ARTICLES
https://doi.org/10.1038/s42256-021-00304-3

nature machine intelligence

Check for updates

Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices

Rui Qiao[1], Ngoc Hieu Tran[2], Lei Xin[3], Xin Chen[3], Ming Li[2], Baozhen Shan[3]✉ and Ali Ghodsi[1]✉

De novo peptide sequencing is the key technology for finding novel peptides from mass spectra. The overall quality of sequencing results depends on the de novo peptide sequencing algorithm as well as the quality of mass spectra. Over the past decade, the resolution and accuracy of mass spectrometers have improved by orders of magnitude and higher-resolution mass spectra have been generated. How to effectively take advantage of those high-resolution data without substantially increasing the computational complexity remains a challenge for de novo peptide sequencing tools. Here we present PointNovo, a neural network-based de novo peptide sequencing model that can robustly handle any resolution levels of mass spectrometry data while keeping the computational complexity unchanged. Our extensive experiment results show PointNovo outperforms existing de novo peptide sequencing tools by capitalizing on the ultra-high resolution of the latest mass spectrometers.

De novo peptide sequencing is the process of reconstructing a peptide sequence directly from a tandem mass spectrum and peptide mass. In the past 20 years, different de novo peptide sequencing tools have been proposed and successful applications have been shown in assembling monoclonal antibody sequences[1] and identifying tumour-specific antigens, especially those resulting from a non-coding region or alternative splicing[2,3]. However, it still remains challenging for a de novo peptide sequencing tool to discriminate between amino acids pairs that have similar masses, for example, glutamine (Q) and lysine (K), or methionine sulfoxide (M(Oxi)) and phenylalanine (F). For instance, when evaluating the accuracy of de novo peptide sequencing, some previous studies[4,5] considered a predicted amino acid matching a real amino acid if their mass difference is smaller than 0.1 Da and if the prefix masses before them differ by less than 0.5 Da. This means, for example, if a de novo sequencing tool reports a Q for a ground-truth K, it will still be labelled as correct by the evaluation criteria as the mass difference between Q and K is smaller than 0.05 Da; however, for antibody sequencing applications or tumour-specific antigen finding, it is important for the de novo sequencing tool to be able to reconstruct the exact sequence of a peptide. Otherwise an amino acid difference could result in an ineffective drug or vaccine. With recent advances in mass spectrometers, the mass accuracy could be improved to around 1 ppm. For a fragment ion of mass 1,000 Da, this means the measurement error is smaller than 0.001 Da. Such high-resolution data allow accurate de novo peptide sequencing.

On the other hand, most existing de novo sequencing tools were developed back in the days when the mass error was greater than 100 ppm. It is not trivial for those tools to take full advantage of the higher precision provided by the latest generation of mass spectrometers. For spectrum graph-based methods[6–8], a higher precision means that less nodes are merged and the generated spectrum graph has more vertices, which directly leads to a higher computational complexity. Similarly, the complexity of dynamic programming-based methods such as PEAKS[9] and Novor[4] are sensitive with respect to the spectrum resolution. For instance, the computational complexity of the dynamic programming proposed by ref. [10] is inversely proportional to the cube of the finest calibration of the mass spectrometer. Furthermore, current existing neural network-based de novo sequencing models such DeepNovo[5] and SMSNet[11] need to first discretize a spectrum to an intensity vector (for example, DeepNovo uses a vector with a length of 150,000 to represent a spectrum when the spectrum resolution parameter is set to 50). Long intensity vectors require considerable memory and CPU time to create and process. In fact, the GPU is often not fully utilized in the original implementation of DeepNovo as the program needs to wait for the CPU to build and process such vectors. Both DeepNovo and SMSNet need to discretize spectra with a higher spectrum resolution parameter (R) to take advantage of the improved precision offered by higher-resolution spectra. The computation and memory demands grow linearly with respect to R for these models (that is, complexity of $O(R)$).

To fully benefit from the high precision that the latest mass spectrometers offer, we present PointNovo, a neural network-based de novo peptide sequencing tool that does not vectorize the mass spectrum. PointNovo is ready to be applied to higher-resolution data that may be generated in the future, without any added complexity. Moreover, our experiment results show that PointNovo also considerably outperforms previous state-of-the-art methods. PointNovo achieves this by directly representing a spectrum as a set of m/z values and intensity pairs, and through the use of an order-invariant network structure[12] to learn from the data of such a structure. Figure 1a demonstrates how PointNovo represents input spectrum and the extraction of features. More details about the model can be found in the Methods.

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada. [2]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. [3]Bioinformatics Solutions, Waterloo, Ontario, Canada. ✉e-mail: bshan@bioinfor.com; ali.ghodsi@uwaterloo.ca

