Journal of Proteomics 223 (2020) 103820

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

PB-Net: Automatic peak integration by sequential deep learning for multiple reaction monitoring

Zhenqin Wu[a,c], Daniel Serie[a], Gege Xu[a], James Zou[a,b,*]

[a] InterVenn Biosciences, United States of America
[b] Department of Biomedical Data Science, Stanford University, United States of America
[c] Department of Chemistry, Stanford University, United States of America

ARTICLE INFO

Keywords:
Mass spectrometry
Glycoproteomics
Machine learning
Deep learning

ABSTRACT

Mass spectrometry (MS) based proteomics has become an indispensable component of modern molecular and cellular biochemistry analysis. Multiple reaction monitoring (MRM) is one of the most well-established MS techniques for molecule detection and quantification. Despite its wide usage, there lacks an accurate computational framework to analyze MRM data, and expert annotation is often required, especially to perform peak integration. Here we propose a deep learning method PB-Net (Peak Boundary Neural Network), built upon recent advances in sequential neural networks, for fully automatic chromatographic peak integration. To train PB-Net, we generated a large dataset of over 170,000 expert annotated peaks from MS transitions spanning a wide dynamic range, including both peptides and intact glycopeptides. Our model demonstrated outstanding performances on unseen test samples, reaching near-perfect agreement (Pearson's r 0.997) with human annotated ground truth. Systematic evaluations also show that PB-Net is substantially more robust and accurate compared to previous state-of-the-art peak integration software. PB-Net can benefit the wide community of mass spectrometry data analysis, especially in applications involving high-throughput MS experiments. Codes and test data used in this work are available at https://github.com/miaecle/PB-net.

Significance: Human annotations serve an important role in accurate quantification of multiple reaction monitoring (MRM) experiments, though they are costly to collect and limit analysis throughput. In this work we proposed and developed a novel technique for the peak-integration step in MRM, based on recent innovations in sequential deep learning models. We collected in total 170,000 expert-annotated MRM peaks and trained a set of accurate and robust neural networks for the task. Results demonstrated a substantial improvement over the current state-of-the-art software for mass spectrometry analysis and comparable level of accuracy and precision as human annotators.

1. Introduction

Multiple reaction monitoring (MRM) is a technique utilized in tandem mass spectrometry (MS) which allows for highly sensitive and specific detection of proteins, lipids, and post-translational modifications (PTMs), among other analytes [3]. MRM's application of sequential mass-to-charge ratio (m/z) filters is exemplified by triple-quadrupole (QqQ) instruments, which select for a precursor ion m/z in Q1, fragment in q2, and further select for a specific product ion m/z in Q3, prior to detection. These steps yield increased dynamic range and sensitivity in targeted quantification, in comparison to techniques such as data independent acquisition (DIA), where all resulting fragments are analyzed and biomarker discovery is often the goal. MRM technology holds great promise for use in new clinical assays and diagnostics, but the lack of precise analysis software remains a major bottleneck.

Traditional analysis of MRM experiments involved choosing the start and stop of a chromatographic peak by hand, working from transitions previously characterized by data-dependent acquisition or other discovery techniques. A typical implementation can be seen in Agilent's MassHunter software, where intensities are plotted over a prespecified range of retention times (RT) for a given precursor and product mass-to-charge ratio (referred to as XIC graphs [23] in the following texts). Selection of the beginning and end of the peak yields an observed RT, peak width, and integrated abundance value. Person to person variability in assessment, human error, and a large time investment render this method inadequate for high-throughput usage.

In recent years several software packages [20] have attempted to fill

* Corresponding author at: R258, 350 Jane Stanford Way, Stanford, CA, United States of America.
E-mail address: jamesz@stanford.edu (J. Zou).

https://doi.org/10.1016/j.jprot.2020.103820
Received 23 October 2019; Received in revised form 17 April 2020; Accepted 9 May 2020
Available online 13 May 2020
1874-3919/

