Received: 23 June 2023 | Revised: 23 August 2023 | Accepted: 24 August 2023

DOI: 10.1002/pmic.202300032

Proteomics
Proteomics and Systems Biology

REVIEW

(Re-)use and (re-)analysis of publicly available metabolomics data

Michael Witting[1,2]

[1] Metabolomics and Proteomics Core, Helmholtz Zentrum München, Neuherberg, Germany

[2] Chair of Analytical Food Chemistry, TUM School of Life Sciences, Freising-Weihenstephan, Germany

Correspondence
Michael Witting, Metabolomics and Proteomics Core, Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neuherberg 85764, Germany.
Email: michael.witting@helmholtz-munich.de

Abstract

Metabolomics, the systematic measurement of small molecules (<1000 Da) in a given biological sample, is a fast-growing field with many different applications. In contrast to transcriptomics and proteomics, sharing of data is not as widespread in metabolomics, though more scientists are sharing their data nowadays. However, to improve data analysis tools and develop new data analytical approaches and to improve metabolite annotation and identification, sharing of reference data is crucial. Here, different possibilities to share (metabolomics) data are reviewed and some recent approaches and applications regarding the (re-)use and (re-)analysis are highlighted.

KEYWORDS
bioinformatics, data, databases, metabolomics, processing and analysis, technology

1 | INTRODUCTION

New scientific discoveries are built on previous results and findings, both positive and negative. Especially nowadays, the pace of publishing new results is exceptionally high. Simultaneously, datasets are becoming increasingly complex, particularly in omics-type data characterized by high dimensions. One such recent addition to the omics analysis is metabolomics, the systematic measurement of all small molecules (<1000 Da) in a biological sample, for example, cells, tissues, or biological fluids. While public data sharing is common practice in other omics fields, such as genomics, transcriptomics, and proteomics, with scientific journals often enforcing it, this is still not the case in metabolomics, although many journals encourage public data sharing [1]. Metabolites are not encoded in the genome and are highly dependent on the environment (e.g., food, exposure, lifestyle, etc.). Therefore, no standardized methodology and way of data sharing is available. Still, data sharing is vital for the further development of metabolomics to improve data analysis strategies and new software tools and metabolite annotation and identification.

Data sharing can be practiced on different levels in metabolomics, and sharing of raw data from studies is only one possibility. Shared data enables the comparison of study outcomes and potentially even integrating various studies to enhance statistical power. Several tips for comparing public metabolomics studies have recently been suggested [2]. In addition, many other data-sharing possibilities exist and are briefly summarized here.

However, data sharing is only one side of the coin. The (re)use and potential (re)analysis is the other. To advance metabolomics approaches, data analysis tools and software sharing of reference data is required. This includes not only the data itself but also related metadata, which allows the correct interpretation and use. While this article focuses on the sharing of high-resolution liquid chromatography-tandem mass spectrometry (LC-MS/MS) data, several aspects discussed can also be applied to other methodologies, such as targeted or nuclear magnetic resonance spectroscopy (NMR)-based metabolomics.

Abbreviations: HMDB, Human Metabolome Database; ChEBI, Chemical Entities of Biological Interest; GNPS, Global Natural Products Social Molecular Networking; RT, Retention Time; CCS, Collisional Cross Section; TWIMS, Traveling Wave Ion Mobility Spectrometry; DTIMS, Drift Tube Ion Mobility Spectrometry; TIMS, Trapped Ion Mobility Spectrometry.

2 | TYPES OF DATA FOR RE-USE AND DATABASES

Data sharing in metabolomics facilitates re-use across multiple levels, enabling the scientific community to derive greater benefits. Increased data sharing leads to enhanced collective knowledge. Machine learning and artificial intelligence are more commonly used in metabolomics, necessitating a larger pool of training data from different areas and application fields. Below, we provide a concise summary of the different types of data which can be shared in metabolomics, along with corresponding resources.

2.1 | Metabolite structures

Although several metabolite structure databases exist, grow, and are curated, the further sharing of metabolite structures needs to be encouraged. This is particularly crucial for newly identified metabolites with novel structures. While these structures are often part of the articles or the associated supplementary information, structures cannot be found in machine-readable formats such as Simplified Molecular-Input Line-Entry System (SMILES) or International Chemical Identifier (InChIs). A positive example here is SMID-DB.org, which stores the structures and information on secondary metabolites from *Caenorhabditis elegans* and other related nematodes, which have been identified in different publications, including SMILES and, if available, reference spectra [3]. Additionally, sharing such structures in larger, more general databases such as Chemical Entities of Biological Interest (ChEBI) [4, 5], PubChem [6], ChemSpider, or LipidMaps [7] make them accessible to a broader audience.

