De novo peptide sequencing by deep learning

Ngoc Hieu Tran[a,1], Xianglilan Zhang[a,b,1], Lei Xin[c], Baozhen Shan[c], and Ming Li[a,2]

[a]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada; [b]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China; and [c]Bioinformatics Solutions Inc., Waterloo, ON N2L 6J2, Canada

Edited by John R. Yates III, Scripps Research Institute, La Jolla, CA, and accepted by Editorial Board Member David Baker June 26, 2017 (received for review April 6, 2017)

De novo peptide sequencing from tandem MS data is the key technology in proteomics for the characterization of proteins, especially for new sequences, such as mAbs. In this study, we propose a deep neural network model, DeepNovo, for de novo peptide sequencing. DeepNovo architecture combines recent advances in convolutional neural networks and recurrent neural networks to learn features of tandem mass spectra, fragment ions, and sequence patterns of peptides. The networks are further integrated with local dynamic programming to solve the complex optimization task of de novo sequencing. We evaluated the method on a wide variety of species and found that DeepNovo considerably outperformed state of the art methods, achieving 7.7–22.9% higher accuracy at the amino acid level and 38.1–64.0% higher accuracy at the peptide level. We further used DeepNovo to automatically reconstruct the complete sequences of antibody light and heavy chains of mouse, achieving 97.5–100% coverage and 97.2–99.5% accuracy, without assisting databases. Moreover, DeepNovo is retrainable to adapt to any sources of data and provides a complete end-to-end training and prediction solution to the de novo sequencing problem. Not only does our study extend the deep learning revolution to a new field, but it also shows an innovative approach in solving optimization problems by using deep learning and dynamic programming.

deep learning | MS | de novo sequencing

Proteomics research focuses on large-scale studies to characterize the proteome, the entire set of proteins, in a living organism (1–5). In proteomics, de novo peptide sequencing from tandem MS data plays the key role in the characterization of novel protein sequences. This field has been actively studied over the past 20 y, and many de novo sequencing tools have been proposed, such as PepNovo, PEAKS, NovoHMM, MSNovo, pNovo, UniNovo, and Novor among others (6–19). The recent "gold rush" into mAbs has undoubtedly elevated the application of de novo sequencing to a new horizon (20–23). However, computational challenges still remain, because MS/MS spectra contain much noise and ambiguity that require rigorous global optimization with various forms of dynamic programming that have been developed over the past decade (8–10, 12, 13, 15–19, 24).

In this study, we introduce neural networks and deep learning to de novo peptide sequencing and achieve major breakthroughs on this well-studied problem. Deep learning has recently brought about a revolution in many research fields (25), repeatedly breaking state of the art records in image processing (26, 27), speech recognition (28), and natural language processing (29). It now forms the core of the artificial intelligence platforms of several technology giants, such as Google, Facebook, and Microsoft, as well as many startups in the industry. Deep learning has also made its way into biological sciences (30) [for instance, in the field of genomics, where deep neural network models have been developed for predicting the effects of noncoding single-nucleotide variants (31), predicting protein DNA and RNA binding sites (32), protein contact map prediction (33), and MS imaging (34)]. The key aspect of deep learning is its ability to learn multiple levels of representation of high-dimensional data through its many layers of neurons. Furthermore, unlike traditional machine learning methods, those feature layers are not predesigned based on domain-specific knowledge, and hence, they have more flexibility to discover complex structures of the data.

The task of de novo peptide sequencing is to reconstruct the amino acid sequence of a peptide given an MS/MS spectrum and the peptide mass. A spectrum can be represented as a histogram of intensity vs. mass (more precisely, m/z) of the ions acquired from the peptide fragmentation inside a mass spectrometer (Fig. 1*A*). The problem bears some similarity to the recently trending topic of "automatically generating a description for an image." In that research, a convolutional neural network (CNN; i.e., a type of feed-forward artificial neural network consisting of multiple layers of receptive fields) is used to encode or "understand" the image. Then, a long short-term memory (LSTM) recurrent neural network (RNN) (35) is used to decode or "describe" the content of the image (36, 37). The research is exciting, because it tries to connect image recognition and natural language processing by integrating two fundamental types of neural networks, CNN and LSTM.

In our de novo sequencing problem, the research is carried to the next extreme, where exactly 1 of 20^L amino acid sequences can be considered as the correct prediction (L is the peptide length, and 20 is the total number of amino acid letters). Another challenge is that peptide fragmentation generates multiple types of ions, including a, b, c, x, y, z, internal cleavage, and immonium ions (38). Depending on the fragmentation methods, different types of ions may have quite different intensity values (peak heights), and yet, the ion type information remains unknown from spectrum data. Furthermore, there are plenty of noise peaks mixing together with the real ions. Finally, the predicted amino acid sequence should have its total mass approximately equal to the given peptide mass. This challenge points to a complicated problem of pattern recognition and global optimization on noisy and incomplete data. The problem is typically handled by global dynamic programming (8–10, 12, 13, 15–19, 24), divide and conquer (11),

Significance

Our method, DeepNovo, introduces deep learning to de novo peptide sequencing from tandem MS data, the key technology for protein characterization in proteomics research. DeepNovo achieves major improvement of sequencing accuracy over state of the art methods and subsequently enables complete assembly of protein sequences without assisting databases. Our model is retrainable to adapt to any sources of data and provides a complete end-to-end training and prediction solution, an important feature given the growing massive amount of data. Our study also presents an innovative approach to combine deep learning and dynamic programming to solve optimization problems.

