# MCnebula: Critical chemical classes to classify and boost identification by visualization for untargeted LC-MS/MS data analysis

**Lichuang Huang1, 2 #, Qiyuan Shan1, 2 #,** **Qiang Lyu1, Shuosheng Zhang3, Lu Wang1, 2\*, Gang Cao1, 2\***

1 School of Pharmacy, Zhejiang Chinese Medical University, Hangzhou, China

2 Jinhua Institute, Zhejiang Chinese Medical University, Hangzhou, China

3 College of Chinese Materia Medica and Food Engineering, Shanxi University of Chinese Medicine, Jinzhong, China.

#These authors contributed equally to this work

\*Address correspondence to:

**Lu Wang, Ph.D.**

School of Pharmacy, Zhejiang Chinese Medical University, No. 548 Binwen Road, Hangzhou, Zhejiang 310053, China; E-mail: luwang0520@163.com

**Professor Gang Cao, Ph.D.**

School of Pharmacy, Zhejiang Chinese Medical University, No. 548 Binwen Road, Hangzhou, Zhejiang 310053, China; Tel/Fax: +86 571 87195895; E-mail: caogang33@163.com

## Abstract

Untargeted mass spectrometry is a robust tool for biology, but it usually requires much time on data analysis especially for system biology. We established a framework called MCnebula (Multiple-Chemical nebula) to facilitate mass spectrometry data analysis process by focusing on critical chemical classes and visualization in multiple dimensions. It consisted of three vital steps: (1) abundance-based classes (ABC) selection algorithm, (2) critical chemical classes to classify ‘features’ (compounds), (3) visualization as multiple Child-Nebulae (network graph) with annotation, chemical classification and structure. Notably, MCnebula can be applied to explore classification and structural characteristic of unknown compounds that beyond the limit of spectral library. What’s more, it is intuitive and convenient for pathway analysis and biomarker discovery due to its function of ABC selection and visualization. MCnebula was implemented in the R language. We provided a series of tools in the R packages to facilitate downstream analysis in a MCnebula-featured way, including feature selection (statistical analysis of binary comparisons), homology tracing of top features, pathway enrichment analysis, heat map clustering analysis, spectral visualization analysis, chemical information query and output analysis reports, etc. In order to illustrate the broad utility of MCnebula, we investigated a human-derived serum dataset for metabolomics analysis. The results indicated that ‘Acyl carnitines’ were screened out by tracing structural classes of biomarkers which was consistent with the reference. We also investigated a plant-derived dataset of herbal *E. ulmoides* to achieve a rapid unknown compound annotation and discovery.

**Keywords:** Mass spectrometry, visualization, chemical classes, identification, MCnebula

## Introduction

Analyzing untargeted liquid chromatography/tandem mass spectrometry (LC-MS/MS) dataset is complicated, due to the massive of data volume, complexity of spectra and structural diversity of compounds. In the past decades, a lot of researchers attempted to address the issues. Many technical software or web-based interfaces were developed to provide a one-stop bulk solution for data analysis[1](#ref-2020p)–[4](#ref-2016a). These solutions applied or suggested flexible mass spectra processing tools or analogous algorithms[5](#ref-2012d)–[8](#ref-2010). To reduce false-positive and false-negative results, more algorithms have been implemented for peak deconvolution, feature selection or statistical filtering[9](#ref-2017f)–[12](#ref-2022b). Every feature corresponding to a compound within sample or parallel samples was prevalently equipped with fragmentation spectra for identification. In this context, researchers have to be confronted with a problem: how to identify so many compounds accurately and quickly?

Until today, several strategies have been developed for identifying compounds with fragmentation spectra. **1)** Spectral library matching. A number of public available databases were built to settle that via achieving re-usability of reference fragmentation spectra, such as MassBank, MassBank of North America (MoNA), Global Natural Products Society molecular networking (GNPS)[4](#ref-2016a). In the meanwhile, these fragmentation spectra are available via their web servers, third-party platform (e.g., [CompMass](http://prime.psc.riken.jp/compms/msdial/main.html#MSP%3E)) or specific tools (MASST)[13](#ref-2020cm). However, compared with structure database (PubChem harbours over 100 million records), spectral library is too small in size that limit the application of mass spectrometry. To cross this barrier, **2)** *In silico* simulation by fragmentation spectra. *In silico* tools have been increasingly developed for simulating fragmentation spectra[14](#ref-2010c)–[17](#ref-2017aq). Some databases such as MoNA collated *in silico* fragmentation spectra and were available for public[18](#ref-2013w). **3)** *In silico* prediction with matching learning. At present, the algorithms made machine training from reference mass dataset or libraries, then ‘learned’ how to predict chemical fingerprints or principles so as to retrieve the correct structure from molecular structure database[19](#ref-2012ab)–[21](#ref-2018ay).

*In silico* methods are developing quickly. Currently, the cutting-edge technology, called SIRIUS 4[22](#ref-duhrkop_sirius_2019), integrates many advanced artificial intelligence algorithms and has achieved an accuracy rate of 70% when retrieving from molecular structure database. This method helps to identify metabolites beyond the scope of spectral libraries. While *in silico* tools boost chemical identification, it is still lack of a proper framework that could incorporate and leverage SIRIUS 4 into user-friendly way for biological research, such as biomarker discovery and pathway analysis of mass spectral dataset. Compounds annotation and screening of biomarkers manually are both time-consuming and the results are impressed by subjective factors. Molecular networking is more and more popular due to its visualization and data transparency. Molecular networking was a spectral correlation and visualization method that can detect spectra from related molecules (so-called spectral networks), even when the spectra were not matched to any known compounds[4](#ref-2016a). Based on the concept of molecular networking, we proposed an idea, clustering features for visualization of chemical classification probably contribute to the discovery of biomarkers and metabolic pathway analysis.

The history of classification in chemistry dated back to the middle of the last century. The Chemical Fragmentation Coding system was firstly developed by Derwent World Patent Index (DWPI) in 1963. Until recent years chemical classification like Gene Ontology (GO)[23](#ref-2000g) which was organized with taxonomy and ontology was proposed more systematically[24](#ref-2016). ClassyFire is popular for compound annotation in LC-MS dataset analysis due to its computation available and systematicness[25](#ref-2019bt)–[28](#ref-2019bq). The taxonomy and ontology is robust and useful for chemistry. For example, a hierarchical classification-based method, called Qemistree, was proposed to analyze mass spectrometry data by expressing molecular relationships as a tree, which could be represented in the context of sample metadata and chemical ontologies[29](#ref-2021b).

Untargeted metabolomics is a field of omics science that uses cutting-edge analytical chemistry techniques and advanced computational methods to characterize complex biochemical mixtures aimlessly. LC­MS-based untargeted metabolomics is very popular due to its high sensitivity, small sample volume and direct injection without separation etc.[30](#ref-2016aq). With the help of statistical methodologies, researchers could screen and identify more-informative disease biomarkers from thousands of LC-MS features, to aid the design or development of improved treatments and to better assess health outcomes[31](#ref-2016ar). These statistical approaches mainly involved classical statistic and artificial intelligence models(e.g., random forests)[32](#ref-2019bv). Both approaches were inevitable to introduce specific biases, owing to the complexity of feature set or algorithmic stability[33](#ref-2017i). Furthermore, analyzing at feature level (quantification of the peak area relating to the ‘features’) was unable to profile systematic effects on metabolites unbiased[34](#ref-duhrkop_systematic_2021). In this view, analyzing at chemical classified level (classifying of ‘features’ with the knowledge of systematic chemical classification) may be a better settlement. However, it shouldn’t be ignored that the differences of metabolites at the same classified level. For example, small-molecules belonging to ‘Indoles and derivatives’ had structural dependent affection on aryl hydrocarbon receptor (AHR)[35](#ref-2019c). Different structural characteristics will lead to diverse activities. The settlement for that is integrating both ‘feature’ level statistic and classified level assessment.

