human mass spectrometry library (PHL, Rosenberger et al., 2014) which contains target and decoy peptides were used to aid the DIA identification. To optimize the identification performance, parameter 'mz_extraction_window' in OpenSWATH was proportional to the gradient suggested by Demichev et al., 2020, i.e. 600s, 1200s, 1800s and 2400s for HeLa-0.5h, HeLa-1h, HeLa-1.5h and HeLa-2h, respectively. The remaining options employed by OpenSWATH (v.2.4.0) were: '-batchSize 1000 -readOptions cacheWorkingInMemory -mz_extraction_window 20 -ppm -mz_correction_function quadratic_regression_delta_ppm -min_rsq 0.8'. PyProphet (v.0.24.1) used the default parameters.

For all dataset in this study, only peptides with fixed modification of carbamidomethylation on cysteine and variable modification of oxidation on methionine were considered.

## 2.2 Model

For fragment presence prediction based on peptide sequence, the predictor should be able to capture the effects of all amino acids on the fragmentation and be able to handle sequential input with a variable length. Similar to Prosit, Alpha-Frag also used bi-directional gated recurrent memory

units (Bi-GRU, Chung et al., 2014) to address these requirements. Fig. 2a is the architecture overview of Alpha-Frag. A detailed description of the model is presented below.

**Inputs**: Including peptide sequence and precursor charge. Each amino acid of peptide and precursor charge will be embedded to a vector of dimension 32. Alpha-Frag supports precursor charge with no more than 4, as well as 20 natural amino acids and variable modification of oxidation on methionine. Each cysteine is treated as fixed modification of carbamidomethylation on itself.

**Model**: Alpha-Frag connects embedding layer to two layers of Bi-GRU. Dropout (Srivastava et al., 2014) is set to 0.5 between the two Bi-GRU layers. A self-attention layer (Lin et al., 2017) is used to merge the Bi-GRU outputs to a vector of dimension 64. Concatenating this vector with charge embedding vector, Alpha-Frag represents the precursor as a vector of dimension 96 in latent space. Afterwards, the 96-dimensional vector is converted to dimension 116 (Alpha-Frag considers peptide length maximum to 30 and only y/b fragment ions with no more than 2 charges. The maximum number of fragment ions is: 2 * 2 * 29 = 116) by a fully connected layer and a sigmoid layer.

**Outputs**: The output 116-dimensional vector is ordered as follows: y1+, y1++, b1+, b1++, y2+ and so on. Each value of the output vector represents the occurrence probability of the corresponding fragment ion. Alpha-Frag adopts 0.5 as the existence threshold. Only fragment ion with a predicted probability greater than 0.5 will be considered presence. Fragment ions at impossible dimensions (such as y17 for an 8-mer peptide) are set to zero directly.

Of note, we experimented with different dropout and dimension of each layer. However, this did not result in any significant gains on the performance. Therefore, we used the common parameter values as described above.

### 2.3 Training and test

PyTorch (v.1.1.0, https://pytorch.org) was selected to implement and train Alpha-Frag on one GPU Nvidia GTX 1060. We used the Adam optimizer with an initial learning rate of 0.001 and 1024 samples per batch. The loss function in training was binary cross entropy loss (BCE loss):

$$\mathcal{L}oss_{BCE} = \frac{1}{N}\sum_{i=0}^{N}[y_i \log x_i + (1-y_i)\log(1-x_i)] \tag{1}$$

where $x_i$ means the label of the corresponding fragment ion either 0 or 1, $y_i$ is the predicted probability, and $N$ is the length of the output vector. It should be noted that the fragment ions out-of-range or out-of-charge were masked and not involved in the calculation of loss. For the test dataset, we calculated the intersection over union (IoU, Fig. 2b) to evaluate the accuracy of prediction. Given a set of experimental present fragment ions $\mathcal{A}$ and a set of predicted present fragment ions $\mathcal{B}$, the IoU is defined as follows:

$$\mathrm{IoU}(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}\cup\mathcal{B}|} \tag{2}$$

For example, the experimental fragment ions is ‘y2+, y3+, y4+, y5++, b2+, b3+’, the predicted fragment ions is ‘y1+, y2+, y4+, y5+, b3+’, then the intersection between them is ‘y2+, y4+, b3+’ and the union is ‘y1+, y2+, y3+, y4+, y5+, y5++, b2+, b3+’. Hence, the IoU value is equal to 3/8. The larger the IoU, the more consistent the predicted fragment ions with the experimental fragment ions are.

# 3 Results

## 3.1 Bi-GRU based deep learning model

The details of the neural deep learning network have been described separately in method part (Fig. 2). For each epoch, we calculated the averaged BCE loss and the averaged IoU for the training and test dataset, respectively. As the epoch increased, the averaged BCE loss decreased and the averaged IoU increased, shown in Fig. 2c. After 20 epochs, the model converged and the training was terminated to balance the learning and the potential overfitting. At last, the averaged IoU reached 0.815, indicating that Alpha-Frag predicts fragment ion presence well for the test dataset.

