Spetrum Prediction

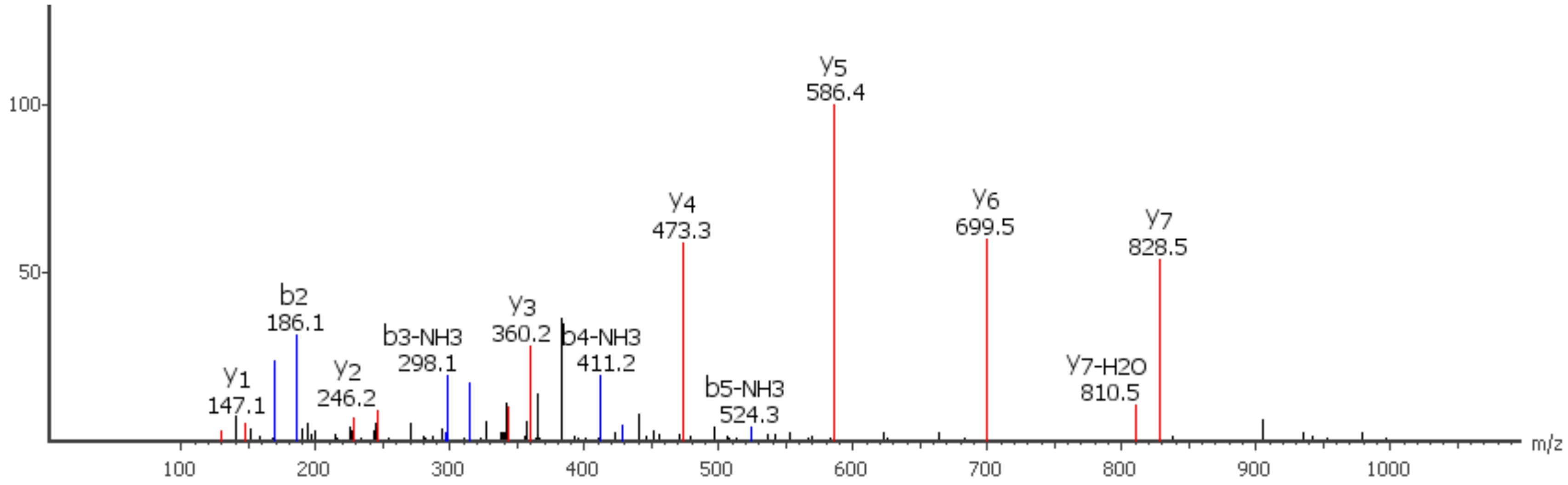
Spectrum Prediction

ANELLLNVK

MS/MS

Peptide Identification

Spectrum Prediction



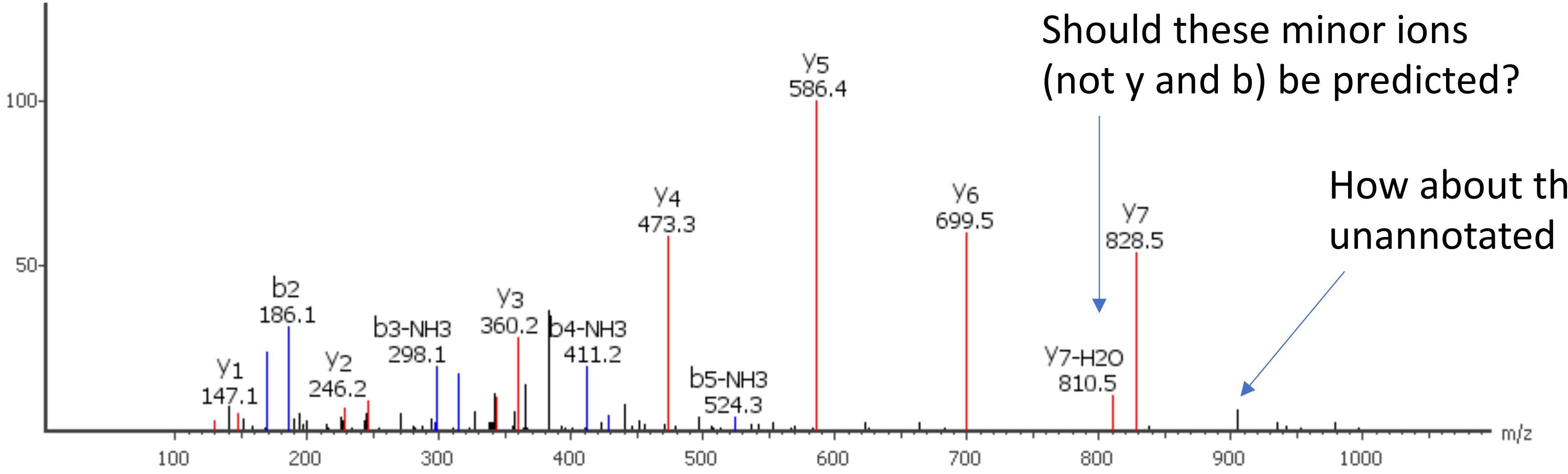
Spectrum Prediction

ANELLLNVK

Spectrum Prediction

Two lines of research:

