Metabolomics (2024) 20:41
https://doi.org/10.1007/s11306-024-02090-6

REVIEW ARTICLE

Perspective: use and reuse of NMR-based metabolomics data: what works and what remains challenging

Goncalo Jorge Gouveia[1,2] · Thomas Head[1,3] · Leo L. Cheng[1,4] · Chaevien S. Clendinen[1,5] · John R. Cort[1,6] · Xiuxia Du[1,7] · Arthur S. Edison[1,8] · Candace C. Fleischer[1,9] · Jeffrey Hoch[1,10] · Nathaniel Mercaldo[1,11] · Wimal Pathmasiri[1,12] · Daniel Raftery[1,13] · Tracey B. Schock[1,14] · Lloyd W. Sumner[1,15] · Panteleimon G. Takis[1,16,17] · Valérie Copié[1,18] · Hamid R. Eghbalnia[1,10] · Robert Powers[1,19]

Received: 20 October 2023 / Accepted: 12 January 2024 / Published online: 13 March 2024

Abstract

Background The National Cancer Institute issued a Request for Information (RFI; NOT-CA-23-007) in October 2022, soliciting input on using and reusing metabolomics data. This RFI aimed to gather input on best practices for metabolomics data storage, management, and use/reuse.

Aim of review The nuclear magnetic resonance (NMR) Interest Group within the Metabolomics Association of North America (MANA) prepared a set of recommendations regarding the deposition, archiving, use, and reuse of NMR-based and, to a lesser extent, mass spectrometry (MS)-based metabolomics datasets. These recommendations were built on the collective experiences of metabolomics researchers within MANA who are generating, handling, and analyzing diverse metabolomics datasets spanning experimental (sample handling and preparation, NMR/MS metabolomics data acquisition, processing, and spectral analyses) to computational (automation of spectral processing, univariate and multivariate statistical analysis, metabolite prediction and identification, multi-omics data integration, etc.) studies.

Key scientific concepts of review We provide a synopsis of our collective view regarding the use and reuse of metabolomics data and articulate several recommendations regarding best practices, which are aimed at encouraging researchers to strengthen efforts toward maximizing the utility of metabolomics data, multi-omics data integration, and enhancing the overall scientific impact of metabolomics studies.

Keywords NMR · Mass spectrometry · Metabolomics · Best practices · Use and reuse of metabolomics datasets

1 Introduction

Metabolomics is an inherently interdisciplinary science, leveraging multiple analytical chemistry technologies and requiring carefully crafted experimental designs, complex data processing and statistical analyses, and intricate data interpretation regarding biological and clinical significance. The large amount of data and metadata generated at each stage of an experimental metabolomics workflow are highly modular yet thoroughly interconnected, which can tremendously impact results, data interpretation, and reproducibility. Metabolomics data are recorded and tabulated in a variety of formats and follow different ontologies from diverse scientific fields, requiring broad and considerable expertise to evaluate their merits effectively and to utilize them meaningfully. Metabolomics Workbench (https://www.metabolomicsworkbench.org/), now known as the National Metabolomics Data Repository (NMDR)), funded by the National Institutes of Health's (NIH) Common Fund (Smirnov et al., 2021; Sud et al., 2016), is an extensive data repository that includes metadata and experimental metabolomics data from 2000 + studies. Tremendous progress has been made in building this repository since its inception in 2013. While the NMDR has significantly contributed to making metabolomics data accessible to the scientific community, technical issues remain that limit data reuse in this and other repositories. For example, the complexity of metabolomics experiments presents significant challenges for consistent data reporting, archiving, retrieval, and reuse. Community-led efforts such as coordination of standards in metabolomics (COSMOS, (Salek et al., 2015), metabolomics standards initiative (MSI, (Sansone et al.,

Extended author information available on the last page of the article

2007b)), metabolomics quality assurance and quality control consortium (mQACC, https://www.mqacc.org/, (Beger et al., 2019)), and the Metabolomics Association of North America (MANA, https://www.metabolomicsna.org/), among others, have sought to establish minimum reporting criteria to address some of the challenges faced by metabolomics data generators, depositors, and users. Nevertheless, comprehensive data architecture and robust, flexible experimental workflows are still lacking. There is also a dearth of templates and incentives for compliance with data formatting and deposition, and a deficit of reliable mechanisms to facilitate the ease of reuse and re-analysis of deposited metabolomics data. These gaps take on added significance considering the recent updates to NIH's data management policy that enables investigators to implement and maintain research data on in-house systems during the grant period. Investigator adherence to standardized practices and community-defined policies is difficult to measure, which further hinders progress.

Herein, we highlight current challenges facing the metabolomics community and provide recommendations to strengthen the field's adherence to Findable, Accessible, Interoperable, and Reusable (FAIR) data principles (Wilkinson et al., 2016). We advocate for minimum scientific reporting standards by placing an emphasis on "what should be reported" to ensure results can be reproduced and data reused reliably. A technology-neutral and format-agnostic approach upholds scientific rigor, while also establishing essential baseline requirements that remain relevant in light of the fast-paced development of novel technologies and computational tools. An ultimate adherence to these standards will enhance the value of metabolomics data and their potential for long term use and reuse.

