# Towards the interpretation of tandem mass spectra with self-supervised machine learning

Roman Bushuiev1,2, Tomáš Pluskal1

1Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nám. 2, 160 00 Praha 6, Czech Republic

2Czech Technical University in Prague, Faculty of Information Technology, Thákurova 9, 160 00 Praha 6, Czech Republic

The identification of unknown molecules is the main bottleneck in biological and biochemical applications of mass spectrometry. State-of-the-art computational approaches deduce desired structural properties of molecules from their tandem mass spectra, relying on previously annotated spectral libraries. However, the diversity of mass spectrometry conditions and limited scope of the spectral libraries limit the capabilities of such methods. In our project, we propose to break the limitation with the paradigm of self-supervised machine learning. Firstly, we extract a high-quality subsample of unannotated tandem mass spectra from the MassIVE repository (hundreds of millions of spectra). Secondly, we use the obtained data to pre-train MSBERT - a version of the prominent BERT language model adapted to operate on mass spectra. For this purpose, we employ an analog of the “masked language model” training objective, where we hide random peaks in the spectra and train the model to predict the hidden peaks. Thirdly, we fine-tune MSBERT end-to-end to predict structural properties of small molecules (e.g. number of Carbon atoms or presence of Nitrogen), using annotated spectral libraries as training data. In preliminary experiments, we have shown that self-supervised pre-training helps MSBERT to solve the downstream tasks. It implies that, similarly to BERT in natural language, MSBERT has the potential to become a working horse for a wide range of mass spectrometry problems. Therefore, we are currently designing an advanced architecture of the model and estimating its optimal training hyperparameters to shift its inductive bias from the natural language to mass spectrometry.