MS²DIP: Highly accurate MS2 spectrum prediction for modified peptides

Ralf Gabriels[1,2], Robbin Bouwmeester[1,2], Jasper Zuallaert[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, Ghent, Belgium

Accurate MS2 spectrum predictions enable drastic improvements in peptide identification workflows. This identification improvement is particularly useful for challenging proteomics experiments where conventional identification software often falls short. Notable examples of such cases are proteogenomics, data independent acquisition, and open modification searches. The latter also implicitly requires models that can account for residue modifications, but current state-of-the-art MS2 spectrum predictors cannot take these into account. Instead, the corresponding mass shift is introduced, and peak intensities are simply presumed to remain the same for modified and unmodified forms. We here therefore introduce a novel peptide spectrum predictor, called MS²DIP, which can provide accurate predictions for peptides carrying residue modifications, including for modifications not seen during training.

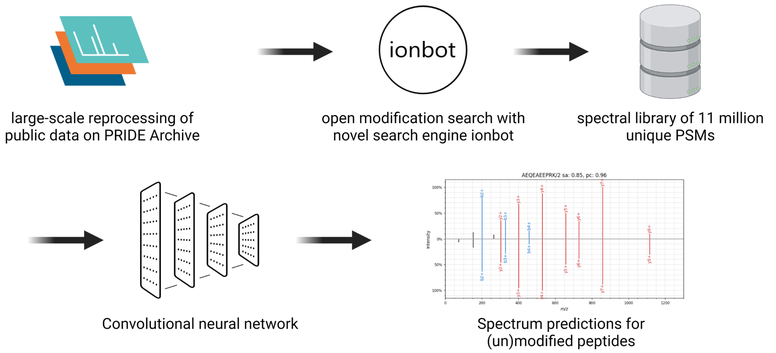


Figure 1. Schematic overview of the MS²DIP development workflow. A large amount of raw mass spectrometry data was downloaded from PRIDE Archive and processed with the novel open modification search engine ionbot, which resulted in a spectral library containing 11 million unique PSMs (by sequence, modifications, and charge). This dataset of modified and unmodified peptide spectra forms the basis for training a convolutional neural network spectrum predictor.

Methods

MS²DIP leverages a state-of-the-art deep learning architecture that enables it to predict spectra for unmodified and modified peptides, by learning the resulting MS2 peak intensities from the atomic composition of each (modified) residue. This, combined with suitable training data, allows MS²DIP to generalize its model across all amino acids, as well as any residue modification, even previously unseen ones. The training data consists of more than 11 million unique combinations of sequence, modifications, precursor charge, and collision energy, originating from open modification searches of a large amount of public proteomics data. This diverse dataset ensures applicability across experimental conditions, and provides an ideal representation of modifications commonly found in open searches (Figure 1).

Preliminary results

Current prototype models of MS²DIP already drastically outperform our previous spectrum prediction tool, MS²PIP, on both modified as well as unmodified peptides, with median Pearson correlations of 0.907 for modified, and 0.943 for unmodified peptides. MS²DIP also outperforms the out-of-the-box version of pDeep3, which shows median Pearson correlations of 0.856 for modified, and 0.924 for unmodified peptides (Figure 2).

