

AnnoMe: user-defined classification of HR-MS/MS spectra for natural product discovery

Christoph Bueschl^{1,*}, Tomas Rypar^{1,2}, Lenka Molcanova³, Juraj Markus³, Bernhard Seidl⁴, Maria Doppler⁴, David Ruso⁴, Christina Maisl¹, Karel Smejkal³, Rainer Schuhmacher¹

¹Department of Agricultural Sciences, Institute of Bioanalytics and Agro-Metabolomics (iBAM), BOKU University, Vienna, 1180, Austria

²Department of Chemistry and Biochemistry, Mendel University in Brno, Brno, 613 00, Czech Republic

³Department of Natural Drugs, Faculty of Pharmacy, Masaryk University, Brno, 612 00, Czech Republic

⁴Core Facility Bioactive Molecules: Screening and Analysis, BOKU University, Tulln, 3400, Austria

*Corresponding author. Department of Agricultural Sciences, Institute of Bioanalytics and Agro-Metabolomics (iBAM), BOKU University, Vienna, 1180, Austria. E-mail: christoph.bueschl@boku.ac.at

Associate Editor: Nurcan Tunçbağ

Abstract

Summary: Annotation of HR-MS/MS spectra is a complex task that can be tackled either by expert interpretation or machine learning models that rely on large spectral/structural databases for training. Frequently, users want to find novel compounds of a particular substance class they are already familiar with. This requires the classification of detected compounds as “relevant” (i.e. belonging to the compound class of interest) or not (i.e. “other”). For such applications, the python-based AnnoMe software is presented that allows users to classify their experimental HR-MS/MS spectra according to their aims. By leveraging a user-curated dataset of “relevant” and “other” reference HR-MS/MS spectra alongside structure-informed embeddings (MS2DeepScore), the package enables rapid and accurate prediction of “relevant” compounds with custom-trained classification models and a majority vote, facilitating exploration of the complex chemical space inherent to LC-HRMS/MS data. This software is demonstrated by predicting putative prenylated flavonoids for prioritization in natural product discovery.

Availability and implementation: Code, documentation, and datasets are available at <https://github.com/chrboku/AnnoMe> and <https://zenodo.org/records/16322488>.

1 Introduction

Natural product discovery strategies rely on prioritization of candidate molecules from liquid chromatography high resolution tandem mass spectrometry (LC-HR-MS/MS) datasets. Researchers often seek “more of the same,” e.g. new metabolites of a particular chemical class, because structural similarity often correlates with bioactivity, mechanistic hypotheses, and downstream validation feasibility (Wolfender *et al.* 2019). However, screening for class-consistent candidates at scale is challenging as spectral libraries are incomplete (Hoffmann *et al.* 2022) and manual, rule-based searches are labor-intensive. Consequently, there is a need for flexible tools that help accelerate semi-targeted dataset exploration. In addition, as resolving the complete chemical structure of novel compounds is laborious and generally requires orthogonal analytical techniques like NMR spectroscopy, the prioritization of the detected metabolites is of utmost importance in natural product discovery (Kim *et al.* 2010).

To tackle this challenge, machine learning (ML) has developed into an essential tool in bioinformatics and is also rapidly

transforming the field of analytical chemistry, particularly for tackling the complexity of LC-HR-MS/MS data. Using analytical characteristics such as mass-to-charge ratio (m/z), HR-MS/MS fragmentation spectra, and retention time, ML enables to predict and interpret compound or sample classifications, and also facilitates structure elucidation where traditional methods are limited by manual annotation challenges (Kim *et al.* 2010, Hong *et al.* 2025, Liu *et al.* 2025).

ML approaches have shown potential to overcome key challenges in compound annotation via *in silico* HR-MS/MS spectra predictions (Wang *et al.* 2021, Murphy *et al.* 2023), structure prediction from fragmentation spectra (Dührkop *et al.* 2019, Stravs *et al.* 2022) to find motifs (i.e. HR-MS/MS fragment subsets) of structural features/substructures (Ortega *et al.* 2025), or improving spectral matching through automated pattern extraction (Huber *et al.* 2021, Guo *et al.* 2023, Bushuiev *et al.* 2026). MS2DeepScore has achieved notable success in clustering large spectral datasets and improving library matching by learning structure-informed spectral embeddings from fragmentation patterns (Huber *et al.* 2021).

