DeepRTAlign: toward accurate retention time alignment for large cohort mass spectrometry data analysis

Yi Liu[1,2], Yingying Zhang[2,3], Yuanjun Zhai[2], Fuchu He[2,4], Yunping Zhu[2*], Cheng Chang[2,4*]

1. Faculty of Environment and Life, Beijing University of Technology, Beijing 100023, China.
2. State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.
3. Chongqing Key Laboratory on Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
4. Research Unit of Proteomics Driven Cancer Precision Medicine, Chinese Academy of Medical Sciences, Beijing 102206, China.

* To whom correspondence should be addressed:

Yunping Zhu, Email: zhuyunping@gmail.com

Cheng Chang, Email: changchengbio@163.com

ABSTRACT

Retention time (RT) alignment is one of the crucial steps in liquid chromatography mass spectrometry (LC-MS)-based proteomic and metabolomic experiments, especially for large cohort studies. And it can be achieved using computational methods; the most popular methods are the warping function method and the direct matching method. However, the existing tools can hardly handle monotonic and non-monotonic RT shifts simultaneously. To overcome this, we developed a deep learning-based RT alignment tool, named DeepRTAlign, for large cohort LC-MS data analysis. It firstly performs a coarse alignment by calculating the average time shift between any two samples and then uses RT and intensity as the main features to train its deep learning-based model. We demonstrate that DeepRTAlign has improved performances on several proteomic and metabolomic datasets especially when handling complex samples by benchmarking it against current state-of-the-art approaches. Ultimately, we show that DeepRTAlign can improve the identification sensitivity of MS data without compromising the quantitative precision compared to MaxQuant, FragPipe and DIA-NN with match between runs. In a single-cell data-independent acquisition MS dataset, DeepRTAlign can align 298 (42.7%) more peptides on average than the existing popular tool DIA-NN in each cell.

Introduction

Liquid chromatography (LC) is usually coupled with mass spectrometry (MS) in proteomics experiments to separate the complex samples. The retention time (RT) of each analyte in MS data usually more or less have shifts due to multiple reasons including matrix effects and instrument performances[1]. Thus, in any experiment involving multiple samples, corresponding analytes must be mapped before quantitative, comparative, or statistical analysis. This process is called correspondence[2]. In other words, this problem can be defined as finding the "same compound" in multiple samples. Generally, in proteomics, correspondence can be done based on peptide identifications. However, taken the test data in this study as an example, only 10%-15% precursors have the corresponding identifications due to the data-dependent ion selection process in data-dependent acquisition (DDA) mode. Even for the data-independent acquisition (DIA) data, there remains a number of precursors (potential peptides) unidentified, which are not able to be considered in the subsequent analysis due to the complex MS/MS spectra[3]. Most existing tools for DDA and DIA data analysis, such as MaxQuant[4], PANDA[5], FragPipe[6,7] and DIA-NN[8], perform RT alignment using the match between runs (MBR) function (also called cross-align function) to transfer the identified sequences to the unidentified precursors between any two LC-MS runs. Although, MBR can increase the total number of identifications to some extent, it is integrated in specific software tools and relies on the identified peptides, which limits its further application in clinical proteomics research to explore new biomarkers from unidentified precursors. In metabolomics, feature alignment is a prerequisite for identification and quantification. In theory, the accuracy of feature alignment depends on the m/z and RT information in MS data. Currently, high-resolution mass spectrometers can limit the m/z shift to less than 10 ppm. Thus, RT alignment is especially important for accurate analysis of large-scale data in proteomics and metabolomics research.

There are two types of computational methods for RT alignment. One is called the warping method. Warping models first correct the RT shifts of analytes between runs by a linear or non-linear warping function[2,9]. There were several existing popular alignment tools based on this method, such as XCMS[10], MZmine 2[11] and OpenMS[12]. However, this warping method is not able to correct the non-monotonic shift because the warping function is monotonic[9]. Another kind of method is the direct matching method, which attempts to perform correspondence only based on the similarity between specific signals from run to run, without warping function. The representative tools include RTAlign[13], MassUntangler[14] and Peakmatch[15]. The performances of the existing direct matching tools are reported inferior to the tools using warping functions due to the uncertainty of MS signals[2]. Either way, these tools mentioned above can hardly handle both monotonic and non-monotonic RT shifts. Thus, machine learning or deep learning techniques are applied to solve this issue. Li et al. applied Siamese network for accurate peak alignment in gas chromatography-mass spectrometry data from complex metabolomic samples[16]. But we found the Siamese network could not perform well without MS/MS information (**Supplementary Notes**), limiting its application to the precursors without the corresponding identifications, which usually do not have MS/MS spectra or only have low-quality MS/MS spectra.

Here, we present a deep learning-based RT alignment tool, named DeepRTAlign, for large cohort LC-MS data analysis. Combining a coarse alignment (pseudo warping function) and a deep learning-based model (direct matching), DeepRTAlign can deal with non-monotonic shifts as well as monotonic shifts. We have demonstrated its high accuracy and sensitivity in several proteomic and metabolomic datasets compared with existing popular tools. Further, DeepRTAlign allows us to apply MS features directly and accurately to downstream biological analysis, such as biomarker discovery or prognosis prediction, which can be considered as a complement to the traditional protein-centric methods (**Supplementary Notes**).

Methods and Materials

Datasets

As shown in **Table S1**, all the datasets used in training and testing the deep learning model in DeepRTAlign (i.e., one training set and seven test sets) were collected from the following papers: (1) The training set HCC-T is from the tumor samples of 101 early-stage hepatocellular carcinoma (HCC) patients in Jiang et al.'s paper[17]. The corresponding non-tumor dataset (HCC-N) was considered as a test set in this study. (2) Dataset HCC-R was from the tumor samples of 11 HCC patients with liver transplantation in the paper of Bhat et al.[18]. (3) Datasets UPS2-M and UPS2-Y were from the paper of Chang et al.[19], which were constructed by spiking 48 UPS2 standard proteins (Proteomics Dynamic Range Standard, Sigma-Aldrich) into mouse cell digestion mixture (UPS2-M) and yeast digestion mixture (UPS2-Y), respectively. (4) Dataset EC-H was from the paper of Shen et al.[20], which was constructed by spiking *Escherichia coli* cell digestion mixture into human cell digestion mixture. (5) Dataset AT was about the *Arabidopsis thaliana* seeds (AT) from the paper of Ginsawaeng et al.[21]. (6) Dataset SC was a single-cell (SC) proteomic dataset including 18 HT22 cells without the treatment of nocodazole from the paper of Li et al.[22].

Further, we tested the generalizability boundary of DeepRTAlign on the seven test sets and seven other datasets (five metabolomic datasets and two proteomic datasets). The five metabolomic datasets include (1) Dataset NCC19, a large-scale plasma analysis about SARS-CoV-2 from the paper of Barberis et al.[23]; (2) Dataset SM1100, standard mixtures consisting of 1100 compounds, from the paper of Li et al.[24]; (3) Dataset MM, which was from the *Mus musculus* samples in the paper of Wase et al.[25]; (4) Dataset SO, which was from soil samples in the paper of Swenson et al.[26]; (5) Dataset GUS, which was about global untargeted serum metabolomic samples from the paper of Gibson et al.[27]. The two proteomic datasets are (1) Dataset MI, which was about mouse intestinal proteomes from the paper of Lichtman et al.[28]; (2) Dataset CD, which was obtained from the gut microbiota of patients with Crohn's disease in the paper of Mottawea et al.[29].