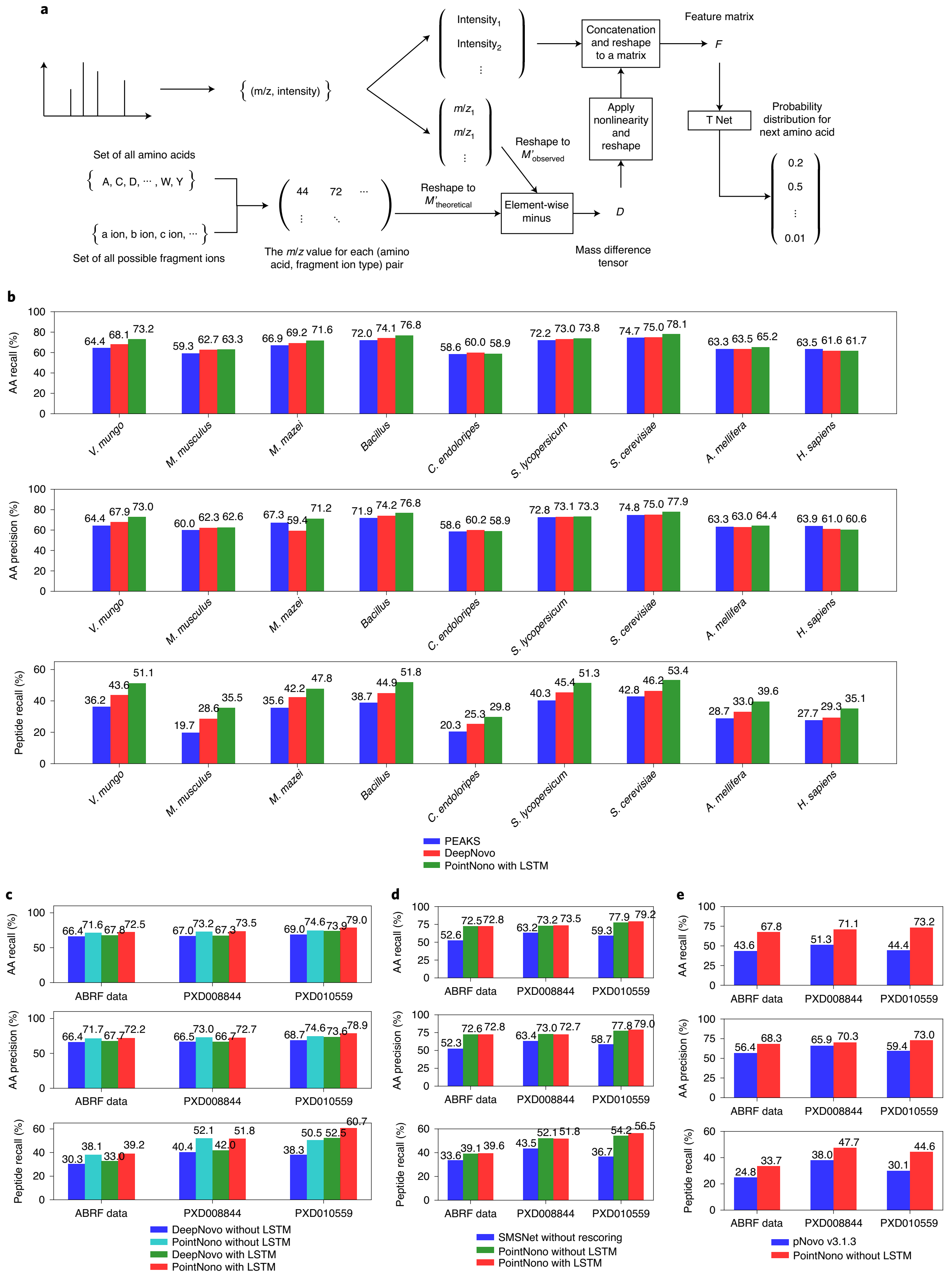


Fig. 1 | a, Spectrum representation and feature extraction in PointNovo. **b**, Accuracy comparisons between the nine species data published by DeepNovo. **c**, Accuracy comparisons between PointNovo and DeepNovo on three high-resolution MS/MS datasets. **d**, Accuracy comparisons between PointNovo and SMSNet on three high-resolution MS/MS datasets. **e**, Accuracy comparisons between PointNovo and pNovo3 on three high-resolution MS/MS datasets.

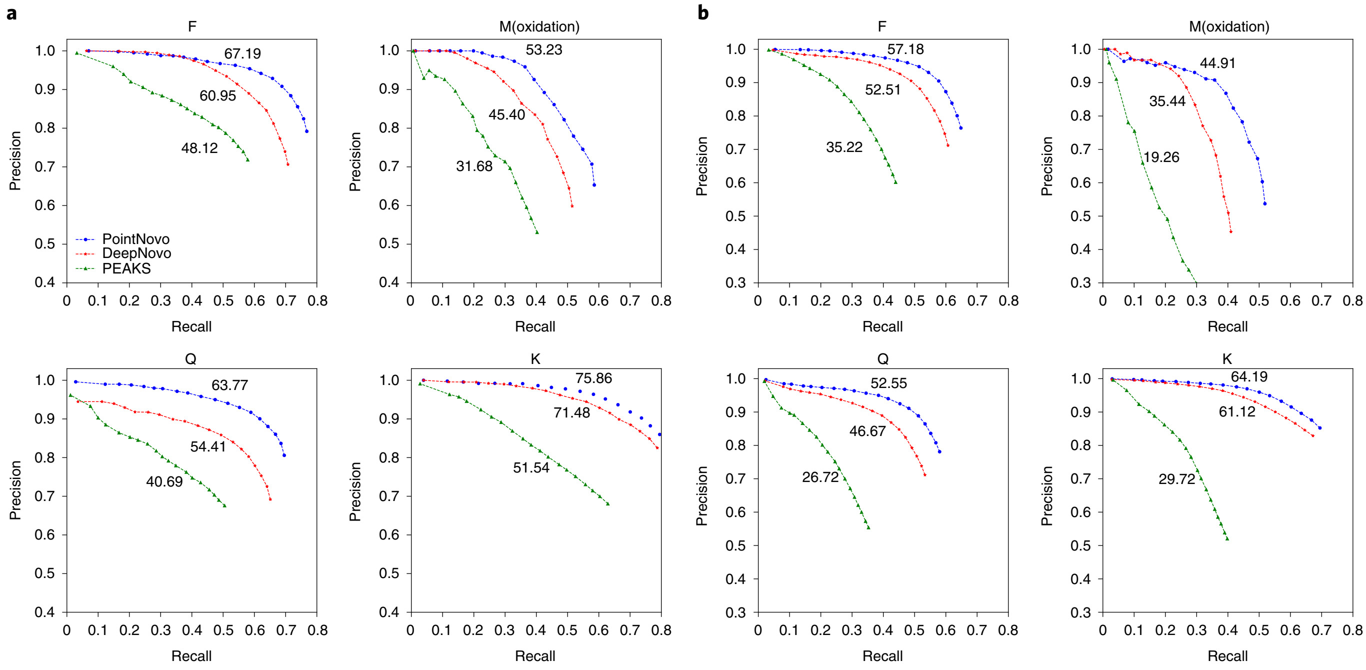


Fig. 2 | a, Precision–recall curves for amino acids pairs with similar mass on test spectra from PXD008844. **b**, Precision–recall curves for amino acids pairs with similar mass on test spectra from PXD010559. The values labelling the curves are average precision values.