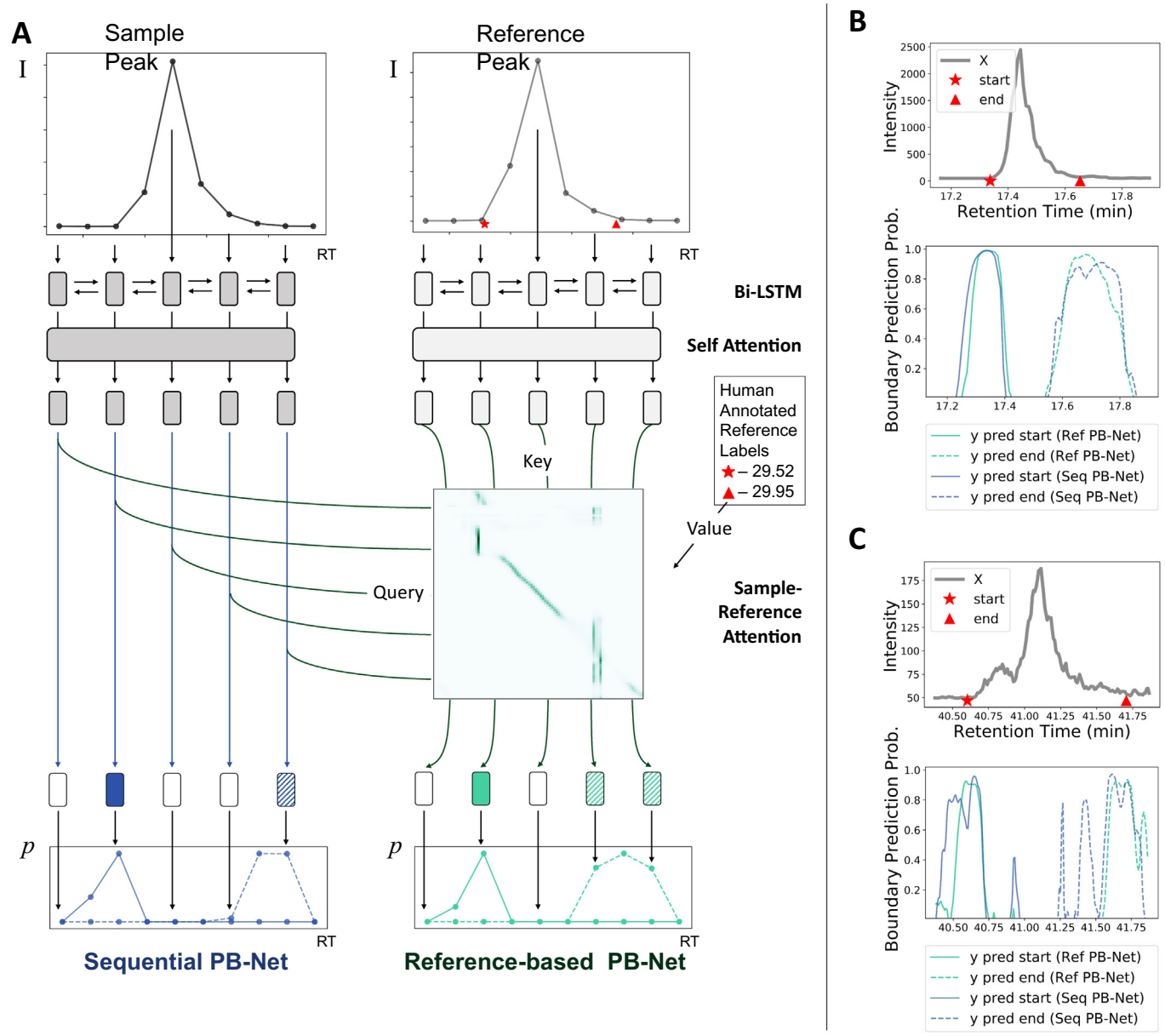

Fig. 1. Model structures and sample inputs/outputs of PB-Net A: Illustration of models for peak boundary prediction. Left half shows the sequential PB-Net composed of bi-LSTM and self attention layers on sample input, with output directly mapped to predictions (blue). Right half shows the encoding of reference peak with same network structures. Encodings of sample, reference and human annotated peak start/end (marked as red star and triangle) are cross-linked through the sample-reference attention layer, generating the predictions for reference-based PB-Net (green). B and C: Example inputs and predictions from the two models. Note that lower signal-noise ratio in input will cause more noisy boundary predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

this gap by automatically integrating peaks from a list of transitions, among them OpenMS [22], DIA-Umpire [27], and Skyline [17,23]. Skyline has the largest user base and is arguably the gold standard for targeted MS quantification of peptides, accepting input formats from a wide range of MS vendors. While it exhibits good performance on high abundance peaks (especially peptides from abundant proteins), it can be less accurate in integrating low-abundant and highly-heterogeneous species such as glycosylated peptides. Due to the highly variable peak shape, low signal-noise ratio and complex sample component, it will be in principle hard to have a comprehensive and robust deterministic algorithm for peak integration in these cases. Other external issues, including retention time shift, will further add to the complexity.

In parallel, significant advances in machine learning and deep learning [13] techniques have been witnessed in recent years, especially in the field of computer vision [5] and natural language processing [25,28]. This naturally raises opportunities in applications of chemistry/biology related problems, such as genome analysis [33], biomedical image analysis [15] and molecular property predictions [30]. In the field of mass spectrometry, data-assisted techniques have also been applied to numerous tasks. Zohora et al. [32] proposed using a convolutional neural network (CNN) to scan LC-MS maps for peptide feature detection. Tran et al. [26] used a combination of CNN and long short-term memory (LSTM) to predict peptide sequence purely from tandem MS data. Zhou et al. [31] and Ma et al. [16] applied deep neural networks on the reverse task of predicting experimental mass spectra patterns and retention time of peptides, respectively. Demichev et al. [4] applied neural networks in data-independent acquisition for signal filtering. These works focused on a wide range of different aspects and

procedures of mass spectrometry-based proteomics, but none is directly applicable to the peak quantification task as we are looking at. In this work, we develop new deep learning models to solve the long-standing task of chromatographic peak quantification for high-throughput MRM.

To train our deep learning model, we generated a large new dataset of over 170,000 expert annotated peaks from 210 human serum samples. This data covers MS transitions spanning a wide dynamic range, including both peptides and intact glycopeptides. The sequence of intensities within each transition's RT window served as our input features, and the human-annotated start and stop RT were our labels. Together they form a training dataset for a bi-directional recurrent neural network [18] (RNN) with LSTM [9] units. To the best of our knowledge, this is the first work to apply deep neural nets in such a large dataset to the problem of targeted MS quantification.

2. Materials and methods

2.1. Dataset preparation

Two sets of mass spectra experiments were utilized in this study. We used the first experiment for training and model tuning (validation), and the second set for test-time evaluation, accepting only the precursor and product mass-to-charge ratios and RT window as input. Details of experiments are elaborated in Supplementary Information. Set one consisted of human serum samples and raw data were provided from a collaborator's lab at UC Davis (Carlito L, unpublished data), while set two were commercially-available serum from ovarian cancer and benign mass patients purchased from Indivumed GmbH in mid-2018. Both sets of samples underwent the same experimental protocol(outlined in Supplementary Information). Post-run, Agilent.D format files were converted to mzML via msConvert within the Proteowizard 3.0 software suite [10] for bioinformatic processing.