2 Generation and deposition of metabolomics datasets

Members of the MANA NMR Interest group and the NMR community have extensive experience generating numerous NMR and MS metabolomics datasets. While some metabolomics datasets have been deposited in public repositories such as NMDR or MetaboLights (https://www.ebi.ac.uk/metabolights/), most metabolomics datasets remain within individual laboratories or research groups (Kale et al., 2016; Steinbeck et al., 2012). The final destination of data created by scientists outside of the metabolomics community is unclear.

The fact that metabolomics data repositories are a relatively new resource is a prime reason for the limited number of currently deposited datasets. Our recent survey of the 2020 scientific NMR metabolomics literature revealed that only 11% of published papers deposited their metabolomics datasets (Powers et al., 2024). Furthermore, several practical barriers to depositing metabolomics data into public repositories contribute to these low numbers. Depositing a metabolomics dataset into a repository tends to be time-consuming, cumbersome, and frustrating without any perceived benefit to the investigator. There may be an unwillingness or inability to deposit data due to the substantial time requirement, intellectual property concerns, institutional or Health Insurance Portability and Accountability Act (HIPAA) restrictions, or other regulations. Simply stated, strong community incentives to deposit data are missing since most journals and granting agencies do not require the deposition of reusable metabolomics data into publicly accessible repositories to complete the publication of a manuscript or to adhere to funding requirements.

2.1 Our recommendations

The data deposition process should be streamlined, uniform, and incentivized. For example, it would be extremely useful for the information and the process needed to deposit data in the NMDR repository to be more similar or the same as other data repositories (e.g., MetaboLights or BioMagResBank (BMRB) Metabolomics (Romero et al., 2020)), or the conversion of data formats between different repositories be automated and seamless. Deposition of complete and accurate metadata along with the raw and processed spectral data is equally important and burdensome, but will require alternative and creative solutions to ensure harmonization across repositories. For example, while standardization of protocols may not be a true reality due to the complexity of scientific questions and the diversity of instrumentation, it is still likely that consensus protocols will become agreed upon as the field of metabolomics continues to mature. Accordingly, data deposition may be easily and automatically cross-referenced with existing experimental parameter sets that have been validated. Overall, we encourage funding agencies and scientific journals to support requirements of reusable data deposition during funding reviews and before publications.

3 Use of deposited metabolomics datasets

Most metabolomics experts typically only have experience in a single metabolomics platform. For example, NMR experts often have limited experience with MS data, and vice versa. Therefore, the use and reuse of MS data by NMR experts (and vice versa) is often complex and requires specific expertise. Encouragingly, due to the complementary nature of NMR and MS metabolomics data, there is a small (but slowly growing) group of investigators with experience utilizing diverse analytical platforms, including both NMR

and MS (i.e., multi-platform metabolomics, (Jeppesen & Powers, 2023)).

It is becoming increasingly important to adopt a multi-platform approach that combines NMR and MS data to advance the value of metabolomics research. This requires provisions of tools and guidance to seamlessly integrate diverse types of acquired metabolomics data and training of researchers to utilize these resources effectively. Public data repositories are not currently configured to manage the complexity of a multi-platform metabolomics study. Instead, the focus tends to be directed toward a single analytical method. The growth towards multi-platform and multi-omics (e.g., metabolomics, lipidomics, proteomics) studies further exacerbate the situation.

3.1 Our recommendations

There is an urgent need for deposited metabolomics datasets to be annotated and accompanied by comprehensive metadata and instructions enabling non-experts to reuse the data and remain cognizant of the potential pitfalls and limitations of the study. It is also important for public repositories to provide technical support to help inexperienced data retrievers with data reuse and to provide detailed protocols on how best to use computational tools necessary for accurate re-analysis of deposited data. Finally, repositories need to be designed with a flexible architecture to easily accommodate multiple data types associated with a single study, including liquid chromatography (LC)-MS, gas chromatography (GC)-MS, capillary electrophoresis (CE)-MS, one-dimensional (1D) and two-dimensional (2D) NMR. This may necessitate the cross-referencing of datasets across two or more repositories dedicated to a single analytical method or restructuring existing repositories to be data-type agnostic. More challenging than simply linking disparate data types is the need to archive distinct sets of experimental and processing parameters associated with each data type. Here too, streamlining and automating the inclusion of metadata from published manuscripts and/or cross-referenced to protocol databases will be essential to ensure reproducibility and to enable the proper use and reuse of metabolomics datasets.

4 Reuse of metabolomics datasets currently deposited in public data repositories

There are currently substantial barriers to the reuse of metabolomics data that have been publicly deposited. This may be because the raw or interpreted data, the original experimental parameters, processing protocols, and/or relevant software details are often missing, not defined, poorly annotated, or unavailable. Also missing could be details describing the statistical analysis methods employed, the criteria for identifying statistical significance, details about experimental design and quality control (QC) strategies, and whether relative or absolute concentration changes were measured. Without the necessary metadata, for example, detailed experimental design information, experimental or processing parameters, or data analysis protocols, it is impossible to both replicate the original processing and analysis of the deposited data and to perform valid secondary assessments, data processing, and/or statistical analysis. Thus, the proper interpretation of existing analyses accessible in current repositories is limited.