Evaluation metric and datasets

For performance evaluation, we downloaded the nine species data used by the original publication of DeepNovo (MassIVE dataset identifier: MSV000081382) and applied our model to these data. We implement the same leave-one-out cross-validation scheme as described in ref. [5], that is, all except one of the nine datasets were used to train PointNovo and the trained model is tested on the remaining dataset. We used the same evaluation metric adopted by DeepNovo and Novor when calculating the amino acid precision, amino acid recall and peptide recall, that is, a predicted amino acid matching a real amino acid if their mass difference is smaller than 0.1 Da and if the prefix masses before them vary by less than 0.5 Da. For a fair comparison, we used PointNovo with a long short-term memory (LSTM) module[13], as DeepNovo includes an LSTM module by default. The test results and comparison with DeepNovo are shown in Fig. 1b. PointNovo outperforms DeepNovo consistently at peptide level by a large margin of 13.01–23.95%. We note here that DeepNovo reports a slightly lower amino acid recall but a higher peptide recall rate than PEAKS in the cross-species training for humans; similar results are also observed for PointNovo. We suggest that this is due to some peptides from the test set also appearing in the training set in the cross-species training scheme. The LSTM modules in PointNovo and DeepNovo will be trained to predict the sequences that existed in the training set. It might be a desired property in some applications (for example, training an allele-aware de novo sequencing model for human leucocyte antigen (HLA) peptides[14]), but it is not the best practice for evaluating machine learning models. To better compare our proposed model with DeepNovo, SMSNet and pNovo[15], we collected three high-resolution tandem mass spectrometry (MS/MS) datasets provided by different laboratories (HeLa samples from the Association of Biomolecular Resource Facilities (ABRF), ProteomeXchange dataset with the identifiers of PXD008844[16] and PXD010559[17]). We first ran a database search using PEAKS X on each of the three datasets. The post-translational modifications (PTMs) settings are included in the Methods. The peptide-spectrum matches (PSMs) identified at 1% false discovery rate in each dataset are split into training, validation and test sets at a ratio of 8:1:1. We ensured that no common peptide sequences were shared among the aforementioned sets during the split. Two PointNovo models (with and without LSTM) and two DeepNovo models (with and without LSTM) were then trained from scratch on the training set for each of the three high-resolution MS/MS datasets. The weights that show the best validation loss during training were saved as the trained model weights. Finally, trained models were evaluated on the test set. The amino acid level accuracy, amino acid level recall and peptide level recall on the test set are reported in Fig. 1c. PointNovo improves at peptide level recall by 15.05–23.32% when an LSTM module is included, whereas PointNovo outperforms DeepNovo by 25.61–31.94% when an LSTM module is not included. In a procedure similar to the above experiments, we also compared PointNovo with SMSNet (the results are shown in Fig. 1d). As PointNovo does not contain any post-processing, we applied SMSNet without rescoring in this comparison[11]. Due to a limitation of SMSNet, all PSMs that contain PTMs other than carbamidomethylation of C or oxidation of M are removed from our datasets. As a result, the training, validation and test sets are slightly different from previous experiments and that is the reason why accuracies of PointNovo reported in Fig. 1d are different from those reported in Fig. 1c. In cases without rescoring, we notice that SMSNet sometimes predicts exceptionally long sequences for spectra of poor quality. The existence of such long sequences undermines the amino acid level accuracy; the peptide level recall metric therefore shows a better performance comparison between the two models. Nevertheless, PointNovo outperforms SMSNet (without rescoring) at peptide level by over 17%. We would like to point out that the contribution of sequence-mask-search made by SMSNet is orthogonal to the improvement made by PointNovo. A similar post-processing could be applied to the output of PointNovo. In Fig. 1e we show the comparison results between PointNovo and pNovo. To the best of our knowledge, pNovo3 with spectral prediction[15] has not been released for users; we downloaded the most recent pNovo release (v.3.1.3) and compared it with PointNovo. As pNovo is distributed as pretrained software, we cannot adopt the same training procedure as the previous experiments as that would give PointNovo an unfair advantage. To make a fair comparison, we collected four other high-resolution MS/MS datasets: PXD008808[18], PXD011246[19], PXD012645[20] and PXD012979[21]. We trained PointNovo without an LSTM model on the identified PSMs of these four datasets and applied the trained model to the test sets of the ABRF, PXD008844 and PXD010559 datasets. In this experiment, we again exclude all PSMs that contain PTMs other than carbamidomethylation of C or oxidation of M from the training and test sets as we need to apply the same trained model to all three test sets. Figure 1e shows that our trained PointNovo without-LSTM model outperforms pNovo by more than 25.5% at peptide level. More interestingly, the PointNovo performance gap between Fig. 1d and Fig. 1e gives us an estimate of the generalizability of our proposed model. The metrics reported in Fig. 1d represent the performance in the best-case scenario, where the training spectra are acquired in the same experimental setting as the test spectra (for example, different fractions of the same sample). The results of Fig. 1e also represent the performance in the normal scenario, in which training spectra are collected from multiple experiments conducted by different laboratories. Above all, our results as shown in Fig. 1b–e demonstrate that PointNovo consistently outperforms DeepNovo, SMSNet (without rescoring) and pNovo on all three different test sets. Here we would like to again explain that although the results shown in Fig. 1c–e are from the same test datasets, these results cannot be merged as the three experiments are conducted in different settings (that is, different PTMs included, different training datasets) for the purpose of making a fair comparison.

Better discrimination of amino acids of similar masses

To further demonstrate that our proposed PointNovo model could take full advantage of the high-resolution data and better discriminate between amino acids pairs that have similar masses, we calculate the precision and recall for amino acid pairs F and M(Oxi) (the mass difference is smaller than 0.035 Da), Q and K. In this analysis, a predicted amino acid is considered as matching the ground-truth amino acid in the target sequence if (and only if) the amino acids are exactly the same and the prefix masses before them vary by less than 0.5 Da. Both DeepNovo and PointNovo are trained without the LSTM modules, as we want to compare their ability of learning from spectra and not their ability to remember the sequence patterns. The precision–recall curves for two datasets are shown in Fig. 2a,b. PointNovo improves the average precision for all four amino acids, which are notorious for being hard to discriminate. Specifically, for amino acid Q and M(Oxi), we observe a significant improvement of more than 15%. Extended Data Fig. 1 shows Venn diagrams of the peptide sets identified by PEAKS X (database search), predicted by PointNovo and DeepNovo on the ABRF, PXD008844 and PXD010559 datasets. Following the practice introduced in ref. [22], we filtered the de novo peptides on the basis of their peptide scores given by the models. Peptide score cut-offs are selected so that the amino acid accuracy is 90%. The intersection between two sets represents peptides of the exact same amino acid sequence. As can be seen from Extended Data Fig. 1, PointNovo's prediction always covers more peptides identified by PEAKS X than DeepNovo.

Finally, to show that PointNovo can potentially benefit from the improved precision of higher-resolution spectra generated in the future, we simulate low-resolution spectra of the ABRF, PXD008844 and PXD010559 datasets. PointNovo's performance on these spectra is reported in Extended Data Fig. 2. The low-resolution spectra are generated by adding random parts-per-million errors ($\epsilon \approx U(-10{,}10)$) to the m/z value of every peak in original spectra datasets. PointNovo is then trained and tested on the jittered training and test spectra. The comparison results in Extended Data Fig. 2 demonstrate that we could indeed expect better performance on higher-resolution spectra with PointNovo.

The above results demonstrate that PointNovo outperforms previous state-of-the-art de novo peptide sequencing tools by a significant margin and could better discriminate between similar amino acids pairs. Also, unlike previous neural network-based de novo peptide sequencing tools, PointNovo does not include any spectrum vectorization. It is thus ready to be applied to the more precise mass spectrometry data generated in the future.

Methods

Spectrum representation. In DeepNovo and SMSNet, spectra are represented as intensity vectors, where each index of the vectors represents a small m/z bin and the value represents the sum of intensities of all peaks that fall into that bin. This representation of spectra naturally has the problem of accuracy and speed/memory trade-off. In PointNovo, we propose to directly represent a spectrum as a set of (m/z, intensity) pairs. For each spectrum we select the top N most intense peaks (by default $N = 1{,}000$) and represent the spectrum as $\{(m/z_i, I_i)\}_{i=1}^{N}$. Further, we denote $\mathbf{M}_{\text{observed}} = (m/z_1, \cdots, m/z_N)$ as the observed m/z vector and $\mathbf{I} = (I_1, \cdots, I_N)$.

Feature extraction. Aside from the 20 amino acid residues and their PTMs, we also include three special tokens denoted start, end and padding in our model's vocabulary set. We denote the number of tokens (including amino acid residues and PTMs) as v and the number of ion types as k. PointNovo uses the twelve types of ion ($k = 12$): b, y, a, b(2+), y(2+), a(2+), b-H_2O, y-H_2O, a-H_2O, b-NH_3, y-NH_3 and a-NH_3. At each prediction step, we compute the theoretical m/z values for each token and ion-type pair. The result is a matrix of shape (v,k), which is denoted $\mathbf{M}_{\text{theoretical}}$. We next expand the dimension of $\mathbf{M}_{\text{observed}}$ to make it a three-dimensional tensor of shape (N,1,1), and then repeat $\mathbf{M}_{\text{observed}}$ on the second and third dimensions v times and k times, respectively. The result is denoted as $\mathbf{M}'_{\text{observed}}$, which is a tensor of shape (N,v,k). Similarly, we expand $\mathbf{M}_{\text{theoretical}}$ to the shape of (1,v,k), repeat it on the first dimension N times and denote the result as $\mathbf{M}'_{\text{theoretical}}$. We can then compute the m/z difference tensor ($\mathbf{D}$), in which each element represents the difference between the m/z value for an observed peak and the theoretical m/z for a token and ion-type pair.

$$\mathbf{D} = \mathbf{M}'_{\text{observed}} - \mathbf{M}'_{\text{theoretical}}$$

It is worth noting that the above equation could be computed efficiently by using the broadcast behaviour in popular deep learning frameworks such as Tensorflow[23] and PyTorch[24].