In addition to chemical classifying and statistical analysis, clustering visualization was also a popular tool for untargeted mass spectrometry data analysis. Over the last decade, Global Natural Products Social Molecular Networking (GNPS) is more and more popular as a clustering visualization tool based on MS dataset. GNPS applied molecular networking connecting mass spectra of molecules based on the similarity of their fragmentation patterns[36](#ref-2012a). Unfortunately, molecular networking of GNPS mainly depend on on spectral similarity instead of compounds structural or classified similarity. For example, flavonoids consist of an aromatic ring joined to an oxygenated heterocyclic ring linked to a phenyl group,which were expected to be clustered together since its specific class and structural similarity. However, , it was reported that some compounds belonging to flavonoids happened to be absent from the cluster of other flavonoids compounds in previous research[34](#ref-duhrkop_systematic_2021). Thus, clustering visualization in a classified level is a better choice for untargeted mass spectra dataset. Earlier in 2012, the concept of molecular networking with visualization for mass data analysis was proposed for the first time[36](#ref-2012a) but *in silico* tools for predicting compound classification by fragmentation spectra were not available at that time. Nowadays, with the development of automatic classified *in silico* tools[24](#ref-2016), it is time for a revolution of the visualization strategy with higher confidence in classified level.

For above consideration, we proposed a comprehensive framework, named MCnebula (<https://mcnebula.org/>), for untargeted LC-MS/MS dataset analysis. MCnebula integrated a new abundance-based classes (ABC) selection algorithm for chemical classes selection. The principle of ABC selection algorithm: (1) applied an initial filtering to thousands of chemical classes based on the predicted probability, (2) regarded all ‘features’ as a whole, examined the number and abundance of ‘features’ of each chemical classification (classification at different levels, classification of sub-structure and dominant structure), and then selected representative classes, (3) these chemical classes were followed by goodness assessment (about identification of its classified compounds) and identicality assessment (the extent to which these chemical classes are distinguished from each other in the context of MS/MS spectra). The final chemical classes would serve to the subsequent analysis: visualized as Child-Nebulae and focus on these chemical classes (or Nebulae) for biomarker or chemistry discovery. The top ‘features’ based on statistical analysis could be set as tracer to discover more homology compounds of chemical structure or spectral similarity or chemical class. MCnebula can be used to explore unknown compounds because of the annotation module and the cutting-edge technology of SIRIUS software[20](#ref-duhrkop_searching_2015),[22](#ref-duhrkop_sirius_2019),[34](#ref-duhrkop_systematic_2021),[37](#ref-bocker_sirius_2009)–[39](#ref-ludwig_database-independent_2020), which exceeded the limitations of spectral library matching. MCnebula was implemented in the R language and can be easily integrated into the diverse biological analysis workflow of R. MCnebula (updated to MCnebula2, which included more tools such as ABC selection algorithm, Nebula visualization, statistical analysis, and output report etc.) was written primarily in S4 system of object-oriented programming. It allowed all data for one-button analysis from the beginning to the end, which facilitated data processing. In addition to the basic function of MCnebula), we provided an additional ‘exMCnebula2’ package for downstream analysis, which contained all the analysis tools used in this study such as pathway enrichment analysis, heatmap clustering analysis, spectral visualization analysis, chemical information query, etc. Downstream analysis of untargeted LC/MS-MS is complex and varies from data to data. The additional tools in exMCnebula2 package could provide a prototype for the expanded application of MCnebula.

In this article, two datasets were applied in MCnebula in order to demonstrate the broad utility of our method. One was a human-derived serum dataset that correlated with mortality risk profiling of staphylococcus aureus Bacteremia (SaB); the other was a plant-derived herbal dataset that related to the traditional processing of herbal medicine.

## Experimental section

### MCnebula algorithm

**Overview**. The analysis of untargeted LC-MS/MS datasets typically starts with feature detection. Annotating these features is time-consuming and extracting useful information for further biological research can also be challenging. To address these issues, we present MCnebula, with an abundance-based classes (ABC) selection algorithm, to boost mass spectrometry data analysis by focusing on critical chemical classes and visualization in multiple dimensions.

**Molecular formula and chemical structure prediction**. The analysis of MS/MS spectra involve a process of inference and prediction. We deduce the molecular formula based on the molecular weight of elements composition and possible fragmentation pattern of the MS/MS spectrum by SIRIUS; We then search for the exact chemical structure from compound structure databases. However, this process is often uncertain as several factors can affect the reliability of MS/MS data and the correctness of inference. SIRIUS generates a list of candidates for the potential molecular formula, chemical structure, and chemical classification behind each MS/MS spectral feature. In such cases, MCnebula extract these candidates and determine the unique molecular formula and chemical structure for each MS/MS spectrum based on the highest score of prediction.

**Top candidate selection by multiple score systems**. After the process of compound prediction, we get candidates for molecular formula, structure, and chemical classes. Some candidates are correct while others are not. While the correct predictions for molecular formula and chemical structure are unique, the chemical classification may have multiple correct predictions belonging to different classes of hierarchy. The scoring method depends on personalized research purpose. We can use scores based on various criteria such as isotopes, mass error, structural similarity, or chemical classes to rank and filter the candidates. With numerous score systems available, we filter out low-scoring candidates and focus on those with higher scores, which are more likely to be the correct compound. However, in most cases, the top candidates from the three scoring systems are not always consistent. So, we choose a ‘specific candidate’ as a reference in the user-selected scoring system and retrieve its chemical information from the other systems for data integration. We obtain unique molecular formulae and chemical structure for the reference by scoring and ranking, but for chemical classes, more work is needed.

**Chemical classification**. Compounds with MS/MS spectrum can be classified based on their overall structure or local structure, which we refer to as the dominant structure and substructure, respectively. Therefore, in the chemical classification system, we can classify compounds not only based on their dominant structure, but also based on their substructure. When the dominant structure is unavailable, or the MS/MS spectral fragment is insufficient, we classify the compounds using substructure information to gain knowledge of the compound. Note: Classifying compounds based on their dominant structure is straightforward. For example, we classify Taxifolin as a Flavone, not a phenol, even though its local structure contains a phenol substructure. We prefer to classify compounds based on their dominant structure because it is more concise and provides more information. However, during the MS/MS spectral analysis, we sometimes can only classify compounds based on their substructure.

**ABC selection**. The ABC selection algorithm evaluates all features collectively in an untargeted LC-MS/MS dataset, by examining the number and abundance of features for each chemical classification at different levels with both substructure and dominant structure. Based on this analysis, representative classes are selected for subsequent analysis (as shown in Fig. 6).

* Create Stardust Classes (Inner filter). By using the posterior probability of classification prediction (PPCP) data for each feature, simple threshold or absolute conditions are set to filter the chemical classes, and refer to as ‘inner’ filtering.
* Cross filter Stardust Classes. This involves combining the data of the chemical classes and their classified features (i.e. Stardust Classes data), grouping them based on the chemical classes, and then performing statistics on the features within each group. Statistics may also be performed on these data in conjunction with features annotation data, or to compare groups with each other. This method involves crossover of attributes for filtering, hence it is referred to as ‘cross’ filtering. (See details in the next subsection about Cross filter Stardust Classes.)