In the first experiment, peaks labeled with zero/low abundances or high signal-noise ratios were excluded. Samples with large annotation differences were also excluded. In total, 106,355 peaks (210 serum samples, 716 transitions, 70.7% of all samples) were generated. This set was split into a training set with 572 transitions and a validation set with 144 transitions. The transitions employed were characterized in previous work [14,19] and had peptide precursors from 5 to 56 amino acids in length, spanning 65 distinct serum proteins. They included low abundance, glycosylated proteoforms, and the collected transitions had an overall dynamic range of four orders of magnitude.

The second experiment was conducted in-house on 135 serum samples collected from ovarian cancer and benign mass patients. In total, 503 transitions were evaluated and used exclusively for testing. A minimum amount of filtering was applied (similar procedure as training/validation dataset), yielding 67,672 XIC graphs (99.7% of all samples) for testing. Note that 313 test set transitions were not present in the train/validation set.

Raw data from mzML were processed to facilitate model training and calculation: on each transition, all signals within a retention time window from reference retention time start - 0.2 min to reference retention time stop $+0.2$ min and a mass-charge ratio window of $\pm$ 0.1 around the desired precursor and product m/z were collected. Signals were summed along the mass-charge ratio window first, producing extracted-ion chromatograms (XIC). Human annotators and all models except for Skyline are presented with the XIC curves around the target transition (see Fig. 1B and C, and Fig. 3B for some example XIC inputs). Skyline was configured to directly calculate on mzML and a transition list recording precursor m/z, product m/z and rough retention time window. Note that all filtering and processing steps were performed before any model training and testing.

Twelve human annotators were employed to label the peak boundaries based on a reference labeled by a mass spectrometrist, and a small set of transitions (in total 1619 peaks) were further labeled independently by all annotators to test for consistency. Complete details of the MS experiments and dataset preparation pipeline are provided in the Supplementary Information.

2.2. Sequential neural networks for peak quantification

We extended a bi-directional LSTM, a widely applied method in time-series applications [7] and natural language processing [24], to build our peak integration model PeakBoundaryNet (PB-Net). The framework of the training/prediction process is illustrated in Fig. 1A (left), in which inputs (XIC graphs) are encoded through two bi-LSTM layers and a self-attention layer [28] for relating separate positions. The prediction is performed point-wise, generating a sequence output $\hat{y} \in \mathbb{R}^{N\times 2}$ for each input curve. Note that the boundary prediction task is separated into two independent components: the predictions of peak start and peak end. In other words, PB-Net outputs two probabilities for each point on the intensity curve, whether it marks the start and the end of the peak, forming distributions of boundaries. The model is trained end-to-end using Adam optimizer [11] on the cross entropy loss between our prediction and the smoothed label. Minimum hyper-parameter search is applied to optimize performance on the leave-out validation set. Details of network structures and train/validation set performances are elaborated in Supplementary Information. Results from the test set are presented in Results sections.

2.3. Reference-based sequential model

Due to the highly variable shape and width of peaks, a unified prediction model may not generalize well to unseen transitions. Based on the observation that samples from different patients on the same transition are consistent in shape, we proposed another variant of PB-Net, which utilized reference peaks (peaks collected from pooled serum and annotated by a mass spectrometrist) to refine predictions.

The model is illustrated in Fig. 1A (right). With the same framework as the vanilla sequential model, the reference-based PB-Net incorporated two encoding networks that take a query sample and reference as inputs, respectively. The encoded features for both are compared and merged in the sample-reference attention layer, generating predictions. Provided with query peak features, reference peak features and reference labels, the attention layer will compare query and reference point by point such that points in the query which are highly similar to the start/end marks in the reference will obtain higher predicted probabilities. Given two consistent samples, the ideal attention map of this operation will be an “identity matrix,” in which points with the same surroundings are regarded as highly similar (also see Fig. S1). We added a divergence term between the attention map and a retention time mapping matrix to achieve this regularization.

Note that in the setting of reference-based prediction, we are no longer trying to identify the noise-signal transition point, but rather a good encoding mechanism of the point and its context in order to optimize the similarity mapping. In applications, reference-based PB-Net acts similarly to few-shot learning models [29]. It is trained on pairs of query/reference, aiming at accurately mapping reference labels to predictions on the query sample. This naturally allows the model to perform better on unseen transition peaks as long as a correctly labeled reference (one or a few samples) is provided.

2.4. Baseline models

The models introduced above were compared with two baseline predictors in this study. Skyline [17], an open-source software designed for data analysis of mass spectrometry applications including multiple reaction monitoring, was applied to calculate peak areas (abundances) of test samples. We also adapted a deterministic method from a previous work [2], which will be referred to as the “Rule-based” method in the following text. We tuned its parameters to maximize train and validation set performances and applied the same set of parameters during

test set evaluations. Implementation details and parameters of the method are elaborated in Supplementary Information.