Based on the expert knowledge of de novo peptide sequencing[9], we design an activation function σ:

$$\sigma(\mathbf{D}) = \exp\{-|\mathbf{D}| \times c\}$$

Here the exponential and absolute operations are all element-wise operations. The intuition for σ is that an observed peak could only be considered matching a theoretical m/z location if the absolute m/z difference is small. For example, if we set $c = 100$, then an observed peak that is 0.02 Da away from a theoretical location would generate a signal of $e^{-2} \approx 0.135$, which is only one-seventh of the signal of a perfect match. In our experiments we tried setting c to a trainable parameter and updating it through backpropagation. It shows similar performance with setting $c = 100$. We set $c = 100$ in all of the experiments reported in this manuscript for better model interpretability; however, setting c to a learnable parameter would require less past knowledge about the resolution of training spectra and might be preferable in certain cases.

We next reshape the N by v by k tensor $\sigma(\mathbf{D})$ to a matrix $\sigma(\mathbf{D})'$ of shape N by vk, reshape $\mathbf{I}$ to a N by 1 vector $\mathbf{I}'$. Finally, the feature matrix $\mathbf{F}$ used for predicting the next amino acid is simply the concatenation of $\sigma(\mathbf{D})'$ and $\mathbf{I}'$:

$$\mathbf{F} = \sigma(\mathbf{D})' \oplus \mathbf{I}'$$

Here $\oplus$ represents concatenation along the second dimension. The output $\mathbf{F}$ is a matrix of shape N by $vk + 1$.

A spectrum is a set of (m/z, intensity) pairs, which means the order of peaks should be irrelevant. The prediction network should therefore have an order-invariant property with respect to the first dimension of $\mathbf{F}$. To the best of our knowledge, T Net (Structure is shown in Extended Data Fig. 3) is the first model designed for this kind of order-invariant data. It demonstrated state-of-the-art performance on the point-cloud classification task[12]. We therefore apply T Net to learn from the feature matrix $\mathbf{F}$. The global max pooling operation in T Net guarantees that the output would not change for any row permutations of $\mathbf{F}$.

We experimented using parts per million m/z difference instead of absolute difference in matrix $\mathbf{F}$. The experimental result shown in Extended Data Fig. 4 suggests that on these Fusion Lumos datasets that we collected, the parts-per-million difference method and absolute difference method gives very similar results.

Initial state for LSTM. The LSTM module is an optional component in PointNovo. In some applications (for example, training an allele-aware de novo sequencing model for HLA peptides) it might be desirable for the model to remember some peptide sequence patterns. In such cases we can include an LSTM module in PointNovo. The full model structure of PointNovo (both with and without an LSTM module) is shown in Extended Data Fig. 5.

We need to initialize the hidden states of LSTM with information from the original spectrum. Inspired by the success of positional embedding introduced by Vaswani and colleagues[25], we choose to embed each peak into a vector. The input spectrum is first discretized at 0.1 Da resolution. When applied to the without-LSTM case, the discretization step is not needed.

We next create a sinusoidal m/z positional embedding matrix $\mathbf{E}$ in the way suggested by[25]:

$$\mathbf{E}_{(\mathrm{loc},2j-1)} = \sin\left(\mathrm{loc}/10,000^{\frac{2j-2}{512}}\right)$$

$$\mathbf{E}_{(\mathrm{loc},2j)} = \cos\left(\mathrm{loc}/10,000^{\frac{2j-2}{512}}\right)$$

$$\forall j \in \{1,\ 2,\ \ldots,\ 256\}$$

Here loc represents the m/z index after discretization. We use $\mathbf{E}_l$ to denote the lth row vector $\mathbf{E}$. The sinusoidal embedding has a desired property that for any distance d, $\mathbf{E}_{\mathrm{loc}+d}$ could be represented as a linear function of $\mathbf{E}_{\mathrm{loc}}$. This property is important because in mass spectra the m/z differences between observed peaks contains useful information that indicates which amino acids possibly exist. For an input spectrum $\{(m/z_i, I_i)\}_{i=1}^{N}$, we denote loc_i to represent the index of m/z_i after discretization and we use $I_i\mathbf{E}_{\mathrm{loc}_i}$ as the vector representation of the ith peak. A spectrum representation vector $\mathbf{S}$ can then be generated by taking the summation of the vector representations of all peaks:

$$\mathbf{S} = \sum_{i=1}^{N} I_i \mathbf{E}_{\mathrm{loc}_i}$$

We multiplied the intensities with the embedded peak vectors because we think the effect of a single peak, in the representation of a spectrum, should be proportional to its intensity. Finally, the hidden states of the LSTM module are initialized to $\mathbf{S}$.

Post-translational modifications settings. For the ABRF dataset we set carbamidomethylation of C as fixed modification, oxdiation of M and deamidation of NQ as variable modification. For the PXD008844 dataset we set carbamidomethylation of C as fixed modification, oxdiation of M as variable modification, whereas for PXD010559 we set carbamidomethylation of C as fixed modification and oxdiation of M, deamidation of NQ and phosphorylation of STY are set as variable modification.

Training and searching. As suggested in ref. [22], we used focal loss[26] instead of cross-entropy loss when training the model. We trained PointNovo with the Adam algorithm[27] and an initial learning rate of 10^{-3}. After every 300 training steps, the loss on validation set is computed. If the validation loss has not achieved a new low in the recent ten evaluations then the learning rate would be dropped by half. We applied the beam search algorithm for the searching part. Similar to DeepNovo, PointNovo uses knapsack algorithm to reduce the search space.

Applying trained PointNovo model on a dataset with different peptide patterns. In the without-LSTM mode, PointNovo extract most information directly from the MS/MS spectra. It is thus possible (although not the best practice) to apply a trained PointNovo model to a dataset with peptides of totally different sequence patterns. We downloaded a HLA peptide dataset to demonstrate this[28]. A PointNovo without-LSTM model is trained on the identified PSMs of patient Mel 15 data and then applied to patient Mel 16. The two patients do not share any common HLA alleles, which means the sequence patterns of the HLA peptides are different. Extended Data Fig. 6 demonstrated that although the PointNovo model is trained on a relatively small dataset with different sequence patterns from the test set, it can still achieve a comparable performance with PEAKS de novo. Moreover, we downloaded another HeLa sample dataset processed by multiple different enzymes[29] and report the cross-enzyme testing performance in Extended Data Fig. 7. We need to point out here that although the PointNovo without LSTM model does not remember sequence patterns, it would still learn the overall amino acid distribution from training peptides. The best practice in application is therefore to train a separate model for each enzyme.

Speed of PointNovo. On an RTX 2080 Ti GPU, a training step (batch size 16) takes around 0.4 s; for inference (that is, de novo peptide sequencing), PointNovo (with LSTM) can process around 20 spectra per second. Without LSTM, PointNovo can perform de novo peptide sequencing on more than 70 spectra in one second.

Data availability

The source data for all experiments reported by this paper are accessible through the following link: https://zenodo.org/record/3998873 (ref. [30]).

Code availability

The source code of PointNovo is in this github repo: https://github.com/volpato30/PointNovo (ref. [31]).

