# FragHub: A mass spectral libraries data integration workflow

Axel Dablanc[1,2]‡, Solweig Hennechart[1,2,3]‡, Amélie Perez[1,2], Guillaume Cabanac[3,4], Yann Guitton[5], Nils Paulhe[6], Bernard Lyan[6], Emilien Jamin[7,2], Franck Giacomoni[6], Guillaume Marti[1,2]*

[1]Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, 24 Chemin de Borde Rouge, Auzeville, 31320, Auzeville-Tolosane, France

[2]MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, Toulouse, France

[3]Université Toulouse 3 – Paul Sabatier, IRIT UMR 5505 CNRS, Toulouse, France

[4]Institut Universitaire de France (IUF), Paris

[5]Oniris,INRAE, Laberca, 44300 Nantes, France

[6]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand F-63000, France

[7]Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France

‡These authors contributed equally

**Abstract**

Open mass spectral libraries (OMSL) 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. 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 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 https://doi.org/10.5281/zenodo.11057687.

*Keywords*

*Open mass spectral library, metabolomics, dereplication, Mass spectrometry, database*

## 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 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 condition of matrix chemical profiling and is further limited to pure standards 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), 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, sequencing, or minimal requirements. This lack of standardization restricts OMSL compatibility with open-source processing software, making them 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 similarities measures.

For metadata enrichment related to chemical identifiers, another Python package MSMetaEnhancer have 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 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 support multiple OMSL formats (.MSP, .MGF, .JSON, .CSV, .XML) and harmonizes metadata using RDKit[15] and internal dictionaries. It allows for user-defined filtering options and handle 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 API (available at https://fraghub.peakforest.org/).

## 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 open mass spectral libraries. These operations involve detailed metadata normalization steps using RDKit, ensuring that data entries from disparate sources become interoperable. To validate and benchmark our approach, we utilized datasets encompassing over 790,000 spectra, demonstrating FragHub's ability to efficiently process and refine these entries for better usability in metabolic studies.

1. Open Mass Spectral Library 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). Additionally, an in-house database was created using MZMine3 to test outputs compatibility with FragHub. A step-by-step tutorial to create an in-house library is available in supplementary data. The dataset gathered for this work comprises 794,985 MS/MS spectra with the associated metadata.

**Tableau 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 | 2024.01 GNPS only | CC-0 1.0 | 63,935 |
| MassBank | https://github.com/MassBank/MassBank-data | .MSP | 2023.11 | CC-BY 4.0 | 164,261 |

## 2. FragHub Workflow

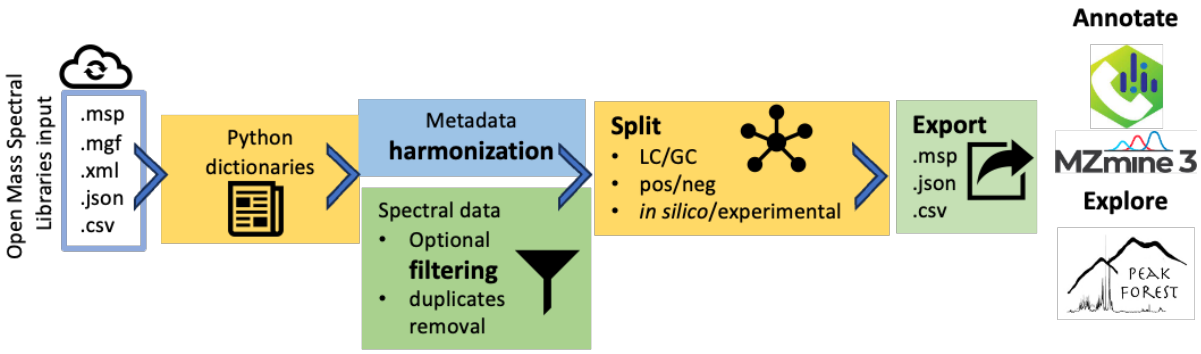


**Figure 1: FragHub workflow showing the 4 steps from OMSLs input to export files**

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 a unique hashing key (SHA-256) for each spectrum using the InChIKey and fragmentation spectra; if an InChIKey is unavailable, the hashing key is derived from all available spectral data. This unique identifier, termed 'FragHubID', simplifies the tracking and elimination of duplicate spectra both within and across OMSLs. FragHubIDs are recorded in the “update.json” file, which helps in maintaining a

repository of processed spectra, ensuring that only new spectra are processed upon the addition of new OMSL entries, as configured by the user.