The resulting dataset is called Nebula-Index, which records multiple chemical classes and their associated features. Each chemical class is considered as a ‘Nebula’ and its classified ‘features’ are the components of the Nebula. These Nebulae are visualized as networks, with Parent-Nebula representing all features combined and Child-Nebulae representing individual classes with their ‘features’.

**Details of Cross filter Stardust Classes.** This method involves integrating three modules, as shown in Fig. 6:

*Cross-filtering by ‘quantity’* (abundance selection): The ‘features’ quantity limitation is set for each group (a chemical class with its classified ‘features’). The minimum quantity of ‘features’ within the group and the maximum proportion of ‘features’ quantity within the group versus all ‘features’ (unique) quantity of all groups are used as thresholds. Chemical classes with too many or too few ‘features’ are filtered out.

*Cross-filtering by ‘score’* (Goodness assessment): This step associates the Stardust Classes data with ‘features’ annotation data. For each group, the Goodness assessment is performed for each target attribute (continuous attribute, generally a scoring attribute of compound identification, such as ‘Tanimoto similarity’). If the group satisfies all the expected Goodness, the chemical class is retained. The Goodness () is calculated as follows: , where is the quantity of ‘features’ of which the target attributes satisfy the cut-off, and is the quantity of all ‘features’.

The assessment of Goodness is related to the parameters of ‘tolerance’ and ‘cutoff’: the expected Goodness value of ‘tolerance’ and the actual Goodness, which is related to the parameter ‘cutoff’.

Goodness assessment can be given to multiple target attributes. Note that the chemical class is retained only if it passes the Goodness assessment of all target attributes. The main purpose of this step is to filter out chemical classes with too many ‘features’ of low structural identification.

*Cross-filtering by ‘identical’* (identicality assessment): This step involves a similarity assessment of chemical classes. A hierarchical range is set for chemical classification, and groups within this range are compared for similarity. If the classified ‘features’ of two groups are almost identical to each other, the chemical class represented by one of the groups is excluded. The degree of identicality between two groups (A and B) is assessed, as well as the value of the parameter ‘identical\_factor’ () :

: ratio of the classified ‘features’ of A belonging to B

: ratio of the classified ‘features’ of B belonging to A

: value of parameter ‘identical\_factor’

If and , the two groups are considered identical, and the group with fewer ‘features’ is discarded. The purpose of this step is to filter out classes that may incorporate each other and are similar in scope. The *in silico* prediction approach may not be able to distinguish which class the potential compound belongs to from the LC-MS/MS spectra.

**Network graph presentation**. The features and their annotations are integrated as Nebulae based on the Nebula-Index. These Nebulae are represented as network-type graph data. The feature annotation data includes top candidates for chemical formula and structure. The MS/MS spectral similarity of the features is calculated and used to generate the edge data for the network graph.

**Visualization system**. MCnebula integrates various R packages to format data, including the ‘ggplot2’ package. To make visualization easier for users, we developed the ‘ggset’ data class, which stores pre-defined ggplot2 plotting functions and parameters for visualizing Nebulae. Users can customize the visualization according to their specific needs or the requirements of the publisher.

**Statistical analysis**. MCnebula integrates the functions of the ‘limma’ package for differential expression analysis of RNA-sequence and microarray data[40](#ref-gentleman_limma_2005-1), and package them for differential analysis of metabolomics data. The gene expression matrix and feature quantification matrix of LC-MS are similar, both have phenotypic variables (sample information) and dependent variables (gene expression or feature quantification values). Our method can be appropriate for statistical analysis of feature quantification of experimental designs in which explanatory variables are factorial variables and the design matrix is without an intercept[41](#ref-law_guide_2020).

**Feature detection**. Feature detection is a kind of algorithm for detecting peaks from mass data file, and most mass spectrometry processing tools have a similar function. Users can implement this process with any tool, but to access the MCnebula workflow, .mgf (long list file containing MS/MS information) and .csv files (or other formatted table file of feature quantification) were required for output. In this study, all processing of Feature detection were implemented in MZmine2 (version 2.53). But now, the R package MCnebula2 has provided some convenient tools which integrated XCMS (R package) methods or functions for Feature detection. See details in: <https://mcnebula.org/>.

**Data structure**. MCnebula was primarily developed using the R S4 system of object-oriented programming. All data including ‘features’ annotation data and visualization data is stored in a single object (class ‘mcnebula’), which simplifies the application, makes data management and analysis easier to perform and repeat.

**Reporting system**. MCnebula includes a reporting system that enables the analysis process to be output as a PDF document or in other formats. The reporting system is based on the ‘report’ data class, which stores each step of the analysis as a section and can be easily modified according to the user’s requirements. Furthermore, the ‘rmarkdown’ R package[42](#ref-xie_r_2020) is incorporated in the reporting system to generate reports.

**Code Compatibility**. MCnebula performs downstream analysis by extracting data from the pre-computed SIRIUS project, which is the primary data source for MCnebula 2. The SIRIUS software is continually updated and enhanced. From SIRIUS version 4 to version 5 (https://github.com/boecker-lab/sirius), the data structure and attribute names in the project directory have been modified. To ensure that MCnebula is not affected by version problems, we have designed its application programming interface (API) for various SIRIUS versions.

### MCnebula evaluation

**Spectra dataset for evaluation**. To evaluate the performance of MCnebula, the spectra from the GNPS MS/MS library was used (http://prime.psc.riken.jp/compms/msdial/main.html#MSP). To prevent overfitting during library match evaluation, ‘noise’ was added to the MS/MS spectra[43](#ref-2021). Two models of noise were simulated: medium noise and high noise. The simulation involved a global mass shift, individual mass deviations, intensity variations, and additional ‘noise peaks.’ Isotope patterns were also simulated using the ‘get.isotopes.pattern’ function within the ‘rcdk’ R package[44](#ref-2007j). The mass and abundance of isotopes were considered for the adduct type to increase or decrease exact mass. The ‘isotope peaks’ were merged into the MS1 list of compounds, and all spectra collections were formatted into mgf or csv file for the continuing MCnebula and benchmark analysis.

**Evaluation method.** MCnebula and benchmark workflow was conducted for all the three simulated datasets. SIRIUS 4 command-line interface (CLI) (version 4.9.12) was applied for computation, and filtered out MS/MS spectra with empty fragmentation peaks. In total 7524 out of 8782 compounds were left for evaluation. ClassyFire was used to assess the classification accuracy[24](#ref-2016). After *in-silico* annotation, we obtained structure annotation, International Chemical Identifier Key (InChIKey), and other metadata of these compounds. Considering that ClassyFire only supports chemical identities those structures have been classified in its server previously, we used the first hash block of InChIKeys (InChIKey planar, which represents the molecular skeleton) to query the PubChem application programming interface (API) (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest)[45](#ref-2022ak). This provided us with all the possible InChIKeys of isomerism (stereo, isotopic substitution)[46](#ref-2012e). Classification of small molecules depends on their molecular skeleton, so chemical identities that share the same InChIKey planar are identical in classification. The InChIKey list was imported into ClassyFire to obtain chemical classification. In our R script, once any InChIKey of isomerism matched the classified data in the database, we turned off the acquisition status for this molecular skeleton. In the end, we collated, integrated, and assigned all these chemical annotations as standard reference.

