MAW - The Reproducible Metabolome Annotation Workflow for Untargeted Tandem Mass Spectrometry

Mahnoor Zulfiqar[1*], Luiz Gadelha[1], Christoph Steinbeck[1*], Maria Sorokina[1,2], Kristian Peters[3,4,5*]

[1]Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University, Lessingstrasse 8, 07743 Jena, Germany

[2]Research & Development, Bayer AG, 13353 Berlin, Germany

[3]iDiv - German Centre for Integrative Biodiversity Research, Halle-Jena-Leipzig, Germany

[4]Geobotany and Botanical Gardens, Martin-Luther University of Halle-Wittenberg, 06108 Halle, Germany

[5]Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany

***Correspondence:**
Mahnoor Zulfiqar
mahnoor.zulfiqar@uni-jena.de
Kristian Peters
kristian.peters@ipb-halle.de
Christoph Steinbeck
christoph.steinbeck@uni-jena.de

Author Contributions
Mahnoor Zulfiqar (ORCID: 0000-0002-8330-4071): Development of Metabolome Annotation Workflow (MAW), writing the manuscript, making MAW and relevant data finable, accessible and reproducible, modifying Docker images
Luiz Gadelha (ORCID: 0000-0002-8122-9522): Conceptualizing and making MAW FAIR, making Docker images, parallelizing R functions, revising the draft
Christoph Steinbeck (ORCID: 0000-0001-6966-0814): Project supervision and obtaining the funds, revising the draft
Maria Sorokina (ORCID: 0000-0001-9359-7149): Work Conception, Project Supervision, revising the draft
Kristian Peters (ORCID: 0000-0002-4321-0257): Provision of Bryophytes dataset, Project Supervision, revising the draft

All authors reviewed the manuscript.

Abstract

Mapping the chemical space of compounds to chemical structures remains a challenge in metabolomics. Despite the advancements in untargeted Liquid Chromatography Mass Spectrometry (LCMS) to achieve a high-throughput profile of metabolites from complex biological resources, only a small fraction of these metabolites can be annotated with confidence. Many novel computational methods and tools have been developed to enable chemical structure annotation to known and unknown compounds such as *in-silico* generated spectra and molecular networking. We have developed an automated and reproducible Metabolome Annotation Workflow (MAW) for untargeted metabolomics data implemented in R and Python to assist the metabolic annotation further by combining Tandem Mass Spectrometry (MS^2) input data pre-processing and different computational approaches to give a reliable structure candidate. MAW takes the LC-MS^2 spectra as input and generates a list of putative candidates from spectral and compound libraries. The libraries are integrated via an R package called Spectra and through the CLI (Command-Line-Interface) version of the metabolite annotation tool SIRIUS as part of the R segment of the workflow (MAW-R). The final candidate selection is performed using the cheminformatics tool RDKit in the Python segment (MAW-Py). Furthermore, each feature is assigned a chemical structure and can be imported to a chemical structure similarity network. Following FAIR (Findable, Accessible, Interoperable, Reusable) principles, MAW is available as Docker images. The source code and documentation are available on the GitHub repository zmahnoor14/MAW and the version of the workflow presented in this article has a DOI (10.5281/zenodo.7148450) associated with it.

Keywords: untargeted metabolomics, workflow, tandem mass spectrometry, FAIR, metabolite annotation

1 Introduction

Omics technologies have gained prevalence in the biological domain, over the last few decades. One of the recent developments in omics science is metabolomics, which is a collection of high-throughput analytical platforms, used to represent the metabolic profile of a specimen (Pinu et al., 2019). Among metabolomics approaches, liquid chromatography coupled to mass spectrometry (LC-MS) is the most approved technique for the identification of lower molecular weight molecules (typically below 1000 Da), as it provides a broader coverage of chemically diverse metabolites (Creek et al., 2014). However, this chemical heterogeneity poses a challenge to the number of metabolites that can be identified with metabolomics (Alseekh et al., 2021).

To support such metabolomics challenges, different data sources such as spectral, compound, and pathway databases have been integrated into metabolomics data processing pipelines and routines (Verhoeven et al., 2020). Typically, an LC-MS metabolomics pipeline involves the following steps: (1) conversion of RAW spectral data to an open-source machine-readable format such as .mzML, (2) peak detection, (3) peak alignment, (4) retention time correction, (5) isotope detection, (6) adduct annotation, (7) peak annotation, and (8) data analysis. To acquire more structural information about the samples, tandem mass spectrometry is performed, where two (MS^2) or more (MS^n) analyzers are

combined with a collision chamber which performs fragmentation of the peaks picked in MS^1 (Niessen, 2017). The fragmentation spectra from MS^2 are then used to perform metabolite annotation. To provide a distinction between different levels of confidence for these metabolite annotations, the Metabolomics Standards Initiative (MSI), in 2007, proposed four levels of identification (Sumner et al., 2007) which were later expanded (Schymanski et al., 2014) to five levels. These levels of identification are described in Table 1.

Many software tools and R-packages have been developed for metabolite annotation that use LC-MS^2 as input data. A few of the most popular examples are SIRIUS (Dührkop et al., 2019) and MetFrag (Ruttkies et al., 2016). Some other relatively new tools are MetaboAnnotatoR (Graça et al., 2022), CliqueMS (Senan et al., 2019), CFM-ID 4.0 (Wang et al., 2021), DEREPLICATOR(+) (Mohimani et al., 2018), molDiscovery (Cao et al., 2021), CluMSID (Depke et al., 2019), and MS2Query (Jonge et al., 2022). MS annotation tools generally annotate compounds based on the contents of public databases. Many of these tools are integrated into Metabolomics Workflows such as PhenoMeNal (Peters et al., 2019), Workflow4Metabolomics (W4m) (Giacomoni et al., 2015), and GNPS (Wang et al., 2016). They utilize workflow management systems, such as Galaxy (Afgan et al., 2018), KNIME (Berthold et al., 2009), TidyMass (Shen et al., 2022), or work as independent software such as MS-Dial (Tsugawa et al., 2015), or mzMine (Pluskal et al., 2010) which can be integrated into any Metabolomics pipeline or workflow but currently do not support many of the FAIR (Wilkinson et al., 2016) criteria (see below).

Customization in these metabolomics workflows is a very important aspect due to the variations in the experimental design and source specimen. Annotation, for instance, is also usually a manual process to select the best candidate by going through the candidate list provided by different annotation tools, which takes effort and time. In untargeted metabolomics studies and in phenome centers where large amounts of data are generated repeatedly, it is preferred to run the workflow (Cohen-Boulakia et al., 2017; Wratten et al., 2021) on large-scale data in an automated manner. For such large datasets, reproducibility (Schloss, 2018) can be another issue where the intermediate results, in an easy-to-read and open format, should be tracked to check the final outcome. In other words, the provenance (Herschel et al., 2017) of these intermediate results should be recorded. In addition to reproducibility, best practices defined in the FAIR principles (Goble et al., 2020; Wilkinson et al., 2016) should be employed within workflow systems to improve the Findability, Accessibility, Interoperability, and Reusability of the research objects.

In this work, we introduce a reproducible untargeted metabolomics annotation workflow tool, Metabolome Annotation Workflow (MAW). MAW performs (1) MS^2 data preprocessing, (2) spectral database dereplication, (3) compound database dereplication, (4) candidate selection, and (5) chemical compound classification using various tools integrated into the workflow. It provides a list of top-scoring candidates as a network as well as a final candidate for each feature. MAW follows the FAIR principles and guidelines. It uses open-source databases and tools. The complete MAW source code is available on GitHub (Zulfiqar, 2022a). It is developed in the languages R and Python and can be downloaded with all its functionalities and dependencies as two docker containers for MAW-R and MAW-Py from DockerHub (Zulfiqar, 2022b, 2022c).

2 Methods

2.1 Input Data Pre-requisites

The objective of this workflow is metabolome annotation from the data acquired from Liquid Chromatography Tandem Mass Spectrometry (LC-MS^2) with Data-dependent Acquisition (DDA) mode (Stahl et al., 1996). Raw instrument data should be converted to the open mzML format (Martens et al., 2011). Ideally, the input files with .mzML format should contain MS^1 features such as precursor mass (m/z), retention time, and polarity, along with the MS^1 and MS^2 peaks. The RAW files obtained after data acquisition can be converted to .mzML format using the MS-Convertor from the ProteoWizard suite (Chambers et al., 2012) or Mass Spectrometry File conversion on GNPS. The profile peak data should be converted to centroid peak data during the conversion.

