# Deep Learning-based Comprehensive Prediction of Peptide Tandem Mass Spectra to Generate Proteome-Scale Mass Spectral Libraries

Qian Dong, Joel Lapin, Xinjian (Eric) Yan, MSDC, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

## Overview

In contrast to the rule-based traditional models, deep learning (DL) techniques, a subset of machine learning that uses artificial neural networks to mimic the learning process of the human brain, have emerged as a more advanced solution for spectral prediction. The recent advancements in this field have been fueled by modern shotgun proteomics generating large volumes of high-throughput peptide tandem mass spectrometry data, which provide ample data for training DL models. However, current DL approaches have largely excluded a substantial fraction of higher energy-induced collisional dissociation (HCD) fragment ions beyond the typical sequence b and y ion series, mainly due to the challenges of understanding and annotating complex fragment ion types under collision-induced dissociation. This study aims to enhance peptide identification by predicting a broader range of HCD fragment ion types, ultimately providing a more accurate and in-depth interpretation of peptide spectra.

## Data and Methods

We generated a high-quality DL dataset collected from the NIST reference peptide spectral libraries[1], consisting of approximately 1.8 million unique tandem mass spectra from around 800,000 protonated peptides. The spectral data are preprocessed and tailored for DL model development, in which each raw mass spectrum was filtered, and collision energy recalibrated. By automatically analyzing all annotated fragment ions in these libraries, we constructed a comprehensive HCD fragmentation dictionary, encompassing 7,919 isotope peaks originating from 3,143 distinct sequence ions, neutral losses, internal, immonium, and amino acid fragment ions. We implemented an attention-based deep neural network. The architecture of our model starts with an input tensor that embeds peptide sequence, modification, charge, and collision energy and ends with a one-dimensional output vector of prediction intensities for the 7,919 isotope peaks in our dictionary.

## Results

**1. Data Strategy.** NIST peptide libraries are comprehensive, curated mass spectral reference collections from various organisms and proteins useful for the rapid matching and identification of acquired MS/MS spectra. Can we directly combine all these different mass spectral libraries for model training?

**Table 1.** NIST peptide libraries of 5 million unique HCD tandem mass spectra

| Experiment Type | Library Name | Spectrum | Collision Energy | Spectra | Peptide | Year Built |
|---|---|---|---|---|---|---|
| label free | Human Proteome | Consensus | Mixed NCE and eV | 911,783 | 605,677 | 2020 |
| label free | Human Proteome | Selected | NCE and eV | 911,783 | 605,677 | before 2020 |
| label free | Human Proteome (Synthetic) | Selected | NCE | 696,692 | 188,805 | 2017 |
| label free | Human (Phospho) | Selected | NCE and eV | 66,922 | 27,400 | 2019 |
| label free | Human (Skin and Hair) | Consensus | NCE and eV | 27,971 | 20,131 | 2021 |
| label free | Mouse CPTAC Tumor | Selected | NCE | 17,851 | 10,026 | 2014 |
| label free | Chinese Hamster Ovary | Selected | NCE and eV | 158,026 | 74,509 | 2018 |
| iTRAQ-4 | Human CPTAC Tumor | Selected | NCE | 1,201,632 | 390,009 | 2014 |
| iTRAQ-4 | Human CPTAC Phospho | Selected | NCE | 223,340 | 67,533 | 2014 |
| TMT-10 | Human CPTAC Tumor | Selected | NCE | 597,548 | 386,224 | 2019 |
| iTRAQ-4 | Mouse CPTAC Tumor | Selected | NCE | 91,068 | 42,095 | 2014 |
| iTRAQ-4 | Mouse CPTAC (Phospho) | Selected | NCE | 15,746 | 7,026 | 2014 |

Note: A total of > 5 million HCD spectra are included in these libraries with each spectrum uniquely defined by sequence, modification, charge, and collision energy. Libraries vary in how they are constructed, and the underlying data used; some with chemical labeling, others rely on label-free spectra, and they can be created using consensus or selected methods.