Received: 21 October 2025. Revised: 11 March 2026. Accepted: 12 April 2026

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Numerous supervised classifiers have been reported for diverse tasks such as predictions of chemical class (Picache *et al.* 2020), or even antimicrobial resistance directly (Weis *et al.* 2022). Others can predict bioactive scaffolds using molecular fingerprints (Brittin *et al.* 2025), or find bioactive compounds (Brown *et al.* 2022).

In this study, we present the python-based software AnnoMe for supervised classification of HR-MS/MS spectra. By leveraging a user-curated dataset consisting of a set of diverse reference spectra (i.e. derived from “relevant” and “other” compounds) alongside MS2DeepScore embeddings, it enables accurate prediction of novel compounds, facilitating exploration of the complex chemical space according to the chemical class of interest to the user. Among others, explorative studies can aim at prioritizing (i) novel terpenoids, (ii) phenylpropanoids, or (iii) indole alkaloids. Here, an example of the workflow is presented with the search for novel prenylated flavonoid metabolites in *Paulownia tomentosa* fruit extracts.

2 Workflow overview and methods

The AnnoMe package is designed to assist the user in prioritizing metabolic features (i.e. HR-MS/MS spectra) by assigning them to one of two orthogonal classes. The first class “relevant” represents compounds the user is interested in, while compounds the user is not interested in are assigned to the “other” class. What classifies as “relevant” is to be defined by the user and can be, e.g. the presence of a particular substructure. After the ML models have been trained, novel spectra of unknown compounds can also be classified as either “relevant” or “other” guiding

further exploration of the dataset toward the user’s “relevant” compounds. The individual steps of the workflow (Fig. 1) are:

- The user provides sets of HRMS/MS spectra representing the two classes “relevant” and “other” (MGF files). These spectra can be obtained by (i) filtering public spectral repositories, or (ii) be compiled from in-house spectral libraries. For substructure-based filtering, the AnnoMe package provides functions and a graphical user interface to filter based on SMARTS strings (<https://daylight.com/dayhtml/doc/theory/theory.smarts.html>).
- Once the training and testing datasets have been established, MS2Deepscore embeddings are calculated for each HR-MS/MS spectrum. Instead of relying solely on peak intensity overlaps (classical cosine similarity), MS2DeepScore learns to map HR-MS/MS spectra into a 500-dimensional numerical representation (embedding) that reflects the underlying chemical structure. For more details the reader is referred to Huber *et al.* (2021).
- In the next step, different machine learning models (e.g. LDA, SVM, NeuralNet; [Supplementary Information SI1](#), available as [supplementary data](#) at *Bioinformatics Advances* online) are trained on the calculated embeddings, to optimize for correct prediction of the classes based on the training data. Cross-fold validation (the actual strategy is to be defined by the user, stratified CV per default) is carried out and different performance metrics (balanced accuracy, F1 score, confusion matrices) alongside diagnostic plots are automatically generated for final review.
- The user can optionally specify a test dataset consisting of HR-MS/MS spectra of known “relevant” and “other”