There are two reference datasets used in this study to run MAW. The first dataset consists of standards from a study on hypersalinity in diatoms (Nikitashina et al., 2022). The standards are Betaine, Pipecolinic acid, Cysteinolic acid, Methionine sulfoxide, N,N-dimethylarginine, O-acetyl-L-carnitine, O-propanoyl-L-carnitine, O-butanoyl-L-carnitine, and Isovalerylcarnitine. The second dataset is from a study on nine different Bryophyte species (Peters et al., 2021). The .mzML files of the entire study on Bryophytes are available at MetaboLights repository with ID MTBLS709. Both datasets are obtained using $LCMS^2$ in DDA mode. The mzML files used in this study have been made available on Zenodo - Diatom Standards Dataset (Zulfiqar, 2022d); Bryophytes Dataset (Zulfiqar and Peters, 2022).

To use MS^2 Spectral databases in MAW, we downloaded GNPS, HMDB (Wishart et al., 2007), and MassBank (Horai et al., 2010). These databases were stored as an R object using MsBackend, which is a virtual class defined in the R package Spectra (Rainer et al., 2022) to store and retrieve the mass spectrometry data. The databases are available on Zenodo (Zulfiqar, 2022e).

2.2 Metabolome Annotation Workflow

The Metabolome Annotation Workflow (MAW) is executed in R and Python. Figure 1 illustrates the overview of the Workflow.

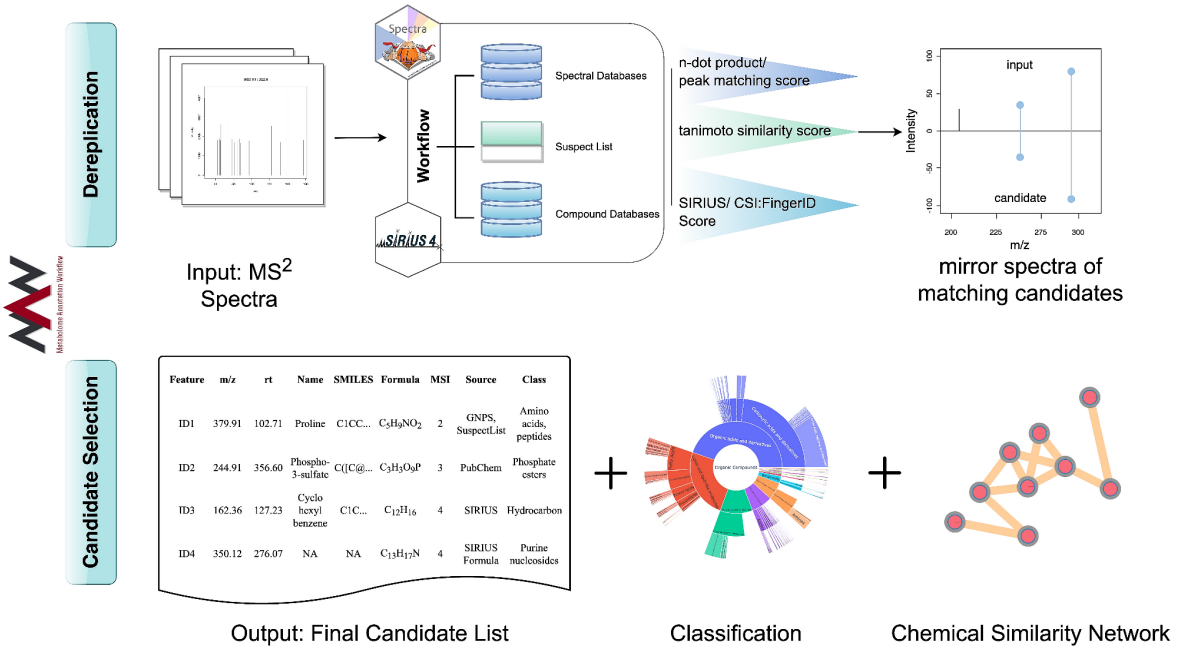


Figure 1: Overview of MAW. MAW-R starts with the input data files and searches the spectra in spectral and compound databases with the R package Spectra and SIRIUS respectively. Once the dereplication is completed and each file and precursor m/z has results from different databases, MAW-Py performs candidate selection from these sources, generates classes, and creates a table for chemical similarity network, which can be visualized in Cytoscape (Shannon et al., 2003).

2.2.1 MAW-R Segment

The workflow starts with the preprocessing of MS^2 spectra using the package Spectra. The function `spec_Processing` in MAW-R, reads the .mzML files and extracts all precursor masses (m/z). This step is common for both spectral and compound database dereplication.

2.2.1.1 Spectral Database Dereplication

To start spectral database dereplication, the `spec_dereplication_file` module is used which is described step-wise in Figure 2. This module requires GNPS, HMDB, and MassBank spectra stored as R objects. The first function within this module is the `spec2_Processing` function which extracts all spectra with a given precursor m/z value, normalizes the intensity, and removes low intensity peaks (<0.05) This function is applied to both query and database candidate spectra so that they are comparable.

For GNPS, the fragment peaks are defined as matching if the difference between their m/z values and the difference between these adjusted m/z values is less than the given tolerance (15 ppm by default). This scoring parameter is set to 0.85 (1.0 being the maximum) as the threshold. For HMDB and MassBank, the default settings are used which is a comparison of the two spectra using the dot product. The default is set to 0.70 (1.0 being the maximum). The top candidates are also given a normalized score of similarity between m/z of individual matching peaks and the intensity of these peaks by function `peakdf`. The ratio of matching peaks by the total peaks is also weighed within these scores.

$$\text{mz_score} = \frac{mz_q \cap mz_d}{n_q + n_d} \qquad \text{(I)}$$

The mz_score is the intersection of the matching peaks among the total number of peaks in the two comparing spectra. $mz_q \cap mz_d$ expression defines the number of matching peaks. n_q is the total number of peaks in the query spectrum and n_d is the total number of matching peaks in the database spectrum.

$$\text{int_score} = \frac{\sum_{i=1}^{n_{qd}} 1-(diff_{qdi} / 100)}{n_{qd}} \qquad \text{(II)}$$

The int_score is the average of the summation of the difference between individual matching peaks defined by $diff_{qdi}$. Since the score is normalized to 1, the $diff_{qdi}$ values are divided by 100 and then subtracted from 1 to give more weightage to peaks with a lesser difference in intensity.

2.2.1.2 Compound Database Dereplication

To use SIRIUS CLI, a parameter file is required as an input for each precursor m/z. The `sirius_param` function is used to generate these parameter files in SIRIUS readable .ms format. The parameter file contains the id of the feature, precursor m/z, charge, median of retention time, isotopic peaks if present, and the MS^1 and MS^2 peaks. In order to extract this information, `ms2_peaks` and `ms1_peaks` functions are required. `ms2_peaks` extracts and combines the fragment peak lists for each precursor m/z. R package CAMERA is used to extract MS^1 peaks and perform isotope annotation with default parameters. To run SIRIUS, the `run_sirius` function is applied based on the isotopic peak annotation. CANOPUS (Dührkop et al., 2021) is also used to add classification. Figure 2 explains the steps for Compound Database Dereplication with SIRIUS.

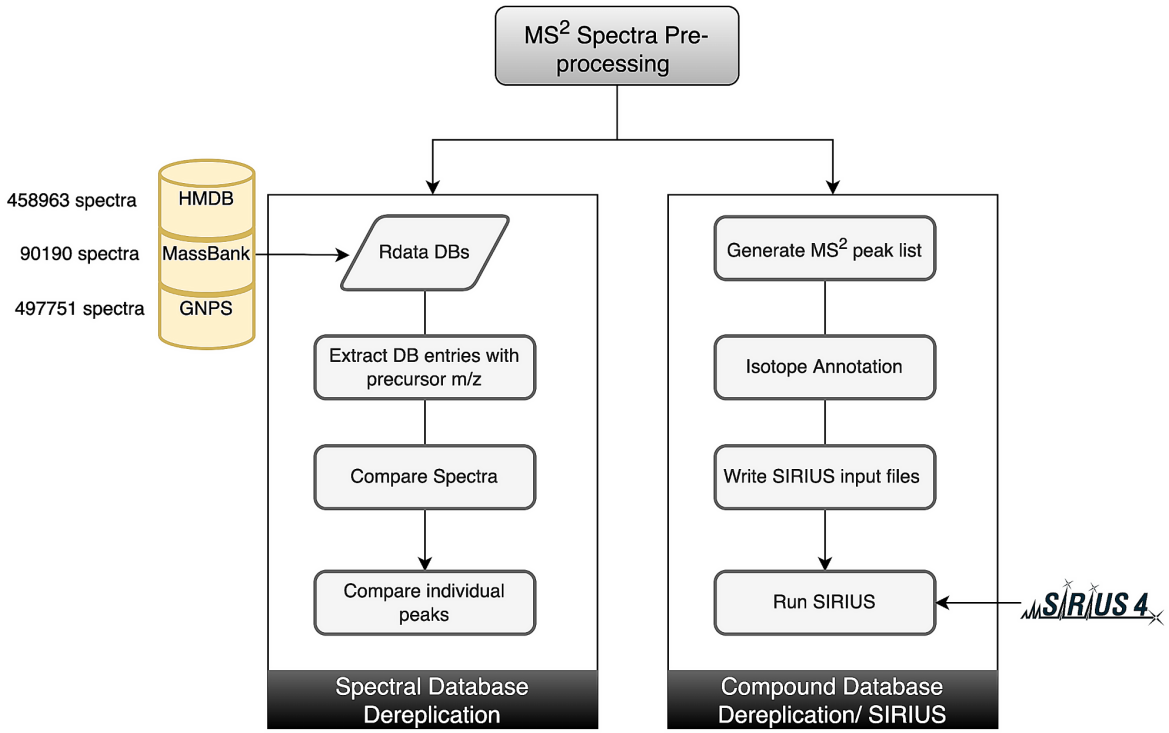


Figure 2: MAW-R performs spectral database dereplication with HMDB, MassBank, and GNPS for each precursor m/z. SIRIUS is used for compound database dereplication with SIRIUS CLI version 4.9.12

