FragHub: A Mass Spectral Library Data Integration Workflow

Axel Dablanc, Solweig Hennechart, Amélie Perez, Guillaume Cabanac, Yann Guitton, Nils Paulhe, Bernard Lyan, Emilien Jamin, Franck Giacomoni, Guillaume Marti

analytical chemistry

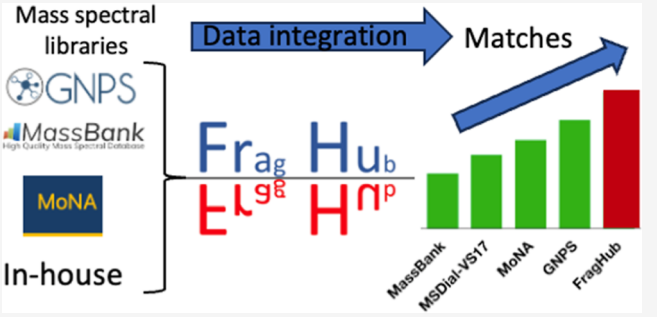
pubs.acs.org/ac
Article

FragHub: A Mass Spectral Library Data Integration Workflow

Axel Dablanc,[▽] Solweig Hennechart,[▽] Amélie Perez, Guillaume Cabanac, Yann Guitton, Nils Paulhe, Bernard Lyan, Emilien L. Jamin, Franck Giacomoni, and Guillaume Marti*

Cite This: *Anal. Chem.* 2024, 96, 12489−12496
Read Online

ABSTRACT: Open mass spectral libraries (OMSLs) are critical for metabolite annotation and machine learning, especially given the rising volume of untargeted metabolomic studies and the development of annotation pipelines. Despite their importance, the practical application of OMSLs is hampered by the lack of standardized file formats, metadata fields, and supporting ontology. Current libraries, often restricted to specific topics or matrices, such as natural products, lipids, or the human metabolome, may limit the discovery potential of untargeted studies. The goal of FragHub is to provide users with the capability to integrate various OMSLs into a single unified format, thereby enhancing the annotation accuracy and reliability. FragHub addresses these challenges by integrating multiple OMSLs into a single comprehensive database, supporting various data formats, and harmonizing metadata. It also proposes some generic filters for the mass spectrum using a graphical user interface. Additionally, a workflow to generate in-house libraries compatible with FragHub is proposed. FragHub dynamically segregates libraries based on ionization modes and chromatography techniques, thereby enhancing data utility in metabolomic research. The FragHub Python code is publicly available under a MIT license, at the following repository: https://github.com/eMetaboHUB/FragHub. Generated data can be accessed at 10.5281/zenodo.11057687.



1. INTRODUCTION

Liquid chromatography−mass spectrometry (LC−MS) chemical profiling provides hundreds to thousands of features (m/z × RT pairs) from a single biological matrix. The process of dereplication, which involves annotating all detected spectral signatures, is a major bottleneck in LC−MS-based metabolomics.[1] Annotations rely on a "body of evidence" approach initially formalized by the Metabolomics Standards Initiative, stratified into four confidence levels: level 1, identified metabolites using authentic standard compounds; level 2, putatively annotated metabolites using public/commercial spectral libraries; level 3, putatively characterized metabolites based on diagnostic ions and/or partial spectral similarities to known compounds of a chemical class; and level 4, unknown metabolites.[2] These confidence levels have been further refined to include new strategies such as the mass spectral similarity network or low library match score (level 2b), in silico-based annotation (level 3), molecular formula match (level 4), and unknown spectral signals (level 5).[3] A comprehensive dereplication may maximize annotation level 1 but involve a LC−MS/MS spectral library setup in identical analytical conditions of matrix chemical profiling and is further limited to pure standard availability. Actually, authentic standard-centric annotation may identify only 1 to 10% of all detected signals in a biological matrix but can be enriched using open mass spectral library (OMSL) resources to fill gaps with annotation level 2.[4]

Many OMSLs are freely available such as GNPS, MassBank, MoNA, RIKEN, and HMDB[5−8] and immensely valuable for dereplication purposes. However, dealing with these resources is challenging due to the lack of standardized file formats and architecture. These libraries encompass a variety of file structures for mass spectral data, including ASCII-based formats like Mascot Generic Format (.MGF) and NIST MSP (.MSP), as well as MassBank records, JavaScript Object Notation (.JSON), and Extensible Markup Language (.XML) or in the form of an SQLite database.[9] While these formats generally follow a similar organizational schema—detailing compound spectra with core metadata on chemical identifiers (SMILES, INCHI, name, or adduct forms), experimental conditions (collision energy, ionization mode, polarity, or instrument type), and extended metadata for experimental measurements (m/z values, MS/MS fragments, and their intensities)—there is no uniformity in metadata field names,

Received: April 28, 2024
Revised: June 16, 2024
Accepted: June 24, 2024
Published: July 19, 2024

Table 1. OMSL List Used to Develop FragHub

spectral library name	URL	file format	version	license	spectra
MoNA	https://mona.fiehnlab.ucdavis.edu/downloads	.JSON	2024.01	CC-BY 4.0	190,359
MS-Dial-VS17	http://prime.psc.riken.jp/compms/msdial/main.html#MSP	.MSP	2022.08	CC-BY 4.0	376,430
GNPS	https://gnps-external.ucsd.edu/gnpslibrary	.MGF	2024.01GNPS only	CC-0 1.0	63,935
MassBank	https://github.com/MassBank/MassBank-data	.MSP	2023.11	CC-BY 4.0	164,261

sequencing, or minimal requirements. This lack of standardization restricts OMSL compatibility with open-source processing software, making it prone to parsing and reading errors. For instance, OpenMS[10] only supports .MGF format, while MS-Dial[11] manages generic .MSP or MassBank records and MZmine[12] imports as .JSON, .MGF, and .MSP files but may face parsing issues. Additionally, each OMSL favors a unique file format with its own metadata structure, based on undocumented and unversioned data models, limiting interoperability among LC−MS processing software and hindering the integrated use of multiple databases. MassBank is one of the few resources to offer guidelines describing these records based on a versioned repository (V2.6.0).