Received: 15 December 2019; Accepted: 26 January 2021;
Published online: 18 March 2021

References

1. Tran, N. H. et al. Complete de novo assembly of monoclonal antibody sequences. *Sci. Rep.* **6**, 1–10 (2016).
2. Faridi, P. et al. A subset of HLA-I peptides are not genomically templated: evidence for *cis*- and trans-spliced peptide ligands. *Sci Immunol.* **3**, eaar3947 (2018).
3. Laumont, C. M. et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).
4. Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass. Spectrom.* **26**, 1885–1894 (2015).
5. Tran, H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl Acad. Sci. USA* **114**, 8247–8252 (2017).
6. Dancík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6**, 327–342 (1999).
7. Chen, T., Kao, M. Y., Tepel, M., Rush, J. & Church, G. M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **8**, 325–337 (2001).
8. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
9. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
10. Ma, B., Zhang, K. & Liang, C. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *J. Comput. Syst. Sci.* **70**, 418–430 (2005).
11. Karunratanakul, K., Tang, H., Speicher, D. W., Chuangsuwanich, E. & Sriswasdi, S. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Mol. Cell. Proteom.* **18**, 2478–2491 (2019).
12. Qi C. R., Su H., Mo K. & Guibas L. J. PointNet: deep learning on point sets for 3D classification and segmentation. In *Proc. IEEE Conference On Computer Vision and Pattern Recognition* 652–660 (IEEE, 2016).
13. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
14. Tran N. H. et al. Identifying neoantigens for cancer vaccines by personalized deep learning of individual immunopeptidomes. *Nat. Mach. Intell.* **2**, 764–771 (2019).
15. Yang, H., Chi, H., Zeng, W., Zhou, W. & He, S. pNovo3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics.* **35**, i183–i190 (2019).
16. Zhu, Y. et al. Spatially resolved proteome mapping of laser capture microdissected tissue with automated sample transfer to nanodroplets. *Mol. Cell. Proteom.* **17**, 1843–1874 (2018).
17. Shears, M. J. et al. Proteomic analysis of plasmodium merosomes: the link between liver and blood stages in malaria. *J Proteome Res.* **18**, 3404–3418 (2019).
18. Sobolesky, P. et al. Proteomic analysis of non-depleted serum proteins from bottlenose dolphins uncovers a high vanin-1 phenotype. *Sci. Rep.* **6**, 33879 (2016).
19. Benitez-Amaro, A. et al. Molecular basis for the protective effects of low-density lipoprotein receptor-related protein 1 (LRP1)-derived peptides against LDL aggregation. *Biochim. Biophys. Acta Biomembr.* **1861**, 1302–1316 (2019).
20. Sim, S. Y. et al. In-depth proteomic analysis of human bronchoalveolar lavage fluid toward the biomarker discovery for lung cancers. *Proteom. Clin. Appl.* **13**, 1900028 (2019).
21. Haythorne, E. et al. Diabetes causes marked inhibition of mitochondrial metabolism in pancreatic β-cells. *Nat. Commun.* **10**, 2474 (2019).
22. Tran, N. H. et al. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63–66 (2019).
23. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* 265–283 (USENIX, 2016).
24. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 8026–8037 (NeurIPS, 2019).

25. Vaswani, A. et al. Attention Is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (NeurIPS, 2017).
26. Lin, T., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. In *Proc. IEEE International Conference on Computer Vision* 2980–2988 (IEEE, 2017).
27. Kingma, D. P. & Ba, L. J. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations* (ICLR, 2015).
28. Bassani-Sternberg, Michal et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun*. **7**, 13404 (2016).
29. Bekker-Jensen, DorteB. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
30. Qiao, R. *Source Data for PointNovo* (Zenodo, 2020); https://doi.org/10.5281/zenodo.3998873
31. Ma, Z. *Volpato30/PointNovo: First Release (Version v0.0.1)* (Zenodo, 2020); https://doi.org/10.5281/zenodo.3960823

Acknowledgements

We thank Z. Ma for discussions on order-invariant networks. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) (funding reference no. RGPIN-2019-04824), China's National Key Research and Development Program under grant no. 2018YFB1003202 and NSERC grant no. OGP0046506. This work was performed while R.Q. was visiting Bioinformatics Solutions.

Author contributions

R.Q. and A.G. conceived the research idea and the prototype of the model. R.Q. implemented the proposed algorithm and analysed the data. N.H.T, M.L, B.S, X.C. and L.X. contributed to model design and data analysis. N.H.T, M.L., A.G. and R.Q. wrote the manuscript. A.G. and M.L. supervised the research project.

Competing interests

The authors have filed a patent application for the PointNovo model in the USPTO Provisional Application (US Provisional Patent Application no. 62/833,959) by Bioinformatics Solutions, Waterloo, Canada. The authors are named inventors in the patent application. L.X., X.C. and B.S. are employees of Bioinformatics Solutions.

Additional information

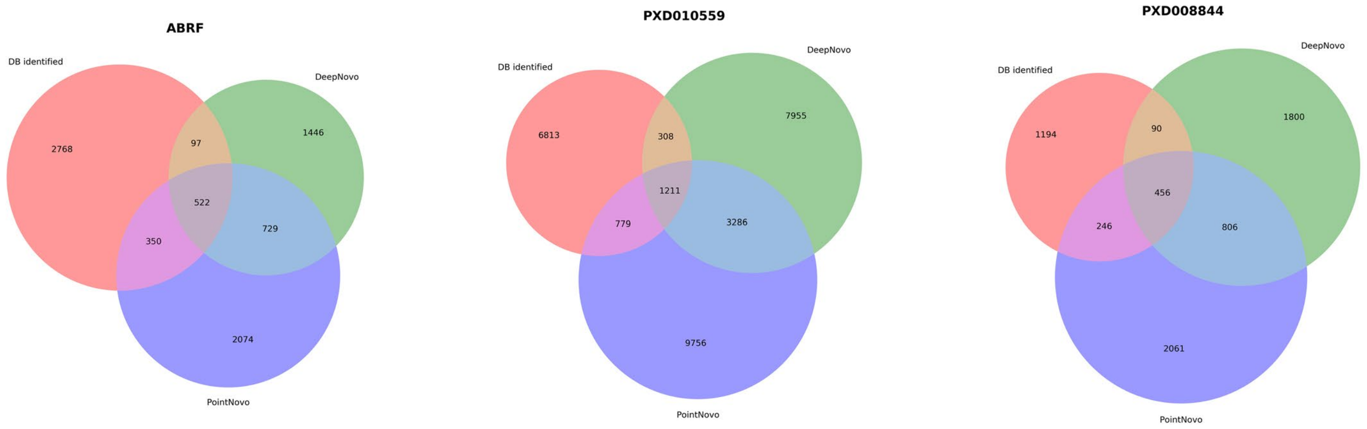
Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00304-3.

Correspondence and requests for materials should be addressed to B.S. or A.G.