2.2.2 MAW-Py Segment

2.2.2.1 Post-processing

The spectra files from GNPS, and HMDB miss some data and metadata; the `spec_postproc` function in MAW-Py adds these data entries. For GNPS, some entries do not have SMILES and have non-conventional names for the compound. `spec_postproc` removes any naming anomalies and adds SMILES using PubChemPy. For HMDB, the results only contain the HMDB database ID. To obtain the names, formulae, and SMILES, structures.sdf file is downloaded from HMDB Downloads. The mz_score, int_score, and the ratio of matching peaks should be >= 0.50. To post-process the results from SIRIUS, a function `sirius_postproc` is defined. This function adds the top-scoring candidates from SIRIUS. Top-scoring candidates are kept for calculation of the Maximum Common Substructure (MCSS). If there were no structure candidates predicted, then the function adds the top-ranking formulae as the candidates.

2.2.2.2 Candidate Selection

The module `CandidateSelection_SimilarityandIdentity` takes results from spectral and compound databases and runs tanimoto similarity score (keeping a threshold of 0.85) to search for a common candidate among the top predicted candidates from all these sources. The results from this chemical similarity calculation can be visualized in the chemical similarity network. This can help in the selection of the most probable candidate structure.

To select one candidate among these top candidates, without manually visualizing the network or going through the list of top candidates, further calculations are performed, with a tanimoto similarity score of 0.99. This calculation selects one candidate that is present among the top candidates from all or a few of the database sources or SIRIUS. In case of a disagreement, we provide a prioritization scheme with results from GNPS given higher priority, as compared to SIRIUS, MassBank, and HMDB. The result is a list of features and corresponding candidates with MSI confidence levels of identification. Figure 3 explains the step-wise processing of candidate selection.

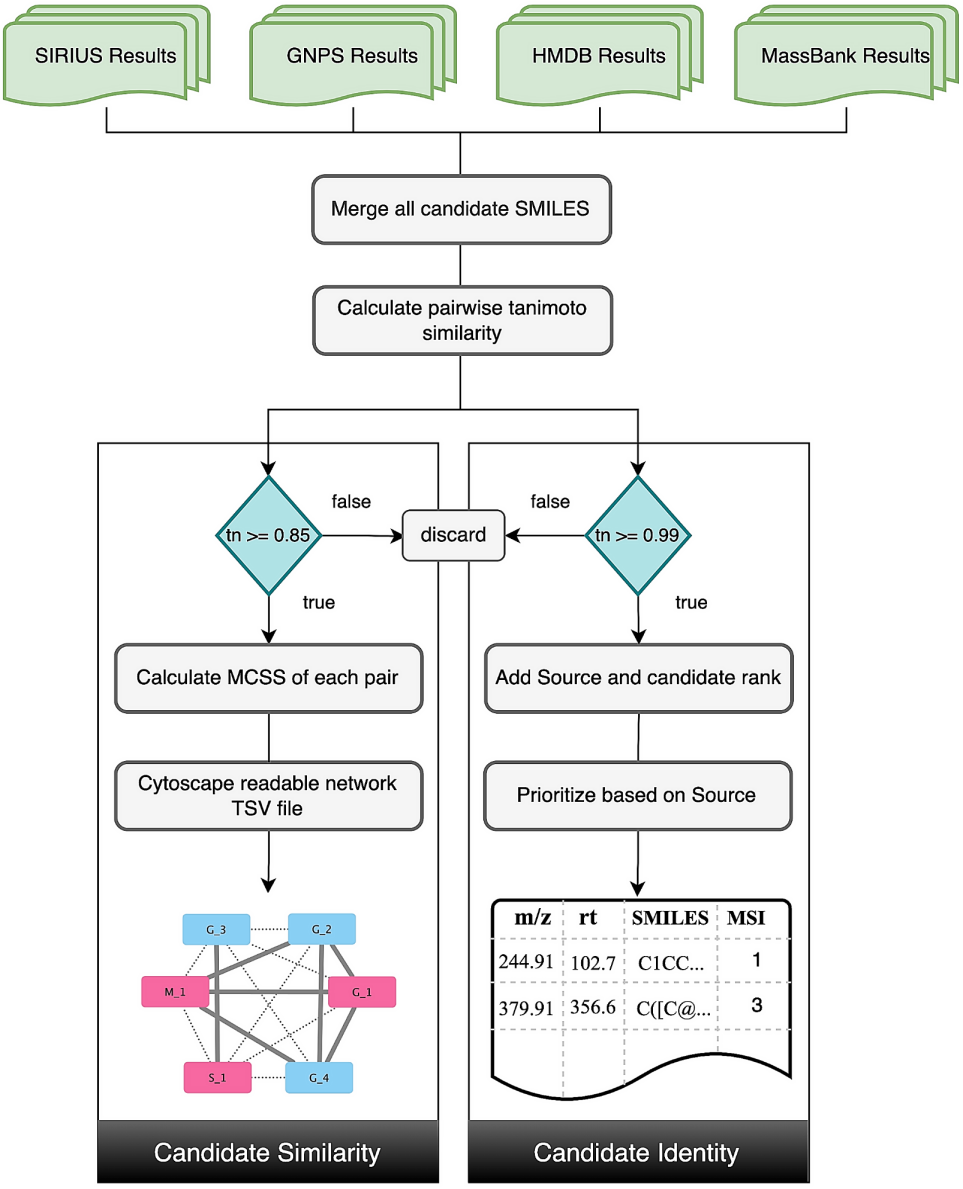


Figure 3: Candidate Selection takes results of top-scoring candidates from all sources and calculates chemical similarity and chemical identity calculation to provide a list of top-scoring candidates that could be the most probable candidate, or one top-scoring candidate per feature, respectively.

2.2.2.3 Compound Classification

Annotation for compound classes can be achieved by a function called `classification`. CANOPUS is a chemical classification prediction tool, integrated with SIRIUS and annotates a chemical class using the $LCMS^2$ spectra. To provide classification to annotations that have a source annotation other than SIRIUS (such as GNPS, HMDB, or MassBank), ClassyFire (Djoumbou Feunang et al., 2016) is integrated into the workflow which takes SMILES as input. Both CANOPUS and ClassyFire are based on ChemOnt Ontology(Djoumbou Feunang et al., 2016). To use ClassyFire in Python, pybatchclassyfire (Allard, 2020) package is used.

2.3 Towards a FAIR and Reproducible MAW

Computational Workflows like MAW can be described as research objects (Bechhofer et al., 2010), to which we can apply the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. They have been followed to make MAW compatible with the latest good practices for scientific data management. To make a workflow reproducible, in addition to following the FAIR principles, there are many proposals and we have implemented the ideas which were more suitable for MAW. MAW is on GitHub, all the source code is openly shared via zmahnoor14/MAW, and a DOI has been provided from Zenodo to the first version of MAW via the MAW GitHub repository. These practices improve findability and accessibility. MAW has separate Docker images available on DockerHub for MAW-R and MAW-Py, this improves interoperability and reuse. In the *Future Developments* subsection, we detail additional practices that will be implemented to further support reproducibility and the FAIR principles.

3 Results

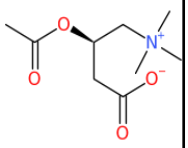
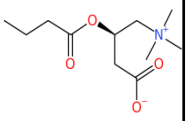
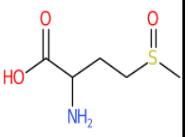
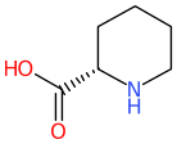
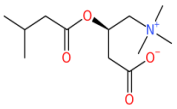
3.1 Annotations from Tandem Mass Spectrometry Data

To evaluate the performance of MAW, we utilized two datasets, a list of standard molecules from the Diatoms and untargeted $LC\text{-}MS^2$ spectra from the Bryophytes dataset. Each high-scoring candidate is given an MSI-Level of confidence. A detailed overview of these MSI-Levels is presented in Table 1.