Recently, the Python package MatchMS[13] has proposed a pipeline to harmonize metadata and clean experimental values but focus mainly on data exploration using various MS/MS similarity measures. For metadata enrichment related to chemical identifiers, another Python package MSMetaEnhancer has been added to MatchMS satellites tools.[14] Another shortcoming arises when using an OMSL: extracting a subset of interesting data proves difficult, given that most downloadable files are a concatenation of the two ionization modes, several collision energy methods, several instrument types, and a mix of predicted and experimental data. As a result, despite the great value of using one or several OMSLs, this appears challenging for the dereplication of tandem mass spectra in daily work.

To bridge this gap, we introduce FragHub, a workflow that integrates diverse mass spectral libraries to streamline and enhance the annotation process. FragHub supports multiple OMSL formats (.MSP, .MGF, .JSON, .CSV, and .XML) and harmonizes metadata using RDKit[15] and internal dictionaries. It allows for user-defined filtering options and handles outputs from MZmine's spectral library generation module, ensuring seamless integration of in-house databases. FragHub not only concatenates libraries from diverse sources into a unified format but also classifies the spectra according to chromatographic methods (GC/LC−MS), ionization modes (positive/negative), and data origin (predicted/experimental). Available as a Python package with a straightforward user interface, FragHub supports flexible parameter settings.

The processed libraries are compatible with metabolomics data processing software such as MS-Dial, MZmine3, or Flash Entropy Search[16] but also interoperable with spectral data management software such as PeakForest.[17] A PeakForest instance for FragHub is accessible online, providing tools for viewing, browsing, and filtering spectral data through a Web portal or application programming interfaces (APIs) (available at https://fraghub.peakforest.org/).

2. MATERIALS AND METHODS

FragHub's workflow was meticulously designed to parse and standardize spectral data across various formats, including .MSP, .MGF, .JSON, .CSV, and .XML, as derived from several widely utilized OMSLs. These operations involve detailed metadata normalization steps using RDKit, ensuring that data entries from disparate sources become interoperable. Simply concatenating databases is not satisfactory due to differing metadata architectures and quality. The integration process leverages this diversity to fill gaps and increase the overall information for each compound, improving the accuracy of the annotations. To validate and benchmark our approach, we utilized data sets encompassing over 790,000 spectra, demonstrating FragHub's ability to efficiently process and refine these entries for better usability in metabolic studies.

2.1. OMSL Resources. The workflow was tested with a diversity of public software libraries (different data formats, diversity of metadata); four OMSLs were selected and downloaded in early January 2024 (see Table 1). However, end users may integrate any database, including licensed ones or in-house, if they are in MSP, MGF, JSON, or CSV format. For this purpose, an in-house database was created using MZmine3 to test output compatibility with FragHub. A step-by-step tutorial to create an in-house library is available in Supporting Information (S6: MZmine tutorial). The data set gathered for this work comprises 794,985 MS/MS spectra with the associated metadata.

2.2. FragHub Workflow. The initial step of the FragHub workflow involves parsing various data file formats, such as .MSP, .MGF, .JSON, .CSV, and .XML, into field names and their corresponding values, as delineated in Table 1. The workflow employs a mapping dictionary to translate current keys into standardized keys that adhere to GNPS naming conventions, thereby ensuring compatibility with data reprocessing software like MS-Dial, MZmine, and Flash Entropy Search which utilizes MSP and JSON for annotation.

To effectively manage duplicates and facilitate further data processing, FragHub generates SPLASH keys for each spectrum according to the algorithm of Wohlgemuth and coworkers.[18] All SPLASH keys are recorded in the "update.json" file, which helps in maintaining a repository of treated spectra, ensuring that only new spectra are processed upon the addition of new OMSL entries, as configured by the user. Additionally, this deduplication procedure allows for faster peak annotation by limiting the number of queries within LC−MS processing software.

The workflow conducts a thorough cleaning and normalization of the compound metadata and spectral data. It verifies the accuracy of SMILES, InChI, and InChIKey assignments, reallocating them as needed, and eliminates any spectra lacking both InChI and SMILES. RDKit is utilized to standardize chemical identifiers and calculate both exact and average molecular masses. Unparsable identifiers are removed, and any missing "name" data are substituted with the corresponding molecule's InChI, where applicable. Nonspecific values such as "RT: 0.0" or "adduct: unknown" are replaced with the placeholder "UNKNOWN". The workflow also updates adduct values, ion mode keys, and MS levels using a comprehensive mapping dictionary from the data directory and tentatively

https://doi.org/10.1021/acs.analchem.4c02219
Anal. Chem. 2024, 96, 12489−12496

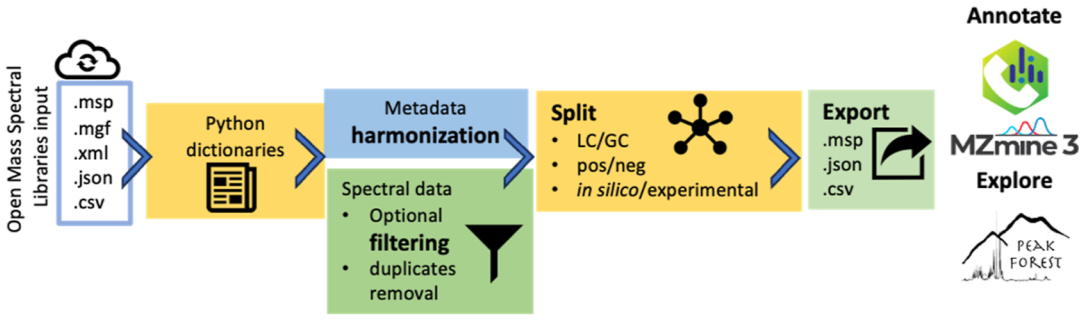


Figure 1. FragHub workflow showing the 4 steps from OMSLs input to export files.

calculates empty m/z precursor values based on the exact mass and identified adduct.

