Updated MS²PIP web server delivers 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, VIB, Ghent, Belgium
2. Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

* To whom correspondence should be addressed:
Tel: +32 9 264 93 58;
Email: lennart.martens@vib-ugent.be

Address:
VIB-UGent Center for Medical Biotechnology
A. Baertsoenkaai 3
B9000 Ghent, Belgium

ABSTRACT

MS²PIP is a data-driven tool that accurately predicts peak intensities for a given peptide's fragmentation mass spectrum. Since the release of the MS²PIP web server in 2015, we have brought significant updates to both the tool and the web server. Next to the original models for CID and HCD fragmentation, we have added specific models for the TripleTOF 5600+ mass spectrometer, for TMT-labeled peptides, for iTRAQ-labeled peptides and for iTRAQ-labeled phosphopeptides. Because the fragmentation pattern is heavily altered in each of these cases, these additional models greatly improve the prediction accuracy for their corresponding data types. We have also substantially reduced the computational resources required to run MS²PIP, and have completely rebuilt the web server, which now allows predictions of up to 100.000 peptide sequences in a single request. The MS²PIP web server is freely available at https://iomics.ugent.be/ms2pip/.

INTRODUCTION

In high throughput tandem mass spectrometry (MS^2), peptides are identified by analyzing their fragmentation spectra. These spectra are obtained by collision induced dissociation (CID) or higher-energy collisional dissociation (HCD), where peptides are made to collide with an inert gas, or by electron-transfer dissociation (ETD) or electron-capture dissociation (ECD), in which electrons are transferred to peptides. After fragmentation, the mass-to-charge ratios (m/z) and intensities of the resulting fragment ions are measured, making up the two dimensions of a fragmentation spectrum. While the fragment ions' m/z can easily be calculated for any given peptide, their intensities have proven to follow extremely complex patterns (1).

In 2013, we therefore developed the data-driven tool MS^2PIP: MS^2 Peak Intensity Prediction (2), which can predict fragment ion intensities. By applying machine learning algorithms on the vast amounts of data present in public proteomics repositories such as the PRIDE Archive (3, 4), we could create generalized models that accurately predict the expected normalized MS^2 peak intensities for a given peptide. While the first iteration of MS^2PIP outperformed the then state-of-the art prediction tool PeptideART (5), it was originally only trained for CID fragmentation spectra. As HCD fragmentation became more popular in the field, we therefore expanded MS^2PIP with prediction models for HCD spectra. In 2015, we built the MS^2PIP web server to make these models easily available to all potential users, regardless of their computational resources (6).

Over the past few years, MS^2PIP has been used by researchers to create proteome-wide spectral libraries for proteomics search engines (including Data Independent Acquisition), for the selection of discriminative transitions for targeted proteomics (7, 8), and for the validation of interesting peptide identifications (e.g. biomarkers) (9, 10). Moreover, we have also shown that MS^2PIP predictions can be used to improve upon and even replace proteomics search engine output when rescoring peptide-to-spectrum matches (11).

Because of the great interest in, and steadily increasing relevance of MS^2 peak intensity prediction, we have continued updating and improving upon MS^2PIP and the MS^2PIP web server. We have updated MS^2PIP to be more computationally efficient, rebuilt the MS^2PIP web server to handle up to 100.000 peptide sequences per request instead of 1.000, and we have added specific models for the TripleTOF 5600+ mass spectrometer and isobaric labeled peptides.

NEW IN THE 2019 VERSION OF MS^2PIP

More efficient MS^2PIP code

Rapid advances in Machine Learning research combined with larger and more diverse training datasets have allowed for more accurate MS^2PIP predictive models. The Random Forest algorithm employed in the original MS^2PIP has made room for a Gradient Tree Boosting algorithm (12) that, combined with more training data, improves prediction accuracy. This is especially the case for peptides with higher charge states, where the large performance differences between charge 2+ and 3+ observed for the original MS^2PIP models have been significantly reduced in the new version (Supplementary Figure 1).

In addition, we have drastically reduced the required computational resources, while also improving the already fast prediction speeds. The large memory footprint of the original version (requiring several gigabytes) is now reduced to just a few 100 megabytes, depending on input request size. When run locally on a normal four core laptop, MS^2PIP can predict peak intensities for a million peptides in less than 5 minutes.

