Deep learning benchmark data for *de novo* peptide sequencing

Joon-Yong Lee[1*], Lisa Bramer[2], Nathan Hodas[2], Courtney D. Corley[2], Samuel H. Payne[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory
[2]National Security Directorate, Pacific Northwest National Laboratory

[*]Email: joonyong.lee@pnnl.gov

Deep learning has been quickly adapted to various applications in biological context after having shown very successful achievements in computer vision and natural language processing. In the big data era, deeper models powered by a vast number of datasets can extract and recognize complex pattern driven from image and text data. In proteomics, recently, there have been a few attempts to take advantage of deep learning approaches to solve historically challenging problems.[1] One such problem is the database-free peptide identification, which is critical for microbiome research. In addressing this challenge with deep learning, researchers require very large and publicly accessible datasets. However, such a large and well-curated benchmark dataset do not yet exist. We present a benchmark dataset for developing a deep learning approach in proteomics. It totally contains 5.76 million spectra obtained by high resolution mass spectrometry representing 1 million unique peptides using rigorous quality controls. Using our benchmark dataset, we tested a deep learning framework introduced for *de novo* peptide sequencing. We discovered that the size of training data significantly affects to the performance in the previous deep learning model for *de novo* sequencing. We also improved the deep learning model with hyperparameter optimization.

References:

1. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci.* **114,** 8247–8252 (2017).