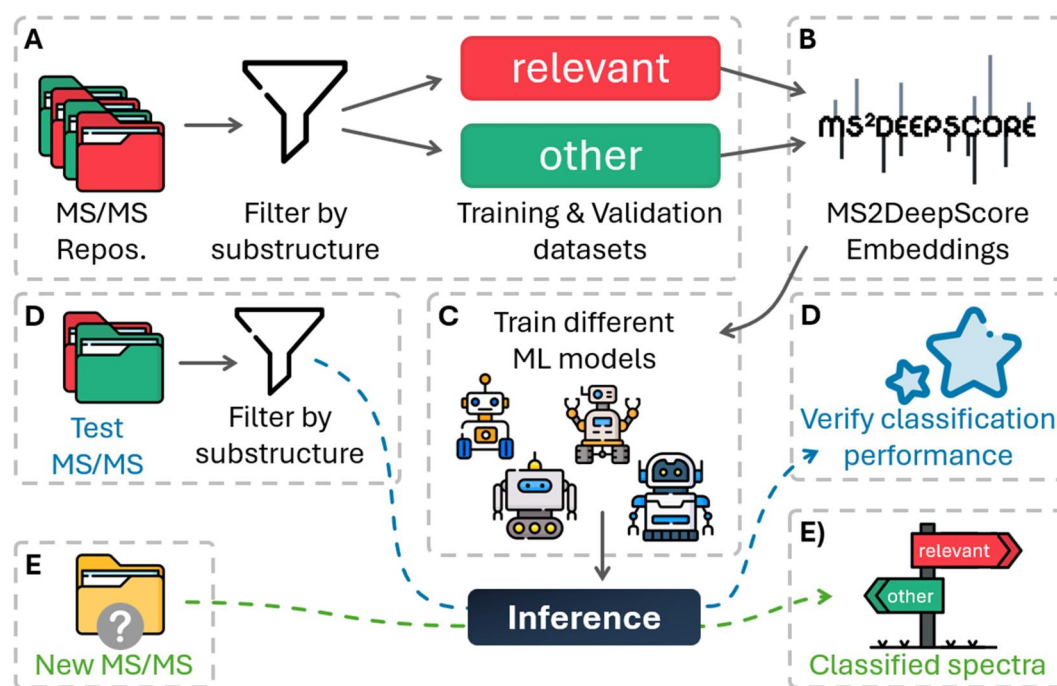


Figure 1 Overview of the workflow. (A) Reference HR-MS/MS datasets of “relevant” and “other” compounds are established. (B) Structural-embeddings are calculated via MS2DeepScore. (C) A series of machine learning models is trained to differentiate between the two classes. (D) An optional test dataset is used to assess the model’s performance. (E) HR-MS/MS spectra of unknowns are classified as “relevant” or “other.”

compounds to test the model's performance on data not used during training.

E. Finally, unclassified, experimental spectra (e.g. from an explorative study) are also classified utilizing the trained model.

The AnnoMe package is implemented in the Python programming language. It utilizes MS2DeepScore for structural embeddings, and the scikit-learn package for classification tasks (Huber *et al.* 2021). It offers modules for downloading common, public HR-MS/MS libraries (e.g. MoNA, MassSpecGym, MSnLib), parsing and overview generation (polarity, fragmentation energies, etc.), and substructure-based filtering via SMARTS strings. The user has the possibility to set hyperparameter values for the classifiers, the cross-validation strategy, and configure the majority vote to avoid overfitting. The software can be used programmatically (e.g. in a Jupyter notebook) or interactively (i.e. graphical user interfaces for (i) substructure-based filtering and (ii) classification training and inference), with the latter also conveniently visualizing matching and non-matching structures for SMARTS string filters and the classifier's performance (F1, accuracy). The software does not require high performance computing resources and has been tested successfully on a standard laptop (Intel Core Ultra 5 125U, 12 cores; 16GB memory; SSD; Win11) with 655 000 loaded HR-MS/MS spectra. The package, documentation, and a comprehensive tutorial (exemplified with prenylated flavonoids) are available at <https://github.com/chrbooku/AnnoMe/releases/tag/v1.0.0>.

3 Results and discussion

3.1 Demonstration of (iso-)flavonoid prediction trained on public repository data

The developed AnnoMe package was tasked with classifying (iso-)flavonoids as an example. For this, the MSnLib (Brungs 2025, <https://zenodo.org/records/15683662>, version v6) dataset was filtered with respective SMART strings (Supplementary Information SI2, available as [supplementary data](#) at *Bioinformatics Advances* online), resulting in 655 (iso-)flavonoids (class “relevant”), while 10 000 non-(iso-)flavonoids were used as the “other” class. Furthermore, another 192 (iso-)flavonoid and 249 “other” compounds were measured in-house and used to test the classifier's performance. Of these, 74 (iso-)flavonoids were unique to the test dataset. For details of the LC-HRMS/MS measurements, please refer to [Supplementary Information SI3-5](#), available as [supplementary](#)

[data](#) at *Bioinformatics Advances* online. Separate classifiers were trained for ionization polarity (pos/neg) and collision energy settings (20 eV, 30 eV, 60 eV, stepped collision energy [20, 45, 70 eV]). After training, the test datasets showed average balanced accuracy rates of 0.866 (sd 3%, min 0.818, max 0.908) for the training set and 0.82 (sd 4%, min 0.76, max 0.87) for the test set across the spectral subsets, with true-positive rates above 55% and 82% on average, while the false-positive rates were below 10% and 5.2% on average, meaning that the classifier did not incorrectly predict a truly “other” compound as a “relevant” one (Table ST1, available as [supplementary data](#) at *Bioinformatics Advances* online), suggesting that the classifier successfully learned to differentiate between (iso-)flavonoid and other compounds.