- To predict only the main fragment ions (e.g. b and y ions).
- To predict the full spectrum.

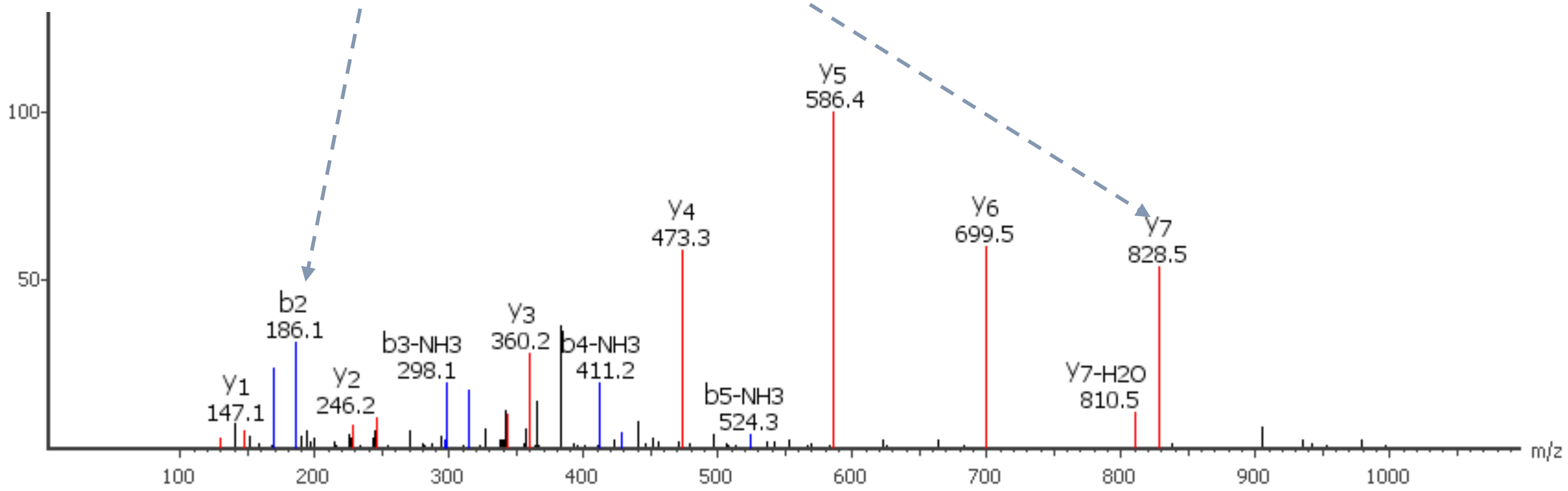


Should these minor ions (not y and b) be predicted?

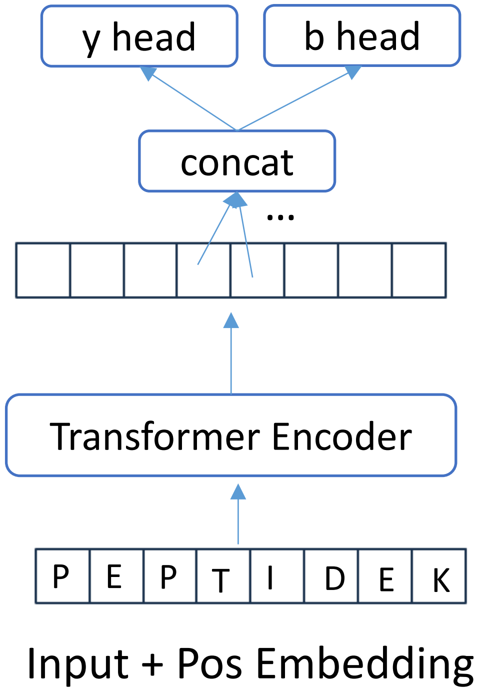
How about these unannotated ones?