LC-MS/MS is often not able to identify full structural details such as the position of hydroxyl groups in complex metabolites such as flavonoids or the position and stereochemistry of double bonds in lipids. To address this limitation, ChEBI, for example, allows the submission and storage of partial structures. Submission of partial structures and the associated molecular formula makes it possible to increase the chemical space covered. Subsequently, when the full structures are identified, they can be linked to the partial structure via the ChEBI ontology (e.g., 1,2-dihexanoly-sn-glycero-3-phosphocholine [CHEBI:72999] is a phosphatidylcholine 32:0 [CHEBI:66850]).

However, the actual structure of metabolites and the information on the organisms that produce them is crucial. This organism-specific information aids dereplication during metabolite identification and helps to filter spurious identifications that are unlikely to be present in the studied organism. A recent example is the LOTUS database, which contains taxonomical information on organisms producing the respective natural product [8]. LOTUS is completely linked to Wikidata and is built entirely from open data. Additionally, other databases, such as ChEBI or LipidMaps, also store associations between molecules and organisms that produce them. In ChEBI, specific entries, such as CHEBI:78804 – *C. elegans* metabolite or CHEBI:75771 – mouse metabolite, have been generated, and metabolites can be linked to them. Such information on the presence of metabolites in specific species or taxa can be used for improved annotation of metabolites [9, 8]. Furthermore, species-specific metabolites and reference spectra databases can be constructed from this information. Beside the organism-specificity, in case of multicellular organisms the tissue or cell-type origin of a metabolites is of great importance for correct metabolite annotation, for example, having a role such as "mouse lung metabolite". However, currently this information is not part of most metabolite structure databases. The Human Metabolome Database (HMDB) represents an exception curating this information for several metabolites, for example location in biospecimen or tissues [10]. A summary of all mentioned databases with their respective URL can be found in Table 1.

2.2 | Reference mass spectra

Sharing of reference mass spectra is one of the most obvious and important factors in advancing LC-MS/MS-based metabolomics. Laboratories can't hold a reference standard for each known metabolite, resulting in limited focus and size of their in-house reference libraries. Though, to be able to annotate metabolites beyond these libraries, it is essential to incorporate diverse reference spectra from different analytical platforms (e.g., different MS types, Orbitrap, QToF, IT). Though, according to different identification schemes, these reference spectra do not provide definitive identifications (which requires a reference standard to be measured under the identical analytical condition), their availability dramatically helps narrow the list of putative [11, 12]. Though more and more laboratories share their reference libraries in the public domain, only a small growth in novel compounds is observed.

In most cases, laboratories initially focus on constructing in-house libraries with common metabolites like amino acids, organic acids, and fatty acids or rely on commercially available chemical libraries. Also, to further boost advances *in-silico* methods beyond their current state, more chemical diversity is required. Fragmentation spectra of novel compounds identified shall be deposited in electronic databases (and not only included in the supplementary information of articles).

Different platforms for sharing MS data have evolved over the last years, and it is becoming more and more standard to upload reference spectra of substances measured in in-house libraries. MassBank [13], MassBank of North America, and GNPS are primary databases that can store MS data [14]. All of them offer different functionalities on top of actually storing the spectra. For example, the GNPS ecosystem offers the Mass Search Tool (MASST), which allows searching reference libraries and public datasets for similar spectra [14, 15]. Different variants of this search tool now exist, for example, FoodMASST, microbeMASST, or plantMASST [16]. To enable such tools, a combination with taxonomically informed metabolite libraries is required (see above). Beyond the purpose of annotation, reference spectra can be used to develop and evaluate novel in-silico approaches for the analysis of MS^2 data, for example, CSI:FingerID [17, 18], CFM-ID [19, 20], MetFrag, etc. [21]. Such tools enable advances beyond classical library and spectral matching for metabolite annotation, opening new avenues for analysis and interpretation.

TABLE 1 Resources to share (novel) metabolite structures.

Resource	Weblink	Comment
Chemical Entities of Bioloigcal Interest (ChEBI)	https://www.ebi.ac.uk/chebi/	- Submission of partial structures possible - Rich ontology to link chemical structures
LipidMaps	https://lipidmaps.org/	- Lipid-centric database also storing taxonomic information
LOTUS	https://lotus.naturalproducts.net/	- Natural product database also storing taxonomic information
PubChem	https://pubchem.ncbi.nlm.nih.gov/	- General chemical database
HMDB	https://hmdb.ca/	- Stores tissue and biospecimen location for human metabolites

2.3 | Retention time and collisional cross section data

Retention times (RTs) and collisional cross sections (CCSs) are valuable orthogonal parameters that can be used to identify metabolites. Above mentioned identification schemes require such an orthogonal parameter of chemical reference standards matched to a metabolite feature for the highest level of identification [11, 12]. While CCS values are almost instrument-independent, RTs strongly depend on the employed chromatographic system and instrumentation. Even though approaches for the normalization of RTs have been suggested [22, 23], substantial variations persist between different column brands, and sharing of retention data is not widespread. It is important to note that data sharing should encompass metadata sharing, as the (re-)use of RTs heavily relies on the available metadata [24]. RTs alone are practically useless without the information on the employed column, eluents, flow rate, temperature, and other relevant parameters. Despite this, RT collections are becoming more available. One example is PredRet, which represents an RT collection, but also offers a tool for projecting RTs across different chromatographic systems [25, 26]. In the future, larger collections of RTs will enable the development of novel machine-learning models for the prediction of RTs to enhance metabolite identification [27].