Specific models for isobaric labeled peptides and the TripleTOF 5600+ mass spectrometer

One of the most important changes in this new version of MS²PIP is the addition of specialized models for specific types of peptide spectra. The type of mass spectrometer, fragmentation method and certain peptide modifications (such as isobaric labels and phosphorylation) can heavily alter peptide fragmentation patterns. We have therefore now also trained specialized models for the TripleTOF 5600+ mass spectrometer, for TMT-labeled peptides (13), for iTRAQ-labeled peptides (14), and for iTRAQ-labeled phosphopeptides. Each of these models were trained and evaluated on publicly available spectral libraries or experimental datasets, ranging in size from 130.000 to 1.6 million peptide spectra. Final validation on every model was based on wholly independent data sets, ranging in size from 9.000 to 50.000 unique peptide spectra (Table 1). Spectral libraries were filtered for unique peptides and then converted to MS²PIP input format. For experimental datasets, original peptide identifications as provided by the data submitter were used, where available. When not available, we performed identification ourselves using the MS-GF+ (15) search engine in combination with Percolator (16) for post-processing.

Table 1. Train-test and evaluation datasets used for specific MS2PIP models.

Model	Use	Dataset	# Unique peptides
CID	Train-test	NIST CID (17)	340 356
	Evaluation	NIST CID Yeast (17)	92 609
HCD	Train-test	MassIVE-KB (18)	1 623 712
	Evaluation	PXD008034 (19)	35 269
TripleTOF 5600+	Train-test	PXD000954 (20)	215 713
	Evaluation	PXD001587 (21)	15 111
TMT	Train-test	Peng Lab TMT Spectral Library (22)	1 185 547
	Evaluation	PXD009495 (23)	36 137
iTRAQ	Train-test	NIST iTRAQ (17)	704 041
	Evaluation	PXD001189 (24)	41 502
iTRAQ phospho	Train-test	NIST iTRAQ phospho (17)	183 383
	Evaluation	PXD001189 (24)	9 088

Redesigned, more robust web server

Along with the heavily updated MS²PIP models, we also rebuilt the web server from the ground up. Like the previous version, this web server is built using the Flask framework (https://flask.pocoo.org) with a front-end built upon Bootstrap (https://getbootstrap.com).

In this newly built web server we have implemented a robust queueing system that handles concurrent tasks. This allowed us to increase the maximum number of peptide sequences *per* request from 1.000 to 100.000. Apart from submitting a single task through the website, users can also automate their requests through MS²PIP's updated RESTful API, for which we provide an example Python script. One request of 100.000 peptide sequences takes less than five minutes to complete, including up- and download time. Predictions for 1,000 peptide sequences are returned in less than three seconds.

On the user-friendly webpage, users can select one of the available models and upload a csv file with peptide sequences, precursor charges, and modifications. After uploading this input file, a progress bar displays the status of the request and a URL is displayed to which the user can return at any time to check the status of their request in case they would close the browser window. When the predictions are ready, the user can inspect the results through several interactive plots, and the predicted spectra can be downloaded in Mascot Generic File (MGF) format or as comma-separated values (CSV).

PERFORMANCE OF THE SPECIFIC MODELS

We can evaluate MS²PIP model performances by predicting peak intensities for peptides present in the external evaluation datasets and by comparing these predictions to their corresponding empirical spectra. This comparison is performed by calculating the Pearson correlation between the predicted and experimental spectra. The resulting Pearson correlation distributions for each of the specific models are shown in Figure 1A.