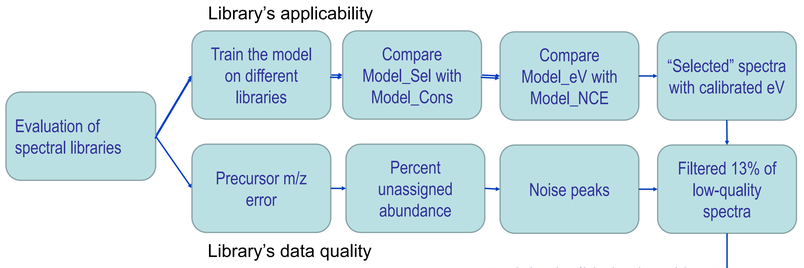


**Figure 1.** Data preprocessing strategies are used for improving the spectral data's relevancy and consistency, resulting in a unified dataset suitable for training, validation, and testing in deep learning.

Training/Validation/Tesdting Datasets: four "selected"-type, high-quality label-free HCD spectral libraries consisting of ~1.8 million tandem mass spectra of ~ 800,000 peptides

**2. Implementation of a deep learning attention model architecture.** We present an attention-based deep neural network designed to predict the intensities of a wide range of fragment ion series generated under higher energy collision-induced dissociation (HCD) conditions, intended for peptide and protein identification in shotgun proteomics experiments.

A total of 3,143 unique HCD fragment ions from 7,919 isotope peaks

| Ion types | Counts |
|---|---|
| a-ions | 226 |
| b-ions | 2026 |
| y-ions | 3643 |
| parent ions | 131 |
| internal ions | 1836 |
| immonium/others | 57 |

Note: Including neutral loss and isotopic peaks.

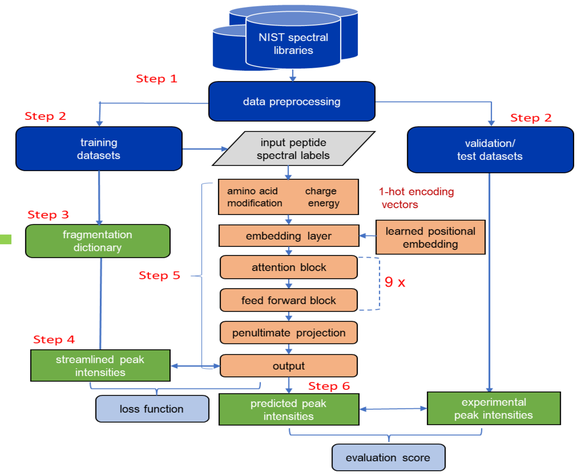


**Figure 2.** The six-step workflow involves (1) data preprocessing, (2) splitting all data into training, validation, and testing datasets, (3) creating a fragmentation dictionary, (4) transforming data into streamlined peak lists, (5) building and training the model, and (6) predicting MS/MS spectra on the evaluation data (validation/test/external sets).

**3.1. Prediction of tandem mass spectra of unmodified and covalently modified peptides.** We provide the CSS scores for the five major peptide classes produced by trypsin digestion. In general, modified peptides add more complexity to the structure and fragmentation than unmodified peptides. Overall, all score differences between Class 1 and the others are within 0.04 for the training and test data, showing that the model can handle a wide variety of different peptides. (Table 2).

| Class | Description | TestCommon | TestUniq |
|---|---|---|---|
| 1 | Tryptic_unmodified | 0.964 | 0.947 |
| 2 | Tryptic_modified | 0.942 | 0.933 |
| 3 | Miscleaved_unmodified | 0.936 | 0.919 |
| 4 | Miscleaved_modified | 0.924 | 0.905 |
| 5 | semitryptic | 0.925 | 0.918 |

**Table 2.** The median CSS was calculated for each peptide class in TestCommon (most similar to the training data) and TestUniq (dissimilar to the training data).