2.5. Featurization for neural network inputs

Inputs (XIC curves) were formulated as sequences of points on the intensity-retention time plane, with no fixed start/stop or length. To generate a uniform representation, we applied featurization on each sample. Retention times for all points in a single XIC curve were first centered at the curve's apex position, then expanded by 128 equally spaced Gaussian bins, ranging from −1 min to 1 min. Intensities were similarly discretized by applying 256 Gaussian functions on the range from 0 to 500, assuming that any points with intensities larger than the top threshold should be regarded as part of the peak. The full input for each sample curve after the transformations described above was formed as $x \in \mathbb{R}^{N\times 384}$, in which length of the curve N was a variable ranging from 50 to 300. In reference-based PB-Net training/prediction, each sample was paired with its reference. The pair of inputs had the same featurization process as above and were formed as two matrices, length of which should be close but not necessarily equal.

2.6. Training settings

Both PB-Net and PB-Net with reference sample were implemented in pytorch [1]. The vanilla sequential variant contained two bi-LSTM layers and a multi-head attention layer, output of which was passed through a fully-connected layer and mapped to two tasks, each with two classes. Softmax was applied to generate probability curves of peak start/end throughout the sequence.

The reference-based PB-Net contained two sets of sequential network structure (till the self-attention layer) identical to above. Weights were not shared. A sample-reference attention layer was applied on the sample encodings (as queries), reference encodings (as keys) and reference labels (as values), and output was directly used as model predictions for peak start/end probabilities. Models were trained by cross entropy loss between smoothed labels and the softmax predictions on the train subset and lightly optimized on the valid set. Detailed structures and training hyper-parameters can be found in the Supplementary Information. All related codes for neural network construction and training, as well as test data for evaluation are available at https://github.com/miaecle/PB-net.

3. Results

3.1. Boundary prediction performances

We tested PB-Net on the leave-out test set for evaluation. Predictions were evaluated on two criteria, accuracy in predicting boundaries (peak start/end) and accuracy in predicting abundances (peak area). While the former is directly related to the setup of the problem, the latter will be of much higher importance to downstream applications, which utilize analyte abundance for analysis.

In predictions, points in the curve with maximum predicted probabilities for the two tasks (peak start/end) are used as markers for boundaries respectively, which will be referred to as “argmax boundaries” in the following text. We calculated mean-absolute-error (MAE) between argmax boundaries and the human annotations, along with an accuracy score defined as ratio of samples whose boundary predictions were within an error threshold of 1.2 s (2 points in a sequence) around the ground truth annotations. Given that the peak duration was 20 ± 8 seconds (33 ± 14 points in a sequence) in the test set, the error threshold was small enough so that any boundary prediction within the window did not significantly change the abundance.

Table 1 presents performance scores of the three tested methods. Note that Skyline generated abundance predictions only and will not be evaluated and compared on boundary tasks. Sequential PB-Net demonstrated significant performance boosts over the rule-based method and Skyline. Reference-based PB-Net, with the aid of extra reference data, achieved top scores. As rule-based method is tuned to optimize training/validation set performances(see Table S2), strong overfitting was observed. In comparison, thanks to the better sample quality, the two variants of PB-Net achieved MAE of 2.33 and 1.56 s on the test set respectively, two times smaller than the rule-based method. Likewise, in accuracy score performance the reference-based PB-Net took the lead with a 30% advantage over the rule-based method. Fig. 2A and B show a bar plot and unnormalized cumulative distribution of boundary prediction error, in which PB-Net presented distributions with lower error as well as less outliers.

Table 1
Boundary prediction performance on independent test set (67,672 peaks).

Model	MAE(second)	Boundary Accuracy[a]
Rule-based	4.79	0.285
Sequential PB-Net	2.33	0.432
Reference-based PB-Net	1.56	0.584

[a] Proportion of samples whose boundary predictions are within 1.2 s error.

3.2. Abundance prediction performances

For evaluation of abundances, we calculated baseline-adjusted integrals between ground truth and predicted boundaries respectively. In this step, other than directly applying the argmax boundaries, we further tested a weighted abundance calculation method that regards the prediction curve of peak boundaries as two probability distributions and derive the abundance as an average of multiple peak start/end pairs. This method mainly helped in stabilizing the abundance calculation and eliminate certain type of outliers. Performances were reported as numbers in parenthesis and its details were discussed in the Supplementary Information.

Spearman's rank correlation coefficient (Spearman's r) and Pearson correlation coefficient (Pearson's r) were evaluated between ground truth and predictions on each individual transition. Due to the wide range of abundances ($1 \sim 10^{15}$), Pearson's r will be dominated by samples with large abundances, so we reported correlations calculated on log-abundances instead to avoid bias. Average scores over all transitions are reported in Table 2. Furthermore, for indication of practical usage, we also reported an accuracy score defined as ratio of samples whose predicted abundances were within ± 5% of the annotated abundances. Note that this accuracy metric has certain caveat: samples with deviated boundary predictions may appear to have same abundances if the deviations are on the same direction.

Abundance predictions presented similar results: sequential PB-Net achieved better and more robust performances on evaluating peak abundances (Fig. 2C), with an over 20% increase on accuracy(± 5%). Note that in this problem the influence from errors in boundary prediction will be largely alleviated as boundary points typically contribute little to the overall integral. Correlation scores indicated that PB-Net performed significantly better than the current off-the-shelf mass spec data analysis software, with a near-perfect Spearman's correlation coefficient at 0.984 and 0.992. Intriguingly, the rule-based method also achieved a higher correlation score than the Skyline predictions. We speculate that this is due to the complex composition of test samples, which generated co-eluting peaks or compounds with close retention time. As we fix our detection window closely around the desired transition (which is also the setting for human annotators), Skyline might accidentally pick peaks from different range, resulting in significantly worse performance.