4.1 Our recommendations

We suggest that metabolomics data repositories should work toward addressing the following issues:

(1) Request that all raw data in both vendor-specific proprietary format and open data formats (e.g., nmrML, mzML, mzXML, netCDF, mzTab) from each study be deposited;
(2) Ensure that all the raw data files in the archive match with samples listed on the repository's website and are clearly annotated with sample identification numbers and metadata that are part of the experimental design;
(3) Instrument vendors and/or data repositories should provide user-friendly tools to convert raw (e.g., LC–MS, GC–MS, or 1D, or 2D NMR) metabolomics data into open format data such as mzML, mzXML, CDF, nmrML;
(4) Request that information be provided that demonstrates the confidence in metabolite identification included in the study and that clearly provide the evidence for each metabolite assignments (Alseekh et al., 2021; Kirwan et al., 2022; Peter et al., 2021; Sumner et al., 2007);
(5) Include raw data and detailed information of in-house physical reference standards or publicly available compound library information that were used in the reported studies to identify or annotate metabolites;
(6) Develop and implement data quality processing pipelines (i.e., from raw data to statistical models and annotated metabolites) to provide benchmarks to assess the quality of the deposited data and analysis;
(7) Undertake extensive curation prior to making the deposited data publicly accessible to ensure complete and validated datasets;
(8) Develop powerful query and visualization tools to assess the quality of raw and processed NMR and MS metabolomics deposited datasets;
(9) Request raw data from QC samples such as blanks, pool samples, standard reference material samples etc.;

(10) Request point of addition, raw data, and concentration of internal standards;
(11) Request sample injection orders for large scale studies where hundreds or thousands of samples are involved; and
(12) Provide version information for software tools and a comprehensive list of user defined parameters that have been used to produce the results deposited in public repositories.

Data repositories will not be able to achieve these ambitious goals on their own. It is critical for data repositories to coordinate and formalize data deposition requirements as well as receive a strong commitment from the entire scientific community. Simply put, journals and funding agencies need to adopt policies requiring the deposition of reusable metabolomics data into publicly accessible repositories as part of the publication and grant reporting processes.

5 Sample and experimental metadata required for effective use and reuse of metabolomics data

A key component of impactful metabolomics studies is the inclusion of relevant metadata regarding the collected samples. For example, the quality of a clinical study is directly correlated with the quality of the collected and reported metabolite data. As metabolite levels can be affected by many factors, including, for human studies, gender, body mass index, age, alcohol status, etc. (Navarro et al., 2023; Tolstikov et al., 2020), this information is important to include for proper analysis. Sample treatment, including drug therapies or other procedures, as well as collection methods and procedures, also affect metabolite profiles and must be reported. This is true for cells, animals, humans as well as environmental samples. Fortunately, the metabolomics community has the opportunity to adopt tools developed by the proteomics field for reporting biological and technical metadata, and for connecting samples with corresponding metadata to yield an understandable dataset available for reanalysis and interpretation (Claeys et al., 2023; Deutsch et al., 2023).

Significant efforts have been invested in building consensus reporting standards for metabolomics and exposomics (non-targeted analysis) data (Alseekh et al., 2021; Kirwan et al., 2022; Peter et al., 2021; Sansone et al., 2007a; Sumner et al., 2007). Complete and thorough metadata is of equal importance to the raw and processed spectral datasets, which includes experimental design, sample and, when relevant, clinical, and demographic information, experimental and processing parameters, statistical models, and analysis criteria. It is impossible to assess the quality and usability of metabolomics datasets and to enable their reuse, (re-process and/or reanalyze) without access to detailed metadata.

5.1 Our recommendations

As mentioned, often incomplete datasets are deposited in metabolomics data repositories. In this context, we suggest the mandatory provision of parameters necessary for use and reuse in data uploads, including the raw data from QC and blank samples. Ideally, data would be provided in open formats, such as mzML, mzXML, or CDF, as this would maximize the use and reuse of the raw data. Depositing the data in vendor formatted files (Bruker, Agilent, JEOL, etc.) would be an acceptable alternative considering automated data conversion processes are routinely implemented in databases or repositories.

We also strongly recommend that carefully curated metadata (Table 1) be included as part of the data deposition process. Metadata should include:

(a) Accurate indexing of sample identification numbers to specific raw data and/or processing filenames;
(b) Employing standardized nomenclature or ontologies to facilitate the task of automation and reuse of data;
(c) Providing detailed information on experimental design and study factors that may impact the results or interpretations;
(d) Detailed and clear information on the number of experimental and control groups, number of biological replicates, number of technical replicates per group, number and type of control samples, as well as clinical and demographic information when relevant and appropriately available;
(e) Details on the type of instruments and automation used, including manufacturer, model, software version, spectrometer frequency, nucleus, NMR probe, acquisition parameters (including NMR pulse programs) type of mass analyzer, HPLC platforms, etc.;
(f) Detailed and extensive description of sample type, sample conditions, sample handling and preparation, as well as detailed description of QC strategies; and
(g) Detailed description of spectral processing and data matrix pre-processing parameters.