Instrument details (e.g., model types such as QTOF or FT) and ionization modes (such as electrospray ionization or APCI) are normalized using the HUPO PSI mass spectrometry controlled vocabulary via an in-house hierarchical decision tree available in the data directory.

Spectra lacking essential information like SMILES, InChI, or a valid precursor m/z value as well as those failing to meet user-specified filter criteria are excluded. A detailed list of discarded spectra is compiled, highlighting the reasons for their removal.

Furthermore, FragHub annotates the "predicted" field to distinguish between experimental and predicted spectra and normalizes retention times to minutes. Following metadata normalization, user-defined filters are applied through the graphical user interface to refine the peak list (Table S2).

Finally, the workflow segregates the spectra by ion detection mode (positive/negative) and separation techniques (LC or GC) and categorizes them as experimental or predicted, removing any potential duplicates based on similar InChIKeys and their fragment lists. The entire process is efficiently completed in less than 20 min on a desktop computer equipped with an Intel Core i9-13900 and 128 GB RAM DDR5, handling over a million spectra in various test formats (see Figure 1).

2.3. OMSL Benchmarking for Annotation. In order to benchmark each OMSL for annotation purposes on a real data set, raw data from Nicolle et al.[19] were used (10.5281/zenodo.8421008). Quality control (pool of whole *Arabidopsis thaliana* extracts) and blank thermo RAW data were imported into MS-Dial v5.231120. Chromatograms were deconvoluted and aligned using the same parameters as Nicolle et al. and then filtered with the help of integrated MS-CleanR[20] with a blank ratio of 0.8; incorrect mass and ghost peak were removed; a relative standard deviation of 40% and a relative mass defect between 50 and 3500 ppm. The alignment result was submitted to MS/MS-based annotation using each OMSL processed by FragHub applying all default filters and exported in .MSP format. The default MS-Dial parameters were used for spectral matches: Dot product score >600; weighted dot product >600; reverse dot product >800; matched spectrum percentage >25%; and minimum number of matched peaks = 3. End users may tune them according to their needs.

2.4. Chemical Space Representation. Chemical classes were deciphered using the NPclassifier API.[21] PathwayNP and superclassNP were kept for each compound for figure coloration. The t-distributed stochastic neighbor encoding (t-SNE) dimensionality reduction was calculated from PubChem fingerprints using a perplexity of 30 and an exaggeration of 1.

2.5. PeakForest Database. PeakForest is a multiplatform digital infrastructure for interoperable metabolite spectral data and metadata management. It captures and stores different types of metabolomics data from mass spectrometry and nuclear magnetic resonance, providing users with valuable insights into metabolite identification and annotation processes. The infrastructure consists of a structured database, API, a Web interface, and Web services offering tools for browsing, managing, and curating spectral data and metadata. Standardized procedures and formats have been implemented to guarantee information quality and interoperability. These features provide users with intuitive access to spectral data, facilitating efficient data annotation and analysis workflows. Finally, PeakForest is designed to facilitate the centralized storage of data at the laboratory level and to facilitate sharing between laboratories and public databases.

3. RESULTS

FragHub, developed in Python programming language, leverages four widely used OMSLs for LC−MS-based metabolomic analysis. In this study, we specifically utilized GNPS-tagged databases in the .MGF format, comprising 13,507 compounds and 63,935 spectra. MoNA (MassBank of North America) significantly enriches our data set with 21,839 unique compounds across 190,359 spectra, available in .MSP, .SDF, and .JSON formats. MassBank stands out for its spectral diversity, offering over 164,261 spectral data sets associated with 8358 compounds. MS-Dial-VS17 represents a unique integration, merging several databases and in-house acquired spectra accounting for 376,430 spectra and 22,282 compounds. This data set is the only library presplit into positive ionization (PI) and negative ionization modes. For these latter two databases, the .MSP format was utilized within FragHub. To showcase FragHub's adaptability, multiple formats were processed (as detailed in Table 1). The integration of these four OMSLs yields a combined total of 794,985 spectra for 35,673 unique chemical identifiers. The FragHub data integration workflow refines this further to 602,744 spectra for 32,193 unique chemicals, as illustrated in Figure 2. Detailed logs of the spectra excluded during the OMSL processing are listed in Table S4.