With the advent of ion mobility spectrometry and the more widespread application in metabolomics and lipidomics, CCS databases are becoming more critical. Ion mobility enables the separation of ions based on their shape enabling the potential separation of isobaric and isomeric structures. Since deviations between instruments are typically relatively small, CCS values obtained in different laboratories can be used for metabolite annotation [28, 29]. One example of a CCS database is the CCS Compendium storing CCS values from different instruments (TWIMS, DTIMS, TIMS) [30]. Besides the CCS compendium, different collections exist and enable the prediction of CCS values [31–34].

2.4 | Entire datasets (raw data)

Besides sharing individual mass spectra, entire LC-MS/MS runs or datasets can be shared. They often include processed feature tables that provide information about metabolite quantities, peak intensities, or areas. While single feature tables are often included in the supplementary information of published articles or generic data-sharing platforms such as Zenodo, there are dedicated platforms for sharing of metabolomics raw data, such as MetaboLights [35], Metabolomics Workbench [36], or MassIVE/GNPS [14]. Sharing of such raw data allows other scientists to evaluate the results of the specific study but also to develop new algorithms for peak picking, ion deconvolution, etc. This is especially important when new analytical methods or approaches are becoming available (such as data-independent acquisition [DIA] or ion mobility in the past).

Furthermore, in theory, data from different sources can be fused and compared to increase statistical power. However, in reality, the diversity of data from different laboratories makes direct comparisons challenging, as different mass spectrometric setups might have different responses to a specific metabolite. An essential factor for the (re-)use of such datasets is the comprehensive capture of metadata, including information about the organism, experimental conditions, and other relevant details. As an example, Harrieder et al. recently checked metadata associated with different datasets in Metabolights and Metabolomics Workbench for the completeness of chromatographic metadata [24]. They found that 70% of all data was incomplete and missed important information. Lastly, if data is stored in metabolomics-centric repositories, any information regarding identified metabolites can be easily retrieved without manually searching within articles or their supplementary information. Furthermore, most of these repositories allow to specify for example organism and tissue of origin, which allows to reconstruct specific metabolomes, even including unknown metabolites. A summary of all mentioned repositories can be found in Table 2.

2.5 | (Spatial) Distribution of metabolites

Certain metabolites are only produced in specific organs, tissues or even cells. Information on the spatial distribution of metabolites is important for better understanding of biological functions. The METASPACE project (https://metaspace2020.eu/) offers an annotation platform for spatial metabolomics based on MS imaging (MSI). The webportal represents a repository for high-resolution MSI data sets. Annotation on the MS1 level can performed using several of the mentioned metabolite structure databases [37]. Images are associated with rich metadata such instrumentation and origin of samples. Beside images other database exist, for example, the MetaboAtlas21 (https://metaboatlas21.metabolomics.fgu.cas.cz/), which allows to browse dis-

TABLE 2 Repositories for metabolomics (raw) data.

Repository	Weblink	Comment
MetaboLights	https://www.ebi.ac.uk/metabolights/	- Database for storing of raw data and associated results - Cross-species (allows to browse for species-specific metabolites)
Metabolomics Workbench	https://www.metabolomicsworkbench.org/	- Hosts RefMet (Reference List of Metabolite Names)
GNPS/MassIVE	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp	- Rich ecosystem for (re-)use and (re-)analysis of metabolomics data
METASPACE	https://metaspace2020.eu/	- Ecosystem for annotation and sharing of MSI data

tribution of metabolites and lipids in different mouse tissues. Such atlases will become more important in future.

3 | (RE-)USE AND (RE-)ANALYSIS OF METABOLOMICS DATA

3.1 | Metabolite identification

The different presented types of data allow a different level of re-use and re-analysis. The most straightforward way to re-use public data is through mass spectral libraries for metabolite identification. Publicly shared spectra can be matched against measured spectra from own experiments to aid annotation of metabolites not covered in in-house databases. This provides putative annotations and can help to narrow down potential candidates for further structural elucidation. Besides the actual library matching, high-quality reference spectra are required for the development of in silico annotations tools, such CSI:FingerID, CFM-ID, and others [19, 17, 38]). For a more detailed review, see [39].

Submission of novel structures to chemical reference databases such as ChEBI, PubChem, or others expands the search space for the aforementioned in-silico tools. Together with the information on organisms producing metabolites, this can narrow down potential candidates. However, great care needs to be taken. Ideally, manual curation and data verification must be performed since automatic methods and meta-scores can potentially result in an artificially high increase in “true positive results” [40]. Furthermore, metabolite structure databases can serve as input for the annotation of metabolites in MSI experiments [37].