Annotation-Level	**MSI-Level**
Standard is given (information provided by the user)	1
Candidate present in either of spectral databases or in more than 1 source	2
Candidate only suggested by SIRIUS, and/or has a chemical class prediction	3
Molecular formula predicted	4
Only features measured with MS^2	5

Table 1: MAW operates on MSI-Levels and categorizes annotations based on the rules given above.

The MSI-Levels 1 and 2 represent known compounds that are present in spectral databases and in the case of MSI-Level 1, a standard is utilized to confirm their identification. Hence, results from the first dataset of standards are used to validate the results from MAW. Out of 9 standard metabolites, 6 were correctly identified as the top 1st candidates with MAW as shown in Table 2. Three particular cases in these 9 standards are: (a) the reduced form of betaine was annotated (TrimethylGlycine). (b) Cysteinolic acid wasn't detected in Spectral databases, rather 1-Chlorobenzotriazole was annotated by both GNPS and MassBank. SIRIUS identified Cysteinolic Acid and hence this annotation received the rank 3, according to our current database prioritization scheme, described in the method section. (c) Valeryl L-carnitine was detected as the top candidate by GNPS and MassBank, while the correct compound Isoveleryl-L-carnitine was detected as the top candidate from SIRIUS and the second top candidate by GNPS. Some of the correct matches could have been discarded due to a scoring threshold. For example, Butanoyl Carnitine search in MassBank had no results because the threshold set for MassBank score is 0.70. The correct annotation for Butanoyl Carnitine within MassBank had a score of 0.44, which is way lower than the set threshold.

Precursor m/z	Rt [sec]	Molecular Formula	Molecule	Chemical Class	Name	MSI-Level	Rank (MAW)
204.123	148.99	$C_9H_{17}NO_4$		Fatty Acyls	Acetyl-L-Carnitine	1	1
232.154	102.83	$C_{11}H_{21}NO_4$		Fatty Acyls	Butanoyl-L-Carnitine	1	1
166.053	203.69	$C_5H_{11}NO_3S$		Carboxylic acids and derivatives	Methionine Sulfoxide	1	1
130.086	131.48	$C_6H_{11}NO_2$		Carboxylic acids and derivatives	Pipecolic Acid	1	1
246.170	87.76	$C_{12}H_{23}NO_4$		Fatty Acyls	Isovaleryl L-carnitine	1	2

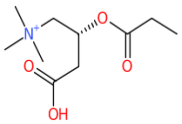
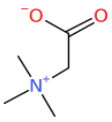
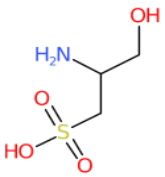
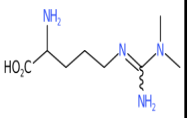
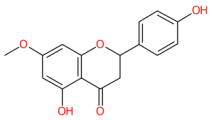
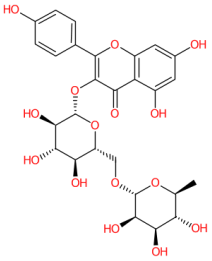
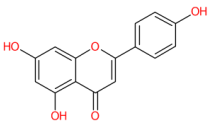
218.138	123.34	$C_{10}H_{20}NO_4$ +		Fatty Acyls	Propanoyl-L-Carnitine	1	1
118.085	145.57	$C_5H_{11}NO_2$		Carboxylic acids and derivatives	Betaine (annotated :Trimethyglycine)	1	1
154.018	155.28	$C_3H_9NO_4S$		Organic sulfonic acids and derivatives	Cysteinolic Acid	1	3
203.150	348.67	$C_8H_{18}N_4O_2$		Carboxylic acids and derivatives	N,N-dimethyl arginine	1	1

Table 2: List of standards from the diatom dataset and top candidates from MAW

For the Bryophytes datasets, earlier study annotated 48 features with chemical structure and the rest were annotated with a chemical class using MetFamily classifier. With MAW, we used only the MS^2 spectra from this study compiled in one MGF file available on Zenodo. MAW identified 887 out of 933 features (keeping default thresholds for Spectral Databases and with no thresholds for SIRIUS). A few 1st ranking annotations are given in table 3.

Precursor m/z	Rt [sec]	Molecular Formula	Molecule	Chemical Class	Name	MSI-Level
165.055	47.16	$C_{16}H_{14}O_5$		Flavonoids	(RS)-Sakuranetin	2
287.056	318.26	$C_{27}H_{30}O_{15}$		Flavonoids	Nicotiflorin	2
271.062	380.48	$C_{15}H_{10}O_5$		Flavonoids	Apigenin	2

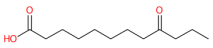
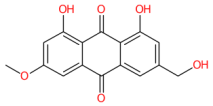
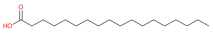
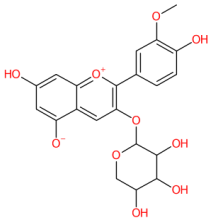
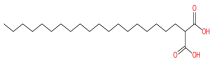
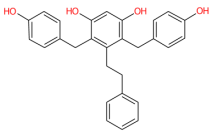
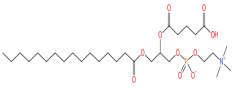
213.148	608.26	$C_{12}H_{22}O_3$		Fatty Acyls	9-oxododecanoic acid	3
255.065	517.52	$C_{16}H_{12}O_6$		Anthracenes	Phallacinol	2
283.263	942.50	$C_{18}H_{36}O_2$		Fatty Acyls	octadecanoic acid	3
301.071	430.43	$C_{21}H_{20}O_{10}$		Flavonoids	Peonidin-3-O-alpha-arabinoside	2
369.297	902.84	$C_{22}H_{42}O_4$		Fatty Acyls	Eikosandicarbonsaure	3
425.176	607.49	$C_{28}H_{26}O_4$		Phenols	Shancigusin B	3
544.339	662.92	$C_{29}H_{56}NO_{10}P$		Glycerophospholipids	1-Palmitoyl-2-glutaryl phosphatidylcholine	2

Table 3: Describes some of the features and annotations in the Bryophytes dataset.

CANOPUS and ClassyFire (ChemONT ver 2.1) were used to analyze the chemical diversity in the MS^2 spectra of the Bryophytes dataset in MAW. The most prevalent superclasses are Lipids, such as Glycerophospholipids for the stability of membranes; Organic Acids representing amino acids; class of Flavonoids that providing protection against the Reactive Oxygen Species (ROS); and Benzenoids for response against biotic and abiotic stress. MAW annotated many Flavonoids with high MSI-Levels. The classification results are in agreement with the results from the previous study obtained using the MetFamily classifier (ChemONT ver 2.1). Figure 5 explains in detail the annotated superclasses, classes, and subclasses identified in MAW.

