

PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management

Nils Paulhe, Cécile Canlet, Annelaure Damont, Lindsay Peyriga, Stéphanie Durand, Catherine Deborde, Sandra Alves, Stephane Bernillon, Thierry Berton, Raphael Bir, et al.

▶ **To cite this version:**

Nils Paulhe, Cécile Canlet, Annelaure Damont, Lindsay Peyriga, Stéphanie Durand, et al.. PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management. Metabolomics, 2022, 18 (6), pp.40. 10.1007/s11306-022-01899-3 . hal-03695453

HAL Id: hal-03695453
https://hal.inrae.fr/hal-03695453

Submitted on 14 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Metabolomics (2022) 18:40
https://doi.org/10.1007/s11306-022-01899-3

SOFTWARE/DATABASE

PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management

Nils Paulhe[1] · Cécile Canlet[2] · Annelaure Damont[3] · Lindsay Peyriga[4] · Stéphanie Durand[1] · Catherine Deborde[5] · Sandra Alves[3] · Stephane Bernillon[5] · Thierry Berton[5] · Raphael Bir[1] · Alyssa Bouville[2] · Edern Cahoreau[4] · Delphine Centeno[1] · Robin Costantino[2] · Laurent Debrauwer[2] · Alexis Delabrière[3] · Christophe Duperier[1] · Sylvain Emery[1] · Amelie Flandin[5] · Ulli Hohenester[3] · Daniel Jacob[5] · Charlotte Joly[1] · Cyril Jousse[1] · Marie Lagree[1] · Nadia Lamari[5] · Marie Lefebvre[5] · Claire Lopez-Piffet[1] · Bernard Lyan[1] · Mickael Maucourt[5] · Carole Migne[1] · Marie-Francoise Olivier[3] · Estelle Rathahao-Paris[3] · Pierre Petriacq[5] · Julie Pinelli[5] · Léa Roch[5] · Pierrick Roger[3] · Simon Roques[5] · Jean-Claude Tabet[3] · Marie Tremblay-Franco[2] · Mounir Traïkia[1] · Anna Warnet[3] · Vanessa Zhendre[5] · Dominique Rolin[5] · Fabien Jourdan[2] · Etienne Thévenot[3] · Annick Moing[5] · Emilien Jamin[2] · François Fenaille[3] · Christophe Junot[3] · Estelle Pujos-Guillot[1] · Franck Giacomoni[1]

Received: 22 February 2022 / Accepted: 22 May 2022
© The Author(s) 2022

Abstract

Introduction Accuracy of feature annotation and metabolite identification in biological samples is a key element in metabolomics research. However, the annotation process is often hampered by the lack of spectral reference data in experimental conditions, as well as logistical difficulties in the spectral data management and exchange of annotations between laboratories.

Objectives To design an open-source infrastructure allowing hosting both nuclear magnetic resonance (NMR) and mass spectra (MS), with an ergonomic Web interface and Web services to support metabolite annotation and laboratory data management.

Methods We developed the PeakForest infrastructure, an open-source Java tool with automatic programming interfaces that can be deployed locally to organize spectral data for metabolome annotation in laboratories. Standardized operating procedures and formats were included to ensure data quality and interoperability, in line with international recommendations and FAIR principles.

Results PeakForest is able to capture and store experimental spectral MS and NMR metadata as well as collect and display signal annotations. This modular system provides a structured database with inbuilt tools to curate information, browse and reuse spectral information in data treatment. PeakForest offers data formalization and centralization at the laboratory level, facilitating shared spectral data across laboratories and integration into public databases.

Conclusion PeakForest is a comprehensive resource which addresses a technical bottleneck, namely large-scale spectral data annotation and metabolite identification for metabolomics laboratories with multiple instruments. PeakForest databases can be used in conjunction with bespoke data analysis pipelines in the Galaxy environment, offering the opportunity to meet the evolving needs of metabolomics research. Developed and tested by the French metabolomics community, PeakForest is freely-available at https://github.com/peakforest.

Keywords Curation · Database · FAIR · Interoperability · Metabolite identification · Spectral library

1 Introduction

Over the last 20 years, untargeted metabolomics has developed into a powerful phenotyping tool to better understand biological systems and identify associated biomarkers. This approach, based on multiple analytical platforms, generates massive and complex data that need appropriate analyses to

✉ Franck Giacomoni
Franck.Giacomoni@inrae.fr

Extended author information available on the last page of the article

Published online: 14 June 2022

extract biologically-meaningful information (Alonso et al., 2015). In particular, downstream analysis of metabolomics data requires annotation and identification of features in metabolic profiles. In order to move towards standardized identification methods, the metabolomics community has proposed a definition of metabolite identification accuracy ranges from unknown compounds to confidently-identified compounds (Sumner et al., 2007). This classification undergoes regular amendments, led by the Metabolomics Society and international consortia (Creek et al., 2014; Malinowska & Viant, 2019), in order to reduce ambiguities, better account for chemical structures and improve metabolite annotation confidence. However, metabolite annotation remains a major bottleneck in untargeted mass spectrometry (MS) and nuclear magnetic resonance (NMR) metabolomics (Dona et al., 2016; Nash & Dunn, 2019), and the development of workflows and dedicated tools is critical for accurate metabolite identification (Misra, 2021).

In high-resolution MS or NMR datasets, the first annotation step generally consists of matching experimental accurate masses or chemical shifts with those contained in public and commercial databases. Vendors and associated companies offer a large number of solutions to mine MS or NMR spectral libraries directly from their own instruments (e.g. NIST™) and proprietary databases such as Chenomx™. However, these solutions can lack interoperability with academic bioinformatic tools and often require specific acquisition conditions that oblige operators to adopt operating procedures that differ from experimental conditions. Public libraries have the advantage of hosting large-scale data from different organisms and technical instruments, and include a multiplicity of cross-references concerning biological or chemical information (Vinaixa et al., 2016). Valuable metabolomics resources in the field include MS libraries such as Wishart laboratory databases (Wishart et al., 2018), the Global Natural Product Social Molecular Networking (GNPS, Wang et al., 2016), LIPID MAPS (Fahy et al., 2009), MassBank (Horai et al., 2010), Metlin (Guijas et al., 2018), MoNA or mzCloud™. The NMR community shares resources with MS, such as the human metabolome database (Wishart et al., 2009), and NMR-specific banks such as the BioMagResBank (Ulrich et al., 2007), the Birmingham Metabolite Library (Ludwig et al., 2012), and more generalist banks including NMRShiftDB (Kuhn & Schlörer, 2015) for organic compounds. This relative abundancy of resources represents a large data heterogeneity, from "in silico" spectra derived from modeling and/or references to highly-curated spectra obtained from pure compounds, and remains far from containing experimental data of all known metabolites (Vinaixa et al., 2016). As with commercial resources, the exchange and interoperability of annotations from one laboratory to another can be difficult due to different formatting requirements. Moreover, existing databases are not always easy to increment with new compounds or spectra from external users due to logistical constraints or lack of recommendations (Johnson & Lange, 2015; Spicer et al., 2017).

Database interoperability simplifies the mining of multiple databases, promotes efficient use of metabolomics data and is at the heart of FAIR guidelines ("Findable, Accessible, Interoperable, Reusable"). At the metadata level, common vocabulary and consensual description levels in data collecting steps are required (Alseekh et al., 2021). At the computing level, common formats and application programming interfaces (API) are needed to enable data exchange between databases and connect databases to data treatment tools (Anwar et al., 2021; Merlet et al., 2016). Recent calls in the metabolomics community emphasize the need to develop adapted informatics infrastructures for laboratories based on FAIR principles in order to improve the exchange and interoperability of annotations from one laboratory to another and the sharing (and inclusion) of local spectral libraries with reference databases (Haug et al., 2017; Sansone et al., 2012). However, laboratory-based systems can be limited in their capacity to centralize all the descriptive and analytical characteristics of reference compounds, and may face difficulties in exporting spectral data in recommended formats such as mzML (Martens et al., 2011) or nmrML (Schober et al., 2018).

In this paper we present PeakForest, an open-source and open access infrastructure which hosts for the first time both NMR and mass spectra, with an ergonomic Web interface and Web services to support metabolite annotation and laboratory data management. This resource, deployable locally, facilitates the production of high-quality spectral records and their submission to international repositories such as MassBank Europe[1] or MassBank of North America[2] (MoNA). Building on the expertise and the experience of members of the French national metabolomics and fluxomics infrastructure (MetaboHUB), PeakForest integrates database interoperability at the metadata- and the computing-level, in order to further the implementation of FAIR spectral databases within the metabolomics community.

2 Materials and methods

PeakForest is a modular framework including a database, a graphical user interface and a Web API. It is designed to ensure interoperability, easy deployment and code sustainability. Particular care has been taken to ensure resource security and intellectual property (data author, licensing,

[1] https://massbank.eu.

[2] https://mona.fiehnlab.ucdavis.edu.

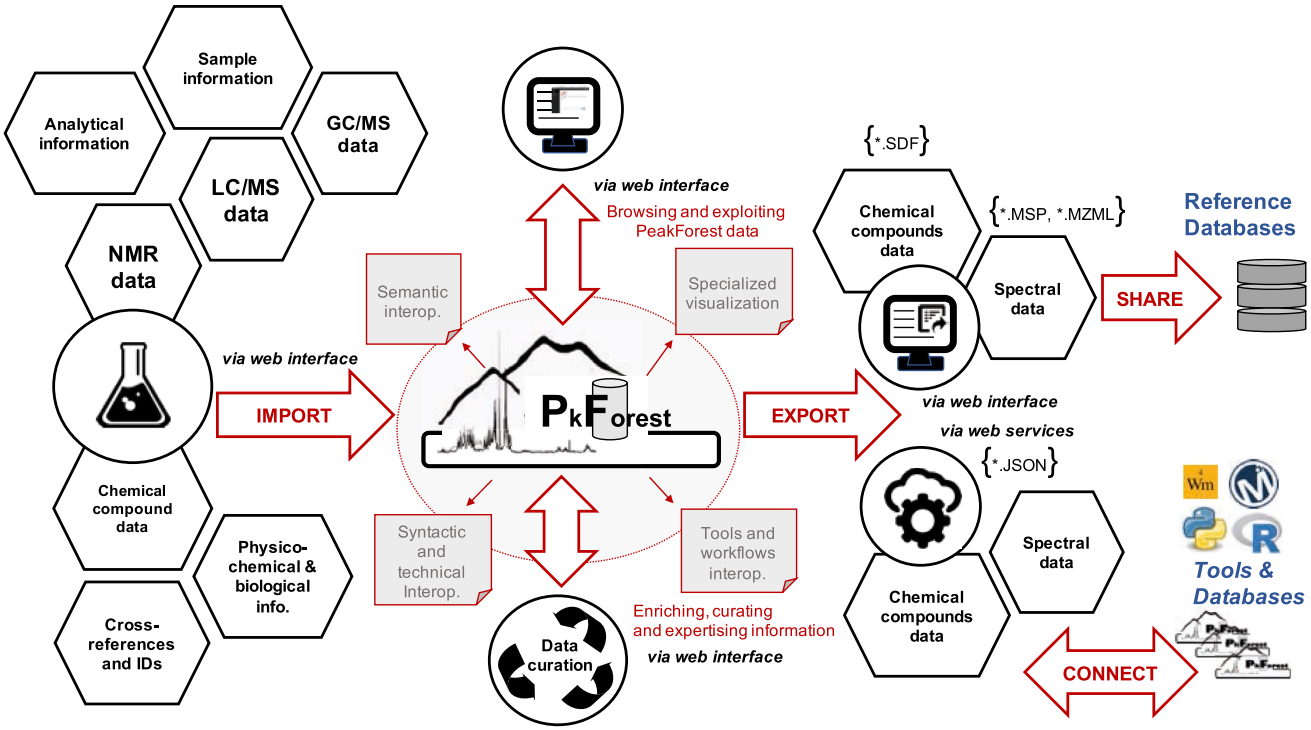


Fig. 1 PeakForest database inputs and outputs

bibliographic references). PeakForest manages metabolomic data during the metabolite identification process, from the acquisition of chemical standard spectra, the annotation of spectra in biological matrices via external peak matching tools, to the spectra linkage with external resources for biological interpretation or publication (PeakForest has been specifically designed to be easily interfaced with external tools). Users can import compounds and spectral data manually extracted from raw data, curate information and exploit this knowledge via data export or PeakForest Web services connected tools (Fig. 1). PeakForest stores metadata related to raw data rather than raw data itself to minimize storage space usage.

2.1 PeakForest framework technical specifications

PeakForest is a Java Web application (version 8) and has been designed as an API's toolbox (Online Resource 1). The architecture leverages a complete data model, integrating entities as chemical compounds and spectra.

PeakForest proposes representational state transfer (*REST*) protocol-based Web services allowing programmatic data access to external resources and tools. Based on the OpenApi (v3.0) standard, any developer can generate a PeakForest Web services client in common programming languages (Python, R…). Centralizing the REST specifications in a unique OpenAPI compliant file allows consistent documentation, up-to-date code and exchanges between servers and clients.

All PeakForest components use the same Java APIs. The Web application and the REST documentation are hosted on a Tomcat server (version 7) allowing the application to be run using a basic Java virtual machine (version 8). Full details and technical aspects are described in the official install documentation[3] and short tutorials.[4] In order to facilitate the deployment of local PeakForest databases within laboratories, users can be attributed different privileges and permissions in the PeakForest system (Online Resource 1). PeakForest is available on DockerHub,[5] making possible the run of several specialized databases on a unique server.

2.2 Source and code project

PeakForest is a free and open-source project under the CECILL-2.1[6] license, published on GitHub.[7] Issues and incidents can be reported on this PeakForest official public

[3] https://peakforest.org/local_install.
[4] https://peakforest.org/howto.
[5] https://hub.docker.com/u/metabohub.
[6] https://cecill.info/licences/Licence_CeCILL_V2.1-en.html.
[7] https://github.com/peakforest.

repository. The PeakForest database model and project code have been deposited on APP,[8] a European organization for the protection of authors and publishers of digital creations (ID: *IDDN.FR.001.180009.000.S.C.2021.000.31230*).

3 Results

PeakForest manages metabolomics data including chemical compound descriptors and different types of analytical spectra. Main Web interface functionalities are detailed below:

3.1 Building a PeakForest database

3.1.1 Compounds data and metadata inputs

Adapted for compounds already present in public databases, the import module allows the addition of individual chemical compounds, attributes a specific internal identifier, and creates the associated compound card from an InChIKey or a common name (Fig. 2). An "addition assistant" checks if the compound is already present in the database, and proposes compound candidates based on the Fiehn 'Chemical Translation Service' (Wohlgemuth et al., 2010) and the PubChem PUG-REST service (Kim et al., 2018) where necessary to complete the missing lnChIKey data. Using the same Web services, this module fills missing information such as synonyms, structural representation of compounds and provides external cross-references where available. The module computes accurate mass, formulae and SMILES (simplified molecular-input line-entry system) as well as running molecule structure image depiction with *OpenBabel*.[9] A batch system is also available to import a large compound list (Online Resource 2) and create associated compound cards based on minimal compound information found in the imported file (common name, InChI and InChIKey are mandatory for this function to operate). Use of InChIKey identifiers avoids difficulties associated with ambiguous compound names.

Each compound card is organized in different sections. The first section presents general compound information (common name, synonyms, molecular formula, accurate and average masses). The second section is related to structural data with common numerical molecule representations in MDL Molfile (Dalby et al., 1992), Canonical SMILES (O'Boyle, 2012), InChI and InChIKey formats (Goodman et al., 2021; Southan, 2013) and *2D* and *3D* molecule images. A "cross-reference identifiers" section includes four selected external chemical compounds bank references, with ChEBI (Hastings et al., 2016), PubChem (Kim et al., 2021), KEGG (Kanehisa et al., 2016), and HMDB (Wishart et al., 2018) and a modular system to add any Web hyperlinks from specific biological knowledge banks or metabolic networks databases. The compound card also gives direct access to the full list of spectral cards available on the database instance, grouped by analytical techniques.

3.1.2 Spectral data and metadata inputs

The spectral module allows the addition of new spectra to the local PeakForest database. PeakForest supports a large range of spectral types, in line with common metabolomics analytical technologies: NMR-1D (^{1}H, ^{13}C), NMR-2D (JRES, COSY, TOCSY, NOESY, HSQC, HMBC), LC–MS(/MS^{n}), FIA-MS(/MS^{n}) and GC–MS. During the spectral import process, all spectra are associated with a chemical compound, and qualified as either a chemical standard, part of a chemical standard-mix, present in a reference biological matrix (e.g. NIST plasma) or in a biological (or environmental) matrix. Spectral data models have been designed based on expert advice using standardized metadata describing NMR and MS acquisition methods. Heterogeneity of analytical instruments is considered, and unique descriptors[10] are proposed to enable metadata sharing within the metabolomics community. PeakForest uses IUPAC nomenclature and MassBank consortium proposal[11] for mass spectrometry (Murray et al., 2013) and suggests the use of standardized and chemically-consistent ion annotation procedure following the recommendations of Damont et al., 2019 for user-defined peak attributions (Damont et al., 2019). The database structure also supports information on sample preparation and species origins.

An interactive and adaptive Web form is provided for spectrum addition, with four major steps for a LC–MS spectral import example: spectrum type, sample type, liquid chromatography condition and MS analyser information. An excel-like file template can also be generated with prefilled and predefined data and metadata based on user analytical methods (example templates in Online Resources 3–6). A batch system exists for large-scale imports of new spectra. All imported spectra generate spectrum cards with a short description of sample preparation (centrifugation, purification, dilution or derivatization conditions). Layout of spectrum cards depends on spectral type (Table 1, Fig. 3). Each spectrum is provided with a specific internal identifier and a splash identifier is also computed for LC–MS data (Wohlgemuth et al., 2016). The summary of compounds and spectral data origins is available in Online Resource 7.

[8] https://www.app.asso.fr/en.

[9] http://openbabel.org/.

[10] https://peakforest.org/descriptors.

[11] http://www.mssj.jp/english/about/pdf/MassBank_manual-en.pdf.

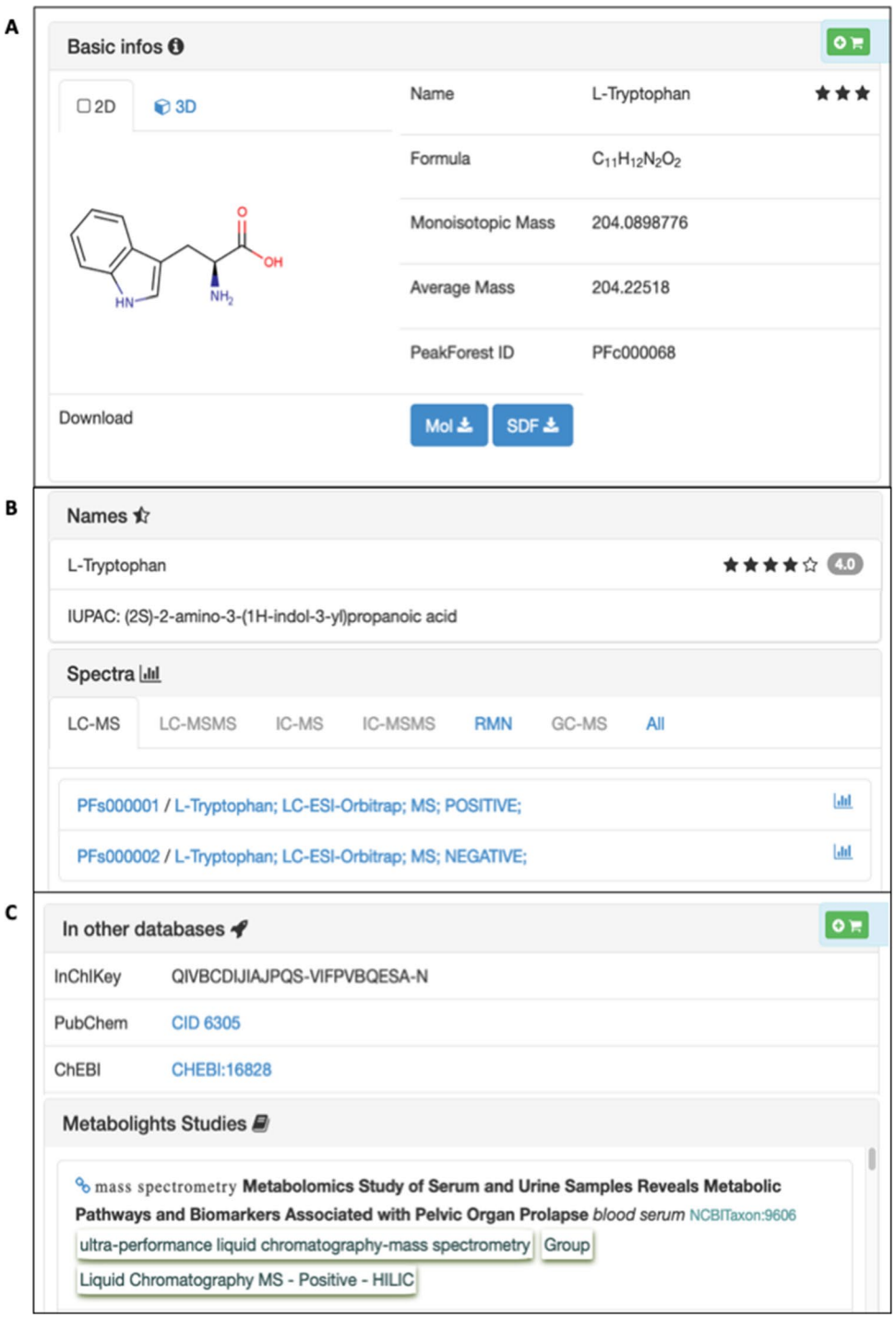


Fig. 2 Example of a chemical compound card (L-tryptophan) with compound basic information (**A**), related names and spectra parts (**B**) and externals identifiers and related Metabolights studies (**C**)

3.2 Data quality and curation

3.2.1 Standardized operating procedures to ensure data quality

Different standard operating procedures (SOPs) provide comprehensive information about the data that can be stored in PeakForest, allowing users to achieve high-quality data levels in local PeakForest databases. All SOPs and publication references are publicly and freely-available on the PeakForest Web portal peakforest.org. To illustrate Web interfaces and MetaboHUB SOPs application, a PeakForest demonstrator is available at https://demo.peakforest.org. This instance contains a dataset of 96 chemical compounds selected for their biological interest and their relatively-easy identification in biological matrices or biofluids. The

Table 1 Spectral data and metadata group view with links of corresponding templates (available in the supplementary material section) and potential usage for users

Metadata group label	Context and template part link	Metadata usage
Sample metadata	All spectra templates "sample" sheet	Informs users about the spectrum's type (single chemical compound, mix of chemical compounds, NIST plasma or biological matrix) – Extra information is available for NMR spectra like the optional isotopic labelling – Information about sample preparation will be available in a future planned release
Liquid chromatography	Only fullscan and fragmentation LC spectra (LC–MS / LC–MS/MS) "chromatography" sheet	Informs users about chromatography data (column brand, type, name, length, diameter, flow rate, injection volume, gradient…) – The column characteristics can be used as filter in Web service
Gas chromatography	Only full scan GC spectra (GC–MS) "chromatography" sheet	Informs users about chromatography data (column brand, type, name, length, diameter, …) – The column characteristics can be used as filter in Web service
Ion chromatography	Only fullscan IC spectra (IC-MS) "chromatography" sheet	Informs users about chromatography data (column brand, type, name, length, diameter, …)
Ionization method	All mass spectra "MS_analyzer" sheet "GCMS_analyzer" sheet for LC and IC spectra	Informs users about instrument characteristics and settings for the acquisition
Ion analyzer	All mass spectra "MS_analyzer" sheet for LC and IC spectra "GCMS_analyzer" sheet for LC and IC spectra	Informs users about instrument characteristics and settings for the acquisition
NMR instrument *NMR processing software*	For all NMR spectra "NMR_analyzer" sheet	Informs users about instrument and software characteristics and settings for the acquisition and processing parameters
"Other" metadata	For all spectra "Other" sheet	Informs users about the spectrum's authors, ownership, raw file, …

demonstrator also contains a spectra collection acquired on different analytical platforms for a range of analytical conditions by the MetaboHUB consortium, including flow injection analysis (FIA), reversed-phase (C18) chromatography and hydrophilic interaction chromatography (HILIC) coupled to an Exactive or an Orbitrap Fusion (Thermo Fisher Scientific) or an Impact HDII Quadrupole-Time Of Flight (QToF) (Bruker, Daltonics) mass spectrometer; gas chromatography coupled to Accurate Mass QToF 7200 mass spectrometer (Agilent Technologies, Inc); Avance III NMR spectrometers (Bruker, Biospin) with different environments (magnetic field from 500 to 800 MHz for proton frequency and probes at room temperature to cryoprobes).

3.2.2 In silico data checking and automatic/manual curation

The PeakForest data model specifies mandatory data and metadata to be included during import phases, with inbuilt routines to limit or identify mistakes and data inconsistency. For example, PeakForest manages chemical compound unicity, based on the InChI/InChIKey set. If a user imports information about a compound already present in the database, new common names are added to existing ones as synonyms but properties such as the accurate mass are not recomputed. If different external identifiers (e.g. ChEBI ID) are provided, the system does not update them but creates a "curation message" associated with the compound. This curation message text will indicate the nature of the conflict and "curator" users will be able to manually update the entry if a correction is required. Automatic data enrichment within in-silico metadata is also possible. For example, imported compound information can be enriched with in-silico generated physico-chemical properties such as LogP, computed by the OChem Web services (Sushko et al., 2011), and the endogenous mammalian status of molecules are determined using the integrated BioSM tool (Hamdalla et al., 2013).

PeakForest is designed to allow manual data curation with a manual scoring system to rank compound names and a star grading system to distinguish different statuses in compound curation (initial compound import to final validation). Curation is not only feasible on compound descriptors but also on spectral data where, for instance, peak assignments may be added, modified or refined by

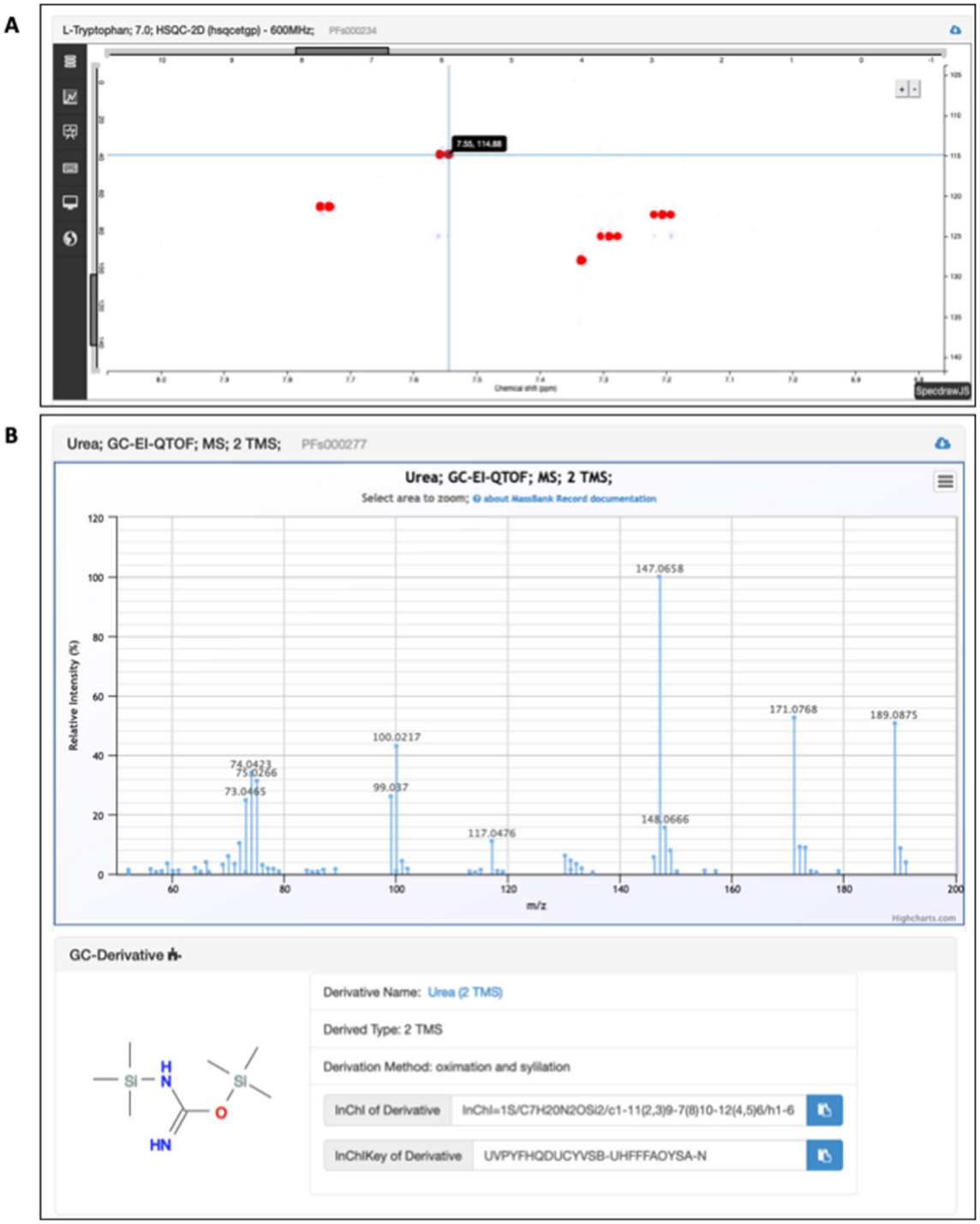


Fig. 3 Spectral card examples related to L-Tryptophan; pH 7.0; HSQC-2D (hsqcetgp)—600 MHz card (**A**) and Urea; GC-EI-QTOF; MS; 2 TMS card (**B**)

analytical chemists. Authenticated users on a local PeakForest database can submit messages and raise issues on compound or spectral cards. All these messages are grouped in the PeakForest curation message centre; users with a curator role can manage all reported issues and access the edition mode of all compounds, spectral data and metadata. Users can also associate any scientific publications with a particular chemical compound card through the simple and original paper's digital object identifier (DOI) or its PubMed identifier.

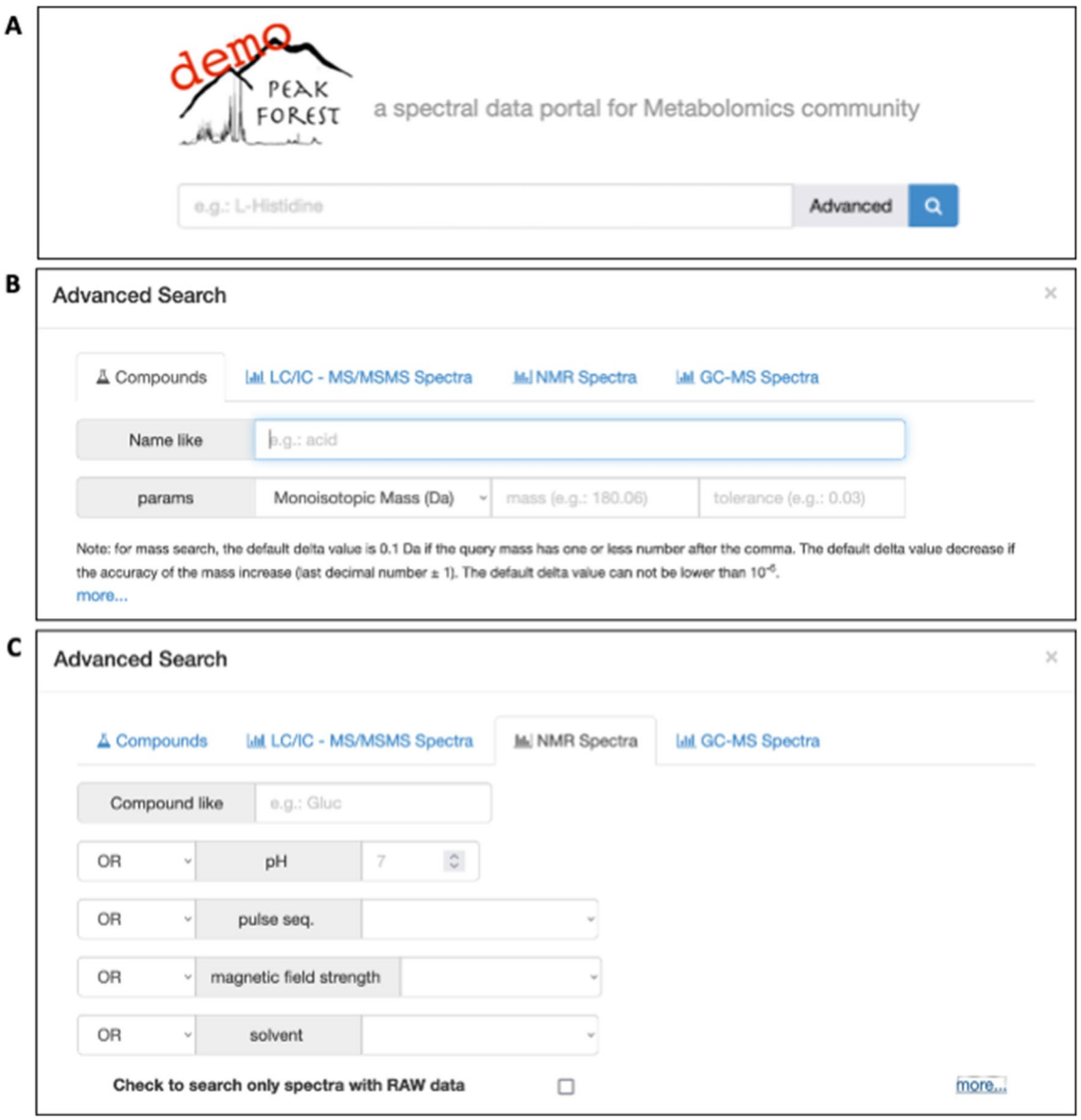


Fig. 4 Screenshot of the PeakForest search modules with the "quick search" tool (**A**), and the advanced search tool for compounds (**B**) and for NMR data (**C**)

3.3 Browsing and using PeakForest data

3.3.1 PeakForest search interfaces

PeakForest offers specific modules to browse and mine compound and spectral data. The "quick search" Web module uses properties or identifiers (Fig. 4), whereas the "advanced search" module is able to interpret natural language and support advanced keyword queries. A filter system, integrating basic logic gates ('AND', 'OR') allows users to query the database, and a range of queries are possible depending on spectral type.

3.3.2 Exporting data from PeakForest

Since sharing and opening data are as important as building and organizing individual laboratory data and metadata, PeakForest proposes two ways to export local collections into current metabolomics standards. This functionality guarantees the interoperability of PeakForest data collections with common bioinformatics tools, as well as the export into international metabolomics databases (MassBank, MONA, HMDB, GNPS, …) and repositories (MetaboLights, Metabolomics Workbench, …). All compounds and their properties can be exported as a "comma separated value" formatted file, and individual. compound cards can also be

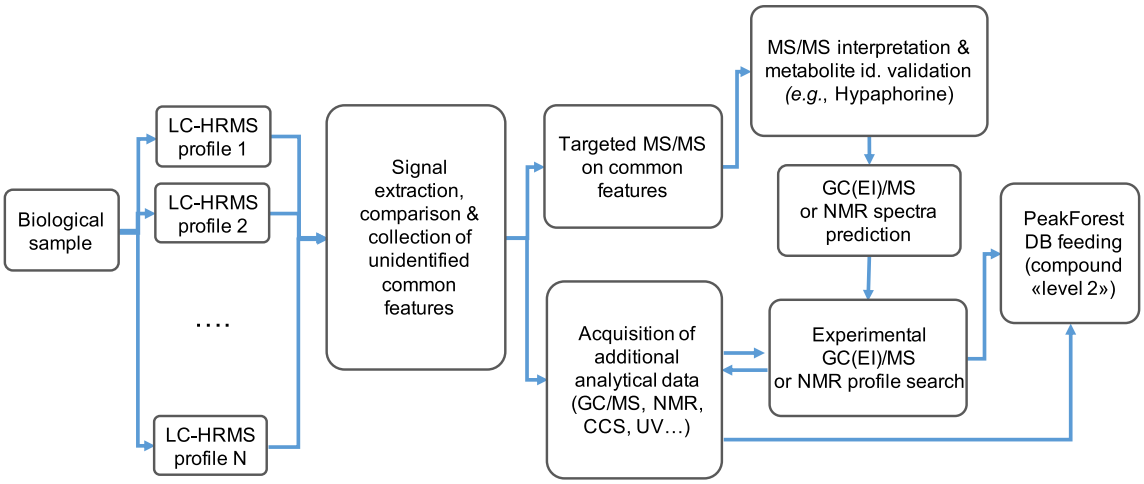


Fig. 5 Example of an inter-platform workflow strategy used to generate LC-HRMS/MS, NMR and other orthogonal data related to unknown metabolites in biological matrices, and subsequent PeakForest database enrichment with biological compounds identified with a confidence "MSI level 2"

exported manually as a SDF (Dalby et al., 1992) formatted file. Spectra are exportable in the MassBank text format and/or specific open formats adapted to different measurement techniques (e.g. nmrML (Schober et al., 2018) for 1D-NMR spectra and MSP for mass spectra) easily linked to common softwares and tools including MS-DIAL (Lai et al., 2018) and GCMS NIST MSsearch.[12]

3.3.3 Programmatic interface

The RESTful Web services associated with each local PeakForest infrastructure provide direct access to the complete local data collection. A set of methods are designed to give simple access to compound or spectral cards by exact or partial matching and return information in JSON formatted files. A range of options is also available to facilitate queries including searching for a specific mass-to-charge ratio (m/z) or chemical shift. This Web services layer allows bioinformatics programmers to easily and quickly integrate an existing PeakForest instance as a dynamic data resource with secure access enabled by an authentication token system. PeakForest-compatible tools are currently available as part of the Galaxy project Toolshed[13] and are hosted on the Workflow4Metabolomics platform.[14] The complete REST API documentation is available for each local PeakForest database and a generic version is available on the peakforest.org portal. The PeakForest team is committed to develop this API based on queries and contributions submitted via the GitHub repository.

3.3.4 PeakForest for spectral data collection of metabolites absent from the chemical library

PeakForest offers the possibility to integrate and share physico-chemical characteristics of not fully elucidated compounds seen recurrently in biological matrices of interest. The identification of these not fully elucidated compounds is facilitated by the aggregation of convergent data from different techniques and analytical platforms (Fig. 5). As a test case, a common and non-assigned signal at m/z *247.1441* that is recurrently detected following LC-HRMS analysis of NIST plasma in positive mode of ionization on two different instruments, an Orbitrap Tribrid Fusion (Thermo Fisher Scientific) and a Q-ToF Impact II (Bruker, Daltonics), led us to further investigations. Additional HRMS/MS acquisitions performed on both instruments allowed to conclude to the presence of a metabolite annotated as hypaphorine in this sample and to list its characteristics (RT, HRMS, MS/MS, etc.) in PeakForest retrieved from each analytical system. With the referencing of these analytical data in PeakForest, hypaphorine (or lenticin) proved to be also detected in human urine (Garcia-Aloy et al., 2020) and could be annotated with a confidence "level 2" with spectra stored in the MoNA public database.[15] Additionally, signals obtained with NMR spectra prediction of hypaphorine could be searched for in the experimental NMR profiles of the same

[12] https://www.nist.gov/system/files/documents/srd/nistms.pdf.

[13] https://toolshed.g2.bx.psu.edu/.

[14] https://workflow4metabolomics.usegalaxy.fr/.

[15] https://mona.fiehnlab.ucdavis.edu/spectra/display/VF-NPL-QEHF001173.

biological sample (NIST plasma) in order to confirm its identity (confidence level 2) and enrich the NMR set of data of biological compound in PeakForest. Eventually, when available, the authentic standard may be purchased, analysed and reported in PeakForest, to reach "level 1" annotation.

4 Discussion

Improving confidence in the outputs of data annotation and metabolite identification in biological samples is a priority for the metabolomic field (Wishart et al., 2022). Here we present the first comprehensive web-based infrastructure which can simultaneously handle MS and NMR data, and is designed to organize spectral data for metabolome annotation. PeakForest is able to capture and store experimental spectral data (peak lists) as well as associated annotations, and offers the possibility to curate information over time, to be browsed and reused in data treatment and, ultimately, shared across laboratories. PeakForest has the added value of providing a structured system with onsite data centralization and security for small laboratories with limited IT support.

Over the last 15 years, a number of database structures and tools have been proposed to generate local data repositories for either MS or NMR spectral collections, promoting accuracy in metabolite identification (Ferry-Dumazet et al., 2011; Horai et al., 2010; Palmer et al., 2018). PeakForest builds on the principles of these tools, and incorporates the latest developments in web frameworks and applications to provide a unique data model compatible with multiple research questions, depending on laboratory focus and/or instruments. In addition to storing metabolites/spectral metadata, our data model can describe information on sample preparation, biological species identifiers and biological matrices, providing a complete compound profile for future use. For example, a dedicated PeakForest database can be implemented for a particular disease or class of biological species, based on concepts of "sample type-specific databases". Indeed it has been shown that the use of a restricted search domain increases the precision of annotations (Reisdorph et al., 2019). A key innovation in PeakForest is the capacity to support both MS and NMR spectral data for a multitude of low- and high-resolution configurations and dimensions (i.e., 1D and 2D NMR, MS and MS^2 data), chromatographic retention times, and seeking to integrate additional structural information or others molecular descriptors, such as ultraviolet–visible spectra and ion mobility mass spectrometric data (collision cross sections, CCS) in the near future. Furthermore, recent studies have called for combined mass and NMR analyses of biological samples in order to improve metabolome annotation and enhance metabolome coverage (Marshall & Powers, 2017). This is particularly relevant for the identification of carbohydrates and their derivatives, which are poorly-detected in LC-HRMS but fully-characterized by GC-HRMS and quantified with NMR (Comte et al., 2021). By centralizing a large diversity of metabolites and profiles, PeakForest proposes a well-adapted database for studies based on a multi-platform untargeted strategy.

Studies in metabolomics are increasingly using high-throughput screening technologies and large sample batches; these approaches generate "big data" and bring new challenges regarding data management and security. The key to efficient data management involves setting up appropriate data stewardship during production, treatment, mining and knowledge dissemination. In this context, global guidelines are now available to help specific metabolomics communities to enhance their data usage and new knowledge creation (Savoi et al., 2021). However, applying these recommendations can be complex and often requires changes in traditional work practices, new skills and/or training (Griffin et al., 2018). PeakForest complies with "best practice" data management guidelines via SOPs which comprehensively address the data production workflow, from sample collection and preparation, chemotype analysis and data treatment to metabolite annotation. It is designed as a complementary component of effective information systems, and can be used in conjunction with laboratory information management systems (LIMS) which provide high traceability for sample data and metadata (Hunter et al., 2017). PeakForest is also compatible with modern data analysis workflows-based platforms which generate metadata reproducibility for data analyses (Giacomoni et al., 2015; Guitton et al., 2017; Huan et al., 2017; Pang et al., 2021; Tautenhahn et al., 2012; Xia & Wishart, 2011).

In line with the open-science movement, the metabolomics community has made significant efforts to align their standards with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles through the creation and the use of open data formats and online resources (Mendez et al., 2019). PeakForest adopts FAIR practices (Johnson & Lange, 2015; Wilkinson et al., 2016) in order to ensure that data at the laboratory level is consistent with, and can be used by, the international metabolomics community. PeakForest uses standardized vocabulary, including accurate and relevant attributes. Metabolomics annotation recommendations have been implemented, such as assigning cross reference ID on metabolites and making sharable reference spectra (Kind et al., 2018; Redestig et al., 2010), and use of unambiguous and persistent identifiers for metabolites and LC–MS data (Wohlgemuth et al., 2016) and will facilitate the future integration of such data in data contextualization tools (Cottret et al., 2018; Delmas et al., 2021). Interoperability with data analysis and data mining tools is ensured by open communication protocols (REST), and data export functions with conversion

into common open standards formats ensure interoperability with academic databases and open repositories in the metabolomics community. Overall, the PeakForest framework is designed to encourage greater openness of laboratory databases, facilitating integration into public databases and knowledge sharing between laboratories, and ultimately promoting effective metabolomics research (fewer redundancies across laboratories, faster annotation of unknowns).

PeakForest has been extensively beta-tested by the members of the French metabolomics community. It is fully-operational (available at https://peakforest.org/), and already currently used in research laboratories in France and abroad. In order to minimize server storage of unnecessary data, PeakForest is designed for metabolite annotation report data rather than storage of raw spectral data. Technical installation of PeakForest in any laboratory information system requires personnel with computing or bioinformatics skills, but once installed, the interface is user-friendly and the infrastructure is adapted for use by MS and NMR scientists with no specific computer skills. Of course, a minimum number of compounds and spectra need to be imported by users in order to make local PeakForest databases fully-functional. In addition, PeakForest does not include data clean-up or extraction tools per se and requires metadata to be manually entered in the database; users supply completely-cleaned data following SOPs edited by the PeakForest team, and PeakForest tutorials are available to assist in this step. Data and metadata within PeakForest can be changed (curated) with ease; traceability of requested modifications is ensured via curation modules. This feature is ideally-suited to local data facilities under the supervision of a data manager, instrumental experts or associated scientists, but is less easy to deploy in larger-scale public repositories as it requires centralised and standardized highly-reactive error detection and correction processes.

As with all databases, initial construction of the relevant database and spectral collection in line with laboratory needs will take some time. Deployment of PeakForest is an iterative process, and initial investment in the database will yield increasing returns in terms of accuracy and efficiency as the spectral collections are incremented. At present a number of basic peak-matching tools are available as add-ons via the Galaxy toolshed, but additional developments are needed to enhance the MS–MS and NMR-2D matching capabilities, such as interoperability with MS-Dial (Tsugawa et al., 2015). For laboratories with access to bioinformatics support, integration of bespoke tools to the PeakForest infrastructure can contribute the optimization of local data pipelines analysis and metabolite annotation. This adaptability of PeakForest system offers multiple perspectives, and the opportunity to meet the evolving needs of the metabolomics research community.

5 Conclusion

This paper describes a novel digital infrastructure for the development of "new generation", structured and interoperable databases. PeakForest is intended to overcome a technical bottleneck, namely large-scale collaborative spectral data annotation and metabolite identification for metabolomics laboratories with multiple instruments. The PeakForest database is a user-friendly data management solution, which can be used to build metabolite and spectral collections, and has in-built modules which allow users to curate and mine annotation mass and NMR data via Web interfaces and external tools. By generating an ecosystem of interoperable databases, PeakForest represents a significant advance for promoting open science in the field of metabolomics, maintaining and promoting good practice and scientific rigour, as well as increasing the shareability and findability of data. The PeakForest digital infrastructure is associated with a public portal proposing technical guides, tutorials and SOPs. Integration of PeakForest in the Galaxy workflow environment facilitates the adoption and personalization of local databases by bioinformatics developers associated with metabolomics facilities, and the emergence of new tools for the scientific community based on a common standard. Finally, PeakForest also offers opportunities for future alignments with additional existing biological and chemical ontologies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11306-022-01899-3.

Acknowledgements The 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 "Investissements d'Avenir". The authors thank all MetaboHUB nodes and French metabolomics facilities for their investment in the development project and their collaboration in spectral collection, production and curation. We also thank Professor D. Rolin for his continual support of the project and its team, Dr. Y. Guitton for critical reading and Dr. J.M.G. Bloor for significant improvements to the manuscript.

Author contributions DR, FJ, CJ, EPG, FG and ET conceived and coordinated research. NP, FG, DJ, MTF, PR, AD and MLe designed database architecture and API, performed informatic development and tests. CD, NP, FG designed the PeakForest informatics environment. NP, FG, AM, CDe, CLP, DJ, SB, EC, LP, MT, AB, EJ, LD, ADa, UH, AW, FF, JCT, MFO, SA, EPR, BL, SDu, CJol, DC, CM, ML, CJou and CC conducted data and metadata standardisation, and designed spectral templates and SOPs. NL, SR, TB, AF, PP, RC, AB, UH, CJol, DC, RB, SE, JP, LR, VZ, MM, CLP and ML provided data and performed demo database curation. NP, ADa, CC, JCT, AM, FF, FJ, CJ, EPG and FG drafted the manuscript. All authors read, reviewed and approved the final manuscript.

Data availability This paper does not present data but a digital solution. A PeakForest demonstrator is available at https://demo.peakforest.org.

This instance contains an example dataset of 96 chemical compounds and 400 spectra.