3.2 | Reference datasets for the development of new workflows

Entire datasets can be used to develop new bioinformatics tools and approaches. This includes every possible step, from peak picking to feature grouping and metabolite identification. Bioinformatics laboratories working on such tools often do not have the capacity to generate required datasets on their own and rely on publicly available datasets.

For instance, the MetaboLights datasets MTBLS235 and MTBLS234 contain reference data for developing peak picking and assembling into features [41]. Notably, it contained a synthetic dataset for which the ground truth is known (known number of metabolites or features and their identity, which is typically not the case for biological datasets). Another example is the dataset MTBLS1108 submitted to MetaboLights, which contains data from data-dependent (DDA) and data- DIA, which was used for the development of the DIAMetAlyzer workflow [42]. In addition to sharing the complete dataset, the workflow, and associated code are also made available (https://openms.de/application/diametalyzer/ and https://github.com/oliveralka/DIAMetAlyzer_additional_code). This enables direct benchmarking of new processing methods for DIA data and the comparison against an established workflow.

3.3 | Reanalysis of metabolomics data at a repository scale

Publicly shared data can be used for reanalysis, including new statistical analysis, search for novel compounds described, or comparison with other datasets. ReDu was developed for exactly this purpose allowing the extraction of specific knowledge from public datasets [43]. ReDu allows establishing associations between compounds and different metadata,for example, sex, life stage, etc.

Advancements in computational power and improved algorithms allow metabolomics data analysis at a repository scale with hundreds to thousands of LC-MS/MS runs and spectra. One example was performed for testing of a novel confidence score for metabolite annotation beyond spectral libraries [18]. Over 2500 LC-MS/MS runs from different human sources were annotated, including novel compounds not present in HMDB [44]. Another example is the creation of new suspect spectral libraries [45]. Spectra of new structures have been inferred from nearest neighbors of spectra with reference matches in molecular networks, for example, for novel acylcarnitine species.

4 | PROBLEMS AND OPPORTUNITIES

Metabolomics is generally still very much technology driven; as such, no universally accepted analysis method exists (if ever possible). Different laboratories use different types of equipment (e.g., Orbitraps vs. ToFs) and different chromatographic methods [46]. While the integration of targeted metabolomics data based on absolute concentrations or known and identified metabolites might be possible, it becomes more challenging for non-identified metabolites. Instrumentation variations, such as differences in dynamic range and ionization efficiencies due to variations in ionization sources, result in varying relative abundances of adducts and in-source fragmentation, which are compound-dependent. Furthermore, different chromatographic methods will result in different RTs. Approaches such as PredRet can partially help to establish correspondence between datasets [26]. The use of MS^2 additionally aids information for mapping. However, differences in collision energy between different instrumentation and experimental settings can lead to differences in fragmentation spectra. The use of merged or ramped spectra might overcome this in future. More research is required to better understand how differences between analytical setups are evolving and if there are ways to overcome and normalize them. In case of lipidomics analysis, it has been recently shown that shared reference materials can improve harmonization of different methods [47]. Besides the actual technical differences, several differences in the semantics of metabolites exist, for example, identifiers for metabolites are not harmonized. Metabolite names can often be ambiguous, and systematic IUPAC names are often not used because of their lengths and complexity, and trivial names are preferred (e.g., (2S)-2-amino-3-(1H-indol-3-yl)propanoic acid vs. L-Tryptophan). The most unambiguous identifier for a metabolite is its structure, which can be reported using a SMILES, InChI, or InChIKey. Several approaches have been published to overcome this issue, for example, bridgeDB or RefMet [48, 49]. However, metabolite nomenclature is a re-occurring issue [50].

Nevertheless, several opportunities are given by sharing metabolomics data. Different metabolomics datasets covering the same or similar biological questions can be combined to increase the statistical power of studies. However, since metabolomics is far from a standardized technology, integration of datasets might be complicated if collected on different platforms. Standardized targeted metabolomics methods and kits can help to generate data that can be easily merged to improve statistical power [51, 52]. Results from such studies will represent the first line of large-scale integration of data for broader data analysis and enable new findings. However, the knowledge of the metabolism of different organisms is still scattered. Public sharing of metabolomics datasets also allows the data-driven reconstruction of organism metabolomes. For example, the repository MetaboLights allows to search for organism-specific studies and compounds. Compounds are retrieved from the annotated and identified compounds in the datasets deposited and linked to a specific species. Together with in-silico reconstructions of metabolism (also known as genome-scale metabolic models), the knowledge can be continuously updated and enhanced to create a more fine-grained picture.

Besides scientific questions, public datasets can be used to educate the next generation of metabolomics scientists.

