Bioinformatics Solutions Inc.

# Accuracy and score distributions comparison of PEAKS *de novo* and *DeepNovo* using HLA datasets

Qing Zhang, MSc, Software Systems Manager
Kyle Hoffman, Phd, Applications Manager

Bioinformatics Solutions Inc., Waterloo, Canada

## Abstract:

*DeepNovo* is a deep learning based algorithm for *de novo* sequencing that predicts the peptide from the MS/MS scan by iteratively predicting amino acids consecutively. The goal of this note is to compare PEAKS *de novo* and *DeepNovo* approaches by evaluating the accuracy and scoring of HLA peptidome data.

## Inroduction:

Over a decade ago, PEAKS *de novo* sequencing was introduced as a tool to assist peptide spectral matching to a protein sequence database. This approach increases both sensitivity and accuracy of peptide identifications. *De novo* sequencing derives the peptide sequence directly from the MS/MS spectrum, with PEAKS *de novo* algorithm computing amino acid sequences with the local confidence scores for each position as well as the confidence scores for the entire sequence [1]

Earlier this year, we introduced a workflow, called *DeepNovo* Peptidome, for the analysis of immunopeptides. Here, in addition to the classic database and homology searches, PEAKS uses a novel and unique *de novo* sequencing approach, named DeepNovo [2,3]. Importantly, while PEAKS *de novo* sequencing algorithm produces a score based on the concept of the score gap of the Peptide-Spectrum-Match (PSM) by mutating each amino acid in the peptide, *DeepNovo* takes a very different approach to evaluate the amino acid confidence. DeepNovo predicts the *de novo* peptide by iteratively predicting one amino acid after another. At each iteration, based on the spectrum and the amino acid sequence already generated, *DeepNovo* predicts the next amino acid and its score. This score is a value from 0 to 100 and represents the probability of a particular amino acid, out of 20 possible amino acid candidates considered by PEAKS (or more candidates if PTMs are included), to be present at a position within the peptide sequence. The score of the predicted peptide is then calculated as the average of its amino acid scores.

The goal of this note is to compare PEAKS *de novo* and *DeepNovo* approaches by evaluating the accuracy and scoring of HLA peptidome data.

## Methods and Results:

For evaluating PEAKS *de novo* and *DeepNovo* algorithms, we used immunopeptidome data acquired by timsTOF SCP (Bruker) [4] and Orbitrap (Thermo) [5] mass spectrometers to perform PEAKS Database (DB) search, PEAKS *de novo* search, and *DeepNovo* search. We selected one sample from each dataset for the data analysis.

For timsTOF SCP data, a total of 13,322 PSMs was obtained from the PEAKS DB search under 0.1% Peptide FDR. These PSMs were then used to evaluate the de novo results as follows. For each PSM, the database identified peptide was considered as the target, and the de novo peptides predicted by PEAKS *de novo* or *DeepNovo* were compared to the target peptides to determine the peptide and amino acid accuracies. The de novo score cutoffs for both PEAKS *de novo* and *DeepNovo* were set to 0 in order to consider all peptide sequence predictions.

The calculated accuracies of PEAKS *de novo* and *DeepNovo* predictions for the resulting PSMs are shown in Table 1. Here, *DeepNovo* outperforms PEAKS *de novo* by 20.5% at the amino acid level and 16.6% at the peptide level. Furthermore, Figure 1 shows the overlap between PEAKS DB, PEAKS *de novo*, and *DeepNovo* peptides, as well as the overlap of the amino acids predicted by PEAKS *de novo* and *DeepNovo*. While PEAKS *de novo* and *DeepNovo* shared a large amount of amino acid assignments (up to 64%), they shared much fewer common peptides (32.5%). More importantly, compared to PEAKS *de novo*, *DeepNovo* shared significantly more peptides with PEAKS DB.

| | Accuracy | |
|---|---|---|
| **Tools** | **Amino Acid** | **Peptide** |
| PEAKS *de novo* | 60.9% | 34.6% |
| *DeepNovo* | 81.4% | 51.2% |

Table 1. timsTOF Accuracy of PEAKS *de novo* and *DeepNovo* predictions of PSMs at the amino acid and peptide levels.

| | Accuracy | |
|---|---|---|
| **Tools** | **Amino Acid** | **Peptide** |
| PEAKS *de novo* | 72.0% | 42.5% |
| *DeepNovo* | 77.5% | 53.1% |

Table 2. Orbitrap Accuracy of PEAKS *de novo* and *DeepNovo* predictions of PSMs at the amino acid and peptide levels.

