

# Deep learning benchmark data for *de novo* peptide sequencing

Joon-Yong Lee<sup>1</sup>, Lisa Bramer<sup>2</sup>, Nathan Hodas<sup>2</sup>, Courtney D. Corley<sup>2</sup>, Samuel H. Payne<sup>1</sup>

<sup>1</sup>Biological Sciences Division and <sup>2</sup>National Security Directorate  
Pacific Northwest National Laboratory, Richland WA

## Overview

### Questions

- How can we annotate mass spectrometry data without a protein database?
- Can deep learning improve this *de novo* sequencing?

### Challenges

- Deep learning requires very large and publicly accessible datasets.
- Large and well-curated benchmark datasets do not yet exist.

### Solution

- We present a benchmark dataset for deep learning in proteomics.
- Utilizing our massive dataset, we improve best-in-class algorithms by 3-5x

## Introduction

- Peptide identification via database search is a popular and statistically robust method to annotate mass spectrometry data, but only works if the sequence database are available.
- *De novo* sequencing, which is the database-free peptide identification, is critical for microbiome and environmental research.
- Vast sequence space ( $20^n$  possible combinations) makes *de novo* sequencing very challenging. Current *de novo* peptide sequencing methods average 10% accuracy. This lack of accuracy prevents broad adoption.

## Methods

### Building benchmark data

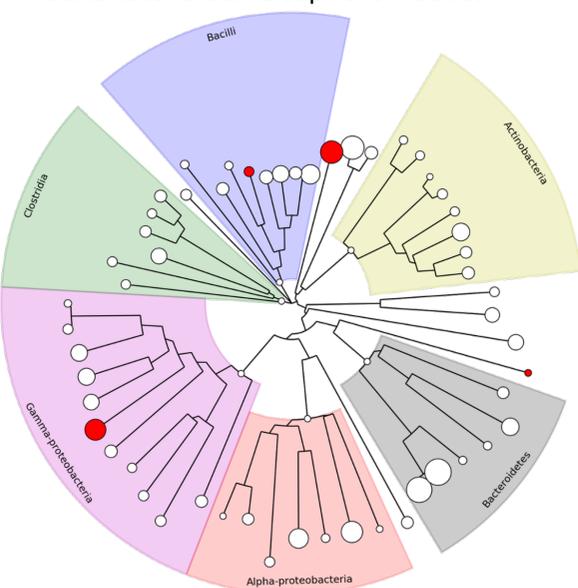
- We generate proteomics data from 55 diverse bacteria.
- All spectra are annotated with industry standard MSGF+ algorithm (database search).
- Results are filtered to  $q < 0.001$ .
- Final dataset contains 5.76 million spectra representing 1 million unique peptides.
- Data was extracted and regularized for easy input to DeepLearning packages

### Testing benchmark data

- DeepNovo is a deep learning framework introduced for *de novo* peptide sequencing, written in TensorFlow (doi.org/10.1073/pnas.1705691114).
- We evaluated the accuracy of DeepNovo with our benchmark dataset
- We re-trained DeepNovo models and we investigated the effect of the size of training datasets.
- Hyperparameter optimization for DeepNovo improved performance.
- The DeepLearning benchmark dataset can facilitate the train/test to improve models.

## Results

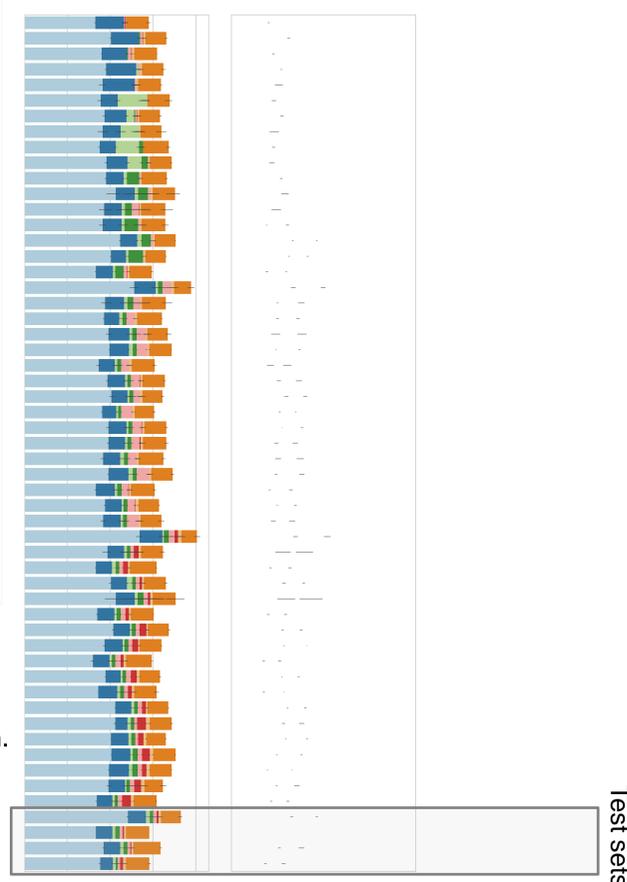
- We gathered 5.76 million spectra from 55 bacteria to re-train DeepNovo models.



† Clade sizes are proportional to the number of spectra. Red and white colors indicate the test and train sets, respectively.

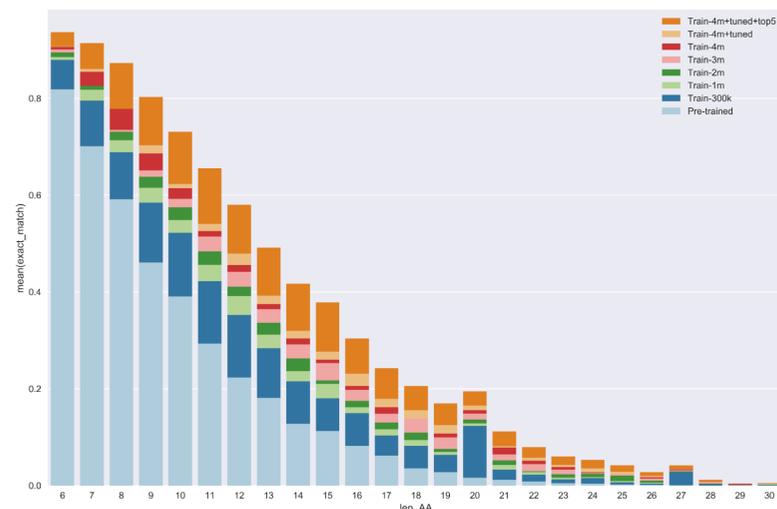
### Test results

- Histogram of accuracy levels according to all species

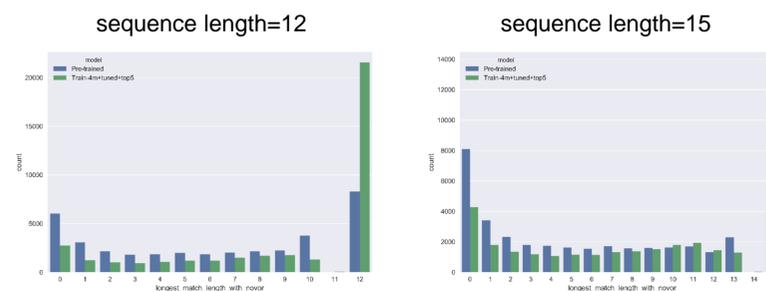


### Test accuracy comparisons by AA length

- Interestingly, for peptides less than 15 AA length, it performs over 40% accuracy levels which is reasonably better than any other non-DL tools.

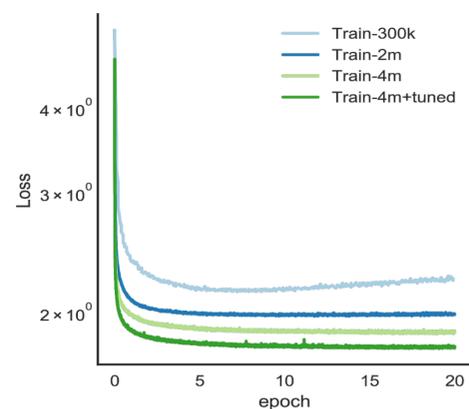


- The length of longest correct subsequences is dramatically improved.

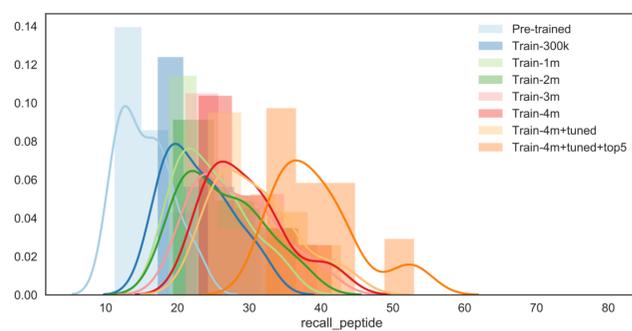


### Training results

- The below figures show the loss trends for validation sets in training model within 20 epoch.
- Increased training data led to lower Loss during training. Training with less than 2M spectra suffered from overfitting.



- Histogram of accuracy levels for test datasets



## Conclusion

- We achieved 3~5x improved accuracy with deep learning models feeding a large volume of data.
- Hyperparameter optimization will be promising to create the better model with our benchmark dataset.

## Acknowledgments

This work was supported by PNNL's DeepScience LDRD, the U.S. Department of Energy, Office of Biological and Environmental Research, Early Career Research Program, and the NIH/NIGMS (GM103493). Work was performed in the Environmental Molecular Science Laboratory, a DOE national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA. Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830.

