PepNet: A Fully Convolutional Neural Network for De novo Peptide Sequencing

Kaiyuan Liu
Indiana University Bloomington https://orcid.org/0000-0002-3404-2802
Yuzhen Ye
School of Informatics and Computing, Indiana University Bloomington
Haixu Tang (hatang@indiana.edu)
Indiana University Bloomington

Methods Article

Keywords:

Posted Date: February 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1341615/v1

PepNet: A Fully Convolutional Neural Network for *De novo* Peptide Sequencing

Kaiyuan Liu, Yuzhen Ye, and Haixu Tang*

Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana 47405, United States

E-mail: hatang@indiana.edu

Abstract

The *de novo* peptide sequencing, which does not rely on a comprehensive target sequence database, provided us a way to identify novel peptides from tandem mass (MS/MS) spectra. However, current *de novo* sequencing algorithms suffer from lower accuracy and coverage, which hinders their applications in proteomics. In this paper, we present *PepNet*, a fully convolutional neural network (CNN) for high accuracy *de novo* peptide sequencing. It takes an MS/MS spectrum (represented as a high dimensional vector) as input, and outputs the optimal peptide sequence along with its confidence score. Our model was trained using a total of 30 million high-energy collisional dissociation (HCD) MS/MS spectra from multiple human peptide spectral libraries. The evaluation results show that PepNet significantly outperformed currently best-performing *de novo* sequencing algorithms (e.g. PointNovo and DeepNovo) at both peptide level accuracy and positional level accuracy. In addition, PepNet can sequence a large fraction of spectra that were not identified by database search engines, and thus could be used as a complementary tool of database search engines for peptide identification in proteomics.

Introduction

The past decade has witnessed the great advances of mass spectrometry techniques, in particular liquid chromatography coupled tandem mass spectrometry (LC-MS/MS). With enhanced throughput and sensitivity, LC-MS/MS has become one of the most commonly used approaches to functional studies of proteins at the whole proteome scale across various physiological (e.g., diseases) conditions in higher organisms including human.

In a typical proteomics experiment, after MS/MS spectra are acquired, the first and arguably most important step is to identify the peptides from these spectra. Numerous algorithms have been developed to

address this problem, which mostly falls into three categories: protein database searching, spectral library searching, and *de novo* sequencing. Pioneered the peptide sequence tag method[1] and the Sequest algorithm,[2] protein database searching is the predominant approach used for peptide identification. Till now many peptide searching engines have been developed, including commonly used tools such as Mascot ,[3] X!Tandem ,[4] OMSSA,[5] MyriMatch,[6] Protein Prospector[7,8] and MSGF+.[9] Those methods compare the experimental spectra with the theoretical spectra generated from the peptides in a protein database and report those likely true peptide-spectrum matches (PSMs).

In contrast, the spectral library search approach compares newly acquired MS/MS spectra against a library containing previously characterized experimental spectra that were used in early computational analysis .[10,11] Thanks to the improved repeatability and reproducibility of MS/MS data as well as the increasing availability of massive experimental spectra (e.g., from the proteomics data repository [12] and large-scale synthetic peptides projects [13]), the spectral library search approach has become more increasingly adopted and is implemented in software tools such as X!hunter,[14] SpectraST[15] and msSLASH .[16]

Finally, *de novo* sequencing algorithms attempt to derive a peptide sequence directly from its MS/MS spectrum without using references such as a spectral library or a protein sequence database.[17] Many *de novo* sequencing algorithms adopted a graph theoretical formulation to compute the longest path in the *spectrum graph* by employing a dynamic programming algorithm[18,19] and adaptive scoring schemes .[20–22] With the advancement of high-resolution MS instruments, the performance of *de novo* sequencing algorithms improves significantly ,[23,24] in particular with more sophisticated scoring schemes. More recently, DeepNovo [25–27] and its successor PointNovo[28] was developed using deep learning algorithms, which automatically learn the fragment ion patterns relevant to peptide sequences from massive MS/MS spectra of peptides and reported improved performance.

Results

Accurate HCD-MS/MS spectra *de novo* sequencing by deep learning

We present a deep learning algorithm, *PepNet*, that directly outputs the peptide sequence from a given HCD-MS/MS spectra with high accuracy. As depicted in Figure 1, the input for our model is a $20,000$ dimensional vector representation of the spectrum (for details see the Method section). The input vector will go through six continuing temporal convolutional network (TCN) blocks[29] and down-sampling layers to capture the relationships between observed peaks, as depicted in the TCN branch of Fig. 1. Apparently, those TCN blocks working on different *resolution* levels. Thus we add a merging branch (top-down branch