The alternative weighted abundance calculation method generated more robust estimates than the argmax boundaries, especially in reference-free cases, raising Pearson's r from 0.989 to 0.997. Increase in

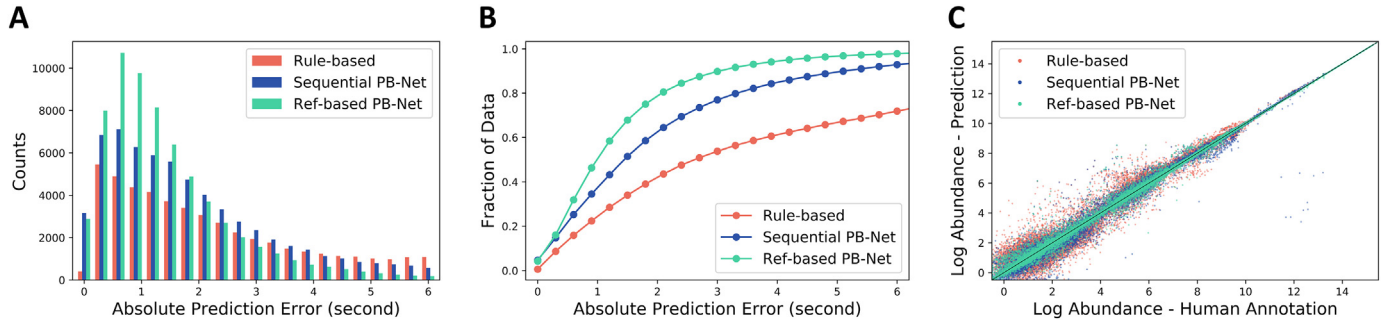

Fig. 2. Boundary and abundance prediction performances of PB-Net. A: Bar plot of absolute error in boundary predictions on test set. B: Cumulative distributions of boundary prediction errors. Note that 19,039 (28.1%) samples are out of range (> 6 seconds) in rule-based predictions, 4824 (7.1%) samples in Sequential PB-Net, 1470 (2.2%) samples in Reference-based PB-Net. C: Scatter plot of peak abundance prediction/annotation on the test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Abundance prediction performance on independent test set (67,672 peaks). Values in parentheses are calculated through the weighted average process discussed in the text.

Model	Log-Abd Pearson's r	Spearman's r	Abundance Accuracy[a]
Skyline	0.674	0.734	0.527
Rule-based	0.988	0.971	0.585
Sequential PB-Net	0.989 (0.997)	0.984 (0.990)	0.689 (0.652)
Reference-based PB-Net	0.997 (0.998)	0.992 (0.993)	0.800 (0.762)

[a] Proportion of samples whose abundance predictions are within ± 5% error.

the reference-based PB-Net was relatively minor. As shown in Fig. S2, this step mainly helped in avoiding outliers whose argmax boundaries caused significantly lower abundances, which typically suffer from abnormally high predicted boundary probabilities within the range of peak (at valley or shoulder points).

3.3. Improved accuracy and consistency by peak references

Between the two PB-Net variants, thanks to the inclusion of extra reference data, reference-based model demonstrated further improvement over the vanilla sequential model, offering 15% and 10% increases on the accuracy of boundary and abundance predictions. Individually, we observed the reference-based PB-Net presenting a better boundary error distribution with fewer outliers (2.2% versus 7.1%, Fig. 2B), consistent with the observation in abundance correlation plots: green points are more concentrated around the center line while blue points include a few outlying predictions (Fig. 2C).

Apart from improved performance across the mean metrics, the reference-based PB-Net provided another advantage over the standalone predictor: higher consistency in predictions on the same transitions across multiple samples. Especially for transitions whose boundaries could vary due to shoulder and valley points, the reference-based model tends to follow the same prediction form of its reference (Fig. 3A), which was the main objective of its model structure design. This is highly desirable in high throughput experiments, in that abundances are generated based on the very same standard and hence allow for more accurate comparison and quantification. It should also be noted that the strong prior assumption in the model may also hurt its performance in cases of significant retention time shift or strong batch effect, but incorporating data augmentation with regard to these issues could largely alleviate their influences (see Fig. S7 for details).

Lastly, it should be emphasized that the reference-based PB-Net would be most applicable and cost-efficient for high-throughput experiments on the same set of transitions in which manual annotations of a few high quality reference peaks are reasonable prerequisites. On most discovery or small scale experiments, the standalone reference-free PB-Net would be more suitable.

3.4. Comparison with human annotators

To further validate the practical usage of PB-Net, we compared model predictions against a group of human annotators calculating peak abundances. Within the test set, we analyzed a subset consisting of 12 transitions for all 135 serum samples, and had all 12 annotators mark the peak start/end independently. Combined with the original test set labels, 13 sets of human annotations were prepared for this subset (note that two sets were annotated by the same individual but at different time, which are still regarded as independent sets). Then we calculated relative standard deviations for each sample peak across the 13 annotations as a proxy for variations between different annotators, and compared this with the model predictions' relative errors. Results are summarized in Fig. 4.

High consistency over annotators was observed over most samples, as demonstrated by the grey bars in the figure. Mean RSD over all sample peaks for human annotators was 2.5%. In parallel, both sequential PB-Net and reference-based PB-Net performed well on the subset, achieving mean relative error of 3.5% and 2.1% respectively compared with human annotator average. This suggests that the differences between the algorithmic and human annotations are comparable to the variation between human annotators. The distributions of error, as shown in the blue and green bars, were similar but slightly worse than the RSD across annotators.