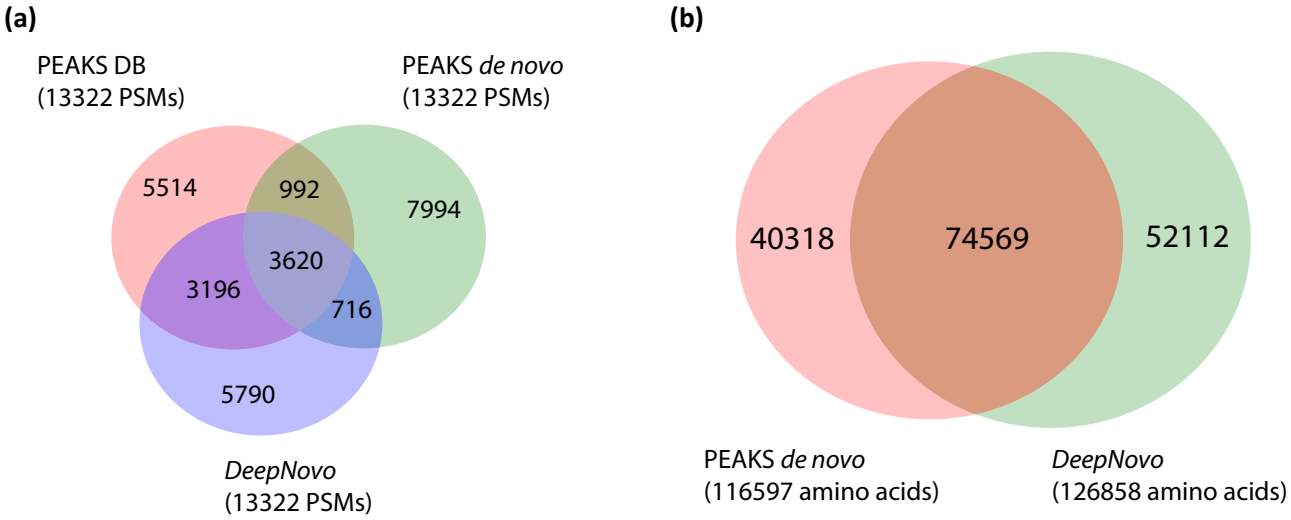


Fig 1. Venn diagrams showing an overlap of PEAKS DB, PEAKS *de novo* and *DeepNovo* peptides (left) and an overlap of amino acids predicted by PEAKS *de novo* and *DeepNovo* (right).

Next, we investigated the distributions of the amino acid accuracies versus the *de novo* scores to determine the score cutoffs. We varied the score cutoffs from 0 to 100, and for each cutoff, we counted the *de novo* predictions above each score to calculate the amino acid accuracy. The distributions of PEAKS *de novo* and *DeepNovo* scores are shown in Figure 2. For example, for a score ≥55 for *DeepNovo*, the predicted peptides will have the amino accuracy of ~95%. Similarly, for the peptides predicted by PEAKS *de novo* having the amino acid accuracy of around 95%, the *de novo* score cutoff needs to be set around 80.