## 3.2 Performance evaluation of fragment presence prediction

We hypothesized that fragment presence prediction based on intensity prediction is not sufficient to recover the fragment presence. We justified by the illustration of a representative precursor EPPLLLGVLHPNTK(3+) for its experimental spectrum, the predicted spectrum by Prosit (unless otherwise specified, Prosit works with the optimal NCE 35 in this study) and its predicted fragment ions by Alpha-Frag in Fig. 3. Although the Pearson correlation coefficient (PCC) between the experimental spectrum and the predicted spectrum by Prosit reached as high as 0.90, the predicted spectrum contained 3.06 (46 vs. 15) times as many fragment ions as the number of matched y/b-ions in the experimental spectrum. In other words, ~67.3% of predicted fragment ions by Prosit were not present. By contrast, Alpha-Frag predicted 14 fragment ions and 13 of which were matched to experimental spectrum. Therefore, the IoU between the experimental present fragment ions and the predicted present fragment ions by Alpha-Frag reached 0.81, illustrating the strong agreement between prediction and observation for fragment presence in this case.

To evaluate the performance of fragment presence prediction of Alpha-Frag comprehensively, we benchmarked it against theoretical enumeration model (hereinafter referred to as Enumeration model), pDeep3 (Tarn et al., 2020) and Prosit model using the three datasets: Ammar's, Huang's and

Bekker-Jensen's dataset. The Enumeration model outputs all possible y/b-ions with less length and no more charges (up to 2) than precursors. For the two intensity predictors pDeep3 and Prosit, we selected all non-zero intensity y/b-ions based on the result of intensity prediction. For Alpha-Frag, a probability of 0.5 was chosen as the threshold to distinguish for presence and absence of ions. The selection of threshold 0 for intensity cutoff and threshold 0.5 for probability cutoff were intuitive and dataset independent.

First, we compared the difference between the number of predicted fragment ions and the number of measured fragment ions both normalized by the number of all possible fragment ions. Considering the fact that y-ions and b-ions have different presence characteristic, the comparison were performed by three categories: y/b-ions, only y-ions and only b-ions. Meanwhile, to simplify the presentation, we defined normalized number of fragment ions (NNF) as follows:

$$\mathrm{NNF} = \left\{ \frac{\#\ \text{fragment ions}}{i \times (\text{length - 1}) \times \text{charge}} \right.$$

where length means peptide length, charge is the fragment charge up to 2 and no more than precursor charge, and $i$ is 2 for y/b-ions category, 1 for y-ion and b-ion categories. We plotted the distribution of NNF for the predicted fragment ions and the measured ions by boxplot in Fig. 4. As can be seen, in all three datasets, the result of Enumeration model was equal to one, and the median number of measured y-ions (an average of ~0.37) were more than b-ions (an average of ~0.15) both of which were expected. Meanwhile, both pDeep3 and Prosit overestimated the number of measured fragment ions, and Prosit generated more b-ions than pDeep3 for all three datasets. On the other hand, Alpha-Frag predicted less ions than pDeep3 and Prosit and was the closest to the number of measured ions, even less than it in the Bekker-Jensen's dataset. The comparison of NNF implied that Alpha-Frag is superior in predicting the number of fragment ions.

Besides the number of fragment ions, IoU can be used to check the consistency between the experimental fragment ions and the predicted fragment ions. As indicated by the violin plot in Fig. 4 for the IoU density distribution of the four predictors in three datasets, it is observed that the Enumeration model had the lowest median of IoU as it fails to filter ions that are not likely to be present. Prosit was better at predicting y-ions but worse at predicting b-ions compared to pDeep3. Alpha-Frag outperformed markedly over all the other predictors across the three datasets. In terms of median IoU of y/b-ions, Alpha-Frag achieved an average of ~204.7%, ~56.7% and ~44.1% improvements than Enumeration model, pDeep3 and Prosit, respectively. Moreover, we compared the running time of the predictors to evaluate the computing efficiency. In our device with CPU Intel i7-7700K 4 cores, Win10, 64 bit, 64 GB memory and GPU Nvidia GTX 1060, Enumerator model, pDeep3, Prosit and Alpha-Frag spent ~0s, 250.6s, 32.5s and 5.2s to predict fragment presence of 315,974 precursors of Berken-Jensen's dataset. In a nutshell, Alpha-Frag predicts fragment presence accurately and quickly.

### 3.3 Presence prediction improves peptide identification of DIA data

A major application of Alpha-Frag is to provide a priori information by predicting fragment presence for scoring in DIA identification. We conducted a workflow as shown in Fig. 5a to incorporate Alpha-Frag into the pipeline of OpenSWATH and PyProphet. Once OpenSWATH determined and scored the peak groups based on the spectra library, Alpha-Frag was performed to predict the fragment ions based on the peptide sequence and precursor charge. Then, depending on the apex retention time of the peak group, the single MS/MS spectrum was picked out whose retention time was closest to the apex. Afterward, auxiliary consistency scores were calculated based on the picked spectrum and the predicted fragment ions and further fed into PyProphet with the other scores (blue lines in Fig. 5a). At last, PyProphet merged all scores and reported the identification result.