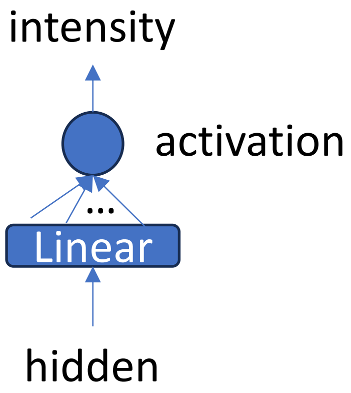
Predict the y/b Ion Intensity

b_1	**A NELLLNVK**	y_8
b_2	**AN ELLLNVK**	y_7
b_3	**ANE LLLNVK**	y_6
b_4	**ANEL LLNVK**	y_5
b_5	**ANELL LNVK**	y_4
b_6	**ANELLL NVK**	y_3
b_7	**ANELLLN VK**	y_2
b_8	**ANELLLNV K**	y_1



A Straightforward Model



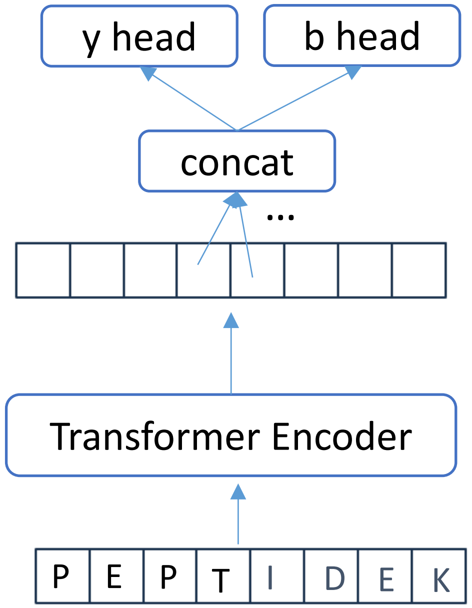


output head

Label = Normalized intensity

Loss = MSE (Mean Square Error)

Possible Options



y head

b head

concat

...

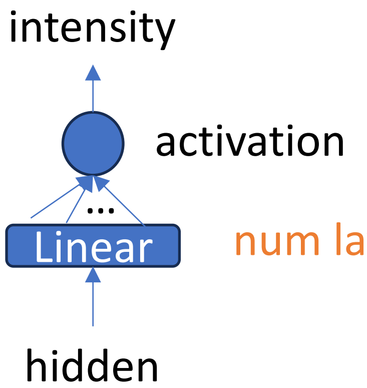
Transformer Encoder

P E P T I D E K

Input + Pos Embedding

num layers
d_model

Encode meta info (e.g. precursor charge)



intensity

activation

...

Linear

hidden

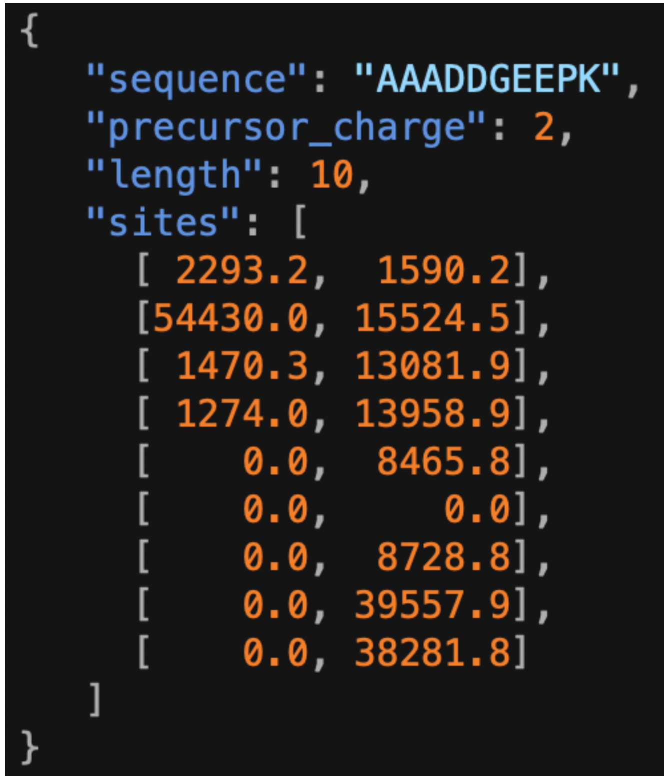
num layers

output head

Use a transformer decoder as well?