Approximately 45% of chemicals are shared between two or more OMSLs, highlighting the interconnected nature of these resources. Conversely, 19,419 compounds are exclusive to a single OMSL (see Figure 3). The FragHub workflow effectively reduces redundancy by eliminating about 200,000 duplicate spectra from an initial pool of 794,985, underscoring

https://doi.org/10.1021/acs.analchem.4c02219
Anal. Chem. 2024, 96, 12489−12496

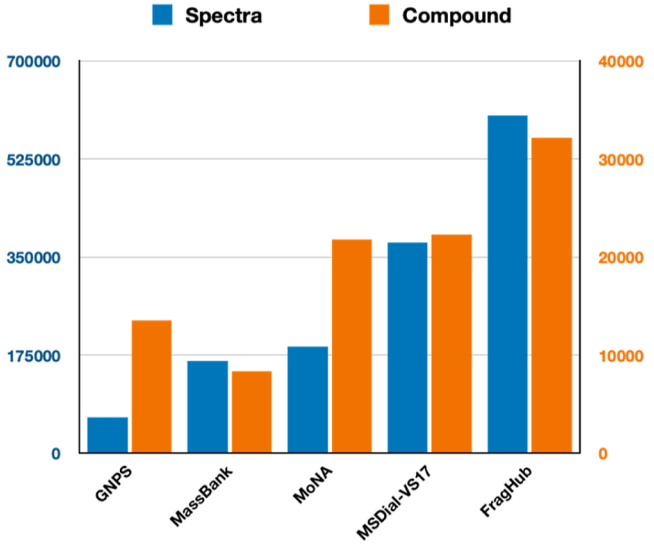


Figure 2. Integration output analysis bar plot displaying the counts of MS/MS spectra and unique InChIKeys derived from each OMSL. The left *y*-axis represents the number of spectra, while the right *y*-axis shows the number of unique chemical identifiers. This visualization underscores the harmonization capabilities of FragHub, demonstrating its efficacy in integrating and deduplicating spectral data from diverse libraries.

the diverse chemical compositions and experimental conditions—such as collision energy, instrument type, and adduct forms of isolated pseudomolecular ions—that characterize each library. The median number of spectra per compound ranges from 2 in GNPS to 12 in MassBank, illustrating significant spectral redundancy that can be tailored based on user preferences.

For example, applying a filter to remove spectra with fewer than three MS/MS signals results in a 15% reduction in the number of entries, as depicted in Figure 4. Further refinement

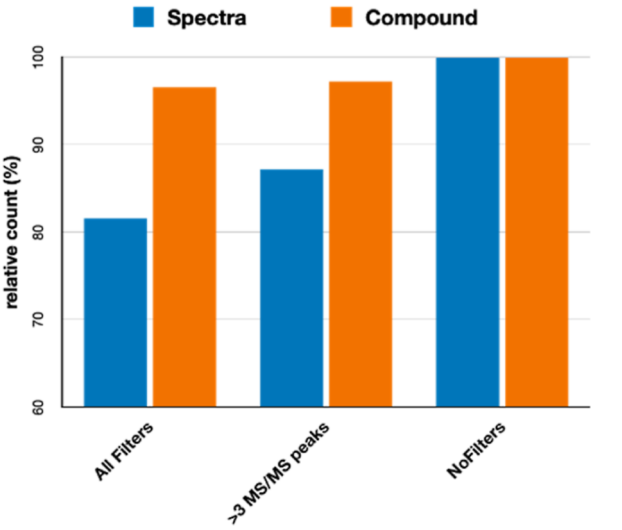


Figure 4. Filter impact analysis bar plot quantifying the impact of applying default FragHub filters on the spectral and compound data retained from integrated OMSLs. The plot compares the percentages of spectra and compounds retained with and without filtering, showcasing the effectiveness of filters in enhancing the data quality without a significant loss of chemical diversity.

is achieved through a second filter, which excludes spectra unless they meet a minimum threshold of three signals and two

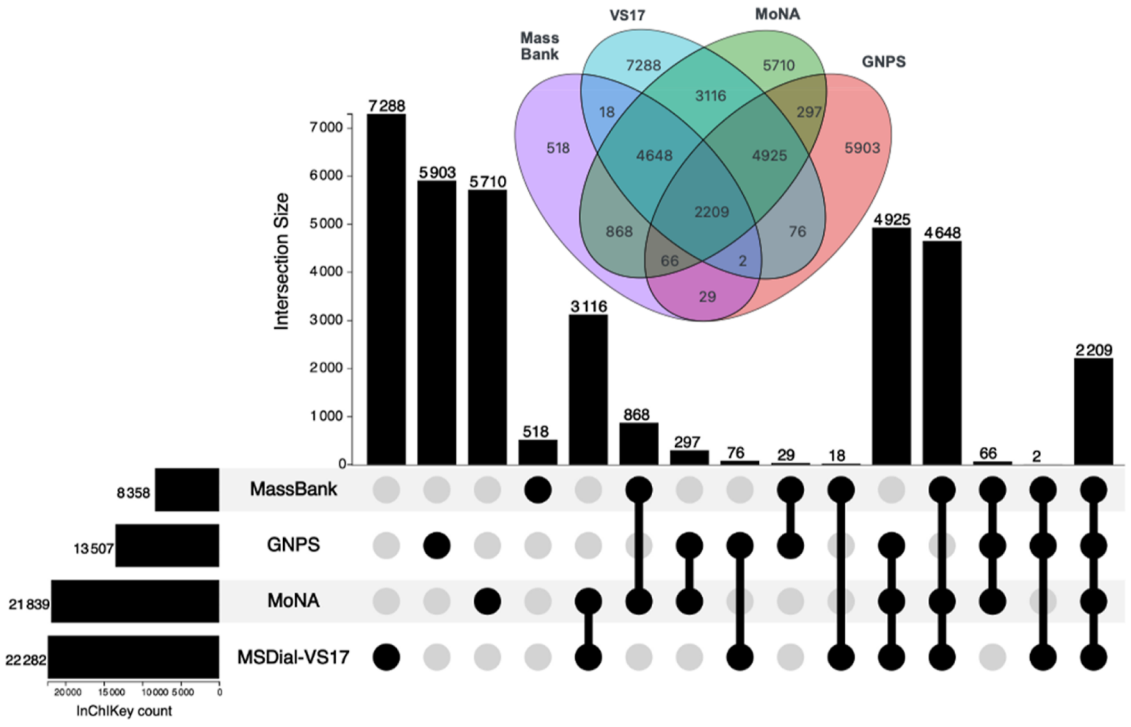


Figure 3. Compound overlap among OMSLs' Venn diagram and UpSet plot illustrating the intersection of unique compounds across various OMSLs. Each bar indicates the number of unique compounds exclusive to a single library or shared between multiple libraries, highlighting the complementary nature of the integrated libraries in covering a broader chemical space.

https://doi.org/10.1021/acs.analchem.4c02219
Anal. Chem. 2024, 96, 12489−12496

MS/MS peaks with intensities above 5%. This stringent criterion retains 81% of the total spectra while incurring a substantial loss of compounds, amounting to 3.5%, thereby optimizing the data set for higher-quality annotations. However, these parameters serve as an example but users may customize or deactivate filtering options to meet specific needs.

