# Determining Appropriate Score Thresholds using DeepNovo

Rui Qiao, Lei Xin, Shengying Pan, Xiaodong Wei, Xin Chen, Jonathan R. Krieger, Baozhen Shan
Bioinformatics Solutions Inc., Waterloo, Canada

## Abstract

*De novo* peptide sequencing is the key technology for finding novel peptides from mass spectrometry (MS) data. It has been applied to important applications including antibody sequencing and neoantigen discovery. Recent advances in *de novo* sequencing often involve deep neural networks to enhance prediction accuracy. The adoption of deep learning has significantly improved sequencing accuracy at both the amino acid and peptide level. However, the probability scores given by these neural-networks-based models have different distributions from the widely used PEAKS de novo average local confidence (ALC) scores. In this application note, we help users determine an appropriate score threshold when using DeepNovo.

## Introduction

In 2017, Tran et al. first introduced deep learning technology into *de novo* peptide sequencing of mass spectrometry data [1] and reported a significant improvement in accuracy. Since then, there have been many publications using deep neural networks to improve *de novo* peptide sequencing methods and solve other important problems in mass spectrometry [3-6]. More recently, Qiao et al. proposed a PointNet [7] based de novo peptide sequencing algorithm [8] that further improves the peptide level accuracy by a significant margin of 15%. Moreover, this algorithm can process more than 70 spectra per second on a

single commercial grade GPU, making it possible to perform real time *de novo* peptide sequencing on a GPU workstation [8]. The speed and accuracy of this technology naturally draw interests in turning it into this product we present as DeepNovo.

The positional amino acid score given by DeepNovo represents the log probability of an amino acid appearing at a certain position within a peptide sequence. With this, the average log probability of all positions is used as a peptide level score. The different physical meanings of DeepNovo scores makes it necessary to adjust previously widely used heuristics like the 50% peptide ALC score cutoff [2].

## Methods and Results

To propose appropriate score thresholds to use with DeepNovo, we study the relationship between DeepNovo scores and the well-known PEAKS de novo ALC scores.

We prepared two in-house datasets run on a timsTOF Pro instrument. A 24 fraction human single species dataset and a single shot mixed species dataset containing proteins from Human, *S. cerevisiae* and *E.coli* (HYE). The DeepNovo model was trained on the 24 fraction single species dataset, and the model was evaluated on a single shot of the mixed species dataset.

To evaluate the model, we first extracted the peptide spectrum matches (PSMs) above 1% FDR reported by database search (DB search) using PEAKS Online. We then compared the de novo sequences given by DeepNovo on those same PSMs recovered from DB search (Table 1). To assess the accuracy of the model, we adopted an established evaluation metric previously used by DeepNovo. Essentially, at each amino acid, two criteria need to be met for a peptide to be considered a real target match: (i) The mass difference between the de novo generated amino acid and the database amino acid must be smaller than 0.1 Da, and (ii) the mass difference between the prefix masses (the partial peptide generated up to this particular amino acid) in the de novo generated sequence and database search must have a delta mass of less than 0.5Da.

Table 1. Accuracy of DeepNovo *de novo* sequenced peptides

| | DeepNovo |
|---|---|
| Amino acid accuracy | 64.71% |
| Amino acid recall | 65.17% |
| Peptide recall | 35.53% |

Due to the nature of MS data, *de novo* peptide sequencing results are known to be prone to errors. A commonly used technique is to filter *de novo* peptides by a score threshold, retaining only the high confidence results for further analysis. To assess the appropriate score for DeepNovo, we analyzed the effects of different DeepNovo score thresholds and an amino acid level accuracy versus score thresholds plot is shown in Figure 1.

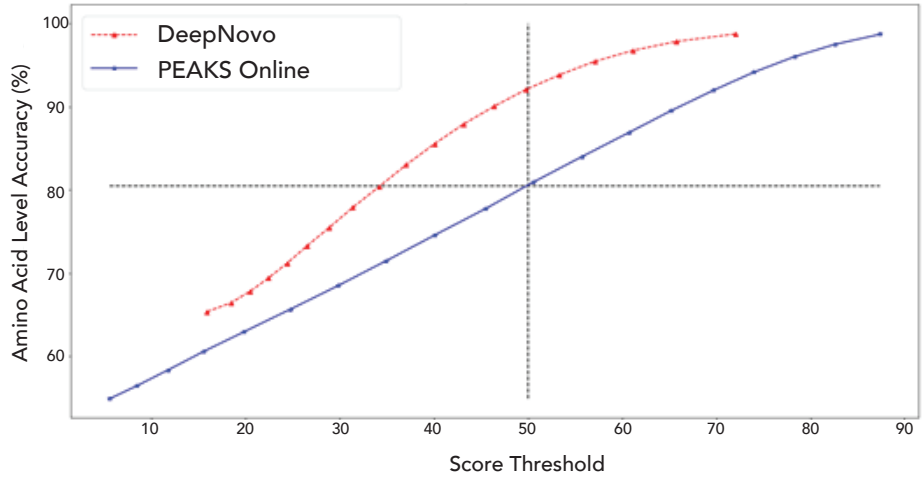


Figure 1: Amino acid accuracy versus de novo score threshold

Traditionally, a PEAKS *de novo* ALC score threshold of 50% can help users to retain PEAKS *de novo* peptides that on average have an amino acid level accuracy of 80%. A similar approach can be achieved with DeepNovo when a score threshold of around 35% is used. Therefore, we suggest that a 35% DeepNovo peptide score is good replacement for the 50% ALC score filtering heuristic.

## Conclusion

A DeepNovo score of >35% can be considered as an accurate score, and is comparable to the PEAKS de novo 50% ALC score.

## References

[1] Tran, Ngoc Hieu, et al. "De novo peptide sequencing by deep learning." Proceedings of the National Academy of Sciences 114.31 (2017): 8247-8252.

[2] Ma, Bin, et al. "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry." Rapid communications in mass spectrometry 17.20 (2003): 2337-2342.

[3] Karunratanakul, Korrawe, et al. "Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework." Molecular & Cellular Proteomics 18.12 (2019): 2478-2491.

[4] Tran, Ngoc Hieu, et al. "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry." Nature methods 16.1 (2019): 63-66.

[5] Gessulat, Siegfried, et al. "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning." Nature methods 16.6 (2019): 509-518.

[6] Bulik-Sullivan, Brendan, et al. "Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification." Nature biotechnology 37.1 (2019): 55-63.

[7] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Qiao, Rui, et al. "Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices." Nature Machine Intelligence (2021): 1-6.

**Bioinformatics Solutions Inc.**
204-470 Weber St.N., Waterloo, ON, Canada
Tel: 1-855-885-8288

Follow us on social media
@Bioinformatics-Solutions-Inc-
@PEAKSProteomics
@Bioinfor