Workflow of DeepRTAlign

<Figure 1>

The whole workflow of DeepRTAlign is shown in **Figure 1**, which can be divided into two parts, i.e., the training part and the application part. The training part contains the following steps.

(1) Precursor detection and feature extraction. Taking raw MS files as input, precursor detection and its feature extraction were performed using an in-house developed tool XICFinder, which is a MS feature extraction tool similar to Dinosaur[30]. The algorithm of XICFinder is based on our quantitative tool PANDA[5]. Using the Component Object Model (COM) of MSFileReader, it can handle Thermo raw files directly. XICFinder first detects isotope pattern in each spectrum. Then, the isotope patterns detected in several subsequent spectra are merged to a feature. A mass tolerance of 10 ppm was used in this step. The precursor ions with more than one MS/MS spectrum were stored in each MS file for subsequent analysis.

(2) Coarse alignment. First, the RT in all the samples will be linearly scaled to a certain range (e.g., 80 min in this study, as the RT range of training dataset HCC-T is 80 min). Second, for each m/z, the feature with the highest intensity is selected to build a new list for each sample. Then, all the samples except an anchor sample (we considered the first sample as the anchor in this study) will be divided into pieces by user-defined RT window (we used 1 minute in this study). All the features in each piece are compared with the features in the anchor sample (mass tolerance: 0.01 Da). If the same feature does not exist in the anchor sample, this feature is ignored. Then, the RT shift is calculated for each feature pair. For all the features in each piece, the average RT shift is calculated. Each piece is aligned with the anchor sample by adding its average RT shift.

(3) Input vector construction. Only the RT and intensity values of each feature are considered when constructing the input vector. As shown in **Figure 1a**, we consider two adjacent features (according to RT) before and after the target feature corresponding to a peptide. Part 1 and part 4 are the original values of RT and

intensity. Part 2 and part 3 are the difference values between two samples. Thus, each feature-feature pair will be transferred to a 5×8 vector as the input of the deep neural network (**Figure S1**).

(4) Deep neural network (DNN). The DNN model in DeepRTAlign contains three hidden layers (each has 5000 neurons), which is used as a classifier that distinguishes between two types of feature-feature pairs (i.e., the two features should be or not be aligned). Finally, a total of 200,000 feature-feature pairs were collected from the HCC-T dataset based on the Mascot identification results (mass tolerance: ± 10 ppm, RT tolerance: ± 5 min). 100,000 of them are collected from the same peptides, which should be aligned (labeled as positive). The other 100,000 are collected from different peptides, which should not be aligned (labeled as negative). These 200,000 feature-feature pairs were used to train the DNN model. It should be noted that it is not necessary to know the peptide sequences corresponding to the features when performing feature alignment. The identification results of several popular search engines (such as Mascot, MaxQuant and FragPipe) are only used as ground truths when benchmarking DeepRTAlign.

(5) The hyperparameters in DNN. BCELoss function in Pytorch is used as the loss function. Batch size is 1000, and the number of epochs is 400. The initial learning rate is set to 0.001, and is multiplied by 0.1 every 100 epochs. All other parameters are kept by default in Pytorch.

(6) Parameter evaluation. Network parameters used were examined on the training set (HCC-T) by 10-fold cross validation and the best parameters were selected based on the cross-validation results (**Table S2** and **Table S3**). And the trained model was evaluated on several independent test sets (**Table S4**). These results demonstrated that there is no overfitting in the DNN model.

In the application part (**Figure 1b**), DeepRTAlign directly supports the results of four MS feature extraction tools, i.e., Dinosaur[30], MaxQuant, OpenMS and XICFinder as input. Feature lists from other tools (such as MZmine 2) can be used after format conversion refer to the formats of the four tools mentioned above. In this part, feature lists will first go through the coarse alignment step and the input vector construction step as same as those in the training part. Then, the constructed input vectors will be fed into the trained DNN model. According to the classification results of the DNN model, DeepRTAlign will output an aligned feature list for further analysis.

Machine learning models for evaluation

To make a systematical evaluation of our DNN model's performance, we compared it with several popular machine learning methods, i.e., random forests (RF), k-nearest neighbors (KNN), support vector machine (SVM) and logistic regression (LR). The parameters for each machine learning methods were set after optimization. The parameters of the RF model are max feature number 0.2, number of estimators 100, max depth 20 and the tree number in the forest 10. For KNN, the k is set to 3. For SVM, the kernel function is set to non-linear kernel. For LR, the penalty is L2. All the other parameters are default values in scikit-learn. In total, we trained four machine learning models (named as RF, KNN, SVM and LR) as references in this study.

Tools for alignment comparison

As shown in **Table S5**, all the alignment tools can be classified into three types based on the input information required. The representative tools in each type were compared with DeepRTAlign in this study.

First, two existing popular alignment tools (MZmine 2 and OpenMS) were used for alignment comparison because the two tools showed the best precision and recall in assessments of Lange et al.'s paper[31] and Pluskal et al.'s paper[11]. The recommended parameters in the official user manuals of MZmine 2 and OpenMS were used. When comparing with MZmine 2 or OpenMS, we considered the Mascot identification results corresponding to MZmine 2 or OpenMS features (mass tolerance: ± 10 ppm, RT tolerance: ± 5 min) as the ground truth, respectively.

The precision formula is:

$$\text{Precision} = \frac{1}{N}\sum_{k=1}^{N}\frac{A_k \cap G_k}{A_k}$$

The recall formula is:

$$\text{Recall} = \frac{1}{N}\sum_{k=1}^{N}\frac{A_k \cap G_k}{G_k}$$

N is the sample pair number. A_k is the aligned feature number in the k_{th} sample pair. G_k is the ground truth of the k_{th} sample pair.

Second, Quandenser[32], as an alignment method that requires MS/MS information, was also compared with DeepRTAlign on UPS2-M and UPS2-Y datasets. The peptide identification results (Mascot) after quality control (false discovery rate < 1% at both peptide and protein levels) were considered as our ground truth. PepDistiller[33] was used to perform quality control for Mascot identification results.

Third, we compared DeepRTAlign with MaxQuant, FragPipe and DIA-NN for DDA and DIA data analysis, respectively. For DDA data (UPS2-M, UPS2-Y and EC-H), the peptide identification results (Mascot, MaxQuant or FragPipe) after quality control (false discovery rate < 1% at both peptide and protein levels) were considered as the ground truth. For DIA data, we used a single-cell proteomic dataset obtained from 18 HT22 cells without nocodazole treated in Li et al.'s paper[22]. All the parameters are kept the same as described in Li et al.'s paper. We used Dinosaur to extract MS features in each cell and DeepRTAlign to align the features in all the 18 cells, compared with the aligned results of DIA-NN with and without MBR function.

Dataset simulation for generalizability evaluation

We further generated multiple simulated datasets with different RT shifts (considered as noise) for each real-world dataset. Here is the dataset simulation procedure. (1) All the features are extracted in each dataset using OpenMS to form an original feature list. Features with charges 2-6 are considered for proteomic data and features with charges 1-6 are considered for metabolomic data. (2) A RT shift based on a normal distribution with increasing standard deviations (σ = 0, 0.1, 0.3, 0.5, 0.7, 1, 3, 5) for each mean value (μ = 0, 5 and 10 minutes) are added in each feature to form a new feature list by modifying the featureXML file generated by OpenMS. (3) The new feature list with artificial RT shifts is aligned to the original feature list by DeepRTAlign and OpenMS. In theory, each feature with a RT shift in the new feature list should be aligned to the same feature in the original list. (4) The precision and recall values were calculated to evaluate the generalizability boundary of DeepRTAlign.