3.2 Demonstration of prenylated flavonoid prediction trained on public repository data

AnnoMe's performance is further demonstrated for prenylated flavonoids, which are bioactive metabolites, produced by plants such as *Paulownia tomentosa* trees (Cheng *et al.* 2019). The MSnLib was filtered for prenylated flavonoids (class “relevant”), while structures not matching any of the filters were assigned to the class “other.” Furthermore, an independent test set consisting of HR-MS/MS spectra obtained from 64 prenylated flavonoids (class “relevant”) and HR-MS/MS spectra of a wheat ear extract (class “other”; wheat does not produce prenylated flavonoids) was compiled and used to test the performance of the trained ML models (Supplementary Information SI3-5, available as [supplementary data](#) at *Bioinformatics Advances* online). A minor subset of 17 prenylated flavonoids was present in both libraries, while 174 and 47 prenylated flavonoids were unique for the MSnLib or the in-house dataset respectively, making the test set largely distinct from the training set with respect to the “relevant” compounds (Supplementary Information SI6, available as [supplementary data](#) at *Bioinformatics Advances* online, shows different plots of the complementarity of the training and test sets as well as dimensionality reduction plots). The resulting training and test sets consisted of 30 703 and 7312 spectra, respectively. Training was done separately for each ionization polarity. With the graphical user interfaces, most steps (e.g. compound and meta-information filtering) completed in a couple of minutes, while the main steps (calculation of embeddings and training) took 3:45 hours and 40 minutes respectively on a standard laptop (Intel Core Ultra 5 125U, 12 cores; 16GB memory; SSD; Win11).

Table 1 Prediction performance for the prenylated flavonoids trained on the MSnLib.^a

Split	pol.	n_rel.	n_other	TP	FN	TN	FP	bal.acc.
Train	neg	71	10 095	100.0%	0.0%	97.5%	2.5%	0.9874
	pos	28	20 509	100.0%	0.0%	98.4%	1.6%	0.992
Validation	neg	115	3154	76.5%	23.5%	83.5%	16.5%	0.8002
	pos	115	3928	77.4%	22.6%	94.7%	5.3%	0.8606

^a pol: polarity, n_rel: number of “relevant,” n_other: number of “other,” TP: true-positive, FN: false-negative, TN: true-negative, FP: false-positive, bal.acc: balanced accuracy.

Obtained model performance metrics are shown in [Table 1](#). In the training set, the models showed both high sensitivity and specificity, while in the validation set, these values were reduced. Moreover, a test consisting of 155 non-prenylated (iso-) flavonoid compounds from the in-house library served as a challenging ground-truth negative control ([Table ST2](#), available as [supplementary data](#) at *Bioinformatics Advances* online). Here, with the trained classifiers 58% and 67% (negative and positive modes respectively) of the HR-MS/MS spectra were correctly assigned to non-prenylated compounds (“other” class). In summary, the trained models successfully predicted the classes correctly for most of the tested compounds but struggled with structurally similar compounds.

A comparison with SIRIUS ([Dührkop et al. 2019](#), version 6.2.2, parameters are listed in [Supplementary Information S17](#), available as [supplementary data](#) at *Bioinformatics Advances* online) on the test dataset showed that more of the prenylated flavonoids (ground truth) were classified correctly with the AnnoMe-established prediction models (average 77% in the two test sets, [Table ST2](#), available as [supplementary data](#) at *Bioinformatics Advances* online), which is higher than that obtained with SIRIUS (41% of the spectra were predicted as prenylated flavonoids within the ClassyFire ontology, category “most specific class”). The results demonstrate that the trained classifier is suited to predict compounds belonging to the class of prenylated flavonoids, when trained on publicly available reference datasets only. However, it should also be noted that SIRIUS predicts different compound classes simultaneously and is therefore more versatile for comprehensive metabolomics dataset annotation, while AnnoMe can be used as a complementary tool for data exploration and compound prioritization.

3.3 Demonstration with in-house prenylated flavonoids

Similar to using HR-MS/MS spectra from public repositories, classifiers can also be trained on in-house libraries (e.g. different fragmentation method, particular substance class of interest that is under-represented in public libraries). However, as in-house libraries are typically smaller in size than public repositories, care needs to be exercised to make sure that the class “other” is adequately covered.

For this demonstration, an in-house library consisting of HR-MS/MS spectra derived from some 417 natural products, mainly from plants and fungi, plus 64 prenylated flavonoids were used. Additionally, HR-MS/MS spectra obtained from specialized metabolites of a flowering wheat ear were used to enrich the dataset with the class “other.”

