Updated MS^{2}PIP web server supports cutting-edge proteomics applications

Arthur Declercq, Robbin Bouwmeester, Cristina Chiva, Eduard Sabidó, Aurélie Hirschler, Christine Carapito, Lennart Martens, Sven Degroeve, Ralf Gabriels

W338–W342 *Nucleic Acids Research, 2023, Vol. 51, Web Server issue* *Published online 4 May 2023*
https://doi.org/10.1093/nar/gkad335

Updated MS^{2}PIP web server supports cutting-edge proteomics applications

Arthur Declercq [1,2], Robbin Bouwmeester [1,2], Cristina Chiva [3,4], Eduard Sabidó [3,4], Aurélie Hirschler [5], Christine Carapito [5], Lennart Martens [1,2], Sven Degroeve [1,2,*] and Ralf Gabriels [1,2]

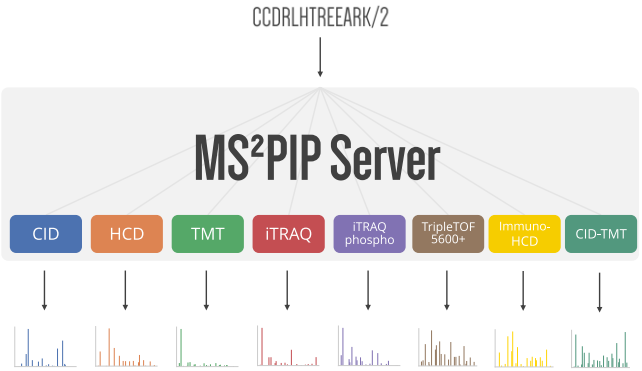
[1]VIB-UGent Center for Medical Biotechnology, VIB, Belgium, [2]Department of Biomolecular Medicine, Ghent University, Belgium, [3]Proteomics Unit, Universitat Pompeu Fabra, 08003, Barcelona, Spain, [4]Proteomics Unit, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08003, Barcelona, Spain and [5]Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), Université de Strasbourg, CNRS, France

Received February 26, 2023; Revised April 04, 2023; Editorial Decision April 12, 2023; Accepted April 25, 2023

ABSTRACT

Interest in the use of machine learning for peptide fragmentation spectrum prediction has been strongly on the rise over the past years, especially for applications in challenging proteomics identification workflows such as immunopeptidomics and the full-proteome identification of data independent acquisition spectra. Since its inception, the MS^{2}PIP peptide spectrum predictor has been widely used for various downstream applications, mostly thanks to its accuracy, ease-of-use, and broad applicability. We here present a thoroughly updated version of the MS^{2}PIP web server, which includes new and more performant prediction models for both tryptic- and non-tryptic peptides, for immunopeptides, and for CID-fragmented TMT-labeled peptides. Additionally, we have also added new functionality to greatly facilitate the generation of proteome-wide predicted spectral libraries, requiring only a FASTA protein file as input. These libraries also include retention time predictions from DeepLC. Moreover, we now provide pre-built and ready-to-download spectral libraries for various model organisms in multiple DIA-compatible spectral library formats. Besides upgrading the back-end models, the user experience on the MS^{2}PIP web server is thus also greatly enhanced, extending its applicability to new domains, including immunopeptidomics and MS3-based TMT quantification experiments. MS^{2}PIP is freely available at https://iomics.ugent.be/ms2pip/.

GRAPHICAL ABSTRACT



INTRODUCTION

Over the past decade, the ever-broadening scope of diverse proteomics workflows has engendered greatly increased interest in the field. However, these new applications each come with their specific challenges. For example, immunopeptidomics must address the non-tryptic nature of immunopeptides, whereas isobaric labelling for quantification can result in reduced identification efficiency (1,2). These specialized approaches, which build on sample preparation and proteomics acquisition innovations, therefore also require novel developments in data analysis to maximally exploit the value of the corresponding data.

One data analysis innovation that has impacted nearly all of the new proteomics workflows is the machine learning-based prediction of accurate peptide fragmentation spectra, as pioneered by MS^{2}PIP and others (3,4). Indeed, we have previously showcased the wide applicability of MS^{2}PIP predictions (5–7) (https://iomics.ugent.be/ms2pip), and how these can be leveraged to boost the yields from various proteomics identification strategies (8). Interesting use cases of these predictions include the rescoring

*To whom correspondence should be addressed. Tel: +32 9 224 98 54; Email: sven.degroeve@vib-ugent.be

of peptide-spectrum matches (PSMs) (8,9), the creation of proteome-wide spectral libraries for data-independent acquisition (DIA) (10,11) and streamlining the design of targeted proteomics experiments (12,13). While MS^{2}PIP already supported a wide variety of fragmentation methods, instruments, and labelling techniques, the development of various novel impactful proteomics workflows resulted in a clear demand for additional, specialized MS^{2}PIP models.