5 | CONCLUSION

Data sharing in metabolomics can be conducted on different levels, from submitting novel chemical structures to structural databases and sharing reference spectra and libraries to entire datasets. Such sharing is essential for the growth of the field of metabolomics. Though different obstacles and problems associated with metabolomics need to be solved (e.g., common identifiers, comparable methods), each new dataset, reference spectrum, or novel structure increases our knowledge of the metabolism of different organisms and biological systems and is therefore valuable and important.

However, the field of metabolomics is far from being standardized and requires more vigorous control of metadata related to experimentation and instrumentation. Without meaningful metadata, shared data is only of partial use. For example, an RT without a description of the employed chromatographic system represents just a single number or a reference spectrum without information on the chemical structure cannot be used for training purposes.

As technological advancements highly influence metabolomics, it is crucial to make new types of data for the community to keep up with these developments. With the introduction of ion mobility instruments, there has been a significant release of CCS databases and collections, which is expected with novel and alternative fragmentation modes, such as electron activated dissociation (EAD) or ultraviolet photodissociation (UVPD). Both have been shown to be valuable tools for the detailed analysis of lipids allowing them to determine double bond and sn-positions in glycerophospholipids [53, 54]. Furthermore, new data types are needed, such as for the prediction of quantities, ionization efficiency, or adduct formation [55, 56].

It is important to acknowledge that metabolomics is still behind fields like genomics, transcriptomics, and proteomics in terms of data sharing, and new standards need to be established. Nevertheless, big parts of the metabolomics community realized the value of sharing data on different scales, and data becomes more available. Facilitating easy integration and uploading to the metabolomics repository will help to streamline this process further. Current software tools often allow the export to common open data formats, such as .mzML and mzTab [57, 58, 59]. Once automatic upload and (re) data analysis become feasible; metabolomics will flourish and be used by a wider range of scientists, including non-experts. Until then: Share your data!

ACKNOWLEDGMENTS

The author wants to acknowledge all maintainers and curators of different repositories and databases; you are the unsung heroes of metabolomics.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The author has declared no conflict of interest.

DATA AVAILABILITY STATEMENT

not applicable

ORCID

Michael Witting https://orcid.org/0000-0002-1462-4426

REFERENCES

1. Spicer, R. A., & Steinbeck, C. (2017). A lost opportunity for science: Journals promote data sharing in metabolomics but do not enforce it. *Metabolomics*, *14*(1), 16. https://doi.org/10.1007/s11306-017-1309-5
2. Stancliffe, E., & Patti, G. J. (2022). Quick tips for re-using metabolomics data. *Nature Cell Biology*, *24*(11), 1560–1562. https://doi.org/10.1038/s41556-022-01019-2
3. Artyukhin, A. B., Zhang, Y. K., Akagi, A. E., Panda, O., Sternberg, P. W., & Schroeder, F. C. (2018). Metabolomic “Dark Matter” dependent on peroxisomal β-oxidation in Caenorhabditis elegans. *Journal of the American Chemical Society*, *140*(8), 2841–2852. https://doi.org/10.1021/jacs.7b11811
4. Hastings, J., De Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., & Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013 2013. *Nucleic Acids Research*, *41*(D1), D456–D463. https://doi.org/10.1093/nar/gks1146
5. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031
6. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(Database issue), D1202–D1213. https://doi.org/10.1093/nar/gkv951
7. Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., Merrill, A. H., Murphy, R. C., Raetz, C. R. H., Russell, D. W., & Subramaniam, S. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, *35*(Suppl 1), D527–D532. https://doi.org/10.1093/nar/gkl838
8. Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J. G., Stephan, R., Page, R., Vondrášek, J., Steinbeck, C., Pauli, G. F., Wolfender, J.-L., Bisson, J., & Allard, P.-M. (2022). The LOTUS initiative for open knowledge management in natural products research. *eLife*, *26*(11), e70780. https://doi.org/10.7554/eLife.70780
9. Rutz, A., Dounoue-Kubo, M., Ollivier, S., Bisson, J., Bagheri, M., Saesong, T., Ebrahimi, S. N., Ingkaninan, K., Wolfender, J.-L., & Allard, P.-M. (2019). Taxonomically informed scoring enhances confidence in natural products annotation. *Frontiers in Plant Science*, *10*, https://doi.org/10.3389/fpls.2019.01329
10. Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., & Gautam, V. (2021). HMDB 5.0: The human metabolome database for 2022. *Nucleic Acids Research*, *50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062
11. Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science, & Technology*, *48*(4), 2097–2098. https://doi.org/10.1021/es5002105
12. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., & Viant, M R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, *3*(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2
13. Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., & Nishioka, T. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, *45*, https://doi.org/10.1002/jms.1777
14. Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, *34*(8), 828–837. https://doi.org/10.1038/nbt.3597
15. Wang, M., Jarmusch, A. K., Vargas, F., Aksenov, A. A., Gauglitz, J. M., Weldon, K., Petras, D., Da Silva, R., Quinn, R., Melnik, A. V., Van Der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Nothias, L. F., Aceves, C. M., Panitchpakdi, M., Brown, E., Di Ottavio, F., Sikora, N., Elijah, E. O., & Dorrestein, P. C. (2020). Mass spectrometry searches using MASST. *Nature Biotechnology*, *38*(1), 23–26. https://doi.org/10.1038/s41587-019-0375-9
16. West, K. A., Schmid, R., Gauglitz, J. M., Wang, M., & Dorrestein, P. C. (2022). foodMASST a mass spectrometry search tool for foods and beveragess and beverages. *npj Science of Food*, *6*(1), 22. https://doi.org/10.1038/s41538-022-00137-3
17. Dührkop, K., Shen, H., Meusel, M., Rousu, J., & Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 12580–5. https://doi.org/10.1073/pnas.1509788112
18. Hoffmann, M. A., Nothias, L.-F., Ludwig, M., Fleischauer, M., Gentry, E. C., Witting, M., Dorrestein, P. C., Dührkop, K., & Böcker, S. (2022). High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology*, *40*(3), 411–421. https://doi.org/10.1038/s41587-021-01045-9
19. Djoumbou-Feunang, Y., Pon, A., Karu, N., Zheng, J., Li, C., Arndt, D., Gautam, M., Allen, F., & Wishart, D S. (2019). CFM-ID 3.0: CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites*, *9*(4), 72. https://www.mdpi.com/2218-1989/9/4/72
20. Allen, F., Greiner, R., & Wishart, D. (2015). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, *11*(1), 98–110. https://doi.org/10.1007/s11306-014-0676-4
21. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, *8*(1), 3. https://doi.org/10.1186/s13321-016-0115-9
22. Aalizadeh, R., Nikolopoulou, V., & Thomaidis, N. S. (2022). Development of liquid chromatographic retention index based on cocamide diethanolamine homologous series (C(n)-DEA). *Analytical Chemistry*, *94*(46), 15987–15996. https://doi.org/10.1021/acs.analchem.2c02893
23. Stoffel, R., Quilliam, M. A., Hardt, N., Fridstrom, A., & Witting, M. (2022). N-Alkylpyridinium sulfonates for retention time indexing in reversed-phase-liquid chromatography-mass spectrometry-based metabolomics. *Analytical and Bioanalytical Chemistry*, *414*(25), 7387–7398. https://doi.org/10.1007/s00216-021-03828-0
24. Harrieder, E.-M., Kretschmer, F., Dunn, W., Böcker, S., & Witting, M. (2022). Critical assessment of chromatographic metadata in publicly available metabolomics data repositories. *Metabolomics*, *18*(12), 97. https://doi.org/10.1007/s11306-022-01956-x
25. Low, D. Y., Micheau, P., Koistinen, V. M., Hanhineva, K., Abrankó, L., Rodriguez-Mateos, A., Da Silva, A. B., Van Poucke, C., Almeida, C., Andres-Lacueva, C., Rai, D. K., Capanoglu, E., Tomás Barberán, F. A., Mattivi, F., Schmidt, G., Gürdeniz, G., Valentová, K., Bresciani, L., Petrásková, L., … Manach, C. (2021). Data sharing in PredRet for accurate prediction of retention time: Application to plant food bioactive compounds. *Food Chemistry*, *357*, 129757. https://doi.org/10.1016/j.foodchem.2021.129757

26. Stanstrup, J., Neumann, S., & Vrhovsek, U. (2015). PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry*, *87*(18), 9421–9428. https://doi.org/10.1021/acs.analchem.5b02287
27. Witting, M., & Böcker, S. (2020). Current status of retention time prediction in metabolite identification. *Journal of Separation Science*, *43*(9-10), 1746–1754. https://doi.org/10.1002/jssc.202000060
28. Hinnenkamp, V., Klein, J., Meckelmann, S. W., Balsaa, P., Schmidt, T. C., & Schmitz, O. J. (2018). Comparison of CCS values determined by traveling wave ion mobility mass spectrometry and drift tube ion mobility mass spectrometry. *Analytical Chemistry*, *90*(20), 12042–12050. https://doi.org/10.1021/acs.analchem.8b02711
29. Nye, L. C., Williams, J. P., Munjoma, N. C., Letertre, M. P. M., Coen, M., Bouwmeester, R., Martens, L., Swann, J. R., Nicholson, J. K., Plumb, R. S., Mccullagh, M., Gethings, L. A., Lai, S., Langridge, J. I., Vissers, J. P. C., & Wilson, I. D. (2019). A comparison of collision cross section values obtained via travelling wave ion mobility-mass spectrometry and ultra high performance liquid chromatography-ion mobility-mass spectrometry: Application to the characterisation of metabolites in rat urine. *Journal of Chromatography A*, *1602*, 386–396. https://doi.org/10.1016/j.chroma.2019.06.056
30. Leaptrot, K. L., May, J. C., Dodds, J. N., & Mclean, J. A. (2019). Ion mobility conformational lipid atlas for high confidence lipidomics. *Nature Communications*, *10*(1), 985. https://doi.org/10.1038/s41467-019-08897-5
31. Zhou, Z., Luo, M., Chen, X., Yin, Y., Xiong, X., Wang, R., & Zhu, Z.-J. (2020). Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics. *Nature Communications*, *11*(1), 4334. https://doi.org/10.1038/s41467-020-18171-8
32. Zhou, Z., Shen, X., Tu, J., & Zhu, Z.-J. (2016). Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical Chemistry*, *88*(22), 11084–11091. https://doi.org/10.1021/acs.analchem.6b03091
33. Zhou, Z., Tu, J., Xiong, X., Shen, X., & Zhu, Z.-J. (2017). LipidCCS: Prediction of collision cross-section values for lipids with high precision to support ion mobility–mass spectrometry-based lipidomics. *Analytical Chemistry*, *89*(17), 9559–9566. https://doi.org/10.1021/acs.analchem.7b02625
34. Zhou, Z., Xiong, X., & Zhu, Z.-J. (2017). MetCCS predictor: A web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics*, *33*(14), 2235–2237. https://doi.org/10.1093/bioinformatics/btx140
35. Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S.-A., Griffin, J. L., & Steinbeck, C. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, *41*(D1), D781–D786. https://doi.org/10.1093/nar/gks1004
36. 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*(D1), D463–D470. https://doi.org/10.1093/nar/gkv1042
37. Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., Kovalev, V., Fuchser, J., Nikolenko, S., Pineau, C., Becker, M., & Alexandrov, T. (2017). FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, *14*(1), 57–60. https://doi.org/10.1038/nmeth.4072
38. Allen, F., Greiner, R., & Wishart, D. (2015). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, *11*, https://doi.org/10.1007/s11306-014-0676-4
39. Bittremieux, W., Wang, M., & Dorrestein, P. C. (2022). The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics*, *18*(12), 94. https://doi.org/10.1007/s11306-022-01947-y
40. Hoffmann, M. A., Kretschmer, F., Ludwig, M., & Böcker, S. (2023). MAD HATTER correctly annotates 98% of small molecule tandem mass spectra searching in PubChem. *Metabolites*, *13*(3), 314. https://www.mdpi.com/2218-1989/13/3/314
41. Kenar, E., Franken, H., Forcisi, S., Wörmann, K., Häring, H.-U., Lehmann, R., Schmitt-Kopplin, P., Zell, A., & Kohlbacher, O. (2014). Automated label-free quantification of metabolites from liquid chromatography–mass spectrometry data*. *Molecular, & Cellular Proteomics*, *13*(1), 348–359. https://doi.org/10.1074/mcp.M113.031278
42. Alka, O., Shanthamoorthy, P., Witting, M., Kleigrewe, K., Kohlbacher, O., & Röst, H. L. (2022). DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. *Nature Communications*, *13*(1), 1347. https://doi.org/10.1038/s41467-022-29006-z
43. Jarmusch, A. K., Wang, M., Aceves, C. M., Advani, R. S., Aguirre, S., Aksenov, A. A., Aleti, G., Aron, A. T., Bauermeister, A., Bolleddu, S., Bouslimani, A., Caraballo Rodriguez, A. M., Chaar, R., Coras, R., Elijah, E. O., Ernst, M., Gauglitz, J. M., Gentry, E. C., Husband, M., … Dorrestein, P. C. (2020). ReDU: A framework to find and reanalyze public mass spectrometry data. *Nature Methods*, *17*(9), 901–904. https://doi.org/10.1038/s41592-020-0916-7
44. Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., … Scalbert, A. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, *46*(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089
45. Bittremieux, W., Avalon, N. E., Thomas, S. P., Kakhkhorov, S. A., Aksenov, A. A., Gomes, P. W. P., Aceves, C. M., Caraballo Rodríguez, A. M., Gauglitz, J. M., Gerwick, W. H., Jarmusch, A. K., Kaddurah-Daouk, R. F., Kang, K. B., Kim, H. W., Kondić, T., Mannochio-Russo, H., Meehan, M. J., Melnik, A. V., Nothias, L.-F., … Dorrestein, P. C. (2022). Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *bioRxiv*, 2022.2005.2015.490691. https://doi.org/10.1101/2022.05.15.490691
46. Harrieder, E.-M., Kretschmer, F., Böcker, S., & Witting, M. (2022). Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics. *Journal of Chromatography B*, *1188*, 123069. https://doi.org/10.1016/j.jchromb.2021.123069
47. Triebl, A., Burla, B., Selvalatchmanan, J., Oh, J., Tan, S. H., Chan, M Y., Mellet, N. A., Meikle, P. J., Torta, F., & Wenk, M. R. (2020). Shared reference materials harmonize lipidomics across MS-based detection platforms and laboratories. *Journal of Lipid Research*, *61*(1), 105–115. https://doi.org/10.1194/jlr.D119000393
48. Fahy, E., & Subramaniam, S. (2020). RefMet: A reference nomenclature for metabolomics. *Nature Methods*, *17*(12), 1173–1174. https://doi.org/10.1038/s41592-020-01009-y
49. Van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., & Evelo, C. T. (2010). The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, *11*(1), 5. https://doi.org/10.1186/1471-2105-11-5
50. Koistinen, V., Kärkkäinen, O., Keski-Rahkonen, P., Tsugawa, H., Scalbert, A., Arita, M., Wishart, D., & Hanhineva, K. (2023). Towards a Rosetta stone for metabolomics: Recommendations to overcome inconsistent metabolite nomenclature. *Nature Metabolism*, *5*(3), 351–354. https://doi.org/10.1038/s42255-023-00757-3
51. Floegel, A., Kühn, T., Sookthai, D., Johnson, T., Prehn, C., Rolle-Kampczyk, U., Otto, W., Weikert, C., Illig, T., Von Bergen, M., Adamski, J., Boeing, H., Kaaks, R., & Pischon, T. (2018). Serum metabolites and risk of myocardial infarction and ischemic stroke: A targeted metabolomic approach in two German prospective cohorts. *European Journal of Epi-*

demiology, *33*(1), 55–66. https://doi.org/10.1007/s10654-017-0333-0
52. Iqbal, K., Dietrich, S., Wittenbecher, C., Krumsiek, J., Kühn, T., Lacruz, M. E., Kluttig, A., Prehn, C., Adamski, J., Von Bergen, M., Kaaks, R., Schulze, M. B., Boeing, H., & Floegel, A. (2018). Comparison of metabolite networks from four German population-based studies. *International Journal of Epidemiology*, *47*(6), 2070–2081. https://doi.org/10.1093/ije/dyy119
53. Campbell, J. L., & Baba, T. (2015). Near-complete structural characterization of phosphatidylcholines using electron impact excitation of ions from organics. *Analytical Chemistry*, *87*(11), 5837–5845. https://doi.org/10.1021/acs.analchem.5b01460
54. Williams, P. E., Klein, D. R., Greer, S. M., & Brodbelt, J. S. (2017). Pinpointing double bond and sn-positions in glycerophospholipids via hybrid 193 nm ultraviolet photodissociation (UVPD) mass spectrometry. *Journal of the American Chemical Society*, *139*(44), 15681–15690. https://doi.org/10.1021/jacs.7b06416
55. Costalunga, R., Tshepelevitsh, S., Sepman, H., Kull, M., & Kruve, A. (2022). Sodium adduct formation with graph-based machine learning can aid structural elucidation in non-targeted LC/ESI/HRMS. *Analytica Chimica Acta*, *1204*, 339402. https://doi.org/10.1016/j.aca.2021.339402
56. Palm, E., & Kruve, A. (2022). Machine learning for absolute quantification of unidentified compounds in non-targeted LC/HRMS. *Molecules*, *27*(3), 1013. https://www.mdpi.com/1420-3049/27/3/1013
57. Deutsch, E. W. (2010). Mass spectrometer output file format mzML. *Methods in Molecular Biology (Clifton, N.J.)*, *604*, 319–331. https://doi.org/10.1007/978-1-60761-444-9_22
58. Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.-W., del Toro, N., Pérez-Riverol, Y., Ghali, F., Bandeira, N., … Hermjakob, H. (2014). The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience*. *Molecular, & Cellular Proteomics*, *13*(10), 2765–2775. https://doi.org/10.1074/mcp.O113.036681
59. Hoffmann, N., Rein, J., Sachsenberg, T., Hartler, J., Haug, K., Mayer, G., Alka, O., Dayalan, S., Pearce, J. T. M., Rocca-Serra, P., Qi, D., Eisenacher, M., Perez-Riverol, Y., Vizcaíno, J. A., Salek, R. M., Neumann, S., & Jones, A. R. (2019). mzTab-M: A mzTab-M: A Data standard for sharing quantitative results in mass spectrometry metabolomics. *Analytical Chemistry*, *91*(5), 3302–3310. https://doi.org/10.1021/acs.analchem.8b04310

How to cite this article: Witting, M. (2023). (Re-)use and (re-)analysis of publicly available metabolomics data. *Proteomics*, *23*, e2300032. https://doi.org/10.1002/pmic.202300032

AUTHOR BIOGRAPHY

Michael Witting studied at the Georg-Simon-Ohm University of Applied Sciences in Nuremberg, Germany, conducted his research as PhD student at the Research Unit Analytical BioGeoChemistry, Helmholtz Munich and received his PhD from the Technical University of Munich. Since 2021 he is heading the metabolomics part of the Metabolomics and Proteomics Core, Helmholtz Munich. His primary research interest is the development and application of novel LC-MS/MS methods for targeted and non-targeted metabolomics and lipidomics, with a specific focus on improved metabolite identification.