Title: A guide to reverse metabolomics – a framework for big data discovery strategy.

Authors

Vincent Charron-Lamoureux[1,2], Helena Mannochio-Russo[1,2], Santosh Lamichhane[3], Shipei Xing[1,2], Abubaker Patan[1,2], Paulo Wender Portal Gomes[1,2], Prajit Rajkumar[1,2], Victoria Deleray[1,2], Andrés Mauricio Caraballo-Rodríguez[1,2], Kee Voon Chua[4], Lye Siang Lee[4], Zhao Liu[4], Jianhong Ching[4,5], Mingxun Wang[6], Pieter C. Dorrestein[1,2,*]

Affiliations

[1]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA, [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA, [3]Turku Bioscience Center, University of Turku and Åbo Akademi University, 20520 Turku, Finland, [4]Cardiovascular and Metabolic Disorders Programme, Duke-NUS Medical School, Singapore, [5]KK research Centre, KK Women's and Children's Hospital, Singapore, [6]Department of Computer Science, University of California Riverside, Riverside, CA, USA

*Author to whom correspondence should be addressed.

Abstract

Untargeted metabolomics is evolving into a field of big data science. There is a growing interest within the metabolomics community in mining MS/MS-based data from public repositories. The theme of this protocol, reverse metabolomics, is a data science strategy that differs from the traditional LC-MS/MS-based untargeted metabolomics approach. In traditional untargeted metabolomics, we first collect the samples to address a predefined question and then collect LC-MS/MS data. We then identify metabolites associated with a phenotype (e.g., disease vs. healthy), and elucidate or validate their structural details (e.g., molecular formula, structural classification, substructure, or complete structural annotation or identification). Reverse metabolomics, however, does not necessarily involve collecting new data or requiring the structural characterization of molecules. Instead, we start with MS/MS spectra for known or unknown molecules and discover phenotype-relevant information such as organ/biofluid distribution, disease condition, intervention status (e.g., pre- and post-intervention), organisms (e.g., mammals vs. others), geography, and any other biologically relevant associations available in public repositories. This protocol guides the reader through the step-by-step process of utilizing available MS/MS data and discovering repository-scale associations of the associated MS/MS spectra. As example, we utilize MS/MS spectra from three small molecules: phenylalanine-cholic acid (a microbially conjugated bile acid), phenylalanine-C4:0, and histidine-C4:0 (two *N*-acyl amides). We leverage the GNPS-based framework to explore

the microbial producers of these molecules and their associations with health conditions and organ distributions in humans and rodents.

Table 1: Glossary of terms

Term	Definition
Reverse metabolomics	A big data science strategy that takes a MS/MS first approach to search public data to uncover file-specific metadata driven organism, organ and biofluid, biological phenotypes, organism and other associations.
GNPS	Global Natural Products Social Networking is a community-driven infrastructure for mass spectrometry data analysis, storage, and for knowledge dissemination
MassIVE	Mass spectrometry Interactive Virtual Environment is a community resource for data deposition of mass spectrometry data
Metabolights	Metabolights is an open data repository with metadata for metabolomics studies.
Metabolomics Workbench	Metabolomics Workbench's National Metabolomics Data Repository is a public repository for metabolomics data storage and analysis
PRIDE	Proteomics Identification Database contains libraries for tools for computational proteomics
LC-MS/MS	Liquid Chromatography-tandem mass spectrometry is a hyphenated analytical technique that combines the separation of molecules based on their affinity with the mobile and stationary phases (LC) and their mass-to-charge ratio (MS)

GC-MS	Gas-Chromatography Mass Spectrometry is a hyphenated analytical technique that separates analytes in gas phase (GC) and their mass-to-charge ratio (MS)
MS/MS	Mass spectrometry/mass spectrometry. Also known as tandem mass spectrometry, MS2, and daughter ion
m/z	Mass-to-charge ratio
Cosine score	The cosine similarity measures the cosine of the angle between two vectors. In mass spectrometry, the cosine score is used to evaluate the similarity between two spectra. It ranges from 0 to 1, where 1 represents identical spectra and 0 denotes no similarity between the spectra
MGF	Mascot Generic Format files, a text formatted representation of MS and MS/MS information
mzML	Open-source text based XML-based format for mass spectrometry files
mzXML	A XML extensive Markup Language for mass spectrometry data
MASST	Mass Spectrometry Search Tool is a web-based search engine that uses a tandem MS spectrum to search against public metabolomics repositories
FASST	A faster implementation of MASST, FASST stands for Fasst mass Spectrometry Search Tool
USI	Universal Spectrum Identifier is a virtual path to the MS/MS spectrum information that is embedded and stored
ReDU	Reanalysis of Data User interface is a database that captures sample information (metadata) with controlled vocabularies and ontologies

Queried spectrum	Selected MS/MS spectrum by the users to use in the fast search tool
Reference spectrum	MS/MS spectrum found in public metabolomics repositories
foodMASST	An ontology informed search tool for known and unknown MS/MS spectra of food-derived molecules
plantMASST	A taxonomically informed search tool for known and unknown MS/MS spectra of plant-derived molecules
microbeMASST	A taxonomically informed search tool for known microbeMASST unknown MS/MS spectra of microbe-derived molecules
MassQL	Mass spectrometry Query Language is a universal language capturing mass spectrometry data patterns

The goal of this protocol

We aimed to provide a step-by-step procedure to the community for performing reverse metabolomics analysis that leverages public metabolomics data with information by matching a MS/MS spectrum of known or unknown molecules.

Introduction

Reverse metabolomics is a discovery-based data science strategy dedicated to the analysis of metabolomics data sourced from thousands of studies simultaneously [1,2]. In reverse metabolomics, researchers can analyze MS/MS spectra to uncover metadata associations, recovering phenotypic characteristics, identify the organisms that produce or catabolize the molecules of interest, determine their organ distributions and other characteristics in a biological system (**Fig. 1**). This process is achieved by linking the obtained data with the metadata associated with publicly available datasets, bringing liquid chromatography tandem mass spectrometry-based (LC-MS/MS) untargeted metabolomics data analysis truly into the realm of big data. Reverse metabolomics is possible because in the last decade more and more untargeted MS/MS metabolomics data has been deposited in the public domain, mainly in repositories such as Metabolights[3], Metabolomics Workbench's National Metabolomics Data Repository (NMDR)[4] and The Global Natural Products Social Molecular Networking[5]/Mass Spectrometry Interactive Virtual Environment (GNPS/MassIVE), but other sources of public data also exist[6,7]. As these repositories are continuously expanding, currently with

approximately 2 million LC-MS/MS runs and roughly 2 billion mass spectrometry tandem spectra, they provide an unprecedented and underutilized opportunity for biological discoveries. Ongoing efforts to standardize publicly deposited data formats, including metadata vocabularies, are underway. These endeavors, coupled with the development of advanced search and data filtering engines, are opening new avenues to identify and prioritize crucial metabolites or metabolite classes. In this protocol we outline a strategy on how one can begin to utilize these public resources.

Although the principle of reverse metabolomics, using the repository search tool Mass Spectrometry Search Tool (MASST) was first employed in 2020 to find files that contained specific MS/MS spectra to uncover disease associations[8], this required manual inspection and interpretation. There are now two specific studies that have used reverse metabolomics to discover new biology and biochemistry from repository information. In one study, the integration of reverse metabolomics with organic synthesis (a way to obtain MS/MS spectra for searching) led to the discovery of 800 molecules found in data derived from human samples, including microbial metabolites associated with fecal samples from people living with Crohn's disease[1]. Another study combined reverse metabolomics with a Mass spectrometry Query Language called MassQL (MassQL is another strategy by which one can obtain MS/MS spectra to search with, details in **Box 1** which provided evidence that there are thousands of modifications that bile acids can undergo, and that many of these are introduced by the microbiota and altered based on the diet[2]. In addition, it was possible to observe that these microbially-derived bile acids are distributed throughout the body, providing support for the potential existence of a microbiome encoder and decoder communication highway[2,9].

Given that the reverse metabolomics strategy relies on a suite of recently introduced tools and resources, its implementation is primarily confined to the scientists who developed this ecosystem. As it is often perceived as a complex task for individuals outside the immediate circle of developers, we aim to provide detailed step-by-step instructions not only to demystify the process of reverse metabolomics but also to empower other scientists with the capability to apply this task to their own work. The goal is to facilitate discovery and formulation of hypotheses generated by metadata associations obtained through reverse metabolomics. Additionally, we want to provide a foundation for others to learn from this approach, to help them to think about how to build their own capabilities that leverage metabolomics data from other data repositories and perhaps improve upon them in the future.

We will carry the reader through the four parts of reverse metabolomics (**Fig. 1**). The first part is obtaining the MS/MS spectra that are to be queried (**Fig. 1 – Part 1**), the second part uses Mass Spectrometry Search Tool (MASST)[10] searches to find the files associated with the MS/MS that are in available databases (**Fig. 1 – Part 2**). A MASST search may also include domain-specific MASSTs searches such as foodMASST[11], microbeMASST[12], plantMASST[13], and other domain-specific MASSTs that have curated

ontologies. These domain-specific MASST searches can be leveraged to understand the link of the MS/MS data to food, microbes and plants, respectively. The third part of reverse metabolomics is to link the files found with MASST to their metadata (**Fig. 1 – Part 3**). This is accomplished utilizing the ReDU framework, i.e., Reanalysis Data User interface[14]. ReDU is designed to harmonize vocabularies for metadata. This facilitates the data science based summaries of the results allowing the formulation of hypotheses. Finally, it is important to validate the observations obtained through reverse metabolomics (**Fig. 1 – Part 4**). While reverse metabolomics can provide new biological hypotheses, the investigator must think about how to further validate the observations that have been made. This can be performed through synthesis of standards when new molecules are proposed to match the MS/MS and retention times or validating observations with additional orthogonal cohorts and/or experiments that can distinguish isomers.

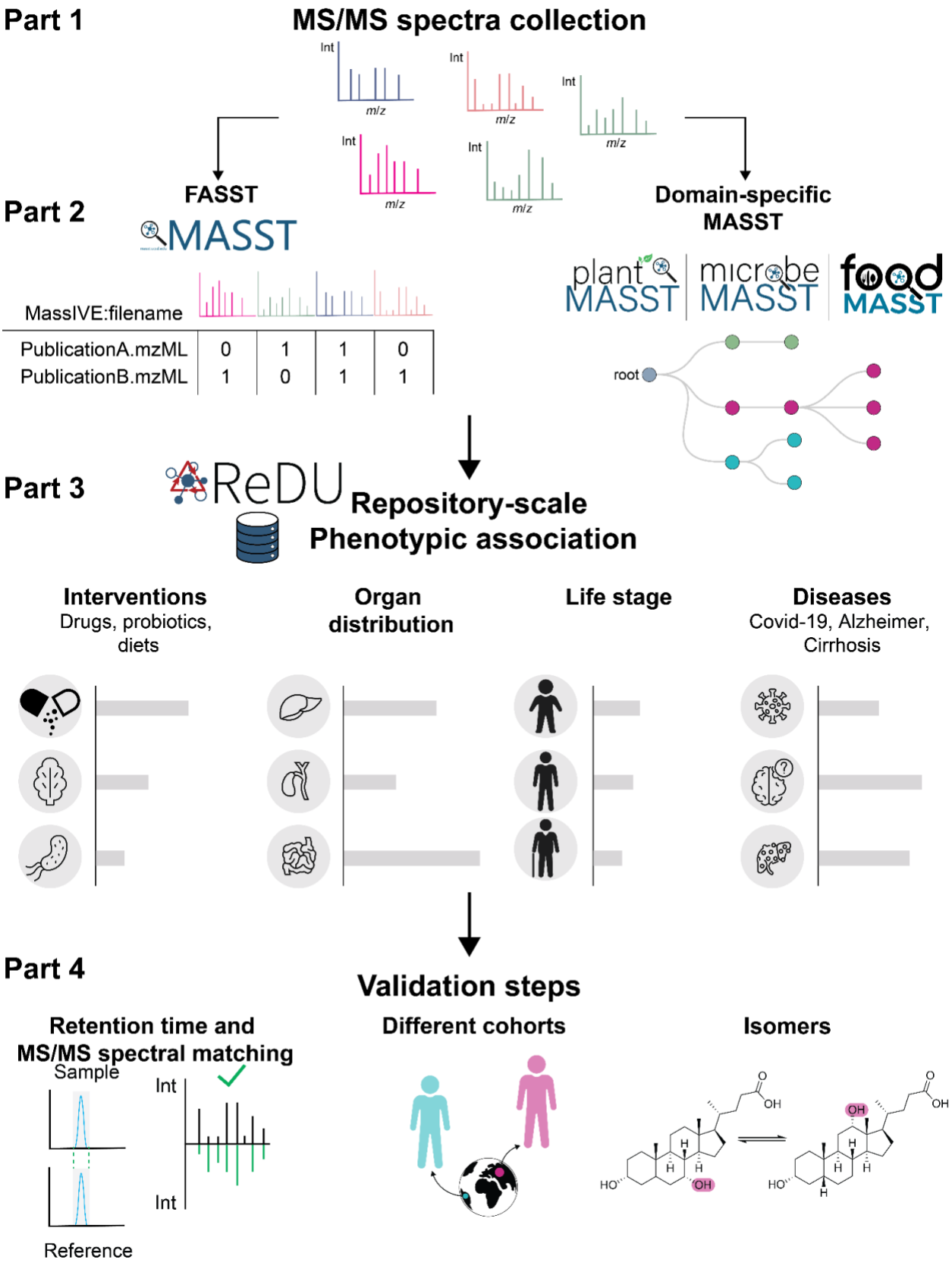


Fig. 1 | An overview of the reverse metabolomics workflow. The initial step involves accessing MS/MS spectra (**Part 1**). A fast MASST (FASST) search of tandem mass spectra is performed to collect identical or similar structures in the GNPS/MassIVE repository (**Part 2**). Domain-specific MASSTs can be used to assess if molecules of interest are microbial-, food-, or plant-derived. Metadata information is linked to each file by incorporating ReDU metadata and summary statistics can be performed (**Part 3**). Examples of validation steps to confirm the phenotypic association of the queried molecules (**Part 4**).

Background needed to understand reverse metabolomics

Due to improving data repository infrastructures, peer review pressure during review for publications, and funding body mandates, there is a noticeable rise in the deposition of untargeted metabolomics data in dedicated repositories, doubling the rate of growth every 2-3 years[3–5]. This exponential growing trend anticipates the creation of robust discovery resources, with tens of millions of files in the foreseeable future, and strategies will be needed to leverage such resources and to benefit society. We anticipate that as the scientific community recognizes the potential value of making discoveries with publicly available data, it motivates additional metabolomics researchers to deposit their data in dedicated repositories.