Due to differences in algorithms and classified results, we evaluated the MCnebula and benchmark methods separately. Since sub-structural classification was not available for the benchmark method, we excluded these classes during the evaluation analysis. Nevertheless, some compounds within the remaining classes may still be classified into sub-structural classes. We assigned three levels for evaluation: ‘True,’ ‘Latent,’ and ‘False.’ ‘True’ indicated that the classified classes were consistent with those of ClassyFire. ‘Latent’ indicated that the classified classes were not consistent with ClassyFire, but their parent classes at the ‘class’ level were consistent. ‘False’ indicated that the classified classes were completely inconsistent with ClassyFire. Then, the true positive (TP) was assigned as: ‘True’ + ‘Latent’; the false positive (FP) was assigned as ‘False’.

To evaluate the identification of classes or structures, we merged the results with a standard reference by InChIKey planar. For the evaluation of chemical structure identification, we considered a structure as ‘True’ if it matched the chemical structure identified by InChIKey planar. However, this evaluation neglected stereochemistry due to the limit of LC-MS/MS detection.

### Other information

More methodological details regarding the MCnebula assessment, handling of samples in the study (serum and phytochemicals), data processing, and acquisition of available data and codes are documented in the supplemental file.

## Results and discussion

### Results

#### Overview

The MCnebula workflow was dedicated to analyzing LC-MS/MS datasets from the beginning, i.e. starting from the raw data obtained from the sample and going through the various stages of analysis to obtain a complete analysis report (Fig. 1). The analysis process followed the general MCnebula analysis templates, from filtering candidates of chemical formula, structural formulae, chemical classes, to creating visual Nebulae; it also allowed for custom advanced analysis, with the help of chemical class focused Nebulae, to perform statistic analysis, features selection, focus on critical metabolites (compounds) and their structural characteristics, pathway enrichment, querying compound synonyms, etc. One of the main characteristics of the MCnebula workflow was the visualization system. The mass dataset is finally presented as Parent-Nebula and Child-Nebula: Parent-Nebula is a simple network visualization of a mixture of all the features, which contains too much information for the user to extract useful information intuitively; Child-Nebulae were multiple network visualization for LC-MS/MS data analysis highlighted the advantages of the ABC selection algorithm for fast filtering and selection of chemical classes. We evaluated workflow of MCnebula in function with other popular public available methods. In terms of the evaluation of our chosen indicators which cover identification, classifying, etc., MCnebula has a wider scope of applicability. (Tab. S1)

#### Method evaluation

**Classified accuracy.** We used a public available reference spectral library to assess the accuracy of classifying by MCnebula. The direct use of such a reference spectral library may lead to over-fitting during the evaluation. We have taken the approach of simulating noise to eliminate this consequence. Simulating noise, i.e., adding invalid noisy data to the reference spectrum or numerically shifting the existing data, also simulates data acquisition similar to a real scenario: due to the different acquisition conditions, the spectral data in the real case will be more noisy compared to the reference spectrum. By adding noise to the reference spectral library, we now have three datasets for evaluation (origin, medium noise and high noise dataset) (commonly 7524 compounds (spectra)). All three datasets were analyzed using MCnebula. Due to the richness of the compounds in the reference spectra, for the origin dataset, we obtained a total of 152 chemical classes (each with a corresponding compound to be evaluated) via using ABC selection algorithm. These 152 chemical classes include both chemical classes refined on the basis of dominant structures and chemical classes refined on the basis of substructures. To facilitate comparison with other methods, we selected only chemical classes that are likely to be dominant structures for evaluation. There were 37 such chemical classes that were selected for evaluation. To evaluate MCnebula more objectively, we chose the molecular networking provided by GNPS (Global Natural Products Social Molecular Networking), with the modules of Feature-based molecular networking (FBMN) and MolNetEnhancer, as the benchmark method to provide a visualized clustering analysis of mass spectra data. GNPS is a typical and popular spectral library-based mass spectrometry annotation method. In principle, it first calculates spectral similarity by conducting mirror match with public spectral library, identifies compounds with the exact chemical structures, and then determines the chemical class based on the annotated chemical structure.

We uploaded the three datasets to the GNPS server and then obtained the results and used them for evaluation. For origin dataset, GNPS resulted in a total of 44 chemical classes (parallel to MCnebula, with at least 50 compounds per chemical class). There were 19 common classes in total. These classes were selected to compare MCnebula and GNPS in parallel in terms of classified number, stability, precision and recall. For the classified number (TP + FP), MCnebula outperformed GNPS in three datasets (MCnebula: 199, 178, 160; GNPS: 162, 95, 81) (Fig. 2a). For the stability of the classifying after adding noise, MCnebula outperformed GNPS in two dataset (MCnebula: 89.5%, 81.2%; GNPS: 59.3%, 49.9%) (Fig. 2a). For the indicators of precision, to assess the performance of classifying, it combined the level of the stability to calculate the relative false rate, rather than the absolute false rate, then it was used to estimate precision. The relative false rate better simulated the actual application to LC-MS/MS analysis, since the actual spectral data contains not only noise but also many unknown compounds that cannot be identified by spectral matching. In this context, MCnebula outperformed the GNPS in the evaluation of the precision in three datasets (MCnebula: 69.8%, 67.1%, 67.4%; GNPS: 48.1%, 51.2%, 52.4%) (Fig. 2a). The recall was estimated using relative false rate as well. As shown in Fig. 2a, MCnebula (82.2%, 81.6%, 81.6%) outperformed than GNPS (51.2%, 52.7%, 53.3%). In addition to the three indicators mentioned above, we also compared MCnebula and GNPS at the individual level for the 19 chemical classes (Fig. 2b). Remarkably, MCnebula was more stable to noise than GNPS.

**Identified accuracy.** Using MCnebula workflow, the origin dataset containing 8057 compounds (precursor ions m/z < 800), all of these compounds were predicted with chemical molecular formulae, and of these, 6610 compounds were predicted with chemical structure. Those chemical structure were evaluated for accuracy in a classified context. For the 37 chemical classes (Fig. 2c), the average false rate of identification was 37%; the average identified compounds number was 156. Among them, most of the identified false rate were between 30% to 40%, however, some classes were quite low, such as ‘Long-chain fatty acids’ or ‘Lignans, neolignans and related compounds’. The reliability of the predicted chemical structure can be assessed in terms of a score. Tanimoto similarity provides such a score for each predicted chemical structure (it provides the matching degree of chemical fingerprints with structures). When Tanimoto similarity sets the cutoff value to 0.5, the average false rate of identification was 29.4%; the average identified compounds number was 139 (Fig. 2c). Above we evaluated the accuracy of the identification of compounds in each chemical class obtained by MCnebula. It should be noted that MCnebula itself does not contain any module for identification, it only utilized the top scoring candidate from the SIRIUS predicted results for annotation. For more evaluation on identification please refer to the publication and our previous related work[22](#ref-duhrkop_sirius_2019),[47](#ref-lai_deep_2022).

#### Data analysis

**Serum metabolic analysis.** To illustrate the application of MCnebula in metabolism, we re-analyzed the serum data from Wozniak et al.[48](#ref-2020s). The serum samples were collected from patients in-hospital infected with Staphylococcus aureus bacteremia (SaB) or not and healthy volunteers. Overall, the samples were divided into 1) control groups, involving NN (non-hospital, non-infected) and HN (hospital, non-infected); 2) infection groups, involving HS (hospital, survival), HM (hospital, mortality).

