# Prosit Transformer: A transformer for prediction of MS2 spectrum intensities

Markus Ekvall[1] Wassim Gabriel[2] Mathias Wilhelm[2] Lukas Käll[1]

October 13, 2021

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology − KTH, Box 1031, 17121 Solna, Sweden
[2]Computational Mass Spectrometry, Technical University of Munich (TUM), Freising, Germany

**Abstract**

Machine learning has been an integral part of the interpretation of data from mass spectrometry-based proteomics for a long time. Relatively recently, a machine-learning structure appeared that has been successful in other areas of bioinformatics, Transformers. They have been proven particularly useful for transfer learning, i.e., adapting networks trained for other tasks to new functionality with fewer training examples than it otherwise would require.

Here, we implemented a Transformer based on the pre-trained model TAPE to predict MS2 intensities. TAPE is a general model trained to predict missing residues from protein sequences. Despite being trained for a different task, we could modify its behavior by adding a prediction head at the end of the TAPE model and fine-tune it using the spectrum intensity from the training set to the well-known predictor Prosit.

We demonstrate that the predictor, which we call Prosit Transformer, outperforms the recurrent neural network-based predictor Prosit, increasing the median angular similarity on its hold-out set from 0.908 to 0.929.

We believe that transformers will significantly increase prediction accuracy for other types of predictions within mass spectrometry-based proteomics.

# Introduction

Just as in many other areas involving analysis of large and complex datasets, modern analysis of mass spectrometry-based proteomics data is tremendously helped by different types of machine learning [10,13]. For example, we nowadays can use machine learning to predict tryptic digestion [20], chromatographic retention time [9, 11, 14], collisional cross-section [12], the accuracy of peptide-spectrum matches [8], the accuracy of transitions in DIA data [4], and de Novo Interpretation of spectra [17] are tasks that utilize machine learning.

One task that has gained traction in the last couple of years is predicting MS2 spectra from peptide sequences [3, 6]. Such predictors can predict relative intensities of the $b$- and $y$-ions of a given peptide sequence. Together with the m/z values of the ions, which one can derive from first principle, one can subsequently form a full MS2 spectrum. MS2 spectrum prediction has in a short time established itself as a means to rescore peptide spectrum matches [2], increase the sensitivity in large search spaces [19], and target-decoy strategies for DIA interpretation [16].

Many types of frameworks are available for training a predictor, such as Support Vector Machines and Recurrent Neural Networks (RNNs) used within mass spectrometry-based proteomics. However, in the last couple of years, a structure first in natural language processing [5] known as Transformers [18], has successfully been employed within other areas of bioinformatics, e.g., structure prediction [1, 15].

Transformers are, like RNNs, designed to handle sequential input data and do so through attention mechanisms, i.e., mechanisms that enhance the essential parts of the input sequence for its output. However, unlike RNNs, the Transformers do not use recurrence, thus enabling a significant speed-up by parallelizing their training. Transformers are based on an encoder-decoder structure, where both the encoder and decoder adopt the multi-headed attention mechanism [18].

Particularly, the Tasks Assessing Protein Embeddings (TAPE) model [15] is exciting; a Transformer-based autoencoder of protein sequences that is formed by withholding one amino acid at a time in a large set of protein sequences and subsequently predicting which is the missing amino acid. This model can subsequently be employed for higher-level tasks by plugging them into some extra layers of neurons in a process known as transfer learning [1, 15].

Here, we argue that Transformers can be a great aid within mass spectrometry-based proteomics. We demonstrate that a TAPE, can be used for the prediction of MS2 spectrum intensities from peptide sequences. We are using the training and test sets of the popular Prosit [6] predictor, and demonstrate that the transformer-based predictor, which we named Prosit Transformer, drastically outperforms the old implementation of Prosit.

# Methods

## Data

We downloaded the Prosit training data from `https://figshare.com/projects/Prosit/35582`. The Prosit data had to be converted from HDF5 to LMDB to be compatible with the Tape framework. The LMDB data files used during training and validation are accessible at `https://figshare.com/articles/dataset/LMDB_data_Tape_Input_Files/16688905`.

## Architecture

The TAPE model consists of twelve 768 hidden units attention layers, both with the attention dropout (DropHead) rate [21] and regular dropout rate set to 0.1. The Prosit-specific transformer has the same parameter but consists of 9 attention layers. The meta-data layer is a multilayer perceptron (MLP) with two layers of size 512 units followed by a dropout rate of 0.1 each. The final prediction layer has the same structure, except for no dropout after the final layer. The activation function is ReLU, except for the prediction layer where the first layer uses a ReLU6 [7] i.e. a max(0,min(6,x)) function as an activation function, and the final layer uses a linear layer.