A key obstacle in leveraging public metabolomics data is the diversity of data formats (with more than 30 vendor-specific formats). To enable data science, it is critical that data are in the same format. The GNPS/MassIVE data analysis ecosystem has addressed this issue by converting LC-MS/MS data to an open format (generally MGF or mzML) if they are not already formatted as such[15,16]. This ready conversion facilitates the use of the data through search and filtering tools such as MASST and MassQL.

Of those search tools, MASST is integral to reverse metabolomics. The input for MASST can be provided either through manual entry of the MS/MS spectrum or using universal spectrum identifiers (USIs)[17]. Originally designed as a digital pathway to MS/MS spectra for proteomics[18], USI is now utilized within the GNPS/MassIVE to generate unique identifiers for datasets, files, and point to individual MS/MS spectra of small molecules[17]. While the utilization of USI is an integral part of the GNPS ecosystem, it is worth noting that these identifiers can also be obtained from other data repositories such as PRoteomics IDEntifications database[19] (PRIDE), Massbank[20], MetaboLights[3], Metabolomics Workbench[4], Zenodo[21] or from in-house data and is anticipated to be used in other future repositories that have compatible application programming interface (APIs). The output from MASST is a table of spectral files with their USIs, which can be used to trace back to the original dataset, file, and scan number that matched the parameters specified in the MASST search.

Another challenge in leveraging public metabolomics data for (big) data science applications is the absence of harmonized metadata. Therefore, inspecting and

interpreting results from the tables obtained through MASST searches poses additional challenges. To enhance analysis at the repository scale, the integration of controlled vocabularies is imperative. Initiatives like ReDU[14] focus on capturing metadata as controlled vocabularies within the GNPS ecosystem and have recently expanded to other repositories such as MetaboLights[3] and Metabolomics Workbench[4,22].

Despite these efforts, challenges persist in efficiently capturing all vocabularies and previously deposited data. To further enhance metadata, community curation initiatives have emerged, leading to the development of foodMASST[11], microbeMASST[12], and plantMASST[13], linking files to metadata fields such as food ontology, microbial taxonomy, and plant taxonomy. Together with ReDU, these metadata curation efforts facilitate the visualization of metadata associations, including global distributions, body distributions, organism associations, phenotype, and experimental interventions.

The initial implementation of MASST involved precomputing a global molecular network, which was time-consuming to search through as the volume of MS/MS spectra increased. For instance, in the original implementation searching 110 million spectra took 10 to 20 minutes, and the search time grew as more data was added. Recent strategies, including hyperdimensional computing in graphics processing units (GPUs)[23–25] and the adoption of indexing of spectra strategies, have been introduced to expedite spectral searches[26,27]. The current version of MASST uses the FASST indexing approach[27]. FASST creates a two-dimensional index of MS/MS peaks and intensities, enabling swift retrieval and comparison of query MS/MS to all publicly indexed MS/MS in parallel. Now, it takes seconds to search 2 billion MS/MS spectra. This acceleration in the search process facilitates reverse metabolomics, enabling descriptive summary statistics retrieved from the matches to public data files (**Fig. 1 – Part 2**). This can be achieved for structurally defined and undefined metabolites across diverse organisms, tissues, and diseases to prioritize what data science and/or visualization should be performed (**Fig. 1 – Part 3**).

Anticipated applications of reverse metabolomics

The breadth of applications for reverse metabolomics is nearly limitless, especially as data repositories continue to expand and strategies for making metadata and data ready for data science applications continue to grow. We envision this protocol to provide a valuable hypothesis-driven approach spanning various research domains, achieved through the linkage of sample information (ReDU metadata) with an MS/MS spectrum. Reverse metabolomics can be leveraged for source-tracking environmental contaminants, understanding the biotransformation of different compounds (e.g., drugs, xenobiotics) and identifying their producers. This approach includes identifying where specific molecules are detected (e.g., bacteria, plants, fungi, humans, rodents), discover their locations within tissues and biological fluids (e.g., brain, liver, gallbladder, feces), their geographical distribution (e.g., Europe, United States, Asia), biological sex (e.g.,

male, female) and other observed phenotypes in a biological system. In clinical research, this protocol can facilitate large-scale quantitative meta-analyses, integrating datasets from multiple cohorts to identify potential biomarkers associated with health phenotypes (e.g., obesity, hypertension, inflammatory bowel disease). As metabolomics repositories continue to grow with improved strategies to capture metadata in a data science ready format, more data associations will be uncovered, further empowering this protocol to uncover new biology and uncharted biochemistry.

Comparison to other methods or approaches

We predict that many methods and approaches that enable the uncovering of new biology and biochemistry from this growing public resource will be developed in the upcoming decade, as the value of public mass spectrometry-based metabolomics data is just beginning to be realized. However, at this time, no alternative strategy leverages repository information to uncover new biology that can search known and structurally uncharacterized MS/MS spectra.

Expertise needed to implement the protocol

To use the current implementation to reverse metabolomics, one must become familiar with GNPS, a community-driven ecosystem designed to facilitate data sharing and re-use, and to provide an interface for processing tandem mass spectrometry data[5] and have a basic working knowledge in R or Python. GNPS is suitable for beginners and expert users in the field of mass spectrometry-based metabolomics, and the documentation is available at https://ccms-ucsd.github.io/GNPSDocumentation/. Knowledge of MS/MS fundamentals is required and include mass tolerance, *m/z*, ion intensity/abundance. Also, being able to understand what a MS/MS spectrum is, including being able to judge a good quality vs. lower quality MS/MS spectrum is required (**Box 2**). This protocol also requires the users to have skills to install programs, rename files and create folders. The script used in this protocol can be adapted to the user's needs with bioinformatic skills. We provide the script in two widely used programming languages (Python and R). In this protocol, we use the R script for the step-by-step instructions.

Materials/resources

Software

- Computer with internet access; this protocol was tested using Apple MacBook Pro (specifications: Apple M2 max, 64 GB RAM, 38 cores GPU, 12 cores CPU) and Windows (specifications: 13th Gen Intel(R) core (TM) i7-13850HX, 2100 Mhz, 32 GB RAM, 20 cores, 28 logical processors).
- Web browser (Safari, Google Chrome, Firefox, Microsoft Edge) to access GNPS (https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp?redirect=auth)
- MassQL (https://github.com/mwang87/MassQueryLanguage)

- R and R studio (https://posit.co/download/rstudio-desktop/)
- FASST (https://fasst.gnps2.org/fastsearch/)
- Domain-specific MASSTs; microbeMASST (https://masst.gnps2.org/microbemasst/), plantMASST (https://masst.gnps2.org/plantmasst/), and foodMASST (https://masst.gnps2.org/foodmasst2/)
- Metabolomics Spectrum Resolver (Metabolomics USI)

Required files

- MS/MS spectrum (USI or fragment ions with matching intensities)
- FASST output tables
- ReDU metadata (https://redu.gnps2.org/)

Overview of the method

This protocol aims to provide researchers with a step-by-step guide to contextualize MS/MS spectra that are obtained by fragmentation – mostly through collision induced dissociation- for specific ion forms of molecules[28], whether they are known or structurally yet-to-be-defined (also referred to in the literature as daughter ion spectrum, fragment ion spectrum, MS2, tandem mass spectrum).

To illustrate the approach, we will guide the reader through the process using a single MS/MS spectrum of a microbiome-derived metabolite, specifically an amino acid conjugated bile acid: phenylalanine-cholic acid (Phe-CA). At the time of its discovery, no phenotypic information existed for Phe-CA, limiting our ability to understand the biological implication of the newly-discovered bile acid. Using reverse metabolomics, Phe-CA was found to be prevalent in Crohn's disease patients[8]. Following that, we will present two additional examples, which involve *N*-acyl amides metabolites. *N*-acyl amides are often important signaling molecules in humans. For these two there is limited knowledge about the producers (is it produced by microbes, hosts or found to be part of food), what organs they might be found in and no knowledge about connections to health conditions or interventions. The aim is to learn such associations about known (or unknown) molecules to enable the formulation of hypotheses that cannot be formulated from reading the literature alone (especially in the case for newly discovered molecules or unknowns) and this is an important application of reverse metabolomics. The two *N*-acyl amides are conjugated with the short chain four carbon fatty acid butyrate and no double bonds (C4:0 per lipid nomenclature[29]) – Phenylalanine-C4:0 and Histidine-C4:0 (Phe-C4:0, His-C4:0, where C4:0 denotes a fatty acid with four carbons and 0 double bonds), showcasing how one can leverage and visualize the metadata association discovered through metadata

summaries of files found to match in MASST searches. Finally, we offer general suggestions for steps to further validate the results.

Step-by-Step Procedure

A visual cheatsheet guide of the step-by-step procedure is provided as **supporting information** to help the user. The protocol begins by gaining access to an MS/MS spectrum **(Fig. 1 – Part 1)**. **Part 2** involves using MASST to search a selected database with the obtained MS/MS spectrum. In **part 3**, a data table linking metadata is created and focuses on formatting the data table appropriately for data science applications on the curated reverse metabolomics data table. Subsequently, data analysis and visualization can be conducted **(Fig. 1)**.

Accessing the MS/MS spectrum for searching GNPS/MassIVE

Timing: ~ 5 min**.**

Important: Manual inspection of the spectra before proceeding to step 2 (FASST) is encouraged to ensure that one searches with spectra that have as little low-intensity ions, which are often background noise of the instrument, as possible.

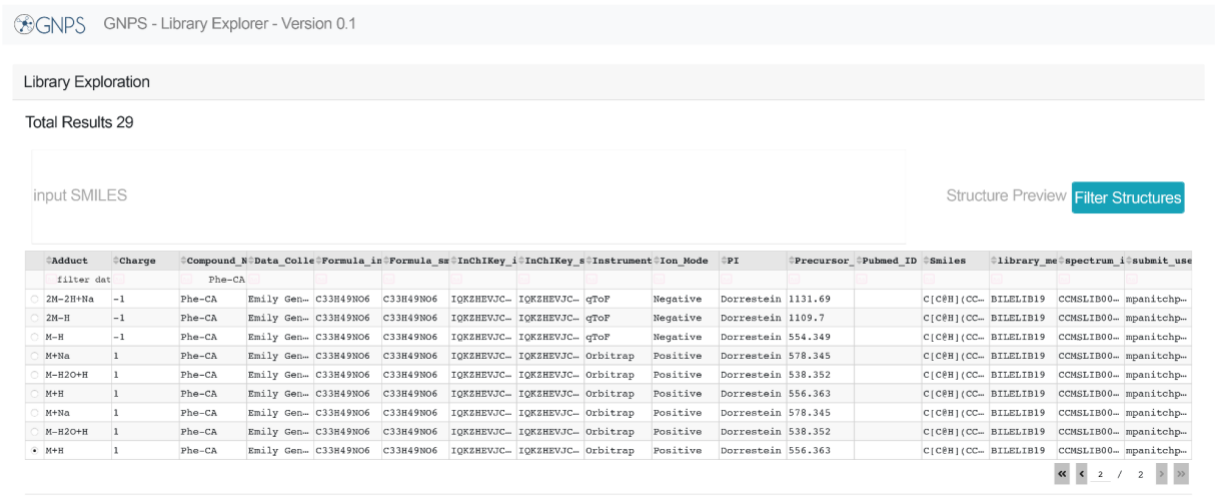
To perform a search with the fast mass spectrometry search tool, one must possess the MS/MS spectrum that one would like to query (**Box 3).** While manual entry of the MS/MS is feasible by entry of the precursor mass, the fragment ions and their relative intensities, here we utilize the USI to ensure that all the data embedded in the MS/MS spectrum is leveraged in the search. In this example, we obtain an USI and use an MS/MS spectrum from the reference MS/MS library of known molecules that is found within the GNPS ecosystem.

1. The GNPS reference library can be accessed through this link: https://ccms-ucsd.github.io/GNPSDocumentation/gnpslibraries/.
    - Click on "view" and then select the desired library. In this case, we choose the "All Public Spectra at GNPS" redirecting to https://library.gnps2.org/.
    - Add the compound name "Phe-CA" to the compound name column (**Fig. 2a**) and press return (in Mac) or enter (in Windows). **Note:** This action filters the library for all compounds containing that name. Frequently, analogs of compounds, as well as different ion forms, such as MS/MS derived from in-source fragment ions, proton, Na+, K+, or other adducts or multimers, may be present[28,30–34]. The prevalence of different ion forms depends on experimental conditions. Users should be aware that multiple spectra may exist for the same molecules because they could have been acquired using different instruments and/or different collisional energies and/or different ion forms (e.g., different adducts, multimers,

in-source fragment ions) which results in differences not only in the observed MS/MS spectra but also the final results. If one wants to be as comprehensive as possible, it is encouraged to search with as many of the MS/MS spectra available for a given compound.