Since there is no direct relationship between a peptide and a spectrum due to the DIA multiplexing nature, the calculation of the consistency between the picked out spectrum and the predicted fragment ions only takes the ions of the intersection of the two into account. Hence, we constructed the consistency scores as defined below (we called them qualitative scores because they did not correlate with fragment intensity):

$$\text{Qualitative scores} = \begin{cases} \text{NNF}_{\text{matched-predicted y/b-ions}} \\ \text{NNF}_{\text{matched-predicted y-ions}} \\ \text{NNF}_{\text{matched-predicted b-ions}} \end{cases}$$

Also, these qualitative scores were built by Enumeration model and Prosit (pDeep3 was no longer used as a comparison as it was inferior to Prosit in term of $\text{NNF}_{\text{y/b-ions}}$ shown in Fig. 4). Besides, quantitative scores (involving fragment intensity) based on the picked experimental MS/MS spectrum and the predicted spectrum by Prosit were also constructed as a comparison. Here, spectral angle (SA, Toprak et al., 2014; Gessulat et al., 2019) was used as a high sensitive similarity measure of two spectra:

$$\text{SA} = 1 - 2\frac{\cos^{-1}(\boldsymbol{S}_e \bullet \boldsymbol{S}_p)}{\pi}$$

where vector $\boldsymbol{S}_e$ and $\boldsymbol{S}_p$ represent the experimental spectrum and the predicted spectrum both with L2 normalization. The length of $\boldsymbol{S}_e$ and $\boldsymbol{S}_p$ are equal to the number of all theoretically possible fragment ions whose charge up to 2 and no more than precursor charge. Thus, the quantitative scores were defined as follows that were parallel to the qualitative scores:

$$\text{Quantitative scores} = \begin{cases} \text{SA}_{\text{y/b-ions}} \\ \text{SA}_{\text{y-ions}} \\ \text{SA}_{\text{b-ions}} \end{cases}$$

We then analyzed the DIA dataset using the scores defined as above with a multiple sets of comparison, including ‘OpenSWATH’, ‘OpenSWATH + Qual.(Enumeration)’, ‘OpenSWATH + Quan.(Prosit)’, ‘OpenSWATH + Qual.(Prosit)’, ‘OpenSWATH + Qual.(Alpha-Frag)’,

‘OpenSWATH + Qual.(Prosit) + Quan.(Prosit)’, ‘OpenSWATH + Qual.(Alpha-Frag) + Quan.(Prosit)’, where ‘OpenSWATH’ means raw OpenSWATH scores, ‘Qual.(Enumeration)’ means qualitative scores by Enumeration and so on. The number of identified precursors against the FDR was plotted in Fig. 5b. It is observed that: 1. Qualitative scores by Enumeration model could not lead to any improvements compared with the raw scores. This indicates the necessity in using qualitative scores of accurate fragment presence prediction to promote DIA identification. 2. Both qualitative scores and quantitative scores by Prosit were benefit for identification. At 1% FDR, ‘Qual.(Prosit)’ brought about 24.6%, 1.6%, 2.1% improvements over ‘Quan.(Prosit)’ for HeLa-0.5h, HeLa-1h, HeLa-1.5h, respectively, while in HeLa-2h, ‘Quan.(Prosit)’ brought about 5.3% over ‘Qual.(Prosit)’. It seems that the qualitative scores by Prosit was more profitable in short gradient compared to the quantitative scores. 3. At 1% FDR, ‘Qual.(Alpha-Frag)’ outperformed ‘Qual.(Prosit)’ and achieved 29.3%, 20.2%, 13.1% and 10.1% gains for HeLa-0.5h, HeLa-1h, HeLa-1.5h and HeLa-2h, respectively, and ‘Qual.(Alpha-Frag)’ improved more as the gradient becomes shorter. Since the difference of fragment presence prediction between Prosit and Alpha-Frag was mainly about the low-intensity ions, we confirmed from this result that these low-intensity ions could also influence the identification of DIA. 4. Not surprisingly, a combination of qualitative scores and quantitative scores obtained the greatest improvements and ‘Qual.(Alpha-Frag) + Quan.(Prosit)’ produce 20.0%(HeLa-0.5), 13.8%(HeLa-1h), 9.6%(HeLa-1.55) and 8.6%(HeLa-2h) improvements compared to ‘Qual.(Prosit) + Quan.(Prosit)’. These results revealed that both qualitative and qualitative scores by predictive models could improve the DIA identification and qualitative scores were more fit to short gradient. Particularly, Alpha-Frag yielded 10.1%-29.3% improvements as it makes better fragment presence prediction compared to Prosit.

To assess which subscore of qualitative scores or quantitative scores had more impact on the identification, we carried out a discriminant experiment (Reiter et al., 2011) for the same four DIA files. After identified by ‘PyProphet + Qual.(Alpha-Frag) + Quan.(Prosit)’, the peptides at <1% FDR

were chosen as the positive class, and the equal number of random decoy peptides were as the negative class. Based on the positive and negative class, we plotted the receiver-operating characteristic (ROC) curves for each subscore (Fig. 5c). For HeLa-0.5h, HeLa-1h and HeLa-1.5h, qualitative score of y-ions had the largest area under the curve (AUC) value while quantitative score of y/b-ions had the highest discriminating power for HeLa-2h. Although 'Qual.(Alpha-Frag)' outperformed 'Quan.(Prosit)' at 1% FDR in HeLa-2h, the reverse that 'Quan.(Prosit)' had more identified precursors than 'Qual.(Alpha-Frag)' at 0.1% FDR (59305 vs. 58689). Taken together, we argued that quantitative scores were slightly more important than the qualitative scores for HeLa-2h. These results implied that the subscore of y-ions in qualitative scores and subscore of y/b-ions (excluding short gradient 0.5h) in quantitative scores had the greatest effect than other subscores on DIA identification, and quantitative scores was more suitable for long gradient (2h).