We have therefore further expanded MS^{2}PIP with the requisite new prediction models, which now include support for tryptic- and non-tryptic peptides, for immunopeptides, and for collision-induced dissociation (CID) spectra of peptides treated with tandem-mass-tag (TMT) quantification labels. These new models allow MS^{2}PIP to be applied in alternative digestion experiments, in immunopeptidomics experiments, and in MS3-TMT-based quantification studies. We have updated the MS^{2}PIP web server to include these new prediction models, alongside several new features, such as the integration of our state-of-the-art retention time predictor DeepLC (14), the option to generate proteome-wide spectral libraries starting from only a FASTA file, and the availability of prebuilt, ready-to-download spectral libraries for ten common model organisms in multiple DIA-compatible file formats. These updates will further streamline downstream use of MS^{2}PIP, allowing even wider adoption and utility.

NEW IN THE 2023 VERSION OF THE MS^{2}PIP WEB SERVER

Updated MS^{2}PIP core library with increased availability

Since the previous MS^{2}PIP web server publication, we have drastically improved the availability and usability of MS^{2}PIP's core library. It is now available as a standalone Python package that can be easily installed on all major OS platforms with PyPI, with Bioconda, or as a BioContainer. In addition to the command line interface (CLI), a new Python interface now allows MS^{2}PIP to be easily integrated into other tools and workflows. To compute correlations between observed and predicted spectra, MS^{2}PIP now supports both MGF and mzML spectrum file formats. MS^{2}PIP now also seamlessly integrates the state-of-the-art retention time predictor DeepLC. Furthermore, we have implemented two new operating modes for MS^{2}PIP: (i) the *fasta2speclib* command allows users to generate proteome-wide predicted spectral libraries, starting from only a FASTA proteome file and (ii) the *single-prediction* command allows users to quickly predict a single spectrum directly from the CLI. The MS^{2}PIP core package is open-source under the permissive Apache2 license, and is freely available at https://github.com/compomics/ms2pip/.

Extended and improved MS^{2}PIP web server

For an optimal, user-friendly experience, MS^{2}PIP is made available as an online web server. Since its previous publication, we have significantly extended the MS^{2}PIP web server functionality. First, the web server contains all new features of the MS^{2}PIP core library, most notably including the new prediction models (see below). Second, without any additional configuration, users can opt to include accurate retention time predictions in the predicted libraries from our retention time predictor DeepLC. Third, the web server now also accepts—next to the existing peptide list input—a protein FASTA file with 'search space' settings that define which peptides will be included in the library. Configurable settings include the cleavage rules for *in silico* digestion, the number of allowed missed cleavages, the precursor m/z range, and common residue modifications. Fourth, we now provide ready-to-download spectral libraries for ten common model organism UniProt reference proteomes, including *Homo sapiens*, *Escherichia coli* and *Arabidopsis thaliana*. Each library is available in the MSP and SSL/MS2 file formats, ensuring compatibility with major DIA search engines, such as DIA-NN (15) and Skyline (16).

New prediction models for (non-)tryptic peptides, immunopeptides and MS3 quantification experiments

We have updated MS^{2}PIP with three new prediction models. The 2019 model for HCD fragmentation was originally only trained on tryptic peptides. Non-tryptic peptides, however, lack the basic lysine or arginine on their C-terminus, which heavily influences fragmentation patterns (17). As a result, the existing MS^{2}PIP models performed sub optimally for non-tryptic peptides. To allow MS^{2}PIP to be applied to proteomics workflows that yield non-tryptic peptides, such as alternative-digestion and biopeptidomics experiments, we have trained a new and improved HCD model capable of both tryptic and non-tryptic peptide predictions. This model was validated on external evaluation data sets containing peptides from both trypsin- and chymotrypsin-digestion. Importantly, this new, much more generic model outperforms the previous model on both tryptic and non-tryptic peptides. Additionally, we have trained a specialized model for immunopeptides to be used in immunopeptidomics experiments. This model was validated on both HLA class I and HLA class II peptides.