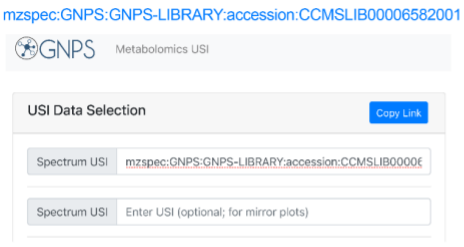
- Navigate and select the M+H ion form, and the circle in the left column should turn blue (**Fig. 2a**). After clicking the M+H ion form, scroll below to get the universal spectrum identifier (shown as a blue hyperlink, **Fig. 2b**). This link can either be copied directly from this page or by clicking on the hyperlink, which redirects to the spectral viewer (**Fig. 2c**). From the spectral viewer, the USI can be copied as indicated in red (Spectrum USI, **Fig. 2b**). An USI for the M+H of the Phe-CA example is mzspec:gnps:GNPS-LIBRARY:accession:CCMSLIB00006582001. **Note:** The 'CCMSLIB' in the USI refers to the specific accession number for spectral libraries in the UCSD Center for Computational Mass Spectrometry, indicating that the MS/MS spectrum can be found in GNPS and the GNPS library. Examples of other USIs obtained from different repositories can be found in Bittremieux et al.[17]

a



Adduct	Charge	Compound_N	Data_Colle	Formula_in	Formula_sm	InChIKey_i	InChIKey_s	Instrument	Ion_Mode	PI	Precursor_	Pubmed_ID	Smiles	library_me	spectrum_i	submit_use
filter dat		Phe-CA														
2M-2H+Na	-1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	qToF	Negative	Dorrestein	1131.69		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
2M-H	-1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	qToF	Negative	Dorrestein	1109.7		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M-H	-1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	qToF	Negative	Dorrestein	554.349		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M+Na	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	578.345		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M-H2O+H	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	538.352		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M+H	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	556.363		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M+Na	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	578.345		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M-H2O+H	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	538.352		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…
M+H	1	Phe-CA	Emily Gen…	C33H49NO6	C33H49NO6	IQKZHEVJC…	IQKZHEVJC…	Orbitrap	Positive	Dorrestein	556.363		C[C@H](CC…	BILELIB19	CCMSLIB00…	mpanitchp…

b



c

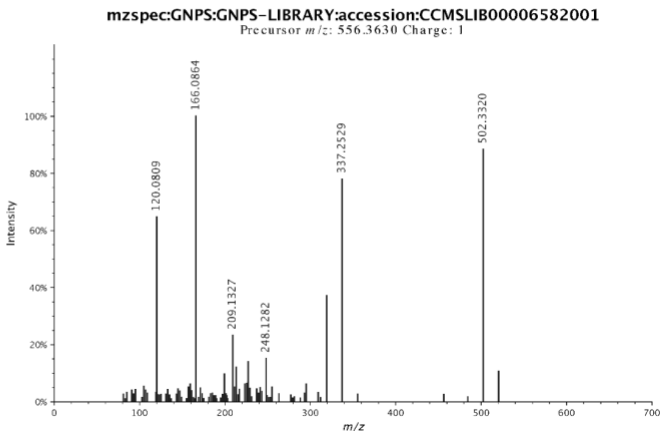


Fig. 2 | Example of GNPS library explorer for Phe-CA. a, Library results from page 2 for Phe-CA in the GNPS library. The library exploration table displays many columns with information such as adduct, charge, compound name, instrument, ion mode, and precursor *m/z*. **b,** Universal Spectrum Identifier information and spectrum USI entry on the metabolomics USI interface. **c,** Visualization of the MS/MS spectrum by clicking on the circle in a (indicated by the dot), or on the USI (blue hyperlink) in b.

Anticipated results: Obtain the USI for the MS/MS spectrum of the molecules of interest. In this example, it is the MS/MS of the M+H ion form acquired in an Orbitrap instrument of Phe-CA from the GNPS library.

Troubleshooting: If the MS/MS of the compound of interest is not available in the GNPS reference library or the user is interested in an USI from an unannotated MS/MS, although it is possible to generate an USI from one's computer, it is encouraged to upload the reference spectra to the GNPS reference library or find the spectra in a file or dataset that is uploaded to GNPS/MassIVE. This will allow the creation of an USI for any MS/MS of interest.

Performing MASST using the fast search tool.
Timing: 2-180 seconds

Fast search enables users to query a single MS/MS spectrum to retrieve identical or structurally related molecules from the GNPS/MassIVE repository **(Fig. 3)**. There are two options to perform FASST search: by providing an USI of the MS/MS spectrum (option A) and manual entry of the MS/MS spectrum (option B) **(Fig. 3a)**. For more information on how to access MS/MS spectra, refer to **Box 3**.

2. **Option A – USI**. Copy and paste the USI from the GNPS library to the spectrum USI section in the fast search tool (https://fasst.gnps2.org/fastsearch/) as shown in **Fig. 3a**. Using the universal spectrum identifier (USI) simplifies the fast search by automatically collecting the precursor ion mass and the charge state information before querying the MS/MS spectrum against the GNPS/MassIVE repository.

3. **Option B – Manual entry of MS/MS spectrum.** To enable manual entry of the MS/MS spectrum, users need to click on the blue hyperlink (No USI? Click to enter peaks manually – **Fig. 3a**) and the MS/MS spectrum should be formatted as a two-column table where each line contains the *m/z* value (mass-to-charge ratio) and the corresponding intensity. Additionally, users must enter the precursor *m/z* and the charge state **(Fig. 3a)**.

This information is important as it defines the precursor ions from which the MS/MS spectrum was generated.

Users must define some settings before launching the MASST search **(Fig. 3a)**. The fast search tool is customizable and the default settings are the following: the precursor ion and the fragment ion tolerance, 0.05 *m/z*; cosine similarity, 0.7; analog search, No (Described below).

- PM (precursor mass) Tolerance (Da): Parent mass peak tolerance. For high-resolution mass spectrometers (orbitraps, qTOF etc.), the recommended starting value is 0.05 *m/z*.

- Fragment Tolerance (Da): Tandem MS peak tolerance. For high-resolution mass spectrometers (orbitraps, qTOF etc.), the recommended value is 0.05 *m/z*.

- Cosine Threshold: A metric that indicates how similar two MS/MS spectra are; a cosine of 1 denotes a perfect match and a cosine score of 0 means no similarity between the two spectra. As default, a cosine of 0.7 is used. This parameter can be adjusted. Careful examination of the reference spectra vs the queried spectra is encouraged to prevent downstream interpretation errors.

- Analog search: This parameter can be enabled to search and find MS/MS of structurally related molecules and a specific range of delta masses between the precursor ions can be defined. This will use a modified cosine that includes all ions that have precursor mass differences[35].

- Library name: Select a library from the dropdown menu. **Note:** Users can search against multiple libraries that are found in the dropdown menus. The libraries include the gnpslibrary (spectral libraries found in GNPS), massivedata_index (search in datasets available on MassIVE) among others. This parameter specifies which set of indexed spectra will be used for the search. In this protocol, we will select the GNPS/MassIVE repository. The indexing of GNPS/MassIVE is updated from time to time, and it is encouraged to put the most up-to-date library (for instance, gnpsdata_index_11_25_23). Current efforts are being made to expand the MASST search to other repositories and will be available by selecting the 'metabolomicspanrepo_index_latest' library.

Note: The ultimate choice of parameters selection is done by the user and the goals they have for the results. More restrictive parameters mean matches will be lost while looser restrictions mean one finds more matches but will also have more incorrect matches –

this is critical to keep in mind when performing the final formulation of a hypothesis with the result summaries.

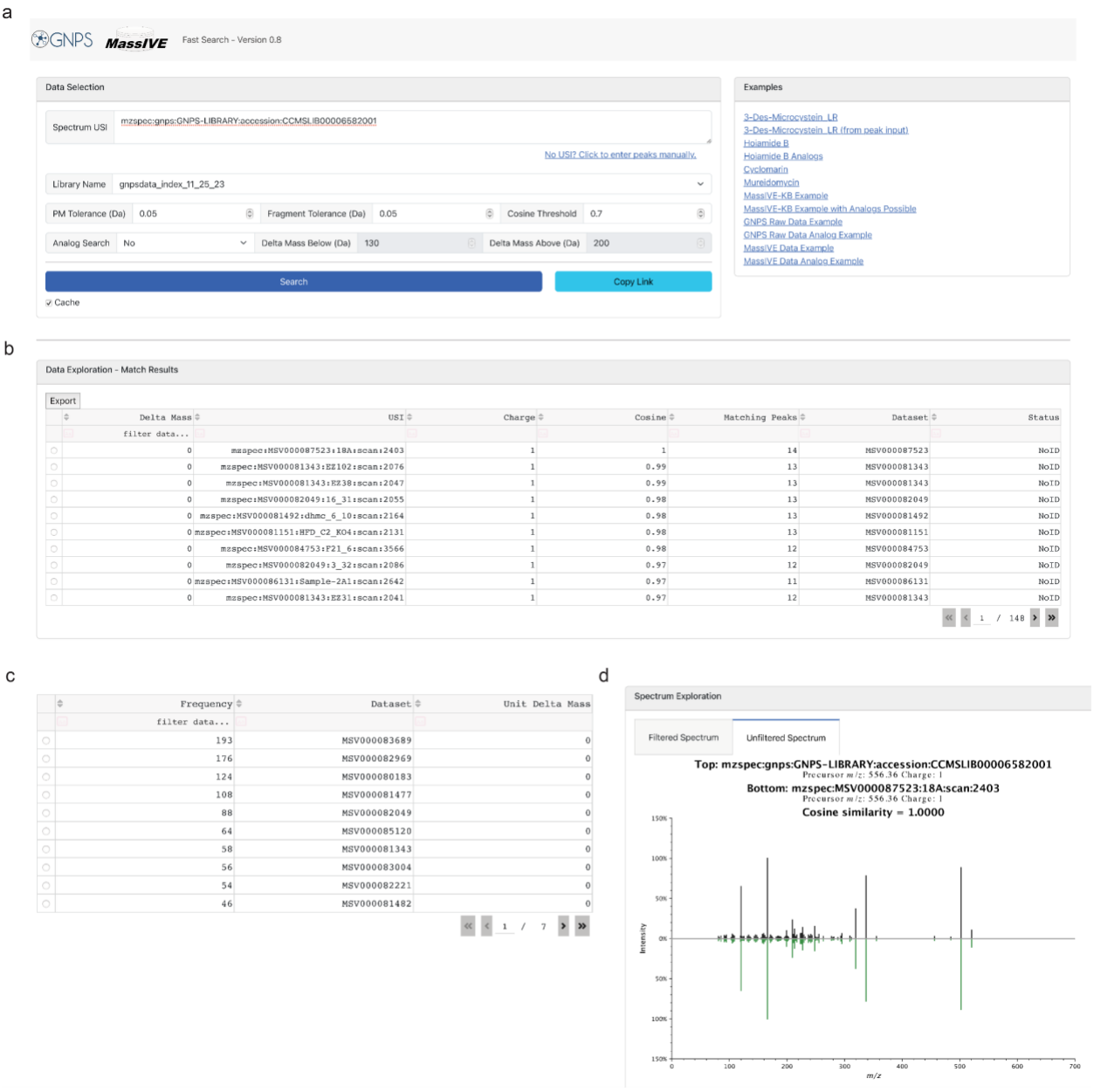


Fig. 3 | Example using the search tool for Phe-CA. a, The fast search tool returns MS/MS spectra within seconds. The Data selection section allows users to input USI or manual entry of the MS/MS spectrum, select the library, and modify parameters for the MASST search. **b,** The Data Exploration section displays the results as a table with information such as delta mass, USI (reference spectrum), charge, cosine similarity, the number of matched peaks, dataset, and status. **c,** A three-column table showing the

frequency found in each dataset and the mass difference from the precursor ion. **d,** Mirror plot between the queried spectrum and the reference spectrum found in GNPS/MassIVE.

Exploration of the MASST results

4. Navigate the FASST search web page to locate 'Data Exploration – Match Results**'.** This section presents the fast search data table and includes columns with information such as delta mass, USI, charge state, cosine similarity score, the number of matching peaks, the MassIVE number ID, and status **(Fig. 3b)**.

5. Users can explore the distribution of delta masses collected from GNPS/MassIVE by clicking on the subsection 'Group by Dataset/Delta Mass'. A table is also generated and provided information about the frequency of the MS/MS spectrum found in each dataset (**Fig. 3c**). **Note:** We advocate for manual inspection of the queried MS/MS spectrum against the reference spectrum from GNPS/MassIVE. Click on the open circle icon within the data table results. When conducting fast search with analog search "ON", particular attention should be paid to the mirror plot to avoid misinterpretation of the spectral data (for more information on mirror plots, see **Box 2**. A mirror plot allows the users to simultaneously visualize the queried and the reference spectrum to evaluate similarities **(Fig. 3d)**. The Metabolomics Spectrum Resolver offers an alternative to visualize the mirror plot (https://metabolomics-usi.gnps2.org/)[17].

6. The 'Spectrum Exploration' section displays the filtered and the unfiltered mirror plot between the queried spectrum (user input) and the reference spectrum (GNPS/MassIVE repository). The number of matched peaks and their intensities are parameters that the users need to consider to evaluate the quality of the queried and the reference MS/MS for a more accurate identification (**Fig. 3d**).

7. Click on the export button (marked in red, **Fig. 3b**) on the top left corner of the 'Data Exploration – Match Results' section to download the table and store it at a known location.

Linking MASST search output to available metadata.

Timing: ~ 30-60 min

Connecting sample information to each scan retrieved from the fast search tool is accomplished through a coding platform. This process requires the transformation, formatting, merging, filtering, and normalization of data tables. Once the data transformation step is complete, phenotypic association results can be visualized using, for instance, heatmaps. Here we focus on finding the tissue and biofluid distribution of three different molecules of interest – Phe-CA (mzspec:gnps:GNPS-LIBRARY:accession:CCMSLIB00006582001), Phe-C4:0 (mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00010010601), and His-C4:0 (mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00011434738) as well as if they are associated with specific human diseases. For this purpose, we used R and Python as they are the two

most common coding languages used in metabolomics data analysis and offer flexibility to the users. This protocol provides a detailed step-by-step instruction using R. To make our protocol more inclusive and accessible to many scientists, we have created a similar workflow using Python (see code availability section).

Fast search tool – FASST batch workflow

The protocol described in this article is designed to efficiently query a small number of MS/MS spectra. Users should note that there is a FASST batch workflow available at this link (https://gnps2.org/workflowinput?workflowname=fasst_batch_workflow), which allows for the search of multiple USIs. However, the script provided in this protocol is not designed to incorporate the output from the batch workflow, but can be adapted with basic coding skills.

8. Download fast search results. The fast search tool output can be downloaded by clicking on the export button on the top left side of the results table which will download a .csv file (see **Fig. 3b**). Earlier, we showcased a fast search tool for Phe-CA. Users need to repeat this process for Phe-C4:0 and His-C4:0 using the USIs provided above. Note: The FASST search is reproducible when the same output table with the same USI's is provided. However, as repositories continue to grow and other repositories will be added in the near future (currently under development), the underlying repository data will continue to expand and therefore one can expect to see additional results not captured yet at the time of this publication.

9. Download ReDU metadata. Go to the following link: https://redu.gnps2.org. Click on “Download Database” in the top right corner.

Once all data tables are downloaded, users should store the files at a known location in a specific folder containing only the .tsv file. This location will be used to define the working directory in R and to import the tables **(Fig. 4 – Steps 1-2)**.

10. Installation of RStudio and preparing the environment. RStudio is the programming interface used throughout this protocol and needs to be installed along with R and can be accessed here https://posit.co/download/rstudio-desktop/. R is available for Linux, macOS, and Windows users. After installing R, users need to download RStudio, which provides a user-friendly interface for analysis. Defining a workspace is the first step and should include all files generated from the fast search. The ReDU metadata should be kept separated from the fast search output.

11. Set working directory (WD). Once RStudio has been downloaded by the users and is operational, we highly recommend creating a new project which will also define a new working directory.