In order to explore the reasons why qualitative scores were more important for shorter gradient, we selected the peptides (48084 precursors) that could be consistently identified by 'PyProphet + Qual.(Alpha-Frag) + Quan.(Prosit)' under the four gradients and examined their qualitative and quantitative scores distribution as well as the corresponding median variation normalized by median value of HeLa-2h. As can be seen from Fig. 6, with gradient decreased from 2h to 1h, the median of qualitative scores slightly increased due to mismatching and the median of quantitative scores consistently declined as interferences. But at gradient 0.5h, the qualitative scores were descending oppositely instead of ascending. We suspected that the fragment suppression occurred at this gradient, and the degree of suppression was higher than that of mismatching. This phenomenon was also found in Tsou et al., 2015 (They found that the number of fragment ions in DIA are less than in DDA as fragment suppression which could also be validated by our Fig. 4 and Fig. 6). We cautiously believe that the degree of mismatching was weaker than the degree of interferences when gradient descending, which may be the reason why qualitative scores were more suitable for relatively short

gradients (0.5h-1.5h). Meanwhile, in the case of 0.5h gradient or below, the applicability of qualitative scores needs further study due to the fragment suppression.

Last but not least, we recorded the running time to finish the auxiliary scoring strategy. ‘Qual.(Alpha-Frag)’ and ‘Quan.(Prosit)’ spent 6.63min and 9.33min, respectively, in the case of HeLa-1h where OpenSWATH generated 818,273 candidate peak groups.

# 4 Conclusion

In this study, we proposed Alpha-Frag, a deep neural network used to predict fragment presence ions and that surpassed substantially other existing predictors. In the validation datasets, Alpha-Frag achieved an average of ~204.7%, ~56.7% and ~44.1% improvements in terms of median IoU of y/b-ions than Enumeration model, pDeep3 and Prosit, respectively. In the accompanying application of Alpha-Frag, qualitative scoring strategy was performed to improve the identification and yielded an increase of 10.1%-29.3% improvements for the DIA dataset (with gradient 0.5h-2h) compared to the same strategy but using Prosit. Furthermore, we explored the performance of qualitative and quantitative scores under different gradients. Our experiments showed that quantitative scores were more suitable for long gradients, while qualitative scores were more profitable for short gradients. Of note, we do not rule out revising this conclusion if there are more accurate scoring strategy for the qualitative and quantitative scores. All these highlight that fragment presence prediction is an effective complement to fragment intensity prediction.

So far, only tryptic peptides, peptides with limited length and with no modifications except for methionine oxidation were considered. Next, as more data for non-tryptic, length longer and modified peptides and chemical labels become available, we will update our model. Further, we plan to divide the spectrum prediction into two phases: fragment presence prediction followed by corresponding intensity prediction to approximate the fragmentation better. We believe this will make the predicted spectrum more accurate and peptide identification more confident for DIA data.

## Acknowledgement

We thank Weigang Ge for checking the code, Dr. Tiannan Guo, Dr. Chunhui, Yuan and Dr. Luang Xu for commenting on the manuscript and Dr. Wenfeng Zeng for his advice on the use of pDeep3.

## Funding

This work was supported by the National Natural Science Foundation of China, No. 61761136005.

Conflict of Interest: none declared.

# References


1. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537, 347–355
2. Gillet L C, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis[J]. Molecular & Cellular Proteomics, 2012, 11(6).
3. Röst H L, Rosenberger G, Navarro P, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data[J]. Nature biotechnology, 2014, 32(3): 219-223.
4. Zhang F, Ge W, Ruan G, et al. Data‐Independent Acquisition Mass Spectrometry‐Based Proteomics and Software Tools: A Glimpse in 2020[J]. Proteomics, 2020, 20(17-18): 1900276.
5. Rosenberger G, Koh C C, Guo T, et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS[J]. Scientific data, 2014, 1: 140031.
6. Wen B, Zeng W F, Liao Y, et al. Deep learning in proteomics[J]. Proteomics, 2020, 20(21-22): 1900335.
7. Gessulat S, Schmidt T, Zolg D P, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning[J]. Nature methods, 2019, 16(6): 509-518.
8. Perkins D N, Pappin D J C, Creasy D M, et al. Probability‐based protein identification by searching sequence databases using mass spectrometry data[J]. ELECTROPHORESIS: An International Journal, 1999, 20(18): 3551-3567.
9. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
10. Zhou X X, Zeng W F, Chi H, et al. pDeep: predicting MS/MS spectra of peptides with deep learning[J]. Analytical chemistry, 2017, 89(23): 12690-12697.
11. Tiwary S, Levy R, Gutenbrunner P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis[J]. Nature methods, 2019, 16(6): 519-525.
12. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
13. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
14. Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International conference on machine learning. 2015: 2048-2057.
15. Toprak U H, Gillet L C, Maiolica A, et al. Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics[J]. Molecular & Cellular Proteomics, 2014, 13(8): 2056-2071.
16. Zolg D P, Wilhelm M, Schnatbaum K, et al. Building ProteomeTools based on a complete synthetic human proteome[J]. Nature methods, 2017, 14(3): 259-262.
17. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification[J]. Nature biotechnology, 2008, 26(12): 1367-1372.
18. Zolg D P, Wilhelm M, Yu P, et al. PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration[J]. Proteomics, 2017, 17(21): 1700263.