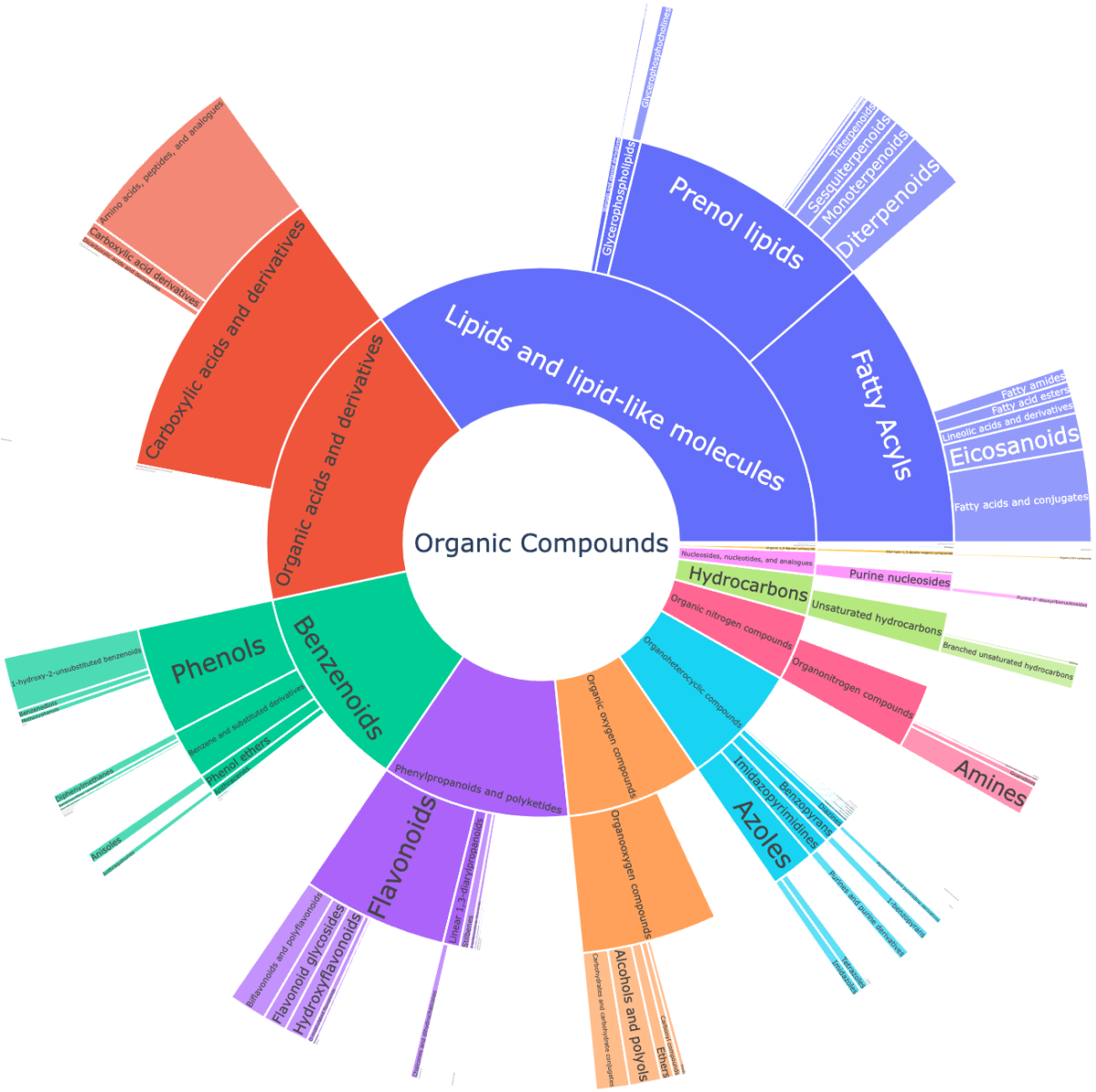


Figure 5: MAW integrates CANOPUS and ClassyFire to annotate chemical classes to features. For the Bryophytes study, the most prevalent subclasses were Lipids (34.6%), Organic acids (16.8%), Benzoids (12.6%), and Phenylpropanoid and Polyketides (10.8%).

3.2 Comparison of MAW with integrated annotation tools using the Standard Dataset

We evaluated MAW's performance in comparison to annotation tools integrated within MAW such as SIRIUS, and Spectra (GNPS, Massbank, and HMDB), using the Standards dataset. MAW identified the correct compounds except for Betaine. Isovaleryl carnitine ranked second and Cysteinolic acid ranked third, as mentioned in Table 2. In the case of using only SIRIUS, fatty acyls (except for Butanoyl Carnitine) were annotated as an isomer or similar structure instead of the actual compound. SIRIUS correctly annotated methionine sulphoxide, N,N-dimethylarginine, and carboxylic acid and its derivatives. GNPS correctly annotated all compounds except for Betaine and Cysteinolic acid. The lowest performances

were shown by MassBank and HMDB. MassBank only annotated N,N-dimethyl arginine, and Acetyl Carnitine with a score threshold of 0.70, and HMDB failed to annotate any correct compound with the same score threshold. We also performed annotations for standards with MetFrag CLI version 2.5.0 and PubChem as the search database. Table 4 gives an overview of the annotations obtained by MAW, Spectra, SIRIUS, and MetFrag.

Annotation-Level	**MAW**	**Spectra (GNPS, HMDB, MassBank)**	**SIRIUS**	**MetFrag (PubChem)**
Total Structural annotations	100% (37)	75.7% (28)	70.2% (26)	21.6%(8)
Chemical Class Prediction	97.3% (36)	-	100% (37)	-
Formula Prediction	100% (37)	-	100% (37)	-

Table 4: Explains the percentage of annotated features in MAW, SIRIUS and Spectra. Total Structural annotations refer to the percentage of annotated compounds that are the same in MAW, Spectra, and SIRIUS. We ran MetFrag and among the top 100 candidates from each feature, 8 features had the same Annotation as MAW. For Chemical classification, MAW annotated 97.3% of the features with a chemical class and all features had a molecular formula prediction.

3.3 Case Study with the flavonoid Nicotiflorin from the Bryophytes Dataset

In recent years ecologically important organisms and metabolites are being extensively researched in the field of ecological metabolomics (Peters et al., 2018). Bryophytes are being actively investigated regarding their unique diversity of natural products. In the chemodiversity study of Bryophytes, many of the features remained unannotated with a chemical structure due to the fact that bryophytes have not received much research as compared to other organism groups such as plants and model organisms such as *Arabidopsis thaliana* (Asakawa et al., 2013). With MAW, we have annotated 887 features with chemical structures.

Here we describe the annotation and evaluation of the flavonoid Nicotiflorin using MAW. The feature ID is file_1M287R318ID319 where File_1 is the file ID, M287 represents the m/z ratio and R318 represents the median retention time in seconds. ID is an individual number assigned to the features in File_1. For ease of understanding, we will refer to this feature as M287R318. The detailed features are provided in Table 5.

Features	**Annotation**	**MSI-Level**
Precursor m/z	287.055902093109	5
Rt (median) [seconds]	318.263296552812	5
Polarity	Negative	5
Molecular Formula	$C_{27}H_{30}O_{15}$	4
Chemical Class	Flavonoids (CHEMONTID:0000334)	3
Final_Candidate_SMILES	c(c5)(O)cc(c(c51)C(C(O[C@H](O3)[C@@H]([C@H]([C@@H]([C@H]3CO[C@@H]([C@@H]4O)O[C@H]([C@@H]([C@H]4O)O)C)O)O)O)=C(c(c2)ccc(O)c2)O1)=O)O	2
Sources	GNPS, HMDB, MassBank	2

Table 5: Features and Annotation for feature ID M287R318 with associated MSI-Level.

MAW gives annotation results in two ways. The first is to generate a chemical similarity network between all top-scoring candidates. MAW generates a TSV file, for each feature, that contains the nodes as candidates and tanimoto similarity scores as edges. Figure 5 shows the chemical similarity network of candidates for M287R318 created with Cytoscape. This visual representation helps in candidate selection by choosing the cluster with the highest number of sources (GNPS, MassBank, HMDB, SIRIUS) and the highest-ranking candidate from these sources.

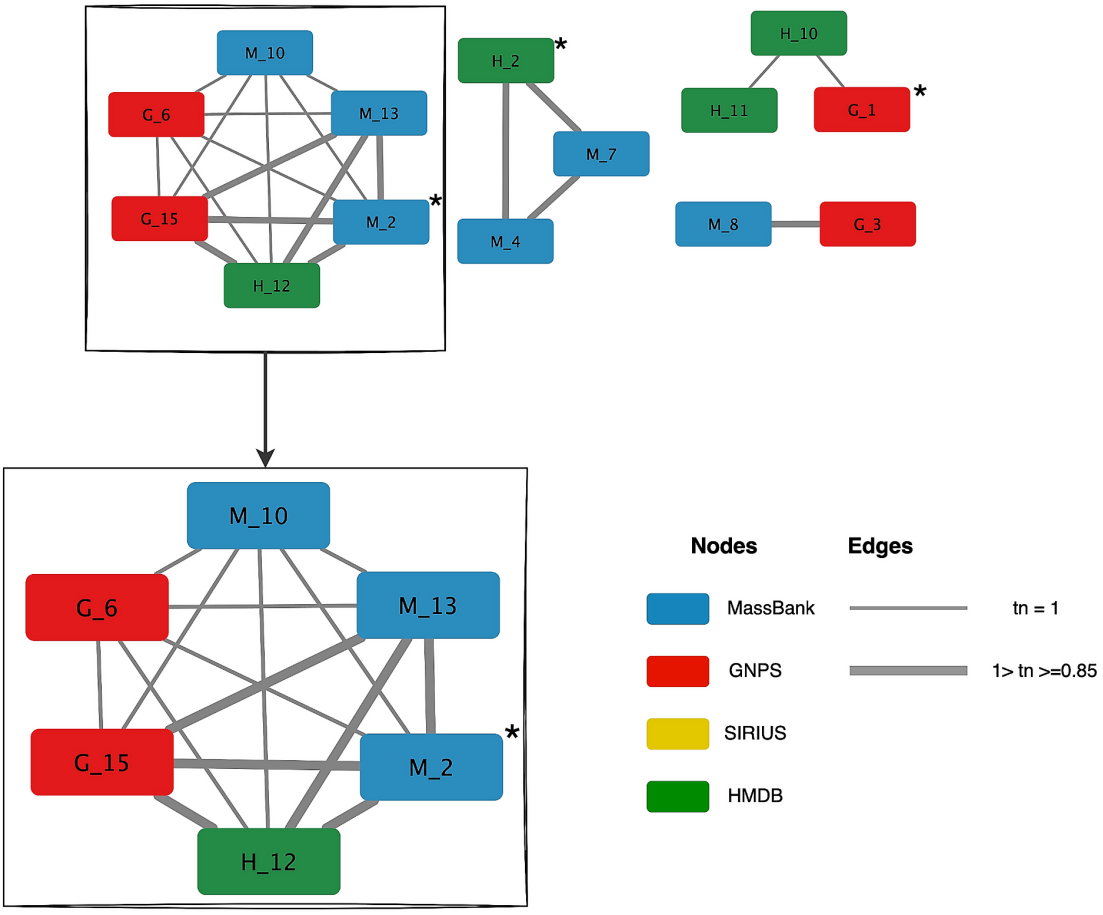


Figure 5: This network represents the chemical structure similarity among the top candidates from GNPS, MassBank, and HMDB (no annotations from SIRIUS are recorded for this feature). The nodes represent a chemical structure (as candidates). The edges represent the tanimoto similarity score (tn) between the two nodes. The asterisk (*) shows the top-ranking candidate from each source that is present in the network. The network with most nodes is zoomed in to visualize that G_15, H_12, M_2, and M_13 have a structure with tanimoto similarity score equal to 1 and hence it is most probable to consider this structure to be annotated to the feature M287R318.