- Open RStudio and (Optional) click on ‘file’, then ‘new project’.
- (Optional) Select ‘New Directory’, followed by ‘New Project’, define a directory name and click on ‘Create Project’.
- Go to the toolbar of RStudio and click on ‘File’ then ‘New File’, and ‘R Script’. We recommend saving the new script before importing all data tables. The working directory folder should encompass all the files required for the analysis.

```
setwd("/yourpath")
```

- Package requirements and installation: this protocol was developed using R version 4.3.1 (2023-06-16). R packages are required to accomplish this protocol and are loaded at the beginning of the script. All packages are available in the CRAN repository.

12. Install R packages from the CRAN Repository using the install.packages() function. Once the packages have been installed in R, lines 5 to 7 of the R script can be inactivated using the hashtag (#) to prevent reinstallation when the script is automatically launched.
    - Data import: *data.table*[36] (version 1.15.4)
    - Data analysis: *tidyverse*[37] (version 2.0.0)
    - Data visualization: *pheatmap*[38] (version 1.0.12)

```
install.packages("data.table", dependencies = TRUE)
install.packages("tidyverse", dependencies = TRUE)
install.packages("pheatmap", dependencies = TRUE)
```

13. Load the R packages using the library() function.

```
library(data.table)
library(tidyverse)
library(pheatmap)
```

14. Data import and merging of fast search results with ReDU metadata. The fast search results should be downloaded and stored in the working directory in a specific folder (see point 8). The files from the fast search must be renamed using the molecule name that was queried. This will be important in Step 16 because the name of the file will be used to automatically fill the dataframe under the “Compound” column.
    - The path leading to the .csv files needs to be defined at the beginning of the script, which will create the “folder_path” object. **Note:** Windows users may have to use backslashes instead of forward slashes as used here.

```
folder_path <- "/folder/subfolder"
```

15. Import the ReDU metadata from the working directory (WD) using the fread() function **(Fig. 4 – Step 3)**.

```
processed_redu_metadata <- "all_sampleinformation.tsv"

if (!file.exists(file.path(getwd(), processed_redu_metadata))) {
    redu_url <- "https://redu.gnps2.org/dump"
    options(timeout = 600)
download.file(redu_url, file.path(getwd(), processed_redu_metadata), mode = "wb")
    redu_metadata <- data.table::fread(processed_redu_metadata)
    } else {
    redu_metadata <- data.table::fread(processed_redu_metadata)
    }
```

16. Get the list of all the .csv files in the subfolder in the WD using list.files() function and read all the .csv files and then, create a new column named “Compound” with the name of the file **(Fig. 4 – Step 4)**

```
file_list <- list.files(folder_path, pattern = "*.csv", full.names = TRUE)

df_list <- lapply(file_list, function(file) {
    df <- read_csv(file)
df$Compound <- tools::file_path_sans_ext(basename(file))
    return(df)})
```

17. Combine all the Fast Search results into a single df using the bind_rows() function **(Fig. 4 – Step 5)**.

```
molecules_interest <- bind_rows(df_list)
```

18. Formatting the Fast Search table and ReDU metadata for data merging. The USI column from the fast search df (molecules_interest) and the “filename” column in the ReDU metadata are targeted as both sharing common information on the MassIVE ID (dataset reference number) and the associated filename.

19. The USI column from molecules_interest is targeted for merging with ReDU. Both tables share the same information (MassIVE ID and the filename) which can be used for combining both dfs.

- Create the function named MassiveID_filename() to extract the second and the third part of the segment in the “USI” column of the molecules_interest df.
- Rename each row in the column “USI” by keeping the MassIVE ID (second part) and the filename (third part) of the USI string.

```
MassiveID_filename <- function(USI) {
    parts <- unlist(strsplit(USI, ":"))
    paste(parts[2], parts[3], sep = ":")}
molecules_interest$USI <- vapply(molecules_interest$USI, MassiveID_filename,
    FUN.VALUE = character(1))
```

- To be compatible for merging, ReDU metadata requires more modification steps. In the column “filename” from ReDU metadata, replace all the “/” by colons “:”.
- Remove the first two characters (f.) by using the substring() function.
- Remove the extension (.mzML and .mzXML) using the gsub() function.

```
redu_metadata_filtered <- redu_metadata |>
dplyr::filter(str_detect(USI, "\\.mzML|\\.mzXML"))

redu_metadata_filtered$USI <- gsub("/", ":", redu_metadata_filtered$USI)
redu_metadata_filtered$USI <- gsub(".mzXML","", redu_metadata_filtered$USI)
redu_metadata_filtered$USI <- gsub(".mzML","", redu_metadata_filtered$USI)
```

- In ReDU metadata, create and apply the ReDU_USI function to keep the first part of the string in the column “filename”, which is the MassIVE ID and the last part of the string (the filename), so it becomes compatible for merging with the df molecules_interest.

```
ReDU_USI <- function(USI) {
    parts <- unlist(strsplit(USI, ":"))
    paste(parts[2], parts[length(parts)], sep = ":")}

redu_metadata_filtered$USI <- vapply(redu_metadata_filtered$USI, ReDU_USI,
    FUN.VALUE = character(1))
```

20. Merge molecules_interest (fast search output) with ReDU metadata by using the left_join() function **(Fig. 4 – Step 6)**.

```
ReDU_MASST <- left_join(molecules_interest, redu_metadata_filtered, by = "USI",
    relationship = "many-to-many")
```

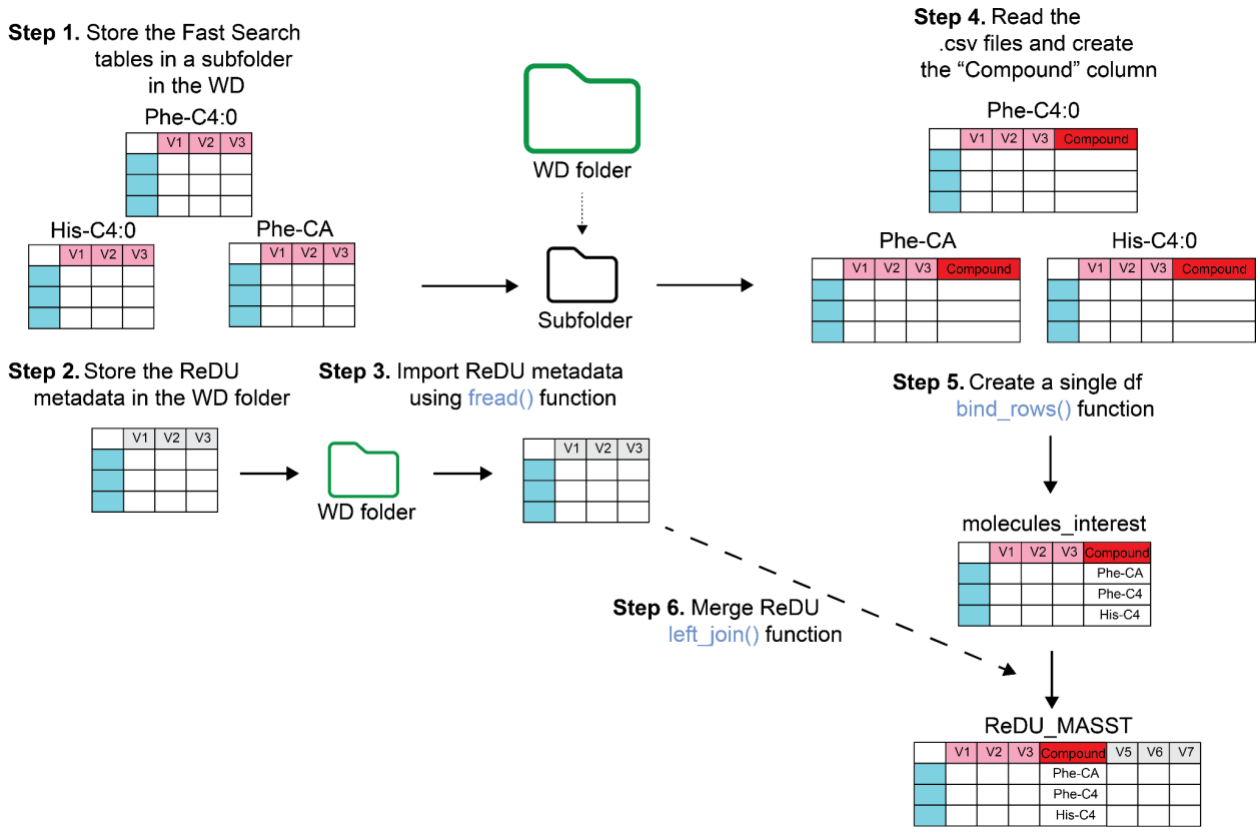


Fig. 4 | A schematic illustration on merging fast search tables with ReDU metadata. The first steps consist of automatically importing the tables after the Fast Search in a subfolder in the working directory (WD) folder (Steps 1-3). Then, a new column "Compound" is created in each df with the name of the molecule that was queried (Step 4). All dfs are combined into a single df (molecules_interest) and merged with ReDU metadata, resulting in the ReDU_MASST df (Step 5-6).

Anticipated results: All files generated from the fast search tool should be combined into a single dataframe, including the addition of the 'Compound' column (column in red in **Fig. 4**). After merging with ReDU metadata, a single dataframe is created and contains both the FASST search tables and ReDU metadata. In the resulting ReDU_MASST df, multiple rows will have an NA, and indicate a missing value due to lack of ReDU metadata. This is expected because not all files in the public domain will have metadata in the ReDU format.

Metadata-driven analysis and visualization.

Timing: ~ 30 – 60 min

In this section, we will illustrate the distribution patterns of His-C4, Phe-C4, and Phe-CA across body parts and biofluids in humans and rodents. Additionally, we will guide the users in evaluating the potential health implications and their prevalence in specific diseases. It is important to highlight that not all files available in public repositories have associated ReDU metadata (sample information) which impedes our ability to fully leverage public data. We strongly encourage the scientific community to make their data available with comprehensive metadata. As more data are being deposited in repositories, more matches will be uncovered and more results will be embedded in heatmaps.

21. (Optional) Body parts and biofluids standardization. To ease analyses, prevent errors, and improve data visualization, body parts and health status can be modified. **Note:** When the names use a mixture of upper vs lower case, the standard library function tolower() can be used to ensure all standardized names are lowercase.

- Concatenate all skin locations to 'skin': skin of trunk, skin of pes, head of neck skin, axilla skin, skin of manus, arm skin, and skin of leg.
- Concatenate serum and plasma to blood.
- Convert uppercase to lowercase for health status: Chronic Illness to chronic illness and Healthy to healthy.

```
ReDU_MASST_standardize <- ReDU_MASST |> dplyr::mutate(
UBERONBodyPartName = str_replace_all(UBERONBodyPartName, 'skin of trunk|skin of pes|head or neck skin|axilla skin|skin of manus|arm skin|skin of leg', 'skin'),
UBERONBodyPartName = str_replace_all(UBERONBodyPartName, 'blood plasma|blood serum', 'blood'), HealthStatus = str_replace(HealthStatus, 'Chronic Illness', 'chronic illness'), HealthStatus = str_replace(HealthStatus, 'Healthy', 'healthy'))
```

22. NCBI taxonomy filtering. Separating humans and rodents results can be used to show translational impact of the observations to assess body part distribution and to associate metabolites to health phenotypes. All human-associated information can be selected using the 'NCBITaxonomy' column and filtered using the filter() function for '9606|Homo sapiens'.

23. Create a new df (df_humans) in which only human-related information will be embedded **(Fig. 5 – Step 7)**.

```
df_humans <- ReDU_MASST_standardize |>
dplyr::filter(NCBITaxonomy == "9606|Homo sapiens")
```

24. Create a new df (df_rodents) for which all different rodent taxonomy identifications are combined **(Fig. 5 – Step 8)**.

- Taxonomy IDs of Rodents: 10088|Mus, 10090|Mus musculus, 10105|Mus minutoides, 10114|Rattus, 10116|Rattus norvegicus.

```
list_rattus_mus <- c('10088|Mus', '10090|Mus musculus', '10105|Mus minutoides',
                     '10114|Rattus', '10116|Rattus norvegicus')

df_rodents <- ReDU_MASST_standardize |>
  dplyr::filter(NCBITaxonomy %in% list_rattus_mus)
```

25. Number of occurrences for organ distribution. Get an overview of the data by counting the number of scans per organ and how they are distributed in humans and rodents.