Author contributions: N.H.T., B.S., and M.L. designed research; N.H.T., X.Z., and L.X. performed research; N.H.T. contributed new reagents/analytic tools; N.H.T., X.Z., and L.X. analyzed data; and N.H.T., X.Z., and M.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.R.Y. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: DeepNovo is publicly available for non-commercial uses. The source code of DeepNovo is stored on GitHub (https://github.com/nh2tran/DeepNovo). All training and testing datasets, pretrained models, and source code of DeepNovo can also be downloaded from the FTP server of the MassIVE database via the following link: ftp://NancyzxII:DeepNovo2017@massive.ucsd.edu (user account: NancyzxII, password: DeepNovo2017).

[1]N.H.T. and X.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: mli@uwaterloo.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1705691114/-/DCSupplemental.

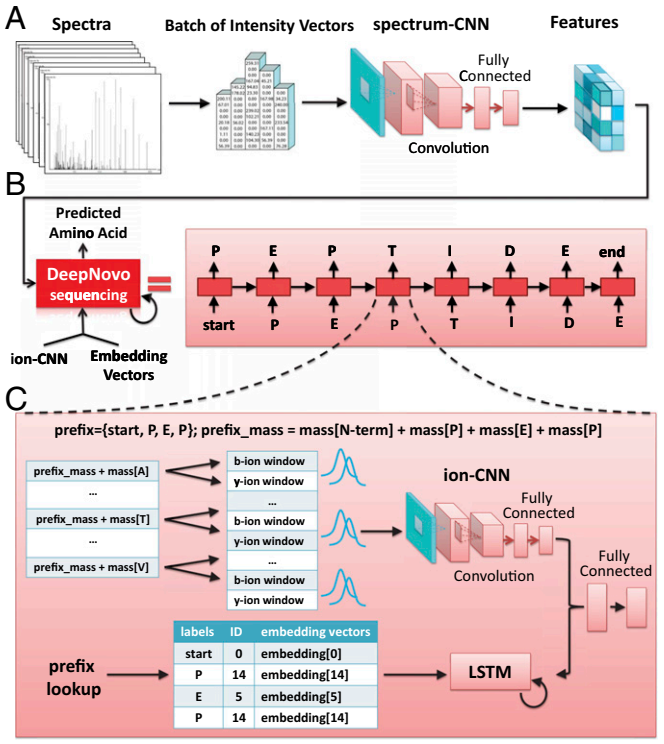

Fig. 1. The DeepNovo model for de novo peptide sequencing. (*A*) Spectra are processed by the CNN spectrum-CNN and then used to initialize the LSTM network. (*B*) DeepNovo sequences a peptide by predicting one amino acid at each iteration. Beginning with a special symbol start, the model predicts the next amino acid by conditioning on the input spectrum and the output of previous steps. The process stops if, in the current step, the model outputs the special symbol end. (*C*) Details of a sequencing step in DeepNovo. Two classification models, ion-CNN and LSTM, use the output of previous sequencing steps as a prefix to predict the next amino acid.

or integer linear programming (14). Hence, a naïve application of existing deep learning architectures does not work directly on this problem. Neural networks are often known to be good at simulating human brain capability, senses and intuition, rather than such precise optimization tasks. Thus, de novo peptide sequencing is a perfect case for us to explore the boundaries of deep learning.

In this study, we have succeeded in designing a deep learning system, DeepNovo, for de novo peptide sequencing. Our model features a sophisticated architecture of CNNs and LSTM networks together with local dynamic programming. DeepNovo has beaten the decade long-standing state of the art records of de novo sequencing algorithms by a large margin of 7.7–22.9% at the amino acid level and 38.1–64.0% at the peptide level. Similar to other deep learning-based models, DeepNovo takes advantage of high-performance computing graphics processing units (GPUs) and massive amounts of data to offer a complete end-to-end training and prediction solution. The CNN and LSTM networks of DeepNovo can be jointly trained from scratch given a set of annotated spectra obtained from spectral libraries or database search tools. This architecture allows us to train both general and specific models to adapt to any sources of data. We further used DeepNovo to automatically reconstruct the complete sequences of antibody light and heavy chains of mouse, an important downstream application of peptide sequencing. This application previously required de novo sequencing, database search, and homology search together to succeed (21) but now can be done by using DeepNovo alone.

Results

DeepNovo Model. The DeepNovo model is briefly illustrated in Fig. 1. The model takes a spectrum as input and tries to sequence the peptide by predicting one amino acid at each iteration (Fig. 1 *A* and *B*). The sequencing process begins with a special symbol "start." At each sequencing step, the model predicts the next amino acid by conditioning on the input spectrum and the output of previous steps. The process stops if, in the current step, the model outputs the special symbol "end." Backward sequencing is performed in a similar way to form the bidirectional sequencing, and the highest-scoring candidate is selected as the final prediction.

Details inside a sequencing step are described in Fig. 1*C*. DeepNovo incorporates two classification models that use the output of previous sequencing steps as a prefix to predict the next amino acid. In the first model, the prefix mass is first calculated as the sum of its amino acids' masses and the corresponding terminal. Then, each amino acid type is appended to the prefix, and the corresponding theoretical b- and y-fragment ions are identified. For each fragment ion, an intensity window of size 1.0 Da around its location on the input spectrum is retrieved. The combined intensity profile of the fragment ions then flows through a CNN, called ion-CNN. The ion-CNN learns local features (the peaks) and summarizes the overall information provided by the fragment ions in the spectrum (*SI Text*).