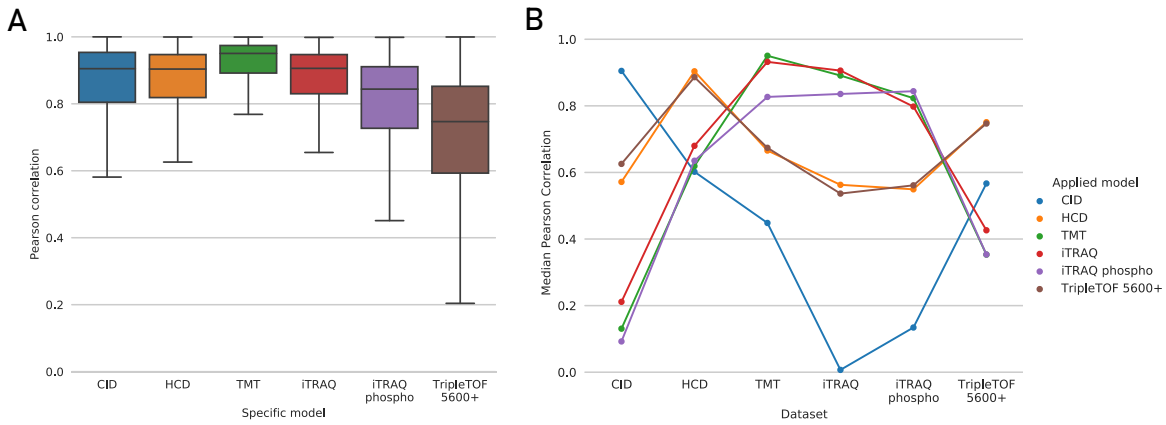

Figure 1. A) Boxplots showing the Pearson correlation coefficients for each of the specific models applied to their respective evaluation dataset. B) Median Pearson correlations when applying all specific models to each evaluation datasets, showing the utility of specialized models. Each dot shows the median Pearson correlation of a specific model applied to a specific evaluation dataset. Dots of the same applied model are connected.

The median Pearson correlations are higher than 0.90 for all models, except for the TripleTOF 5600+ and the iTRAQ phospho models, which have median Pearson correlations of 0.74 and 0.84, respectively. These two lower median correlations might be the result of a higher intrinsic variation in peak intensities of these specific cases. The presence of a phosphorylation alters the fragmentation pattern, and the iTRAQ phospho datasets contain both modified as well as unmodified phosphorylation sites, which leads to a more complex prediction challenge. The lower correlation of the TripleTOF 5600+ model might be explained by a higher variation in peak intensities as well. This because this instrument tends to record spectra with very low peak intensities, which are well below the linear range of the detector, introducing more variation in normalized peak intensities.

When we apply all specific models to each specific evaluation dataset – that is, including mismatched model-dataset combinations, such as applying the TMT model to the HCD evaluation dataset – we observe median Pearson correlations that are consistently substantially higher for correctly matched models and evaluation datasets than for mismatched models and evaluation datasets (Figure 1B). Only the specific TripleTOF 5600+ model is essentially matched by the HCD model when predicting TripleTOF 5600+ spectra. Overall, this figure makes a clear case for the utility of specialized MS²PIP models for specific types of data.

Figure 1B also displays, indirectly, which specific cases show similar fragmentation patterns to one another. The specific models for isobaric-labeled peptides (TMT, iTRAQ and iTRAQ phospho) score similarly across the different evaluation datasets, as is the case for the HCD and TripleTOF 5600+ models. These two groups, however, score very differently for all other evaluation datasets.