The deposition of the metadata listed in Table 1 should be sufficiently flexible to address the needs of both the data depositors and the data users (i.e., informaticians, etc.) while also being simple enough to facilitate the deposition process by avoiding a burdensome task. For example, data depositors may prefer to just upload a single standard text file consisting of the protocol, while data users may wish to avoid data transformations and

Table 1 Recommended Metadata

Metadata class	Metadata types
Study	
Study description	Species and purpose of study. Number and type of experimental groups, number of biological/analytical replicates per group, number and /type of controls
Specimen/sample	Tissues and biofluids (e.g., blood, urine, CSF, etc.). Cell lysates, homogenized tissues, food & beverages, plants, use of isotope labeling, etc
Selection criteria	Study parameters including inclusion/exclusion criteria, population, or group characteristics
Type	Cross sectional, cohort, case–control, retrospective, prospective
Condition/comorbidity	Relevant disease or condition, treatment, gender, age, BMI, etc., using standard language (e.g., ICD9/10)
Concomitant factors	Medications, exposure, compounds (with standard IDs: PubChem)
Other attributes	Study-relevant data. Examples: vitals in human studies, color in a urine sample, clarity in a cell culture
Pre-analytical	
Sample condition	Storage conditions, storage duration, temperature
Equipment/model	Bruker, Agilent, JEOL, Thermo Fisher, Sciex, LECO, Waters, etc
Identification/indexing	Associate sample IDs to specific raw data and processing identifiers
Analytical	
Sample preparation	Lysis or homogenization method (i.e., sonication, bead-beating, etc.), precipitation or filtering method (i.e., removal of biomolecules or debris), extraction method and solvent(s), sample reconstitution
Sample parameters	pH, buffer, solvent, temperature, chemical shift/mass internal reference, isotope labeled standards for quantitation
Validation/QC	Instrument calibration, validation approach, spiking, pooling, replication, QC sample types and frequency
Instrumentation	Manufacturer, model, software version, spectrometer frequency, nuclei, NMR probe type, mass analyzer
Data	Specification, format, auxiliary files
Workflow	SampleJet, automatic tune and match, liquid handlers, robotic systems
Analytical method	1D/2D NMR, LC–MS, GC–MS, CE-MS, FTIR, ionization type, NMR pulse sequence, NMR and MS data acquisition and experimental parameter values
LC/GC details	Column type, column dimensions, solvents, gradient-elution parameters, temperature, separation time
Post-analytical	
NMR spectral processing parameters	Baseline correction, phasing, normalization, scaling, window function, zero-filling, removal of spectral regions, alignment, and referencing
MS spectral processing	Software and version, peak picking, threshold values, filtering, etc
Data matrix feature pre-processing	Binning/bucketing, peak-picking/feature selection criteria (CV, %missing, fold-change), missing data imputation method
Statistical methods	Univariate/multivariate statistics, artificial intelligence, or deep learning methods
Statistical validation	Minimal fold change, reported p-value for significance, false-discovery rate, or multiple hypothesis correction, reported R^2/Q^2 values, proper validation methods reported for supervised statistical models
Software/platform	OS, RAM, CPU, software versions, data processing and analysis parameters or scripts, persistent link to in-house software/tools used for analyzing the data
Metabolite assignments	Procedures for determining metabolite assignments from spectral data, including metabolite harmonization approach (InChI or SMILES keys, KEGG or HMDB IDs, RefMet or other software, etc.)
Other	Links to published papers describing the deposited data, the experimental procedures, or protocols
Results table	Annotated metabolites, metabolic pathways, absolute/relative metabolite concentrations, statistical significance measures (i.e., estimates, confidence intervals, and p-values)

interpretations by using discrete parameters and ontologies that are also readily searchable. Ideally, a minimum reporting standard would define "what should be reported", and the data standard would specify the "how". The fast pace at which new algorithms, methods, software, and processing pipelines are being developed (i.e., peak picking, peak integration, peak deconvolution, AI integration, etc.) necessitates a continually evolving set of minimum criteria to avoid becoming outdated and to enable reliable interconversion of metadata between repositories.

6 Uniform data processing pipeline—a prerequisite for use and reuse of metabolomics data

As mentioned above, complete sets of raw data and metadata are an absolute prerequisite for the effective use and reuse of metabolomics data. In addition to including the experimental design parameters discussed above, it is also critical to provide detailed and complete information about the data pre-processing, statistical methods, and model validation tools employed to identify metabolic differences between two or more categorical (disease status), experimental or treatment groups. Currently, numerous experimental protocols, statistical methods, and software are used by the metabolomics community. For example, our survey of papers from 2010 and 2020 (Powers et al., 2024) identified over 110 unique software packages that were being used to process or analyze NMR metabolomics data sets. This problem is amplified by the lack of publicly available benchmark datasets that makes development, assessment, optimization, and comparison of different steps or pipelines for data analysis difficult if not impossible.

6.1 Our recommendations

Metabolomics researchers and depositors of metabolomics data should be required to include a thorough description of the statistical methods and validation tests conducted in their study. This includes detailed information on the types of univariate and/or multivariate statistical methods used, whether artificial intelligence (AI) or deep learning methods were applied, and if so, what type. The metadata should include, when relevant, the minimal fold change, false discovery rate (FDR) corrected p-values, R^2/Q^2 values, and results from validation tests such as permutation tests, CV-ANOVA, area under the receiver operating characteristic curve (AUROC), and random forest analyses (Szymańska et al., 2012; Worley & Powers, 2013; Xi et al., 2014). We also recognize that not every individual lab or researcher has the statistical tools readily accessible to execute or reconstruct these analyses. It will be imperative to make software tools employed in each metabolomics study publicly available, as these are essential for the re-processing and re-analysis of the published data. Publicly available does not necessarily mean freely available as commercial software is likely to be used, but such software should still be accessible. The efforts of the NMRBox consortium (https://nmrbox.nmrhub.org/) to achieve this goal of archiving and distributing all versions of NMR-related software broadly used by the community are commendable (Maciejewski et al., 2017). In this regard, we encourage metabolomics researchers to collaborate with the NMRBox consortium and to contribute old and new software versions, data processing, and analysis scripts. Importantly, including data processing tools and pipelines within metabolomics data repositories could be a valuable step toward establishing standards and best practices for metabolomics studies.