## Metrics

We measure the accuracy of the predicted intensities with the angular similarity, which is defined as $1 - \frac{2}{\pi}\cos^{-1}\left(\frac{A \cdot B}{(||A|| \cdot ||B||)}\right)$. Here $A$ is the vector of predicted intensities and $B$ is the vector of predicted intensities for the ion series included in the prediction. However, we had to introduce a few changed during training to avoid undefined behavior. Firstly, to avoid undefined values using angular similarity during training, we had to clip the inputs to $\cos^{-1}$ with at $-(1-\epsilon)$ and $(1-\epsilon)$ to avoid getting undefined values. This implementation was necessary since some predictions were too similar to their target after training, resulting in an undefined loss. However, there was no clipping during the evaluation, so it will not affect the final result. Lastly, we also had to introduce a small $\epsilon$ in the denominator in the cosine similarity, i.e., $\max(||A||||B||, \epsilon)$ to assure no undefined behavior during training.

As a measure of the number of erroneous peak predictions, we calculated the $FDR = FP/(FP + TP)$ and $FNR = FN/(FN + TP)$ for each predicted spectrum. Here FP is the number of peaks predicted to be present in a spectrum that was absent in the observed spectrum, $FN$ is the number of peaks predicted to be absent in a spectrum that was present in the observed spectrum, and $TP$ is the number of peaks predicted to be present in a spectrum that was present in the observed spectrum.

## Post-processing of Predicted intensities

For the final result, we use the same post-processing on the predicted spectrum used in Prosit [6]. We set ions with a predicted negative intensity to zero, i.e., a negative intensity indicates an absent peak. Furthermore, we set all ion's intensity that's not obtainable for any given peptide due to too low a charge state or too low peptide length to -1, i.e., the same value used in the experimental data.

## Hardware

The model was trained on the Berzelius SuperPOD, a GPU cluster consisting of 60 NVIDIA DGX A100 systems, linked on a 200 Gbit/second NVIDIA Mellanox InfiniBand HDR network.

# Results

We set out to test whether Transformers are a technology fit for spectrum intensity predictions, i.e., to predict the intensities of the most commonly observed ion-series ($b^{+}, b^{2+}, b^{3+}, y^{+}, y^{2+}, and\ y^{3+}$) of product-ion spectra from peptide fragmentation. The length of the peptides ranged between 7 to 30 amino acids long. As a testbed, we selected to use the train/test data and the preprocessing coming with the Prosit predictor. Prosit's scripts calculating the intensity vectors, adopting meta-data, and calculating predictions' angular similarity has been found robust after years of usage. We also found it straightforward to set up a benchmark, as we could reuse the Prosit test sets just out of the box. To avoid confusion, we will refer to the traditional Prosit predictor as Prosit RNN from hereon.

## Model

We set out to use the setup previously used for training and testing the Prosit model but with a transformer. We used the pre-trained encoder/decoder model TAPE [15] and retrofitted it with a Prosit-specific decoder and some additional application-specific code (See Figure 1). The Tape model will encode the peptide into a 512-dimensional embedding. Furthermore, just as for the original RNN-based Prosit model, we used layers for handling meta-data consisting of the charge state of the spectrum and its collision energy (CE). The charge states range from one to six, represented as 6-dimensional one-hot-encoding. Hence, the meta-data layer has seven inputs nodes to account for both the charge state and CE. The meta-layer transforms the metadata into a 512-dimensional vector that is subsequently is combined with the encoded peptide by element-wise multiplication. Then a Prosit-specific Transformer will decode this combined embedding. Lastly, a two-layered Multilayer perception (MLP) follows the decoding layer, serving as a prediction layer to predict the spectrum intensity. The MLP used activation by a hinge loss function constrained between 0 and 6 (a RELU6 function) as activation between the two final layers to avoid so-called gradient explosion. For the training, the objective function was the mean angular similarity between the observed and predicted spectrum intensity vectors.