Software availability Demo: https://demo.peakforest.org/webapp/; Project Web portal: https://peakforest.org/; Code repository (Major releases): https://github.com/peakforest; Docker files: https://hub.docker.com/u/metabohub.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Research involving human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology, 3*, 23. https://doi.org/10.3389/fbioe.2015.00023

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., Dauria, J., Ewald, J., Ewald, J. C., Fraser, P. D., Giavalisco, P., Hall, R. D., Heinemann, M., Link, H., Luo, J., Neumann, S., Nielsen, J., Perez de Souza, L., Saito, K., Sauer, U., Schroeder, F. C., & Fernie, A. R. (2021). Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nature Methods, 18*(7), 747–756. https://doi.org/10.1038/s41592-021-01197-1

Anwar, A. M., Ahmed, E. A., Soudy, M., Osama, A., Ezzeldin, S., Tanios, A., Mahgoub, S., & Magdeldin, S. (2021). Xconnector: Retrieving and visualizing metabolites and pathways information from various database resources. *Journal of Proteomics, 245*, 104302. https://doi.org/10.1016/j.jprot.2021.104302

Comte, B., Monnerie, S., Brandolini-Bunlon, M., Canlet, C., Castelli, F., Chu-Van, E., Colsch, B., Fenaille, F., Joly, C., Jourdan, F., Lenuzza, N., Lyan, B., Martin, J.-F., Migné, C., Morais, J. A., Pétéra, M., Poupin, N., Vinson, F., Thevenot, E., & Pujos-Guillot, E. (2021). Multiplatform metabolomics for an integrative exploration of metabolic syndrome in older men. *eBioMedicine, 69*, 103440. https://doi.org/10.1016/j.ebiom.2021.103440

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., Poupin, N., Vinson, F., & Jourdan, F. (2018). MetExplore: Collaborative edition and exploration of metabolic networks. *Nucleic Acids Research, 46*(1), 495–502. https://doi.org/10.1093/nar/gky301

Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L. W., Trengove, R., & Wolfender, J.-L. (2014). Metabolite identification: Are you sure? And how do your peers gauge your confidence? *Metabolomics, 10*(3), 350–353. https://doi.org/10.1007/s11306-014-0656-8

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A., & Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences, 32*(3), 244–255. https://doi.org/10.1021/ci00007a012

Damont, A., Olivier, M.-F., Warnet, A., Lyan, B., Pujos-Guillot, E., Jamin, E. L., Debrauwer, L., Bernillon, S., Junot, C., Tabet, J.-C., & Fenaille, F. (2019). Proposal for a chemically consistent way to annotate ions arising from the analysis of reference compounds under ESI conditions: A prerequisite to proper mass spectral database constitution in metabolomics. *Journal of Mass Spectrometry, 54*(6), 567–582. https://doi.org/10.1002/jms.4372

Delmas, M., Filangi, O., Paulhe, N., Vinson, F., Duperier, C., Garrier, W., Saunier, P.-E., Pitarch, Y., Jourdan, F., Giacomoni, F., & Frainay, C. (2021). FORUM: Building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics, 37*(21), 3896–3904. https://doi.org/10.1093/bioinformatics/btab627

Dona, A. C., Kyriakides, M., Scott, F., Shephard, E. A., Varshavi, D., Veselkov, K., & Everett, J. R. (2016). A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and Structural Biotechnology Journal, 14*, 135–153. https://doi.org/10.1016/j.csbj.2016.02.005

Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R. H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M. J. O., & Dennis, E. A. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of Lipid Research, 50*, S9–S14. https://doi.org/10.1194/jlr.R800095-JLR200

Ferry-Dumazet, H., Gil, L., Deborde, C., Moing, A., Bernillon, S., Rolin, D., Nikolski, M., de Daruvar, A., & Jacob, D. (2011). MeRy-B: A web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. *BMC Plant Biology, 11*(1), 104. https://doi.org/10.1186/1471-2229-11-104

Garcia-Aloy, M., Ulaszewska, M., Franceschi, P., Estruel-Amades, S., Weinert, C. H., Tor-Roca, A., Urpi-Sarda, M., Mattivi, F., & Andres-Lacueva, C. (2020). Discovery of intake biomarkers of lentils, chickpeas, and white beans by untargeted LC–MS metabolomics in serum and urine. *Molecular Nutrition & Food Research, 64*(13), 1901137. https://doi.org/10.1002/mnfr.201901137

Giacomoni, F., Le Corguille, G., Monsoor, M., Landi, M., Pericard, P., Petera, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., Goulitquer, S., Thevenot, E. A., & Caron, C. (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31*(9), 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Goodman, J. M., Pletnev, I., Thiessen, P., Bolton, E., & Heller, S. R. (2021). InChI version 1.06: Now more than 99.99% reliable. *Journal of Cheminformatics, 13*(1), 40. https://doi.org/10.1186/s13321-021-00517-z

Griffin, P. C., Khadake, J., LeMay, K. S., Lewis, S. E., Orchard, S., Pask, A., Pope, B., Roessner, U., Russell, K., Seemann, T., Treloar, A., Tyagi, S., Christiansen, J. H., Dayalan, S., Gladman, S., Hangartner, S. B., Hayden, H. L., Ho, W. W. H., Keeble-Gagnère, G., & Schneider, M. V. (2018). Best practice data life cycle approaches for the life sciences. *F1000Research, 6*, 1618. https://doi.org/10.12688/f1000research.12344.2

Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T.,

Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., & Siuzdak, G. (2018). METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry, 90*(5), 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424

Guitton, Y., Tremblay-Franco, M., Le Corguillé, G., Martin, J.-F., Pétéra, M., Roger-Mele, P., Delabrière, A., Goulitquer, S., Monsoor, M., Duperier, C., Canlet, C., Servien, R., Tardivel, P., Caron, C., Giacomoni, F., & Thévenot, E. A. (2017). Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *The International Journal of Biochemistry & Cell Biology, 93*, 89–101. https://doi.org/10.1016/j.biocel.2017.07.002

Hamdalla, M. A., Mandoiu, I. I., Hill, D. W., Rajasekaran, S., & Grant, D. F. (2013). BioSM: Metabolomics tool for identifying endogenous mammalian biochemical structures in chemical structure space. *Journal of Chemical Information and Modeling, 53*(3), 601–612. https://doi.org/10.1021/ci300512q

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research, 44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Haug, K., Salek, R. M., & Steinbeck, C. (2017). Global open data management in metabolomics. *Current Opinion in Chemical Biology, 36*, 58–63. https://doi.org/10.1016/j.cbpa.2016.12.024

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., & Nishioka, T. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*(7), 703–714. https://doi.org/10.1002/jms.1777

Huan, T., Forsberg, E. M., Rinehart, D., Johnson, C. H., Ivanisevic, J., Benton, H. P., Fang, M., Aisporna, A., Hilmers, B., Poole, F. L., Thorgersen, M. P., Adams, M. W. W., Krantz, G., Fields, M. W., Robbins, P. D., Niedernhofer, L. J., Ideker, T., Majumder, E. L., Wall, J. D., & Siuzdak, G. (2017). Systems biology guided by XCMS Online metabolomics. *Nature Methods, 14*(5), 461–462. https://doi.org/10.1038/nmeth.4260

Hunter, A., Dayalan, S., De Souza, D., Power, B., Lorrimar, R., Szabo, T., Nguyen, T., O'Callaghan, S., Hack, J., Pyke, J., Nahid, A., Barrero, R., Roessner, U., Likic, V., Tull, D., Bacic, A., McConville, M., & Bellgard, M. (2017). MASTR-MS: A web-based collaborative laboratory information management system (LIMS) for metabolomics. *Metabolomics, 13*(2), 14. https://doi.org/10.1007/s11306-016-1142-2

Johnson, S. R., & Lange, B. M. (2015). Open-access metabolomics databases for natural product research: Present capabilities and future potential. *Frontiers in Bioengineering and Biotechnology, 3*, 00022. https://doi.org/10.3389/fbioe.2015.00022

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research, 44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021). PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research, 49*(D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971

Kim, S., Thiessen, P. A., Cheng, T., Yu, B., & Bolton, E. E. (2018). An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Research, 46*(W1), W563–W570. https://doi.org/10.1093/nar/gky294

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. (2018). Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews, 37*(4), 513–532. https://doi.org/10.1002/mas.21535

Kuhn, S., & Schlörer, N. E. (2015). Facilitating quality control for spectra assignments of small organic molecules: Nmrshiftdb2—a free in-house NMR database with integrated LIMS for academic service laboratories: Lab administration, spectra assignment aid and local database. *Magnetic Resonance in Chemistry, 53*(8), 582–589. https://doi.org/10.1002/mrc.4263

Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., Ogiwara, A., Meissen, J., Showalter, M., Takeuchi, K., Kind, T., Beal, P., Arita, M., & Fiehn, O. (2018). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods, 15*(1), 53–56. https://doi.org/10.1038/nmeth.4512

Ludwig, C., Easton, J. M., Lodi, A., Tiziani, S., Manzoor, S. E., Southam, A. D., Byrne, J. J., Bishop, L. M., He, S., Arvanitis, T. N., Günther, U. L., & Viant, M. R. (2012). Birmingham Metabolite Library: A publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics, 8*(1), 8–18. https://doi.org/10.1007/s11306-011-0347-7