In quantitative mass spectrometry, MS3 acquisition of TMT-labeled spectra has been gaining popularity over traditional MS2 acquisition (18). However, the combination of CID fragmentation, ion trap acquisition of MS2 spectra, and of TMT-labelling substantially alters fragmentation patterns, which is detrimental for the performance of both the existing CID and HCD-TMT MS^{2}PIP models. Therefore, we have trained and validated a new CID-TMT model to allow for applications of MS^{2}PIP in MS3-TMT-based quantification studies.

Train, test, and evaluation data was downloaded from PRIDE (19,20) and converted to MS^{2}PIP input files (Supplementary Table S1)—except for the CID-TMT training data, which was generated in-house and is available from PRIDE with identifier PXD041002 (see supplementary methods). While not explicitly considered for intensity prediction, the train and test data also included common modifications such as oxidation of M, carbamidomethylation of C and acetylation of the amino termini. To guarantee fully external unseen evaluation data sets, overlapping peptidoforms between train and test sets were removed from the test set. Similar to the 2019 MS^{2}PIP models (7) all new models were trained with a gradient boosting machine learning algorithm (see Supplementary Table S2) as imple-

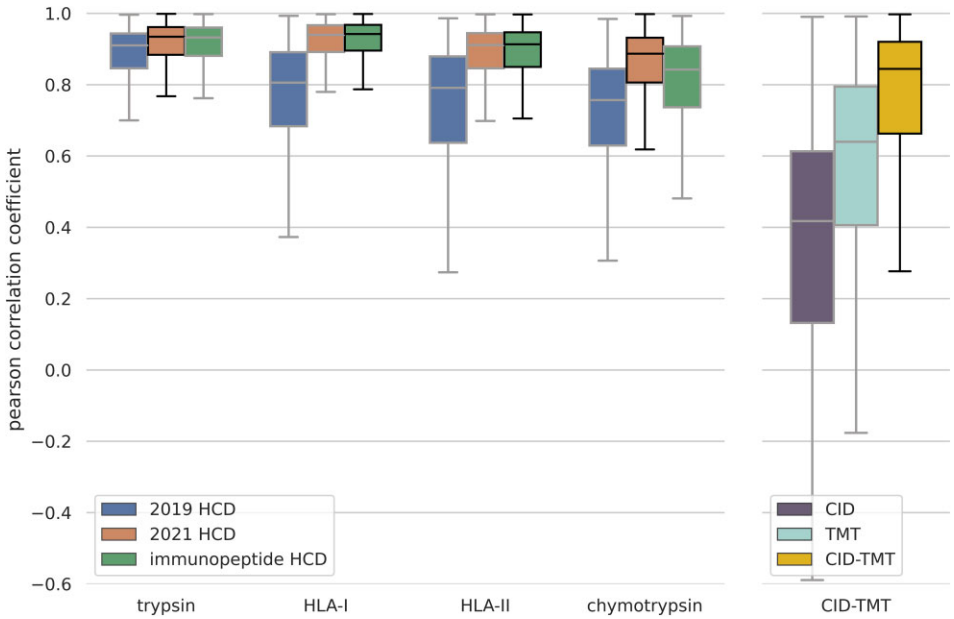


Figure 1. Distribution of Pearson correlation coefficients per spectrum (y-axis) for each newly trained model and the relevant existing models, evaluated on various external unseen data sets (x-axis). Each color represents a model, with the target model for each evaluation data set shown with black borders and the other models shown with grey borders.

mented in the XGBoost Python package (see supplementary methods).

Performance of the new MS^{2}PIP models

To evaluate the newly added MS^{2}PIP models, we selected several unseen, external evaluation data sets to compare the predictions with observed spectra and calculate Pearson correlation coefficients (PCC) per spectrum. The selected orbitrap HCD data sets consist of trypsin-digested, chymotrypsin-digested, HLA class I and HLA class II peptides, respectively. We compared the performance of the new MS^{2}PIP HCD and immunopeptide models with the 2019 MS^{2}PIP model on each of these evaluation data sets. Both new models showed substantial increases in performance on their respective target data sets, with a median PCC of 0.93 and 0.88 for the 2021 HCD model on trypsin and chymotrypsin and a median PCC of 0.94 and 0.91 for the immunopeptide HCD model on HLA-I and HLA-II data (Figure 1). Notably, even when evaluated on a trypsin-digested peptide data set, the new, more generic HCD model still shows an increase in performance, suggesting that combining tryptic and non-tryptic data for training leads to an overall better generalized model. Furthermore, the lower performance of the specialized immunopeptide model on chymotrypsin-digested peptide data indicates that these two types of non-tryptic peptides are likely very different. Indeed, separating predictive performance by peptide length shows a significant drop in accuracy for peptides longer than 17 amino acids for the immunopeptide model, while the new general HCD model shows a consistently high performance across peptide lengths (Supplementary Figure S1). When examining the prediction accuracies for HLA type I and type II in a similar manner, we observe an improved performance across all peptide lengths and for both HLA types, compared to the 2019 HCD model (Supplementary Figure S2).