19. Ammar C, Berchtold E, Csaba G, et al. Multi-reference spectral library yields almost complete coverage of heterogeneous LC-MS/MS data sets[J]. Journal of proteome research, 2019, 18(4): 1553-1566.
20. Huang T, Bruderer R, Muntel J, et al. Combining precursor and fragment information for improved detection of differential abundance in data independent acquisition[J]. Molecular & Cellular Proteomics, 2020, 19(2): 421-430.
21. Bekker-Jensen D B, Kelstrup C D, Batth T S, et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes[J]. Cell systems, 2017, 4(6): 587-599. e4.
22. Bruderer R, Bernhardt O M, Gandhi T, et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results[J]. Molecular & Cellular Proteomics, 2017, 16(12): 2296-2309.
23. Teleman J, Röst H L, Rosenberger G, et al. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data[J]. Bioinformatics, 2015, 31(4): 555-562.
24. Parker S J, Rost H, Rosenberger G, et al. Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry[J]. Molecular & Cellular Proteomics, 2015, 14(10): 2800-2813.
25. Demichev V, Messner C B, Vernardis S I, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput[J]. Nature methods, 2020, 17(1): 41-44.
26. Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
27. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.
28. Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[J]. arXiv preprint arXiv:1703.03130, 2017.
29. Tarn C, Zeng W F, Fei Z C, et al. pDeep3: Towards More Accurate Spectrum Prediction with Fast Few-Shot Learning[J]. bioRxiv, 2020.
30. Reiter L, Rinner O, Picotti P, et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments[J]. Nature methods, 2011, 8(5): 430-435.
31. Tsou C C, Avtonomov D, Larsen B, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics[J]. Nature methods, 2015, 12(3): 258-264.

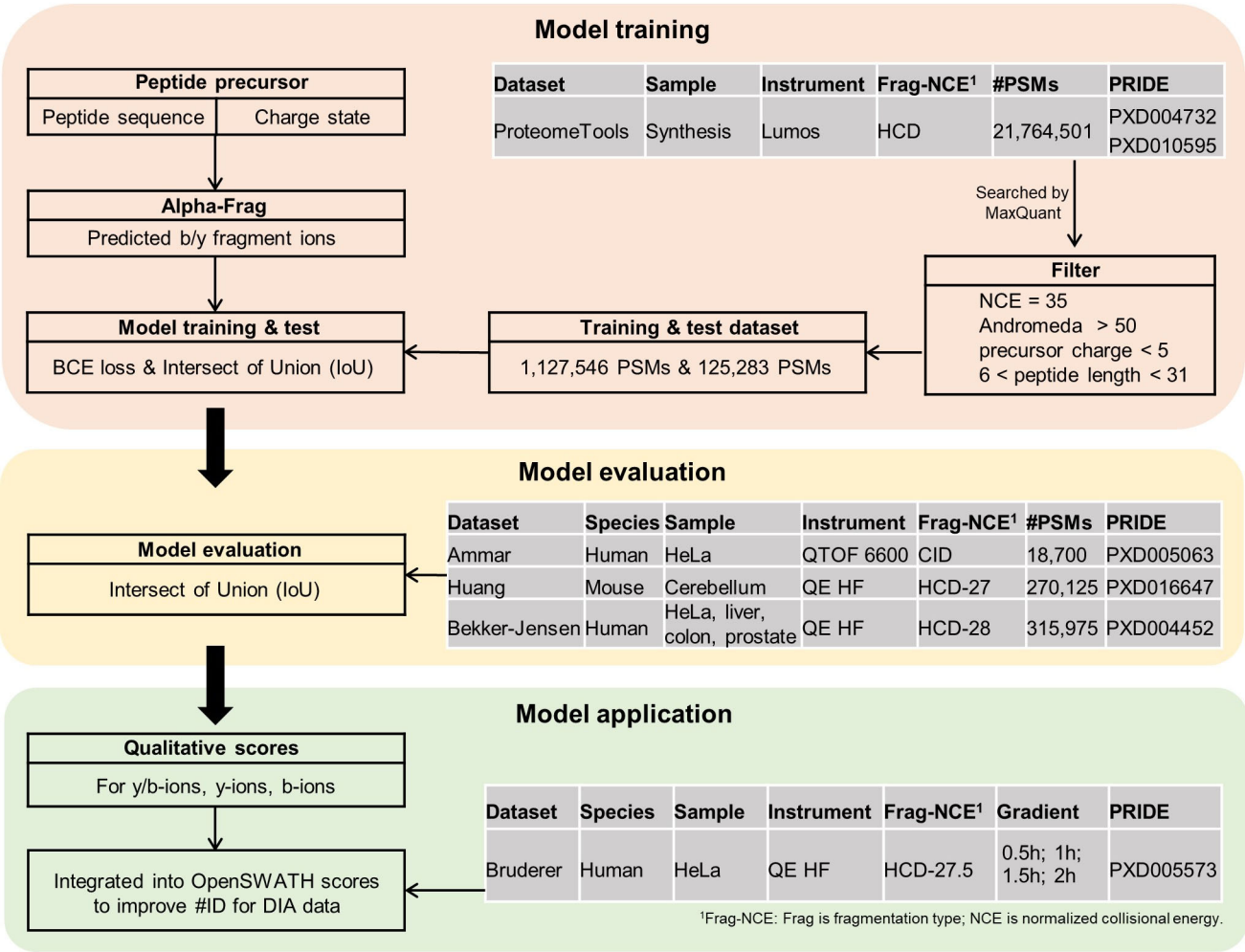


**Fig. 1. Study design.** From top to bottom, the three panels indicate model training, model evaluation and model application, respectively. The table in each panel shows the details of the corresponding datasets.