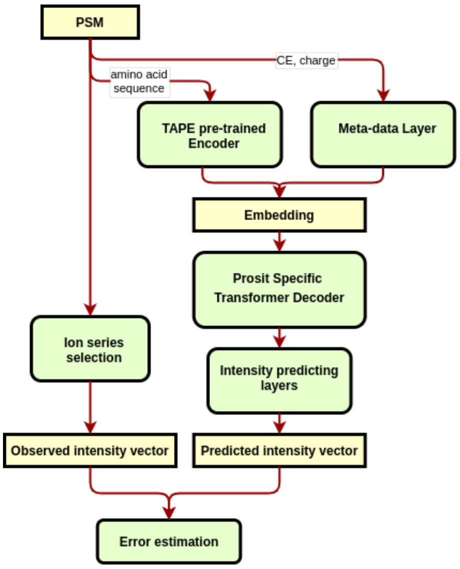

Figure 1: **Architecture of the Prosit Transformer.** The model is dependant on a pre-trained encoder from the TAPE project and uses the design of TAPE for a Prosit Specific decoder. However, our model implements many of the design features of Prosit RNN, i.e., layers handling meta-data and final intensity prediction.

## Training of model

During the training, we used a batch size of 1024, the learning rate of 0.0001, gradient accumulation step of one, and a linear learning rate schedular with 10000 warmup steps. The training proceeded until no further improvement over ten epochs.

To obtain better accuracy in predicting present and absent peaks, we introduced a hyperparameter, $\delta$, setting an artificial offset of the intensities of absent peaks to $\delta_p = \delta/|\text{number of considered peaks}|$. This is to add an extra penalty if the model predicts intensities for absent peaks. By varying the size of $\delta$, we can control the model's propensity to predict peaks as absent, and by such means tune the model's false positive and false negative predictions. We measured the false discovery rate and the false-negative rate of each spectrum and then plotted the average angular similarity, the FDR, and the FNR for different choices of $\delta$. We selected $\delta = 0.34$ for the final training. (See Figure 2)

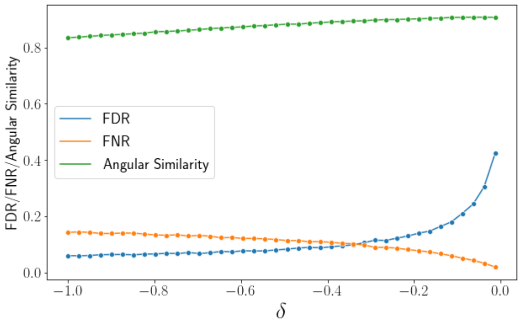

Figure 2: **The effect of adjusting the hyperparameter $\delta$ on the ability to predict absence/presence of individual MS2 peaks.** To obtain better prediction accuracy of present and absent MS2 peaks, we adjusted the intensities of absent peaks from zero to $\delta$. We measured the false discovery rate (FDR) and the false-negative rate (FNR) of each spectrum and then plotted the average angular similarity, the FDR, and the FNR for different choices of $\delta$. We selected $\delta = 0.34$ for the final training. The predicted spectra were not post-processed for the measurements in this figure (See Methods).

## Test of performance

To test the performance of our final Prosit Transformer, we investigated its performance on the same held-out test set as used when initially training Prosit RNN. We calculated the so-called angular similarity between the predicted and observed intensities for both predictors. Overall, we see that the predictions from Prosit Transformer have a higher angular similarity than Prosit RNN and are hence more accurate (Figure 3A). The Prosit Transformer increased the median angular similarity from Prosit RNN's 0.908 to 0.929. We also see that Prosit Transformer obtained a higher angular similarity than Prosit RNN in 75.7% of the spectra, while the opposite was true in 24.3% of the spectra. The same pattern was also true when dividing the PSMs based on their peptide's lengths (Figure 3B). We also wanted to compare the predictors' ability to predict present and absent (zero intensity) fragment peaks. Our choice of hyperparameter delta for Prosit Transformer resulted in a lower fraction of observed absent peaks among the predicted non-zero intensity peaks (Figure 3C) while observing a higher fraction of predicted absent peaks among the observed non-zero intensity peaks (Figure 3D) for Prosit Transformer compared to Prosit RNN.

## Prosit Transformer's ability to model collision energy

We also wanted to test that the improved ability of Prosit Transformer to predict ms2 intensities did not affect the predictor's ability to model collision energy's (CE's) influence on predicted spectra. We hence isolated batches of spectra with CE=0.2, 0.25, 0.3, 0.35, 0.4 and measured the median Angular Similarity when predicting the spectra for a range of different collision energies (Figure 4).