A total of 7680 ‘features’ were detected while running with LC-MS preprocessing on the serum dataset. After predicting the compounds by MS/MS spectra (with SIRIUS software), subsequent analysis was performed using MCnebula. Of these, 6501 ‘features’ were annotated with predicted molecular formula and further, 3449 ‘features’ were annotated with predicted chemical structure. Using ABC selection algorithm, we filtered out more than one thousand chemical classes by applied of ‘inner filter’ module (see method section of ABC selection algorithm); further filtered out 508 chemical classes while performing ‘cross filter’; for the remaining 41 chemical classes, 19 chemical classes were manually filtered out, while leaving the final 22 chemical classes to make up the Nebula-Index, which further visualized as Child-Nebulae. It is worth mentioning that the filtered out 527 (508 + 19) chemical classes could be re-added to the analysis. Herein, with the basic workflow of MCnebula, Parent-Nebula and Child-Nebulae were obtained (Fig. S1, Fig. S2). By interrogating Child-Nebulae, we had a basic insight into the chemical classes contained in the serum dataset. To mine more information from Child-Nebulae: we performed a binary comparison of HS and HM groups, ranking ‘features’ according to Q-value (adjusted P-value); the top 50 ‘features’ were set as ‘tracers’ to mark them in Child-Nebulae (Fig. 3). By combining the features selection algorithm about Q-value, the chemical classes exhibited in Child-Nebulae were reduced. The log2(Fold Change) (log2(FC)) quantification for the HM versus HS groups was visualized in Child-Nebulae (Fig. S3). In Fig. S3, the overall level of ‘Bile acids, alcohols and derivatives’ (BAs) class and ‘Acyl carnitines’ (ACs) (Fig. 4a and b) class increased remarkably, while the overall level of ‘Lysophosphatidylcholines’ (LPCs) class decreased remarkably. Indeed, BAs, ACs and LPCs were reported associated with liver dysfunction, imbalance of intestinal microphylactic homeostasis, and mortality risk[48](#ref-2020s)–[50](#ref-2016at).

By deep annotation of Child-Nebula, all three classes of compounds have similar structural parent nuclei, and their levels in the NN, HN, HS, and HM groups are similar (Fig. 4c, Fig. S4). Subsequently, we performed cluster heat map analysis and pathway enrichment analysis on the compounds of these three classes. As shown in the clustering heat map (Fig. 5), the control group of ACs and BAs were remarkably separated from the infection group, which implied the infection relevance of ACs and BAs to SaB. In contrast, LPCs did not show remarkable SaB infection relevance or mortality relevance, probably owing to the general consistency of this class of compounds for SaB disease. We performed pathway enrichment analysis for these three classes of significant compounds (HS versus HM group, Q-value < 0.05). The results of BAs showed that four compounds exhibited metabolic pathways associated with ‘Bile secretion’, ‘Cholesterol metabolism’, and ‘Primary bile acid biosynthesis’ etc (Fig. S5b). Among them, GCS was a class of compounds with the same parent nucleus. The results for LPCs suggested that compounds with similar parent nucleus structure of LPCs implied association with a series of downstream pathways (Fig. S5c). The significant compounds of ACs were not enriched in the pathway. But, A fundamental role of ACs in tuning the switch between the glucose and fatty acid metabolism was reviewed[51](#ref-2018bi). Their function implemented via bi-directional transport of acyl moieties Between cytosol and mitochondria (Fig. S5a).

In research of Wozniak et al[48](#ref-2020s), five ACs compounds were identified. In addition, four top metabolites (2-Hexadecanoylthio-1-Ethylphosphorylcholine (HEPC); sphingosine-1-phosphate (S1P); decanoyl-carnitine; L-Thyroxine (T4)) were also identified. In our reanalysis, all identifications were in line except for HEPC (see ‘Data and code availability’ section for the report of serum dataset analysis). In our re-analysis, ‘HEPC’ was identified as 1-pentadecanoyl-sn-glycero-3-phosphocholine (LPC15:0) or its stereoisomers. Indeed, HEPC and LPC15:0 are quite similar in structure, but distinct in element constitution (corresponding to C23H48NO5PS and C23H48NO7P respectively). They were clearly distinct in terms of chemical classification. HEPC belongs to ‘Cholines’ (level 5) from ‘Organic nitrogen compounds’ (superclass) family, whereas LPC15:0 belongs to ‘Lysophosphatidylcholines’ (level 5) from ‘Lipids and lipid-like molecules’ family. As a part of MCnebula workflow, sulfur element is detectable for SIRIUS in isotopes pattern with high mass accuracy[37](#ref-bocker_sirius_2009). However, for the MS/MS spectra of ‘HEPC’, there was no candidate formula that containing sulfur element. Overall, we identified more compounds with the MCnebula workflow and many of the results were in line with the analysis of Wozniak et al[48](#ref-2020s). All identified compounds were collated in Tab. S2 (filtered with Tanimoto similarity > 0.5 and de-duplicated with the first hash block of InChIKey (molecular skeleton)). The compounds (top 50 of EFS and MWU) that were not successfully identified via spectral library matching by Wozniak et al. but were identified by our MCnebula workflow for molecular formula or chemical structure were additionally collated (Tab. S3).

**Herbal medicine analysis.** We used MCnebula to interpretate structure diversity and chemical transformation during traditional processing of a representative herbal medicine, *Eucommiae Cortex,* the peel of *Eucommia ulmoides Oliv. (E. ulmoides)*[52](#ref-2021n). After being processed with saline water, it is commonly applied to treat renal diseases for a long time in China but the chemical basis still remained to be explored. With the help of ABC selection algorithm in MCnebula, a total of 29 chemical classes representing the richness of composition of E. ulmoides were obtained. Two groups of quantification data were performed with binary comparison. The top 20 features (Top20) were selected using function ‘select\_features’ (|Log2(Fold change)| > 0.3, Q-value < 0.05, Tanimoto similarity > 0.5) and were traced in Child-Nebulae (Fig. S6). We used MCnebula to draw the mirrored match of MS/MS spectra and extracted ions chromatography (EIC) plots of the Top20 (Fig. S7 and S8). According to Fig. S8, we speculated that the ‘features’ of ID 1642, 1785, and 2321 were newly generated compounds because the peak area levels before the processing were almost zero compared to those after processing. Their chemical structures are showed in Fig. S7. Among them, the ‘feature’ of ID 1642 has a higher probability of correct identification (Tanimoto similarity: 0.69). Based on Fig. S6, we knew that ‘feature’ of ID 1642 belongs to ‘Iridoids and derivatives’ (IAD), the others were ‘Dialkyl ethers’ (DE; ID 1785) and ‘Phenylpropanoids and polyketides’ (PAP; ID 2321). We annotated in depth of the Child-Nebulae of IAD, DE and PAP respectively. The locations of the ‘features’ of ID 1642, 1785 and 2321 in the Child-Nebulae were interrogated (Fig. S9a, b, and c). Only the ‘features’ of ID 1642 had neighboring ‘features’ and their identified chemical structures (ID 2110 and ID 854) had similar parent nuclei. The ‘features’ of ID 2110 and ID 854 were identified with chemical structure (Tanimoto similarity: 0.69 and 0.7 respectively) (Fig. S9d, e, and f); their levels of peak area were decreased and increased after the processing. Based on the chemical structures shown in Fig. S9d and e, we speculated that the compound of ID 2110 was partially converted to the compound of ID 854 after the processing, which may involve chemical changes such as dehydration and rearrangement. Such speculation explained the alteration of the levels of peak area. In addition, the increasement in the level of the compound ID 1642 (its spectra were shown in Fig. S7 and S8) may also be associated with the reduction of the compound ID 2110.