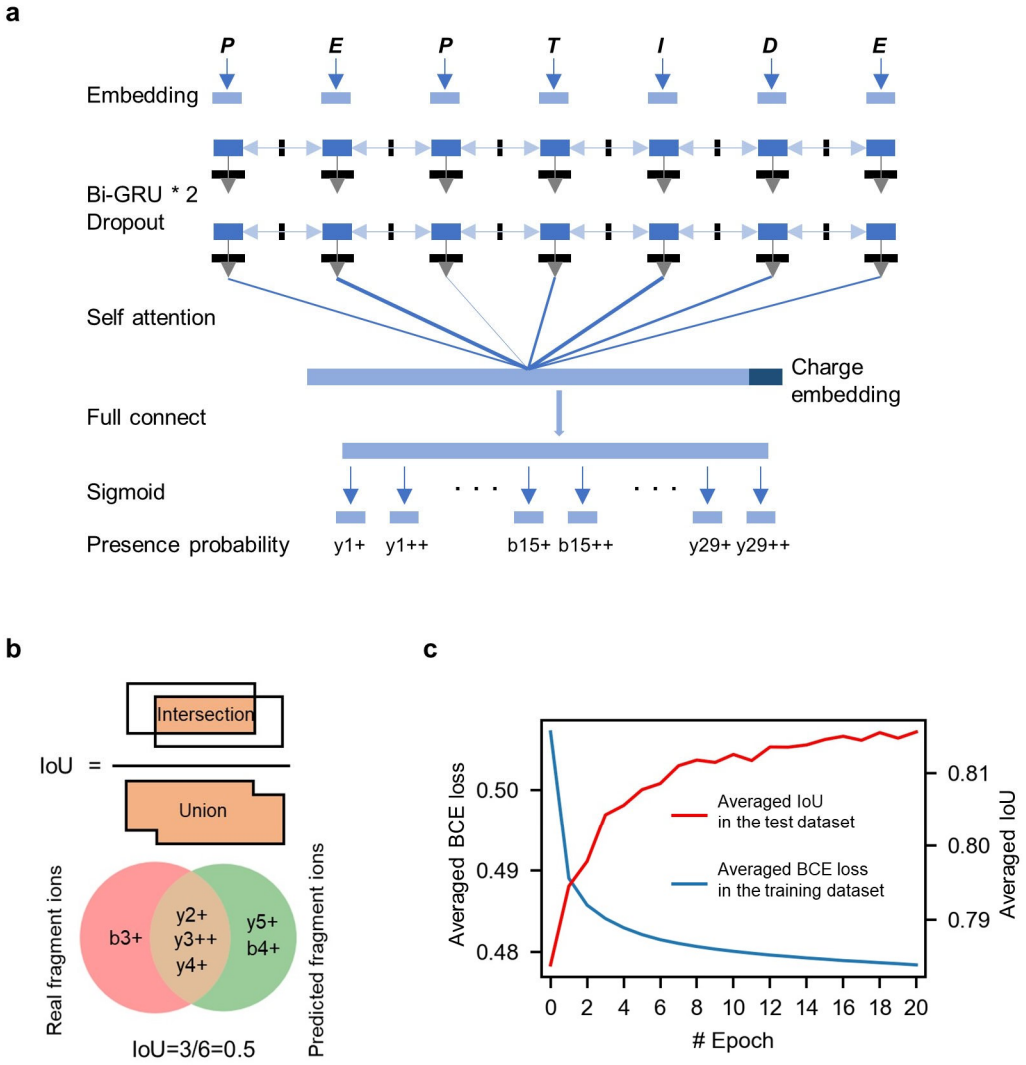


**Fig. 2. The model of Alpha-Frag and its training and test. a)** Bi-GRU based Alpha-Frag model. **b)** Illustration of intersection over union (IoU) which is used to compare the consistency between two sets. **c)** The averaged loss curve in the training dataset and the averaged IoU curve in the test dataset.