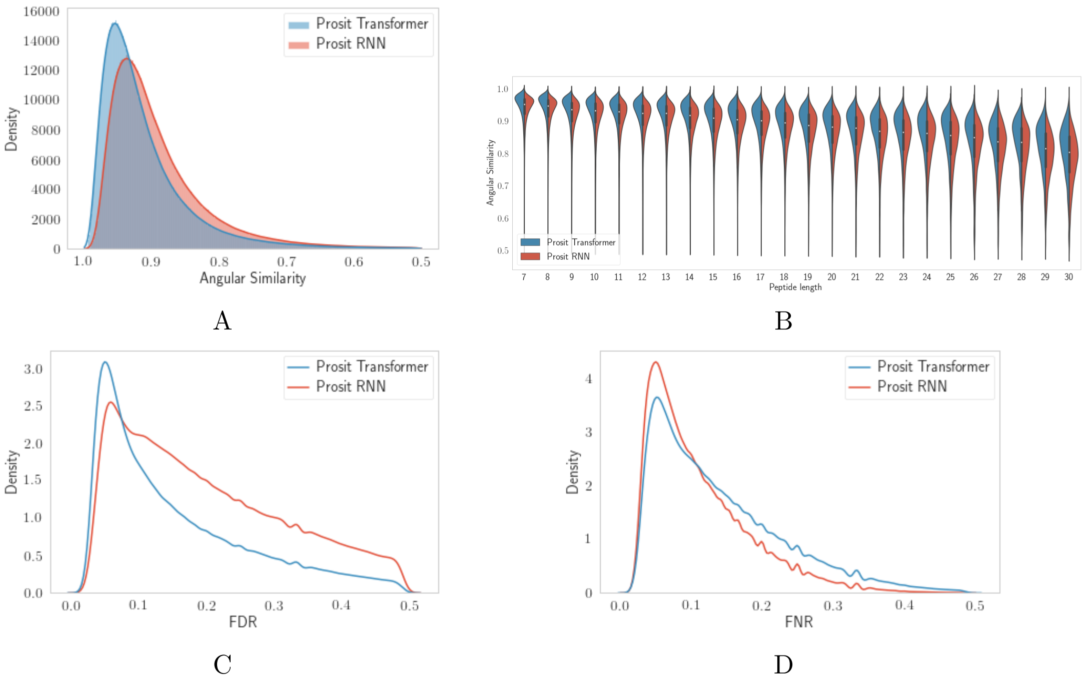

Figure 3: **Comparison of the accuracy of Prosit Transformer and Prosit RNN.** (A) We made separate histograms and smoothed them with a kernel density estimator to observe the distribution of angular similarity for the spectra predicted with Prosit Transformer and Prosit RNN. (B) The same angular similarity was also stratified by the length of peptides (C) We also measured the false discovery rate, i.e. the fraction of observed absent peaks among the predicted non-zero intensity peaks, for each spectrum, and (D) the false-negative rate, i.e. the fraction of predicted absent peaks among the observed non-zero intensity peaks.

The highest angular similarity was found between the observed and predicted spectra when setting CE to the set's actual specified value.

## Discussion

Here we have used a Transformer trained to predict protein sequence and transferred its functionality into predicting intensities of the b- and y- ions of MS2 spectra. The resulting predictor's performance outperformed a predictor built by a classical recurrent neural network. This type of structure can likely be used for other types of peptide property prediction as well.

Here we made use of the framework provided by the original Prosit project. It was essential to be able to access the scripts and data sets provided and hardened by the previous team of algorithm designers. In general, it is of utmost importance to keep this type of resource easy to access. If we want to attract the attention of the machine learning community, which often wants a precise problem formulation and does not like to get into the details of how to generate datasets from scratch, we need to help them.

## Acknowledgements

Vital parts of this manuscript were consieved during the Dagstuhl Seminar 21271 on Computational Proteomics, July 2021.