The methods of MCnebula we have demonstrated for discovering significant compounds and discovering chemical changes can be applied to explore more compounds in Tab. S4, but we would not expand on this description here.

### Discussion

The analysis of LC-MS/MS data is challenging because of its large dataset and the potential information of the unknown compounds as well as the limited of reference spectral library. Researchers often need to take much time on mapping the landscape of all the interesting compounds from this “black box”, and then move to next step in research MCnebula could assist researchers in focusing on potential markers or interesting compounds quickly by combining full-spectrum identification with machine prediction, visualization of sub-nebulae in a multi-dimensional view, and statistical analysis to track top ‘features’ and find analogues. The ABC selection algorithm can summarize a representative chemical class in a dataset and obtain the features to that class, so the overall direction of the study is unbiased. Also, it is an effective guarantee for statistical analysis to produce top features for tracing analysis in next step. The results of statistical analysis based on feature level may cause bias since the loss of information, filtering on the basis of chemical classes level can prevent the bias in some degree. The Child-Nebula, which mapped on the basis of the chemical classes obtained by the ABC selection algorithm, achieved the goal of visualizing the huge untargeted dataset as a single graph. Above all, the parameters of the ABC selection algorithm were subjectively adjustable and they should be determined according to the richness of the chemical class of the studied object. In general, our default parameters used to acquire the chemical classes that are abundant in variety according to the datasets and filtered out those that were too large or too small classes in conceptual scope.

For identification, spectral library matching is still the main method for LC-MS/MS data because of its high accuracy. The general classifying of compounds is also based on it, i.e., the chemical structure is firstly identified by spectral matching, and then its chemical class is evaluated based on the chemical structure. Considering the limit of reference spectral library, The classifying technique like CANOPUS[34](#ref-duhrkop_systematic_2021)incorporated in MCnebula bypassed the first step of identifying the chemical structure but predicted the possible chemical class even if the exact chemical structure was not known. MCnebula combined this cutting-edge technology with ABC selection algorithm and achieve visualization of Child-Nebulae, which make it possible to explore unknown compounds that beyond spectral library. We compared the classifying method of MCnebula with GNPS, of which method relies on chemical structure identification. When different levels of noise were added, the number of classified compounds of GNPS decreased remarkably compared with the stable performance of MCnebula. For the actual acquired MS/MS spectra, they were not as good as the reference spectra and contained some noise. Indeed ,the reality of MS/MS spectra is much closer to the condition with noise. It means MCnebula can resist noise interference in some extent. At the end of the evaluation, we examined the accuracy of the identification by MCnebula. It confirmed that the accuracy of identification fluctuated around 70%, which was the same as SIRIUS[22](#ref-duhrkop_sirius_2019). However, this accuracy is overestimated for some specific chemical classes, such as lipids, which is a result of the limitation of current *in silico* approach in the lipidomics[53](#ref-krettler_map_2021).

Serum metabolomics data was applied to illustrate that MCnebula can be used for pathway analysis and biomarker discovery. Most of our results were consistent with that of reference[48](#ref-2020s). The exciting thing was that that we identified more metabolites that beyond the scope of spectral library matching. Three of the four top metabolites identified by Wozniak et al. were the same as our re-identification, but only one metabolite was controversial. Wozniak et al. mentioned ACs compounds had correlation with SaB disease and ACs compounds were also re-identified in our study. Wozniak et al. used a joint approach of Ensemble Feature Selection (EFS) and Mann-Whitney U (MWU) tests to screen top metabolites[48](#ref-2020s). When we compared the 50 top ‘features’ obtained by the ‘binary comparison’ method integrated in MCnebula with the top 50 metabolites (top 50 of EFS and 50 of MWU) obtained by the joint method of W et al., a total of 37 overlapped metabolites were screened out, including the key metabolite of L-Thyroxine in the reference study. Top ‘features’ were usually different according to the feature selection algorithm. The reliability of the ‘binary comparison’ method was verified again by our ranked results comparing with those of Wozniak et al. In addition to the consistent parts, , more interesting results about other chemical classes associated with SaB disease were revealed by MCnebula. We discovered additional classes, i.e. ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs), that were not concerned in previous study. Acutally, LPCs have been extensively investigated in the context of inflammation and atherosclerosis development[50](#ref-2016at),[54](#ref-2020cv),[55](#ref-2014ao). In a recent review[54](#ref-2020cv), the complex roles of LPCs in vascular inflammation were well described, involving the context-dependent pro- or anti-inflammatory action, impact in innate immune cells and adaptive immune system, etc. Decreasing level of LPCs was associated with wild range of diseases of increasing mortality risk[50](#ref-2016at). The investigation of spesis indicated LPCs concentrations in blood were established correlation with severe sepsis or septic shock[55](#ref-2014ao). LPCs was reported inversely correlated with mortality in sepsis patients[56](#ref-2003n). BAs’ disorder implied a liver dysfunction and imbalance of intestinal microphylactic homeostasis[57](#ref-2021dg). The chemical multiversity of BAs, which were discovered in the BAs’ child-nebula, were determined by the intestinal microbiome and allowed for a complex regulation of adaptive responses in host. In our study, the level of BAs showed higher correlation with SaB infection than ACs. The decreased level of LPCs suggested a mortality risk of SaB infection. From LPCs to BAs, steroids related classes, ‘Lineolic acids and derivatives’, and other fatty acids related classes, showed that liver played a central role in SaB infection and mortality. Liver X receptors (LXRs) played pivotal roles in the transcriptional control of lipid metabolism[58](#ref-2018bd). LXRs modulated membrane phospholipid composition through activation of lysophosphatidylcholine acyltransferase 3 (LPCAT3), which was directly related to LPCs[59](#ref-2021di). The above classes showed correlation with LXRs[58](#ref-2018bd). Unfortunately, LXRs’ specific role in SaB infection or mortality has not been reported and beyond the scope of this research.

In herbal dataset analysis, MCnebula provided a quick annotation of compounds and exploration of chemical changes in Child-Nebulae with a scope of chemical classes. The main components of *E. ulmoides* were lignans, iridoids, phenolics, flavonoids, steroid and terpenoids[60](#ref-huang_traditional_2021). In our study, the chemical classes that obtained by ABC selection algorithm included ‘Lignans, neolignans and related compounds’ (LNARC) and ‘Iridoids and derivatives’ (IAD), as well as ‘Monoterpenoids’ and ‘Terpene glycosides’. The flavonoids were covered by ‘Phenylpropanoids and polyketides’ (PAP)[24](#ref-2016) and phenolics may be found in ‘Methoxyphenols’. The flavonoids were similar to the steroids and were not retained in selected results as ‘Flavonoides’ and ‘Steroids and steroid derivatives’, because they were not as abundant in *E. ulmoides* (bark) as LNARC and IAD. Many of the compounds that identified in chemical classes of LNARC and IAD (Tab. S1) were reported in previous research about LC-MS/MS analysis of *E. ulmoides*[61](#ref-2014w),[62](#ref-2015v). We obtained top features based on statistical comparison of the changes in ‘features’ quantification levels before and after processing. One of the compounds that changed significantly or even was newly produced (ID: 1642) was traced in the Child-Nebulae. We hypothesized that it was related to two structurally similar compounds by transformation. The application of MCnebula in the analysis of plant-derived compounds was well illustrated with this example, particular for quick identification and exploration of chemical changes. Notably, the reference spectral library or database for plant-derived compounds was much more scarce compared with reference spectral library for human-derived metabolites, Although some specific database of plant-derived compounds was constructed[63](#ref-2012ac), there were lack of enough fragmentation spectra for comprehensive library match. With the help of MCnebula, a rapid and reliable resolution of complex compositions of plant-derived can be achieved.