The second way is to select only one highly probable candidate for each feature. Here the chemical similarity network selects the most probable cluster which in this case is presented in Figure 5. The cluster represents 3 sources (spectral databases). This gives a higher probability to the structure in that cluster and can be assigned as the final candidate, in this case, Nicotiflorin. Using the package Spectra, a mirror image of MS^2 Spectra both from the query and the top candidate from either of the spectral databases can be created. An example is given in Figure 6.

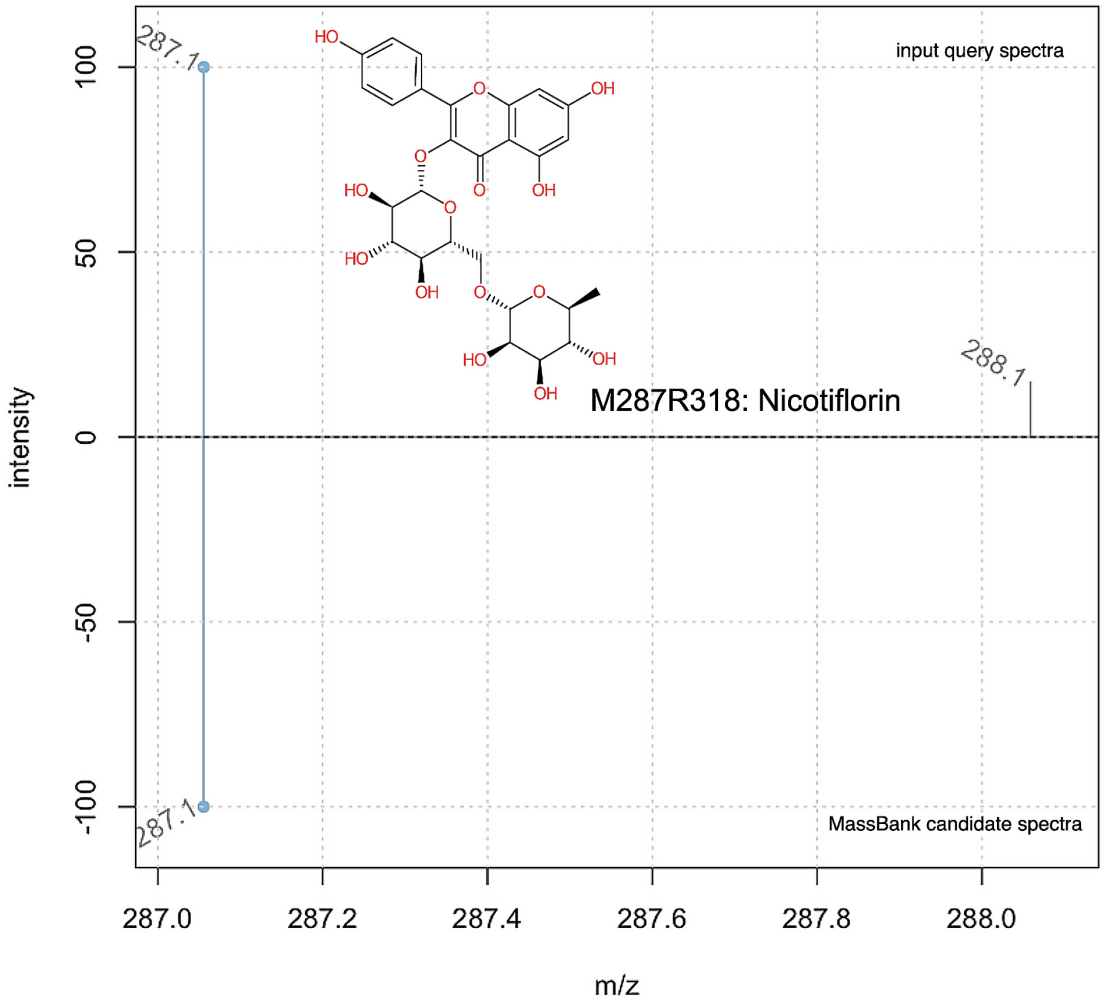


Figure 6: Mirror Spectra of query spectra and M_2 (2nd top candidate from MassBank). The (x, y) plot represents the query spectra with two MS^2 Peaks. The x-axis represents the m/z values of the MS^2 fragment peaks and the y-axis represents the intensity. The (x, -y) plot represents M_2 spectra.

4 Discussion

The described workflow is an automated way to annotate untargeted tandem mass spectrometry metabolomics large-scale datasets. MAW integrates candidate results from spectral and compound databases to annotate each feature according to the defined MSI-Levels which provides an overview of the chemical diversity of a given dataset.

4.1 Evaluation of MAW

Many software suites and R packages have been developed over years to solve the metabolomics annotation challenge, but it still remains the “bottleneck” in metabolomics due

to an infinite chemical dark space (da Silva et al., 2015). MAW is specifically developed to aid the annotation process by automation of candidate selection within this space.

Computational tools such as MetFrag, SIRIUS(CSI:FingerID), and CFM-ID, generate *in silico* spectra of chemical structures from different compound databases based on their specified algorithms. This *in silico* spectral matching provides an annotation of up to MSI-Level 3. However, to provide more confidence in the annotation, the input spectra should be screened against actual experimental spectra. For this purpose, some tools and R-packages have been developed such as Spectra, matchms (Huber et al., 2020), MS2Query (specifically for Analogues), GNPS MASST, which use modified cosine similarity score or deep learning based score to match the MS^2 query spectra against spectral databases. However, there is a huge mismatch between submitted experimental spectra to spectral databases and chemical structures present in compound databases. This leads to a low number of annotated features. MAW combines both *in silico* generated spectral matching and experimental spectral matching to increase the number of confident annotated metabolites, together with an associated MSI-Level. As described in Table 4, the percentage of annotated metabolites with MAW is higher than spectral or compound database approaches alone.

Another advantage of MAW is the ease of candidate selection which allows one to choose the most optimal results. As there are both spectral and compound databases being used, an automated candidate selection is very important. Generally, annotation tools provide a list of ranked molecules based on the associated scoring schemes and the user has to manually go through these candidates or just take the first candidate. This requires a lot of time and effort with big datasets. MAW not only provides these ranked lists but also provides a concise list of highly probable candidates from different databases. The structural similarity between these candidates can be visualized as a chemical similarity network. A function to get only one candidate per feature is also provided. All the intermediate results are stored in an efficiently designed directory system, so the user can always refer back to the origin of the candidate.

A major component behind candidate selection is the scoring scheme followed throughout the workflow. For this particular study and data, we used the scoring schemes provided by Spectra and SIRIUS (CSI:FingerID); and for structural matching, we used the Tanimoto similarity score. The first scoring scheme is the spectral database scores used within the Spectra package. Spectra uses the most common spectral matching metric which is the cosine similarity score that calculates how much overlap is observed between two spectra (Kim et al., 2022). For HMDB and MassBank, this metric is used unaltered. Some good candidates are missed due to lower scores with high spectral matching as in the case of Butanoyl Carnitine and MassBank. For GNPS, this similarity score is altered as described in the method section and the spectral matching with GNPS gives better results as compared to the other two databases. We conclude that spectral matching needs better metrics. The possible cases where cosine similarity metrics fail could be different fragmentation under different experimental setups, or if there is a shift in the fragmentation peaks, the overlap between query and candidate spectra would be low. To deal with this problem in a better way, we also consider the individual peak m/z and individual peak intensity overlap, also counting in the number of matching peaks. All of these MAW-defined metrics should atleast have a 0.5 matching score in addition to a high cosine similarity score. In this way, we could

lower the possibility of getting any false positives but could end up missing candidates with low cosine similarity scores as false negative candidates. This could be solved by integrating Spec2Vec in a future version of MAW. Spec2Vec is a deep learning-based spectral matching score and has shown better spectral matching results as compared to simple or modified cosine similarity scores (Huber et al., 2021). Also, the public spectral databases are not completely standardized and hence can only annotate spectra present in the databases such as in the case of Betaine, where reduced state Trimethylglycine is annotated.