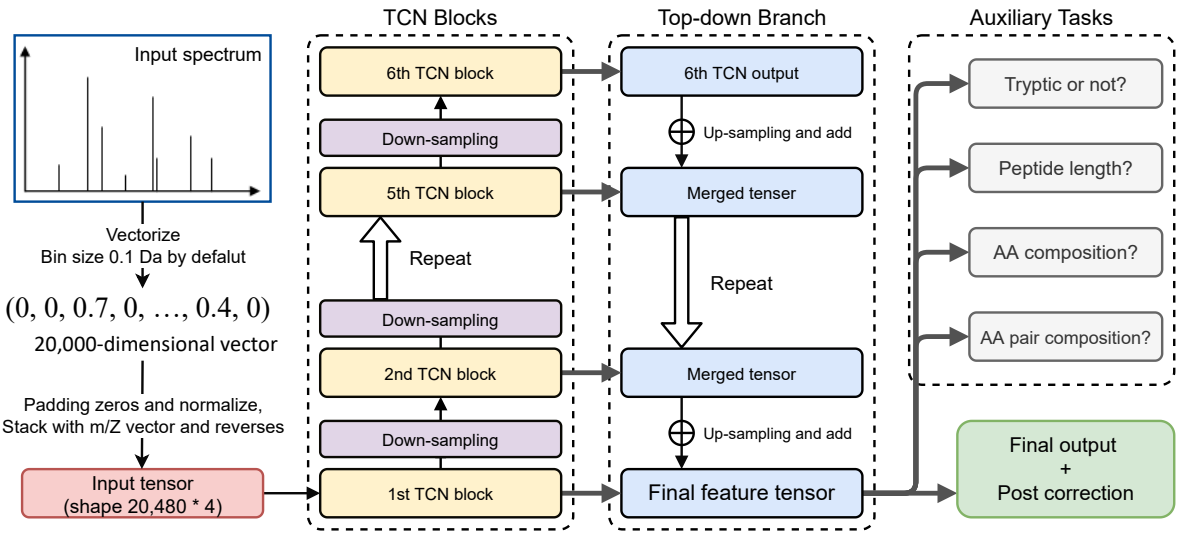

Figure 1: The neural network architecture of PepNet for *de novo* peptide sequencing.

in Fig. 1) that fuse the global and local information into a single feature tensor. Finally, the feature tensor is then converted by a softmax decoding layer into a probabilities matrix of size 30×23, in which each column contains the softmax probabilities of the target amino acid at the corresponding position in the peptide. From the final probabilities matrix, we derive the optimal peptide sequence by choosing the amino acid with the highest probability at each position. Additionally, the feature tensor is also sent to several relatively easy auxiliary tasks (auxiliary tasks branch in Fig. 1) to guide and regulate for the target *de novo* sequencing task.

It's worth noting that, we did not force the model to output a peptide that has the matched precursor mass; instead, we design the loss function to encourage the model to output the peptide sequence containing as many correct amino acids as possible. This approach can prevent the model from outputting a peptide with the desirable amino acid composition, but neglect the sequence information in the input MS/MS spectrum, in particular during the early stage of the training process when the model has not learned sufficient rules and patterns.

We collected the HCD spectra from multiple peptide spectral libraries (see details in the Method section) by retaining spectra with the charges of 1+ to 4+ and from peptides of length no longer than 30, which result in 3,041,570 HCD-MS/MS spectra from 1,066,296 distinct peptides. We implemented using Tensorflow [30] and train the model using the Adam optimizer [31] for 50 epochs (using learning rate of 0.001). The entire PepNet model contains about 10 million parameters. Notably, we did not distinguish spectra that were acquired by different types of instruments during training and testing, as we observed that the HCD spectra acquired using different instruments (e.g., Orbitrap, Fusion, or Q-Exactive) could all be sequenced at high accuracy

by a single model. PepNet is released as open-source software for *de novo* peptide sequencing at GitHub (https://github.com/lkytal/PepNet). We also provided an online tool (https://denovo.predfull.com/) for submitting MS/MS spectra (in MGF format) to *de novo* peptide sequencing.

Evaluation Criteria

We evaluated the performance of PepNet over several proteomics datasets. For these datasets, we used the database searching results from MaxQuant [32] with scores above the threshold of 80 and the precursor mass differences smaller than 100 ppm as the ground truth to compute the accuracy of the *de novo* sequencing results. Then we computed two measurements to evaluate the accuracy of the *de novo* peptide sequencing: the *positional accuracy*, defined as the fraction of correct amino acid residues reported by PepNet, and the *peptide-level accuracy*, defined as the fraction of spectra that the sequenced peptides are completely correct. Note that in both measurements, we do not distinguish the amino acids of Leucine and Isoleucine. The two accuracy measurements, as well as the Precision-Recall (PR) curves of PepNet, were compared with the current state-of-the-art *de novo* peptide sequencing algorithm PointNovo [26] and its predecessor DeepNovo .[25] Finally, we present here only the *de novo* sequencing of 2+ and 3+ HCD spectra of unmodified peptides as they are most common in shotgun proteomics data, although this model can be easily extended to sequencing spectra of other charges or sequencing those from the peptides containing common Post-translational modifications (PTMs).