- Create a function analyze_counts() that will generate new dfs with four columns **(Fig. 5 – Step 9)**.

```
analyze_counts <- function(df, column_interest) {
  df_body_parts <- df |> distinct(across(all_of(column_interest)))
  df_BodyPartName_counts <- df |>
    count(across(all_of(column_interest)), name = "Counts_fastMASST")
  compounds <- df |>
    group_by(across(all_of(column_interest))) |>
    summarise(Compounds = n_distinct(Compound),
      CompoundsList = toString(unique(Compound))) |>
    ungroup()
  combined <- df_body_parts |>
    left_join(df_BodyPartName_counts, by = column_interest) |>
    left_join(compounds, by = column_interest)
  return(combined)}
body_counts_humans <- analyze_counts(df_humans, "UBERONBodyPartName")
head(body_counts_humans)
body_counts_rodents <- analyze_counts(df_rodents, "UBERONBodyPartName")
head(body_counts_rodents)
```

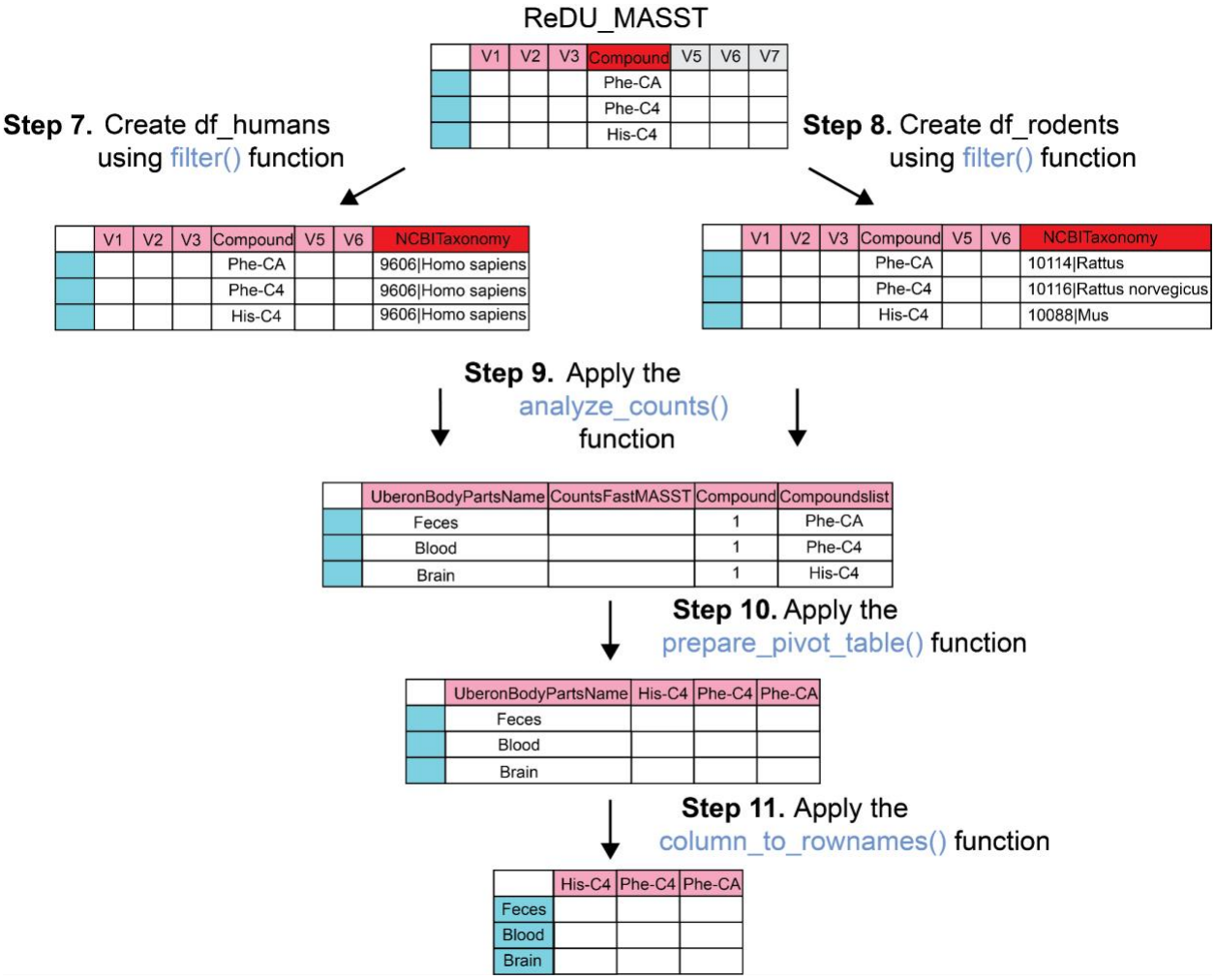


Fig. 5 | A schematic illustration of tables formatting to generate organ distribution heatmaps for humans and rodents. Merged fast search and ReDU tables are filtered to contain only human-related or rodent-related information (Steps 7-8). Functions are created to assess the counts of each queried molecule to a specific UBERON[39] body parts name and are transformed for data visualization (Steps 9-11).

After exploring the data, the next step is to format the data structure for visualization. The aim is to link the number of scans obtained by performing the fast search to evaluate body part distribution. Although the reverse metabolomics workflow focuses on finding the body parts distribution of the queried molecules, other variables such as life stage and biological sex can be incorporated based on user-defined research questions. To accomplish this, we need to create a new table structured as follows: the first column enumerates all unique UBERON body parts and the subsequent column indicates the counts of how many times the MS/MS spectrum of each molecule was retrieved as associated with a specific body part in the FASST searches **(Fig. 5)**.

26. Data visualization using heatmaps. The count table **(Fig. 5 – Step 10)** need to be transformed and using a custom function prepare_pivot_table(), then the first column 'UBERONBodyPartName' must become the row names by applying column_to_rownames() function, already embedded in the *tidyverse* package.

```
prepare_pivot_table <- function(df, column_interest, compound) {
                    grouped_df <- df |>
          group_by(across(all_of(c(compound, column_interest)))) |>
                    summarise(Count = n(), .groups = 'drop')
                    pivot_table <- grouped_df |>
    pivot_wider(names_from = all_of(compound), values_from = Count, values_fill =
                    list(Count = 0))
                    return(pivot_table)}
                    variable <- 'UBERONBodyPartName'
pivot_table_humans <- prepare_pivot_table(df_humans, variable, 'Compound')
pivot_table_rodents <- prepare_pivot_table(df_rodents, variable, 'Compound')
```

- Apply the column_to_rownames() function **(Fig. 5 – Step 11)**. The modified df will be structured as the following: all the body parts are the row names and all columns are the molecule names with the number of counts filling the df.

```
humans_molecules_counts_by_bodypart <- pivot_table_humans |>
          dplyr::arrange(UBERONBodyPartName) |>
    tibble::column_to_rownames("UBERONBodyPartName")
rodents_molecules_counts_by_bodypart <- pivot_table_rodents |>
          dplyr::arrange(UBERONBodyPartName) |>
    tibble::column_to_rownames("UBERONBodyPartName")
```

- Validate that all values in the dfs are of the numerical class.

```
humans_molecules_counts_by_bodypart <- humans_molecules_counts_by_bodypart |>
          dplyr::mutate(across(everything(), as.numeric))
rodents_molecules_counts_by_bodypart <- rodents_molecules_counts_by_bodypart |>
          dplyr::mutate(across(everything(), as.numeric))
```

- Define three colors for heatmap visualization that will be used as the scale gradient. We chose a color gradient from white, light blue, and coral which indicate low, intermediate, and high counts of the molecule of interest per organ. Users can modify these colors based on their preferences.

```
colors_version <- c("#FFFFFF", "#C7D6F0", "#EBB0A6")
color_gradient <- colorRampPalette(colors_version)
gradient_colors <- color_gradient(30)
```

- (Optional) Next, we applied log scale to the value in the df for better visualization.

```
log_humans_molecules_counts_by_bodypart <- log2(1 +
humans_molecules_counts_by_bodypart)
log_rodents_molecules_counts_by_bodypart <- log2(1 +
rodents_molecules_counts_by_bodypart)
```

- Create and export heatmaps showing the organ distribution of the molecules of interest in humans **(Fig. 6 – Step 12)** and rodents **(Fig. 6 – Step 13)**. **Note:** the default clustering method is "complete," however other options are available such as ["ward.D", "ward.D2", "single", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC), or "centroid"].

```
Organ_humans <- pheatmap(log_humans_molecules_counts_by_bodypart,
                         color = gradient_colors,
                         cluster_rows = FALSE,
                         cluster_cols = TRUE,
                         angle_col = 90,
                         main = "Organ distribution in humans",
                         fontsize = 10,
                         cellwidth = 15,
                         cellheight = 15,
                         treeheight_row = 100,
                         fontsize_row = 12,
                         fontsize_col = 12,
                         legend_fontsize = 10,
                         border_color = NA)
Organ_humans

ggsave("Organ_distribution_in_humans.pdf", plot = Organ_humans, width = 10, height
                                        = 10, dpi = 900)

Organ_rodents <- pheatmap(log_rodents_molecules_counts_by_bodypart,
                         color = gradient_colors,
                         cluster_rows = FALSE,
                         cluster_cols = TRUE,
                         angle_col = 90,
                         main = "Organ distribution in rodents",
                         fontsize = 10,
                         cellwidth = 15,
                         cellheight = 15,
```

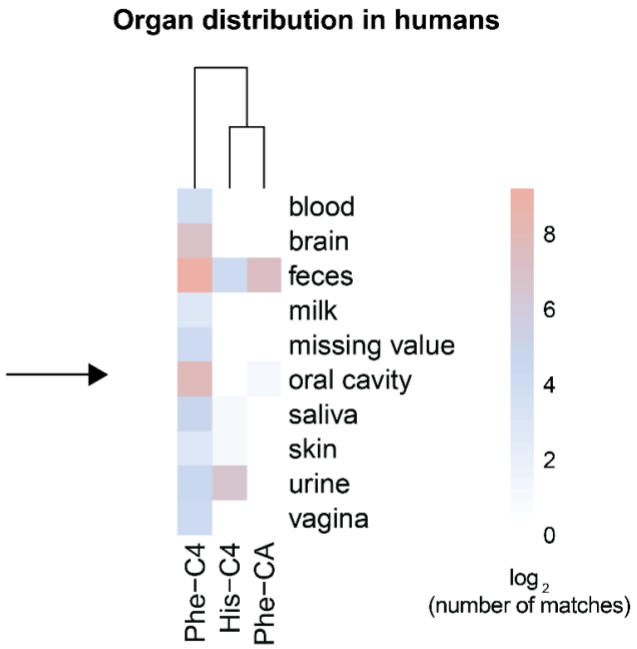
```
                    treeheight_row = 100,
                    fontsize_row = 12,
                    fontsize_col = 12,
                    legend_fontsize = 10,
                    border_color = NA)
Organ_rodents

  ggsave("Organ_distribution_in_rodents.pdf", plot = Organ_rodents, width = 10,
                                        height = 10, dpi = 900)
```

log_humans_molecules_counts_by_bodypart

	His-C4	Phe-C4	Phe-CA
Feces			
Blood			
Brain			

Step 12. Apply the pheatmap() function

```
pheatmap(log_humans_molecules_counts_by_bodypart,
color = gradient_colors,
cluster_rows = FALSE,
cluster_cols = TRUE,
angle_col = 90,
main = "Organ distribution in humans",
fontsize = 10,
cellwidth = 15,
cellheight = 15,
treeheight_row = 100,
fontsize_row = 12,
fontsize_col = 12,
legend_fontsize = 10,
border_color = NA)
```

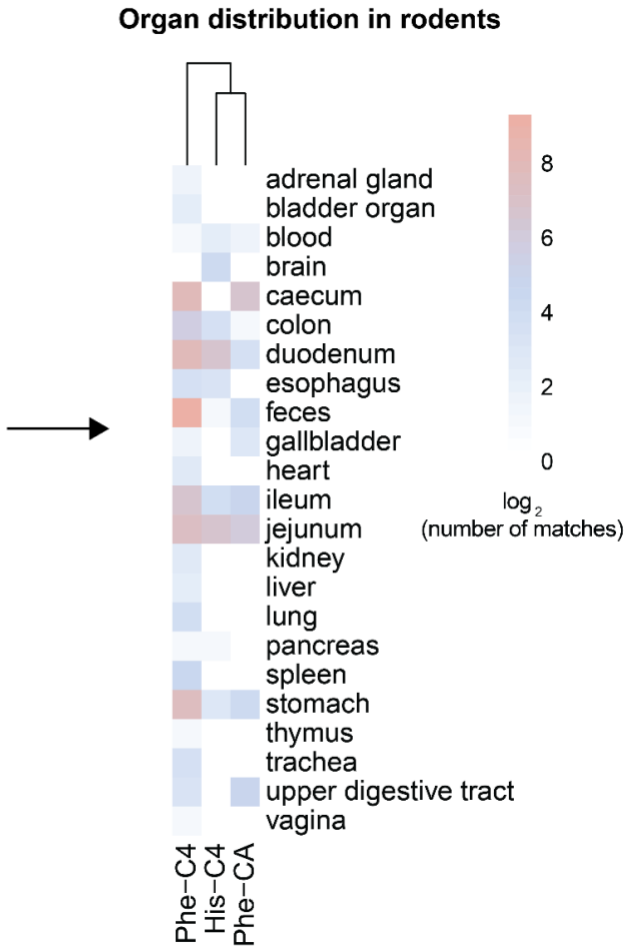


log_rodents_molecules_counts_by_bodypart

	His-C4	Phe-C4	Phe-CA
Feces			
Blood			
Brain			

Step 13. Apply the pheatmap() function

```
pheatmap(log_rodents_molecules_counts_by_bodypart,
color = gradient_colors,
cluster_rows = FALSE,
cluster_cols = TRUE,
angle_col = 90,
main = "Organ distribution in rodents",
fontsize = 10,
cellwidth = 15,
cellheight = 15,
treeheight_row = 100,
fontsize_row = 12,
fontsize_col = 12,
legend_fontsize = 10,
border_color = NA)
```



825

Fig. 6 | Heatmap creation showing organ distribution in humans and rodents of His-C4:0, Phe-CA, and Phe-C4:0 across public repositories with ReDU metadata. The pheatmap() function is used to visualize the organ distribution pattern of the molecules of interest in humans (Step 12). The pheatmap() function is used to visualize the organ distribution pattern of the molecules of interest in rodents (Step 13). **Note:** Missing value indicates no information (in ReDU metadata) in relation to the observed phenotype, however it denotes matches are available in relation to the queried MS/MS spectrum. It should be noted that as the public data with controlled ReDU ontologies grows, the results will be included in the above results (and thus visualization may vary over time).

27. Health phenotype association. Imagine discovering a new mass spectrometry feature or molecule, perhaps found in an animal model or human cohort, with unknown associations to disease, diet patterns, or medical interventions. It is possible that this feature has been detected before in clinical untargeted metabolomics studies but was never reported or discussed in the original article. In the next section, we illustrate how reverse metabolomics can be used to identify associations with health phenotypes, setting the stage for formulating testable hypotheses for follow-up experiments. This workflow aims to associate information to a structurally known or unknown molecule. For instance, more information on biological sex, life stage, disease, and health status can be retrieved from the ReDU metadata.