Another way to reduce the false negative results from spectral database screening is to use compound databases to compensate for all the known compounds which were not assigned a candidate from spectral databases. Here, MetFrag would only be a good option when using a more concise library of molecules, as it gives better results with fewer candidate possibilities. Larger databases such as PubChem can lead to a bigger list of candidates with the actual structure being at a lower rank (e.g: top 100th candidate), such as seen with candidates from Diatom Standards explained in Table 4. So, for the first version of MAW, we have only integrated SIRIUS (CSI:FingerID) for compound databases, a state-of-the-art annotation tool (Stravs et al., 2022) into MAW. For features with no structural annotations, SIRIUS and CANOPUS provide a molecular formula and a chemical class. Thresholds for ranking SIRIUS candidates were considered, however, SIRIUS CLI doesn't provide a deterministic probability score, so providing a threshold would differ from case to case and hence the idea was discarded. Also, the annotation of lipid classes using SIRIUS can lead to false positives due to the complex nature of lipids, and many lipid isomers co-elute in mass spectrometry, making it a challenge to annotate lipids using *in silico* libraries (Koelmel et al., 2020). In the current version of MAW, the database “all” is used in SIRIUS which is a combination of many databases such as PubChem, HMDB, and COCONUT.

All these functionalities make MAW a single infrastructure to provide an overview of the chemical diversity present in any dataset. It also comes with functions to generate graphs to visualize the chemical similarities in a network or the chemical classification in a sunburst diagram. MAW is a FAIRified Workflow that provides access to all the datasets, and libraries used for the development of MAW. The workflow can be operated on any system using two docker containers that confirm interoperability and reusability. The intermediate results can also be tracked and the steps are customized specifically for the annotation step of the metabolomics workflow.

4.2 Future Developments

The future version of MAW would be integrated with Spec2Vec and matchms to enhance the spectral database matching, along with MS2Query for providing structurally similar analogs. We also plan to add MetFrag with PubChemLite and COCONUT in the future to enhance the automated candidate selection. We would also add the option of more database selection in SIRIUS. Another important function to add is the spectral similarity-based Classical Molecular Network (MN) among single-condition datasets or between two conditions and linking these MNs to pathway annotations. We also want to make MAW more user-friendly and to further support the FAIR principles by, for instance, providing a specification of the workflow in the Common Workflow Language (Crusoe et al., 2021),, gathering detailed

provenance of workflow executions, and adding rich metadata about the workflow, its components, and the data used and produced by it.

Conclusion

Structural annotation remains a complicated and time-consuming process. MAW is an automated metabolite annotation workflow. Using both spectral and compound databases, this workflow helps to annotate the query spectral data to a corresponding experimental tandem mass spectra or to an *in silico* generated corresponding spectra in case there are no associated spectra present in spectral databases. Results of annotated metabolites from MAW can be further functionally mapped to metabolic pathways to aid in the structure prediction compounds and can be integrated with other omics data to understand the biological and ecological significance of these metabolites. This could greatly assist metabolite characterization in diverse fields such as clinical metabolomics and natural product discovery.

Availability and Requirements

MAW:

- MAW can be installed with all its functions and dependencies via two docker images submitted to DockerHub (Zulfiqar, 2022b, 2022c).
- Tutorial: Detailed Tutorial for MAW is available on GitHub Wiki, with associated Jupyter Notebooks present on GitHub.
- Source Code + MAW version: The Source Code is available on GitHub and MAW 1.0.0 version is available on Zenodo with the DOI: 10.5281/zenodo.7148450

MS Dataset

- Standards: The LCMS2 standards data generated by Vera Nikitashina are present on the MetaboLights repository with ID: MTBLS3177. These standards are also uploaded to Zenodo as mzML files with DOI: 10.5281/zenodo.7106205
- Bryophytes: The Bryophytes dataset is also available on MetaboLights with the ID: MTBLS709. The MS^2 data file used in this study is available on Zenodo with DOI: 10.5281/zenodo.7107096
- Spectral Databases as Rdata: The currently available versions of the GNPS, HMDB, and MassBank are present on Zenodo with DOI: 10.5281/zenodo.6528931

Acknowledgments

Georg Pohnert (ORCID: 0000-0003-2351-6336) Lab for providing the data for *S. marinoi* dataset, especially **Vera Nikitashina**
Ivo Große Lab for initiating the idea for candidate selection
Kohulan Rajan (ORCID: 0000-0003-1066-7792) for the MAW logo, cleaning code, and helping with setting up the GitHub repository.
Adelene Lai (ORCID: 0000-0002-2985-6473) and Jonas Schaub (ORCID: 0000-0003-1554-6666) for ideas on candidate selection function

Funding Information

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2051 - Project-ID 390713860

Competing Interests

The authors have no competing interests to declare.

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379

Allard, P.-M., 2020. pybatchclassyfire [WWW Document]. GitLab. URL https://gitlab.unige.ch/Pierre-Marie.Allard/pybatchclassyfire (accessed 9.22.22).

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., C. 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., Siuzdak, G., Skirycz, A., Sumner, L.W., Snyder, M.P., Tang, H., Tohge, T., Wang, Y., Wen, W., Wu, S., Xu, G., Zamboni, N., Fernie, A.R., 2021. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat. Methods 18, 747–756. https://doi.org/10.1038/s41592-021-01197-1

Asakawa, Y., Ludwiczuk, A., Nagashima, F., 2013. Phytochemical and biological studies of bryophytes. Phytochemistry, Meinhart H. Zenk Memorial Issue 91, 52–80. https://doi.org/10.1016/j.phytochem.2012.04.012

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I., 2010. Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nat. Preced. 1–1. https://doi.org/10.1038/npre.2010.4626.1

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. KNIME - the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explor. Newsl. 11, 26–31. https://doi.org/10.1145/1656274.1656280

Cao, L., Guler, M., Tagirdzhanov, A., Lee, Y.-Y., Gurevich, A., Mohimani, H., 2021. MolDiscovery: learning mass spectrometry fragmentation of small molecules. Nat. Commun. 12, 3718. https://doi.org/10.1038/s41467-021-23986-0

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P., 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. 30, 918–920. https://doi.org/10.1038/nbt.2377

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Le Bras, Y., Lemoine, F., Mareuil, F., Ménager, H., Pradal,

C., Blanchet, C., 2017. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Gener. Comput. Syst. 75, 284–298. https://doi.org/10.1016/j.future.2017.01.012
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
Crusoe, M., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanić, N., Ménager, H., Soiland-Reyes, S., Gavrilović, B., Goble, C., Community, T., 2021. Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. Commun. ACM 65. https://doi.org/10.1145/3486897
da Silva, R.R., Dorrestein, P.C., Quinn, R.A., 2015. Illuminating the dark matter in metabolomics. Proc. Natl. Acad. Sci. 112, 12549–12550. https://doi.org/10.1073/pnas.1516878112
Depke, T., Franke, R., Brönstrup, M., 2019. CluMSID: an R package for similarity-based clustering of tandem mass spectra to aid feature annotation in metabolomics. Bioinformatics 35, 3196–3198. https://doi.org/10.1093/bioinformatics/btz005
Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R., Wishart, D.S., 2016. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J. Cheminformatics 8, 61. https://doi.org/10.1186/s13321-016-0174-y
Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat. Methods 16, 299–302. https://doi.org/10.1038/s41592-019-0344-8
Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M.A., Petras, D., Gerwick, W.H., Rousu, J., Dorrestein, P.C., Böcker, S., 2021. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat. Biotechnol. 39, 462–471. https://doi.org/10.1038/s41587-020-0740-8
Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., Goulitquer, S., Thévenot, E.A., Caron, C., 2015. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 31, 1493–1495. https://doi.org/10.1093/bioinformatics/btu813
Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M.R., Peters, K., Schober, D., 2020. FAIR Computational Workflows. Data Intell. 2, 108–121. https://doi.org/10.1162/dint_a_00033
Graça, G., Cai, Y., Lau, C.-H.E., Vorkas, P.A., Lewis, M.R., Want, E.J., Herrington, D., Ebbels, T.M.D., 2022. Automated Annotation of Untargeted All-Ion Fragmentation LC–MS Metabolomics Data with MetaboAnnotatoR. Anal. Chem. 94, 3446–3455. https://doi.org/10.1021/acs.analchem.1c03032
Herschel, M., Diestelkämper, R., Ben Lahmar, H., 2017. A survey on provenance: What for? What form? What from? VLDB J. 26, 881–906. https://doi.org/10.1007/s00778-017-0486-1
Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, Kenichi, Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, Ken, Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714. https://doi.org/10.1002/jms.1777
Huber, F., Ridder, L., Verhoeven, S., Spaaks, J.H., Diblen, F., Rogers, S., Hooft, J.J.J. van der, 2021. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLOS Comput. Biol. 17, e1008724. https://doi.org/10.1371/journal.pcbi.1008724