To assess the enhanced utility of integrated OMSLs for annotation tasks, we analyzed chemical fingerprints from *A. thaliana* using MS-Dial. The annotations were performed independently on each OMSL as well as on the integrated data set processed through the FragHub workflow. After the application of MS-CleanR filtration, a total of 435 features were detected in PI mode. The annotation process did not consider the retention time values and relied solely on the accurate mass and MS/MS fragmentation patterns.

The outcomes, depicted in Figure 5, demonstrate a direct relationship between the richness of the compound library in each OMSL and the number of matches achieved: MassBank, with its 41 matches, contrasts with the three other OMSLs, which, containing over 5000 unique compounds each, yielded between 55 and 81 matches. Remarkably, the consolidated file from FragHub, utilizing default filtering criteria, successfully annotated 102 features, corresponding to 24% of the total detected features. FragHub's integration and cleaning steps improve metadata quality and speed up the annotation process by deduplicating spectra, resulting in more reliable annotations.

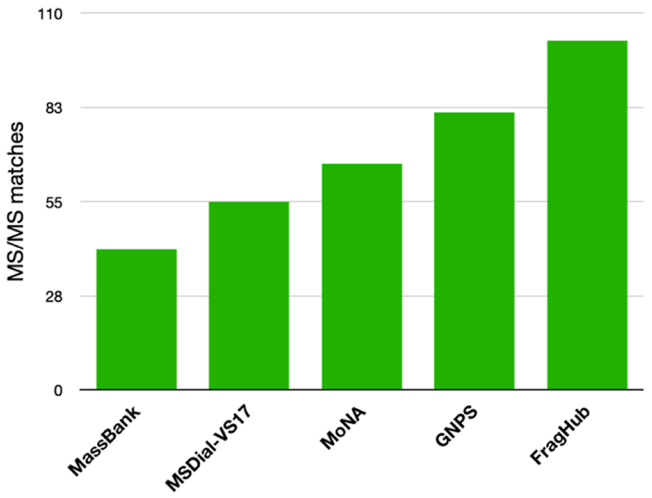


Figure 5. Annotation efficiency comparison bar plot showing the number of features successfully annotated from the Nicolle et al. data set using individual and integrated OMSLs under standard query conditions. This plot demonstrates the increased annotation capabilities achieved through the integrated data set, reflecting FragHub's enhancement for real data set annotation.

The distribution of chemical classes across each OMSL highlights the unique chemical diversity that they cover. Fatty acids predominate in GNPS and MassBank, whereas alkaloids are prominently featured in GNPS and MS-Dial-VS17. MoNA is rich in carbohydrates, amino acids, and peptides. In contrast, shikimates, phenylpropanoids, and terpenoids are more evenly distributed across the OMSLs, as shown in the top panel of Figure 6. The chemical space of each OMSL was analyzed by using a t-SNE dimensionality reduction approach based on PubChem fingerprints. This method effectively reveals local clusters and the overall spatial distribution of compounds, facilitating an intuitive visualization of how different chemical classes aggregate. Typically, compounds within the same class cluster together, with each class occupying distinct regions in the t-SNE plot. GNPS and MS-Dial show denser distributions, particularly in the areas representing terpenoids and alkaloids, whereas MoNA spans a broader area for carbohydrates and MassBank is extensively spread across regions rich in shikimates and phenylpropanoids. Collectively, the integration of these OMSLs through FragHub achieves comprehensive and dense coverage of chemical space across all compound classes.

4. DISCUSSION

The growing number of publications in metabolomics underscores its significance within the omics landscape, yet the relevance of the results stemming from this approach is largely dependent on the quality of annotations derived from spectrometric signals.[22] In this context, OMSLs are key for supporting experimental spectral matching and enhancing annotation rates from untargeted LC−MS fingerprints. The aim of the FragHub workflow is to optimize the use of OMSLs for end users in the field of untargeted LC−MS-based metabolomics. Four OMSLs have been used in various formats to demonstrate the FragHub integration pipeline (Figure 2). A primary challenge in this integration was the normalization of data fields and values from diverse sources. For example, we identified ten distinct keys for ionization states and normalized 487 instrument names and 154 adduct descriptions to 307 and 111, respectively, as detailed in Table S1. The harmonization of collision energies was not addressed due to their varied and nonstandardized measures (around 70 different formats), highlighting the critical need for standardized data practices as recommended in the MassBank documentation, for instance.[23]

Approximately 50% of unique compounds and 20% of spectral duplicates were observed across the OMSLs, indicating that while high redundancy can improve annotation rates, it might also lead to inconsistencies, particularly when using dot product and reverse dot product scoring systems that are highly sensitive to fragment number and intensity. To mitigate these issues, FragHub implements filters that maintain data integrity without compromising compound diversity, as shown in Figure 4. Furthermore, MS/MS data denoising may be applied by plugging FragHub outputs to Libgen[24] or an alternative scoring approach.[25]