Cross-validation showed that the specificity (min 99.4%) was higher than the sensitivity (average 75.7%, min 56.4%, max 87.7%), suggesting that almost all the spectra of the class “other” were correctly predicted ([Table ST3](#), available as [supplementary data](#) at *Bioinformatics Advances* online). Different collision energy setups performed better or worse, thus it is recommended to not merge different collision energy setups for the classification task.

The trained classifier was then utilized to predict prenylated flavonoids in *Glycyrrhiza uralensis* root and *Paulownia tomentosa* fruit extracts ([Table ST4](#), available as [supplementary data](#)

at *Bioinformatics Advances* online). For the two ionization polarities, 231 and 305 HR-MS/MS spectra (25% and 15%) were predicted to be prenylated flavonoids, while for *Paulownia tomentosa* between 171 and 1005 HR-MS/MS spectra (17% to 34%) were predicted to be prenylated flavonoids in the different samples, some of which were subject to be studied further ([Rypar et al. 2025](#)).

3.4 Performance with respect to the number of spectra used for training of the classifier

To estimate the number of HR-MS/MS spectra necessary to achieve high classification performance, the data from public repository training (Section 3.2) was reused and the number of contained spectra randomly reduced in 10% steps. Comparable performance values (on average 62% of the “relevant” class were correctly predicted) were observed for the first five reduction steps (i.e. 100% down to 50% of the HR-MS/MS spectra used as “relevant”), indicating that for the tested dataset, the classifier needed at least some 1000 HR-MS/MS spectra of the “relevant” class to differentiate them from HR-MS/MS spectra of the class “other.” Further details of this investigation are available in [Supplementary Information S18](#), available as [supplementary data](#) at *Bioinformatics Advances* online. It should be noted that the number of 1000 spectra required for highly accurate classification of prenylated flavonoids may not generalize to other substance classes. Thus, users are encouraged to determine the required number of spectra for their specific applications.

4 Conclusion

Many untargeted metabolomics studies are guided by certain chemical structures known to be relevant (i.e. from literature or previous experiments). However, finding such novel compounds is difficult with HR-MS/MS spectra only. The developed AnnoMe software for HR-MS/MS spectra classification offers a practical framework for the metabolomics community. By enabling users to easily filter public datasets, custom-train prediction models, and classify spectra as “relevant” or “other” with high accuracy, the AnnoMe package addresses a common challenge in metabolomics. Validation results demonstrate an average balanced accuracy of 0.84 in correctly predicting compound classes while having a high sensitivity, providing users with confidence in the predictions.

By utilizing public HR-MS/MS spectral repositories, which are nowadays easily accessible, a wealth of data for training is available and SMART codes allow for customizable, fast sampling of relevant structures for training. This significantly improves the classification of compounds and highlights the importance of open data sharing. As many untargeted metabolomics studies are guided by certain chemical structures known to be relevant from literature or previous experiments, annotating such novel compounds is critical.

Acknowledgements

This paper is the result of research conducted within the Interuniversity Research Platform for Agrobiotechnology Tulln (IFA Tulln) by BOKU-University, Vetmeduni, Vienna, and TU Wien. This project was supported by the BOKU Core Facility Bioactive Molecules: Screening and Analysis. The authors thank the metabolomics community for depositing millions of HR-MS/MS spectra to public repositories and their colleagues (Michael Sulyok, Margita Dvorska, Milan Malanik, and Gabriela Salokova) for providing reference standards. Resources from <https://free.pik.com> have been used. The authors also express their gratitude to the anonymous reviewers for their feedback and suggestions to improve the software and manuscript.

Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

Conflicts of interest

None declared.

Funding

L.M., K.S., and J.M. thank for financial support by GAČR 23-04655S “Role of prenylation and glycosylation patterns in anti-inflammatory activity and metabolism of natural phenolic compounds.”

Data availability

Supplementary data and HR-MS/MS spectra of standards, *Paulownia tomentosa*, and *Glycyrrhiza uralensis* extracts are available at *Bioinformatics Advances* online and on Zenodo (<https://dx.doi.org/10.5281/zenodo.16322488>). Raw-data files of the LC-HRMS measurements of the *Paulownia tomentosa* extracts are available on GNPS/MassIVE (<https://massive.ucsd.edu>) under the ID MSV000098757.

Use of AI tools

The authors utilized the OpenAI/GPT-5 and Google/Gemini3 models for implementing the AnnoMe package and for editing the manuscript. The LLM-tools were used for language editing but not to write/draft the manuscript. All AI output was thoroughly reviewed manually prior to use.