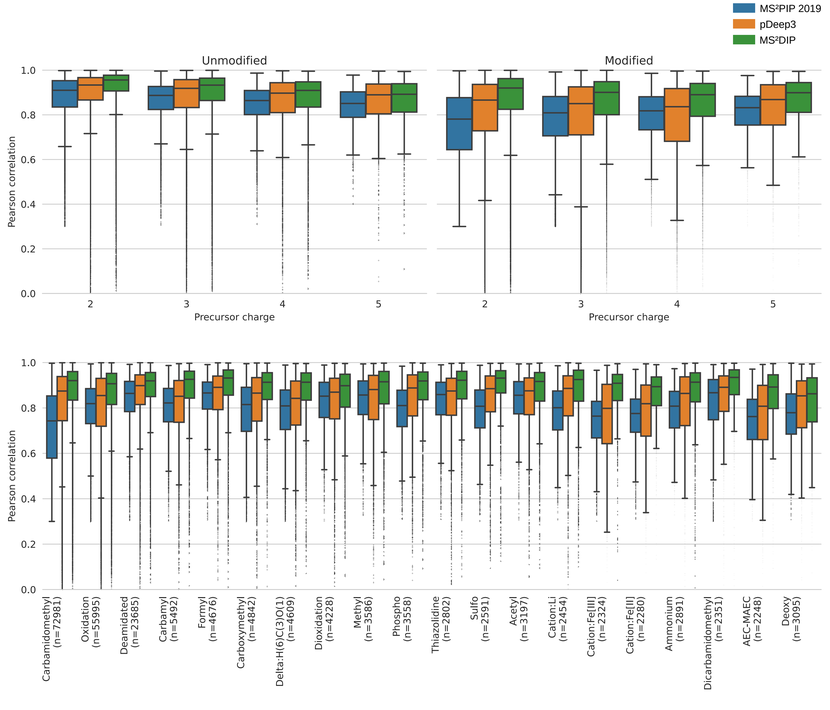


Figure 2. Box plots showing Pearson correlation distributions for MS²PIP 2019 predictions (blue), pDeep3 predictions (orange), and MS²DIP predictions (green). **Top:** Pearson correlations split by unmodified peptides (left) and modified peptides (right) and by precursor charge. **Bottom:** Prediction correlations for the 20 most common modifications in the test dataset, which contains spectra from 81,695 modified and 226,306 unmodified peptides.

We expect further optimizations to the model architecture and hyperparameters to further improve accuracies, allowing MS²DIP to approximate observed technical variance. MS²DIP can easily be integrated into existing as well as novel peptide identification pipelines, such as ionbot, using the Python package or with custom C++ bindings.

MS²Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates

Immunopeptidomics aims to identify immunopeptides, which are presented on major histocompatibility complexes. Existing immunopeptidomics data analysis pipelines have some major hurdles to overcome, which complicates their identification. To enable MS²PIP-based rescoring of immunopeptide PSMs, we have retrained MS²PIP for non-tryptic peptides. Next, the new MS²PIP models, DeepLC, and Percolator were integrated into one software package, called MS²Rescore. Using this integration, we could identify 46% more spectra and 36% more unique peptides at 1% false discovery rate. The integration of the new immunopeptide MS²PIP models, DeepLC, and Percolator into MS²Rescore shows great promise to substantially improve the identification of novel neo- and xeno-epitopes in existing immunopeptidomics workflows.

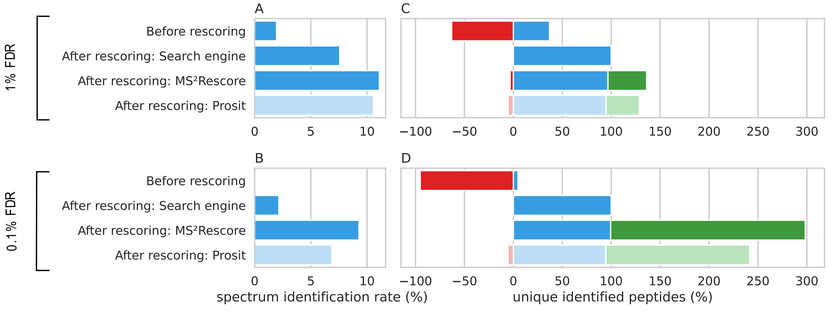


Figure 1. Comparison of rescoring methods at 1% FDR threshold (A, C) and at 0.1% FDR threshold (B, D), evaluated on a large-scale HLA-I immunopeptidomics data set (Sarkizova and Klaeger et al. 2019).

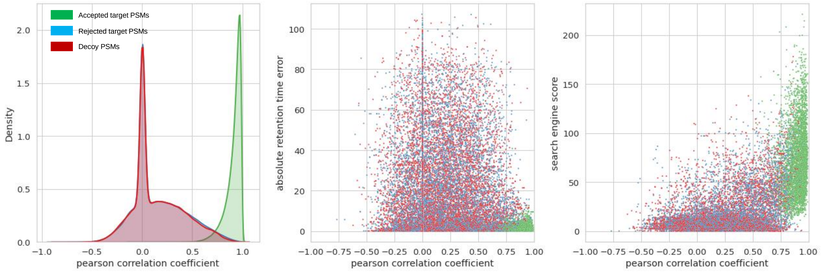
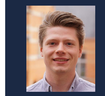


Figure 2. Comparison of the distributions of MS²PIP- and DeepLC features, and the search engine score.



Ralf.Gabriels@UGent.be
@RalfGabriels
@CompOmics
www.compomics.com

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY
GHENT UNIVERSITY
fwo