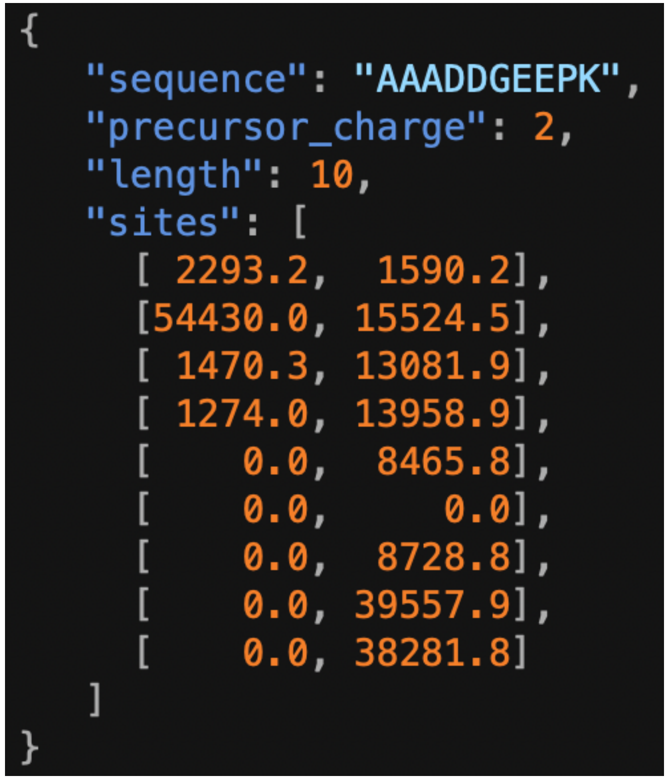
Training Suggestions

- Training data: NIST peptide library (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload)
- Extract data in an intermediate format first (easier for repeated training and manual checking).
- Visualize some results to ensure on the right path.

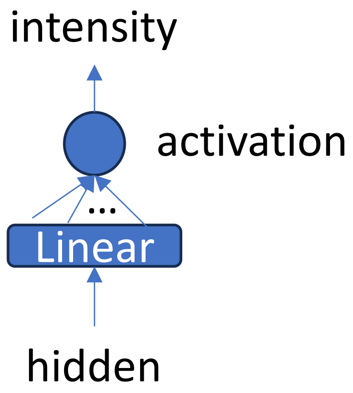


```
{
   "sequence": "AAADDGEEPK",
   "precursor_charge": 2,
   "length": 10,
   "sites": [
     [ 2293.2,  1590.2],
     [54430.0, 15524.5],
     [ 1470.3, 13081.9],
     [ 1274.0, 13958.9],
     [    0.0,  8465.8],
     [    0.0,     0.0],
     [    0.0,  8728.8],
     [    0.0, 39557.9],
     [    0.0, 38281.8]
   ]
}
```

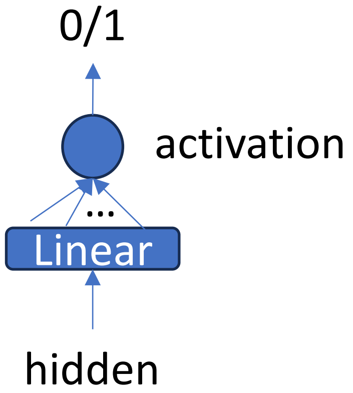
Dealing with Large number of 0



```
{
   "sequence": "AAADDGEEPK",
   "precursor_charge": 2,
   "length": 10,
   "sites": [
     [ 2293.2,  1590.2],
     [54430.0, 15524.5],
     [ 1470.3, 13081.9],
     [ 1274.0, 13958.9],
     [    0.0,  8465.8],
     [    0.0,     0.0],
     [    0.0,  8728.8],
     [    0.0, 39557.9],
     [    0.0, 38281.8]
   ]
}
```