Performance evaluation on a large-scale human proteomics data set

We first evaluated the performance of PepNet using the HCD spectra from a large-scale human proteomics project (ProteomExchange ID: PXD019483) in comparison with PointNovo and DeepNovo. Here, we report the *de novo* sequencing results on the subset of identified spectra from the original study ,[33] in which 468,266 charge 2+ and 87,977 charge 3+ spectra were identified by MaxQuant (i.e., receiving the score 80 or above and mass difference smaller than 100 ppm, as described above), respectively. We will report the *de novo* sequencing results on the remaining unidentified spectra in a separate section below.

As shown in Fig. 2, PepNet reported completely correct peptides on 79.1% 2+ and 51.4% 3+ identified spectra, on which the positional accuracy reaches 92.4% and 77.7%, respectively. In comparison, among the peptides reported by PointNovo, 70.4% and 44.2% are completely correct for 2+ and 3+ spectra, respectively, on which the positional accuracies are 72.6% and 55.3%, respectively; among the peptides reported by DeepNovo, only 53.2% and 24.5% are completely correct for 2+ and 3+ spectra, respectively, on which the positional accuracies are 72.5% and 48.5%, respectively. Also, the precision-recall (PR) curves (Fig. 3)

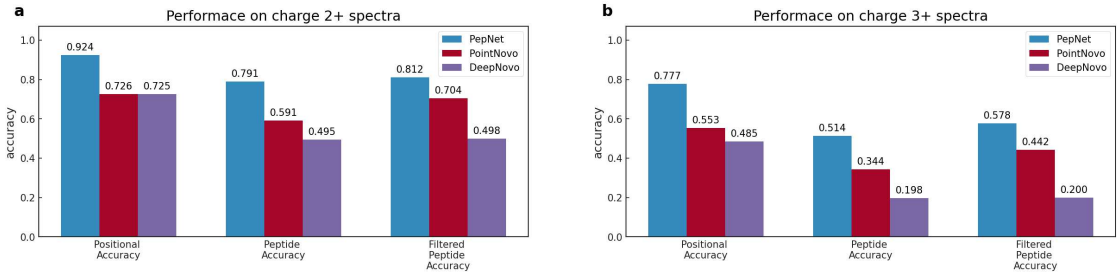

Figure 2: Comparison of the performance of PepNet, PointNovo and DeepNovo on the charge 2+ (a) and charge 3+ (b) spectra in the human proteomics dataset. Here, the *Filtered Peptide Accuracy* is referred to as the peptide-level accuracy on the sequenced peptides after removing those with unmatched precursor masses.

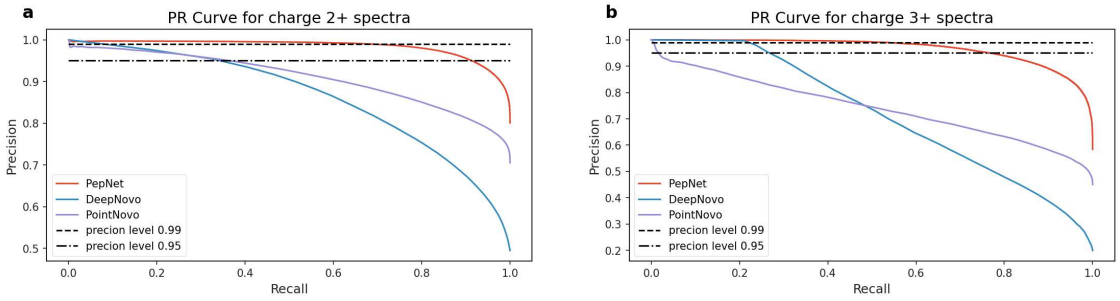

Figure 3: The precision-recall (PR) curves of the peptides sequenced by PepNet, PointNovo, and DeepNovo on the spectra of charge 2+ (a) and charge 3+ (b) in the human proteomics dataset. The two dotted lines represent the precision of 0.99 and 0.95, respectively.

showed that PepNet can sequence most peptides with higher accuracy; in particular, it can sequence the peptides from a majority of spectra (i.e., high recall) even when a high threshold of precision (95% or 99%) is used.

We further investigated the performance of PepNet, PointNovo, and DeepNovo on spectra from the peptides of different lengths, as depicted in Fig. 4. Not surprisingly, the performance of PepNet (as well as PointNovo and DeepNovo) is reduced with the increasing lengths of peptides, perhaps because 1) longer peptides are more challenging to be *de novo* sequenced (with less peptide sequence information in their MS/MS spectra), and 2) the training dataset contains relatively fewer training samples of longer peptides. However, the performance of PepNet is persistently better than PointNovo and DeepNovo, especially for longer peptides.