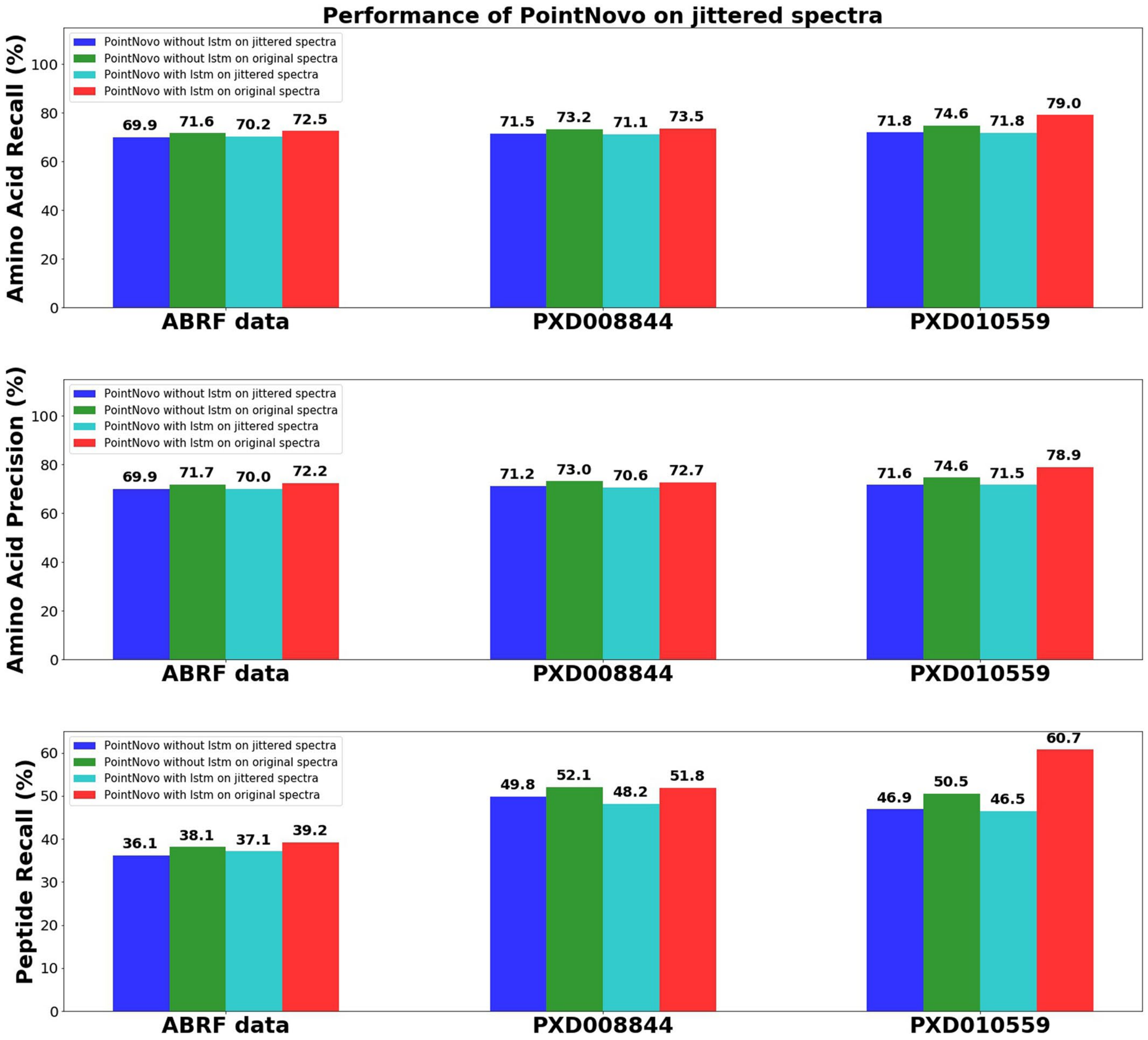
Peer review information *Nature Machine Intelligence t*hanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

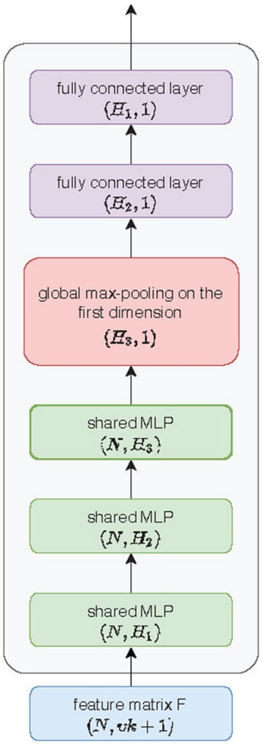
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



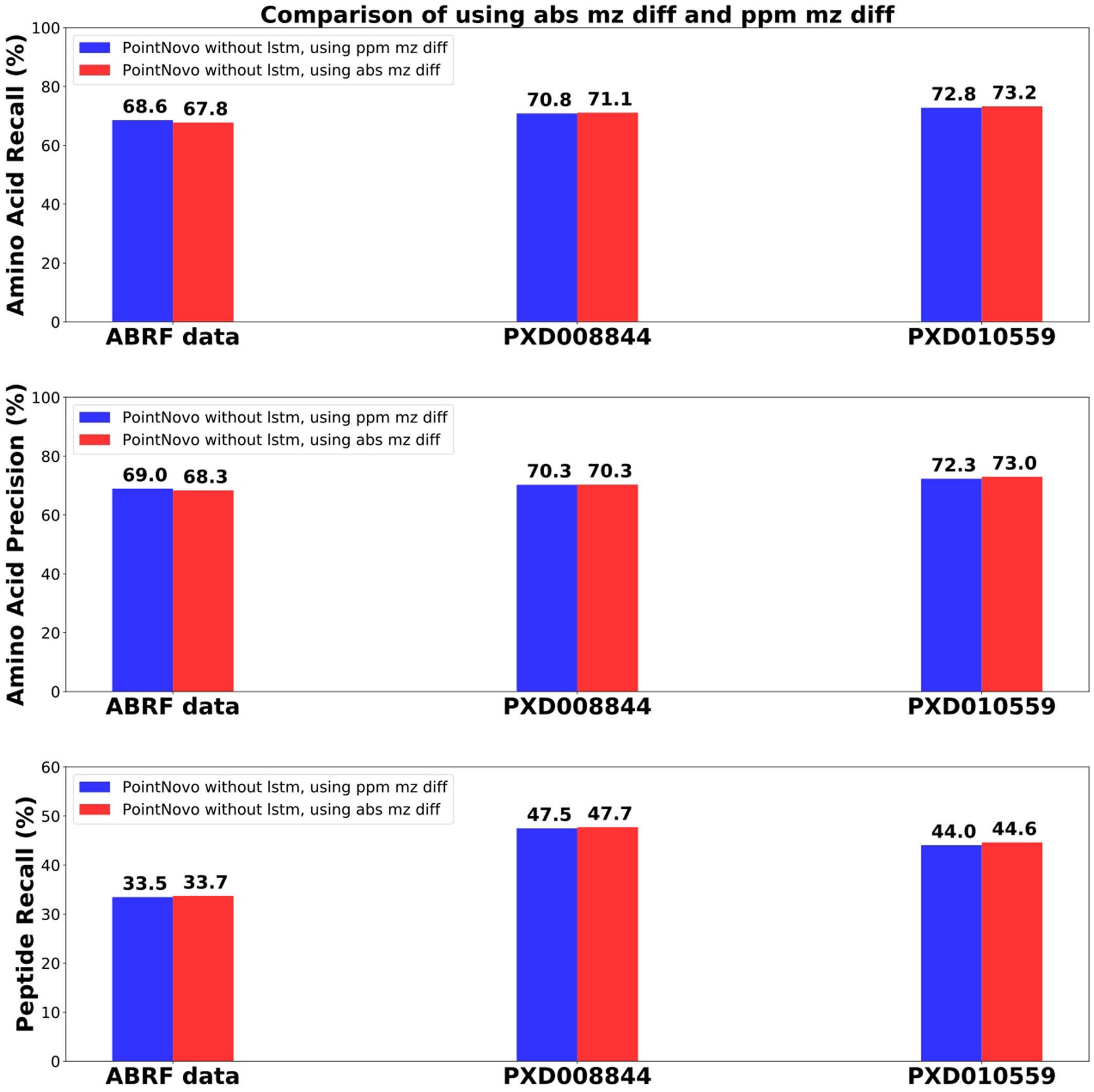
Extended Data Fig. 1 | Set of peptides predicted by PointNovo and DeepNovo, comparing with the set of peptides identified by PEAKS DB. Set of peptides predicted by PointNovo and DeepNovo, comparing with the set of peptides identified by PEAKS DB. Both DeepNovo and PointNovo are trained without the LSTM modules. Peptide score cutoff is applied to the results given by PointNovo and DeepNovo. We select the cutoff scores so that the amino acid accuracy of the remaining predicted peptides is 90%. Here, the overlap between two sets represents the peptides that are exactly the same (that is same amino acid residue sequence). Thus, the peptide recall is different from the number reported in Fig. 1, where a predicted amino acid residue is considered to be correct if the mass difference with the ground truth is smaller than 0.1 Da.