output head



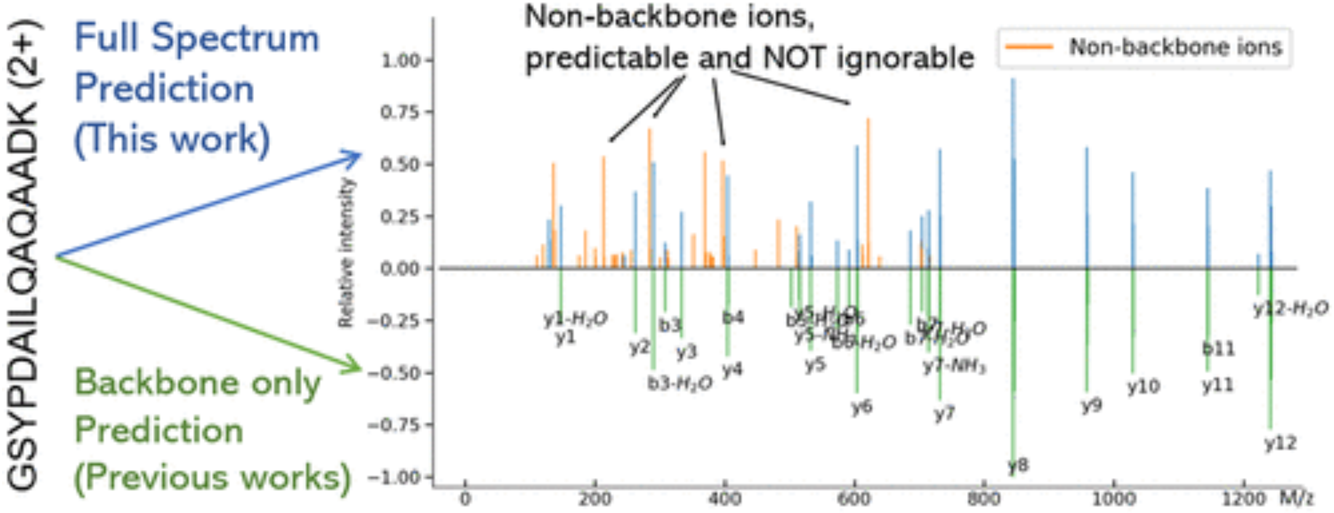
- Output head 2 predicts whether peak exists.
- Loss = BCE + MSE at nonzero positions.

Litearture

- Transformer model to predict y/b ions:
  - Ekvall *et al*. Prosit Transformer: A transformer for Prediction of MS2 Spectrum Intensities. Journal of Proteme Research. 2022.
- CNN model to predict full spectrum:
  - Liu *et al.* Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network. Anal. Chem. 2020, 92, 6, 4275–4283

Full Spectrum Prediction

- *Liu et al.* **Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network**. *Anal. Chem.* 2020, 92, 6, 4275–4283

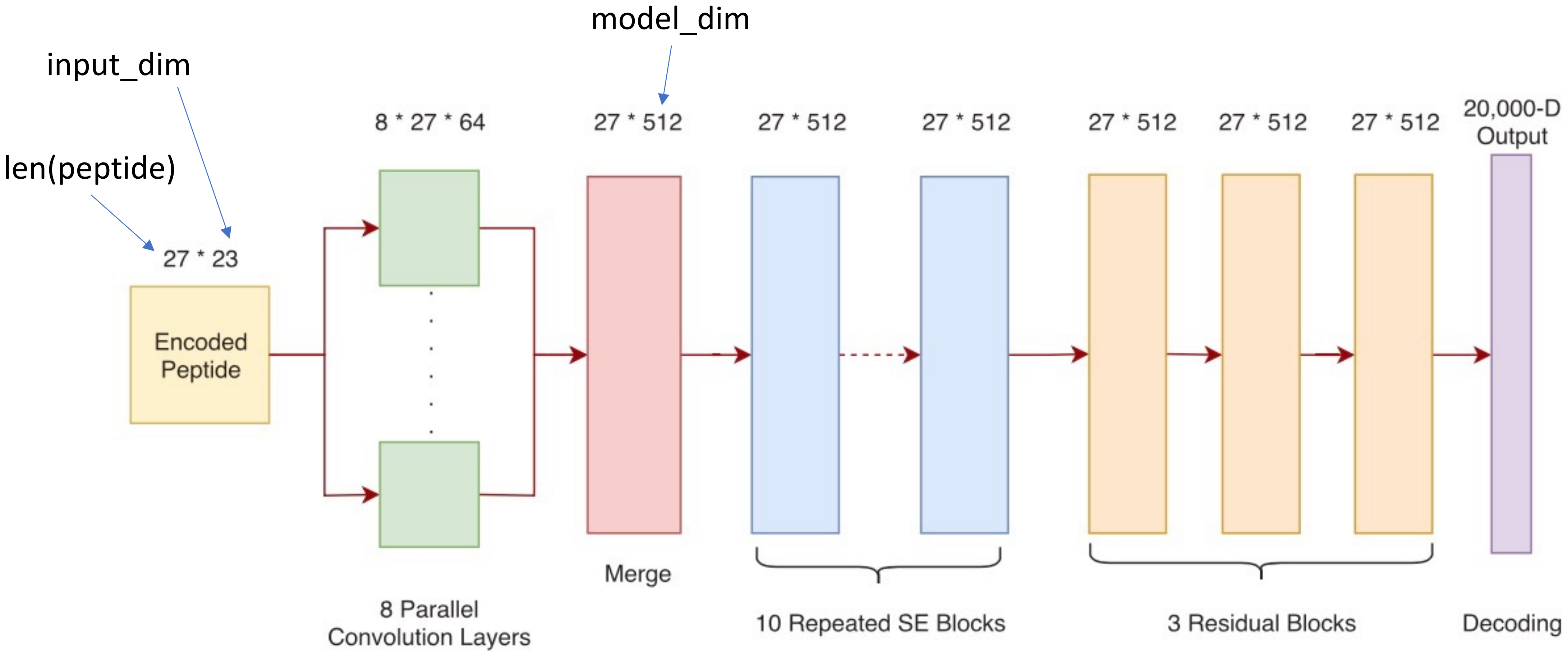


Liu et al. **Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network**. *Anal. Chem.* 2020, 92, 6, 4275–4283

Output Format

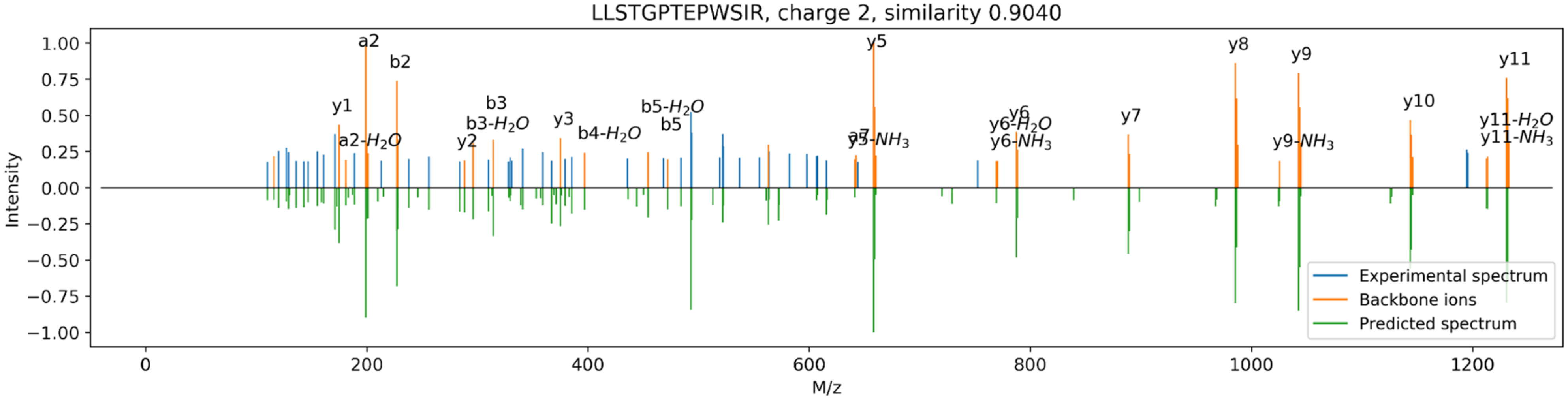
- A spectrum usually has a limited m/z range. The work only predicts peaks between 180-2000.
- The spectrum is represented by a sparse one-dimensional (1-D) vector by binning the m/z range between 0 and 2000 with a given bin width. The value stored is the peak intensity.
- With 0.1 Da bin width, a spectrum becomes a 20,000-dimension vector. The value in each bin is the relative intensity of the tallest peak in the bin. (Note: most dimensions have value 0).
- Prediction is to predict the values in all the bins.