Data availability

Datasets HCC-T and HCC-N can be downloaded from the iProX database[34] under accession number IPX0000937000 or PXD006512. Datasets UPS2-M and UPS2-Y can be downloaded from the iProX database under the accession number IPX00075500 or PXD008428. Datasets HCC-R, EC-H, AT, SC, MI and CD can be downloaded from the PRIDE database[35] under the accession numbers PXD022881, PXD003881, PXD027546, PXD025634, PXD002838 and PXD002882, respectively. Datasets NCC19, SM1100, MM, SO and GUS can be downloaded from the MetaboLights database[36] under the accession numbers MTBLS1866, MTBLS733, MTBLS5430, MTBLS492 and MTBLS650, respectively.

Results and Discussions

Model evaluation on the training set and the test sets

To optimize the network parameters in the DNN model, the 10-fold cross validation results on the training set (HCC-T) were examined and the best parameters were selected based on the cross-validation results (**Table S2** and **Table S3**). Then, to benchmark the DNN model, we additionally trained four models using several popular machine learning methods (RF, KNN, SVM and LR) on the same training set (HCC-T). The test results of our DNN model and all the other machine learning models were shown in **Table S4**. We can find that our DNN model owned the highest AUC compared with other models. Although the DNN model is trained on the HCC-T dataset, it achieved a good generalizability and can be applied to other datasets with different sample types or species. RF shows a slightly higher AUC than DNN on UPS2-Y datasets. We think this is because the RT shift density distribution of UPS2-Y is similar to HCC-T (**Figure S2**). In general, DNN has a better generalization performance and RF can be an alternative solution when computing resources are limited.

Ablation analysis

We performed the ablation analysis to evaluate the DNN model in DeepRTAlign on different test sets (HCC-N, HCC-R, UPS2-M, UPS2-Y, EC-H, AT and SC). For the coarse alignment step, we have shown that there are no obvious differences when using different samples as the anchor sample (**Table S6**). We also tested the performance of DeepRTAlign with or without coarse alignment step (**Table S7**). We can find that our DNN model with the coarse alignment step owned a higher AUC compared with the same DNN model without coarse alignment.

Then, we evaluated the importance of each feature used in the DNN model of DeepRTAlign by replacing them with random values. As shown in **Table S8**, replacing RT (the "RT" column in **Table S8**) with random values obtained a smaller AUC than replacing intensity (the "Intensity" column in **Table S8**), indicating that RT is a more important feature than intensity. Moreover, we found the differences of RT and intensity in the samples to be aligned (the "Difference" column in **Table S8**) are more important than the original RT and intensity values ("Original" column in **Table S8**). These results show that the designs of our input vector (**Figure S1**) and DNN model are reasonable.

Comparison with existing alignment tools

According to the information required, all the alignment tools can be divided into three types (**Table S5**). DeepRTAlign, MZmine 2 and OpenMS only need MS information. Quandenser requires both MS and MS/MS information. In addition, MaxQuant, FragPipe and DIA-NN with MBR functions require identification results for alignment.

Comparison with MZmine 2, OpenMS

First, DeepRTAlign was compared with two other popular MS-based only alignment tools MZmine 2 and OpenMS on proteomic datasets. As shown in **Figure 2**, we can find that DeepRTAlign combining with the feature extraction methods in OpenMS and MZmine 2 showed a further improvement in both precision and recall. Although MZmine 2 showed a slightly higher precision than DeepRTAlign on the EC-H dataset, we think this is due to the fewer features extracted by MZmine 2 (**Table

S9). The noise threshold in MZmine 2 is the key parameter associated with the number of extracted features. But it is difficult for users to choose. Lowering the noise threshold will further make its extraction time to be intolerable. The focus of this work is not to compare the pros and cons of different feature extraction tools, so we chose relatively high noise thresholds (1.0E6 on proteomic data, 1.0E5 on metabolomic data) to ensure the feature quality and control the running time.

<Figure 2>

Then, we also compared DeepRTAlign with MZmine 2 and OpenMS on a public real-world metabolomic dataset SM1100[24]. This dataset was generated from standard mixtures consisting of 1100 compounds with specified concentration ratios. It contains two groups (SA, and SB) and each group has 5 replicates. We used DeepRTAlign (combined with 3 different feature extraction tools: MZmine 2, OpenMS and Dinosaur), MZmine 2 and OpenMS to align features across runs. Thus, it resulted five different algorithm combinations (**Table S10**). Here, to demonstrate the capacity of DeepRTAlign to deal with metabolomic data, every adjacent sample pair in each group (i.e., SA1-SA2, SA2-SA3, SA3-SA4, SA4-SA5 and SB1-SB2, SB2-SB3, SB3-SB4, and SB4-SB5) was aligned using the five algorithm combinations.

For feature extraction, the default parameters in OpenMS and Dinosaur were used. In MZmine 2, the parameter “Noise Level” was set to 1.0E5 to make the extracted feature number similar to those of OpenMS and Dinosaur. Then, the extracted features in each sample were annotated according to the 1100 standard compounds with strict standards (mass tolerance: ± 5 ppm, RT tolerance: ± 0.5 min as suggested in the paper of Li et al.[24]). Based on the annotation results, the precision and recall values for each combination were calculated and shown in **Table S10**. We can see that all the five combinations performed well due to the simpler composition of this metabolic standard dataset (compared with the proteomic datasets).

Comparison with Quandenser

DeepRTAlign was then compared with another popular tool Quandenser[32], which uses both MS and MS/MS information. Quandenser applies unsupervised clustering on both MS1 and MS2 levels to summarize all analytes of interest without assigning identities. Using Dinosaur as the feature extraction method, we compared DeepRTAlign with Quandenser on the UPS2-M and UPS2-Y datasets. The UPS2 peptides in UPS2-M and UPS2-Y datasets are divided into four groups (A-D) with decreasing loading amounts (1 μg, 0.2 μg, 0.04 μg and 0.008 μg) (**Figure S3**). We mapped all the extracted features to the identification results, and only considered the UPS2 peptides without missing values in all the replicates. On the two datasets, we calculated the intensity CVs of these UPS2 peptides among the three replicates. As shown in **Figure 3**, although Quandenser can align more UPS2 features, the CV values of DeepRTAlign in all the group are 47.6% smaller in UPS2-M and 58.3% smaller in UPS2-Y datasets than Quandenser. Similar results can be found on all the extracted features (no matter if there were corresponding identification results) (**Figure S4**).

<Figure 3>

Comparison with MaxQuant, FragPipe and DIA-NN with or without MBR

We compared DeepRTAlign with the alignment methods based on the indentification results, which is currently the most used alignment strategy. MBR is an updated version of these kinds of alignment methods, which can transfer identification results to the un-identification features[37]. We further compared DeepRTAlign with MaxQuant's MBR and FragPipe's MBR on the UPS2-M, UPS2-Y and EC-H datasets (**Figure 4**). MaxQuant or FragPipe was run twice with and without the MBR function keeping the other parameters unchanged.