- Navigating ReDU metadata. The table is imported at the beginning of the script (see point 16).
- Filter the ReDU metadata by separating humans and rodents information in two dfs using the filter() function **(Fig. 7 – Step 14)**.

```
df_redu_humans <- redu_metadata |>
  dplyr::filter(NCBITaxonomy == "9606|Homo sapiens")

df_redu_rodents <- redu_metadata |>
  dplyr::filter(NCBITaxonomy %in% list_rattus_mus)
```

- Subset the ReDU metadata for humans and rodents by defining new dfs based on different information found in ReDU. For instance, DOIDCommonName reports information on the disease ontology, and a new df is created (human_ReDU_DOIDCommonName) where only disease information is embedded **(Fig. 7 – Step 15)**. **Note:** As the public data with controlled ReDU ontologies grows, the results will include the results described here and information of that additional data will then be included as well so the user results may vary.

```
human_ReDU_LifeStage <- df_redu_humans |>
  dplyr::count(LifeStage) |>
  dplyr::rename(LifeStage_counts = n, LifeStage = LifeStage)
human_ReDU_LifeStage$LifeStage_counts <-
  as.numeric(human_ReDU_LifeStage$LifeStage_counts)

human_ReDU_DOIDCommonName <- df_redu_humans |>
  dplyr::count(DOIDCommonName) |>
  dplyr::rename(DOIDCommonName_counts = n, DOIDCommonName =
                  DOIDCommonName)
human_ReDU_DOIDCommonName$DOIDCommonName_counts <-
  as.numeric(human_ReDU_DOIDCommonName$DOIDCommonName_counts)

human_ReDU_HealthStatus <- df_redu_humans |>
  dplyr::count(HealthStatus) |>
  dplyr::rename(HealthStatus_counts = n, HealthStatus = HealthStatus)
human_ReDU_HealthStatus$HealthStatus_counts <-
  as.numeric(human_ReDU_HealthStatus$HealthStatus_counts)

human_ReDU_BiologicalSex <- df_redu_humans |>
  dplyr::count(BiologicalSex) |>
  dplyr::rename(BiologicalSex_counts = n, BiologicalSex = BiologicalSex)
human_ReDU_BiologicalSex$BiologicalSex_counts <-
  as.numeric(human_ReDU_BiologicalSex$BiologicalSex_counts)

```

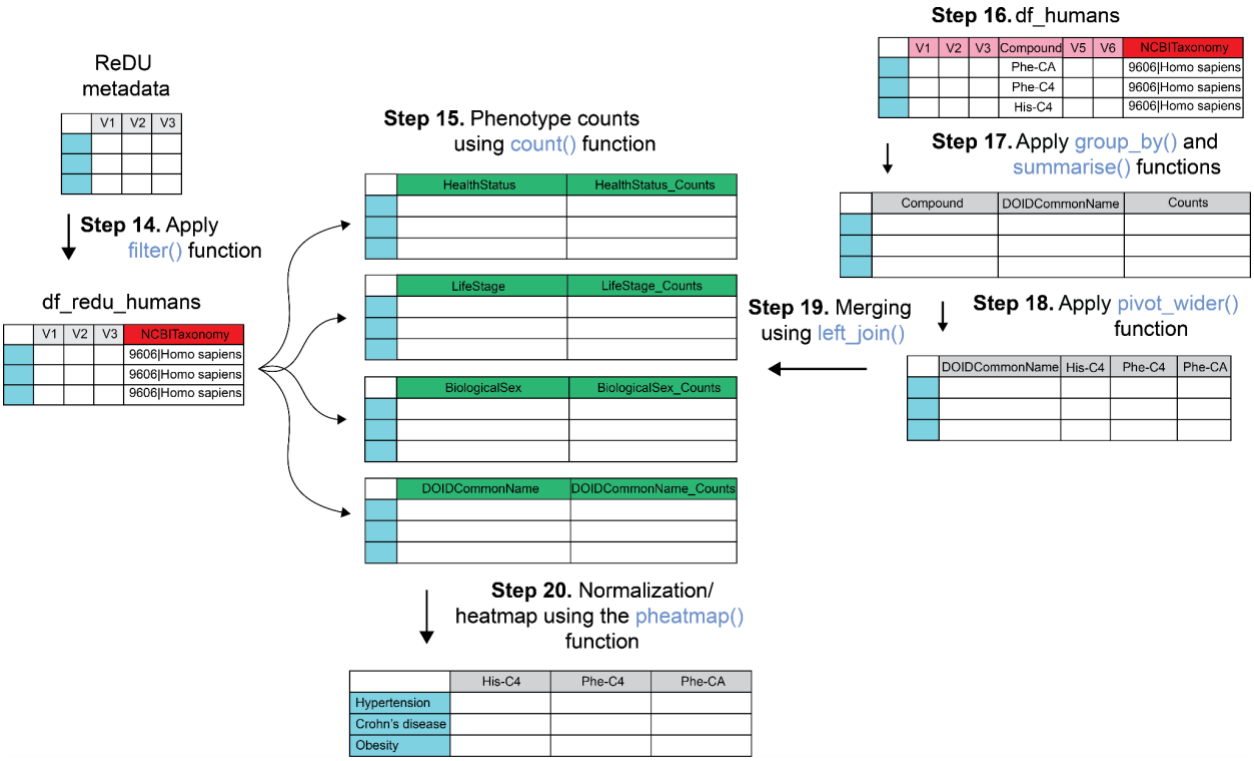


Fig. 7 | Linking MASST results to health phenotype. Steps involved filtering ReDU metadata to only keep human information (Step 14). Further filtering is based on different phenotype information before merging MASST results (Steps 15-19). The table is then normalized based on ReDU information (Step 20).

28. Normalization of fast search results with ReDU metadata. In the previous steps, we showed how many times a spectrum was found in the public domain for a specific UBERON body part or a health phenotype association. However, a normalization of the data can be performed if we want to inspect if there is an association with a particular disease. This is because there is, for instance, a lot of data relative to inflammatory bowel disease (IBD), and if we see more matches to IBD, it does not mean that it is necessarily associated with this particular disease. Therefore, it should take into account all the data available in ReDU for all the different diseases and normalize our MASST results based on this number to potentially help with the interpretation.

- Use the df_humans **(Fig. 7 – Step 16)** that was created at point 23 and perform grouping and summarizing on the data table based on disease ontology (DOIDCommonName) **(Fig. 7 – Step 17)**.

```
grouped_df_humans <- df_humans |>
  group_by(Compound, DOIDCommonName) |>
```

```
    summarise(Count = n()) |>
    ungroup()
```

- Convert a long data table format to a wide format **(Fig. 7 – Step 18)**.

```
grouped_df_humans_pivot_table <- grouped_df_humans |>
  pivot_wider(names_from = Compound, values_from = Count, values_fill = list(Count = 0))
```

- Merging the wide format of the fast search data table with the diseases-subsetted ReDU metadata **(Fig. 7 – Step 19)**.

```
merged_DOID_humans <- left_join(grouped_df_humans_pivot_table,
    human_ReDU_DOIDCommonName, by = "DOIDCommonName")
merged_DOID_humans$DOIDCommonName <- gsub("Crohn's disease", "crohn's disease", merged_DOID_humans$DOIDCommonName)
```

- The compound name columns are targeted for normalization. Calculate the sum, column-wise, to normalize across all diseases by adding the sum into the normalized df **(Fig. 7 – Step 20)**.

```
columns_to_normalize <- setdiff(names(merged_DOID_humans),
    c("DOIDCommonName", "DOIDCommonName_counts"))

normalized_merged_DOID_humans <- merged_DOID_humans |>
    dplyr::mutate(across(all_of(columns_to_normalize), ~ .x / .data$DOIDCommonName_counts)) |>
    dplyr::select(-DOIDCommonName_counts)

sums <- colSums(dplyr::select(normalized_merged_DOID_humans, where(is.numeric)),
    na.rm = TRUE)
    sums_df <- as.data.frame(t(sums))
    sums_df$DOIDCommonName <- 'Sum'
  sums_df <- sums_df[, names(normalized_merged_DOID_humans)]
  merged_sum_humans_DOID <- bind_rows(normalized_merged_DOID_humans,
    sums_df)
  merged_sum_humans_DOID <- merged_sum_humans_DOID |>
  dplyr::filter(!is.na(DOIDCommonName)) |> dplyr::mutate(across(where(is.numeric),
    ~replace_na(.x, 0)))
```

- Divide each numerical value by the sum and multiply by 100 to get the percentage.

```
merged_sum_humans_DOID_percentage <- merged_sum_humans_DOID |>
    dplyr::mutate(across(all_of(columns_to_normalize), ~ .x / .x[n()] * 100))
```

- Remove the column 'sum' that was incorporated to normalize the data. Use the function arrange() for alphabetic ordering and column_to_rownames() **(Fig. 7 – Step 20)** to transfer column 'DOIDCommonName' at the row names to make it compatible with the pheatmap() function **(Fig. 8 – Step 21)**.

```
merged_sum_humans_DOID_percentage <- merged_sum_humans_DOID_percentage
                |> dplyr::filter(DOIDCommonName != "Sum") |>
                    dplyr::arrange(DOIDCommonName) |>
                tibble::column_to_rownames("DOIDCommonName")
```

- Create and export the heatmap showing the prevalence of the molecule of interest in human diseases **(Fig. 8)**.

```
Diseases_humans <- pheatmap(merged_sum_humans_DOID_percentage,
    color = gradient_colors,
    cluster_rows = FALSE,
    cluster_cols = TRUE,
    angle_col = 90,
    main = "Health phenotype association",
    fontsize = 10,
    cellwidth = 15,
    cellheight = 15,
    treeheight_row = 100,
    fontsize_row = 12,
    fontsize_col = 12,
    legend_fontsize = 10,
    border_color = NA)
Diseases_humans

ggsave("Diseases_humans.pdf", plot = Diseases_humans, width = 10, height = 10, dpi
                                        = 900)
```

merged_sum_humans_DOID_percentage_plot

	His-C4	Phe-C4	Phe-CA
rheumatoid arthritis			
crohn's disease			
obesity			

Step 21. Apply the pheatmap() function

```
pheatmap(merged_sum_humans_DOID_percentage_plot,
color = gradient_colors,
cluster_rows = FALSE,
cluster_cols = TRUE,
angle_col = 90,
main = "Health phenotype association",
fontsize = 10,
cellwidth = 15,
cellheight = 15,
treeheight_row = 100,
fontsize_row = 12,
fontsize_col = 12,
legend_fontsize = 10,
border_color = NA)
```

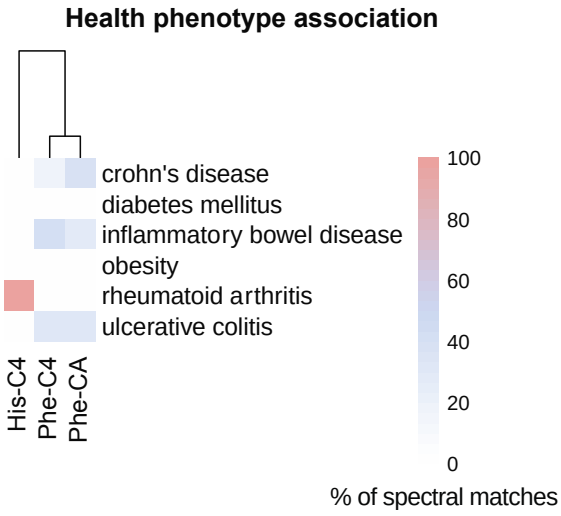


Fig. 8 | Heatmap showing health phenotype association of His-C4:0, Phe-CA, and Phe-C4:0 across public repositories with ReDU metadata. The merged_sum_humans_DOID_percentage df is used to generate the heatmap using the code provided in the grey box. A missing value indicates no information.

Anticipated results: Heatmaps are used to visualize body parts distribution and health phenotypes of the queried molecules. For **Fig. 6**, the number of matches is shown as a $\log_2$ scale and for **Fig. 8**, the percentage of spectral matches is presented.

Troubleshooting: If one is interested in investigating the molecule distribution at a specific skin location or plasma or serum, point 21 should be ignored as it combines all the different skin locations and blood parts. It is possible that the R version installed on the user's computer does not support the pipe operator (|>) used in this script. Alternatively, another pipe operator "%>%" from the *magrittr* package can be used and installed with dependencies. (Optional) install.packages(*magrittr*), then library(*magrittr*).

Validation of observed phenotype association

Timing: ~ days to years

Although many correlations can be found and hypotheses formulated through reverse metabolomics in terms of the discovery of new biochemistry and biology, some general guidelines need to be considered to support any newly formulated hypothesis. Although many other valid scientific strategies can be employed to provide support for a hypothesis, here we described three of them.

Prioritization of reverse metabolomics-based observation. Once an interesting observation is identified, users should proceed with going back to the original studies to which the observations match and perform feature extraction and statistical analysis. MZmine[40,41], MS-DIAL[42], XCMS[43], and OpenMS[44] are LC-MS data processing tools to extract features. Subsequently, using the extracted feature tables, statistical analysis can be conducted using platforms such as MetaboAnalyst[45], the GNPS based feature-based molecular networking stats guide[46], QIIME 2[47] or via custom scripts. For instance, in Gentry et al., based on more frequent detection of MS/MS spectra associated with IBD, we hypothesized that these newly-discovered microbial bile acids were found in higher levels in patients with Crohn's disease[1]. This hypothesis was confirmed by extracting the peak areas and then comparing the peak areas of these bile acids to non-IBD controls. They were found to be statistically significantly different, consistent with the hypothesis formulated based on MS/MS counts. Similarly, in Mohanty et al., we found that MS/MS of polyamines conjugated bile acids were more frequently detected in animals on carnivorous diets. This was confirmed using the extracted ion features, based on peak areas, of the new molecules from the original public dataset, which were statistically higher in carnivores compared to herbivores and omnivores[2]. When validating the results using reverse metabolomics, the users should consider the number of datasets in which the MS/MS matches occur as well as the sample size, to increase confidence in the biological discovery.

Retention time matching. One of the challenges is that when one searches the repositories with an MS/MS spectrum from a library, it can be very similar to other MS/MS spectra of compounds from the same molecular family. Therefore, isomeric compounds

can also match. To provide additional confidence that the samples contain the compound that you believe it contains, isomers can often be separated by chromatographic separation and/or ion mobility. The most straightforward thing one can do is to contact the original data depositors, as our lab has done for bile acids[2], to see if they have some of the samples still available. If samples are not available and they are not easily generated by the original depositors, one will have to find samples that most closely match to the samples of interest. Subsequent LC-MS/MS analyses can be conducted both with the original samples and a standard of the compound of interest that was originally queried against the repository. It is essential that the method of analysis is the same for all the samples, ideally under multiple chromatographic conditions and co-injection of the reference standard to confirm the identity of the compound[48,49]. The compound can be obtained from commercial sources, isolation from natural organisms, or by synthetic approaches. A match in both retention time and MS/MS spectrum between the standard and the compound in the sample confirms the annotation. Further, for the quantification of compounds of interest, users should adhere to the recommendations provided by the metabolomics Quality Assurance and Quality Control Consortium (mQACC) for analytical quality management. These measurements include, but are not limited to, analysis of QC samples such as reference standards, replicate extracted samples, pooled samples, and blanks[50].

Validation with additional cohorts. When an association is found with reverse metabolomics, for instance, between a metabolite and a particular disease, verifying if this association is also observed in different cohorts will significantly strengthen the conclusions. Therefore, when doing reverse metabolomics and to provide additional support for a hypothesis generated by reverse metabolomics, one has to find a way to get additional experimental data on the same or related cohorts. It will further allow for assessment from a molecular family association to a specific compound association. For example, in Gentry et al. we found that the di and tri-hydroxylated bile acid amidate molecular families were associated with IBD data in the public repositories[1]. We then contacted another research group that had recently published an IBD cohort, requesting their collaboration to verify our standards. Given that we now had retention time matches, we were able to accurately identify the specific bile acids that were amidated. Alternatively, had the samples been sent to our lab, we could have confirmed ourselves. This collaborative process not only strengthens support of the association hypothesis but also enhances the generalizability and reliability of our research findings across different populations.