Malinowska, J. M., & Viant, M. R. (2019). Confidence in metabolite identification dictates the applicability of metabolomics to regulatory toxicology. *Current Opinion in Toxicology, 16*, 32–38. https://doi.org/10.1016/j.cotox.2019.03.006

Marshall, D. D., & Powers, R. (2017). Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Progress in Nuclear Magnetic Resonance Spectroscopy, 100*, 1–16. https://doi.org/10.1016/j.pnmrs.2017.01.001

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., & Deutsch, E. W. (2011). MzML—a community standard for mass spectrometry data. *Molecular & Cellular Proteomics, 10*(1), R110.000133. https://doi.org/10.1074/mcp.R110.000133

Mendez, K. M., Pritchard, L., Reinke, S. N., & Broadhurst, D. I. (2019). Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. *Metabolomics, 15*(10), 125. https://doi.org/10.1007/s11306-019-1588-0

Merlet, B., Paulhe, N., Vinson, F., Frainay, C., Chazalviel, M., Poupin, N., Gloaguen, Y., Giacomoni, F., & Jourdan, F. (2016). A computational solution to automatically map metabolite libraries in the context of genome scale metabolic networks. *Frontiers in Molecular Biosciences, 3*, e00002. https://doi.org/10.3389/fmolb.2016.00002

Misra, B. B. (2021). New software tools, databases, and resources in metabolomics: Updates from 2020. *Metabolomics, 17*(5), 49. https://doi.org/10.1007/s11306-021-01796-1

Murray, K. K., Boyd, R. K., Eberlin, M. N., Langley, G. J., Li, L., & Naito, Y. (2013). Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry, 85*(7), 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06

Nash, W. J., & Dunn, W. B. (2019). From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends in Analytical Chemistry*, *120*, 115324. https://doi.org/10.1016/j.trac.2018.11.022

O'Boyle, N. M. (2012). Towards a Universal SMILES representation—A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics, 4*(1), 22. https://doi.org/10.1186/1758-2946-4-22

Palmer, A., Phapale, P., Fay, D., & Alexandrov, T. (2018). Curatr: A web application for creating, curating and sharing a mass spectral

library. *Bioinformatics, 34*(8), 1436–1438. https://doi.org/10.1093/bioinformatics/btx786

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P. -É., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research, 49*(W1), W388–W396. https://doi.org/10.1093/nar/gkab382

Redestig, H., Kusano, M., Fukushima, A., Matsuda, F., Saito, K., & Arita, M. (2010). Consolidating metabolite identifiers to enable contextual and multi-platform metabolomics data analysis. *BMC Bioinformatics, 11*(1), 214. https://doi.org/10.1186/1471-2105-11-214

Reisdorph, N. A., Walmsley, S., & Reisdorph, R. (2019). A perspective and framework for developing sample type specific databases for LC/MS-based clinical metabolomics. *Metabolites, 10*(1), 8. https://doi.org/10.3390/metabo10010008

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., & Hide, W. (2012). Toward interoperable bioscience data. *Nature Genetics, 44*(2), 121–126. https://doi.org/10.1038/ng.1054

Savoi, S., Arapitsas, P., Duchêne, É., Nikolantonaki, M., Ontañón, I., Carlin, S., Schwander, F., Gougeon, R. D., Ferreira, A. C. S., Theodoridis, G., Töpfer, R., Vrhovsek, U., Adam-Blondon, A.-F., Pezzotti, M., & Mattivi, F. (2021). Grapevine and wine metabolomics-based guidelines for FAIR data and metadata management. *Metabolites, 11*(11), 757. https://doi.org/10.3390/metabo11110757

Schober, D., Jacob, D., Wilson, M., Cruz, J. A., Marcu, A., Grant, J. R., Moing, A., Deborde, C., de Figueiredo, L. F., Haug, K., Rocca-Serra, P., Easton, J., Ebbels, T. M. D., Hao, J., Ludwig, C., Günther, U. L., Rosato, A., Klein, M. S., Lewis, I. A., & Neumann, S. (2018). nmrML: A community supported open data standard for the description, storage, and exchange of NMR data. *Analytical Chemistry, 90*(1), 649–656. https://doi.org/10.1021/acs.analchem.7b02795

Southan, C. (2013). InChI in the wild: An assessment of InChIKey searching in Google. *Journal of Cheminformatics, 5*(1), 10. https://doi.org/10.1186/1758-2946-5-10

Spicer, R. A., Salek, R., & Steinbeck, C. (2017). A decade after the metabolomics standards initiative it's time for a revision. *Scientific Data, 4*(1), 170138. https://doi.org/10.1038/sdata.2017.138

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., & Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics, 3*(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., & Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design, 25*(6), 533–554. https://doi.org/10.1007/s10822-011-9440-2

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry, 84*(11), 5035–5039. https://doi.org/10.1021/ac300698c

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & Arita, M. (2015). MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods, 12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., & Markley, J. L. (2007). BioMagResBank. *Nucleic Acids Research, 36*, D402–D408. https://doi.org/10.1093/nar/gkm957

Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., & Yanes, O. (2016). Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry, 78*, 23–35. https://doi.org/10.1016/j.trac.2015.09.005

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology, 34*(8), 828–837. https://doi.org/10.1038/nbt.3597

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., & Scalbert, A. (2018). HMDB 40: The human metabolome database for 2018. *Nucleic Acids Research, 46*(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089

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., & Gautam, V. (2022). HMDB 5.0: The human metabolome database for 2022. *Nucleic Acids Research, 50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., & Forsythe, I. (2009). HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research, 37*, D603–D610. https://doi.org/10.1093/nar/gkn810

Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T., & Fiehn, O. (2010). The Chemical Translation Service—A web-based tool to improve standardization of metabolomic reports. *Bioinformatics, 26*(20), 2647–2648. https://doi.org/10.1093/bioinformatics/btq476

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., & Fiehn, O. (2016). SPLASH, a hashed identifier for mass spectra. *Nature Biotechnology, 34*(11), 1099–1101. https://doi.org/10.1038/nbt.3689

Xia, J., & Wishart, D. S. (2011). Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Current Protocols in Bioinformatics, 34*(1), 10–14. https://doi.org/10.1002/0471250953.bi1410s34

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Nils Paulhe[1] · Cécile Canlet[2] · Annelaure Damont[3] · Lindsay Peyriga[4] · Stéphanie Durand[1] · Catherine Deborde[5] · Sandra Alves[3] · Stephane Bernillon[5] · Thierry Berton[5] · Raphael Bir[1] · Alyssa Bouville[2] · Edern Cahoreau[4] · Delphine Centeno[1] · Robin Costantino[2] · Laurent Debrauwer[2] · Alexis Delabrière[3] · Christophe Duperier[1] · Sylvain Emery[1] · Amelie Flandin[5] · Ulli Hohenester[3] · Daniel Jacob[5] · Charlotte Joly[1] · Cyril Jousse[1] · Marie Lagree[1] · Nadia Lamari[5] · Marie Lefebvre[5] · Claire Lopez-Piffet[1] · Bernard Lyan[1] · Mickael Maucourt[5] · Carole Migne[1] · Marie-Francoise Olivier[3] · Estelle Rathahao-Paris[3] · Pierre Petriacq[5] · Julie Pinelli[5] · Léa Roch[5] · Pierrick Roger[3] · Simon Roques[5] · Jean-Claude Tabet[3] · Marie Tremblay-Franco[2] · Mounir Traïkia[1] · Anna Warnet[3] · Vanessa Zhendre[5] · Dominique Rolin[5] · Fabien Jourdan[2] · Etienne Thévenot[3] · Annick Moing[5] · Emilien Jamin[2] · François Fenaille[3] · Christophe Junot[3] · Estelle Pujos-Guillot[1] · Franck Giacomoni[1]

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

[2] Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, MetaboHUB, 31300 Toulouse, France

[3] Département Médicaments et Technologies pour la Santé (DMTS), Université Paris-Saclay, CEA, INRAE, MetaboHUB, 91191 Gif sur Yvette, France

[4] MetaboHUB-MetaToul, National Infrastructure of Metabolomics & Fluxomics (ANR-11-INBS-0010), 31077 Toulouse, France

[5] Université de Bordeaux, INRAE, Biologie du Fruit et Pathologie, UMR 1332, Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, 71 av E. Bourlaux, 33140 Villenave d'Ornon, France