As shown in **Figure 4**, DeepRTAlign, MaxQuant's MBR and FragPipe's MBR can increase the number of identified peptides without missing values in all the replicates compared to the original identification results without RT alignment. And DeepRTAlign increased the most (16% on UPS2-M, 25% on UPS2-Y and 70% on EC-H for DeepRTAlign (M) in **Figure 4**; 16% on UPS2-M, 26% on UPS2-Y and 65% on EC-H for DeepRTAlign (MF) in **Figure 4**), compared with MaxQuant's MBR (16% on UPS2-M, 21% on UPS2-Y and 42% on EC-H for MBR (M) in **Figure 4**) and FragPipe's MBR (6% on UPS2-M, 3% on UPS2-Y and 46% on EC-H for MBR (F) in

Figure 4), respectively. Furthermore, we find that DeepRTAlign (MM) showed a largest increase in identified peptide number (31% on UPS2-M, 45% on UPS2-Y and 76% on EC-H). Please note that since FragPipe did not provide its extracted feature list, we used the features extracted by MaxQuant for a fair comparison (DeepRTAlign (MF)). Although DeepRTAlign had the largest number of aligned peptides, the peptide intensity CVs did not increase significantly compared with peptides without using DeepRTAlign, indicating that the alignment of DeepRTAlign is accurate. These results demonstrated DeepRTAlign has a better performance for alignment than MBR-applied MaxQuant and FragPipe.

Furthermore, DeepRTAlign also showed a better performance than MBR in DIA-NN on single-cell proteomics DIA data (**Figure S5**). The average number of features at least present in two cells is approximately 35.6 times the average number of peptides, providing the possibility to identify different cell types using the aligned MS features in the future.

<Figure 4>

Generalizability evaluation on simulated datasets

Based on the 14 real-world datasets (9 proteomic datasets: EC-H, HCC-N, UPS2-M, UPS2-Y, HCC-R, AT, MI, SC and CD; 5 metabolomic datasets: NCC19, SM1100, MM, SO and GUS), we generated multiple simulated datasets with different RT shifts for each real-world dataset to evaluate the generalizability boundary of DeepRTAlign.

As shown in **Figure 5** and **Figure 6**, it can be found that in most cases, DeepRTAlign owns higher precision and recall values than OpenMS. And this advantage is more obvious when the feature number increases. For example, when the feature number of a dataset is small (such as 3688 for SM1100), DeepRTAlign and OpenMS show a similar performance (**Figure 6** and **Supplementary Notes**). While, when the feature number is large enough (such as 18064 for NCC19), we can find that DeepRTAlign has significant advantages in precision and recall. It may be due to the coarse alignment step in DeepRTAlign, which requires a certain number of features. Meanwhile, we also find the performances of DeepRTAlign and OpenMS become worse when the standard deviation of the RT shift increases (from 0.1 to 5). Thus, we recommend that for metabolomic and proteomic studies, the distribution of RT shift should be controlled to different levels in practical applications. In most proteomic datasets, when the standard deviation of RT shift is larger than 1 min, the precision and recall drop significantly (**Figure 5**). In most metabolic datasets, a similar phenomenon occurs when the standard deviation is larger than 0.5 min, especially in recall (**Figure 6**). DeepRTAlign has a higher precision and recall than OpenMS on most metabolomic datasets, despite of different chromatography setups. Both proteomic and metabolomic datasets are unaffected when changing the mean of RT shift.

<Figure 5>

<Figure 6>

Performance evaluation using different feature extraction methods

DeepRTAlign is directly compatible with four feature extraction methods (Dinosaur, MaxQuant, OpenMS and XICFinder). Here, we systematically evaluated the performance of DeepRTAlign using these four feature extraction methods on two UPS2 datasets (UPS2-Y and UPS2-M). The Mascot identification results were used as the ground truth. Compared with the methods not using DeepRTAlign, the number of confidently quantified UPS2 peptides have an up to 25% improvement when using DeepRTAlign (**Figure S6a-b**). Moreover, there is no significant difference in the peptide CVs of three technical replicates for the methods with or without DeepRTAlign, indicating that DeepRTAlign did not affect the quantification precision no matter which feature extraction method is used (**Figure S6c-d**).

Conclusions

In summary, we present a deep learning-based tool DeepRTAlign for RT alignment in large cohort proteomic and metabolomic data analysis. DeepRTAlign is based on the basic information of MS spectra (m/z, RT and intensity), which can be applied to all the precursor ions in MS data before identification. We have demonstrated that DeepRTAlign outperformed other existing alignment tools by aligning more corresponding features without compromising the quantification precision and determined its generalizability boundary on multiple proteomic and metabolomic datasets. DeepRTAlign is flexible and robust with different feature extraction tools. Finally, we applied DeepRTAlign to HCC early recurrence prediction (**Supplementary Notes**) as a real-world example and the results showed that aligned MS features have effective information compared with peptides and proteins. DeepRTAlign is expected to be useful in finding low abundant biomarkers which usually only have low-quality MS/MS spectra and play a key role in proteomics-driven precision medicine.

Acknowledgment

We thank Prof. Jing Yang from National Center for Protein Sciences (Beijing) for the helpful suggestions. This work has been supported by the National Key Research and Development Program of China (2021YFA1301603 and 2020YFE0202200), the National Natural Science Foundation of China (32088101), and the CAMS Innovation Fund for Medical Sciences (CIFMS) (2019-I2M-5-063).

Conflict of Interest: none declared.

References

(1) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. *Nat Methods* **2021**, *18* (11), 1363–1369. https://doi.org/10.1038/s41592-021-01301-5.

(2) Smith, R.; Ventura, D.; Prince, J. T. LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review. *Briefings in Bioinformatics* **2013**, *16* (1), 104–117. https://doi.org/10.1093/bib/bbt080.

(3) Fernández-Costa, C.; Martínez-Bartolomé, S.; McClatchy, D. B.; Saviola, A. J.; Yu, N.-K.; Yates, J. R. Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *J. Proteome Res.* **2020**, *19* (8), 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153.

(4) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat Protoc* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.

(5) Chang, C.; Li, M.; Guo, C.; Ding, Y.; Xu, K.; Han, M.; He, F.; Zhu, Y. PANDA: A Comprehensive and Flexible Tool for Quantitative Proteomics Data Analysis. *Bioinformatics* **2019**, *35* (5), 898–900. https://doi.org/10.1093/bioinformatics/bty727.

(6) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(7) Yu, F.; Haynes, S. E.; Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics* **2021**, *20*, 100077. https://doi.org/10.1016/j.mcpro.2021.100077.

(8) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput. *Nat Methods* **2020**, *17* (1), 41–44. https://doi.org/10.1038/s41592-019-0638-x.

(9) Mitra, V.; Smilde, A. K.; Bischoff, R.; Horvatovich, P. Tutorial: Correction of Shifts in Single-Stage LC-MS(/MS) Data. *Analytica Chimica Acta* **2018**, *999*, 37–53. https://doi.org/10.1016/j.aca.2017.09.039.

(10) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779–787. https://doi.org/10.1021/ac051437y.

(11) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics* **2010**, *11* (1), 395. https://doi.org/10.1186/1471-2105-11-395.