**3.2. Prediction of complex HCD fragment ion types.** Predictions based on the validation/test sets demonstrated that our model is able to predict various fragment ion types such as (1) sequence a/b/y ions, (2) neutral losses, (3) internal fragments, (4) immonium ions and side chain fragments, and (5) precursor and related ions. The contribution of each of these ion types to the overall intensity coverage is illustrated in Fig. 3 and an example of predicted spectrum is displayed in Fig. 4.

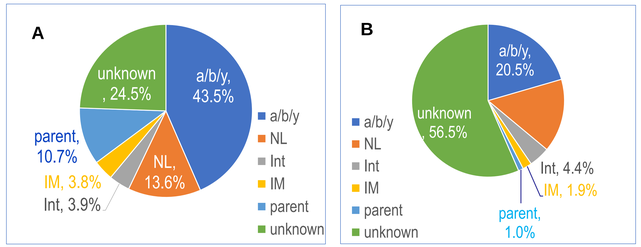


**Figure 3.** Predicted intensity coverage of the six major ion types in HCD spectra in the validation set of 12076 spectra. Note that the remaining 24.5% intensity in the figure is that from unknown abundance contribution, collectively constituting potentially unannotated, contaminant, and noise peaks. **A.** Relative percentage of Ion Abundance, and **B.** Relative percentage of ion counts.

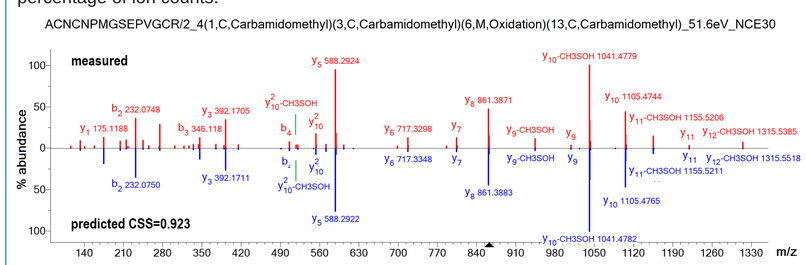


**Figure 4.** Head-to-tail plot showing a predicted peptide HCD spectrum (bottom, blue) of a doubly charged ACNCNPMGSEPVGCR with three Carbamidomethyl cysteine sites and oxidized methionine at 51.6eV/NCE30, and the matching experimental spectrum (top, red).

**4. Benchmarking proteome-scale in-silico spectral libraries on a HeLa dataset.** We tested our predicted libraries on reanalyzing a human HeLa dataset (PXD022287). We compared the results of our in-silico library for peptide identification to other popular data analysis software tools, such as MS-GF+ (i.e., sequence search) and the MS PepSearch/NIST Human Library (i.e., spectral search).

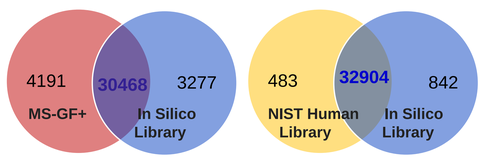


**Figure 5** shows that our in-silico libraries exhibit excellent performance on this dataset, yielding 30,468 (87.9 %) and 32,904 (98.6 %) overlapping identifications with MS-GF+ and the NIST libraries, respectively.

## Conclusion

In this work, we developed an advanced deep neural network system, that excels in predicting a wide array of complex HCD MS/MS fragment ion types. Our model system was demonstrated to accurately predict all known fragment ion types, including sequence a/b/y ions, neutral losses, internal fragments, immonium ions, amino acid side chain fragments, and precursor ions. This approach ultimately broadens the applicability of current predictive models to a wide array of different peptides and modified peptides involved in complex proteomics studies.

**References**

1. NIST Libraries of Peptide Mass Spectra, https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload.