Supporting Information

Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network

Kaiyuan Liu, Sujun Li, Lei Wang, Yuzhen Ye, and Haixu Tang*

School of Informatics, Computing, and Engineering, Indiana University,
Bloomington, Indiana 47405, United States
hatang@indiana.edu

Contents

1 Data Selection

High-quality training data is critical for achieving good performance. Although we do not have a gold standard to distinguish true and false PSMs, we may filter out suspicious PSMs to retain a more promising training set. In this study, we removed all spectra containing fewer than 20 peaks (i.e., under-fragmented) or more than 500 peaks (i.e., over-fragmented). Besides, all PSMs with precursor mass mismatched more than 200 ppm are removed. We also exclude PSMs with peptide length greater than 25 or precursor mass greater than 2000 m/z, as those spectra are relatively rare (e.g., less than 4 percent in our collected HCD spectra dataset).

2 Implementation and Training

The CNN model was implemented in Python using the Keras framework Chollet et al. (2015) with Tensorflow **?** back-end. We also implemented a standalone software named *PredFull* for predicting HCD spectra of given input peptide sequences. The software is released open-source on Github at `https://github.com/lkytal/PredFull` and can also be accessed through a web service at `http://www.predfull.com/`. We also share the whole training and testing set at `http://www.predfull.com/datasets`, including the raw experimental spectra, as well as the predicted spectra of the testing peptides in these datasets.

The model was trained by Adam Kingma and Ba (2015) optimizer at a learning rate of 0.0003, with a batch size of 1024. The training process spans 50 epochs (Supplementary Figure S8), while the learning rate will be decay to 5×10^{-5} at the 30th epoch and 1.25×10^{-5} at the 40th epoch. The training process takes around 12 hours ($\sim 7 \times 10^{-4}$ second per sample) using two NVIDIA GTX 1080ti GPUs, while the prediction takes $\sim 10^{-3}$ second per peptide.

3 Running other Predictors

For pDeep, we download and execute their Github release (`https://github.com/pFindStudio/pDeep/tree/master/pDeep2`) for prediction, setting NCE to 30% and the instrument to QE. For the extended pDeep version, we re-implement it using Keras following the structure described by the original paper Zhou et al. (2017), but extend the model to predict additional backbone ions (including a/x/c/y ions and their neutral loss derivatives) as well. We then train the model with the same training set as this work, using Adam optimizer at a learning rate of 0.0002.

For Prosit, we simply download the Github source code `https://github.com/kusterlab/prosit` for prediction. For DeepMass, we use the Github scripts to pre-process (`https://github.com/verilylifesciences/deepmass/tree/master/prism`) the input and the processed data was sent to their Google Cloud engine (as instructed in their Github pages) for spectrum prediction.

4 Supplementary Figures

4.1 Figure S1

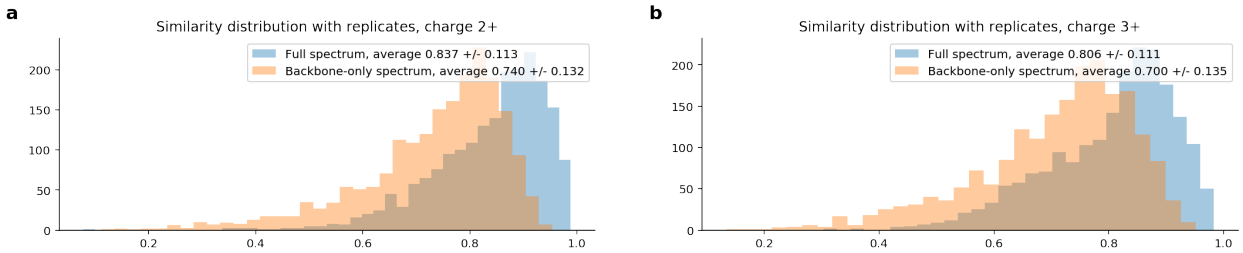
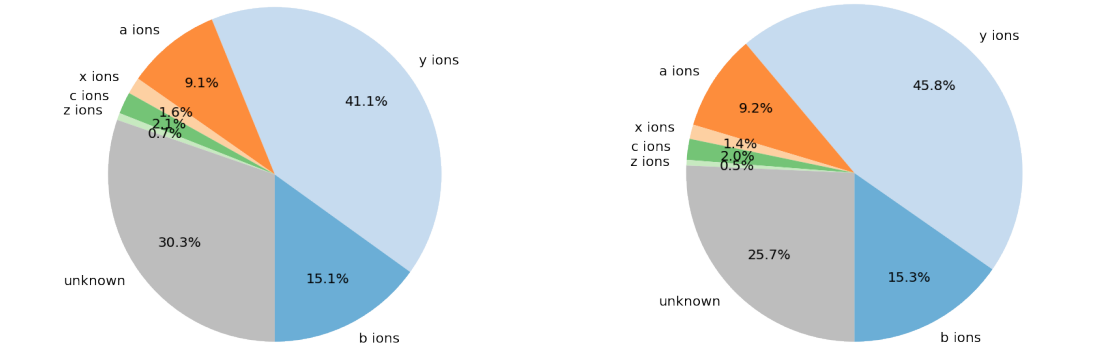
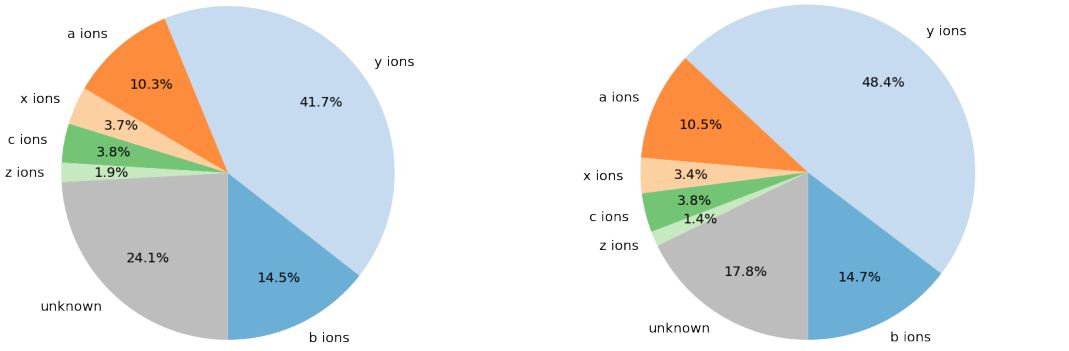


Figure S1: Similarity distribution of full spectrum or backbone-only spectrum with its replicates (a) similarity distribution of charge 2+ HCD spectra (b) similarity distribution of charge 3+ HCD spectra.

4.2 Figure S2



(a)



(b)

Figure S2: Intensity composition of fragment ion types in experimental (left) versus predicted (right) HCD spectra for 2+ (a) and 3+ (b) precursor ions.

4.3 Figure S3

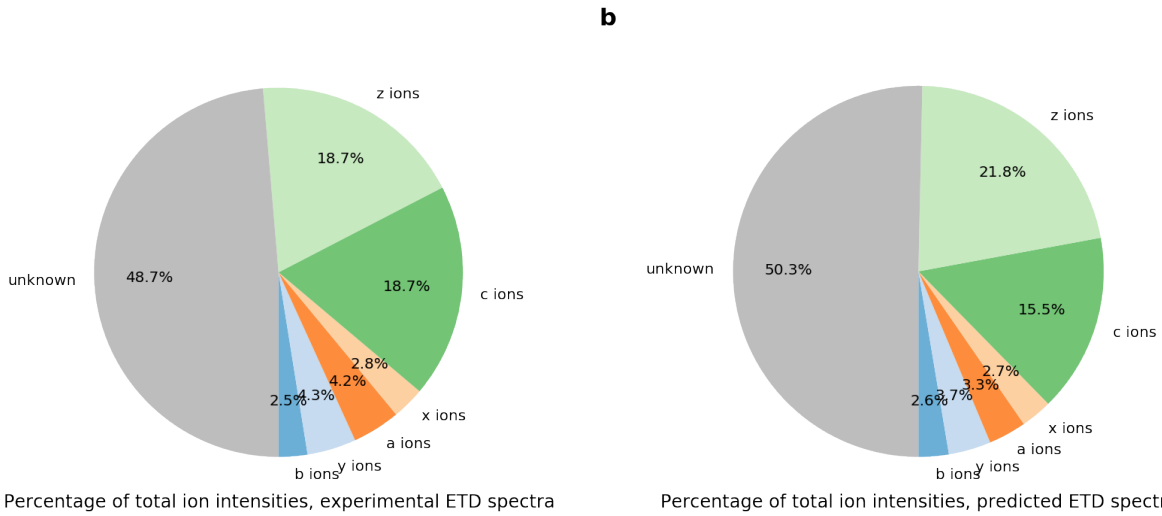


Figure S3: Average intensities of different fragment ions in experimental (a) and predicted (b) ETD spectra of charges 1+ to 4+.

4.4 Figure S4

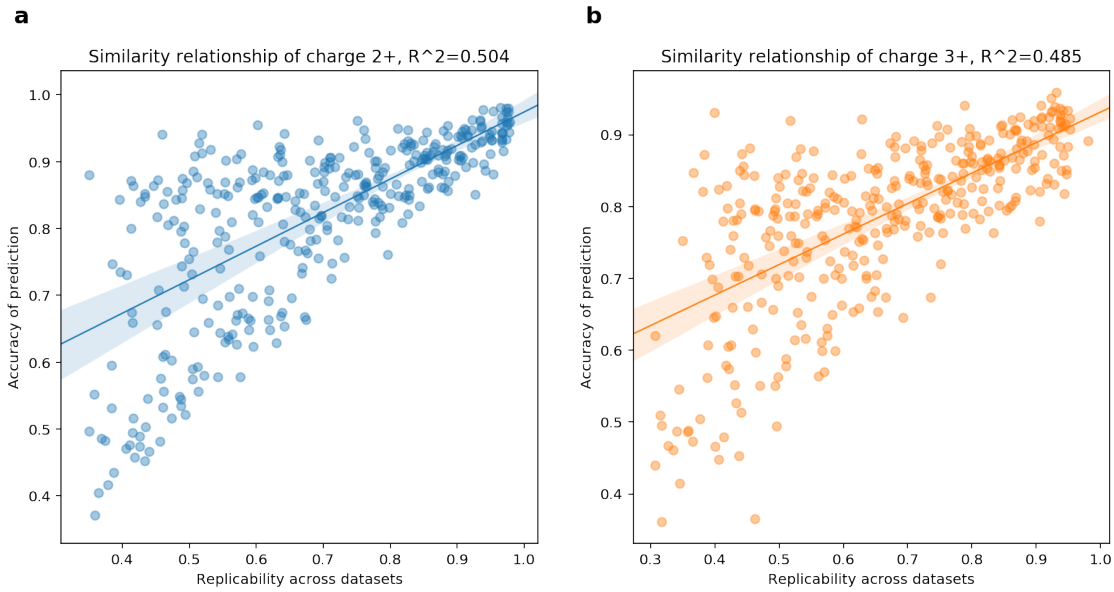
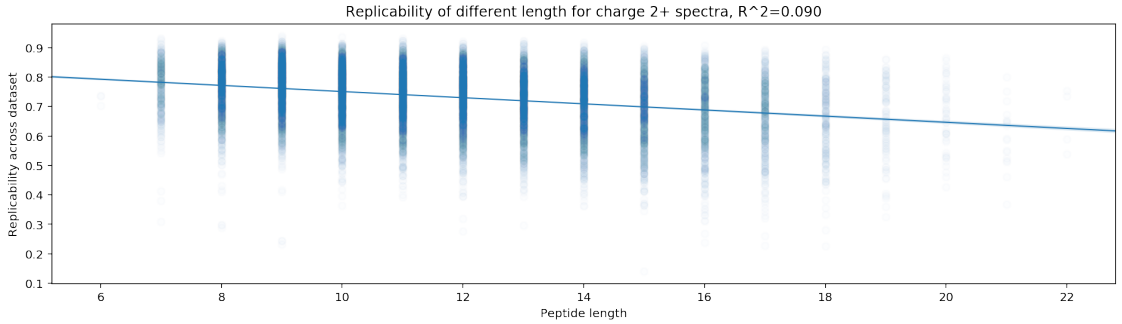
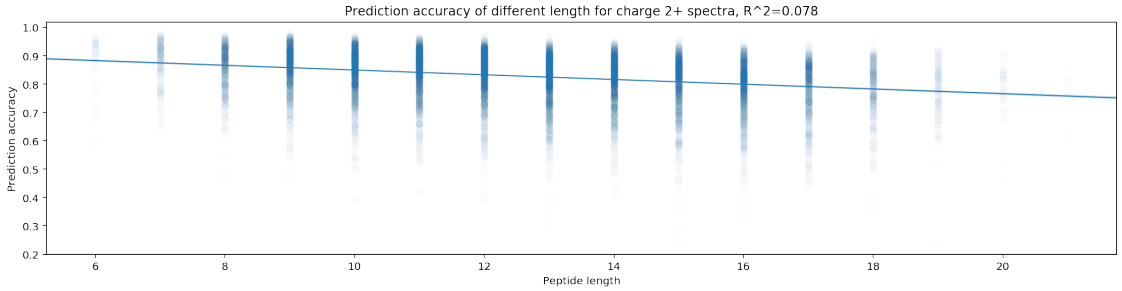


Figure S4: The accuracy of predicted spectra is highly correlated with similarity between replicated spectra across experiments for the same peptides. **a**, Relationship of charge 2+ spectra. **b**, Relationship of charge 3+ spectra.

4.5 Figure S5



(a)



(b)

Figure S5: **a**, The similarities between the experimental and predicted HCD spectra decrease with the increasing peptide length. The statistics were conducted over 10,000 HCD spectra of charge 2+. **b**, The similarities between replicated HCD spectra decrease with the increasing peptide length. The statistics ware conducted over 10,000 randomly sampled experimental HCD spectra of charge 2+.

4.6 Figure S6

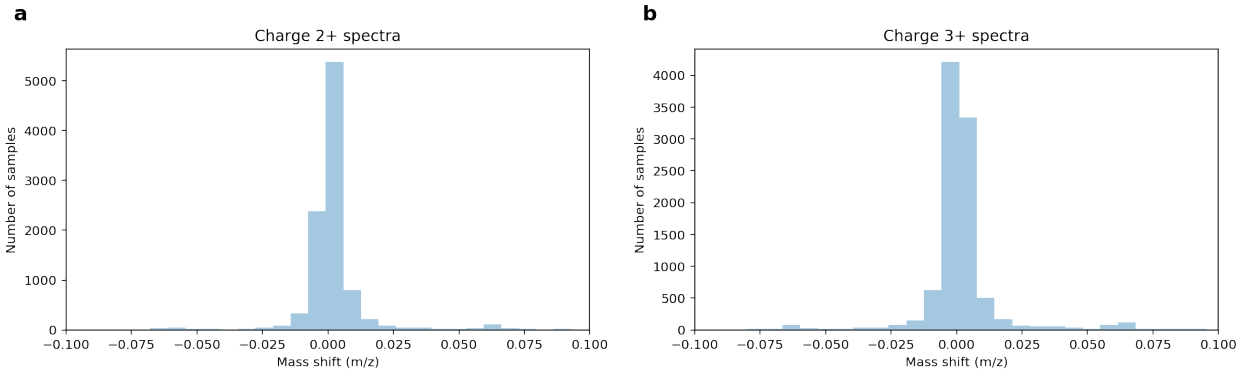


Figure S6: The distribution of m/z shifts between replicated HCD spectra of charge 2+ (a) and 3+ (b). Both statistics were conducted over 10,000 HCD spectra of charge 2+ and charge 3+.