Isomers confirmation. In untargeted mass spectrometry analysis, distinguishing between regio and stereoisomers is challenging due to their nearly identical mass spectra. For instance, compounds like deoxycholic acid and chenodeoxycholic acid

exhibit very similar fragmentation patterns. While future solutions are anticipated, current MASST implementations do not reliably differentiate between them. In such cases, these compounds should be annotated as belonging to molecular families (e.g., all matches for deoxycholate are categorized under dihydroxylated bile acids) to acknowledge the limitations of MS/MS spectral alignments. To conclusively identify the specific regio or stereoisomer in samples, retention time matching with standards is recommended. Alternatively, ion mobility-based mass spectrometry could potentially overcome these limitations, provided there are samples that are accessible for analysis.

Expanded reverse metabolomics – searching domain-specific MASSTs.
Timing: ~ 5-30 min

There is also often metadata organization that is not readily captured with ReDU metadata structure and therefore would be harder to visualize. This includes ontologies associated with microbes, plants, food, and other structured metadata. To aid the interpretation of MS/MS data from the repository, community curation initiatives have led to the development of microbeMASST[12], plantMASST[13], and foodMASST[11]. These three domain-specific MASSTs are integrated into MASST which aims to reveal the origins of a molecule by querying MS/MS spectra against domain-specific datasets. Other domain-specific MASST searches are being created in the future. Using domain-specific MASST searches, one can further provide insights into the biological context of the queried MS/MS spectrum (insight into potential microbial producers, dietary lifestyles, plant-derived metabolites) helping users to formulate hypotheses and design experiments for validation, thereby enriching the reverse metabolomics workflow. Therefore, if one is interested in knowing if the molecules of interest (for example, Phe-CA) are microbial-, animal-, or food-derived, these domain-specific MASSTs can be utilized **(Fig. 9)**. The USIs or the manual entry of the MS/MS spectrum of the molecules of interest are used to search against all three domain-specific MASSTs (**Fig. 9a**). The results are displayed as a tree, indicating that Phe-CA is found in food from animal sources and microbes (**Fig. 9b**). Further analysis can be achieved by clicking on the nodes.

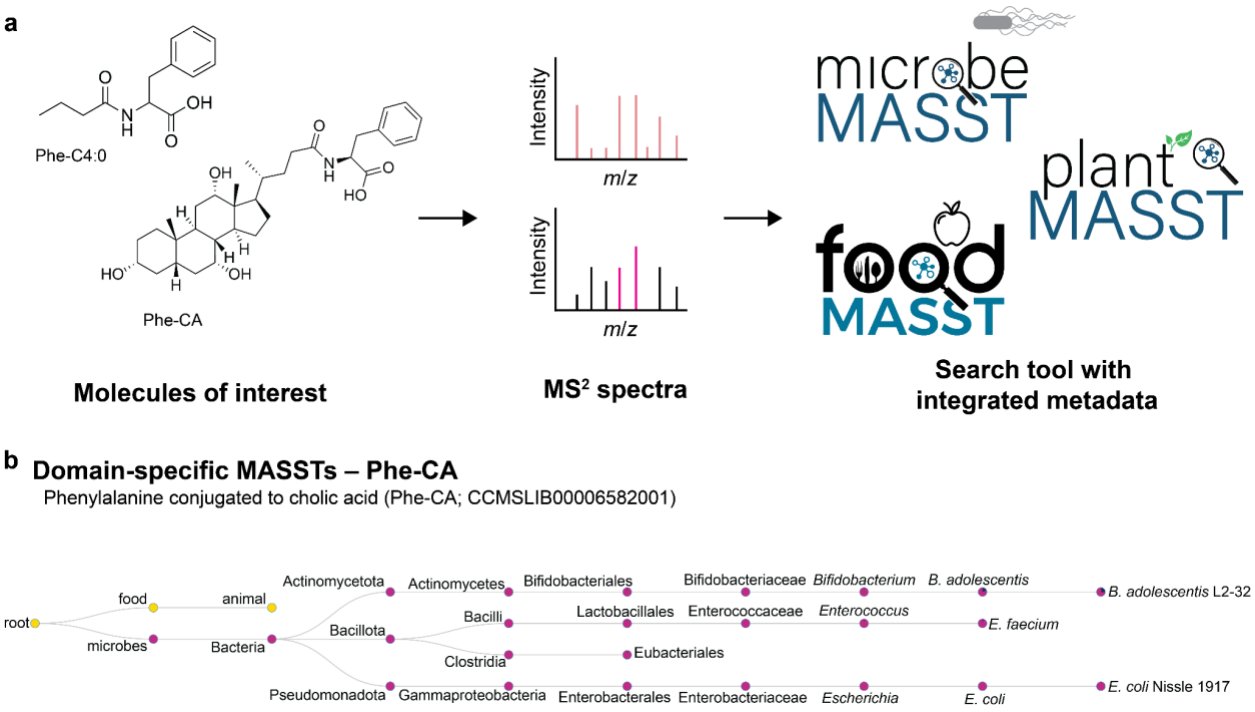


Fig. 9 | Example of domain-specific MASSTs searches of Phe-CA. a, Path undertaken from structurally-known or unknown molecules to launch domain-specific MASSTs with embedded metadata information. **b,** MASST search outputs showing that Phe-CA has been detected in animal and bacterial monocultures (using minimum match ions set to 3). Nodes in the tree display the proportion of MS/MS found against the reference database.

Anticipated results: The domain-specific MASSTs interface displays query results in interactive trees which can be downloaded as HTML files. Each node in the tree has embedded information that is domain-specific and includes the number of matched samples, the total number of available samples, and the frequency of occurrence at that taxonomic level.

Troubleshooting: Matches between the queried MS/MS spectrum and the reference spectrum depend on the availability of the data in the databases. For instance, if a molecule has no matches on microbeMASST, that does not mean that the molecule is not microbial-derived; rather it may indicate that it has not been detected in the data currently available in data from microbial cultures that are part of microbeMASST. In other words, it is possible to get a biological interpretation of what we retrieved as a match in our search, but not about the absence of a match. As more data are being deposited in repositories, more results are expected.

Limitations of reverse metabolomics

Reverse metabolomics uses searches to gather similar or identical MS/MS spectra found in the public domain. This approach is constrained by the type of MS/MS spectra available in public repositories. Notably, more data collected in positive mode ionization are found in public repositories compared to negative mode ionization. When an MS/MS spectrum in both ionization modes exists, it is advisable to select the positive ionization MS/MS spectrum. If the MS/MS spectrum of the molecule of interest does not exist in the GNPS library, it is recommended to collect and deposit the MS/MS spectrum acquired in both positive and negative ionization mode. Uploading the MS/MS spectrum to GNPS will allow one to create an USI which can then be used in reverse metabolomics.

While GNPS does support gas-chromatography mass spectrometry (GC-MS) for library searches and molecular networking[51], the indexing of such data has not been performed as it requires a deconvolution step of all raw data prior to indexing, and deconvoluted spectra would need to be used as input. Therefore, reverse metabolomics is not possible with GC-MS datasets in the GNPS/MassIVE ecosystem at this time. As a complement, BinBase[52] offers a GC-MS-based metabolome database to match spectra and retrieve biological metadata for thermostable small molecules of <650 Da. LC-Binbase is also being created, and it has the potential to be used in reverse metabolomics.

Reverse metabolomics leverages fast search to query each MS/MS spectrum to find structurally related molecules in public mass spectrometry repositories. These queries can be done much faster compared to the original implementation of MASST which relied on a large molecular network. It is now faster due to the pre-filtering and pre-indexing of all the spectra available in the public domain akin to the way Google indexes text. However, additional evaluation of spectral matches is important since spectral matches will be relative to filtered spectra (**Box 2**).

In the majority of cases isomers give rise to nearly identical MS/MS spectra, thus at the level of MASST searches, isomers can often not be distinguished. This will have to be resolved by follow-up experiments using standards of the isomers and extracts of samples. Sometimes, the MS/MS data itself has unique ions and ratios of ions to be able to differentiate post-MASST searches. However, when interpreting reverse metabolomics results, it is important to consider if there are other stereo or regio isomers that are merged in the results and interpret the results accordingly. For example, the Phe-CA, while this stands for Phe cholic acid amidate, should be reported as Phe trihydroxy bile acid amidate as other related isomers to cholic acid have similar MS/MS spectra.

Another important consideration is based on the number of ions that one searches, which are common to affect any MS/MS based batching approaches, irrespective of algorithm and resource that is used, but also hold true for MASST searches. The more

ions that are required to match, the more restrictive the search will be and effectively also reduces the number of matches. Choosing the appropriate settings is always a balance and trade-off between obtaining a larger number of matches vs false discoveries. Based on FDR estimations[53], the maximum number of correct matches are obtained when the minimum number of ions are set to 4 or 5. Higher number of ions gives rise to fewer matches but can be adjusted by allowing a lower cosine score and still obtain the similar FDR's (e.g., towards the cosine match score of 0.5). On the other hand if one lowers the number of ions to match, one has to increase the score threshold (e.g., to cosine of 0.9 or higher). It is discouraged to use fewer than 3 ions in the search as it is essentially impossible to get the FDR in acceptable range even when the score is raised to 0.95 or higher.

Finally, this workflow is contingent on the diversity of studies deposited in repositories together with parsable data science ready metadata. For example, many available datasets do not capture all types of diseases (cancer, infectious), interventions (e.g., antibiotic treatments, probiotics, surgical procedures and recovery process, fecal microbiota transplantation), nutritional state, and resilience factors. This information may be in papers but not yet readily accessible. If we can leverage this data to make discoveries, it will accelerate the full potential of reverse metabolomics type strategies. Therefore, we would advocate for the community to make their data publicly accessible, using controlled vocabularies and ontology metadata where possible, as it will accelerate downstream discoveries. We further envision that community curation efforts, such as was done for the domain-specific MASSTs, large language models and parsing scripts will help to further enhance the metadata associated with public data for data science applications such as reverse metabolomics. This is only the beginning of metabolomics evolution into a Big Data scientific discipline, there are many creative uses that remain to be explored and we expect that the concept of reverse metabolomics will play a key role in showing the value of all the effort put in by scientists all over the world that make their data public.

Data availability: The data used in this protocol are publicly available on GitHub (https://github.com/VCLamoureux/reverse-metabolomics) and already present in GNPS library (https://library.gnps2.org/).

Code availability: The code used for the reverse metabolomics workflow can be accessed on GitHub (https://github.com/VCLamoureux/reverse-metabolomics).

References

1. Gentry, E. C. *et al.* Reverse metabolomics for the discovery of chemical structures from humans. *Nature* **626**, 419–426 (2024).
2. Mohanty, I. *et al.* The underappreciated diversity of bile acid modifications. *Cell* **187**, 1801-1818.e20 (2024).
3. Haug, K. *et al.* MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Research* **48**, D440–D444 (2020).
4. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research* **44**, D463–D470 (2016).
5. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828–837 (2016).
6. Akiyama, K. *et al.* PRIMe: A web site that assembles tools for metabolomics and transcriptomics. *In silico biology* **8**, 339–45 (2008).
7. Lee, B. *et al.* Introduction of the Korea BioData Station (K-BDS) for sharing biological data. *Genomics Inform* **21**, (2023).
8. Quinn, R. A. *et al.* Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
9. Mohanty, I. *et al.* The changing metabolic landscape of bile acids – keys to metabolism and immune regulation. *Nat Rev Gastroenterol Hepatol* 1–24 (2024) doi:10.1038/s41575-024-00914-3.
10. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat Biotechnol* **38**, 23–26 (2020).
11. West, K. A., Schmid, R., Gauglitz, J. M., Wang, M. & Dorrestein, P. C. foodMASST a mass spectrometry search tool for foods and beverages. *npj Sci Food* **6**, 22 (2022).

12. Zuffa, S. *et al.* microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat Microbiol* **9**, 336–345 (2024).
13. Gomes, P. W. P. *et al.* plantMASST - Community-driven chemotaxonomic digitization of plants. 2024.05.13.593988 Preprint at https://doi.org/10.1101/2024.05.13.593988 (2024).
14. Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat Methods* **17**, 901–904 (2020).
15. Martens, L. *et al.* mzML—a Community Standard for Mass Spectrometry Data *. *Molecular & Cellular Proteomics* **10**, (2011).
16. Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J. Proteome Res.* **19**, 537–542 (2020).
17. Bittremieux, W. *et al.* Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. 2020.05.09.086066 Preprint at https://doi.org/10.1101/2020.05.09.086066 (2020).
18. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat Methods* **18**, 768–770 (2021).
19. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
20. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **45**, 703–714 (2010).
21. European Organization For Nuclear Research & OpenAIRE. Zenodo. (2013) doi:10.25495/7GXK-RD71.
22. Abiead, Y. E. *et al.* Enabling pan-repository reanalysis for big data science of public metabolomics data. Preprint at https://doi.org/10.26434/chemrxiv-2024-jt46s (2024).
23. Kang, J., Xu, W., Bittremieux, W., Moshiri, N. & Rosing, T. Accelerating open modification spectral library searching on tensor core in high-dimensional space. *Bioinformatics* **39**, btad404 (2023).

24. J. Kang, B. Khaleghi, T. Rosing, & Y. Kim. OpenHD: A GPU-Powered Framework for Hyperdimensional Computing. *IEEE Transactions on Computers* **71**, 2753–2765 (2022).

25. Li, Y. & Fiehn, O. Flash entropy search to query all mass spectral libraries in real time. *Nat Methods* **20**, 1475–1478 (2023).

26. Mongia, M. *et al.* Fast mass spectrometry search and clustering of untargeted metabolomics data. *Nat Biotechnol* 1–6 (2024) doi:10.1038/s41587-023-01985-4.

27. Batsoyol, N., Pullman, B., Wang, M., Bandeira, N. & Swanson, S. P-massive: a real-time search engine for a multi-terabyte mass spectrometry database. in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* 1–15 (IEEE Press, Dallas, Texas, 2022).

28. Schmid, R. *et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* **12**, 3832 (2021).

29. Liebisch, G. *et al.* Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. *Journal of Lipid Research* **61**, 1539–1555 (2020).

30. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).

31. DeFelice, B. C. *et al.* Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing. *Anal. Chem.* **89**, 3250–3255 (2017).