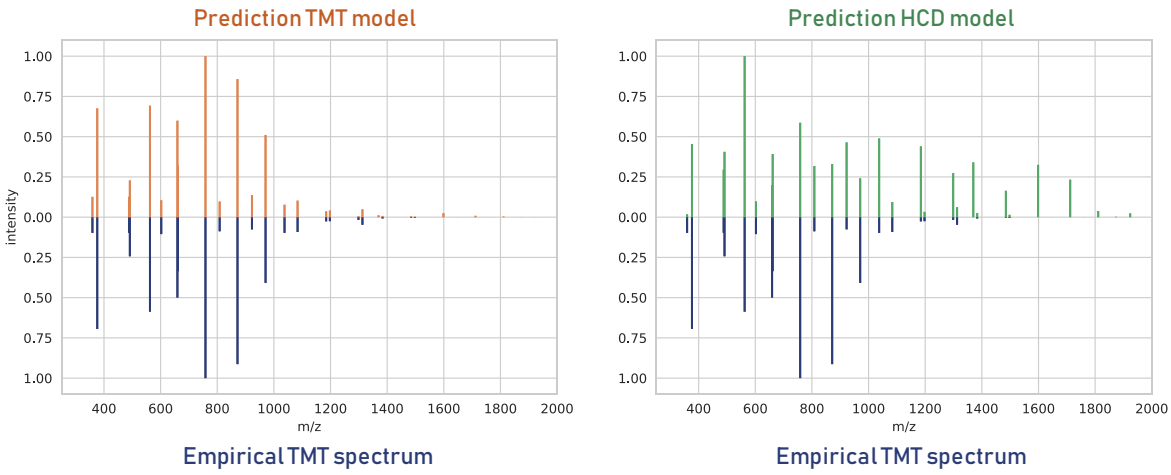

Figure 2. Predictions for the peptide sequence EENGVLVLNDANFDNFVADK, carrying two TMT labels, predicted by the specific TMT model (top left) and HCD model (top right), compared to the empirical spectrum (bottom left and right).

We can also visualize the differences in fragmentation pattern by simply plotting the predictions from two different models for the same peptide sequence and mirroring the empirical spectrum below these predictions. This is shown in Figure 2 for the TMT and HCD models with an example TMT-labeled peptide spectrum. While the TMT model nicely mirrors the empirical TMT spectrum, the HCD model does not match the empirical TMT spectrum at all.

CONCLUSION AND FUTURE PERSPECTIVES

With the advent of novel mass spectrometry methods and new computational pipelines, MS² peak intensity prediction is becoming even more relevant. As one of the front runners in peak intensity prediction, MS²PIP has already been used for a variety of purposes, including the creation of proteome-wide spectral libraries, targeted proteomics applications, the validation of interesting peptide identifications, and the rescoring of search engine output.

With this update, we present our latest efforts in further widening the scope of MS²PIP. The new web server enables researchers to easily obtain more predictions more efficiently, and the new MS²PIP models extend the applicability of MS²PIP to more varied, popular use cases, allowing it to be applied when specific fragmentation methods, instruments, or labeling techniques are employed.

AVAILABILITY

The MS²PIP web server is freely available via iomics.ugent.be/ms2pip. Documentation for contacting the RESTful API is available via iomics.ugent.be/ms2pip/api/. MS²PIP is open source, licensed under the Apache-2.0 License, and is hosted on github.com/compomics/ms2pip_c.

References

1. Barton,S.J. and Whittaker,J.C. (2009) Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom. Rev.*, **28**, 177–187.
https://doi.org/10.1002/mas.20188

2. Degroeve,S. and Martens,L. (2013) MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, **29**, 3199–203.
https://doi.org/10.1093/bioinformatics/btt544
http://www.ncbi.nlm.nih.gov/pubmed/24078703

3. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: The proteomics identifications database. *Proteomics*, **5**, 3537–3545.
https://doi.org/10.1002/pmic.200401303

4. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M., *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
https://doi.org/10.1093/nar/gky1106

5. Arnold,R.J., Jayasankar,N., Aggarwal,D., Tang,H. and Radivojac,P. (2005) A machine learning approach to predicting peptide fragmentation spectra. In *Biocomputing 2006*.pp. 219–230.
https://doi.org/10.1142/9789812701626_0021
http://www.ncbi.nlm.nih.gov/pubmed/17094241

6. Degroeve,S., Maddelein,D. and Martens,L. (2015) MS^2 PIP prediction server: compute and visualize MS^2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.*, **43**, W326–W330.
https://doi.org/10.1093/nar/gkv542
http://www.ncbi.nlm.nih.gov/pubmed/25990723

7. Albrethsen,J., Frederiksen,H., Andersson,A.-M., Anand-Ivell,R., Nordkap,L., Bang,A.K., Jørgensen,N. and Juul,A. (2018) Development and validation of a mass spectrometry-based assay for quantification of insulin-like factor 3 in human serum. *Clin. Chem. Lab. Med.*, **56**, 1913–1920.
https://doi.org/10.1515/cclm-2018-0171

8. Mesuere,B., Van der Jeugt,F., Devreese,B., Vandamme,P. and Dawyndt,P. (2016) The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. *Proteomics*, **16**, 2313–2318.
https://doi.org/10.1002/pmic.201600023

9. Budamgunta,H., Olexiouk,V., Luyten,W., Schildermans,K., Maes,E., Boonen,K., Menschaert,G. and Baggerman,G. (2018) Comprehensive Peptide Analysis of Mouse Brain Striatum Identifies Novel sORF-Encoded Polypeptides. *Proteomics*, **18**, 1700218.
https://doi.org/10.1002/pmic.201700218

