MS²PIP: Fast and accurate MS² peak intensity prediction 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, Ghent, Belgium

An important step in the analysis of high-throughput mass spectrometry-based proteomics is the correct identification of the peptide MS² fragmentation spectra obtained from a sample analysis. Due to incomplete understanding of the fragmentation process and unpredictable machine noise, matching MS² spectra with the correct peptide is far from trivial. We therefore developed MS²PIP: MS² Peak Intensity Prediction, a data-driven tool that accurately predicts the expected MS² spectrum for a given peptide.[1,2] Since its first publication, we have rebuilt MS²PIP from the ground up to be faster and more accurate. We also trained specific MS²PIP models for multiple specific cases: HCD and CID fragmentation, TripleTOF 5600+ instruments, and iTRAQ- and TMT-labeled peptides. In each of these cases, the peak intensities are substantially influenced by the specific instrument or approach. Specialized models therefore greatly improve the accuracy of MS²PIP. [3]

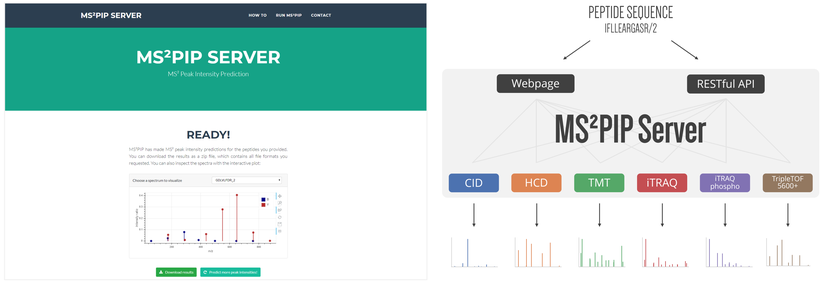

Figure 1. (Left) Screenshot of the MS²PIP Server result page. After a user has uploaded (up to 100 000) peptide sequences and MS²PIP has made its predictions, the results can be inspected through interactive plots. The predicted spectra can be downloaded in CSV, MGF, MSP and SSL/MS2 file formats. **(Right)** Schematic overview of the MS²PIP Server. The web server can be contacted through the user-friendly webpage or through the developer-friendly RESTful API. Peak intensities can be predicted for multiple fragmentation methods, instruments and labeling techniques.

Methods

To train and evaluate specialized MS²PIP models, we downloaded and parsed a multitude of publicly available spectral libraries and experimental datasets. The size of the train-test datasets ranged from 183 000 to 1.6 million unique peptide spectra. The evaluation datasets contained between 9000 and 92 000 unique peptide spectra (Table 1). We can evaluate the models' performance by predicting MS² spectra present in the external evaluation datasets and comparing these predictions to their corresponding empirical spectra. This comparison is done by calculating the Pearson Correlation Coefficient (PCC) of the two spectra, each normalized to their total-ion-current.

All code for training, testing, evaluating and employing MS²PIP models are available on GitHub.com/CompOmics/MS2PIP_c.

Results

The median PCCs between empirical spectra and spectra predicted with the corresponding specialized models are consistently higher then when we apply other models to the same data set. Only the TripleTOF 5600+ model is essentially matched by the HCD model when predicting TripleTOF 5600+ spectra (Figure 2B). Predictions from the correct models yield median PCCs higher then 0.90, except for the TripleTOF 5600+ and the iTRAQ phospho models, which yield median PCCs of 0.74 and 0.84, respectively (Figure 2A).

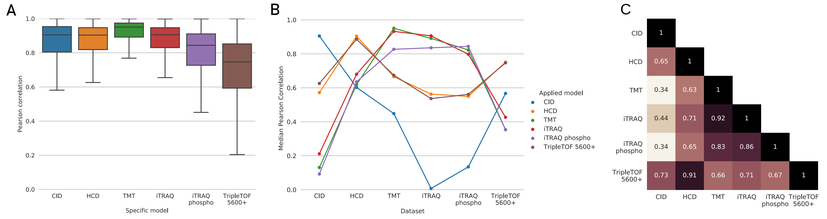

Figure 2. (A) Boxplots showing the Pearson correlation coefficients (PCCs) for each of the specialized models applied to their respective evaluation dataset. **(B)** Median PCCs when applying all specialized models to all evaluation datasets, showing the utility of specialized models. Each dot shows the median PCC of a specialized model applied to a specific evaluation dataset. To improve readability, dots representing performance of a single model are connected. **(C)** Correlation matrix showing median PCCs of the direct comparison of the different model predictions.

It is also noteworthy that models for labeling techniques perform similarly on all datasets, indicating that TMT and iTRAQ labels affect the fragmentation pattern in a comparable fashion. The same is true for the HCD model and the TripleTOF 5600+ model. This is to be expected, as the orbitrap's HCD and the TripleTOF's CID are both beam-type fragmentation methods. The similarities between certain models can be confirmed by directly calculating PCCs between the different model predictions (Figure 2C).