7 Integration of metabolomics data with other types of omics data

A multi-omics approach includes any combination of genomics, transcriptomics, proteomics, metabolomics, and lipidomics datasets originating from multiple analytical platforms. Research projects employing multi-omics methods are slowly gaining popularity since, although challenging methodologically, they greatly enhance the information content for a given study and provide a comprehensive and more accurate view of the system. Multi-omics studies may become a major source of metabolomics datasets; however, due to the complex structures of diverse types of omics datasets, multi-omics data or links to multi-omics datasets are rarely available in current metabolomics data repositories.

7.1 Our recommendations

Creators of data repositories should consider including the deposition of multi-omics datasets or cross-references to linked data. For this to happen easily, it will be necessary to simplify the deposition of multi-omics datasets, which are quite different in types and structures. Another approach is to link multi-omics datasets across multiple data repositories in which each repository only accepts a single type of data. Of course, these approaches would necessitate multiple, separate data deposition activities that may increase the occurrence of errors or lead to missing metadata. It will also require cooperation between the administrators of each data depository to facilitate accurate data file linkage. One potential solution would be the coordinated development and adoption of a single metadata template shared by all omics depositories that would require a single entry of the metadata for depositing the raw and processed multi-omics dataset across multiple repositories. The UK biobank (https://www.ukbiobank.ac.uk/) and the Omics Discovery Index (https://www.omicsdi.org/) provide possible templates for how to structure public data repositories that accommodate multi-modal datasets (Bycroft et al., 2018; Perez-Riverol et al., 2017). In addition to cross-linking multi-omics datasets in repositories, it would be beneficial to include a set of quality controls and standard reference materials (SRM), which are scarce and underdeveloped except for National

Institute of Standards and Technology (NIST) SRM 1950 (human plasma sample with the quantitative analysis and certification of approximately 100 metabolites, (Phinney et al., 2013)) to facilitate the merging and translation of multi-omics datasets.

8 Standardization of metabolomics software and informatics tools

Both commercial and freely available software tools have been developed for processing MS- and NMR-based metabolomics data. Commercial software such as MNova (https://mestrelab.com/software/mnova/), Chenomx (https://www.chenomx.com/), Progenesis QI (Nonlinear Dynamics), and Compound Discoverer (Thermo Fisher), can be quite costly compared to freely available academic software. In addition, the inner workings of these commercial software packages are not typically transparent, making it difficult for users to reliably compare and evaluate results generated by different software tools designed to accomplish the same process. At least one commercial vendor (Bruker) now offers an AI-based identification and quantification model built entirely with proprietary databases and algorithms that has a fee structure and is completely opaque to subscribers. On the other hand, commercial software developed by professionals may be more stable and user-friendly compared to open-source software typically developed in academic laboratories. Academic software is not always well-maintained, can quickly become outdated, or may no longer function or be available due to a lack of resources or funding. This can result in a significant loss of knowledge and a tremendous loss of a return on investment to funding agencies.

8.1 Our recommendations

Funding agencies should strongly encourage the development and sharing of well-documented, open-source software tools for metabolomics. Software vendors should also consider the value of sharing proprietary software or at least to offer tools for open-source data conversion. Software vendors could play a critical role in this effort of software sharing. This will address the need to establish best practices in the field and a uniform data processing pipeline. In the meantime, additional funding is needed (1) to attract and retain talented professional-level software developers in academia to successfully compete against the higher salaries the technology industry offers and (2) for long-term maintenance and growth of software tools. Ideally, the resulting software tools would be made publicly available and hosted on open-access sites such as GitHub (Gilroy & Kaplan, 2019), NMRBox (Maciejewski et al., 2017), and others.

9 Barriers to harmonization across datasets and interoperability

Our collective experience suggests that the non-uniform organizational structure of data archives makes it incredibly challenging to automate harmonization across multiple data sets. The proliferation of redundant databases and depositories is an underlying source of this problem. Instead of constantly "re-inventing the wheel," granting agencies, administrators, and developers of databases and repositories need to streamline and uniformize deposition interfaces as much as possible. Our experience has been that uploading data to publicly accessible data repositories can be excessively time-consuming, burdensome, and may lead to unnecessary delays in the final acceptance of manuscripts in journals. This occurs simply because every database is unique in its interface and deposition process, and in what data and metadata is required. As a result, database errors are routine and caused by the human depositor due in part to the complexity of the deposition process. This issue is exacerbated for studies combining metabolomics, lipidomics, proteomics, transcriptomics, and/or genomics since the diverse types of omics data are stored across multiple data repositories, or worse, because some data components are not publicly available.