(12) Lange, E.; Gröpl, C.; Schulz-Trieglaff, O.; Leinenbach, A.; Huber, C.; Reinert, K. A Geometric Approach for the Alignment of Liquid Chromatography—Mass Spectrometry Data. *Bioinformatics* **2007**, *23* (13), i273–i281. https://doi.org/10.1093/bioinformatics/btm209.

(13) Duran, A. L.; Yang, J.; Wang, L.; Sumner, L. W. Metabolomics Spectral Formatting, Alignment and Conversion Tools (MSFACTs). *Bioinformatics* **2003**, *19* (17), 2283–2293. https://doi.org/10.1093/bioinformatics/btg315.

(14) Ballardini, R.; Benevento, M.; Arrigoni, G.; Pattini, L.; Roda, A. MassUntangler: A Novel Alignment Tool for Label-Free Liquid Chromatography–Mass Spectrometry Proteomic Data. *Journal of Chromatography A* **2011**, *1218* (49), 8859–8868. https://doi.org/10.1016/j.chroma.2011.06.062.

(15) Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E. High-Speed Peak Matching Algorithm for Retention Time Alignment of Gas Chromatographic Data for Chemometric Analysis. *Journal of Chromatography A* **2003**, *996* (1–2), 141–155. https://doi.org/10.1016/S0021-9673(03)00616-2.

(16) Li, M.; Wang, X. R. Peak Alignment of Gas Chromatography–Mass Spectrometry Data with Deep Learning. *Journal of Chromatography A* **2019**, *1604*, 460–476. https://doi.org/10.1016/j.chroma.2019.460476.

(17) Jiang, Y.; Sun, A.; Zhao, Y.; Ying, W.; Sun, H.; Yang, X.; Xing, B.; Sun, W.; Ren, L.; Hu, B.; Li, C.; Zhang, L.; Qin, G.; Zhang, M.; Chen, N.; Zhang, M.; Huang, Y.; Zhou, J.; Zhao, Y.; Liu, M.; Zhu, X.; Qiu, Y.; Sun, Y.; Huang, C.; Yan, M.;

Wang, M.; Liu, W.; Tian, F.; Xu, H.; Zhou, J.; Wu, Z.; Shi, T.; Zhu, W.; Qin, J.; Xie, L.; Fan, J.; Qian, X.; He, F.; Chinese Human Proteome Project (CNHPP) Consortium. Proteomics Identifies New Therapeutic Targets of Early-Stage Hepatocellular Carcinoma. *Nature* **2019**, *567* (7747), 257–261. https://doi.org/10.1038/s41586-019-0987-8.

(18) Bhat, M.; Clotet-Freixas, S.; Baciu, C.; Pasini, E.; Hammad, A.; Ivanics, T.; Reid, S.; Azhie, A.; Angeli, M.; Ghanekar, A.; Fischer, S.; Sapisochin, G.; Konvalinka, A. Combined Proteomic/Transcriptomic Signature of Recurrence Post-Liver Transplantation for Hepatocellular Carcinoma beyond Milan. *Clin Proteom* **2021**, *18* (1), 27. https://doi.org/10.1186/s12014-021-09333-x.

(19) Chang, C.; Zhang, J.; Xu, C.; Zhao, Y.; Ma, J.; Chen, T.; He, F.; Xie, H.; Zhu, Y. Quantitative and In-Depth Survey of the Isotopic Abundance Distribution Errors in Shotgun Proteomics. *Anal Chem* **2016**, *88* (13), 6844–6851. https://doi.org/10.1021/acs.analchem.6b01409.

(20) Shen, X.; Shen, S.; Li, J.; Hu, Q.; Nie, L.; Tu, C.; Wang, X.; Poulsen, D. J.; Orsburn, B. C.; Wang, J.; Qu, J. IonStar Enables High-Precision, Low-Missing-Data Proteomics Quantification in Large Biological Cohorts. *Proc Natl Acad Sci USA* **2018**, *115* (21), E4767–E4776. https://doi.org/10.1073/pnas.1800541115.

(21) Ginsawaeng, O.; Gorka, M.; Erban, A.; Heise, C.; Brueckner, F.; Hoefgen, R.; Kopka, J.; Skirycz, A.; Hincha, D. K.; Zuther, E. Characterization of the Heat-Stable Proteome during Seed Germination in Arabidopsis with Special Focus on LEA Proteins. *IJMS* **2021**, *22* (15), 8172. https://doi.org/10.3390/ijms22158172.

(22) Li, Y.; Li, H.; Xie, Y.; Chen, S.; Qin, R.; Dong, H.; Yu, Y.; Wang, J.; Qian, X.; Qin, W. An Integrated Strategy for Mass Spectrometry-Based Multiomics Analysis of Single Cells. *Anal. Chem.* **2021**, *93* (42), 14059–14067. https://doi.org/10.1021/acs.analchem.0c05209.

(23) Barberis, E.; Timo, S.; Amede, E.; Vanella, V. V.; Puricelli, C.; Cappellano, G.; Raineri, D.; Cittone, M. G.; Rizzi, E.; Pedrinelli, A. R.; Vassia, V.; Casciaro, F. G.; Priora, S.; Nerici, I.; Galbiati, A.; Hayden, E.; Falasca, M.; Vaschetto, R.; Sainaghi, P. P.; Dianzani, U.; Rolla, R.; Chiocchetti, A.; Baldanzi, G.; Marengo, E.; Manfredi, M. Large-Scale Plasma Analysis Revealed New Mechanisms and Molecules Associated with the Host Response to SARS-CoV-2. *IJMS* **2020**, *21* (22), 8623. https://doi.org/10.3390/ijms21228623.

(24) Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. Comprehensive Evaluation of Untargeted Metabolomics Data Processing Software in Feature Detection, Quantification and Discriminating Marker Selection. *Analytica Chimica Acta* **2018**, *1029*, 50–57. https://doi.org/10.1016/j.aca.2018.05.001.

(25) Wase, N.; Gutiérrez, J. M.; Rucavado, A.; Fox, J. W. Longitudinal Metabolomics and Lipidomics Analyses Reveal Alterations Associated with Envenoming by Bothrops Asper and Daboia Russelii in an Experimental Murine Model. *Toxins* **2022**, *14* (10), 657. https://doi.org/10.3390/toxins14100657.

(26) Swenson, T. L.; Karaoz, U.; Swenson, J. M.; Bowen, B. P.; Northen, T. R. Linking Soil Biology and Chemistry in Biological Soil Crust Using Isolate Exometabolomics. *Nat Commun* **2018**, *9* (1), 19. https://doi.org/10.1038/s41467-

017-02356-9.

(27) Gibson, C. L.; Codreanu, S. G.; Schrimpe-Rutledge, A. C.; Retzlaff, C. L.; Wright, J.; Mortlock, D. P.; Sherrod, S. D.; McLean, J. A.; Blakely, R. D. Global Untargeted Serum Metabolomic Analyses Nominate Metabolic Pathways Responsive to Loss of Expression of the Orphan Metallo β-Lactamase, MBLAC1. *Mol. Omics* **2018**, *14* (3), 142–155. https://doi.org/10.1039/C7MO00022G.

(28) Lichtman, J. S.; Alsentzer, E.; Jaffe, M.; Sprockett, D.; Masutani, E.; Ikwa, E.; Fragiadakis, G. K.; Clifford, D.; Huang, B. E.; Sonnenburg, J. L.; Huang, K. C.; Elias, J. E. The Effect of Microbial Colonization on the Host Proteome Varies by Gastrointestinal Location. *ISME J* **2016**, *10* (5), 1170–1181. https://doi.org/10.1038/ismej.2015.187.