The second model of DeepNovo is an LSTM network, the most popular type of RNN (35). The LSTM model represents each amino acid class by an embedding vector [i.e., a collection of parameters that characterize the class; similar to word2vec (39)]. Given a prefix, the model looks for the corresponding embedding vectors and sequentially put them through the LSTM network. Moreover, DeepNovo also encodes the input spectrum and uses it to initialize the cell state of the LSTM network (36, 37). For that purpose, the spectrum is discretized into an intensity vector that subsequently flows through another CNN, called spectrum-CNN, before being fed to the LSTM network (Fig. 1*A* and *SI Text*).

The outputs of the two models are finally combined to produce a probability distribution over the amino acid classes. The next amino acid can be selected as the one with the highest probability or sampled from the distribution. Moreover, given the peptide mass and the prefix mass, DeepNovo calculates the suffix mass and uses the knapsack dynamic programming algorithm to filter out those amino acids with masses that do not fit the suffix mass. This processing guarantees that final candidate sequences will have the correct peptide mass. Combining all together, DeepNovo then performs beam search, a heuristic search algorithm that explores a fixed number of top candidate sequences at each iteration, until it finds the optimum prediction. Additional details of the model can be found in *Methods* and *SI Text*.

Datasets and Benchmark Criteria. We evaluated the performance of DeepNovo compared with current state of the art de novo peptide sequencing tools, including PEAKS [version 8.0 (40)], Novor (19), and PepNovo (12). For performance evaluation, we used two sets of data, low resolution and high resolution, from previous publications. The low-resolution set includes seven datasets (41–47) (Table S1). The first five datasets were acquired from the Thermo Scientific LTQ Orbitrap with the collision-induced dissociation (CID) technique. The other two were acquired from the Thermo Scientific Orbitrap Fusion with the higher-energy collisional dissociation (HCD) technique. The high-resolution set includes nine datasets acquired from the Thermo Scientific Q-Exactive with the HCD technique (48–56) (Table S2). We chose data from a wide variety of species and research groups to ensure an unbiased evaluation. All datasets can be downloaded from the ProteomeXchange database and the Chorus database. More details about the datasets and liquid chromatography (LC)-MS/MS experiments can be found in Tables S1 and S2 and the original publications.

We used PEAKS DB software [version 8.0 (40)] with a false discovery rate of 1% to search those datasets against the UniProt database and the taxon of the sample. The peptide sequences identified from the database search were assigned to the corresponding MS/MS spectra and then used as ground truth for testing the accuracy of de novo sequencing results. Tables S1 and S2 show

the summary of PEAKS DB search results for the low- and high-resolution datasets, respectively.

We performed leave-one-out cross-validations. In each validation, all except one of the datasets were used for training DeepNovo (from scratch), and the remaining dataset was used for testing. Other tools have already been trained by their authors and were only tested on all datasets. It should be noted that the training datasets and the testing dataset come from different species. The cross-validation is to guarantee unbiased training and testing and does not give DeepNovo any advantage. All tools were configured with the same settings, including fixed modification carbamidomethylation, variable modifications oxidation and deamidation, and fragment ion and precursor mass error tolerances (Tables S1 and S2).

To measure the accuracy of de novo sequencing results, we compared the real peptide sequence and the de novo peptide sequence of each spectrum. A de novo amino acid is considered "matched" with a real amino acid if their masses are different by less than 0.1 Da and if the prefix masses before them are different by less than 0.5 Da. Such an approximate match is used instead of an exact match because of the resolution of the benchmark datasets. We calculated the total recall (and precision) of de novo sequencing as the ratio of the total number of matched amino acids over the total length of real peptide sequences (and predicted peptide sequences, respectively) in the testing dataset. We also calculated the recall at the peptide level (i.e., the fraction of real peptide sequences that were fully correctly predicted). Most importantly, all sequencing tools report confidence scores for their predictions. The confidence scores reflect the quality of predicted amino acids and are valuable for downstream analysis [e.g., reconstructing the entire protein sequence from its peptides (21)]. Setting a higher threshold of confidence scores will output a smaller set of peptides with high precision but will leave the rest of the dataset without results, hence leading to lower recall and vice versa. Hence, given the availability of recall, precision, and confidence scores, it is reasonable to draw precision–recall curves and use the area under the curve (AUC) as a summary of de novo sequencing accuracy (57). These measures of sequencing accuracy have been widely used in previous publications (10, 12, 19).

Comparison of De Novo Sequencing Accuracy. Fig. 2 and Fig. S1 show the precision–recall curves and the AUC of de novo sequencing tools on the seven low-resolution datasets. DeepNovo considerably outperformed other tools across all seven datasets. In particular, for *Homo sapiens*, the AUC of DeepNovo was 33.3% higher than that of PEAKS (0.48/0.36 = 1.333) and 11.6% higher than that of Novor (0.48/0.43 = 1.116). PEAKS and Novor often came in the second place, whereas PepNovo fell behind, probably because of not being updated with recent data. We also noticed that Novor performed relatively better on CID data, whereas PEAKS performed relatively better on HCD data. The AUC of DeepNovo was 18.8–50.0% higher than PEAKS, 7.7–34.4% higher than Novor, and overall, 7.7–22.9% higher than the second best tool across all seven datasets. The improvement of DeepNovo over other methods was much better on HCD data than on CID data, probably because the HCD technique produces better fragmentations and hence, more patterns for DeepNovo to learn. The superior accuracy over state of the art sequencing tools on a wide variety of species shows the powerful and robust performance of DeepNovo.