Harmonization of nomenclature (i.e., ontologies) plays a significant role in enabling cross-study comparisons by resolving small molecule identities through a standardized lexicon that recognizes and links molecular representations of isobars and isomers at different levels of structural resolution as well as uncertainty in stereochemistry or regiochemistry (Creek et al., 2014; Villalba et al., 2023). The complexity of factors involved has encouraged application-specific approaches that rely on enhancements to existing nomenclatures to fill the gap (Dashti et al., 2017; Fahy & Subramaniam, 2020; Heller et al., 2015; Weininger, 1988), while new computational approaches have demanded the creation of more novel representations (Wigh et al., 2022). In their absence, metabolite nomenclature is likely the largest barrier to the harmonization of datasets across multiple studies, laboratories, and data depositories. InChI ((Heller et al., 2015), IUPAC standard) and SMILES (Weininger, 1988) are commonly employed attempts to obtain a unique and discrete representation of a chemical structure, but these and other approaches have well-recognized limitations. These simple InChi and SMILES ASCII strings do not capture the three-dimensional structure, stereochemistry, and charge of the molecule. SMILES, which is proprietary, and InChi strings are also not unique and depend on the specific algorithm used by the software to generate these strings. Thus, it may be difficult or impossible to interchange metabolite

information across multiple sources if different software has been used to annotate metabolites. The nomenclature problem is exacerbated if the exact structure is not decipherable from the experimental data. Lower levels of structural resolution will likely lead to ambiguities when interconverting metadata between repositories. Metabolomics databases such as the Human Metabolome Database (HMDB, https://hmdb.ca/) and Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) have partially solved the metabolite nomenclature issue by assigning unique identifiers for each chemical entity in the database. Of course, database identifiers are typically manually assigned by investigators leading to similar concerns regarding accuracy and do not address variations in structural resolution as dictated by the experimental data.

9.1 Our recommendation

We encourage implementing and using a uniform structure for data archiving that would enable automated data processing from different studies. For example, the NIH provides sample repositories with clear submission guidelines for several data types, e.g., genomic data. These guidelines could also be extended to metabolomic data (https://sharing.nih.gov/data-management-and-sharing-policy). Another recommendation is to define and then encourage the use of common ontologies specific to metabolomics in a similar vein as developed by the Proteomics Standard Initiative (Deutsch et al., 2023). Such uniform practices are essential to enable the development of tools to improve metabolomics experiments and data depositions and the cross-linking of multi-omics studies. These capabilities will aid in ensuring compliance with data deposition, facilitate data reuse, and contribute to higher-quality metabolomics studies with increased significance for biology and enhancing our understanding of complex systems.

10 Overall conclusion

Establishing publicly available metabolomics data repositories has been a big step toward increasing the value of metabolomics data sets. Addressing challenges to improve the efficient reutilization, integration, and synchronization of metabolomics data will represent another notable stride toward maximizing the utility and reuse of metabolomics datasets.

While we presented herein several bottlenecks to metabolomics data use and reuse, none of them are insurmountable. With a strong commitment from the metabolomics community in partnership with funding agencies, vendors, and publishers of scientific manuscripts, we are confident that the metabolomics research field can improve the type and quality of data and analysis generated, and thus augment the value of metabolomics findings for the broader scientific community. Our overarching goal is to enhance the accuracy, transparency, reproducibility, and ease of use and reuse of metabolomics data. Website designers and developers of these data repositories are encouraged to seek input from their scientific communities, create user-friendly search engines, facilitate data browsing, and provide versatile and accurate search options. The primary goal of a repository interface should be to assist the investigator in identifying metabolomics studies of interest and to guide the relative ease of reuse and reanalysis of metabolomics data, especially for future metadata studies.

Author contributions All authors contributed to the overall construct and composition of the review. VC, HRE, and RP wrote the original draft of the manuscript. All authors read, revised, and approved the manuscript.

Data availability This article does not contain any metabolomics or metadata.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

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

References

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

Beger, R. D., Dunn, W. B., Bandukwala, A., Bethan, B., Broadhurst, D., Clish, C. B., Dasari, S., Derr, L., Evans, A., Fischer, S., Flynn, T., Hartung, T., Herrington, D., Higashi, R., Hsu, P. C., Jones, C., Kachman, M., Karuso, H., Kruppa, G., … Zanetti, K. A. (2019). Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics, 15*, 1–5. https://doi.org/10.1007/s11306-018-1453-6

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S.,

Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z

Claeys, T., Van Den Bossche, T., Perez-Riverol, Y., Gevaert, K., Vizcaíno, J. A., & Martens, L. (2023). lesSDRF is more: Maximizing the value of proteomics data through streamlined metadata annotation. *Nature Communications, 14*, 6743. https://doi.org/10.1038/s41467-023-42543-5

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

Dashti, H., Westler, W. M., Markley, J. L., & Eghbalnia, H. R. (2017). Unique identifiers for small molecules enable rigorous labeling of their atoms. *Scientific Data, 4*, 170073. https://doi.org/10.1038/sdata.2017.73

Deutsch, E. W., Vizcaíno, J. A., Jones, A. R., Binz, P.-A., Lam, H., Klein, J., Bittremieux, W., Perez-Riverol, Y., Tabb, D. L., Walzer, M., Ricard-Blum, S., Hermjakob, H., Neumann, S., Mak, T. D., Kawano, S., Mendoza, L., Van Den Bossche, T., Gabriels, R., Bandeira, N., … Orchard, S. E. (2023). Proteomics standards initiative at twenty years: Current activities and future work. *Journal of Proteome Research, 22*, 287–301. https://doi.org/10.1021/acs.jproteome.2c00637