4.7 Figure S7

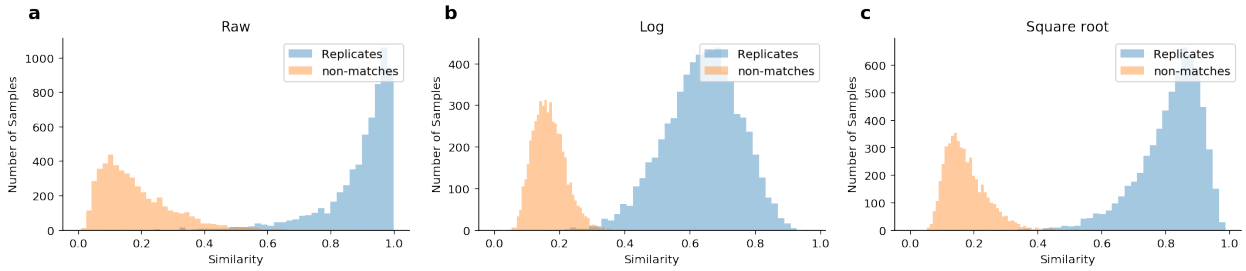


Figure S7: The distributions of similarities between the replicated experimental spectra of the same peptides (blue) versus those between two distinct peptides with the same precursor mass (yellow) when different normalization functions were applied to the intensities of fragment ions. The statistics ware conducted over 5,000 randomly sampled HCD spectra of charge 2+. **a**, Original intensities. **b**, Intensities transformed by *Log*. **c**, Intensities transformed by square root.

4.8 Figure S8

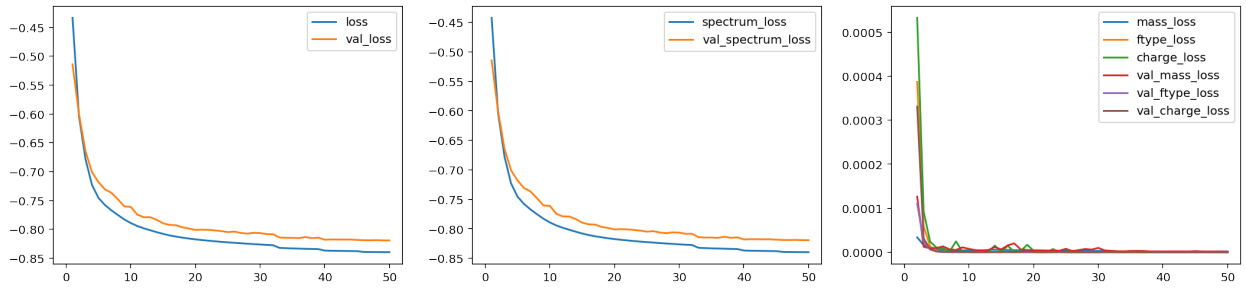


Figure S8: The decreasing of the losses (y-axis) on the training and testing data along with the training history (x-axis: number of epochs). **Left panel**, Total loss; the training and testing losses are close, indicating the model does not over-fit to the training data. **Center panel**, The loss of the spectra prediction task. **Right panel**, Other losses of auxiliary tasks, which quickly drops to nearly zero as expected.

References

Chollet, F., et al. Keras. `https://keras.io`, 2015.

Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). 2015.

Zhou, X.-X.; Zeng, W.-F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S.-M.; Zhang, Z. pdeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical chemistry* **2017**, *89*, 12690–12697.