## Conclusion

The analysis of LC-MS/MS data is challenging because of its large dataset and much information of the unknown compounds as well as the limited of reference spectral library. Thus, we established a framework called MCnebula to facilitate mass spectrometry data analysis by focusing on critical chemical classes and visualization in multiple dimensions. MCnebula was proposed and implemented in the R language with package of MCnebula. As an integrated visualization method, MCnebula may be more popular for researchers without background of bioinformatics and computer science. According to the results of method evaluation, MCnebula had a lower relative false rate of classified accuracy and its accuracy of identification was up to 70%. In order to illustrate the broad utility of MCnebula, we investigated a human-derived serum dataset for metabolomics analysis. The results indicated that ‘Acyl carnitines’ were screened out by tracing structural classes of biomarkers which was consistent with the reference. We also investigated a plant-derived dataset of herbal E. ulmoides to achieve a rapid unknown compound annotation and discovery. MCnebula has a great potential in the field of chemistry and biology. In the future, we hope that fields of application of MCnebula could expand to agriculture, food science, medicine and so on.

## Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 81922073), Zhejiang Province Traditional Chinese Medicine Science and Technology Project (Nos. 2022ZQ033).

# Reference

(1) Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M. A Lipidome Atlas in MS-DIAL 4. *Nature Biotechnology* **2020**, *38* (10), 1159–1163. <https://doi.org/10.1038/s41587-020-0531-2>.

(2) Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Research* **2018**, *46* (W1), W486–W494. <https://doi.org/10.1093/nar/gky310>.

(3) Tsugawa, H. Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. In *Comprehensive Natural Products III*; Elsevier, 2020; pp 189–210. <https://doi.org/10.1016/B978-0-12-409547-2.14645-1>.

(4) 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. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34* (8), 828–837. <https://doi.org/10.1038/nbt.3597>.

(5) 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. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nature Biotechnology* **2012**, *30* (10), 918–920. <https://doi.org/ghh626>.

(6) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nature Methods* **2016**, *13* (9), 741–748. <https://doi.org/f82r32>.

(7) Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78* (3), 779–787. <https://doi.org/b58xrd>.

(8) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics* **2010**, *11* (1), 395. <https://doi.org/bxbwnj>.

(9) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Analytical Chemistry* **2017**, *89* (17), 8696–8703. <https://doi.org/gbrjtm>.

(10) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. Optimization of Metabolomic Data Processing Using NOREVA. *Nature Protocols* **2022**, *17* (1), 129–151. <https://doi.org/10.1038/s41596-021-00636-9>.

(11) Mahieu, N. G.; Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer Than 1000 Unique Metabolites. *Analytical Chemistry* **2017**, *89* (19), 10397–10406. <https://doi.org/10.1021/acs.analchem.7b02380>.

(12) Gloaguen, Y.; Kirwan, J. A.; Beule, D. Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics. *Analytical Chemistry* **2022**, *94* (12), 4930–4937. <https://doi.org/10.1021/acs.analchem.1c02220>.

(13) 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.; Labarta-Bajo, L.; Gentry, E. C.; Shalapour, S.; Kyle, K. E.; Puckett, S. P.; Watrous, J. D.; Carpenter, C. S.; Bouslimani, A.; Ernst, M.; Swafford, A. D.; Zúñiga, E. I.; Balunas, M. J.; Klassen, J. L.; Loomba, R.; Knight, R.; Bandeira, N.; Dorrestein, P. C. Mass Spectrometry Searches Using MASST. *Nature Biotechnology* **2020**, *38* (1), 23–26. <https://doi.org/10.1038/s41587-019-0375-9>.

(14) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In Silico Fragmentation for Computer Assisted Identification of Metabolite Mass Spectra. *BMC Bioinformatics* **2010**, *11* (1), 148. <https://doi.org/10.1186/1471-2105-11-148>.

(15) Allen, F.; Greiner, R.; Wishart, D. Competitive Fragmentation Modeling of ESI-MS/MS Spectra for Putative Metabolite Identification. *Metabolomics* **2015**, *11* (1), 98–110. <https://doi.org/10.1007/s11306-014-0676-4>.

(16) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relaunched: Incorporating Strategies Beyond in Silico Fragmentation. *Journal of Cheminformatics* **2016**, *8*, 3. <https://doi.org/10.1186/s13321-016-0115-9>.

(17) Blaženović, I.; Kind, T.; Torbašinović, H.; Obrenović, S.; Mehta, S. S.; Tsugawa, H.; Wermuth, T.; Schauer, N.; Jahn, M.; Biedendieck, R.; Jahn, D.; Fiehn, O. Comprehensive Comparison of in Silico MS/MS Fragmentation Tools of the CASMI Contest: Database Boosting Is Needed to Achieve 93% Accuracy. *Journal of Cheminformatics* **2017**, *9* (1), 32. <https://doi.org/10.1186/s13321-017-0219-x>.

(18) Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. LipidBlast in Silico Tandem Mass Spectrometry Database for Lipid Identification. *Nature Methods* **2013**, *10* (8), 755–758. <https://doi.org/10.1038/nmeth.2551>.

(19) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. Metabolite Identification and Molecular Fingerprint Prediction Through Machine Learning. *Bioinformatics (Oxford, England)* **2012**, *28* (18), 2333–2341. <https://doi.org/10.1093/bioinformatics/bts437>.

(20) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proceedings of the National Academy of Sciences* **2015**, *112* (41), 12580–12585. <https://doi.org/10.1073/pnas.1509788112>.

(21) Ludwig, M.; Dührkop, K.; Böcker, S. Bayesian Networks for Mass Spectrometric Metabolite Identification via Molecular Fingerprints. *Bioinformatics (Oxford, England)* **2018**, *34* (13), i333–i340. <https://doi.org/10.1093/bioinformatics/bty245>.

(22) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nature Methods* **2019**, *16* (4), 299–302. <https://doi.org/10.1038/s41592-019-0344-8>.

(23) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nature Genetics* **2000**, *25* (1), 25–29. <https://doi.org/10.1038/75556>.

(24) 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. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *Journal of Cheminformatics* **2016**, *8* (1), 61. <https://doi.org/10.1186/s13321-016-0174-y>.

(25) Blaženović, I.; Kind, T.; Sa, M. R.; Ji, J.; Vaniya, A.; Wancewicz, B.; Roberts, B. S.; Torbašinović, H.; Lee, T.; Mehta, S. S.; Showalter, M. R.; Song, H.; Kwok, J.; Jahn, D.; Kim, J.; Fiehn, O. Structure Annotation of All Mass Spectra in Untargeted Metabolomics. *Analytical chemistry* **2019**, *91* (3), 2155–2162. <https://doi.org/10.1021/acs.analchem.8b04698>.