10. Willems,P., Ndah,E., Jonckheere,V., Stael,S., Sticker,A., Martens,L., Van Breusegem,F., Gevaert,K. and Van Damme,P. (2017) N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in Arabidopsis thaliana. *Mol. Cell. Proteomics*, **16**, 1064–1080.
https://doi.org/10.1074/mcp.M116.066662

11. C. Silva,A.S., Martens,L. and Degroeve,S. (2019) Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *bioRxiv*, 10.1101/428805.
https://doi.org/10.1101/428805

12. Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*.
https://doi.org/10.1145/2939672.2939785
http://www.ncbi.nlm.nih.gov/pubmed/22942019

13. Thompson,A., Schäfer,J., Kuhn,K., Kienle,S., Schwarz,J., Schmidt,G., Neumann,T. and Hamon,C. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, **75**, 1895–1904.
https://doi.org/10.1021/ac0262560
http://www.ncbi.nlm.nih.gov/pubmed/12713048

14. Ross,P.L., Huang,Y.N., Marchese,J.N., Williamson,B., Parker,K., Hattan,S., Khainovski,N., Pillai,S., Dey,S., Daniels,S., *et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, **3**, 1154–69.
https://doi.org/10.1074/mcp.M400129-MCP200
http://www.ncbi.nlm.nih.gov/pubmed/15385600

15. Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.
https://doi.org/10.1038/ncomms6277
http://www.ncbi.nlm.nih.gov/pubmed/25358478

16. The,M., MacCoss,M.J., Noble,W.S. and Käll,L. (2016) Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.*, **27**, 1719–1727.
https://doi.org/10.1007/s13361-016-1460-7

17. NIST NIST Libraries of Peptide Tandem Mass Spectra.

18. Wang,M., Wang,J., Carver,J., Pullman,B.S., Cha,S.W. and Bandeira,N. (2018) Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.*, 10.1016/J.CELS.2018.08.004.
https://doi.org/10.1016/J.CELS.2018.08.004

19. Gravina,F., Sanchuki,H.S., Rodrigues,T.E., Gerhardt,E.C.M., Pedrosa,F.O., Souza,E.M., Valdameri,G., de Souza,G.A. and Huergo,L.F. (2018) Proteome analysis of an Escherichia coli ptsN-null strain under different nitrogen regimes. *J. Proteomics*, **174**, 28–35.
https://doi.org/10.1016/J.JPROT.2017.12.006

20. Rosenberger,G., Koh,C.C., Guo,T., Röst,H.L., Kouvonen,P., Collins,B.C., Heusel,M., Liu,Y., Caron,E., Vichalkovski,A., *et al.* (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data*, **1**, 140031.
https://doi.org/10.1038/sdata.2014.31

21. Tsou,C.-C., Avtonomov,D., Larsen,B., Tucholska,M., Choi,H., Gingras,A.-C. and Nesvizhskii,A.I. (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods*, **12**, 258–264.
https://doi.org/10.1038/nmeth.3255

22. Shen,J., Pagala,V.R., Breuer,A.M., Peng,J., Bin Ma,B. and Wang,X. (2018) Spectral Library Search Improves Assignment of TMT Labeled MS/MS Spectra. *J. Proteome Res.*, **17**, 3325–3331.
https://doi.org/10.1021/acs.jproteome.8b00594

23. Mateus,A., Bobonis,J., Kurzawa,N., Stein,F., Helm,D., Hevler,J., Typas,A. and Savitski,M.M. (2018) Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.*, **14**, e8242.

https://doi.org/10.15252/msb.20188242
http://www.ncbi.nlm.nih.gov/pubmed/29980614

24. Beck,F., Geiger,J., Gambaryan,S., Solari,F.A., Dell'Aica,M., Loroch,S., Mattheij,N.J., Mindukshev,I., Pötz,O., Jurk,K., *et al.* (2017) Temporal quantitative phosphoproteomics of ADP stimulation reveals novel central nodes in platelet activation and inhibition. *Blood*, **129**, e1–e12.
https://doi.org/10.1182/blood-2016-05-714048
http://www.ncbi.nlm.nih.gov/pubmed/28060719