The workflow conducts a thorough cleaning and normalization of 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. Non-specific 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 calculates empty m/z precursor values based on the exact mass and identified adduct.

Instrument details (e.g., model types like QTOF or FT) and ionization modes (such as ESI 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), 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 twenty minutes 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.

## 3. OMSL benchmarking for annotation

In order to benchmark each OMSL for annotation purposes on a real dataset, raw data from Nicolle et al.[18] were used (https://doi.org/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, aligned using the same parameters as Nicolle *et al.* Then, filtered with the help of integrated MS-CleanR[19] with a blank ratio of 0.8; incorrect mass and ghost peak removed; a relative standard deviation of 40 and a relative mass defect between 50 and 3500. 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 following 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.

4. Chemical space representation

Chemical classes were deciphered using NPclassifier API[20]. PathwayNP and superclassNP were kept for each compound for figures coloration. The t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction was calculated from PubChem fingerprints using a perplexity of 30 and an exaggeration of 1.

5. PeakForest database

PeakForest is a multi-platform 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 (NMR), providing users with valuable insights into metabolite identification and annotation processes. The infrastructure consists of a structured database, Application Programming Interfaces (API), a web interface and web services offering tools for browsing, managing and curating spectral data and metadata. Standardised 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 centralisation of data at laboratory level and to facilitate sharing between laboratories and public databases.

## Results

FragHub, developed in the Python programming language, leverages four widely used open mass spectral libraries (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 dataset 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 datasets associated with 8,358 compounds. MSDial-VS17 represents a unique integration, merging several databases and in-house acquired spectra accounting for 376,430 spectra and 22,282 compounds. This dataset is the only library pre-split into positive ionization (PI) and negative ionization (NI) modes. For these latter two databases, the .MSP format has been 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 OMSLs processing are maintained in Table S4.

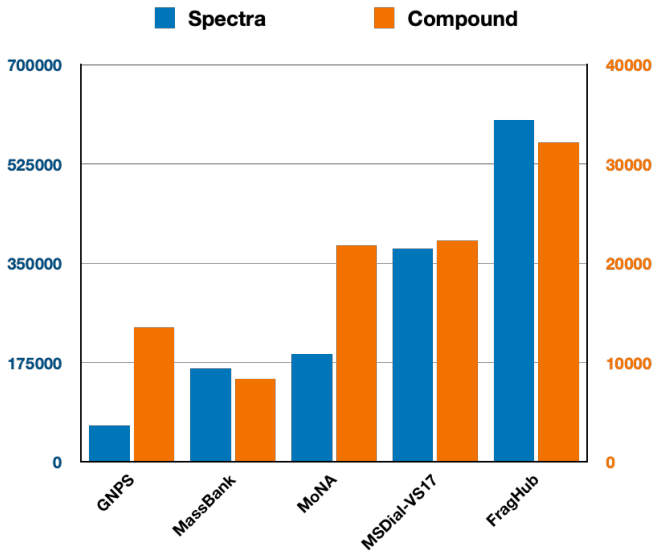


**Figure 2: Integration Output Analysis**

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.

Approximately 45% of chemicals are shared between two or more open mass spectral libraries (OMSLs), highlighting the interconnected nature of these resources. Conversely, 19,419 compounds are exclusive to a single OMSL. The FragHub workflow effectively reduces redundancy by eliminating about 200,000 duplicate spectra from an initial pool of 794,985, underscoring the diverse chemical compositions and experimental conditions—such as collision energy, instrument type, and adduct forms

of isolated pseudo-molecular 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.

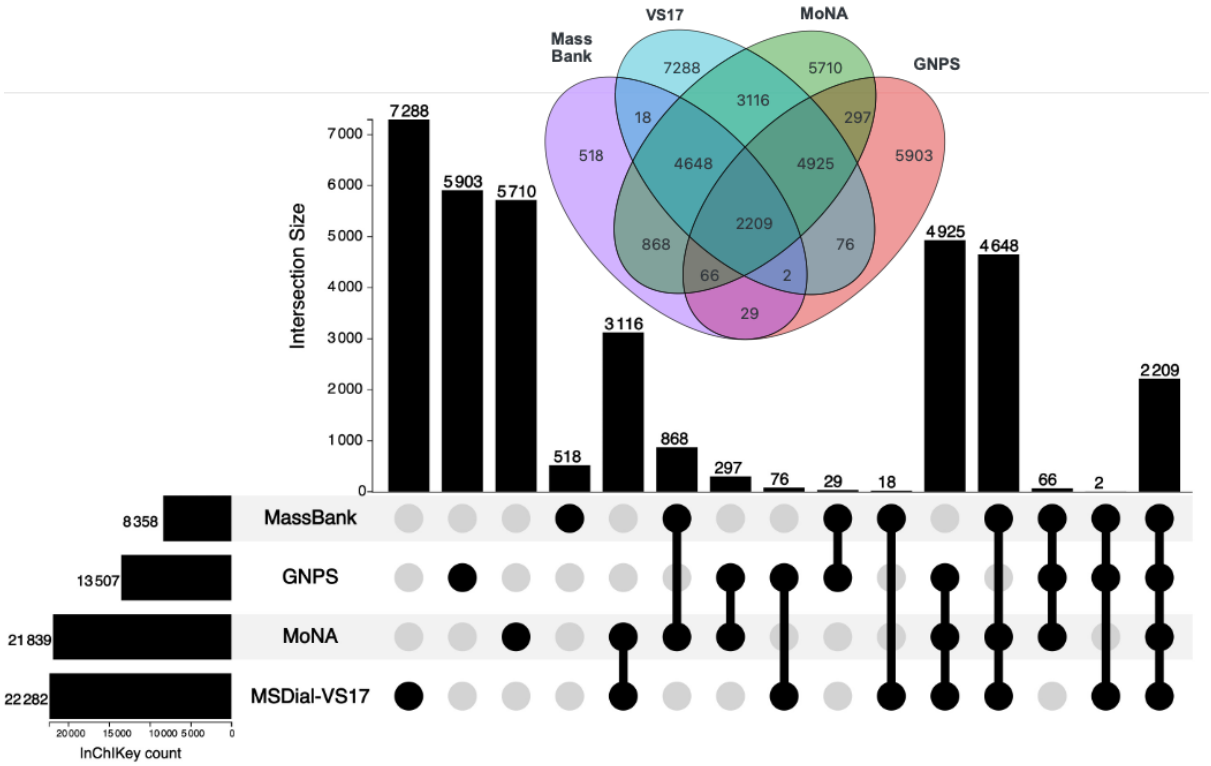


**Figure 3: Compound Overlap among OMSLs**

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 broader chemical space.

For example, applying a filter to remove spectra with fewer than three MS/MS signals results in a 15% reduction in entries, as depicted in Figure 4. Further refinement is achieved through a second filter, which excludes spectra unless they meet a minimum threshold of three signals and two 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 dataset for higher-quality annotations.

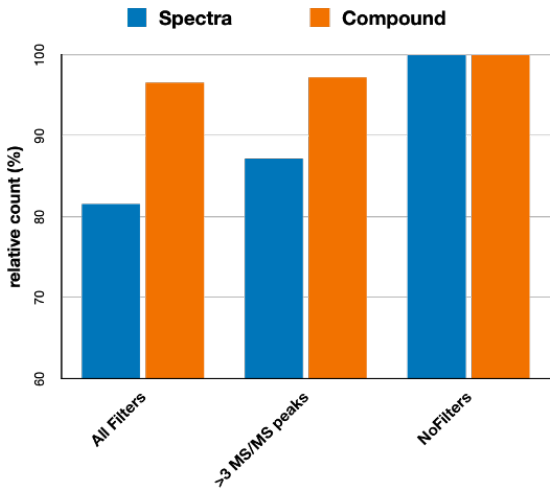


**Figure 4: Filter Impact Analysis**

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 data quality without significant loss of chemical diversity.

To assess the enhanced utility of integrated OMSLs for annotation tasks, we analyzed chemical fingerprints from *Arabidopsis thaliana* using MS-Dial. The annotations were performed independently on each OMSL as well as on the integrated dataset processed through the FragHub workflow. After applying MS-CleanR filtration, a total of 435 features were detected in positive ionization mode. The annotation process did not consider the retention time values and relied solely on 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 5,000 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.

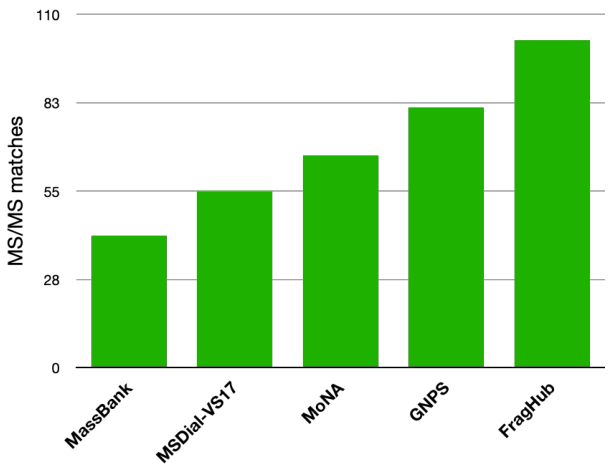


**Figure 5: Annotation Efficiency Comparison**

Bar plot showing the number of features successfully annotated from the Nicolle et al. dataset using individual and integrated OMSLs under standard query conditions. This plot demonstrates the increased annotation capabilities achieved through the integrated dataset, reflecting FragHub's enhancement for real dataset annotation.

The distribution of chemical classes across each OMSL highlights the unique chemical diversity they cover. Fatty acids predominate in GNPS and MassBank, whereas alkaloids are prominently featured in GNPS and MSDial-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 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 MSDial 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 a comprehensive and dense coverage of chemical space across all compound classes.

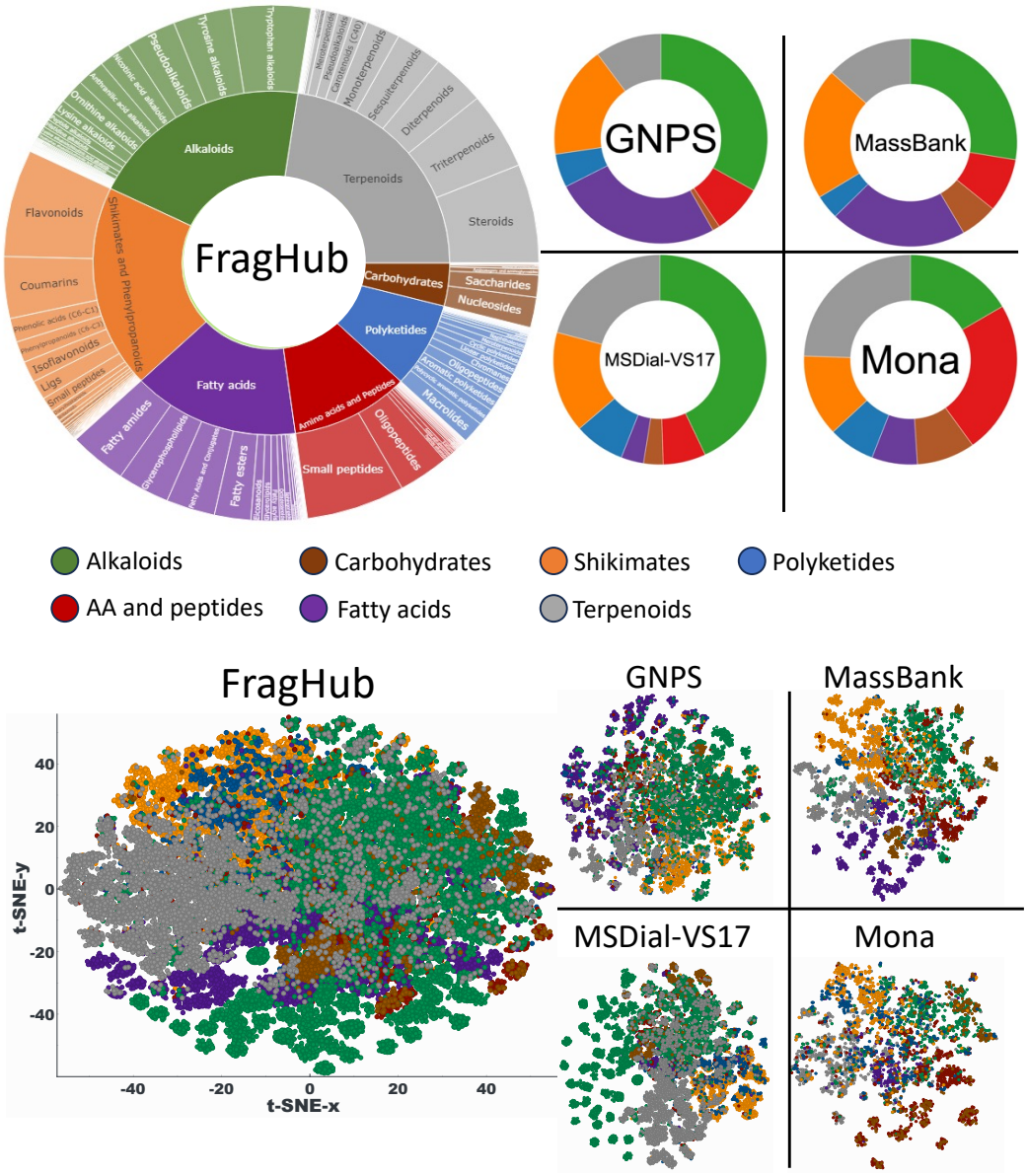


**Figure 6: Chemical Space Coverage**

t-SNE plots overlaid with donut charts depicting the distribution of metabolite classes within each OMSL and the integrated dataset. 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 NPclassifier ontology. The donut charts further detail the proportion of each metabolite class, illustrating the enriched diversity achieved through data integration.

## 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[21]. In this context, OMSLs are key for supporting experimental spectral matching and enhancing annotation rate from untargeted LC-MS fingerprints. The aim of FragHub workflow is to optimize the use of OMSLs for end-users in the field of untargeted LC-MS based metabolomic. 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 non-standardized measures (around 70 different formats), highlighting the critical need for standardized data practices as recommended by MassBank.as recommended in the MassBank documentation for instance[22].

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[23] or alternative scoring approach[24].

The integration of OMSLs used here significantly expands the compound diversity and chemical space coverage and increase annotation rate of untargeted chemical profiling (Figure 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 OMSLs integration which is rather low compared to the diversity of natural products estimated to be several million molecules[25]. Moreover, the chemical classes covered by OMSLs contrast with the distribution of natural products databases such as the Dictionary of Natural Products with an over-representation of alkaloids and polypetides in OMSL, while terpenoids and fatty acids represent the most diverse group in natural product catalogues[26]. This disparity underscores the necessity for orthogonal strategies to fill this gap like raw data digging of mass spectral similarity networks[27] or *in silico* MS/MS prediction tools based on chemical identifiers[28]. 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, then filter on 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-MSMS 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 or

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 has been 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 annotation coverage of LC-MS fingerprints.

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. By addressing the critical challenges of data standardization and compatibility, FragHub provides researchers with powerful tools to unlock the full potential of metabolomic studies.

## References


(1) Wolfender, J.-L.; Marti, G.; Thomas, A.; Bertrand, S. Current Approaches and Challenges for the Metabolite Profiling of Complex Natural Extracts. J. Chromatogr. A 2015, 1382, 136–164. https://doi.org/10.1016/j.chroma.2014.10.091.

(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. Proposed Minimum Reporting Standards for Chemical Analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 2007, 3 (3), 211–221. https://doi.org/10.1007/s11306-007-0082-2.

(3) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ. Sci. Technol. 2014, 48 (4), 2097–2098. https://doi.org/10.1021/es5002105.

(4) Tsugawa, H.; Rai, A.; Saito, K.; Nakabayashi, R. Metabolomics and Complementary Techniques to Investigate the Plant Phytochemical Cosmos. Nat. Prod. Rep. 2021, 10.1039.D1NP00014D. https://doi.org/10.1039/D1NP00014D.

(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. 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.; preprint; 2019. https://doi.org/10.26434/chemrxiv.9333212.v1.

(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. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. J. Mass Spectrom. 2010, 45 (7), 703–714. https://doi.org/10.1002/jms.1777.

(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. A Lipidome Atlas in MS-DIAL 4. Nat. Biotechnol. 2020, 1–5. https://doi.org/10.1038/s41587-020-0531-2.

(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. B.; Greiner, R.; Gautam, V. HMDB 5.0: The Human Metabolome Database for 2022. Nucleic Acids Res. 2022, 50 (D1), D622–D631. https://doi.org/10.1093/nar/gkab1062.

(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. Identification of Small Molecules Using Accurate Mass MS/MS Search. Mass Spectrom. Rev. 2018, 37 (4), 513–532. https://doi.org/10.1002/mas.21535.

(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. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. Nat. Methods 2016, 13 (9), 741–748. https://doi.org/10.1038/nmeth.3959.

(11) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. Nat. Methods 2015, 12 (6), 523–526. https://doi.org/10.1038/nmeth.3393.

(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. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. Nat. Biotechnol. 2023, 41 (4), 447–449. https://doi.org/10.1038/s41587-023-01690-2.

(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. Matchms - Processing and Similarity Evaluation of Mass Spectrometry Data. J. Open Source Softw. 2020, 5 (52), 2411. https://doi.org/10.21105/joss.02411.

(14) Troják, M.; Hecht, H.; Čech, M.; Price, E. J. MSMetaEnhancer: A Python Package for Mass Spectrametadata Annotation. J. Open Source Softw. 2022, 7 (79), 4494. https://doi.org/10.21105/joss.04494.

(15) Landrum, G. Rdkit Documentation. Release 2013, 1 (1–79), 4.

(16) Li, Y.; Fiehn, O. Flash Entropy Search to Query All Mass Spectral Libraries in Real Time. Nat. Methods 2023, 20 (10), 1475–1478. https://doi.org/10.1038/s41592-023-02012-9.

(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. PeakForest: A Multi-Platform Digital Infrastructure for Interoperable Metabolite Spectral Data and Metadata Management. Metabolomics 2022, 18 (6), 40. https://doi.org/10.1007/s11306-022-01899-3.

(18) Nicolle, C.; Gayrard, D.; Noël, A.; Hortala, M.; Amiel, A.; Grat-Simeone, S.; Le Ru, A.; 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. https://doi.org/10.1101/2024.01.20.576418.

(19) 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. MS-CleanR: A Feature-Filtering Workflow for Untargeted LC–MS Based Metabolomics. Anal. Chem. 2020, 92 (14), 9971–9981. https://doi.org/10.1021/acs.analchem.0c01594.

(20) 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. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J. Nat. Prod. 2021, 84 (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.

(21) Theodoridis, G.; Gika, H.; Raftery, D.; Goodacre, R.; Plumb, R. S.; Wilson, I. D. Ensuring Fact-Based Metabolite Identification in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. Anal. Chem. 2023, 95 (8), 3909–3916. https://doi.org/10.1021/acs.analchem.2c05192.

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

(23) Kong, F.; Keshet, U.; Shen, T.; Rodriguez, E.; Fiehn, O. LibGen: Generating High Quality Spectral Libraries of Natural Products for EAD-, UVPD-, and HCD-High Resolution Mass Spectrometers. Anal. Chem. 2023, 95 (46), 16810–16818. https://doi.org/10.1021/acs.analchem.3c02263.

(24) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral Entropy Outperforms MS/MS Dot Product Similarity for Small-Molecule Compound Identification. Nat. Methods 2021, 18 (12), 1524–1531. https://doi.org/10.1038/s41592-021-01331-z.

(25) Medema, M. H.; de Rond, T.; Moore, B. S. Mining Genomes to Illuminate the Specialized Chemistry of Life. Nat. Rev. Genet. 2021, 22 (9), 553–571. https://doi.org/10.1038/s41576-021-00363-7.

(26) Chassagne, F.; Cabanac, G.; Hubert, G.; David, B.; Marti, G. The Landscape of Natural Product Diversity and Their Pharmacological Relevance from a Focus on the Dictionary of Natural Products®. Phytochem. Rev. 2019, 18 (3), 601–622. https://doi.org/10.1007/s11101-019-09606-2.

(27) 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. Open Access Repository-Scale Propagated Nearest Neighbor Suspect Spectral Library for Untargeted Metabolomics. Nat. Commun. 2023, 14 (1), 8488. https://doi.org/10.1038/s41467-023-44035-y.

(28) 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. Artificial Intelligence for Natural Product Drug Discovery. Nat. Rev. Drug Discov. 2023, 22 (11), 895–916. https://doi.org/10.1038/s41573-023-00774-7.

## SUPPORTING INFORMATION

- Supplementary Tables comprising table S1 to S4 in .PDF
- Tutorial for FragHub installation and usage in .PDF
- Tutorial to set-up in-house library using MZMine in .PDF

## AVAILABILITY

FragHub code can be forked, cloned or downloaded on GitHub at the following address: https://github.com/eMetaboHUB/FragHub.

FragHub is available with a pre-built data structure to facilitate the end-user processing. A tutorial is available on GitHub repository and in supplementary data.

OMSLs processed in this study are available on Zenodo repository: https://doi.org/10.5281/zenodo.11057687.

## AUTHOR INFORMATION

### Corresponding Author

* Guillaume Marti, guillaume.marti@univ-tlse3.fr Laboratoire de Recherche en Sciences Végétales, Metatoul-AgromiX Platform, Université de Toulouse, CNRS, INP, 24 Chemin de Borde Rouge, Auzeville, 31320, Auzeville-Tolosane, France.

## Acknowledgment

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.

## Author Contributions

GM proposed the study; AD and SH developed the python package; EJ and BL setup values dictionaries; NP and FG developed the peakforest instance; GM, GC and YG 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.