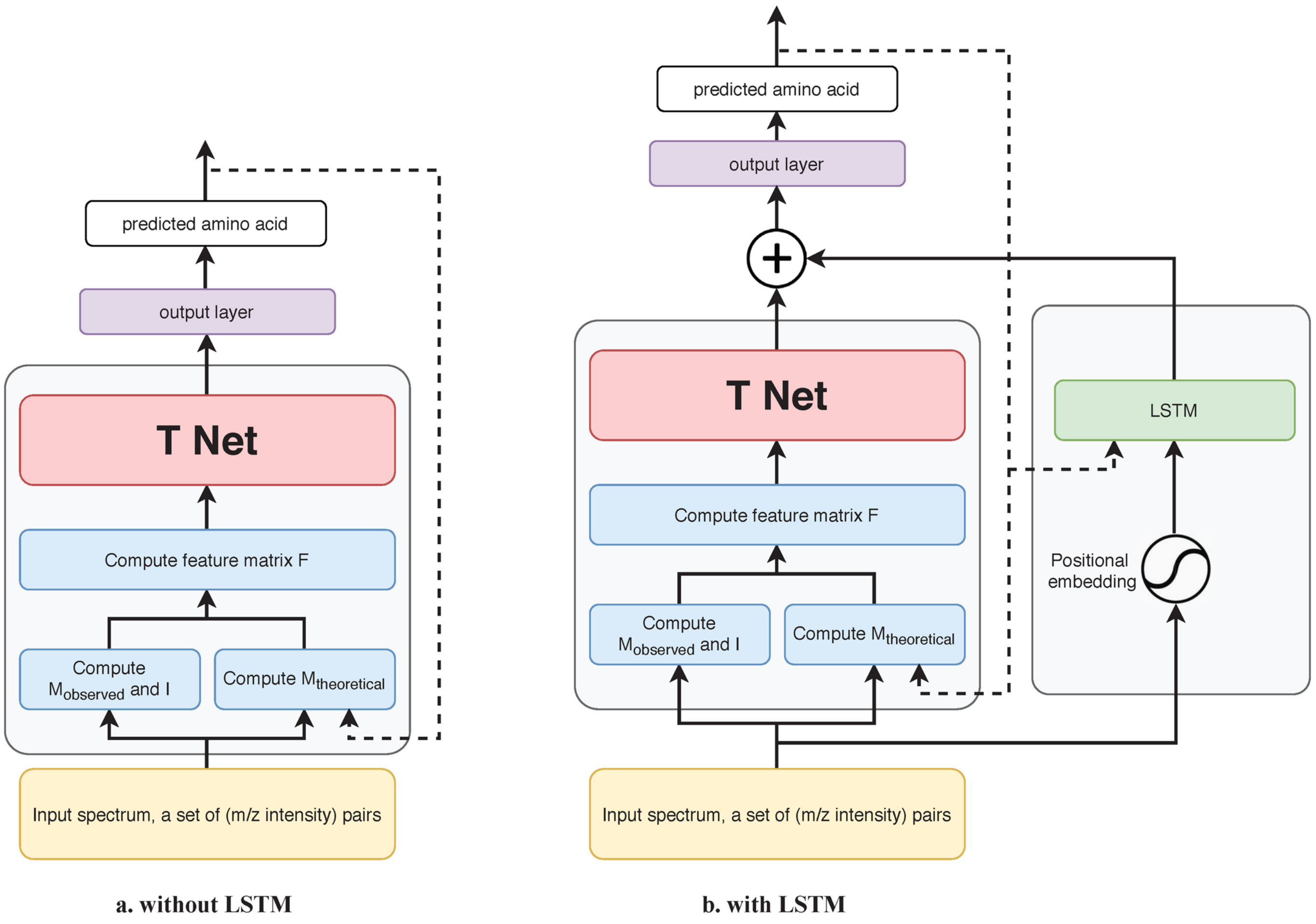
Extended Data Fig. 2 | Performance of PointNovo on jittered spectra. Performance of PointNovo on jittered spectra. To jitter the spectra, we add uniformly distributed random ppm errors to the m/z value of every peak in the original datasets. These jittered spectra could be considered as spectra of lower resolution.



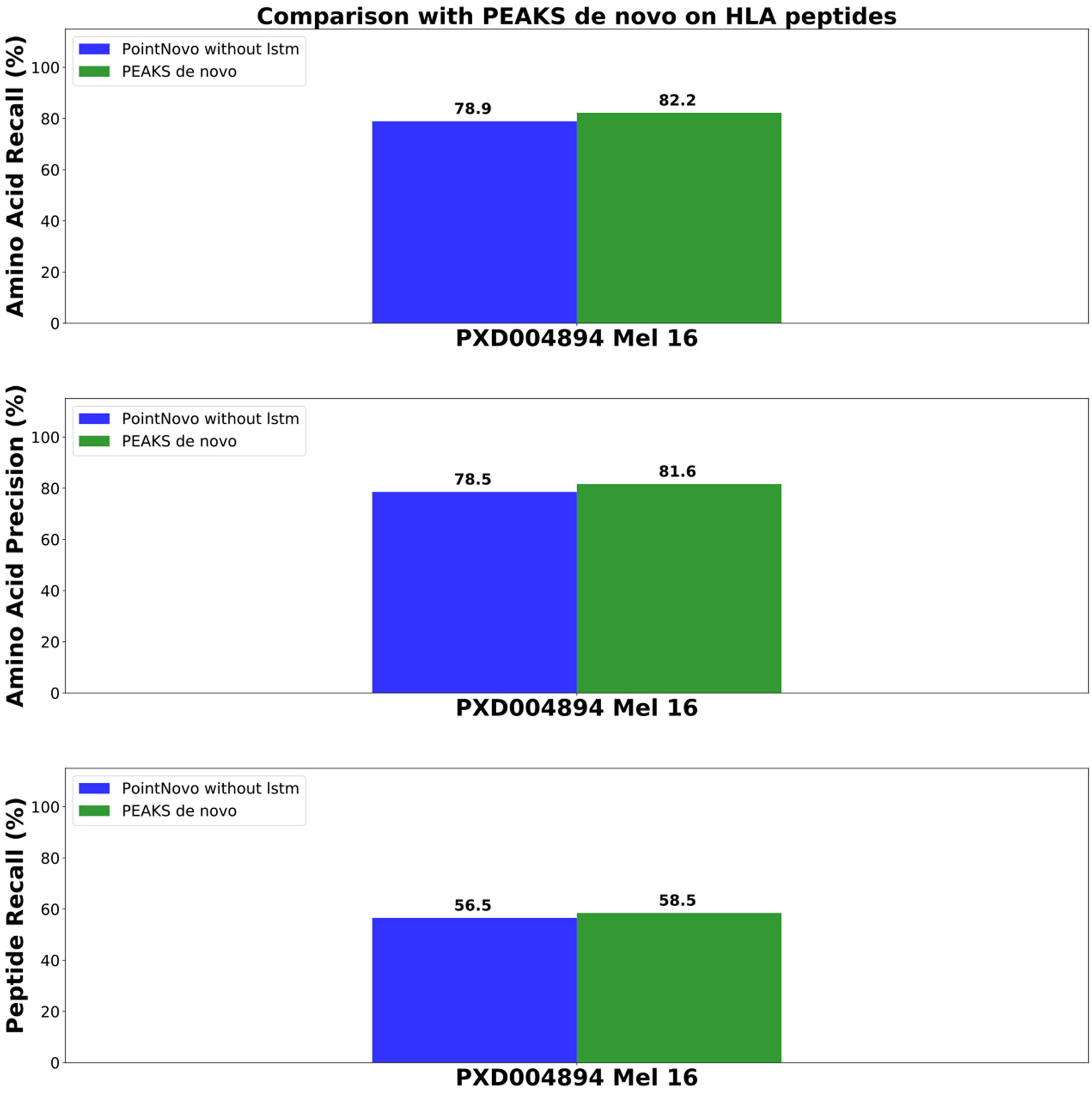
Extended Data Fig. 3 | Structure of T Net. Structure of T Net. The output shape of each layer is annotated below each block. Here *N* denotes the number of data points. *v* and *k* are defined in the feature extraction section of online method. H_i represent the number of hidden neurons in each hidden layer, which are hyper parameters that can be turned by the users.