(29) Mottawea, W.; Chiang, C.-K.; Mühlbauer, M.; Starr, A. E.; Butcher, J.; Abujamel, T.; Deeke, S. A.; Brandel, A.; Zhou, H.; Shokralla, S.; Hajibabaei, M.; Singleton, R.; Benchimol, E. I.; Jobin, C.; Mack, D. R.; Figeys, D.; Stintzi, A. Altered Intestinal Microbiota–Host Mitochondria Crosstalk in New Onset Crohn's Disease. *Nat Commun* **2016**, *7* (1), 13419. https://doi.org/10.1038/ncomms13419.

(30) Teleman, J.; Chawade, A.; Sandin, M.; Levander, F.; Malmström, J. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *J. Proteome Res.* **2016**, *15* (7), 2143–2151. https://doi.org/10.1021/acs.jproteome.6b00016.

(31) Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C. Critical Assessment of Alignment Procedures for LC-MS Proteomics and Metabolomics Measurements. *BMC Bioinformatics* **2008**, *9* (1), 375. https://doi.org/10.1186/1471-2105-9-375.

(32) The, M.; Lukas, K. Focus on the Spectra That Matter by Clustering of Quantification Data in Shotgun Proteomics. *Nature Communications* **2020**, *11*, 3234.

(33) Li, N.; Wu, S.; Zhang, C.; Chang, C.; Zhang, J.; Ma, J.; Li, L.; Qian, X.; Xu, P.; Zhu, Y.; He, F. PepDistiller: A Quality Control Tool to Improve the Sensitivity and Accuracy of Peptide Identifications in Shotgun Proteomics. *Proteomics* **2012**, *12* (11), 1720–1725. https://doi.org/10.1002/pmic.201100167.

(34) Ma, J.; Chen, T.; Wu, S.; Yang, C.; Bai, M.; Shu, K.; Li, K.; Zhang, G.; Jin, Z.; He, F.; Hermjakob, H.; Zhu, Y. IProX: An Integrated Proteome Resource. *Nucleic Acids Res.* **2019**, *47* (D1), D1211–D1217. https://doi.org/10.1093/nar/gky869.

(35) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yılmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Research* **2019**, *47* (D1), D442–D450. https://doi.org/10.1093/nar/gky1106.

(36) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. MetaboLights: A Resource Evolving in Response to the Needs of Its Scientific Community. *Nucleic Acids Research* **2019**, gkz1019.

https://doi.org/10.1093/nar/gkz1019.

(37) Lim, M. Y.; Paulo, J. A.; Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J. Proteome Res.* **2019**, *18* (11), 4020–4026. https://doi.org/10.1021/acs.jproteome.9b00492.

Figure Legends

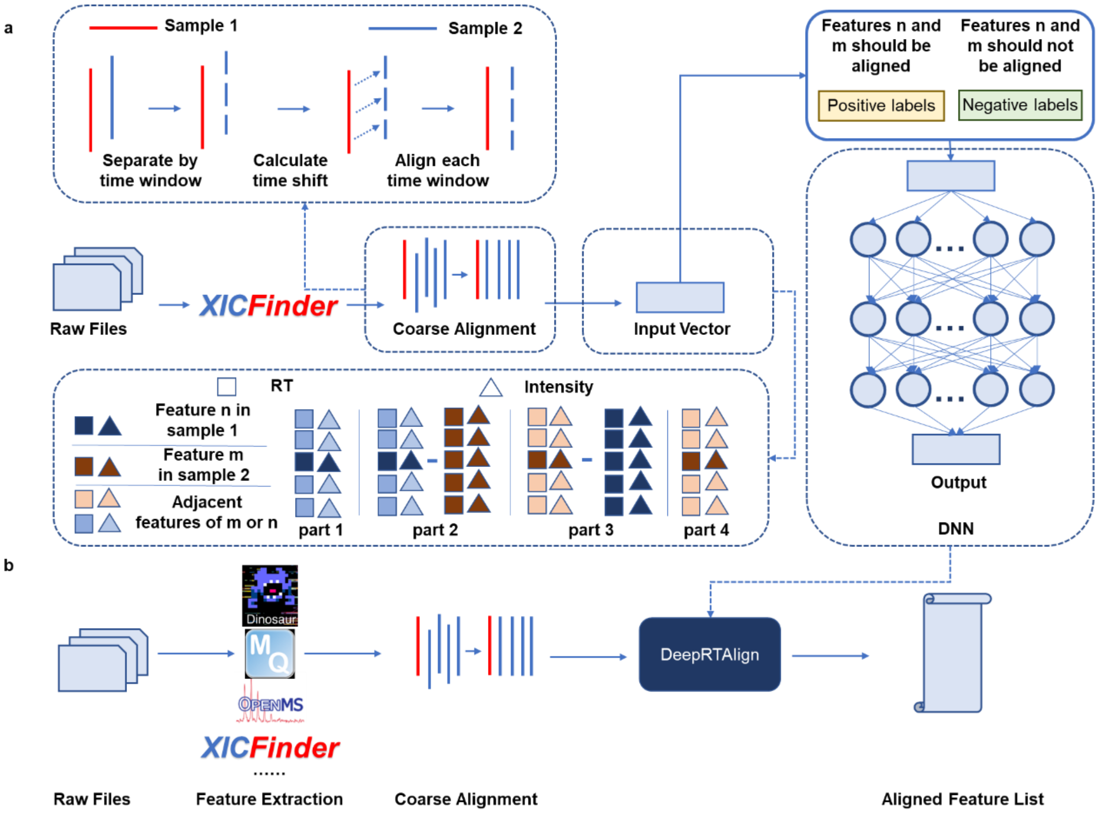

Figure 1. The methods of DeepRTAlign. (a) The training procedures of DeepRTAlign. (b) The workflow for RT alignment using DeepRTAlign. DeepRTAlign supports four feature extraction methods (Dinosaur, MaxQuant, OpenMS and XICFinder), the features extracted will be aligned using the trained model shown in (a). Then, the aligned feature list will be output.