Performance evaluation on proteomics data from non-human organisms

Next, we evaluated the performance of PepNet on the proteomics data from a variety of non-human organisms. We used the HCD spectra from a large-scale proteomics project (ProteomExchange ID: PXD014877) aiming to elucidate the evolutionary landscape of the proteomes in 57 organisms including bacteria, fungi,

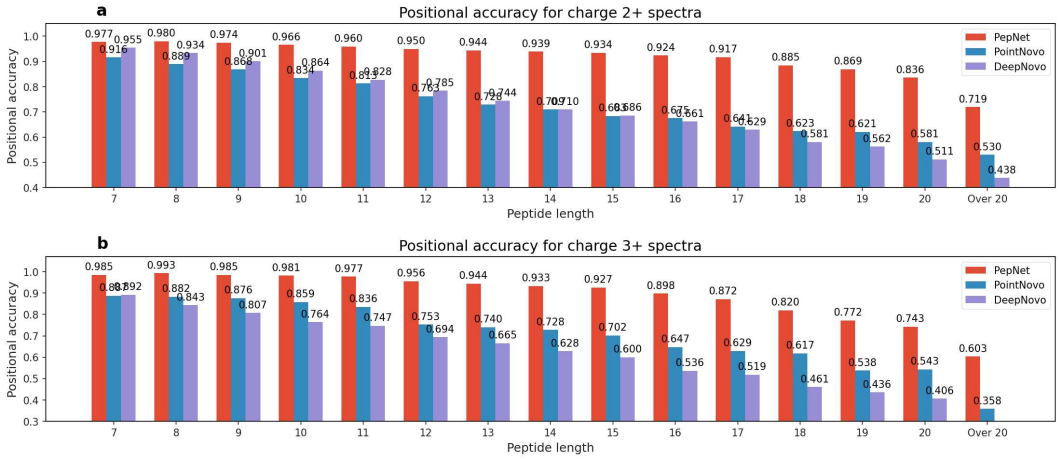

Figure 4: The performance of PepNet, PointNovo, and DeepNovo on the peptides with different length for the spectra of charge 2+ (top) and charge 3+ (bottom) in the human proteomics dataset. Here, *filtered peptide accuracy* indicates the peptide-level accuracies of the sequenced peptides after removing those with un-matched precursor masses.

plants, and animals. Similar to the results shown above, here we report the *de novo* sequencing results on the subset of identified spectra from the original study .[33]

As illustrated in Fig. 5, the performance of PepNet is consistent over various organisms. The positional accuracy on the peptides reported by PepNet is 0.87 ± 0.02 on the sequenced peptides from 2+ spectra, and 0.79 ± 0.07 on the sequenced peptides from 3+ spectra, while the peptide-level accuracy are 0.76 ± 0.03 and 0.49 ± 0.09 on 2+ and 3+ spectra, respectively. The performance of PointNovo is consistently lower than PepNet by about 0.1 for the positional accuracy and by about 0.15 for the peptide-level accuracy, while the performance of DeepNovo is even lower. Notably, the performance of PepNet on the proteomics data from these varieties of organisms is largely comparable with the performance on the human proteome data (see above), indicating that even though PepNet was trained using MS/MS spectra mostly from human peptides, the model is generalized well for the *de novo* sequencing of non-human peptides.

De novo sequencing of spectra not identified by database searching engines

Here, we demonstrate that PepNet can identify a large fraction of MS/MS spectra that cannot be confidently identified by the database searching engines. We applied PepNet to the subset of MS/MS spectra that were not unidentified by MaxQuant in the human proteomics dataset (ProteomExchange ID: PXD019483, as described above). We then computed a *confidence score* for each sequenced peptide as the product of the probabilities of the amino acid at all positions in the peptide. After removing sequenced peptides with unmatched precursor masses or with a confidence score lower than 0.5, PepNet reported 5,772,599 unique peptides from 8,856,090 MS/MS spectra (out of the whole set of 19,360,564 MS/MS spectra). Note that

Figure 5: Performance of PepNet, PointNovo and DeepNovo on the spectra of charge 2+ (a) and charge 3+ (b) in the proteomics datasets acquired from different species.

here the confidence score reflects the probability of the sequenced peptide to be correctly considered by the model; hence, we choose the threshold of 0.5, when the probability of the peptide to be correct is higher than that to be incorrect.

We then searched these sequenced peptides against the human proteins in the UniProt[34] database by using RAPSearch2[35] (without distinguishing amino acids of Leu and Ile). We observed that 375,043 (4.2%; including 54,644 unique peptides) of these sequenced peptides matched perfectly with human proteins, while another 179,368 (2.0%; including 35,083 unique peptides) contain only one substitution comparing with human proteins (Fig. 6). These results suggest that the *de novo* sequencing by PepNet is complementary to the

database searching engines (such as MaxQuant), as further analyses of the *de novo* sequenced peptides that were not identified by database searching engines may lead to the discovery of additional peptides/proteins expressed in the proteome samples, including those not present in the target protein database (e.g., the peptides containing substitutions).