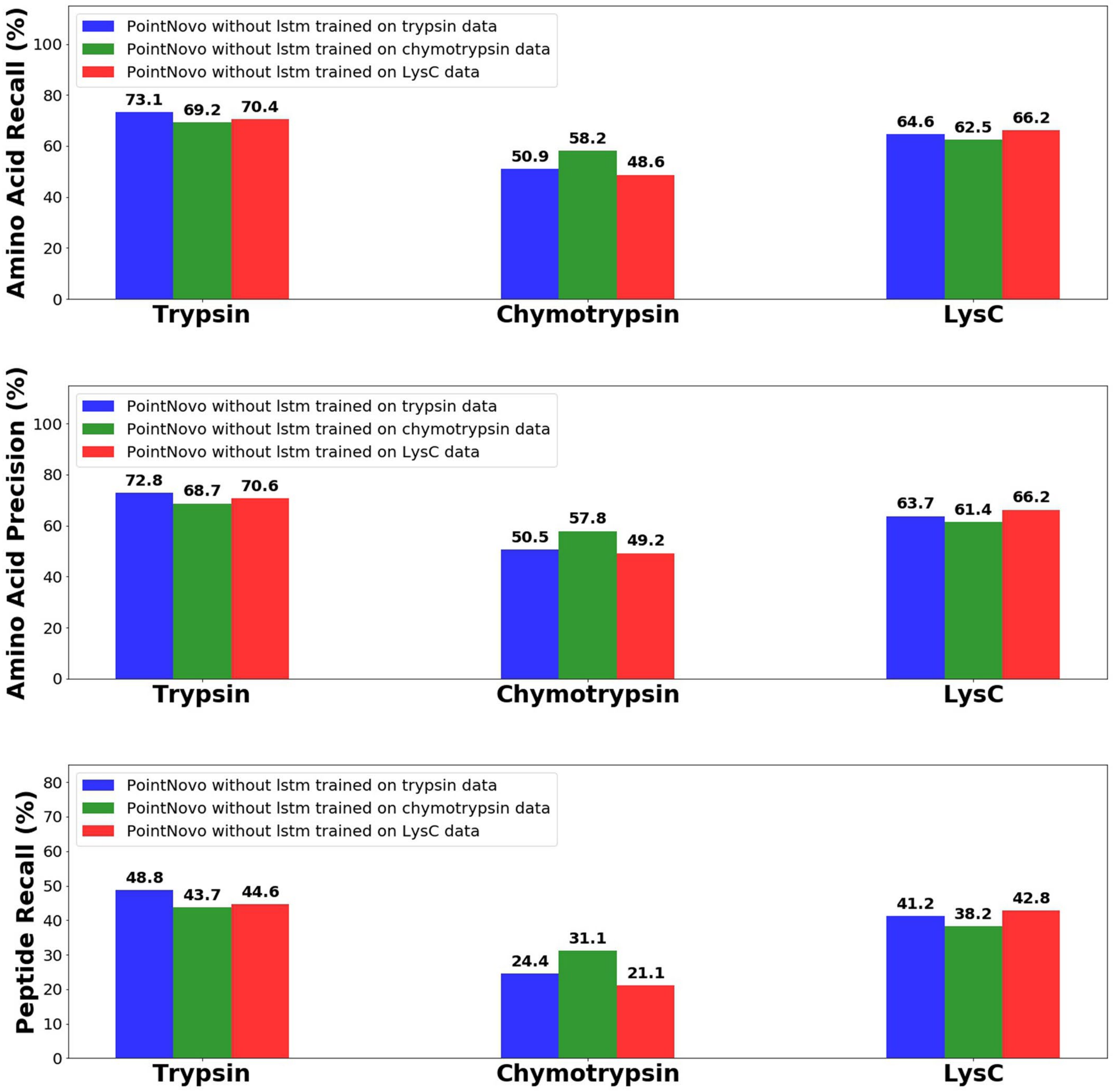
Extended Data Fig. 4 | Comparison of using absolute m/z diff and ppm m/z diff. Here the PointNovo models are trained on the combination of 4 datasets: PXD008808, PXD011246, PXD012645 and PXD012979.


predicted amino acid
output layer
T Net
Compute feature matrix F
Compute $M_{observed}$ and I
Compute $M_{theoretical}$
Input spectrum, a set of (m/z intensity) pairs
a. without LSTM
predicted amino acid
output layer
+
T Net
Compute feature matrix F
Compute $M_{observed}$ and I
Compute $M_{theoretical}$
LSTM
Positional embedding
Input spectrum, a set of (m/z intensity) pairs
b. with LSTM

Extended Data Fig. 5 | Structure of PointNovo. Structure of PointNovo. (a) PointNovo without LSTM. (b) PointNovo with LSTM.



Extended Data Fig. 6 | Comparison with PEAKS de novo on patient Mel 16 data. The PointNovo model here is trained on Mel 15 data, which has different peptide sequence pattern comparing with Mel 16 data.



Extended Data Fig. 7 | Cross-enzyme performance of PointNovo without LSTM model on PXD004452 data. PXD004452 dataset contains Hela samples digested by different enzyme. For each enzyme, we first ran database search peptide sequencing. The identified PSMs at 1% FDR are then split to training, validation and test set according to the ratio of 8:1:1. Separate PointNovo without LSTM models are trained for each enzyme and the cross-enzyme performance on test set is reported here.