References

- Brittin NJ, Anderson JM, Braun DR *et al.* Machine learning-based bioactivity classification of natural products using LC-MS/MS metabolomics. *J Nat Prod* 2025;**88**:361–72. <https://doi.org/10.1021/acs.jnatprod.4c01123>
- Brown KS, Jamieson P, Wu W *et al.* Computation-assisted identification of bioactive compounds in botanical extracts: a case study of anti-inflammatory natural products from hops. *Antioxidants* 2022;**11**:1400.
- Brungs C, Schmid R, Heuckeroth S *et al.* MSnLib: efficient generation of open multi-stage fragmentation mass spectral libraries. *Nat Methods* 2025;**22**:2028–31. <https://doi.org/10.1038/s41592-025-02813-0>
- Bushuiev R, Bushuiev A, Samusevich R *et al.* Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat Biotechnol* 2026;**44**:630–40. <https://doi.org/10.1038/s41587-025-02663-3>
- Cheng CL, Jia XH, Xiao CM *et al.* Paulownia C-geranylated flavonoids: their structural variety, biological activity and application prospects. *Phytochem Rev* 2019;**18**:549–70. <https://doi.org/10.1007/s11101-019-09614-2>
- Dührkop K, Fleischauer M, Ludwig M *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302. <https://doi.org/10.1038/s41592-019-0344-8>
- Guo H, Xue K, Sun H *et al.* Contrastive Learning-Based embedder for the representation of tandem mass spectra. *Anal Chem* 2023;**95**:7888–96. <https://doi.org/10.1021/acs.analchem.3c00260>
- Hoffmann MA, Nothias L-F, Ludwig M *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol* 2022;**40**:411–21. <https://doi.org/10.1038/s41587-021-01045-9>
- Hong Y, Ye Y, Tang H. Machine learning in small-molecule mass spectrometry. *Annu Rev Anal Chem (Palo Alto Calif)* 2025;2025;**18**:193–215. <https://doi.org/10.1146/annurev-anchem-071224-082157>
- Huber F, van der Burg S, van der Hooft JJJ *et al.* MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminform* 2021;**13**:84. <https://doi.org/10.1186/s13321-021-00558-4>
- Kim HK, Choi YH, Verpoorte R. NMR-based metabolomic analysis of plants. *Nat Protoc* 2010;**5**:536–49. <https://doi.org/10.1038/nprot.2009.237>
- Liu Y, De Vijlder T, Bittremieux W *et al.* Current and future deep learning algorithms for tandem mass spectrometry (MS/MS)-based small molecule structure elucidation. *Rapid Commun Mass Spectrom* 2025;**39**:e9120. <https://doi.org/10.1002/rcm.9120>
- Murphy M, Jegelka S, Fraenkel E *et al.* Efficiently predicting high resolution mass spectra with graph neural networks. In: *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, HI: JMLR.org, 2023.
- Ortega LRT, Dietrich J, Wandy J *et al.* Large-scale discovery and annotation of hidden substructure patterns in mass spectrometry profiles. *bioRxiv*, <https://doi.org/10.1101/2025.06.19.659491>, 2025, preprint: not peer reviewed.
- Picache JA, May JC, McLean JA. Chemical class prediction of unknown biomolecules using ion Mobility-Mass spectrometry and machine learning: supervised inference of feature taxonomy from ensemble randomization. *Anal Chem* 2020;**92**:10759–67. <https://doi.org/10.1021/acs.analchem.0c02137>
- Rypar T, Molcanova L, Valkova B *et al.* Targeted isolation of prenylated flavonoids from paulownia tomentosa fruit extracts via

- AI-guided workflow integrating LC-UV-HRMS/MS. *Metabolites* 2025;**15**:616. <https://doi.org/10.3390/metabo15090616>
- Stravs MA, Dührkop K, Böcker S *et al.* MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;**19**: 865–70. <https://doi.org/10.1038/s41592-022-01486-3>
- Wang F, Liigand J, Tian S *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;**93**:11692–700. <https://doi.org/10.1021/acs.analchem.1c01465>
- Weis C, Cuénod A, Rieck B *et al.* Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nat Med* 2022;**28**:164–74. <https://doi.org/10.1038/s41591-021-01619-9>
- Wolfender JL, Litaudon M, Touboul D *et al.* Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery. *Nat Prod Rep* 2019;**36**:855–68. <https://doi.org/10.1039/c9np00004f>