Viewing predictions from each individual annotator as a group, we then calculated pairwise differences between annotators through the average of the relative abundance differences on each sample peak. The outcomes showed a discrepancy between annotators at 0.025 ± 0.009, in line with the RSD value. At the same time, difference between the reference-based PB-Net prediction and the 13 annotators was 0.026 ± 0.007, overlapping with the inter-annotator difference. Corresponding value for the reference-free variant was 0.039 ± 0.005. Given that the model-human difference was not significantly larger, or even close in the reference-based case, it is clear that PB-Nets can serve as good substitutes for manual annotations in this task. We also noted that the human annotators had a peak with selected start and stop to serve as a reference for their reads, making their task directly analogous to the reference-based PB-Net (see Supplementary Information for details).

3.5. Better performances for "confident" predictions

Another advantage of the point-wise prediction structure is its easiness in estimating prediction certainty. In parallel with internal ways of quantifying uncertainty [6,8,12,21], we proposed a heuristic of inferring prediction confidence by inspecting the model prediction curves

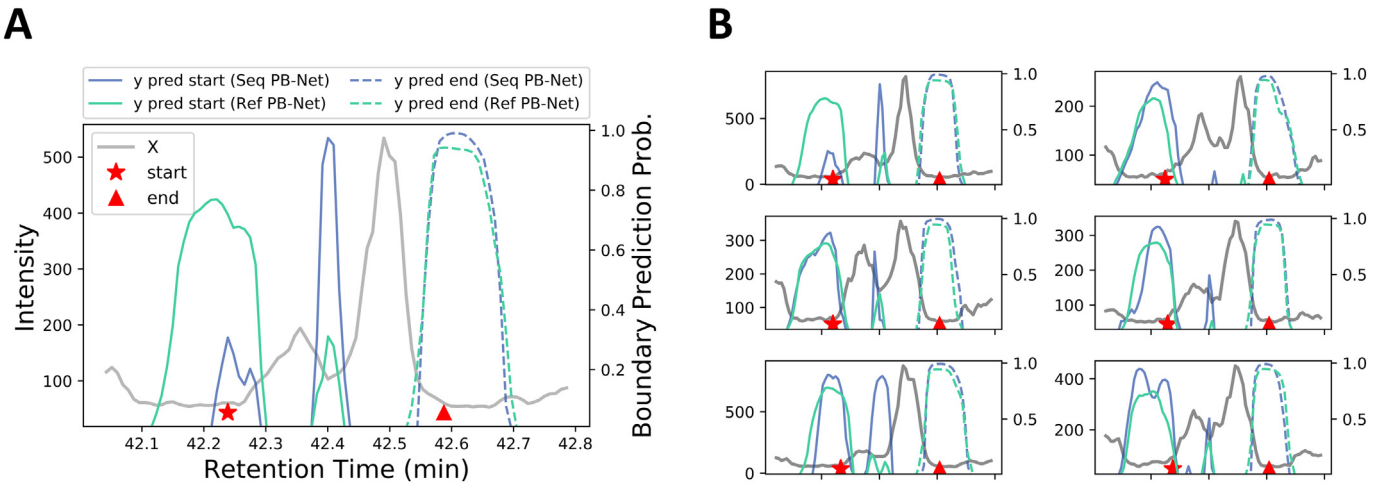

Fig. 3. Reference and sample peaks for a representative transition. A: In reference peak, grey line represents the input curve; red star and red triangle are human annotated peak start and end; blue/green solid line and dashed line indicate predictions from sequential/reference-based PB-Net of peak start and end probabilities. B: Sample peaks for the same transition with different input curves. Note that reference-based models (green) output much more consistent predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for boundaries. Detailed formulas and explanations are elaborated in the Supplementary Information. A graphical illustration of the method is presented in Fig. S3, in which the confidence score is proportional to the shaded area. In short, we would like the prediction values to be high (close to 1) and uni-modal, or close if multiple local maximums appear.

We calculated confidence scores for all predictions from both PB-Nets on the test set and illustrated the histogram in Fig. S4A and C. Across different confidence cutoffs, we observed a clear trend of performance change. As shown in Fig. S4B and D, prediction errors on boundaries steadily decreased with increasing confidence, dropping to near 1 s on the most confident samples.

Correlation of abundances on different confidence intervals are illustrated in Fig. S5. Predictions on high-abundance samples typically had higher confidence and aligned better with human annotations. On the top two bins predictions achieved Pearson's r over 0.999. In contrast, samples with lower prediction confidence yielded worse performance, with higher boundary MAEs and lower abundance correlations. This was in part due to the worse signal-noise ratio of input samples as indicated by the low abundances. Overall, the measurement of confidence served as a good indicator of noise scale and model performance.

4. Discussion

In this work, we demonstrated a sequential neural network built with LSTM and attention blocks for chromatographic peak quantification in multiple reaction monitoring experiments. Two variants were designed for the task: the reference-free sequential PB-Net can solely work on the sample input and provide reasonable estimates of peak start and end; the reference-based PB-Net further improved the prediction accuracy through incorporating human-annotated reference information. In consideration of real world usage, the reference-free version will be universally applicable, especially for experiments on small scale or designed for discovery purposes. The reference-based PB-Net, as restricted by requirement of human-annotated samples, can only be utilized for applications on larger scales, with its higher accuracy and consistency as return.