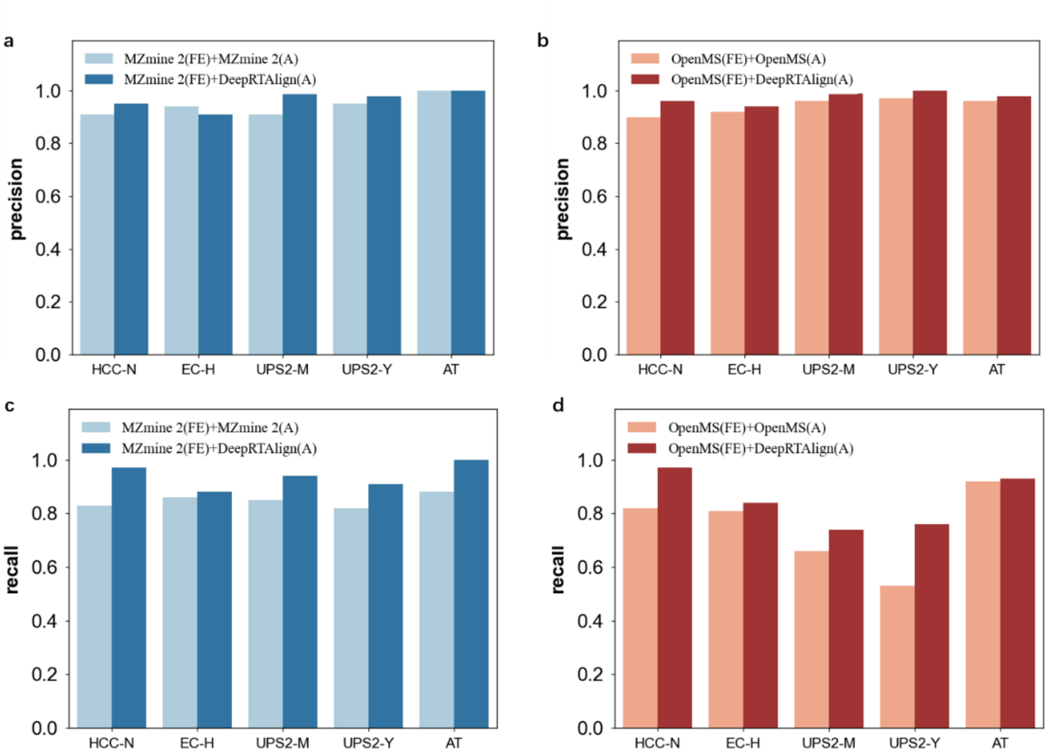

Figure 2. Performance evaluation of DeepRTAlign compared with MZmine 2 and OpenMS. The precision (a) and recall (c) of MZmine 2 and DeepRTAlign on different test sets. The precision (b) and recall (d) of OpenMS and DeepRTAlign on different datasets. “FE” means the feature extraction method. “A” means the RT alignment method. We took the Mascot identification results with FDR<1% as the ground truth in datasets HCC-N, UPS2-M and UPS2-Y and took the MaxQuant identification results with FDR<1% as the ground truth in datasets EC-H and AT. In dataset EC-H we only considered the *E. coli* peptides for evaluation. In datasets UPS2-M and UPS2-Y, we only considered the UPS2 peptides for evaluation.

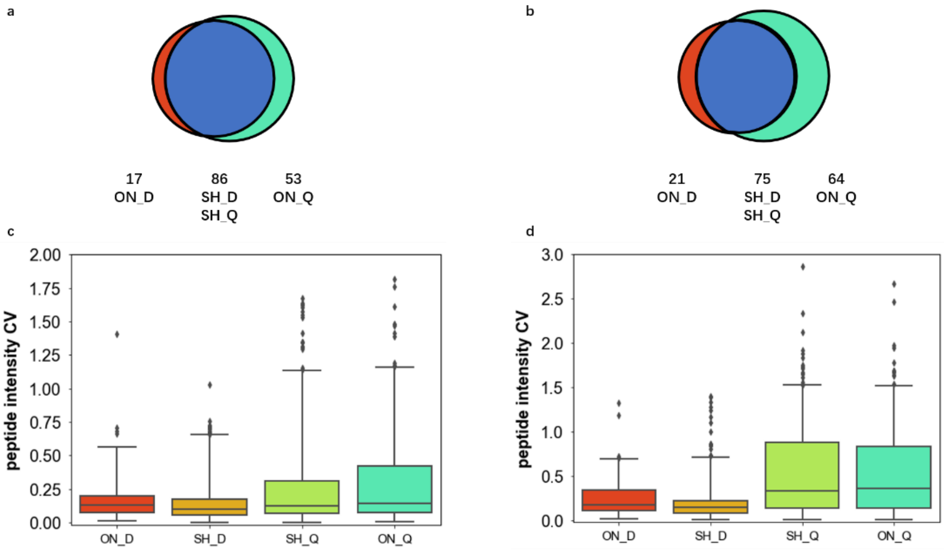

Figure 3. Performance evaluation of DeepRTAlign compared with Quandenser. (a-b) The Venn diagrams of quantified UPS2 peptides in all the three replicates in UPS2-M (a) and UPS2-Y (b) datasets, respectively. (c-d) The boxplots of peptide intensity CVs of the UPS2 peptides in the three replicates in UPS2-M (c) and UPS2-Y (d). ON_D: DeepRTAlign Only. SH_D: DeepRTAlign Shared. ON_Q: Quandenser Only. SH_Q: Quandenser Shared.