Fig. 3 *A* and *B* shows the total recall and precision, respectively, of de novo sequencing results on the seven datasets. Here, we used all sequencing results from each tool, regardless of their confidence scores. Again, DeepNovo consistently achieved both higher recall and precision than other tools. DeepNovo recall was 8.4–30.2% higher than PEAKS and 3.9–22.1% higher than Novor. DeepNovo precision was 2.3–18.1% higher than PEAKS and 2.4–20.9% higher than Novor.

Fig. 3*C* shows the total recall of de novo sequencing tools at the peptide level. MS/MS spectra often have missing fragment ions, making it difficult to predict a few amino acids, especially those at the beginning or the end of a peptide sequence. Hence, de novo-sequenced peptides are often not fully correct. Those few amino acids may not increase the amino acid-level accuracy much, but they can result in substantially more fully correct peptides. As shown in Fig. 3*C*, DeepNovo greatly surpassed other tools; its recall at the peptide level was 38.1–88.0% higher than PEAKS and 42.7–67.6% higher than Novor. This result shows the advantage of the LSTM model in DeepNovo that makes use of sequence patterns to overcome the limitation of MS/MS missing data.

Fig. S2 shows the evaluation results on the nine high-resolution datasets. Novor and PepNovo were not trained with this type of data, and hence, their performance was not as good as PEAKS and DeepNovo. As can be seen from Fig. S2, the AUC of DeepNovo outperformed that of PEAKS across all nine datasets from 1.6 to 33.3%. Fig. S3 shows that the total amino acid recall of DeepNovo was 0.2–5.7% higher than that of PEAKS for eight datasets and 3.1% lower than PEAKS for the *H. sapiens* dataset.

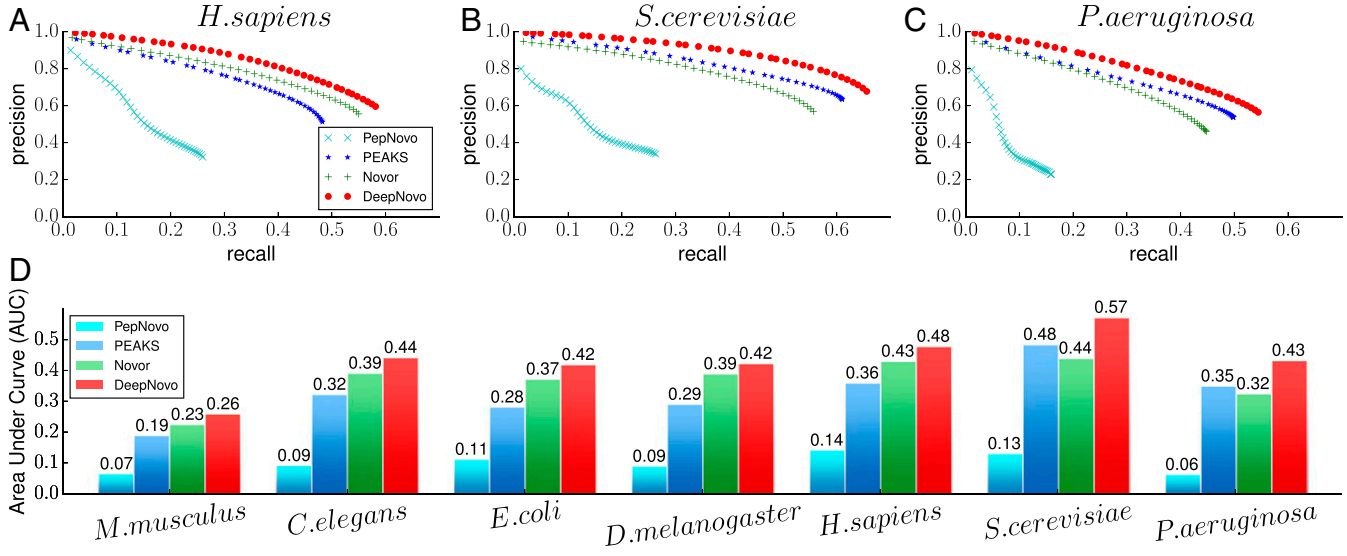

Fig. 2. The precision–recall curves and the AUC of PepNovo, PEAKS, Novor, and DeepNovo. (*A*) Precision–recall curves on *H. sapiens*. (*B*) Precision–recall curves on *Saccharomyces cerevisiae*. (*C*) Precision–recall curves on *Pseudomonas aeruginosa*. (*D*) AUC of four sequencing tools on seven datasets. *C. elegans*, *Caenorhabditis elegans*; *D. melanogaster*, *Drosophila melanogaster*; *E. coli*, *Escherichia coli*; *M. musculus*, *Mus musculus*.