The integration of OMSLs used here significantly expands the compound diversity and chemical space coverage and increases the annotation rate of untargeted chemical profiling (Figures 5 and 6). In the context of holistic approaches, deciphering the interplay of metabolome dynamics across organisms or environments is challenging. The use of large mass spectral libraries extends metabolome coverage outside of expected results enabling a comprehensive understanding of complex systems. We measured 32,193 unique compounds after OMSL integration which is rather low compared to the diversity of natural products estimated to be several million molecules.[26] Moreover, the chemical classes covered by OMSLs contrast with the distribution of natural product databases such as the Dictionary of Natural Products with an over-representation of alkaloids and polypeptides in the OMSL, while terpenoids and fatty acids represent the most diverse group in natural product catalogues.[27] This disparity

https://doi.org/10.1021/acs.analchem.4c02219
Anal. Chem. 2024, 96, 12489−12496

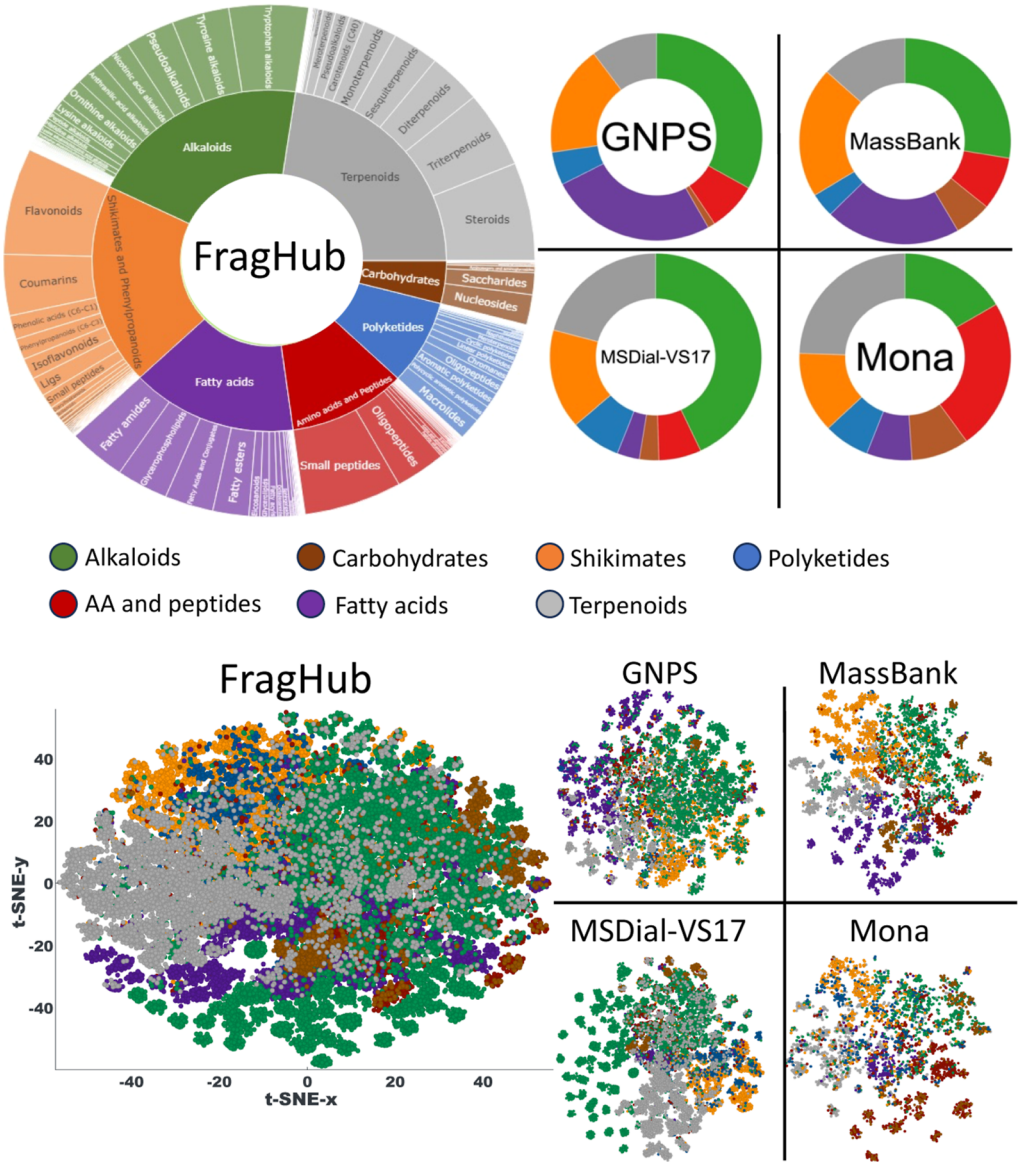


Figure 6. Chemical space coverage t-SNE plots overlaid with donut charts depicting the distribution of metabolite classes within each OMSL and the integrated data set. The t-SNE plots provide a two-dimensional representation of the chemical spaces covered by each library, with colors indicating different chemical classes based on the NPclassifier ontology. The donut charts further detail the proportion of each metabolite class, illustrating the enriched diversity achieved through data integration.

underscores the necessity for orthogonal strategies to fill this gap like raw data digging of mass spectral similarity networks[28] or in silico MS/MS prediction tools based on chemical identifiers.[29] The FragHub integration workflow may help to organize data and explore fragmentation mechanism behavior to set up training sets for deep learning-based strategies.

The FragHub code can handle various input formats and has been multithreaded to process approximately 100,000 spectra per minute (Table S3) which allows the integration of large OMSLs in reasonable time on a personal computer. A simple graphical user interface enables users to select filtering options and data format outputs using distinct profiles. This allows shaping scenarios for specific needs such as in-house database handling or simple .CSV outputs to analyze OMSLs and then filter out specific metadata (e.g., instrument type) and reintegration in .MSP or .JSON formats, for instance.