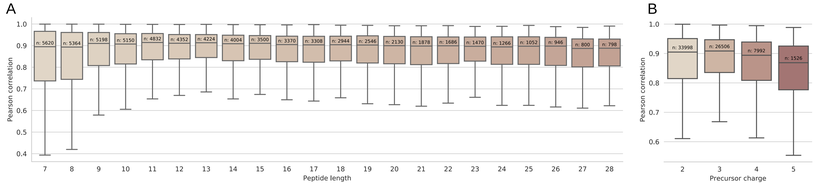

Figure 3. Boxplots showing the Pearson correlation coefficients for the HCD model applied to the HCD evaluation dataset split by **(A)** precursor charge and **(B)** peptide length. Only boxplots containing more than 750 datapoints are plotted. The number in each boxplot displays its number of datapoints.

In this updated version of MS²PIP, we train one model for all precursor charge states and peptide lengths. As a result, we can pool more train data per model, leading to improved prediction accuracies, specifically for longer peptides and peptides with higher charge states (Figure 3).

Table 1: Train-test and evaluation datasets used to generate specialized MS²PIP models

Model	Info	Train-test dataset (# unique peptides)	Evaluation dataset (# unique peptides)
HCD	HCD, orbitrap, tryptic	MassIVE-KB (1 623 712)	PXD008034 (35 269)
CID	CID, linear ion trap, tryptic	NIST CID Human (340 356)	NIST CID Yeast (92 609)
iTRAQ	HCD, orbitrap, tryptic, iTRAQ	NIST iTRAQ (704 041)	PXD001189 (41 502)
iTRAQphospho	HCD, orbitrap, tryptic, iTRAQ, phospho-enriched	NIST iTRAQ phospho (183 383)	PXD001189 (9088)
TMT	HCD, orbitrap, tryptic, TMT	Peng Lab TMT Spectral Library (1 185 547)	PXD009495 (36 137)
TTOF5600	CID, qTOF, tryptic	PXD000954 (215 713)	PXD001587 (15 111)

Discussion and conclusions

Our results confirm that training specialized peak intensity prediction models for specific cases substantially improves the predictions. This can also be confirmed visually when comparing predictions from the HCD and TMT models with an empirical TMT spectrum (Figure 4A).

MS²PIP has already been used for creating proteome-wide spectral libraries for search engines (including Data Independent Acquisition), for selecting discriminative transitions for targeted proteomics[4,5], and for validating interesting peptide identifications (e.g. biomarkers)[6,7]. Recently, we have applied MS²PIP predictions to rescore search engine identifications. This enables us to identify the same number of matches at a 10-fold more conservative false discovery rate, compared to a classic rescoring approach.[8] (Figure 4B)

The new and specialized models extend the applicability of MS²PIP even further, allowing it to be applied to specific fragmentation methods, instruments, or labeling techniques.

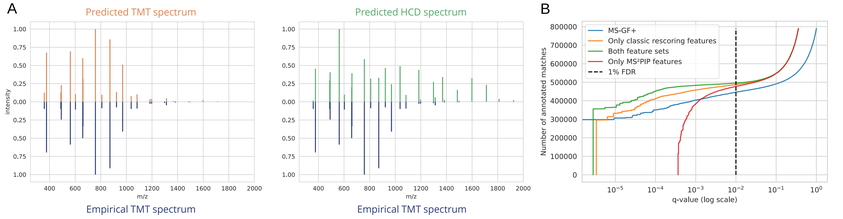

Figure 4. (A) Spectra predicted by MS²PIP TMT model (top left) and HCD model (top right) compared to an empirical spectrum of a TMT-labeled peptide (bottom left and right). **(B)** Number of identified matches at each q-value, for raw search engine output and for different rescoring set-ups (classic rescoring features, only MS²PIP features, and both feature sets), on a HEK-293 sample (PXD001468).

TRY OUT MS²PIP YOURSELF! GO TO OUR USER-FRIENDLY WEBSERVER AND DEVELOPER-FRIENDLY RESTFUL-API AT IOMICS.UGENT.BE/MS2PIP

github.com/RalfG
github.com/CompOmics
twitter.com/RalfGabriels
twitter.com/CompOmics
www.compomics.com
Ralf.Gabriels@UGent.be

1. Sven Degroeve et al. (2013) *Bioinformatics*, doi:10.1093/bioinformatics/btt544
2. Sven Degroeve et al. (2015) *Nucleic Acids Res.*, doi:10.1093/nar/gkv542
3. Ralf Gabriels et al. (2019) *Nucleic Acids Res.*, doi:10.1093/nar/gkz299
4. Jakob Albrethsen et al. (2018) *Clin. Chem. Lab. Med.*, doi:10.1515/cclm-2018-0171
5. Bart Mesuere et al. (2016) *Proteomics*, doi:10.1002/pmic.201600023
6. Harshavardhan Budamgunta et al. (2018) *Proteomics*, doi:10.1002/pmic.201700218
7. Patrick Willems et al. (2017) *Mol. Cell. Proteomics*, doi:10.1074/mcp.M116.066662
8. Ana Silvia C Silva et al. (2019) *Bioinformatics*, doi:10.1093/bioinformatics/btz383