Huber, F., Verhoeven, S., Meijer, C., Spreeuw, H., Castilla, E.M.V., Geng, C., Hooft, J.J. j van der, Rogers, S., Belloum, A., Diblen, F., Spaaks, J.H., 2020. matchms - processing and similarity evaluation of mass spectrometry data. J. Open Source Softw. 5, 2411. https://doi.org/10.21105/joss.02411

Jonge, N.F. de, Louwen, J.R., Chekmeneva, E., Camuzeaux, S., Vermeir, F.J., Jansen, R.S., Huber, F., Hooft, J.J.J. van der, 2022. MS2Query: Reliable and Scalable MS2 Mass Spectral-based Analogue Search. https://doi.org/10.1101/2022.07.22.501125

Kim, S., Kato, I., Zhang, X., 2022. Comparative Analysis of Binary Similarity Measures for Compound Identification in MassSpectrometry-Based Metabolomics. Metabolites 12, 694. https://doi.org/10.3390/metabo12080694

Koelmel, J.P., Li, X., Stow, S.M., Sartain, M.J., Murali, A., Kemperman, R., Tsugawa, H., Takahashi, M., Vasiliou, V., Bowden, J.A., Yost, R.A., Garrett, T.J., Kitagawa, N., 2020. Lipid Annotator: Towards Accurate Annotation in Non-Targeted Liquid Chromatography High-Resolution Tandem Mass Spectrometry (LC-HRMS/MS) Lipidomics Using a Rapid and User-Friendly Software. Metabolites 10, 101. https://doi.org/10.3390/metabo10030101

Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L.-F., Dorrestein, P.C., Pevzner, P.A., 2018. Dereplication of microbial metabolites through database search of mass spectra. Nat. Commun. 9, 4035. https://doi.org/10.1038/s41467-018-06082-8

Niessen, W.M.A., 2017. MS–MS and MSn☆, in: Lindon, J.C., Tranter, G.E., Koppenaal, D.W. (Eds.), Encyclopedia of Spectroscopy and Spectrometry (Third Edition). Academic Press, Oxford, pp. 936–941. https://doi.org/10.1016/B978-0-12-409547-2.05219-7

Nikitashina, V., Stettin, D., Pohnert, G., 2022. Metabolic adaptation of diatoms to hypersalinity. Phytochemistry 201, 113267. https://doi.org/10.1016/j.phytochem.2022.113267

Peters, K., Balcke, G., Kleinenkuhnen, N., Treutler, H., Neumann, S., 2021. Untargeted In Silico Compound Classification—A Novel Metabolomics Method to Assess the Chemodiversity in Bryophytes. Int. J. Mol. Sci. 22, 3251. https://doi.org/10.3390/ijms22063251

Peters, K., Bradbury, J., Bergmann, S., Capuccini, M., Cascante, M., de Atauri, P., Ebbels, T.M.D., Foguet, C., Glen, R., Gonzalez-Beltran, A., Günther, U.L., Handakas, E., Hankemeier, T., Haug, K., Herman, S., Holub, P., Izzo, M., Jacob, D., Johnson, D., Jourdan, F., Kale, N., Karaman, I., Khalili, B., Emami Khonsari, P., Kultima, K., Lampa, S., Larsson, A., Ludwig, C., Moreno, P., Neumann, S., Novella, J.A., O'Donovan, C., Pearce, J.T.M., Peluso, A., Piras, M.E., Pireddu, L., Reed, M.A.C., Rocca-Serra, P., Roger, P., Rosato, A., Rueedi, R., Ruttkies, C., Sadawi, N., Salek, R.M., Sansone, S.-A., Selivanov, V., Spjuth, O., Schober, D., Thévenot, E.A., Tomasoni, M., van Rijswijk, M., van Vliet, M., Viant, M.R., Weber, R.J.M., Zanetti, G., Steinbeck, C., 2019. PhenoMeNal: processing and analysis of metabolomics data in the cloud. GigaScience 8. https://doi.org/10.1093/gigascience/giy149

Peters, K., Worrich, A., Weinhold, A., Alka, O., Balcke, G., Birkemeyer, C., Bruelheide, H., Calf, O.W., Dietz, S., Dührkop, K., Gaquerel, E., Heinig, U., Kücklich, M., Macel, M., Müller, C., Poeschl, Y., Pohnert, G., Ristok, C., Rodríguez, V.M., Ruttkies, C., Schuman, M., Schweiger, R., Shahaf, N., Steinbeck, C., Tortosa, M., Treutler, H., Ueberschaar, N., Velasco, P., Weiß, B.M., Widdig, A., Neumann, S., Dam, N.M. van, 2018. Current Challenges in Plant Eco-Metabolomics. Int. J. Mol. Sci. 19, 1385. https://doi.org/10.3390/ijms19051385

Pinu, F.R., Goldansaz, S.A., Jaine, J., 2019. Translational Metabolomics: Current Challenges and Future Opportunities. Metabolites 9, 108. https://doi.org/10.3390/metabo9060108

Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395.

Rainer, J., Vicini, A., Salzer, L., Stanstrup, J., Badia, J.M., Neumann, S., Stravs, M.A., Verri Hernandes, V., Gatto, L., Gibb, S., Witting, M., 2022. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. Metabolites 12, 173. https://doi.org/10.3390/metabo12020173

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J. Cheminformatics 8, 3. https://doi.org/10.1186/s13321-016-0115-9

Schloss, P.D., 2018. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. mBio 9, e00525-18. https://doi.org/10.1128/mBio.00525-18

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. Environ. Sci. Technol. 48, 2097–2098. https://doi.org/10.1021/es5002105

Senan, O., Aguilar-Mogas, A., Navarro, M., Capellades, J., Noon, L., Burks, D., Yanes, O., Guimerà, R., Sales-Pardo, M., 2019. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. Bioinformatics 35, 4089–4097. https://doi.org/10.1093/bioinformatics/btz207

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. https://doi.org/10.1101/gr.1239303

Shen, X., Yan, H., Wang, C., Gao, P., Johnson, C.H., Snyder, M.P., 2022. TidyMass an object-oriented reproducible analysis framework for LC–MS data. Nat. Commun. 13, 4365. https://doi.org/10.1038/s41467-022-32155-w

Stahl, D.C., Swiderek, K.M., Davis, M.T., Lee, T.D., 1996. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. J. Am. Soc. Mass Spectrom. 7, 532–540. https://doi.org/10.1016/1044-0305(96)00057-8

Stravs, M.A., Dührkop, K., Böcker, S., Zamboni, N., 2022. MSNovelist: de novo structure generation from mass spectra. Nat. Methods 19, 865–870. https://doi.org/10.1038/s41592-022-01486-3

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.N., Lindon, J.C., Marriott, P., Nicholls, A.W., Reily, M.D., Thaden, J.J., Viant, M.R., 2007. Proposed minimum reporting standards for chemical analysis. Metabolomics 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., Arita, M., 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat. Methods 12, 523.

Verhoeven, A., Giera, M., Mayboroda, O.A., 2020. Scientific workflow managers in metabolomics: an overview. Analyst 145, 3801–3808. https://doi.org/10.1039/D0AN00272K

Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., Wishart, D.S., 2021. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. Anal. Chem. https://doi.org/10.1021/acs.analchem.1c01465

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., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom,

A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. 34, 828–837. https://doi.org/10.1038/nbt.3597

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.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., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., Querengesser, L., 2007. HMDB: the Human Metabolome Database. Nucleic Acids Res. 35, D521-526. https://doi.org/10.1093/nar/gkl923

Wratten, L., Wilm, A., Göke, J., 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat. Methods 18, 1161–1168. https://doi.org/10.1038/s41592-021-01254-9

Zulfiqar, M., 2022a. zmahnoor14/MAW. Metabolome Annotation Workflow.

Zulfiqar, M., 2022b. zmahnoor/maw-r [WWW Document]. Docker Image MAW-R. URL https://hub.docker.com/r/zmahnoor/maw-r (accessed 9.22.22).

Zulfiqar, M., 2022c. zmahnoor/maw-py [WWW Document]. Docker Image MAW-Py. URL https://hub.docker.com/r/zmahnoor/maw-py (accessed 9.22.22).

Zulfiqar, M., 2022d. mzML Standard Files from a Diatom [WWW Document]. MzML Stand. Files Diatom. URL 10.5281/zenodo.7106205

Zulfiqar, M., 2022e. Spectral Libraries for MAW [WWW Document]. MAW Database Download. URL 10.5281/zenodo.6528931

Zulfiqar, M., Peters, K., 2022. LCMS2 Spectra from Bryophytes Dataset for Metabolome Annotation Workflow (MAW) [WWW Document]. LCMS2 Spectra Bryophyt. Dataset. URL 10.5281/zenodo.7107096