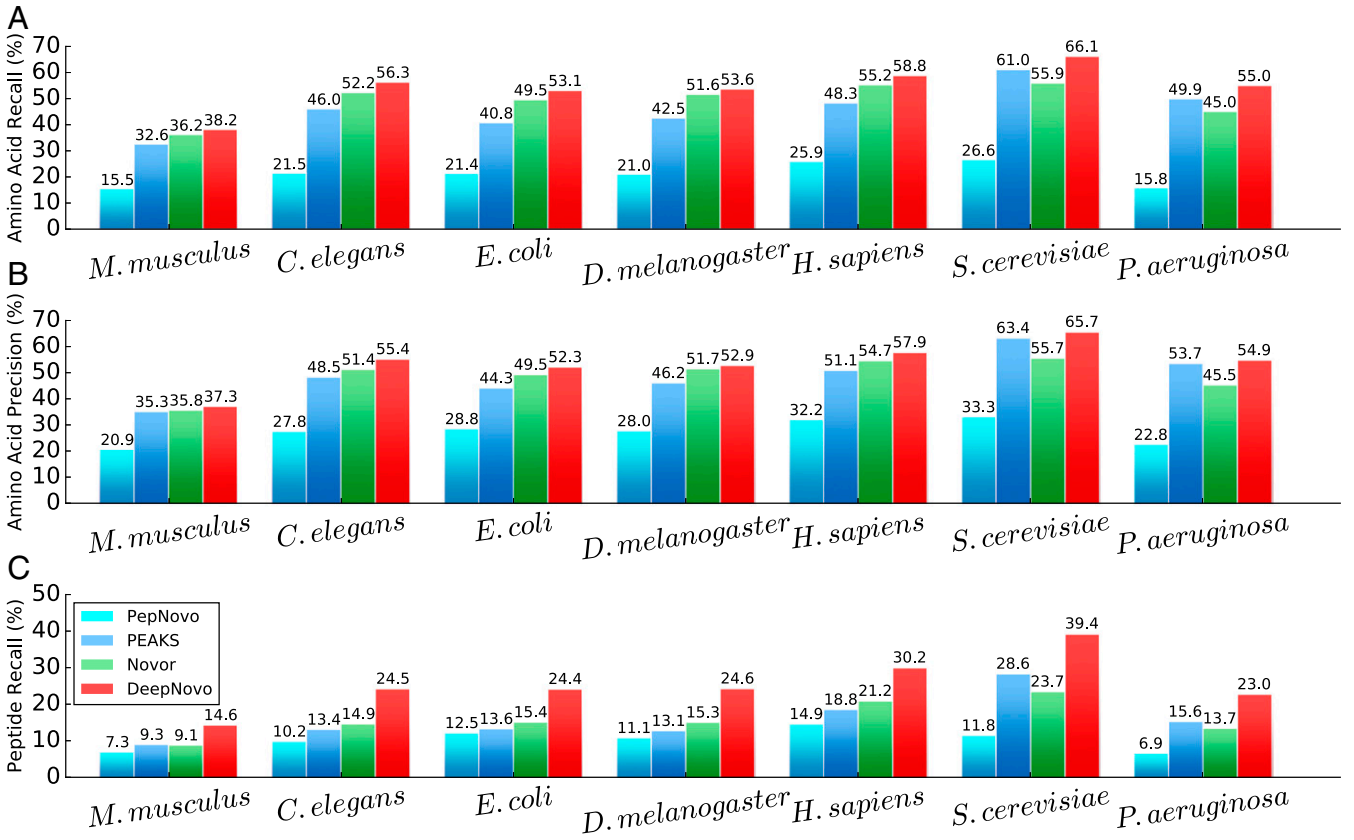

Fig. 3. Total recall and precision of PepNovo, PEAKS, Novor, and DeepNovo on seven datasets. (*A*) Recall at amino acid level. (*B*) Precision at amino acid level. (*C*) Recall at peptide level. *C. elegans*, *Caenorhabditis elegans*; *D. melanogaster*, *Drosophila melanogaster*; *E. coli*, *Escherichia coli*; *M. musculus*, *Mus musculus*; *P. aeruginosa*, *Pseudomonas aeruginosa*; *S. cerevisiae*, *Saccharomyces cerevisiae*.

At the peptide level, the total recall of DeepNovo was 5.9–45.6% higher than PEAKS across all nine datasets.

We also evaluated DeepNovo, Novor, and PEAKS on three testing datasets in the Novor paper (19). The results were consistent with those reported earlier, and DeepNovo achieved 4.1–12.1% higher accuracy than the other tools (Fig. S4).

Performance of Neural Networks Models in DeepNovo. The improvement of DeepNovo over state of the art methods comes from its two classification models, ion-CNN and LSTM, and the knapsack dynamic programming algorithm. Fig. 4 shows a detailed breakdown of how those components contributed to the total recall. DeepNovo has options to use its models individually or collectively, making it very convenient for additional research and development. The neural networks can be trained together, or they can also be trained separately and combined via the last hidden layer, a common training technique for multimodal neural networks. Although it is not a simple cumulative increasing of accuracy when one combines

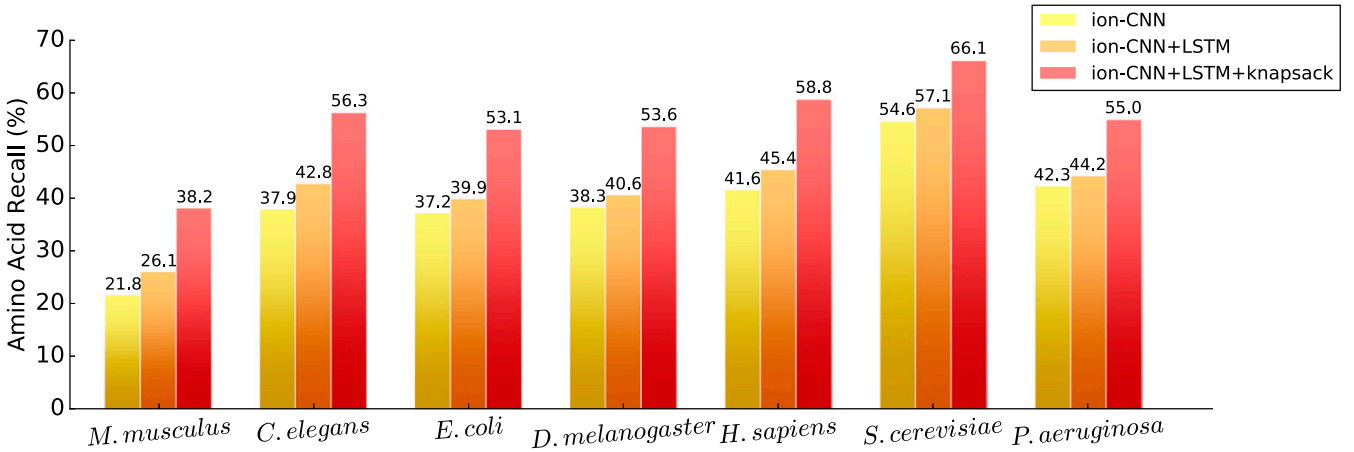

Fig. 4. The contributions of DeepNovo's components to its total recall on seven datasets. *C. elegans*, *Caenorhabditis elegans*; *D. melanogaster*, *Drosophila melanogaster*; *E. coli*, *Escherichia coli*; *M. musculus*, *Mus musculus*; *P. aeruginosa*, *Pseudomonas aeruginosa*; *S. cerevisiae*, *Saccharomyces cerevisiae*.

multiple models, Fig. 4 suggests that there is still much room to improve the LSTM network, and that will be our priority for additional development of DeepNovo.

Reconstructing Antibody Sequences with DeepNovo. In this section, we present a key downstream application of DeepNovo for complete de novo sequencing of mAbs. We trained the DeepNovo model with an in-house antibody database and used it to perform de novo peptide sequencing on two antibody datasets, the WIgG1 light and heavy chains of mouse (21). Note that the two testing datasets were not included in the training database. De novo peptides from DeepNovo were then used by the assembler ALPS (21) to automatically reconstruct the complete sequences of the antibodies (Figs. S5 and S6). For the light chain (length of 219 aa), we were able to reconstruct a single full-length contig that covered 100% of the target with 99.5% accuracy (218/219). For the heavy chain (length of 441 aa), we obtained three contigs together covering 97.5% of the target (430/441) with 97.2% accuracy (418/430). This application of whole-protein sequencing previously required both de novo peptide sequencing and database search together to succeed but now can be achieved with DeepNovo alone. This result further shows the major advantage of DeepNovo for de novo protein sequencing. In addition, we also showed that DeepNovo was able to identify peptides that could not be detected by database search (Fig. S7).

Discussion

De novo peptide sequencing is a challenging and computationally intensive problem that includes both pattern recognition and global optimization on noisy and incomplete data. In this study, we proposed DeepNovo, a deep neural network model that combines recent advances in deep learning and dynamic programming to address this problem. DeepNovo integrates CNNs and LSTM networks to learn features of tandem mass spectra, fragment ions, and sequence patterns for predicting peptides. Our experiment results show that DeepNovo consistently surpassed state of the art records in de novo peptide sequencing.

Interestingly, existing methods for de novo peptide sequencing rely heavily on rigorous global dynamic programming or graph-theoretical algorithms to address the global optimization problem. Here, we use knapsack, a "local" version of dynamic programming, to simply filter amino acids not suitable for the suffix mass, and we do not perform backtracking. This result implies that (*i*) the neural networks in DeepNovo learn better features that can bypass the global optimization problem and that (*ii*) DeepNovo can be further enhanced with more advanced search algorithms.

It should be noted that both method and training data are crucial for the model performance. For example, deep learning models often learn directly from raw data and require a large amount of training data. Some other machine learning models may rely on well-designed features based on domain-specific knowledge and may need less training data. Our study shows that DeepNovo and our training data achieved better de novo sequencing results than other existing methods and their respective training data. A more comprehensive benchmark study of de novo sequencing methods could be done by collecting well-annotated, gold standard training and testing datasets. This benchmark study is a potential question for future research.

Some database search engines and postprocessors, such as MS-GF+ (58) and Percolator (59), allow us to retrain their model parameters to adapt to a particular dataset and hence, increase the peptide identification rate. Similarly, PepNovo (12) includes the option to retrain its scoring models for de novo sequencing. DeepNovo is also retrainable and provides a complete end-to-end training and prediction solution. Retrainable is an important feature given massive amounts of data coming from several types of instruments and diverse species as well as different experiment designs. DeepNovo can be first trained on a huge amount of data to obtain a general model and then, gently retrained on a much smaller yet more targeted data source to reach the final data-specific model. Training data simply include a list of spectra and their corresponding peptides, and such annotated data can be found in spectral libraries, such as the NIST Mass Spectral Library, or retrieved by using database search tools (e.g., PEAKS DB) (40).

Because our work introduces deep learning to de novo peptide sequencing, there were no guidelines on how to design the architecture of neural networks or how to train them with tandem MS data. However, the lack of guidelines also means that there is still a lot of room for improvement. Going deeper is definitely an option. Another interesting topic is that protein sequences from different species may be considered as different languages. Hence, we need to explore how to train the LSTM network for a general model and a species-specific model. We can also train models to target a particular class of instruments or fragmentation techniques. Those all are potential directions for additional research.

Although DeepNovo is presented here in the context of de novo peptide sequencing, the idea can be easily extended to the sequence database search, because both share the same problem of matching a spectrum to a peptide. Moreover, we believe that DeepNovo can be further developed for the analysis of data-independent acquisition, in particular, the problem of inferring multiple sequences from a tandem mass spectra that include fragments from many different peptides. With the LSTM RNN, DeepNovo can learn patterns of peptide sequences in addition to the fragment ion information. The additional information of sequence patterns can offer some help to address the ambiguity of inferring multiple peptides from a spectrum.

After recent breakthroughs of deep learning in image processing, speech recognition, and natural language processing, DeepNovo makes important progress on de novo peptide sequencing, a fundamental and long-standing research problem in the field of bioinformatics. Our work opens a door for combining deep learning with other sophisticated algorithms to solve optimization problems, especially those with complicated mixing of signals and noises. This research will enable more applications of deep learning in the near future.

Methods

The architecture of DeepNovo is described in Fig. 1. More in-depth details of the model together with deep learning background are available in *SI Text*. DeepNovo is implemented and tested on the Google TensorFlow library, Python API, release r0.10.

To train DeepNovo, a dataset is randomly partitioned into three sets: training, validation, and testing. As we mentioned earlier, because of the one to many relationship between peptide and spectra, it is important to make sure that the three sets do not share peptides to avoid overfitting. The training dataset is processed in minibatches. At each training step, a minibatch is randomly selected from the training dataset and fed to the model. The model is provided with a real prefix and asked to predict the next amino acid. The output logits and the real amino acid are then used to calculate the cross-entropy loss function. Next, backpropagation is performed to calculate gradients (partial derivatives) and update the parameters using the Adam optimizer (60). During the training, we periodically calculate the loss function on the validation set and decide to save the model if there is an improvement. More training details can be found in *SI Text*.

ACKNOWLEDGMENTS. We thank Lin He for discussions. We also thank Nicole Keshav, Zac Anderson, and Brian Munro for proofreading the manuscript. This work was partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Grant OGP0046506, China's Key Research and Development Program under Grant 2016YFB1000902, Beijing Advanced Innovation Center for Imaging Technology Grant BAICIT-2016031, and Canada Research Chair program.

1. Johnson RS, Biemann K (1987) The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry* 26:1209–1214.
2. Martin-Visscher LA, et al. (2008) Isolation and characterization of carnocyclin a, a novel circular bacteriocin produced by Carnobacterium maltaromaticum UAL307. *Appl Environ Microbiol* 74:4756–4763.
3. Hatano N, Hamada T (2008) Proteome analysis of pitcher fluid of the carnivorous plant Nepenthes alata. *J Proteome Res* 7:809–816.
4. Catusse J, Strub J-M, Job C, Van Dorsselaer A, Job D (2008) Proteome-wide characterization of sugarbeet seed vigor and its tissue specific expression. *Proc Natl Acad Sci USA* 105:10262–10267.