Previously we have shown that acquisition modes and isobaric labelling techniques can heavily alter peak intensity patterns (7). This is especially the case for ion trap-based CID acquisition of TMT-labelled spectra. Evaluation on a CID-TMT data set shows that neither the existing CID nor the existing HCD-TMT MS^{2}PIP models generalized well for this type of peptide spectra. Interestingly, the HCD-TMT model still outperforms the CID model, suggesting that the labelling method has a larger influence on peak intensity patterns than the fragmentation method (Supplementary Figure S3). This can be confirmed by correlating observed spectra directly for each type. Indeed, observed HCD-TMT spectra correlate slightly better with CID-TMT spectra than with unlabeled CID peptide spectra. Nevertheless, as both correlations are low, there was a need for a specialized CID-TMT prediction model. The newly trained CID-TMT model vastly outperforms current models with a median PCC of 0.84 (Figure 1).

CONCLUSION AND FUTURE PERSPECTIVES

The use of machine learning-based predictive models for analyte behavior has become an indispensable part of proteomics, as is reflected by the number and popularity of machine learning tools – including MS^{2}PIP – that have been published in the past years (3,4). Among these tools, the prediction of fragment intensities and peptide retention times have proven highly valuable useful to improve the confidence in peptide identification (9). While recently many deep learning-based spectrum predictors have been developed, the use of the gradient tree boosting (XGBoost) machine learning algorithm allows us to easily build accurate prediction models for specialized use cases where less

training data might be available. Additionally, MS^{2}PIP does not require graphical processing units and can be run on virtually any computer system. Nevertheless, with the updated MS^{2}PIP web server we aim to make both MS^{2}PIP and DeepLC even more easily accessible to the entire proteomics community. The updated MS^{2}PIP web server is the first to allow users to generate proteome-wide spectral libraries on the fly directly from a FASTA file and additionally provides pre-built spectral libraries for ten model organisms. Furthermore, thanks to the addition of three new, highly performant peptide spectrum prediction models, MS^{2}PIP continues to support and push forward innovations in proteomics and its various established and emerging subfields.

DATA AVAILABILITY

The MS^{2}PIP web server is freely available at https://iomics.ugent.be/ms2pip/. The core library is open source, licensed under the permissive Apache-2.0 license, available as a package on PyPI, Bioconda, and BioContainers, and hosted on https://github.com/compomics/ms2pip/ and published on https://doi.org/10.5281/zenodo.7669701. All scripts for model training, evaluation, and figure generation are available on https://github.com/compomics/ms2pip/tree/v3.11.0/manuscripts/2023/. All training and evaluation data is available on Zenodo at https://doi.org/10.5281/zenodo.7669701. The newly generated CID-TMT data is available from PRIDE with identifier PXD041002 (https://www.ebi.ac.uk/pride/archive/projects/PXD041002).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank all researchers who made their mass spectrometry data publicly available, and we thank Elixir Belgium for featuring MS^{2}PIP as a node service.
Author contributions: Arthur Declercq: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Robbin Bouwmeester: Software, Writing – review & editing, Cristina Chiva: Data curation, Resources, Writing – review & editing, Eduard Sabidó: Data curation, Resources, Writing – review & editing, Aurélie Hirschler: Validation, Writing – review & editing, Christine Carapito: Validation, Writing – review & editing, Lennart Martens: Funding acquisition, Supervision, Writing – review & editing, Sven Degroeve: Conceptualization, Methodology, Supervision, Writing – review & editing, Ralf Gabriels: Methodology, Software, Supervision, Writing – review & editing.