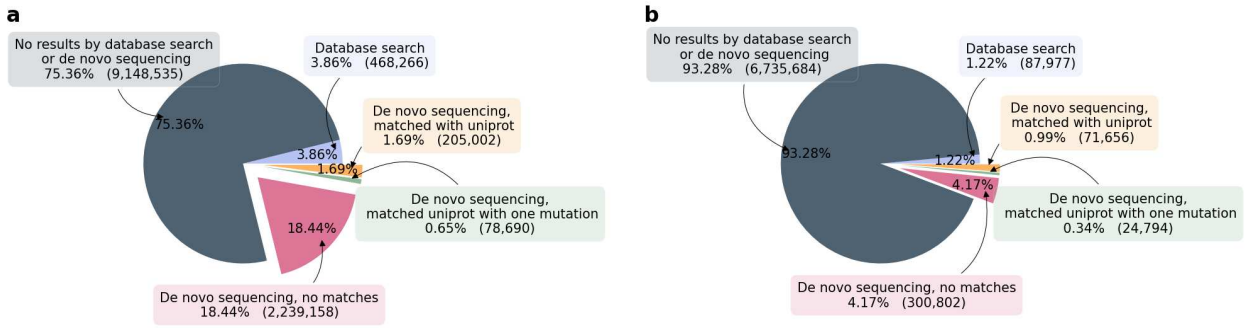

Figure 6: Comparison of *de novo* sequencing results of PepNet (the pulled out parts) with database searching (identified by MaxQuant) on spectra of (a) charge 2+, and (b) charge 3+. Here we only retain *de novo* sequencing results that with matched precursor mass and confidence score > 0.5, which further searched against the Uniprot human database.

Performance of PepNet on proteomics data from Data-Independent Acquisition (DIA)

We further demonstrate that PepNet is also capable of the *de novo* sequencing of the MS/MS spectra derived from the data acquired by using Data Independent Acquisition (DIA). We compared the performance of PepNet against PointNovo and DeepNovo-DIA on a dataset acquired from a human plasma sample (used for the evaluation of DeepNovo-DIA [36]), which contains a total of 57,909 MS/MS spectra derived from DIA data. Notably, DeepNovo-DIA is a refined DeepNovo model using DIA-derived MS/MS spectra as training data; in contrast, we directly applied the PepNet model trained on the HCD-MS/MS spectra from Data Dependent Acquisition (DDA) as described above without any further refinement.

It's not surprising that the performance of PepNet is significantly better than PointNovo, as PointNovo was not trained on the DIA data. Interestingly, PepNet also outperforms DeepNovo-DIA which is especially fine-tuned for the DIA-derived spectra. As shown in Figure 7, PepNet achieved comparable performance on the DIA-derived spectra as on the DDA-acquired spectra, with the positional accuracy of 0.69 and peptide-level accuracy of 0.49 on the combined set of 2+ and 03+ spectra. These results indicate the PepNet model is robust for the *de novo* sequencing of not only the DDA-acquired MS/MS spectra but also the MS/MS spectra derived from DIA data.

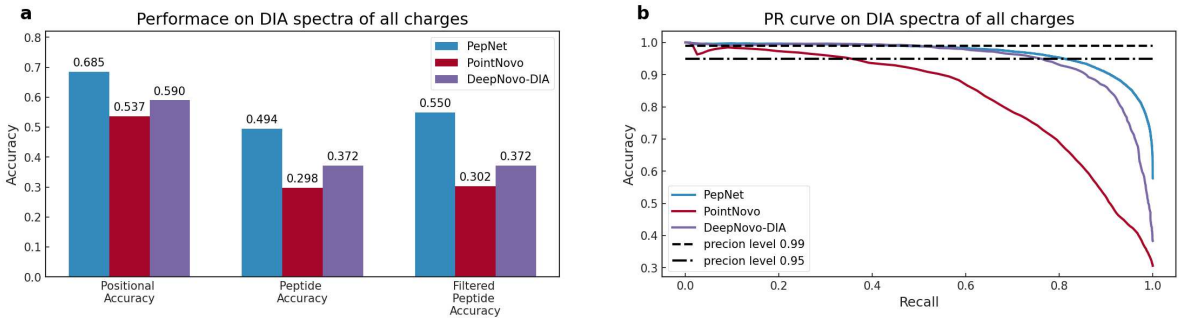

Figure 7: The sequencing accuracy (left) and the PR curve (right) of PepNet, PointNovo, and DeepNovo-DIA on a dataset of DIA-derived MS/MS spectra. The *Filtered Peptide Accuracy* is referred to as the peptide-level accuracy on the sequenced peptides after removing those with unmatched precursor masses. The two dotted lines in the PR curve represent the precision of 0.99 and 0.95, respectively.

Observed factors related to performance

We found multiple factors influencing the performance of PepNet that are worth discussing. We observed that the number of most intensive peaks retained in the input MS/MS spectrum to PepNet has a significant impact on its performance, as shown in Fig. 8; the performance of PepNet significantly increases when more peaks are retained in the spectra given as input the PepNet. This result indicates that the peaks of low intensities, including some non-backbone fragment ions that were considered as "noise" peaks and thus ignored by conventional *de novo* sequencing algorithms, also provide useful information for the PepNet model, especially helpful for determining the residues at some positions where the supportive backbone ions are missing.

