Fast and accurate MS² peak intensity predictions for multiple fragmentation methods, instruments and labeling techniques

Ralf Gabriels[1,2], Lennart Martens[1,2], Sven Degroeve[1,2]

[1] VIB-UGent Center for Medical Biotechnology, Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, Belgium

In mass spectrometry-based proteomics, sequence database search engines have proven to be the gold standard in peptide spectrum identification workflows. However, new demands and novel techniques, such as open modification searches and data-independent acquisition, require a higher resolving power to discriminate good from bad search hits. As the traditional search engines do not fully take advantage of the peak intensity information embedded in peptide spectra, doing so can improve the scoring functions.

We can obtain peak intensity information for virtually every peptide, by training machine learning algorithms on the vast quantities of data present in public proteomics repositories. The machine learning tool MS²PIP (MS² Peak Intensity Prediction) is already capable of doing so with high accuracy. Nevertheless, many post-translational modifications, fragmentation techniques and instruments influence the peak intensities in such a way, that the general MS²PIP models underperform when predicting for these special cases.

Because MS²PIP is a purely data-driven approach, we could train separate models on relevant data sets for CID fragmentation, HCD fragmentation, TMT-labeling, iTRAQ-labeling, and TripleTOF instruments. With the resulting specific models, we were able to obtain MS²PIP accuracies as we would expect for normal peptides, even for specific fragmentation techniques, labeling methods, and instruments.

All MS²PIP models are available on the user-friendly MS²PIP web server (https://iomics.ugent.be/ms2pip). Users can upload up to 100 000 peptide sequences simultaneously, for which MS²PIP predicts MS² spectra in just a few seconds. The resulting spectra can be inspected through interactive plots and can be downloaded in both CSV and MGF file formats.