CNN Architecture for Spectrum Prediction

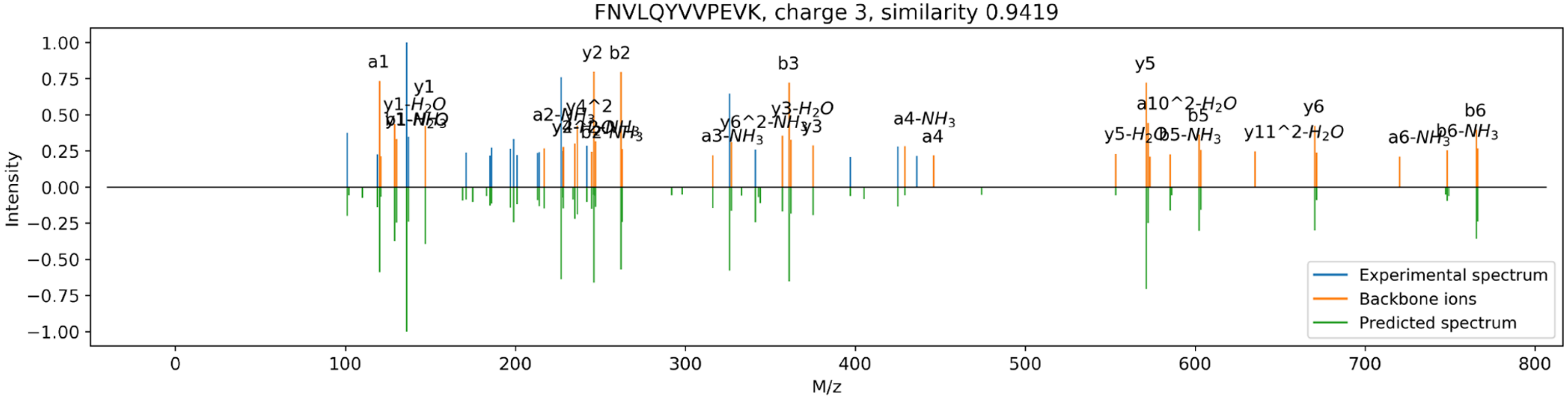


- The 8 parallel convolution layers use kernel size 2 to 9, respectively.
- SE (Squeeze-and-Excitation) blocks are another type of commonly used CNN building blocks.

Predicted Examples

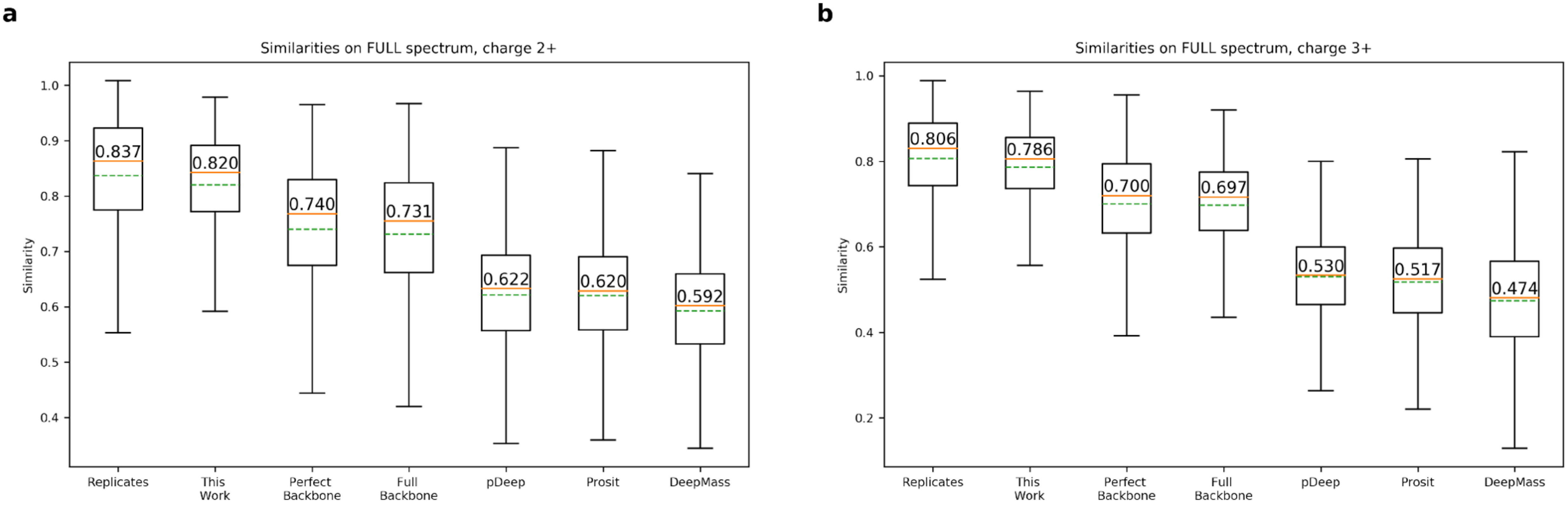


(a)