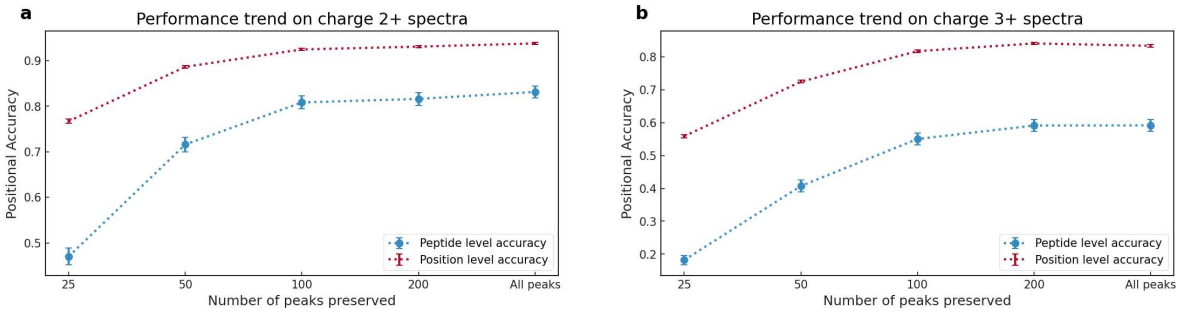

Figure 8: The positional and peptide-level accuracy of PepNet on the input 2+ (a) and 3+ (a) spectra in the testing dataset, on which different number of most intensive peaks are retained. The error bar represent the 99% confidence interval.

Besides, we observed that the positional accuracy by PepNet shared a similar trend for the sequenced peptides of different lengths. As shown in Fig. 9, for the charge 2+ spectra, the positional accuracy is low for the first few amino acid residues at the N- terminus, and becomes very high before decreasing gradually until the last residue at the C-terminus. It's not surprising that the C-terminal residue is determined accurately

because most tested spectra are from tryptic peptides. This trend of sequencing error distribution is largely due to the coverage of the observed fragment ions: the first few residues are easier to determine because the b_1 and b_2 (and even some b_3) ions are often missing in HCD spectra, while y-ions are weaker than b-ions, causing the C-terminal residues are harder to be determined.

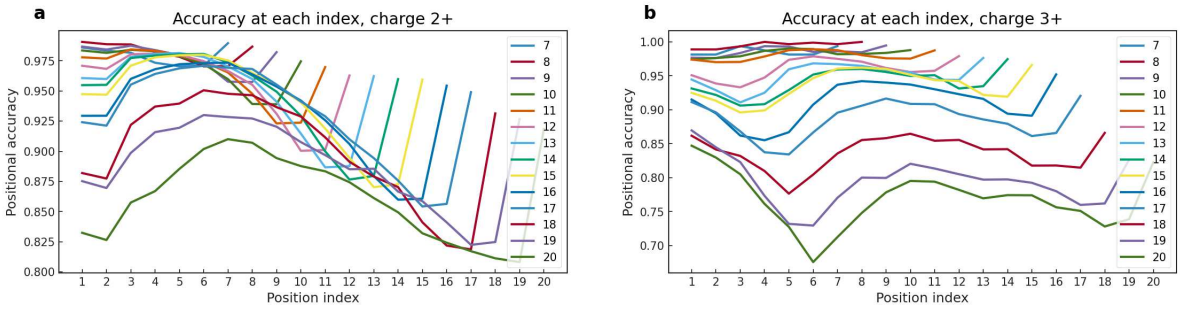

Figure 9: The positional accuracy at each index of the peptides, for sequenced peptide length from 6 to 20. (a) for 2+ spectra. (b) for 3+ spectra.

Discussion

In this paper, we present PepNet, a novel deep learning model for high accuracy *de novo* peptide sequencing from HCD-MS/MS spectra. We first show that PepNet is capable of sequencing human MS/MS spectra with high accuracy (with 92.4% and 77.7% positional accuracy on 2+ and 3+ human spectra from PXD019483, respectively), then we show that the PepNet can also perform constantly well across the proteomes of many non-human organisms (by testing on dataset PXD014877). Furthermore, the *de novo* results on unidentified spectra show that PepNet could discover a comparable number of new identifications as the protein database search engines, and thus can be used as a powerful tool for proteomic data analyses when a comprehensive target protein sequence database is not available (e.g., in metaproteomics [37]).

We believe that the ability to sequence peptides with high accuracy will enable the increasing applications of *de novo* peptide sequencing in proteomics data analyses. In addition to the peptide sequencing for HCD spectra as presented in this paper, PepNet can be extended to the MS/MS spectra acquired by using other fragmentation methods, such as the electron transfer dissociation (ETD), Electron-Transfer/Higher-Energy Collision Dissociation (EThcD), photodissociation (PD) and the infrared multiphoton dissociation (IRMPD). These methods were often considered to result in complex MS/MS spectra, in which those rich information embedded in the complex MS/MS spectra may hopefully improve the accuracy of *de novo* peptide sequencing. Furthermore, the low positional sequencing error rate of PepNet will enable the *de novo* sequencing of whole proteins (e.g., antibodies) based on overlapping peptides (e.g., generated from multiple protease cleavages). Therefore, we anticipate that PepNet will enhance the efficiency in proteomics data analyses and will benefit

the community of life science research.