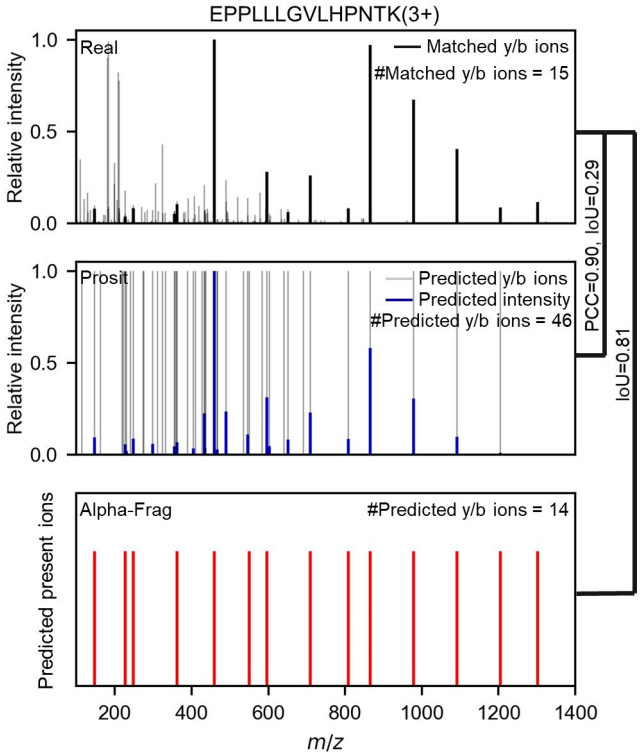


**Fig. 3. An example of the experimental spectrum, the predicted spectrum by Prosit (NCE=35) and the predicted fragment ions by Alpha-Frag of precursor EPPLLLGVLHPNTK(3+).** The matched y/b ions in the experimental spectrum are marked in jet black lines (scan 53773, 01974C_BA1-TUM_missing_first_1_01_01-DDA-1h-R4.raw, PXD010595. These matched y/b ions are determined by MaxQuant at 1% FDR). The predicted spectrum by Prosit contains 46 y/b ions with non-zero intensities which are marked in light black. The blue lines mean the match between the predicted spectrum and the experimental spectrum. We observed that 67.3% of predicted fragment ions are not present in the experimental spectrum, while the Pearson correlation coefficient (PCC) between them reaches 0.90. On the other hand, the IoU between experimental spectrum and predicted fragment ions by Alpha-Frag reaches 0.81.

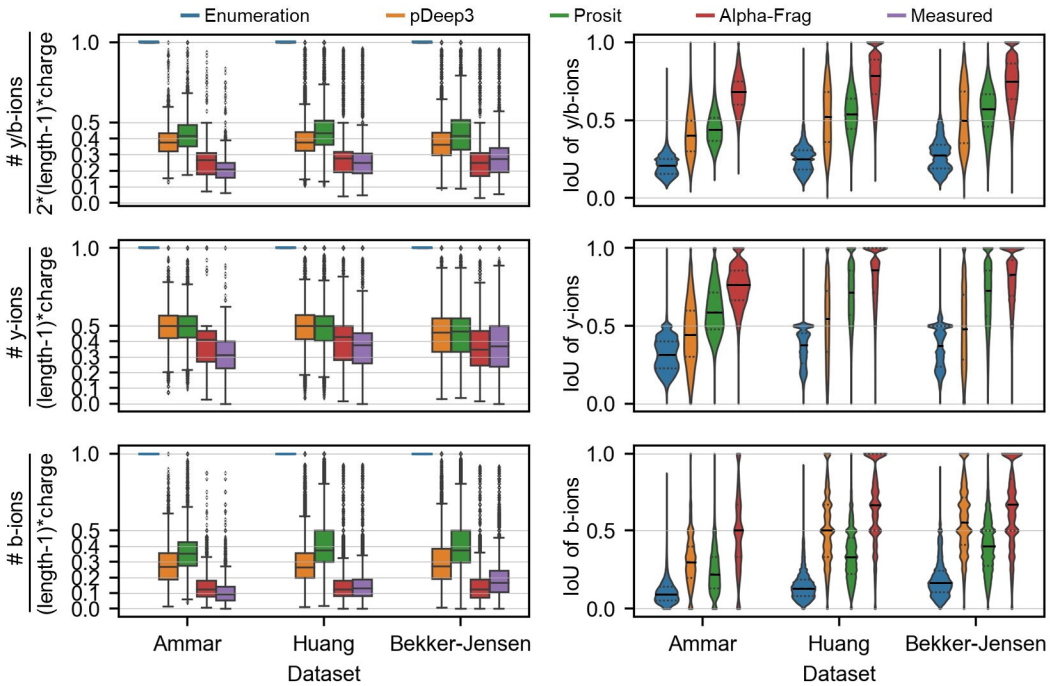


**Fig. 4. Performance comparison of fragment presence predictors. a)** Box plot to show the distribution of the number of predicted fragment ions and the number of measured fragment ions both normalized by peptide length and fragment charge which up to 2 and no more than precursor charge. The box indicates the interquartile range (IQR), its whiskers 1.5× IQR values, and the black line the median. **b)** Violin plot of IoU in the three evaluation datasets, and the black line the median.