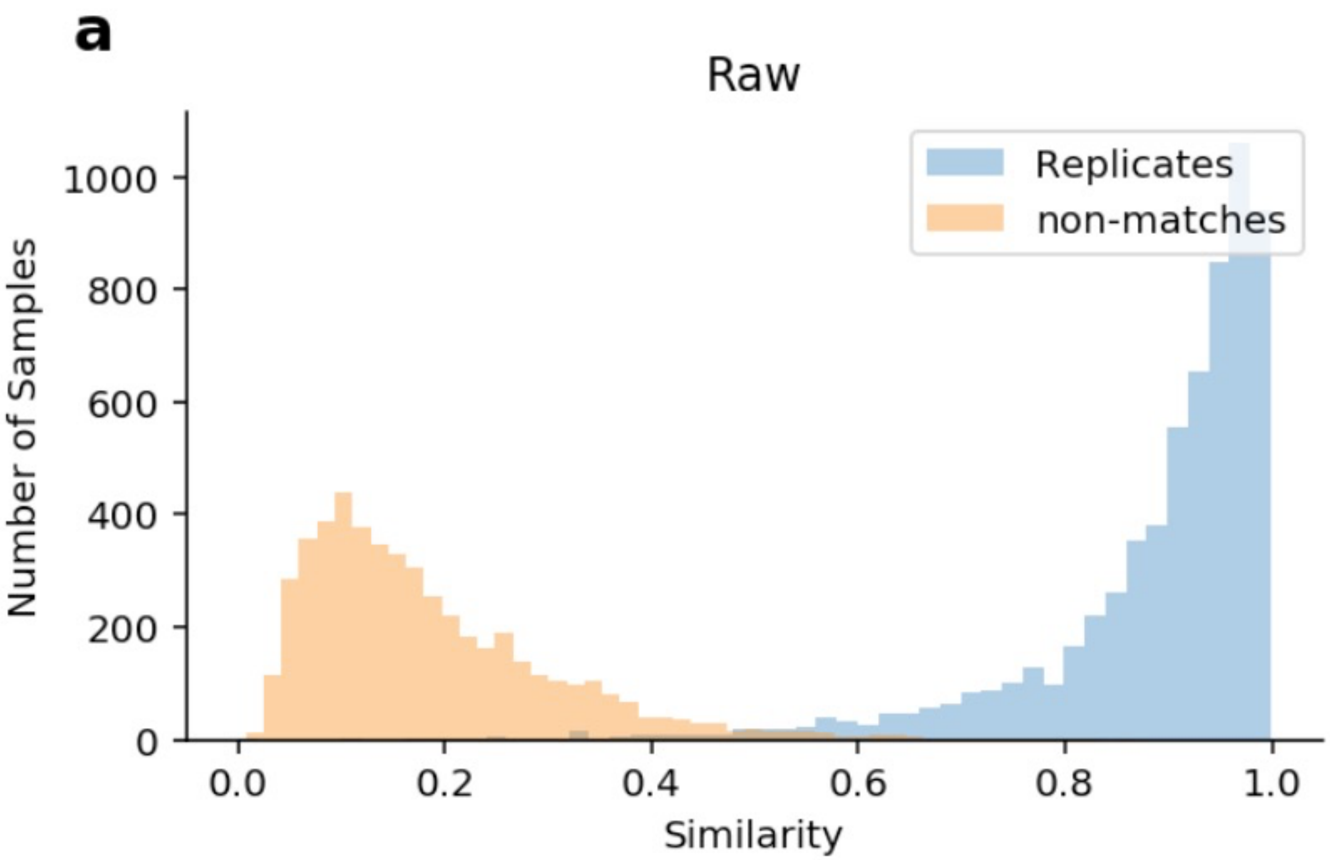


(b)

Comparison to Other Tools



Spectrum Prediction Helps Peptide Identification



If peptide is correct, then its predicted spectrum should match the experimental one with high similarity. Whereas the wrong peptides should have low similarity.