Methods

Representation of the MS/MS spectra

We represent the input MS/MS spectrum as a one-dimensional (1-D) vector by binning the spectrum with a given bin width. We considered only the peaks within the range of mass-to-charge ratio (m/z) between 0 and 2000 as most experimental spectra do not contain peaks with m/z above 2,000. By default, we use a bin width of 0.1, which yields the vector representation of 20,000 dimensions. Based on our experiment, using an even smaller bin size (i.e., higher mass resolution) did not improve the performance of *de novo* sequencing but required longer running times. Finally, we removed the precursor peak in each spectrum, and normalized each spectrum by dividing each peak over the intensity of the maximum peak in that spectrum, similar to our previous work for MS/MS spectrum prediction .[38]

Deep Neural Network Architecture

As depicted in the Result section (Figure 1), the input for our model is a $20,000$ dimensional vector representation of the spectrum as described in the previous section. As depicted in the TCN branch of Fig. 1, six continuing temporal convolutional network (TCN) blocks[29] and down-sampling layers are designed to capture the relationships between observed peaks.

Although those TCN blocks are capable of extracting most information from the input MS/MS spectrum, they work at different levels of *resolution*, and thus retrieve complementary information: the topmost TCN block emphasizes the detailed local structures of the input spectrum whereas the bottom-most block considers mostly the global features of the spectrum. We follow the idea of the Feature Pyramid Networks[39] by adding a top-down branch that merges output from all TCN blocks into a single feature tensor (thus fusing the global and local information together), as depicted in Fig. 1.

The above feature tensor will be concatenated with meta-information (e.g. charge of the spectra, M/z of the precursor, normalized collision energy) to yield the final feature tensor, which is then converted into a final probabilities matrix of the size 30×23 by a softmax decoding layer. As each column in the matrix represents the probabilities of each amino acid at the corresponding position, we can derive the optimal peptide sequence by choosing the amino acid with the highest probability column by column, until we meet the position at which the ending character is of the highest probability. As a post-correction to this strategy, if the theoretical mass of the inferred peptide differs from the experimental precursor mass more than 100

ppm, we attempt to check if any sub-optimal peptide has matched precursor mass: we substitute the amino acid at each position with the one with the second-highest probability; if the resulting peptide has the matched precursor mass, it will be output as the final *de novo* sequencing result; otherwise, the original optimal peptide sequence will still be reported.

In addition to the *de novo* sequencing task, several relatively easy tasks (auxiliary task branch in Fig. 1) are trained simultaneously for archiving better performance, including 1) whether the target peptide is a tryptic peptide; 2) the length of the target peptide; 3) the amino acid composition of the target peptide; 4) the composition of adjacent amino acid pairs in the peptide. These auxiliary tasks serve as guidance and regulations for the *de novo* sequencing task.

Because the PepNet model is fully convolutional, it decodes the entire peptide sequence simultaneously rather than one amino acid at a time as adopted by the Recurrent Neural Network (RNN) in DeepNovo [25,26].

Training Datasets and Process

To compile the training data set, we collected HCD spectra from multiple peptide spectral libraries including the NIST HCD library,[40] the NIST Synthetic HCD library,[40] the Human HCD library from MassIVE,[41] and the synthetic HCD library from ProteomeTools.[13] The number of spectra in these libraries is summarized in Supplementary Table 1. In total, by retaining spectra with the charges of 1+ to 4+ and from peptides of length no longer than 30, we collected 2,962,341 HCD-MS/MS spectra from 1,048,993 distinct peptides.

The whole data set was randomly split into the two subsets, one subset containing 2,908,323 (95%) spectra as the training data, and the other subset containing the remaining 133,247 (5%) spectra serves as the cross-validation set. We ensured that the training and testing data shared no spectra from the same peptide in order to avoid information leakage. The complete training and testing data sets are available in the supplementary files.

We use the Adam optimizer[31] with the learning rate of 0.001 to train the model for 50 epochs (batch size of 64 spectra per GPU). We warm-up the learning rate linearly from 0.0001 to 0.001 within the first 10 epochs for training stability. The complete training process takes around 20 hours using 8 cards of NVIDIA A6000 GPU. We pick the model weight from the epoch that performs best on the cross-validation set after the training finished, which achieved over 0.9 positional accuracy on charge 2+ spectra and over 0.78 positional accuracy on charge 3+ spectra. More details can be found in the *Training performance* section of the Supplementary Materials.

Configuration of PointNovo and DeepNovo

To build the PointNovo and DeepNovo for compassion, we directly used the source code provided by their original publication without modification. For DeepNovo (and also DeepNovo-DIA), we directly used the provided pre-trained weights for all testing tasks. While for PointNovo, as the authors did not provide pre-trained weights, we retrain the model using the intact training dataset provided in the paper. Both PointNovo and DeepNovo were executed by using their default configurations without modifications.

Data availability

The proteomic data of used for this study were taken from previous datasets with the identifier PXD019483 and PXD014877. The MaxQuant searching results provided in these studies were directly reused. The trained models of PepNet were deposited at https://zenodo.org/record/5807120 while the experiment results were deposited at https://zenodo.org/record/6003282.

Code availability

Source code and scripts are available on GitHub at https://github.com/lkytal/pepnet.

Acknowledgement

This work was supported by National Science Foundation (DBI-2011271), National Institutes of Health (1R01AI143254) and Indiana University Precision Health Initiative (PHI).

Competing interests

The authors declare that they have no conflict of interest.

Author contributions

K.L. and H.T. conceived the project. K.L. designed and implemented the deep learning model. K.L., Y.Y., and H.T. analyzed the data. K.L., Y.Y., and H.T. wrote the manuscript. All authors read and revised the manuscript.

References

(1) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical chemistry* **1994**, *66*, 4390–4399.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry* **1994**, *5*, 976–989.

(3) Hirosawa, M.; Hoshida, M.; Ishikawa, M.; Toya, T. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. *Bioinformatics* **1993**, *9*, 161–167.

(4) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(5) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *Journal of proteome research* **2004**, *3*, 958–964.

(6) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research* **2007**, *6*, 654–661.

(7) Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical chemistry* **1999**, *71*, 2871–2882.

(8) Chalkley, R. J.; Baker, P. R.; Medzihradszky, K. F.; Lynn, A. J.; Burlingame, A. In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Molecular & Cellular Proteomics* **2008**, *7*, 2386–2398.

(9) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* **2014**, *5*, 1–10.

(10) Ausloos, P.; Clifton, C.; Lias, S.; Mikaya, A.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O.; Zaikin, V.; Zhu, D. The critical evaluation of a comprehensive mass spectral library. *Journal of the American Society for Mass Spectrometry* **1999**, *10*, 287–299.

(11) Yates, J. R.; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Analytical chemistry* **1998**, *70*, 3557–3565.

(12) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building consensus spectral libraries for peptide identification in proteomics. *Nature methods* **2008**, *5*, 873–875.

(13) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M., et al. Building ProteomeTools based on a complete synthetic human proteome. *Nature methods* **2017**, *14*, 259.

(14) Craig, R.; Cortens, J.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research* **2006**, *5*, 1843–1849.

(15) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655–667.

(16) Wang, L.; Liu, K.; Li, S.; Tang, H. A Fast and Memory-Efficient Spectral Library Search Algorithm Using Locality-Sensitive Hashing. *Proteomics* **2020**, *20*, 2000002.

(17) Allmer, J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert review of proteomics* **2011**, *8*, 645–657.

(18) Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology* **1999**, *6*, 327–342.

(19) Chen, G.-m.; Firth, M.; Rui, O. M. The dynamic relation between stock returns, trading volume, and volatility. *Financial Review* **2001**, *36*, 153–174.

(20) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry* **2005**, *77*, 964–973.

(21) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics* **2012**, *11*.

(22) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R.-X.; Liu, J.; Zeng, W.-F.; Song, C.-Q.; He, S.-M., et al. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *Journal of proteome research* **2013**, *12*, 615–625.

(23) Jeong, K.; Kim, S.; Pevzner, P. A. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **2013**, *29*, 1953–1962.

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

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

(26) Qiao, R.; Tran, N. H.; Xin, L.; Shan, B.; Li, M.; Ghodsi, A. Deepnovov2: Better de novo peptide sequencing with deep learning. *arXiv preprint arXiv:1904.08514* **2019**,

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

(28) Qiao, R.; Tran, N. H.; Xin, L.; Chen, X.; Li, M.; Shan, B.; Ghodsi, A. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence* **2021**, *3*, 420–425.

(29) Bai, S.; Kolter, J. Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* **2018**,

(30) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M., et al. Tensorflow: A system for large-scale machine learning. 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016; pp 265–283.

(31) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). 2015.

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

(33) Müller, J. B.; Geyer, P. E.; Colaço, A. R.; Treit, P. V.; Strauss, M. T.; Oroshi, M.; Doll, S.; Winter, S. V.; Bader, J. M.; Köhler, N., et al. The proteome landscape of the kingdoms of life. *Nature* **2020**, *582*, 592–596.

(34) Consortium, U. UniProt: a hub for protein information. *Nucleic acids research* **2015**, *43*, D204–D212.

(35) Zhao, Y.; Tang, H.; Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **2012**, *28*, 125–126.

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

(37) Maron, P.-A.; Ranjard, L.; Mougel, C.; Lemanceau, P. Metaproteomics: a new approach for studying functional microbial ecology. *Microbial ecology* **2007**, *53*, 486–493.

(38) Liu, K.; Li, S.; Wang, L.; Ye, Y.; Tang, H. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical chemistry* **2020**, *92*, 4275–4283.

(39) Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; pp 2117–2125.

(40) Yang, X.; Neta, P.; Stein, S. E. Extending a Tandem Mass Spectral Library to Include MS 2 Spectra of Fragment Ions Produced In-Source and MS n Spectra. *Journal of The American Society for Mass Spectrometry* **2017**, *28*, 2280–2287.

(41) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell systems* **2018**, *7*, 412–421.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Suppliment.pdf