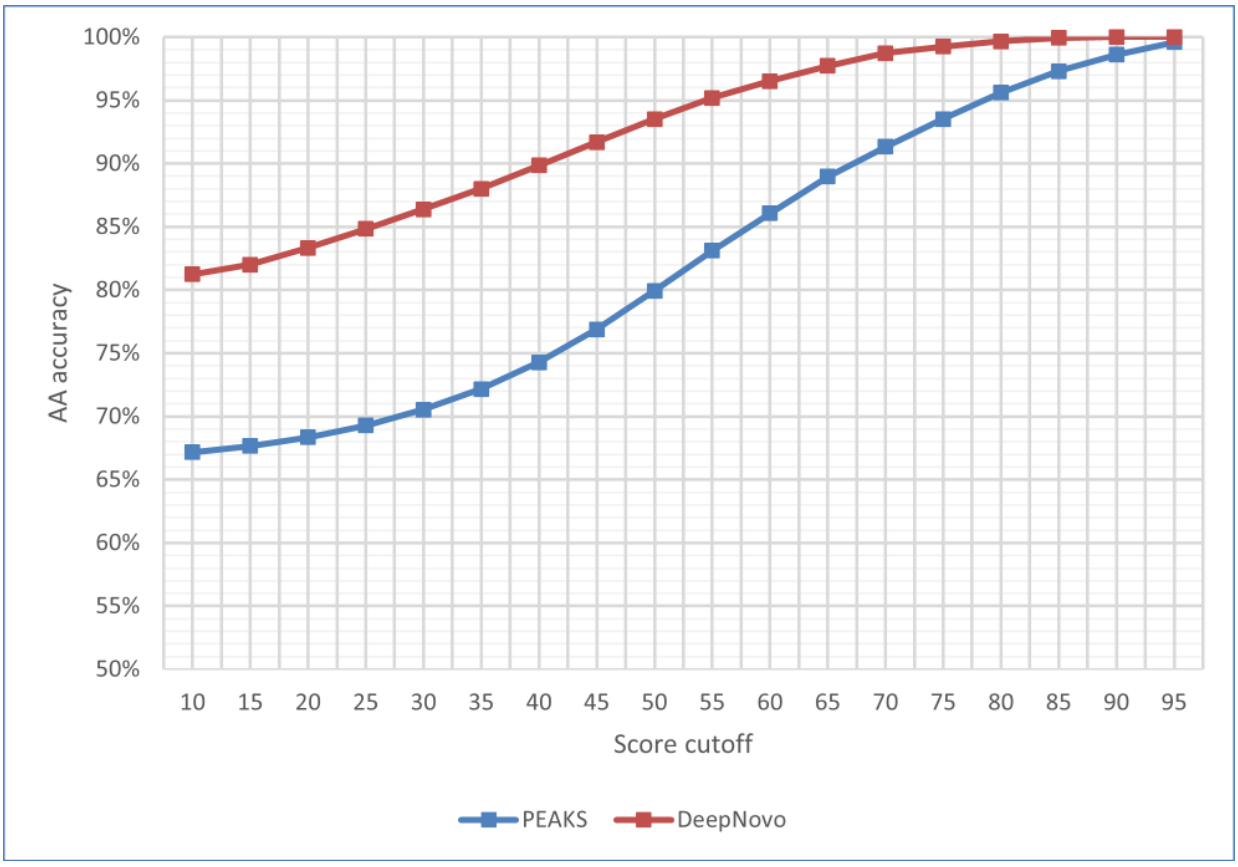


Fig 2. The accuracy-versus-score distributions of PEAKS *de novo* and *DeepNovo* peptides from timsTOF SCP data.

The accuracy-versus-score distribution is hypothesized to be data and instrument dependent. To assess this hypothesis, we looked at the relationship between *de novo* and *DeepNovo* predictions using an Orbitrap dataset following the same approach as described for timsTOF SCP. Table 2 and Figure 3 validated our earlier conclusion that *DeepNovo* performed better than *de novo* in accurate predictions of amino acids and peptide sequences. Indeed, for a similar amino acid accuracy of 95%, the required PEAKS *de novo* score is ~95 while *DeepNovo* is ~60.

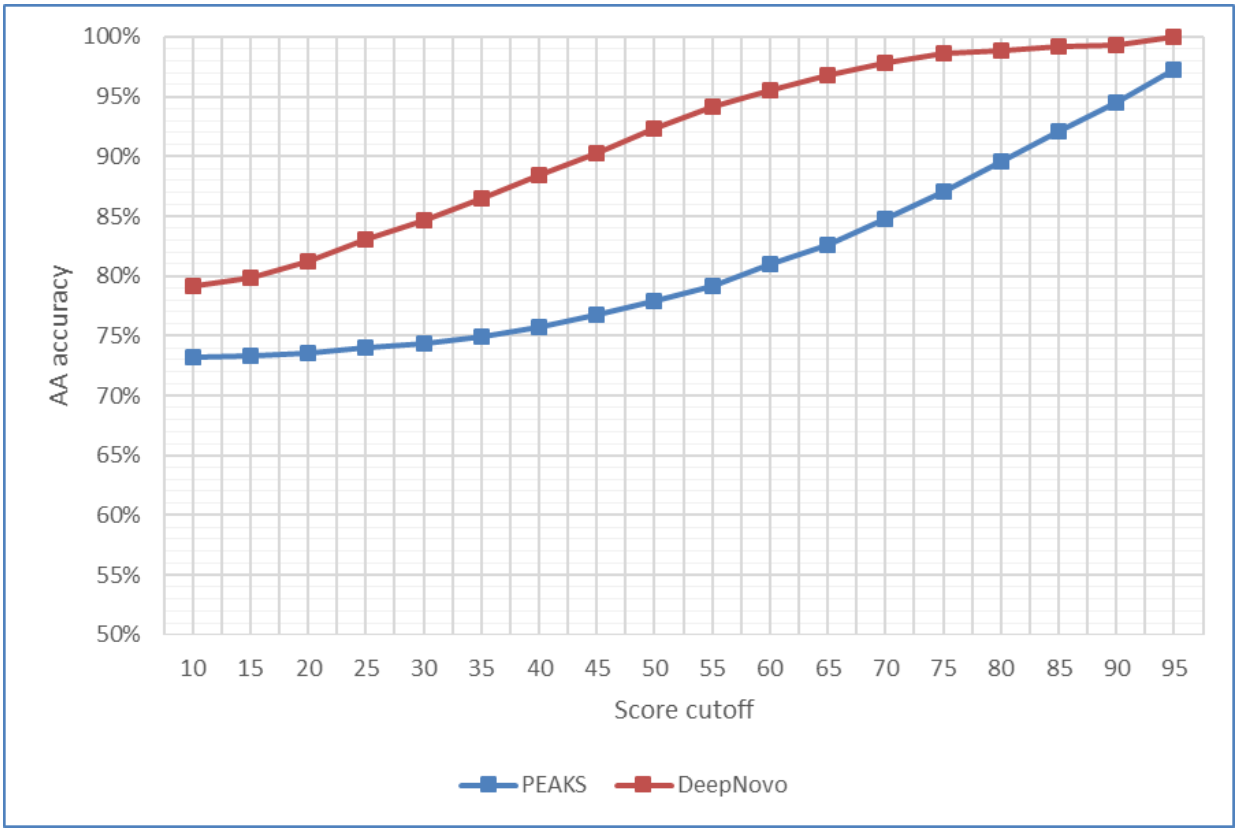


Fig 3. The accuracy-versus-score distributions of PEAKS *de novo* and *DeepNovo* peptides from Thermo Orbitrap data.

## Conclusion:

From our experience with many large public datasets and internal data, we recommend setting a minimum score of 80 when using PEAKS *de novo*, and 55 for *DeepNovo* score (for timsTOF data). For Orbitrap, the recommended cut-off for *DeepNovo* score is 60. Those cutoffs correspond to an average amino acid accuracy of around 95%.

## References:


1. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom., 17, 2337-2342 (2003).
2. Tran, N. H., Zhang, X., Xin, L., Shan, B., & Li, M. (2017). De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences of the United States of America, 114(31), 8247–8252. https://doi.org/10.1073/pnas.1705691114
3. Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., & Li, M. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nature methods, 16(1), 63–66. https://doi.org/10.1038/s41592-018-0260-3
4. Phulphagar, K., et al., Sensitive, high-throughput HLA-I and HLA-II immunopeptidomics using parallel accumulation-serial fragmentation mass spectrometry. bioRxiv, 2023: p. 2023.03.10.532106.
5. Tretter, C., et al., Proteogenomic analysis reveals RNA as an important source for tumor-agnostic neoantigen identification correlating with T-cell infiltration. 2022, Cold Spring Harbor Laboratory.

**Bioinformatics Solutions, Inc.**

140 Columbia St, Suite 202
Waterloo, Ontario N2L 3K8
Canada

Tel: (519) 885-8288
Fax: (519) 885-9075

sales@bioinfor.com
www.bioinfor.com