We trained and tested the above models on two datasets acquired

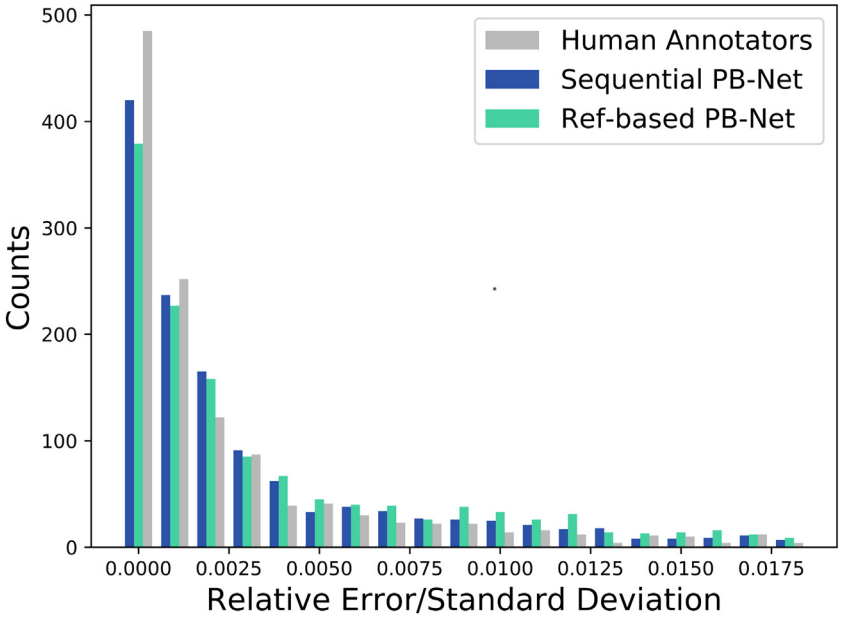

Fig. 4. Relative standard deviations (RSD) of human annotators and relative errors (RE) of PB-Net predictions. Grey bar illustrated RSD between 13 sets of independent human annotations, blue and cyan bars showed RE of sequential and reference-based PB-Net predictions against means of human annotations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from different sources and with different contents (transitions). Outcomes showed that both models significantly outperformed the baseline Skyline prediction and rule-based model on boundary and abundance prediction tasks. Robustness of predictions can be improved by adopting a weighted average step in abundance calculation, which helped in avoiding outliers from predictions with abnormal argmax points. A further comparison with human annotators proved their practicability as their performances lied in the region of normal variations within different annotators. The other contribution of this work consisted in providing a method of interpreting output structure of the sequential models, as we demonstrated a heuristic for estimating prediction confidence. A clear trend of higher performance on more confident predictions was observed, with most peaks with large abundances also being highly confident in their predictions. This measurement provided versatility in implementing workflow quality control, such as setting up a threshold confidence where predictions below the value should be excluded or presented to human experts for re-evaluation.

There are some limitations for PB-Net and its applications. Though predictions on Gaussian, high-abundance peaks are typically accurate, predictions will become relatively noisy in cases of inputs with low signal-noise ratio. The direct consequence would be arbitrary argmax boundaries, as the apex points of predicted probability curves are highly variable under minor changes of input. In the text, we proposed two ways to circumvent the issue: by imposing a reference prior, and by using weighted average abundances, though both come with limitations. Calculating and presenting the confidence could be another solution, such as abstaining when predictions are highly inconfident. A potential improvement will be to propose multiple possible boundaries for a single noisy input and evaluate them either through post-processing (including having experts check) or by calculating voting scores from network's output. As these changes involve complicated changes in either network structure or dataset preparation, they are interesting directions for future work.

The other issue in applications is that a rough retention time window must be specified for the predictions. As this work mainly focuses on determination of precise boundaries given a rough window of a transition, the whole peak must be located within the input window and there should not be any other significant peaks in the same window. In the test set we collected, fixed windows were applied on the set of pre-defined transitions which did not have identified co-eluting components, and experiments were carried out in relatively stable conditions. In general applications, this step of window selection should be done prior to the model predictions, which fits most selective reaction monitoring experiments. In cases when such information is not available, a feasible pipeline will be setting up a larger window and applying detection algorithms to capture the rough positions of peaks prior to predictions. Given the complex composition of serum samples, more exterior information on the surrounding co-eluting peaks, additional transitions, or a very comprehensive reference sample will be necessary to support this step.

A common technique is to utilize multiple transitions to refine the quantification of one component. This has not been discussed in this work, due to the limitations from the setup of our experimental data. Intuitively, this could be achieved by overlaying predictions for multiple transitions in the same retention time period and performing a “pooling” operation to combine the posteriors. This post-processing step should improve both the accuracy and the smoothness/signal-to-noise ratio of model predictions, and we would like to perform a more detailed study for future experiments.

5. Conclusion

In summary, PB-Net provides an accurate and cost-efficient substitute for the conventional manual peak picking pipeline. Maintaining the same level of accuracy, our fully automatic model reduces annotation time cost from 30 s per peak to a matter of milliseconds. Though certain caveats exist, PB-Net demonstrate the utility of deep learning approaches in MS. We believe this contribution will facilitate further advancements in mass spectrometry data analysis, as well as applications involving high-throughput MS experiments.

Declaration of Competing Interest

The work was performed while Z.W., D.S. and G.X. were employees of InterVenn BioSciences. J.Z. is an scientific advisor to InterVenn.

Acknowledgements

The authors would like to thank Carlito Lebrilla and members of his lab at UC Davis for generously providing the raw MS output used for initial training of the bi-directional LSTM. This work was supported by funding from InterVenn BioSciences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2020.103820.