Fahy, E., & Subramaniam, S. (2020). RefMet: A reference nomenclature for metabolomics. *Nature Methods, 17*, 1173–1174. https://doi.org/10.1038/s41592-020-01009-y

Gilroy, S. P., & Kaplan, B. A. (2019). Furthering open science in behavior analysis: an introduction and tutorial for using GitHub in research. *Perspectives on Behavior Science, 42*, 565–581.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics, 7*, 23. https://doi.org/10.1186/s13321-015-0068-4

Jeppesen, M. J., & Powers, R. (2023). Multiplatform untargeted metabolomics. *Magnetic Resonance in Chemistry, 1*, 1–26. https://doi.org/10.1002/mrc.5350

Kale, N. S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., Nainala, V. C., Spicer, R. A., Williams, M., Li, X., Salek, R. M., Griffin, J. L., & Steinbeck, C. (2016). MetaboLights: an open-access database repository for metabolomics data. Current Protocols in Bioinformatics. https://doi.org/10.1002/0471250953.bi1413s53

Kirwan, J. A., Gika, H., Beger, R. D., Bearden, D., Dunn, W. B., Goodacre, R., Theodoridis, G., Witting, M., Yu, L. R., & Wilson, I. D. (2022). Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. *Metabolomics, 18*, 70. https://doi.org/10.1007/s11306-022-01926-3

Maciejewski, M. W., Gryk, M. R., Moraru, I. I., Romero, P. R., Ulrich, E. L., Eghbalnia, H. R., Livny, M., Delaglio, F., & Hoch, J. C. (2017). NMRbox: a resource for biomolecular NMR computation. *Biophysical Journal, 112*, 1529–1534.

Navarro, S. L., Nagana Gowda, G. A., Bettcher, L. F., Pepin, R., Nguyen, N., Ellenberger, M., Zheng, C., Tinker, L. F., Prentice, R. L., Huang, Y., Yang, T., Tabung, F. K., Chan, Q., Loo, R. L., Liu, S., Wactawski-Wende, J., Lampe, J. W., Neuhouser, M. L., & Raftery, D. (2023). Demographic, health and lifestyle factors associated with the metabolome in older women. *Metabolites, 13*, 514.

Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., Carroll, A. J., Spalding, D., Paschall, J., Wang, M., Del-Toro, N., Ternent, T., Zhang, P., Buso, N., Bandeira, N., Deutsch, E. W., Campbell, D. S., Beavis, R. C., Salek, R. M., … Hermjakob, H. (2017). Discovering and linking public omics data sets using the Omics discovery index. *Nature Biotechnology, 35*, 406–409. https://doi.org/10.1038/nbt.3790

Peter, K. T., Phillips, A. L., Knolhoff, A. M., Gardinali, P. R., Manzano, C. A., Miller, K. E., Pristner, M., Sabourin, L., Sumarah, M. W., Warth, B., & Sobus, J. R. (2021). Nontargeted analysis study reporting tool: a framework to improve research transparency and reproducibility. *Analytical Chemistry, 93*, 13870–13879. https://doi.org/10.1021/acs.analchem.1c02621

Phinney, K. W., Ballihaut, G., Bedner, M., Benford, B. S., Camara, J. E., Christopher, S. J., Davis, W. C., Dodder, N. G., Eppe, G., Lang, B. E., Long, S. E., Lowenthal, M. S., McGaw, E. A., Murphy, K. E., Nelson, B. C., Prendergast, J. L., Reiner, J. L., Rimmer, C. A., Sander, L. C., … Castle, A. L. (2013). Development of a Standard reference material for metabolomics research. *Analytical Chemistry, 85*, 11732–11738. https://doi.org/10.1021/ac402689t

Powers, R., Andersson, E. R., Bayless, A. L., Brua, R. B., Chang, M. C., Cheng, L. L., Clendinen, C. S., Cochran, D., Copié, V., Cort, J. R., Crook, A. A., Eghbalnia, H. R., Giacalone, A., Gouveia, G. J., Hoch, J. C., Jeppesen, M. J., Maroli, A. S., Merritt, M. E., Pathmasiri, W., … Wishart, D. S. (2024). Best practices in NMR metabolomics: current state. *TrAC Trends in Analytical Chemistry, 171*, 117478. https://doi.org/10.1016/j.trac.2023.117478

Romero, P. R., Kobayashi, N., Wedell, J. R., Baskaran, K., Iwata, T., Yokochi, M., Maziuk, D., Yao, H., Fujiwara, T., Kurusu, G., Ulrich, E. L., Hoch, J. C., & Markley, J. L. (2020). BioMagResBank (BMRB) as a resource for structural biology. *Methods in Molecular Biology, 2112*, 187–218. https://doi.org/10.1007/978-1-0716-0270-6_14

Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., Correa, E., Reijmers, T., Rosato, A., Tenori, L., Turano, P., Marin, S., Deborde, C., Jacob, D., Rolin, D., Dartigues, B., Conesa, P., Haug, K., Rocca-Serra, P., … Steinbeck, C. (2015). Coordination of standards in metabolomics (COSMOS): facilitating integrated metabolomics data Access. *Metabolomics, 11*, 1587–1597. https://doi.org/10.1007/s11306-015-0810-y