32. Uppal, K., Walker, D. I. & Jones, D. P. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal. Chem.* **89**, 1063–1067 (2017).

33. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).

34. Senan, O. *et al.* CliqueMS: a computational tool for annotating in-source metabolite ions

from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* **35**, 4089–4097 (2019).

35. Bittremieux, W. *et al.* Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Discovery of Structurally Related Molecules. *Journal of the American Society for Mass Spectrometry* (2022) doi:10.1021/jasms.2c00153.

36. Barrett, T. *et al. Data.Table: Extension of `data.Frame`*. (2024).

37. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

38. Kolde, R. pheatmap: Pretty Heatmaps. (2019).

39. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biology* **13**, R5 (2012).

40. Heuckeroth, S. *et al.* Reproducible mass spectrometry data processing and compound annotation in MZmine 3. *Nature Protocols* (2024) doi:10.1038/s41596-024-00996-y.

41. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology* **41**, 447–449 (2023).

42. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **12**, 523–526 (2015).

43. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779–787 (2006).

44. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **13**, 741–748 (2016).

45. Pang, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research* **52**, W398–W406 (2024).

46. Shah, A. K. P. *et al.* The Hitchhiker's Guide to Statistical Analysis of Feature-based Molecular Networks from Non-Targeted Metabolomics Data. Preprint at

https://doi.org/10.26434/chemrxiv-2023-wwbt0 (2023).

47. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**, 852–857 (2019).

48. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211–221 (2007).

49. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).

50. Kirwan, J. A. *et al.* Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. *Metabolomics* **18**, 70 (2022).

51. Aksenov, A. A. *et al.* Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat Biotechnol* **39**, 169–173 (2021).

52. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* **15**, 53–56 (2018).

53. Scheubert, K. *et al.* Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun* **8**, 1494 (2017).

54. Jarmusch, A. K. *et al.* A Universal Language for Finding Mass Spectrometry Data Patterns. 2022.08.06.503000 Preprint at https://doi.org/10.1101/2022.08.06.503000 (2022).

55. Ara, T. *et al.* DDBJ update in 2023: the MetaboBank for metabolomics data and associated metadata. *Nucleic Acids Research* **52**, D67–D71 (2024).

56. Wang, F. *et al.* CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **93**, 11692–11700 (2021).

57. Hong, Y. *et al.* 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* **39**, btad354 (2023).

58. Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Cent. Sci.* **5**, 700–708 (2019).

59. Young, A., Röst, H. & Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell* **6**, 404–416 (2024).

60. Young, A. *et al.* FraGNNet: A Deep Probabilistic Model for Mass Spectrum Prediction. Preprint at https://doi.org/10.48550/arXiv.2404.02360 (2024).

Acknowledgements: V.C.L is supported by Fonds de recherche du Québec - Santé (FRQS) Postdoctoral fellowship (335368). This is supported, in part, by NIH for the NIH collaborative microbial metabolite center U24DK133658; harmonization of metabolomics metadata across repositories R03OD034493; and Alzheimer's gut microbiome project U19AG063744 and BBSRC-NSF award 2152526. A.M.C.-R. and P.C.D. were supported by the Gordon and Betty Moore Foundation, GBMF12120. S.L was supported by the Research Council of Finland and the InFLAMES Flagship Programme of the Research Council of Finland (decision number: 337530). MW is supported by NIH 5U24DK133658-02 and was partially supported by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Disclosures: PCD is an advisor and holds equity in Cybele, BileOmix and Sirenas and a Scientific co-founder, advisor and holds equity to Ometa, Enveda, and Arome with prior approval by UC-San Diego. PCD also consulted for DSM animal health in 2023. MW is a co-founder of Ometa Labs LLC.

Contributions: V.C.L., M.W., P.C.D. conceived the study. V.C.L., S.L., H.M.R., S.X., P.C.D. wrote the manuscript. All authors reviewed and edited the article. All authors have tested and provided feedback on improving the protocol. V.C.L., P.C.D. generated the figures.

Corresponding author: Pieter C. Dorrestein (pdorrestein@health.ucsd.edu)

Key publications using this protocol

Gentry, E. C. *et al.* Reverse metabolomics for the discovery of chemical structures from humans. *Nature* **626**, 419–426 (2024).

Mohanty, I. *et al.* The underappreciated diversity of bile acid modifications. *Cell* **187**, 1801-1818.e20 (2024).

Box 1 – Description of the Mass Spec Query Language (MassQL)

MassQL is a universal query language designed to search, filter data patterns for downstream analysis[54]. It enables filtering MS data based on five mass spectrometry patterns: 1) precursor ions, 2) fragment ions, 3) the mass difference between two fragment ions, 4) the retention time, when chromatography is used and 5) drift time - if ion mobility is employed. The potential for defining patterns in MassQL is vast, including those that take the form of equations. MassQL enables non-computer scientists to find molecules coordinated to specific metal ions, discover drug-associated metabolites, molecules with a particular isotopic pattern, and microbial-derived molecules. Typically, these queries are created by computational experts who are very familiar with the details of the mass spectrometry data under investigation. Once formulated, a query can be applied for other studies or future re-analysis. Queries can target reference MS/MS libraries, single files, complete data sets, or entire repositories. As indicated, MassQL can be used for data filtering, in the context of reverse metabolomics it can be used to find MS/MS of interest that can then be queried using MASST.

An interactive interface is available to guide the user in writing, interpreting, and performing queries with MassQL and includes a query visualization and translation into nine languages to enhance accessibility (MassQL Sandbox). The MassQL compendium includes dozens of example queries and terminology to search patterns in mass spectrometry data for various classes of molecule (https://massql.gnps2.org/compendium/).

For instance, as we had recently demonstrated[2], if one is interested in finding all MS/MS spectra of conjugated trihydroxylated bile acids, a query can be designed based on the MS/MS spectrum. There are two diagnostic fragment ions at *m/z* 337.25 and *m/z* 319.24. Moreover, this query returns an MS/MS peak with a precursor *m/z* of X, and finds a MS/MS peak at X-390.277 with a tolerance of 0.01 *m/z* and a minimum relative intensity of the base peak at 5% - which is relative to the modification in the carboxylate. Thus, a MassQL query is developed to retrieve all the MS/MS spectra as USI's. These MS/MS can then be used to carry out reverse metabolomics to uncover biological associations. A representative query can be found below.

```
QUERY                    scaninfo(MS2DATA)                    WHERE
MS2PROD=337.25                                                  AND
MS2PROD=319.24                                                  AND
MS2PREC=X and MS2PROD=X-390.277:TOLERANCEMZ=0.01:INTENSITYPERCENT=5
```

Box 2 – How to evaluate MS/MS spectra

The quality of a MS/MS spectrum can influence the outcome of reverse metabolomics as FASST searches rely on filtered spectra for spectral matching. The FASST search tool retrieves identical or similar MS/MS spectra found in GNPS/MassIVE repository based on user-defined parameters. In assessing an MS/MS spectrum for FASST, the users should consider the number of peaks and their intensities. An example of a good MS/MS spectrum is provided in **Fig. Box 2a**. By selecting a noisy MS/MS spectrum with few ions, users should be aware that more results from the fast search will be retrieved, hence increasing the false discovery rate[53] (**Fig. Box 2b**). A noisy MS/MS spectrum containing only low and similar intensity peaks should be avoided (**Fig. Box 2c**).

Mirror plot for manual inspection of matching MS/MS spectra

In mass spectrometry, a mirror plot (also known as butterfly plot) can be used to visualize matching spectra (e.g., queried vs. reference spectrum). The results from the fast search can be manually inspected using a mirror plot and the users should consider the cosine score, the number of matched peaks and their intensities (**Fig. Box 2d**). We recommend the users to avoid relying exclusively on the cosine score to evaluate good matches as low number of matched peaks can lead to high cosine score but might increase the false discovery rate of the fast search[53] (**Fig. Box 2e**). Understanding how a MS/MS spectrum has been collected can improve the confidence in the FASST search results. For instance, if a MS/MS spectrum was collected with a low collisional energy, some fragments ions might not be detected (**Fig. Box 2f**).

FASST relies on spectral matching between the queried and the filtered spectrum found in repositories. FASST queries are faster because of the binning and indexing of all the spectra available in the public domain. However, additional inspection of the results, in particular looking at the raw spectrum is highly encouraged since the matches displayed during FASST search will be relative to filtered spectra. **Fig. Box 2g** highlights the difference in cosine score when queried MS/MS spectrum is matched against filtered spectrum and against raw spectrum.

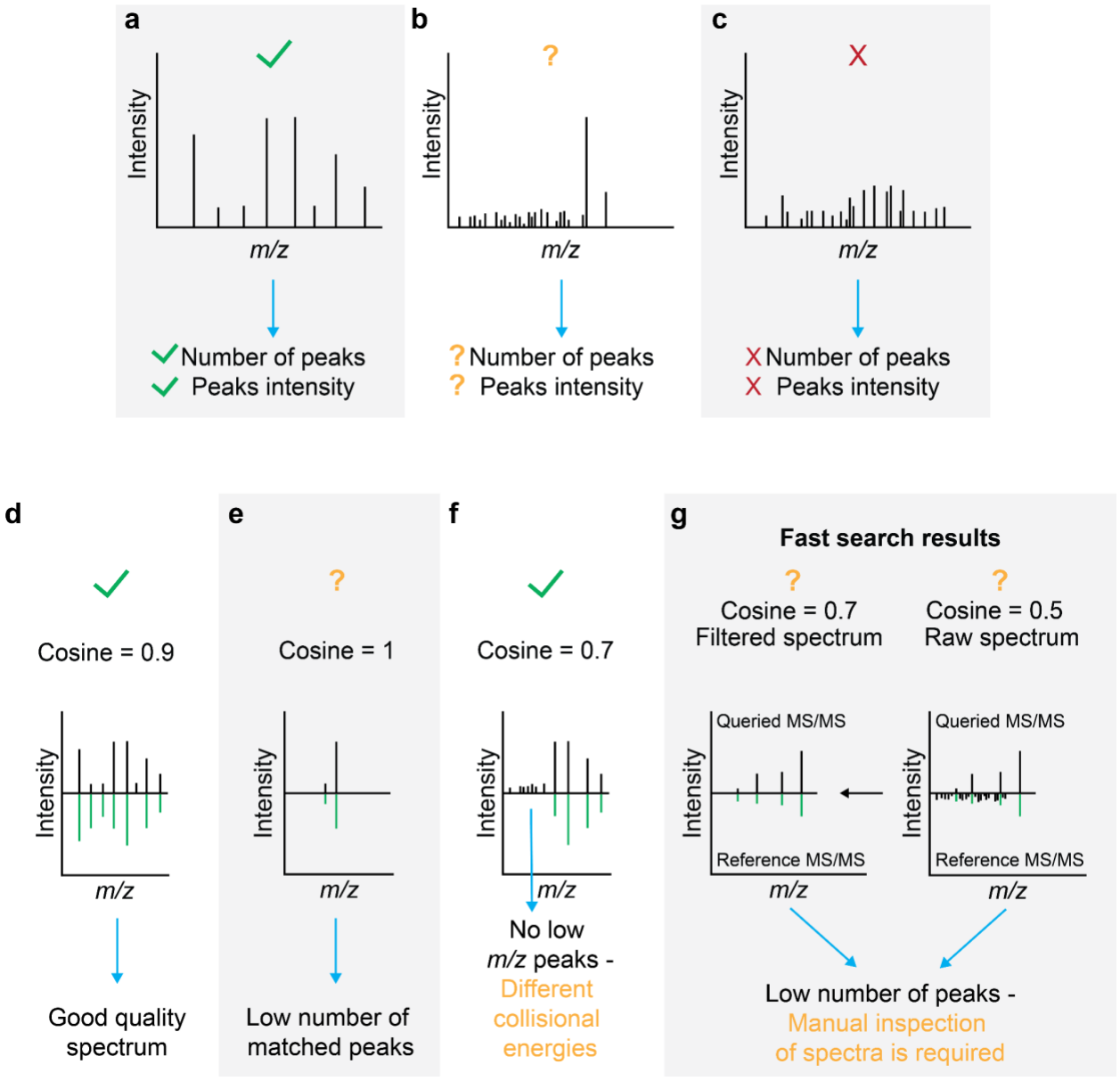


Fig. Box 2 | Visual examples of MS/MS spectra to evaluate quality and fast search results. a, Example of a good MS/MS spectrum. **b,** Example of a noisy MS/MS spectrum that required manual inspection of the results using mirror plots. **c,** MS/MS spectrum that should be avoided for the fast search. **d,** Example of mirror plots showing a high number of matching peaks and their intensities between the query and the reference spectrum. **e,** Example of a mirror plot denoting a high cosine score but low number of peaks. **f,** Example on how mass spectrometry parameters can influence a MS/MS spectrum due to different collision energies or via instrument settings where the lower *m/z* range of the MS/MS data is not collected but what does match between the two spectra has similar

intensities. **g,** Example showing queried MS/MS spectrum against filtered spectrum and against raw spectrum.

Box 3 – Sources of MS/MS spectra

Access to tandem (MS/MS) mass spectra is key in leveraging reverse metabolomics. There are many ways one can source the MS/MS spectrum to be used as input for reverse metabolomics. If a mass spectrometer is available and the researchers have acquired data, the fragments in the spectra can be entered manually (*m/z* values and matching intensities). Alternatively, the universal spectrum identifier (USI) can be used as an input. If the resource where the data is stored does not enable USI creation, linking data to GNPS/MassIVE[5] from other repository including MetaboLights[3], and Metabolomics Workbench[4], MetaboBank[55], MassBank[20] and others[6,7] will enable the generation of USI for the users. When it is not clear how to create an USI, reach out to the developers of those resources. The documentation on how to upload data in GNPS/MassIVE can be found here https://ccms-ucsd.github.io/GNPSDocumentation/datasets/. *In silico* MS/MS prediction (e.g., CFM-ID[56], 3DMolMS[57], NEIMS[58], MassFormer[59], and FraGNNet[60]) is another way of defining the MS/MS spectra used in searches. However, these tools, while useful, still have practical limitations for spectral predictions.

When using such tools for obtaining MS/MS spectra there is the need for expert evaluation of the results to prevent downstream interpretation errors. Additionally, a vast collection of tandem mass spectra are also found within the literature as often part of supplementary information. In such cases one can take a ruler to estimate relative peak intensities for manual input into MASST searches. Finally, there are public and commercial MS/MS reference libraries and repositories and bioinformatic tools like MassQL that can provide the tandem mass spectra from metabolomics repositories that can be used as input for reverse metabolomics.