To demonstrate the potential of this data standardization and structuring work, the compounds and their LC−MS/MS spectra were also imported and stored in a dedicated PeakForest database. The Web application provided enables users, for example, to browse and search for specific chemical names and spectral metadata. It also provides a REST Web service to support massive queries submitted by third-party software or bioinformatics pipelines for metabolomics data annotation. PeakForest was initially developed to store and manage high-quality spectral data in terms of metadata. The FragHub instance of PeakForest can be used to put online a collection of sub-banks in MSP format, compiled, for example, by instrument type. By exploiting the various resources made available by the community and used in the FragHub pipeline, we were able to compile a very large number of MSMS spectra. This work once again highlights the need to open up more and more new spectral data, acquired on recent instruments and supplemented with rich, controlled metadata, in order to increase the annotation coverage of LC−MS fingerprints.

https://doi.org/10.1021/acs.analchem.4c02219
Anal. Chem. 2024, 96, 12489−12496

5. CONCLUSIONS

The integration of multiple mass spectral libraries through FragHub represents a significant advance in the metabolomics field, facilitating a deeper understanding of metabolite environments through enhanced data quality and accessibility. Moreover, FragHub's flexible architecture allows for the rapid incorporation of new data sources, which is critical given the rapid evolution of mass spectrometry libraries. The main objective of FragHub is to allow end users to build customized integrated databases using various resources (in-house, proprietary, or open access). By addressing the key challenges of data standardization and compatibility, FragHub provides researchers with powerful tools to unlock the full potential of metabolomic studies.

■ ASSOCIATED CONTENT

Data Availability Statement

FragHub code can be forked, cloned, or downloaded on GitHub at the following address: https://github.com/eMetaboHUB/FragHub. FragHub is available with a prebuilt data structure to facilitate the end-user processing. A tutorial is available on the GitHub repository and in Supporting Information. OMSLs processed in this study are available on the Zenodo repository: 10.5281/zenodo.11057687.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c02219.

Tables displaying FragHub normalization data, list of FragHub filter functions and their functionality, FragHub benchmarking data, and tracking of spectra deleted by FragHub (PDF)

Tutorial for FragHub installation and usage (PDF)

Tutorial to set up an in-house library using MZmine (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Guillaume Marti − *Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, Auzeville-Tolosane 31320, France; MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse 31000, France;* orcid.org/0000-0002-6321-9005; Email: guillaume.marti@univ-tlse3.fr

Authors

Axel Dablanc − *Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, Auzeville-Tolosane 31320, France; MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse 31000, France*

Solweig Hennechart − *Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, Auzeville-Tolosane 31320, France; MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse 31000, France; Université Toulouse 3—Paul Sabatier, IRIT UMR 5505 CNRS, Toulouse 31062, France*

Amélie Perez − *Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, Auzeville-Tolosane 31320, France; MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse 31000, France*

Guillaume Cabanac − *Université Toulouse 3—Paul Sabatier, IRIT UMR 5505 CNRS, Toulouse 31062, France; Institut Universitaire de France, Paris 75005, France;* orcid.org/0000-0003-3060-6241

Yann Guitton − *Oniris, INRAE, Laberca, Nantes 44300, France;* orcid.org/0000-0002-4479-0636

Nils Paulhe − *Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand F-63000, France*

Bernard Lyan − *Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand F-63000, France*

Emilien L. Jamin − *Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse 31076, France; MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse 31000, France;* orcid.org/0000-0002-4568-9177

Franck Giacomoni − *Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand F-63000, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c02219

Author Contributions

[▽]A.D. and S.H. contributed equally. G.M. proposed the study; A.D. and S.H. developed the Python package; E.J. and B.L. set up value dictionaries; N.P. and F.G. developed the PeakForest instance; A.P. set-up all tutorials; G.M., G.C., and Y.G. benchmarked and reviewed the workflow. The manuscript was written through contributions of all authors and all authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The FragHub and PeakForest project is supported by the French National Facility in Metabolomics & Fluxomics, MetaboHUB (11-INBS-0010), launched by the French Ministry of Research and Higher Education and the French ANR funding agency within the Programme "France 2030". We also acknowledge all contributors to open mass spectral libraries.

■ REFERENCES

(1) Wolfender, J.-L.; Marti, G.; Thomas, A.; Bertrand, S. *J. Chromatogr. A* **2015**, *1382*, 136−164.

(2) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, *3* (3), 211−221.

(3) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48* (4), 2097−2098.

(4) Tsugawa, H.; Rai, A.; Saito, K.; Nakabayashi, R. *Nat. Prod. Rep.* **2021**, *38*, 1729−1759.

(5) Aron, A. T.; Gentry, E.; McPhail, K. L.; Nothias, L. F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F. J. v.d.; J van der Hooft, J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya, C. A.; Martin

H, C.; Gutiérrez, M.; Ulloa, A. M.; Mora, J. A. T.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; I Calderón, A.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. Reproducible Molecular Networking Of Untargeted Mass Spectrometry Data Using GNPS. **2019**, ChemRxiv:9333212.v1. ChemRxiv..

(6) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45* (7), 703−714.

(7) Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M. *Nat. Biotechnol.* **2020**, *38*, 1159−1163.

(8) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H.; Greiner, R.; Gautam, V. *Nucleic Acids Res.* **2022**, *50* (D1), D622−D631.

(9) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. *Mass Spectrom. Rev.* **2018**, *37* (4), 513−532.

(10) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13* (9), 741−748.

(11) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12* (6), 523−526.