FUNDING

Arthur Declercq, Lennart Martens and Ralf Gabriels acknowledge funding from the Research Foundation Flanders (FWO) [12B7123N, G010023N, G028821N, 1SE3722]; Robbin Bouwmeester acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen [HBC.2020.2205]; Sven Degroeve and Lennart Martens acknowledge funding from the European Union's Horizon 2020 Programme (H2020-INFRAIA-2018-1) [823839]; Lennart Martens acknowledges funding from the Ghent University Concerted Research Action [BOF21/GOA/033]. Eduard Sabidó and Cristina Chiva acknowledge support from the Spanish Ministry of Science, Innovation and Universities (PID2020-115092GB-I00), "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208, and "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2017SGR595). The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech). This work has been supported by EPIC-XS, project number 823839, funded by the Horizon 2020 programme of the European Union. Funding for open access charge: the Research Foundation Flanders (FWO) [G028821N].
Conflict of interest statement. None declared.

REFERENCES

1. Faridi,P., Purcell,A.W. and Croft,N.P. (2018) In immunopeptidomics we need a sniper instead of a shotgun. *Proteomics*, **18**, e1700464.
2. Thingholm,T.E., Palmisano,G., Kjeldsen,F. and Larsen,M.R. (2010) Undesirable charge-enhancement of isobaric tagged phosphopeptides leads to reduced identification efficiency. *J. Proteome Res.*, **9**, 4045–4052.
3. Neely,B.A., Dorfer,V., Martens,L., Bludau,I., Bouwmeester,R., Degroeve,S., Deutsch,E.W., Gessulat,S., Käll,L., Palczynski,P. *et al.* (2022) Toward an integrated machine learning model of a proteomics experiment. *J. Proteome Res.*, **22**, 681–696.
4. Bouwmeester,R., Gabriels,R., Van Den Bossche,T., Martens,L. and Degroeve,S. (2020) The age of data-driven proteomics: how machine learning enables novel workflows. *Proteomics*, **20**, 1900351.
5. Degroeve,S. and Martens,L. (2013) MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, **29**, 3199–3203.
6. Degroeve,S., Maddelein,D. and Martens,L. (2015) MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.*, **43**, W326–W330.
7. Gabriels,R., Martens,L. and Degroeve,S. (2019) Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.*, **47**, W295–W299.
8. Silva,C.A.S., Bouwmeester,R., Martens,L. and Degroeve,S. (2019) Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics*, **35**, 5243–5248.
9. Declercq,A., Bouwmeester,R., Hirschler,A., Carapito,C., Degroeve,S., Martens,L. and Gabriels,R. (2022) MS2Rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. *Mol. Cell. Proteomics*, **21**, 100266.
10. Van Puyvelde,B., Willems,S., Gabriels,R., Daled,S., De Clerck,L., Vande Casteele,S., Staes,A., Impens,F., Deforce,D., Martens,L. *et al.* (2020) Removing the hidden data dependency of DIA with predicted spectral libraries. *Proteomics*, **20**, 1900306.
11. Searle,B.C., Swearingen,K.E., Barnes,C.A., Schmidt,T., Gessulat,S., Küster,B. and Wilhelm,M. (2020) Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.*, **11**, 1–10.
12. 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.
13. Van Puyvelde,B., Van Uytfanghe,K., Tytgat,O., Van Oudenhove,L., Gabriels,R., Bouwmeester,R., Daled,S., Van Den Bossche,T., Ramasamy,P., Verhelst,S. *et al.* (2021) Cov-MS: a community-based template assay for mass-spectrometry-based protein detection in SARS-CoV-2 patients. *JACS Au*, **1**, 750–765.

14. Bouwmeester,R., Gabriels,R., Hulstaert,N., Martens,L. and Degroeve,S. (2021) DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods*, **18**, 1363–1369.
15. Demichev,V., Messner,C.B., Vernardis,S.I., Lilley,K.S. and Ralser,M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, **17**, 41–44.
16. Pino,L.K., Searle,B.C., Bollinger,J.G., Nunn,B., MacLean,B. and MacCoss,M.J. (2020) The Skyline ecosystem: informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.*, **39**, 229–244.
17. Wysocki,V.H., Tsaprailis,G., Smith,L.L. and Breci,L.A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.*, **35**.1399–1406.
18. Ting,L., Rad,R., Gygi,S.P. and Haas,W. (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods*, **8**, 937–940.
19. Martens,L., Hermjakob,H., Jones,P., Adamsk,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
20. 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.