Sansone, S. A., Fan, T., Goodacre, R., Griffin, J. L., Hardy, N. W., Kaddurah-Daouk, R., Kristal, B. S., Lindon, J., Mendes, P., Morrison, N., Nikolau, B., Robertson, D., Sumner, L. W., Taylor, C., van der Werf, M., van Ommen, B., & Fiehn, O. (2007a). The metabolomics standards initiative. *Nature Biotechnology, 25*, 846–848. https://doi.org/10.1038/nbt0807-846b

Sansone, S. A., Schober, D., Atherton, H. J., Fiehn, O., Jenkins, H., Rocca-Serra, P., Rubtsov, D. V., Spasic, I., Soldatova, L., Taylor, C., Tseng, A., Viant, M. R., & Members, O. W. G. (2007b). Metabolomics standards initiative: Ontology working group work in progress. *Metabolomics, 3*, 249–256. https://doi.org/10.1007/s11306-007-0069-z

Smirnov, A., Liao, Y., Fahy, E., Subramaniam, S., & Du, X. (2021). ADAP-KDB: a spectral knowledgebase for tracking and prioritizing unknown GC-MS spectra in the NIH's metabolomics data repository. *Analytical Chemistry, 93*, 12213–12220. https://doi.org/10.1021/acs.analchem.1c00355

Steinbeck, C., Conesa, P., Haug, K., Mahendraker, T., Williams, M., Maguire, E., Rocca-Serra, P., Sansone, S. A., Salek, R. M., & Griffin, J. L. (2012). MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics, 8*, 757–760. https://doi.org/10.1007/s11306-012-0462-0

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S., & Subramaniam, S. (2016). 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. https://doi.org/10.1093/nar/gkv1042

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J.

L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics, 3*, 211–221. https://doi.org/10.1007/s11306-007-0082-2

Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*, 3–16. https://doi.org/10.1007/s11306-011-0330-3

Tolstikov, V., Moser, A. J., Sarangarajan, R., Narain, N. R., & Kiebish, M. A. (2020). Current status of metabolomic biomarker discovery: impact of study design and demographic characteristics. *Metabolites, 10*, 224.

Villalba, H., Llambrich, M., Gumà, J., Brezmes, J., & Cumeras, R. (2023). A metabolites merging strategy (MMS): harmonization to enable studies' intercomparison. *Metabolites, 13*, 1167.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*, 31–36. https://doi.org/10.1021/ci00057a005

Wigh, D. S., Goodman, J. M., & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *Wires Computational Molecular Science, 12*, e1603. https://doi.org/10.1002/wcms.1603

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

Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics, 1*, 92–107. https://doi.org/10.2174/2213235x11301010092

Xi, B., Gu, H., Baniasadi, H., & Raftery, D. (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Methods in Molecular Biology, 1198*, 333–353. https://doi.org/10.1007/978-1-4939-1258-2_22

Authors and Affiliations

Goncalo Jorge Gouveia[1,2] · Thomas Head[1,3] · Leo L. Cheng[1,4] · Chaevien S. Clendinen[1,5] · John R. Cort[1,6] · Xiuxia Du[1,7] · Arthur S. Edison[1,8] · Candace C. Fleischer[1,9] · Jeffrey Hoch[1,10] · Nathaniel Mercaldo[1,11] · Wimal Pathmasiri[1,12] · Daniel Raftery[1,13] · Tracey B. Schock[1,14] · Lloyd W. Sumner[1,15] · Panteleimon G. Takis[1,16,17] · Valérie Copié[1,18] · Hamid R. Eghbalnia[1,10] · Robert Powers[1,19]

✉ Robert Powers
rpowers3@unl.edu

Valérie Copié
vcopie@montana.edu

Hamid R. Eghbalnia
heghbalnia@gmail.com

[1] Metabolomics Association of North America (MANA), NMR Special Interest Group, Edmonton, Canada

[2] Institute for Bioscience and Biotechnology Research, National Institute of Standards and Technology, University of Maryland, Gudelsky Drive, Rockville, MD 20850, USA

[3] University of British Columbia, Kelowna, BC V1V 1V7, Canada

[4] Department of Pathology and Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[5] Earth and Biological Sciences Directorate, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352, USA

[6] Earth and Biological Sciences Directorate, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA

[7] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, 9291 University City Blvd, Charlotte, NC 28223, USA

[8] Department of Biochemistry, University of Georgia, Athens, GA, USA

[9] Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA 30322, USA

[10] Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT 06030-3305, USA

[11] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[12] Department of Nutrition, School of Public Health, Nutrition Research Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[13] Department of Anesthesia and Pain Medicine, University of Washington, Seattle, WA 98109, USA

[14] Chemical Sciences Division, National Institute of Standards and Technology (NIST), Charleston, SC 29412, USA

[15] Department of Biochemistry, MU Metabolomics Center, Bond Life Sciences Center, Interdisciplinary Plant Group, University of Missouri, Columbia, MO 65211, USA

[16] Section of Bioanalytical Chemistry, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London SW7 2AZ, UK

[17] Department of Metabolism, Digestion and Reproduction, National Phenome Centre, Imperial College London, London W12 0NN, UK

[18] Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717-3400, USA

[19] Department of Chemistry, Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, 722 Hamilton Hall, Lincoln, NE 68588-0304, USA