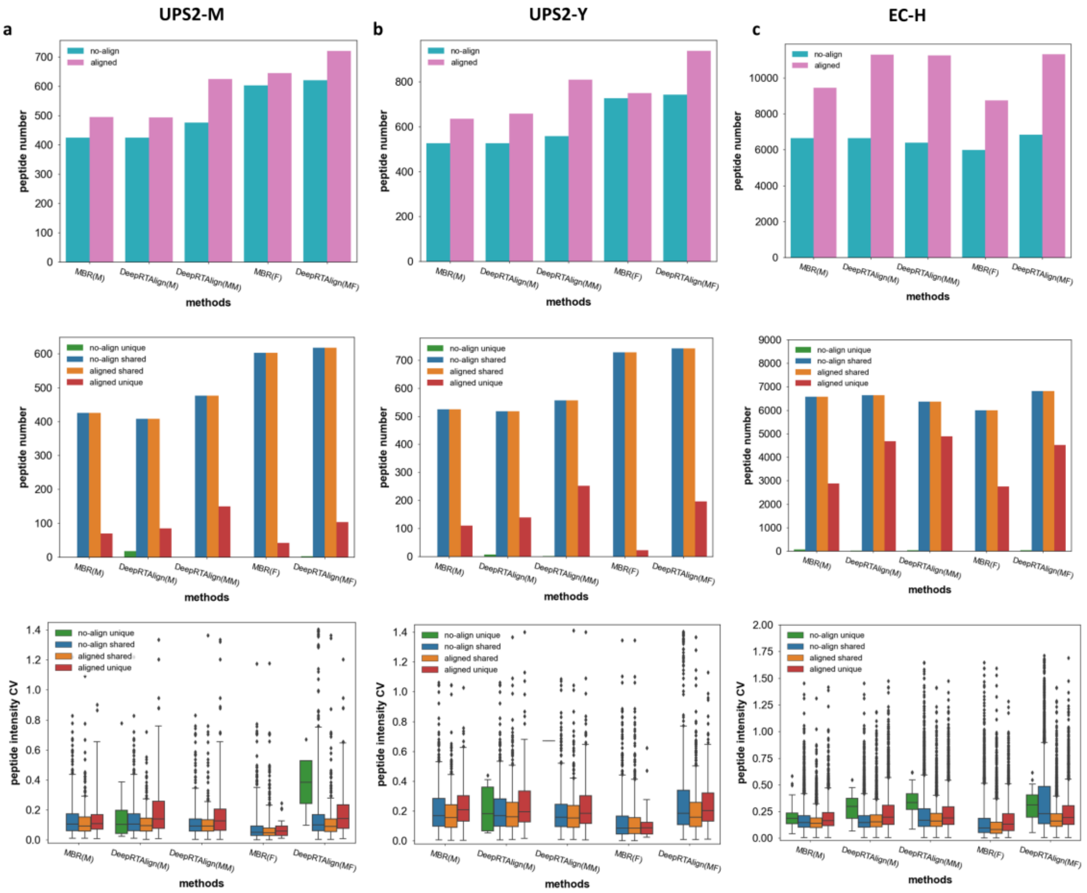

Figure 4. Performance evaluation of DeepRTAlign compared with MBR-applied MaxQuant and FragPipe on UPS2-M dataset (a), UPS2-Y dataset (b) and EC-H dataset (c). The first row is the histograms of identified UPS2 peptides (UPS2-M and UPS2-Y datasets) and *E. coli* peptides (EC-H dataset). "no-align" means no alignment methods (MBR or DeepRTAlign) were used. "aligned" means one alignment method (MBR or DeepRTAlign) was used. The second row is the histograms of identified UPS2 peptides (UPS2-M and UPS2-Y datasets) and *E. coli* peptides (EC-H dataset). "no-align unique" means the peptides uniquely identified without alignment. "no-align shared" and "aligned shared" mean the peptides commonly identified without and with alignment. "aligned unique" means the peptides uniquely identified with alignment. The third row is the boxplots of the intensity CVs of UPS2 peptides (UPS2-M and UPS2-Y datasets) or *E. coli* peptides (EC-H dataset). Only the UPS2 peptides and *E. coli* peptides without missing value among replicates are calculated.

Please note these abbreviations in the figure:

- MBR (M): Features extracted by MaxQuant, aligned by MaxQuant's MBR function and identified by MaxQuant.
- DeepRTAlign (M): Features extracted by MaxQuant, aligned by DeepRTAlign and identified by MaxQuant.
- DeepRTAlign (MM): Features extracted by MaxQuant, aligned by DeepRTAlign and identified by Mascot.
- MBR (F): Features extracted by FragPipe, aligned by FragPipe's MBR function and identified by FragPipe.
- DeepRTAlign (MF): Features extracted by MaxQuant, aligned by DeepRTAlign and

identified by FragPipe.

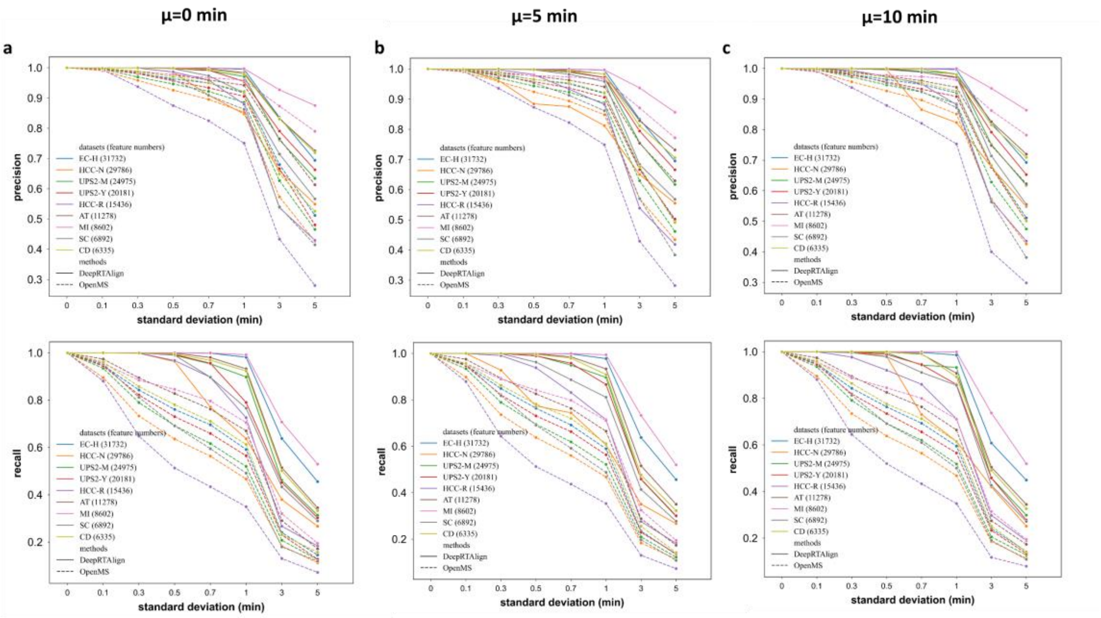

Figure 5. Comparison of DeepRTAlign and OpenMS on multiple simulated datasets generated from 9 real-world proteomic datasets. The simulated datasets were constructed by adding normally distributed RT shifts to the corresponding real-world dataset. (a) μ=0 min. (b) μ=5 min. (c) μ=10 min. The normal distribution has an increasing σ, i.e., σ=0, 0.1, 0.3, 0.5, 0.7, 1, 3, 5 for different μ (0, 5 and 10 minutes), respectively.

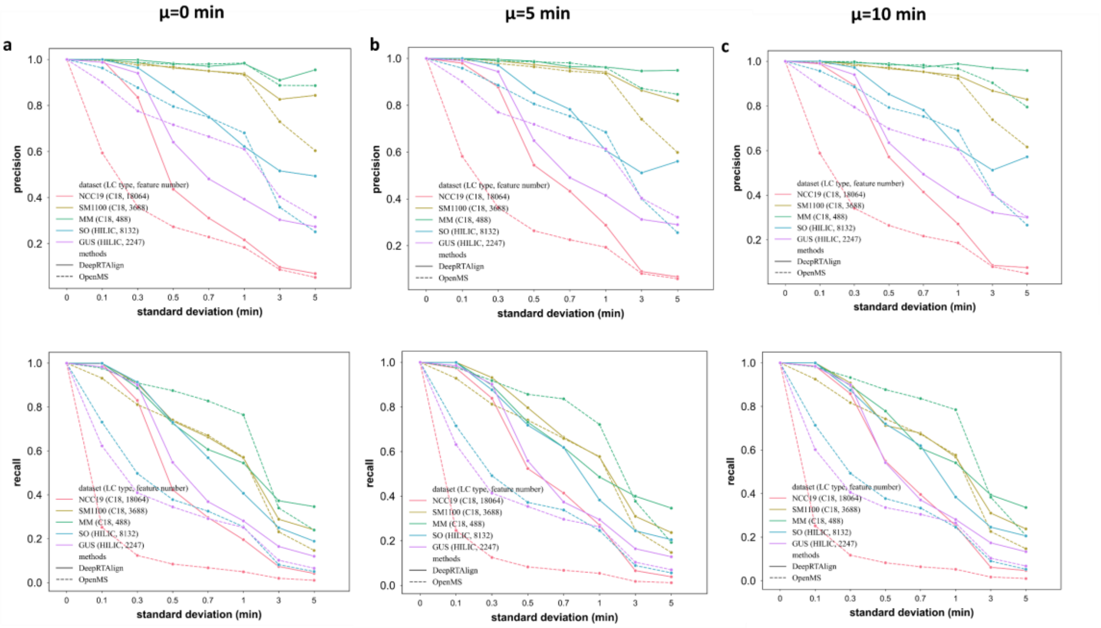

Figure 6. Comparison of DeepRTAlign and OpenMS on multiple simulated datasets generated from 5 real-world metabolomic datasets. The simulated datasets were constructed by adding normally distributed RT shifts to the corresponding real-world dataset. (a) μ=0 min. (b) μ=5 min. (c) μ=10 min. The normal distribution has an increasing σ, i.e., σ=0, 0.1, 0.3, 0.5, 0.7, 1, 3, 5 for different μ (0, 5 and 10 minutes), respectively.