5. Jorrín-Novo JV, et al. (2015) Fourteen years of plant proteomics reflected in Proteomics: Moving from model species and 2DE-based approaches to orphan species and gel-free platforms. *Proteomics* 15:1089–1112.
6. Taylor JA, Johnson RS (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 11: 1067–1075.
7. Taylor JA, Johnson RS (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal Chem* 73:2594–2604.
8. Chen T, Kao MY, Tepel M, Rush J, Church GM (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J Comput Biol* 8:325–337.
9. Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J Comput Biol* 6:327–342.
10. Ma B, et al. (2003) PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 17:2337–2342.
11. Zhang Z (2004) *De novo* peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem* 76:6374–6383.
12. Frank A, Pevzner P (2005) PepNovo: *De novo* peptide sequencing via probabilistic network modeling. *Anal Chem* 77:964–973.
13. Fischer B, et al. (2005) NovoHMM: A hidden Markov model for *de novo* peptide sequencing. *Anal Chem* 77:7265–7273.
14. DiMaggio PA, Jr, Floudas CA (2007) *De novo* peptide identification via tandem mass spectrometry and integer linear optimization. *Anal Chem* 79:1433–1446.
15. Mo L, Dutta D, Wan Y, Chen T (2007) MSNovo: A dynamic programming algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *Anal Chem* 79:4870–4878.
16. Chi H, et al. (2010) pNovo: *De novo* peptide sequencing and identification using HCD spectra. *J Proteome Res* 9:2713–2724.
17. Jeong K, Kim S, Pevzner PA (2013) UniNovo: A universal tool for *de novo* peptide sequencing. *Bioinformatics* 29:1953–1962.
18. Chi H, et al. (2013) pNovo+: *De novo* peptide sequencing using complementary HCD and ETD tandem mass spectra. *J Proteome Res* 12:615–625.
19. Ma B (2015) Novor: Real-time peptide *de novo* sequencing software. *J Am Soc Mass Spectrom* 26:1885–1894.
20. Maggon K (2007) Monoclonal antibody "gold rush." *Curr Med Chem* 14:1978–1987.
21. Tran NH, et al. (2016) Complete *de novo* assembly of monoclonal antibody sequences. *Sci Rep* 6:31730.
22. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR (2008) Automated *de novo* protein sequencing of monoclonal antibodies. *Nat Biotechnol* 26:1336–1338.
23. Guthals A, Clauser KR, Frank AM, Bandeira N (2013) Sequencing-grade *de novo* analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides. *J Proteome Res* 12:2846–2857.
24. Ma B, Johnson R (2012) De novo sequencing and homology searching. *Mol Cell Proteomics* 11:O111.014902.
25. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
26. Ciresan D, Giusti A, Gambardella LM, Schmidhuber J (2012) Deep neural networks segment neuronal membranes in electron microscopy images. *Adv Neural Inf Process Syst* 25:2843–2851.
27. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 25:1097–1105.
28. Hinton G, et al. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Mag* 29:82–97.
29. Sutskever I, Vinyals O, Le Q (2014) Sequence to sequence learning with neural networks. *Adv Neural Inf Process Syst* 27:3104–3112.
30. Rusk N (2016) Deep learning. *Nat Methods* 13:35.
31. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12:931–934.
32. Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33:831–838.
33. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate *de novo* prediction of protein contact map by ultra-deep learning model. *PLOS Comput Biol* 13:e1005324.
34. Inglese P, et al. (2017) Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem Sci (Camb)* 8:3500–3511.
35. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780.
36. Karpathy A, Li FF (2015) Deep visual-semantic alignments for generating image description. *Conf Comput Vis Pattern Recognit Workshops* 2015:3128–3137.
37. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. *Conf Comput Vis Pattern Recognit Workshops* 2015:3156–3164.
38. Steen H, Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5:699–711.
39. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 26:3111–3119.
40. Bioinformatics Solutions Inc. (2016) PEAKS Studio (Bioinformatics Solutions Inc., Waterloo, ON, Canada), Version 8.0.
41. Grosche A, et al. (2016) The proteome of native adult Muller Glial cells from murine retina. *Mol Cell Proteomics* 15:462–480.
42. Marza E, et al. (2015) Genome-wide screen identifies a novel p97/CDC-48-dependent pathway regulating ER-stress-induced gene transcription. *EMBO Rep* 16:332–340.
43. Pettersen VK, Mosevoll KA, Lindemann PC, Wiker HG (2016) Coordination of metabolism and virulence factors expression of extraintestinal pathogenic *Escherichia coli* purified from blood cultures of patients with sepsis. *Mol Cell Proteomics* 15:2890–2907.
44. Hampoelz B, et al. (2016) Pre-assembled nuclear pores insert into the nuclear envelope during early development. *Cell* 166:664–678.
45. Zhang Y, et al. (2015) Tissue-based proteogenomics reveals that human testis endows plentiful missing proteins. *J Proteome Res* 14:3583–3594.
46. Hebert AS, et al. (2014) The one hour yeast proteome. *Mol Cell Proteomics* 13: 339–347.
47. Peng J, Cao J, Ng FM, Hill J (2017) Pseudomonas aeruginosa develops Ciprofloxacin resistance from low to high level with distinctive proteome changes. *J Proteomics* 152:75–87.
48. Paiva AL, Oliveira JT, de Souza GA, Vasconcelos IM (2016) Label-free proteomics reveals that Cowpea severe mosaic virus transiently suppresses the host leaf protein accumulation during the compatible interaction with Cowpea (*Vigna unguiculata* [L.] Walp.). *J Proteome Res* 15:4208–4220.
49. Nevo N, et al. (2017) Impact of cystinosin glycosylation on protein stability by differential dynamic stable isotope labeling by amino acids in cell culture (SILAC). *Mol Cell Proteomics* 16:457–468.
50. Cassidy L, Prasse D, Linke D, Schmitz RA, Tholey A (2016) Combination of bottom-up 2D-LC-MS and semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the Archaeon Methanosarcina mazei. *J Proteome Res* 15:3773–3783.
51. Reuß DR, et al. (2017) Large-scale reduction of the Bacillus subtilis genome: Consequences for the transcriptional network, resource allocation, and metabolism. *Genome Res* 27:289–299.
52. Petersen JM, et al. (2016) Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation. *Nat Microbiol* 2:16195.
53. Mata CI, et al. (2017) In-depth characterization of the tomato fruit pericarp proteome. *Proteomics* 17:1–2.
54. Seidel G, et al. (2017) Quantitative global proteomics of Yeast PBP1 deletion mutants and their stress responses identifies glucose metabolism, mitochondrial, and stress granule changes. *J Proteome Res* 16:504–515.
55. Hu H, et al. (2016) Proteome analysis of the hemolymph, mushroom body, and antenna provides novel insight into honeybee resistance against varroa infestation. *J Proteome Res* 15:2841–2854.
56. Cypryk W, Lorey M, Puustinen A, Nyman TA, Matikainen S (2017) Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon Influenza A virus infection. *J Proteome Res* 16:217–227.
57. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, eds Cohen W, Moore A (ACM, New York), pp 233–240.
58. Kim S, Pevzner PAMS-GF (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5:5277.
59. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 4:923–925.
60. Kingma DP, Ba J Adam: A method for stochastic optimization. arXiv:1412.6980.
61. LeCun Y, et al. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 11:541–551.
62. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. *JMLR Workshop Conf Proc* 15:315–323.
63. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15: 1929–1958.