The training of the Prosit Transformer model was enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

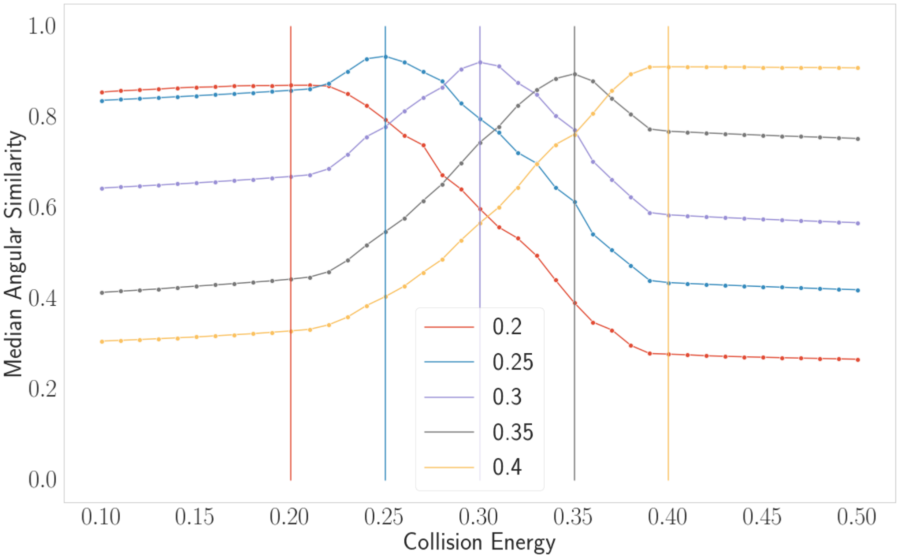

Figure 4: **The mean spectral angle as a function of the collision energy (CE) for spectra acquired with different CEs.**

# Funding

This work has been supported by a grant from the Swedish Foundation for Strategic Research (BD15-0043).

# References

[1] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Syst*, 12(6):654–669.e3, June 2021.

[2] Ana S C Silva, Robbin Bouwmeester, Lennart Martens, and Sven Degroeve. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics*, 35(24):5243–5248, December 2019.

[3] Sven Degroeve, Davy Maddelein, and Lennart Martens. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.*, 43(W1):W326–30, July 2015.

[4] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, 17(1):41–44, January 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods*, 16(6):509–518, June 2019.

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. April 2017.

[8] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–925, November 2007.

[9] Chunwei Ma, Yan Ren, Jiarui Yang, Zhe Ren, Huanming Yang, and Siqi Liu. Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.*, 90(18):10881–10888, September 2018.

[10] Matthias Mann, Chanchal Kumar, Wen-Feng Zeng, and Maximilian T Strauss. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst*, 12(8):759–770, August 2021.

[11] Lennart Martens, Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, and Sven Degroeve. DeepLC can predict retention times for peptides that carry as-yet unseen modifications.

[12] Florian Meier, Niklas D Köhler, Andreas-David Brunner, Jean-Marc H Wanka, Eugenia Voytik, Maximilian T Strauss, Fabian J Theis, and Matthias Mann. Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.*, 12(1):1185, February 2021.

[13] Jesse G Meyer. Deep learning neural network tools for proteomics, 2021.

[14] Luminita Moruz, Daniela Tomazela, and Lukas Käll. Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome Res.*, 9(10):5209–5216, October 2010.

[15] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.*, 32:9689–9701, December 2019.

[16] Brian C Searle, Kristian E Swearingen, Christopher A Barnes, Tobias Schmidt, Siegfried Gessulat, Bernhard Küster, and Mathias Wilhelm. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.*, 11(1):1548, March 2020.

[17] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.*, 114(31):8247–8252, August 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. June 2017.

[19] Mathias Wilhelm, Daniel P Zolg, Michael Graber, Siegfried Gessulat, Tobias Schmidt, Karsten Schnatbaum, Celina Schwencke-Westphal, Philipp Seifert, Niklas de Andrade Krätzig, Johannes Zerweck, Tobias Knaute, Eva Bräunlein, Patroklos Samaras, Ludwig Lautenbacher, Susan Klaeger, Holger Wenschuh, Roland Rad, Bernard Delanghe, Andreas Huhmer, Steven A Carr, Karl R Clauser, Angela M Krackhardt, Ulf Reimer, and Bernhard Kuster. Author correction: Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.*, 12(1):4002, June 2021.

[20] Jinghan Yang, Zhiqiang Gao, Xiuhan Ren, Jie Sheng, Ping Xu, Cheng Chang, and Yan Fu. DeepDigest: Prediction of protein proteolytic digestion with deep learning. *Anal. Chem.*, 93(15):6094–6103, April 2021.

[21] Wangchunshu Zhou, Tao Ge, Furu Wei, Ming Zhou, and Ke Xu. Scheduled DropHead: A regularization method for transformer models, 2020.