References

[1] Paszke Adam, Gross Sam, Chintala Soumith, Chanan Gregory, Edward Yang, D. Zachary, Lin Zeming, Desmaison Alban, Antiga Luca, Lerer Adam, Automatic differentiation in pytorch, Proceedings of Neural Information Processing Systems, 2017.
[2] Ken Aoshima, Kentaro Takahashi, Masayuki Ikawa, Takayuki Kimura, Mitsuru Fukuda, Satoshi Tanaka, Howell E. Parry, Yuichiro Fujita, Akiyasu C. Yoshizawa, Shin-ichi Utsunomiya, et al., A simple peak detection and label-free quantitation algorithm for chromatography-mass spectrometry, BMC Bioinforma. 15 (1) (2014) 376.
[3] Michael S. Bereman, Brendan MacLean, Daniela M. Tomazela, Daniel C. Liebler, Michael J. MacCoss, The development of selected reaction monitoring methods for targeted proteomics via empirical refinement, Proteomics 12 (8) (2012) 1134–1141.
[4] Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, Markus Ralser, Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput, Nat. Methods 17 (1) (2020) 41–44.
[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[6] Yarin Gal, Zoubin Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1050–1059.
[7] Felix A. Gers, Douglas Eck, Jürgen Schmidhuber, Applying lstm to time series predictable through time-window approaches, Neural Nets WIRN Vietri-01, Springer, 2002, pp. 193–200.
[8] Chuan Guo, Geoff Pleiss, Sun Yu, Kilian Q. Weinberger, On calibration of modern neural networks, Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1321–1330.
[9] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[10] Daren Kessner, Matt Chambers, Robert Burke, David Agus, Parag Mallick, Proteowizard: open source software for rapid proteomics tools development, Bioinformatics 24 (21) (2008) 2534–2536.
[11] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations, 2015.
[12] Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.
[13] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[14] Qiongyu Li, Muchena J. Kailemia, Alexander A. Merleev, Gege Xu, Daniel Serie, Lieza M. Danan, Fawaz G. Haj, Emanual Maverakis, Carlito B. Lebrilla, Site-specific glycosylation quantitation of 50 serum glycoproteins enhanced by predictive glycopeptidomics for improved disease biomarker discovery, Anal. Chem. 91 (8) (2019) 5433–5445.
[15] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, Clara I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[16] Chunwei Ma, Yan Ren, Jiarui Yang, Zhe Ren, Huanming Yang, Siqi Liu, Improved peptide retention time prediction in liquid chromatography through deep learning, Anal. Chem. 90 (18) (2018) 10881–10888.
[17] Brendan MacLean, Daniela M. Tomazela, Nicholas Shulman, Matthew Chambers,

Gregory L. Finney, Barbara Frewen, Randall Kern, David L. Tabb, Daniel C. Liebler, Michael J. MacCoss, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics 26 (7) (2010) 966–968.
[18] Danilo P. Mandic, Jonathon Chambers, Recurrent neural networks for prediction: learning algorithms, architectures and stability, John Wiley & Sons, Inc, 2001.
[19] Suzanne Miyamoto, Carol D. Stroble, Sandra Taylor, Qiuting Hong, Carlito B. Lebrilla, Gary S. Leiserowitz, Kyoungmi Kim, L. Renee Ruhaak, Multiple reaction monitoring for the quantitation of serum protein glycosylation profiles: Application to ovarian cancer, J. Proteome Res. 17 (1) (2018) 222–233 (PMID: 29207246).
[20] Lukas N. Mueller, Mi-Youn Brusniak, D.R. Mani, Ruedi Aebersold, An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data, J. Proteome Res. 7 (1) (2008) 51–61.
[21] Radford M. Neal, Bayesian learning for neural networks, 118 Springer Science & Business Media, 2012.
[22] Hannes L. Rst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E. Wolski, Oliver Schilling, Jyoti S. Choudhary, Lars Malmstrm, Ruedi Aebersold, Knut Reinert, Oliver Kohlbacher, Openms: a flexible open-source software platform for mass spectrometry data analysis, Nat. Methods 12 (3) (2015) 258–264.
[23] Birgit Schilling, Matthew J. Rardin, Brendan X. MacLean, Anna M. Zawadzka, Barbara E. Frewen, Michael P. Cusack, Dylan J. Sorensen, Michael S. Bereman, Enxuan Jing, Christine C. Wu, et al., Platform-independent and label-free quantitation of proteomic data using ms1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation, Mol. Cell. Proteomics 11 (5) (2012) 202–214.
[24] Martin Sundermeyer, Ralf Schlüter, Hermann Ney, Lstm neural networks for language modeling, Thirteenth Annual Conference of the International Speech Communication Association, 2012.
[25] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to sequence learning with neural networks, Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
[26] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, Ming Li, De novo peptide sequencing by deep learning, Proceedings of the National Academy of Sciences, 114(31) 2017, pp. 8247–8252.
[27] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, Alexey I. Nesvizhskii, Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics, Nat. Methods 26 (7) (2010) 966–968.
[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., Matching networks for one shot learning, Advances in Neural Information Processing Systems, 2016, pp. 3630–3638.
[30] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, Vijay Pande, Moleculenet: a benchmark for molecular machine learning, Chem. Sci. 9 (2) (2018) 513–530.
[31] Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, Zhifei Zhang, pdeep: Predicting ms/ms spectra of peptides with deep learning, Anal. Chem. 89 (23) (2017) 12690–12697.
[32] Fatema Tuz Zohora, M. Ziaur Rahman, Ngoc Hieu Tran, Lei Xin, Baozhen Shan, Ming Li, Deepiso: A deep learning model for peptide feature detection from lc-ms map, Sci. Rep. 9 (1) (2019) 1–13.
[33] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, Amalio Telenti, A primer on deep learning in genomics, Nat. Genet. 1 (2018).