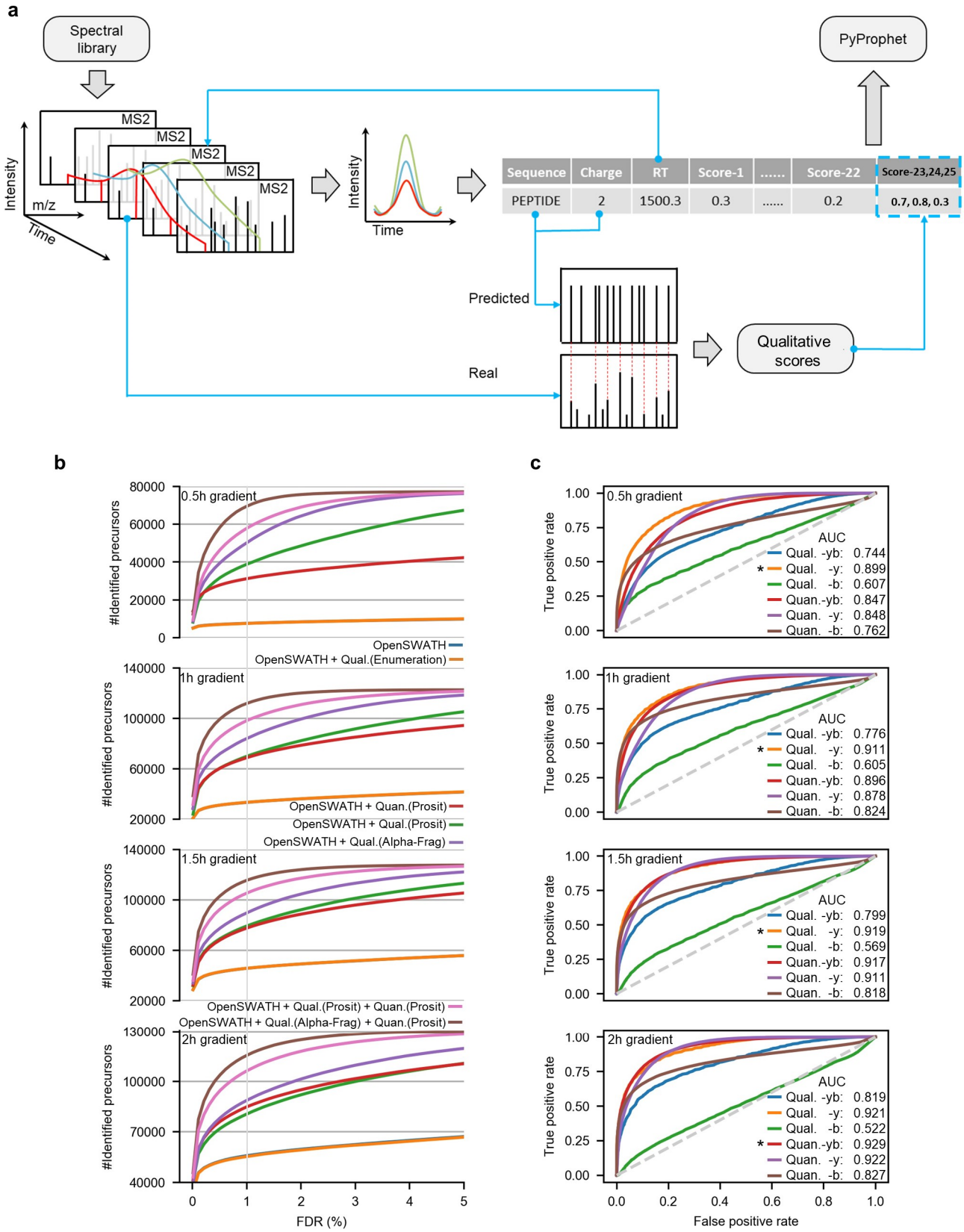


**Fig. 5. Application of fragment presence prediction to DIA identification. a)** Workflow. The non-colored lines represent the workflow of DIA identification by OpenSWATH and PyProphet. The blue lines represent the process of integrating the qualitative scores into OpenSWATH-PyProphet. **b)**

Comparison of identification performance on different gradients. 'OpenSWATH' means the raw OpenSWATH scores. 'Qual.(Enumeration)' means qualitative scores based on Enumeration model and so on. **c)** ROC plots and AUC values for each component within qualitative scores based on Alpha-Frag and quantitative scores based on Prosit. 'Qual. -yb' means qualitative score of y/b-ions and so on. The asterisk indicates the largest AUC.

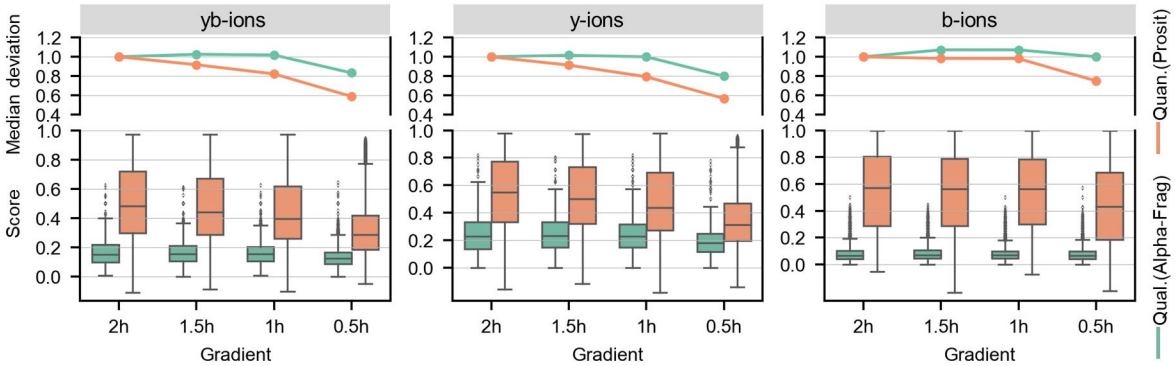


**Fig. 6. Distribution and median deviation of qualitative scores and quantitative scores.** The data was collected based on the peptides (48084 precursors) that could be consistently identified by 'PyProphet + Qual.(Alpha-Frag) + Quan.(Prosit)' under the four gradients. Median deviation is normalized by the median value from 2h gradient.