(12) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; Sarvepalli, A.; Zhang, Z.; Fleischauer, M.; Dührkop, K.; Wesner, M.; Hoogstra, S. J.; Rudt, E.; Mokshyna, O.; Brungs, C.; Ponomarov, K.; Mutabdžija, L.; Damiani, T.; Pudney, C. J.; Earll, M.; Helmer, P. O.; Fallon, T. R.; Schulze, T.; Rivas-Ubach, A.; Bilbao, A.; Richter, H.; Nothias, L.-F.; Wang, M.; Orešič, M.; Weng, J.-K.; Böcker, S.; Jeibmann, A.; Hayen, H.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; Pluskal, T. *Nat. Biotechnol.* **2023**, *41* (4), 447−449.

(13) Huber, F.; Verhoeven, S.; Meijer, C.; Spreeuw, H.; Castilla, E.; Geng, C.; van der Hooft, J.; Rogers, S.; Belloum, A.; Diblen, F.; Spaaks, J. *J. Open Source Softw.* **2020**, *5* (52), 2411.

(14) Troják, M.; Hecht, H.; Čech, M.; Price, E. J. *J. Open Source Softw.* **2022**, *7* (79), 4494.

(15) Landrum, G. *Rdkit Documentation*. Release, 2013.

(16) Li, Y.; Fiehn, O. *Nat. Methods* **2023**, *20* (10), 1475−1478.

(17) Paulhe, N.; Canlet, C.; Damont, A.; Peyriga, L.; Durand, S.; Deborde, C.; Alves, S.; Bernillon, S.; Berton, T.; Bir, R.; Bouville, A.; Cahoreau, E.; Centeno, D.; Costantino, R.; Debrauwer, L.; Delabrière, A.; Duperier, C.; Emery, S.; Flandin, A.; Hohenester, U.; Jacob, D.; Joly, C.; Jousse, C.; Lagree, M.; Lamari, N.; Lefebvre, M.; Lopez-Piffet, C.; Lyan, B.; Maucourt, M.; Migne, C.; Olivier, M.-F.; Rathahao-Paris, E.; Petriacq, P.; Pinelli, J.; Roch, L.; Roger, P.; Roques, S.; Tabet, J.-C.; Tremblay-Franco, M.; Traïkia, M.; Warnet, A.; Zhendre, V.; Rolin, D.; Jourdan, F.; Thévenot, E.; Moing, A.; Jamin, E.; Fenaille, F.; Junot, C.; Pujos-Guillot, E.; Giacomoni, F. *Metabolomics* **2022**, *18* (6), 40.

(18) Wohlgemuth, G.; Mehta, S. S.; Mejia, R. F.; Neumann, S.; Pedrosa, D.; Pluskal, T.; Schymanski, E. L.; Willighagen, E. L.; Wilson, M.; Wishart, D. S.; Arita, M.; Dorrestein, P. C.; Bandeira, N.; Wang, M.; Schulze, T.; Salek, R. M.; Steinbeck, C.; Nainala, V. C.; Mistrik, R.; Nishioka, T.; Fiehn, O. *Nat. Biotechnol.* **2016**, *34* (11), 1099−1101.

(19) Nicolle, C.; Gayrard, D.; Noël, A.; Hortala, M.; Amiel, A.; Grat, S.; Ru, A.L.; Marti, G.; Pernodet, J.-L.; Lautru, S.; Dumas, B.; Rey, T. Root Associated Streptomyces Produce Galbonolides to Modulate Plant Immunity and Promote Rhizosphere Colonisation. **2024**, bioRxiv:2024.01.20.576418, bioRxiv .

(20) Fraisier-Vannier, O.; Chervin, J.; Cabanac, G.; Puech, V.; Fournier, S.; Durand, V.; Amiel, A.; André, O.; Benamar, O. A.; Dumas, B.; Tsugawa, H.; Marti, G. *Anal. Chem.* **2020**, *92* (14), 9971−9981.

(21) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. *J. Nat. Prod.* **2021**, *84* (11), 2795−2807.

(22) Theodoridis, G.; Gika, H.; Raftery, D.; Goodacre, R.; Plumb, R. S.; Wilson, I. D. *Anal. Chem.* **2023**, *95* (8), 3909−3916.

(23) *MassBank Documentation*. https://massbank.github.io/MassBank-documentation/contributor_documentation.html.

(24) Kong, F.; Keshet, U.; Shen, T.; Rodriguez, E.; Fiehn, O. *Anal. Chem.* **2023**, *95* (46), 16810−16818.

(25) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. *Nat. Methods* **2021**, *18* (12), 1524−1531.

(26) Medema, M. H.; de Rond, T.; Moore, B. S. *Nat. Rev. Genet.* **2021**, *22* (9), 553−571.

(27) Chassagne, F.; Cabanac, G.; Hubert, G.; David, B.; Marti, G. *Phytochem. Rev.* **2019**, *18* (3), 601−622.

(28) Bittremieux, W.; Avalon, N. E.; Thomas, S. P.; Kakhkhorov, S. A.; Aksenov, A. A.; Gomes, P. W. P.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Gauglitz, J. M.; Gerwick, W. H.; Huan, T.; Jarmusch, A. K.; Kaddurah-Daouk, R. F.; Kang, K. B.; Kim, H. W.; Kondić, T.; Mannochio-Russo, H.; Meehan, M. J.; Melnik, A. V.; Nothias, L.-F.; O'Donovan, C.; Panitchpakdi, M.; Petras, D.; Schmid, R.; Schymanski, E. L.; Van Der Hooft, J. J. J.; Weldon, K. C.; Yang, H.; Xing, S.; Zemlin, J.; Wang, M.; Dorrestein, P. C. *Nat. Commun.* **2023**, *14* (1), 8488.

(29) Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. *Nat. Rev. Drug Discovery* **2023**, *22* (11), 895−916.