(26) Ernst, M.; Kang, K. B.; Caraballo-Rodríguez, A. M.; Nothias, L.-F.; Wandy, J.; Chen, C.; Wang, M.; Rogers, S.; Medema, M. H.; Dorrestein, P. C.; van der Hooft, J. J. J. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* **2019**, *9* (7). <https://doi.org/10.3390/metabo9070144>.

(27) Lee, J.; da Silva, R. R.; Jang, H. S.; Kim, H. W.; Kwon, Y. S.; Kim, J.-H.; Yang, H. In Silico Annotation of Discriminative Markers of Three Zanthoxylum Species Using Molecular Network Derived Annotation Propagation. *Food chemistry* **2019**, *295*, 368–376. <https://doi.org/10.1016/j.foodchem.2019.05.099>.

(28) Sha, B.; Schymanski, E. L.; Ruttkies, C.; Cousins, I. T.; Wang, Z. Exploring Open Cheminformatics Approaches for Categorizing Per- and Polyfluoroalkyl Substances (PFASs). *Environmental science. Processes & impacts* **2019**, *21* (11), 1835–1851. <https://doi.org/10.1039/c9em00321e>.

(29) Tripathi, A.; Vázquez-Baeza, Y.; Gauglitz, J. M.; Wang, M.; Dührkop, K.; Nothias-Esposito, M.; Acharya, D. D.; Ernst, M.; van der Hooft, J. J. J.; Zhu, Q.; McDonald, D.; Brejnrod, A. D.; Gonzalez, A.; Handelsman, J.; Fleischauer, M.; Ludwig, M.; Böcker, S.; Nothias, L.-F.; Knight, R.; Dorrestein, P. C. Chemically Informed Analyses of Metabolomics Mass Spectrometry Data with Qemistree. *Nature Chemical Biology* **2021**, *17* (2), 146–151. <https://doi.org/10.1038/s41589-020-00677-3>.

(30) Wishart, D. S. Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nature Reviews. Drug Discovery* **2016**, *15* (7), 473–484. <https://doi.org/10.1038/nrd.2016.32>.

(31) Guma, M.; Tiziani, S.; Firestein, G. S. Metabolomics in Rheumatic Diseases: Desperately Seeking Biomarkers. *Nature Reviews. Rheumatology* **2016**, *12* (5), 269–281. <https://doi.org/10.1038/nrrheum.2016.1>.

(32) Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Briefings in Bioinformatics* **2019**, *20* (2), 492–503. <https://doi.org/10.1093/bib/bbx124>.

(33) Neumann, U.; Genze, N.; Heider, D. EFS: An Ensemble Feature Selection Tool Implemented as R-package and Web-Application. *BioData Mining* **2017**, *10* (1), 21. <https://doi.org/10.1186/s13040-017-0142-8>.

(34) 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. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. *Nature Biotechnology* **2021**, *39* (4), 462–471. <https://doi.org/10.1038/s41587-020-0740-8>.

(35) Platten, M.; Nollen, E. A. A.; Röhrig, U. F.; Fallarino, F.; Opitz, C. A. Tryptophan Metabolism as a Common Therapeutic Target in Cancer, Neurodegeneration and Beyond. *Nature Reviews Drug Discovery* **2019**, *18* (5), 379–401. <https://doi.org/gfvk74>.

(36) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. Mass Spectral Molecular Networking of Living Microbial Colonies. *Proceedings of the National Academy of Sciences* **2012**, *109* (26), E1743–E1752. <https://doi.org/10.1073/pnas.1203689109>.

(37) Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A. SIRIUS: Decomposing Isotope Patterns for Metabolite Identification. *Bioinformatics* **2009**, *25* (2), 218–224. <https://doi.org/10.1093/bioinformatics/btn603>.

(38) Dührkop, K.; Böcker, S. Fragmentation Trees Reloaded. In *Research in Computational Molecular Biology*; Przytycka, T. M., Ed.; Springer International Publishing: Cham, 2015; Vol. 9029, pp 65–79. <https://doi.org/10.1007/978-3-319-16706-0_10>.

(39) Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S. Database-Independent Molecular Formula Annotation Using Gibbs Sampling Through ZODIAC. *Nature Machine Intelligence* **2020**, *2* (10), 629–641. <https://doi.org/10.1038/s42256-020-00234-6>.

(40) Smyth, G. K. Limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., Dudoit, S., Eds.; Springer-Verlag: New York, 2005; pp 397–420. <https://doi.org/10.1007/0-387-29362-0_23>.

(41) Law, C. W.; Zeglinski, K.; Dong, X.; Alhamdoosh, M.; Smyth, G. K.; Ritchie, M. E. A Guide to Creating Design Matrices for Gene Expression Experiments. *F1000Research* **2020**, *9*, 1444. <https://doi.org/10.12688/f1000research.27893.1>.

(42) Xie, Y.; Dervieux, C.; Riederer, E. *R Markdown Cookbook*; Chapman and Hall/CRC: Boca Raton, Florida, 2020.

(43) Hoffmann, M. A.; Nothias, L.-F.; Ludwig, M.; Fleischauer, M.; Gentry, E. C.; Witting, M.; Dorrestein, P. C.; Dührkop, K.; Böcker, S. High-Confidence Structural Annotation of Metabolites Absent from Spectral Libraries. *Nature Biotechnology* **2021**. <https://doi.org/10.1038/s41587-021-01045-9>.

(44) Guha, R. Chemical Informatics Functionality in R. *Journal of Statistical Software* **2007**, *18* (6).

(45) Temple Lang, D. *RCurl: General Network (HTTP/FTP/...) Client Interface for R*; 2022.

(46) Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. InChIKey Collision Resistance: An Experimental Testing. *Journal of Cheminformatics* **2012**, *4* (1), 39. <https://doi.org/10.1186/1758-2946-4-39>.

(47) Lai, J.; Huang, L.; Bao, Y.; Wang, L.; Lyu, Q.; Kuang, H.; Wang, K.; Sang, X.; Yang, Q.; Shan, Q.; Cao, G. A Deep Clustering-Based Mass Spectral Data Visualization Strategy for Anti-Renal Fibrotic Lead Compound Identification from Natural Products. *The Analyst* **2022**, *147* (21), 4739–4751. <https://doi.org/10.1039/D2AN01185A>.

(48) Wozniak, J. M.; Mills, R. H.; Olson, J.; Caldera, J. R.; Sepich-Poore, G. D.; Carrillo-Terrazas, M.; Tsai, C.-M.; Vargas, F.; Knight, R.; Dorrestein, P. C.; Liu, G. Y.; Nizet, V.; Sakoulas, G.; Rose, W.; Gonzalez, D. J. Mortality Risk Profiling of Staphylococcus Aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures. *Cell* **2020**, *182* (5), 1311–1327.e14. <https://doi.org/10.1016/j.cell.2020.07.040>.

(49) Chen, Y.-H.; Bi, J.-H.; Xie, M.; Zhang, H.; Shi, Z.-Q.; Guo, H.; Yin, H.-B.; Zhang, J.-N.; Xin, G.-Z.; Song, H.-P. Classification-Based Strategies to Simplify Complex Traditional Chinese Medicine (TCM) Researches Through Liquid Chromatography-Mass Spectrometry in the Last Decade (2011): Theory, Technical Route and Difficulty. *Journal of Chromatography A* **2021**, *1651*, 462307. <https://doi.org/10.1016/j.chroma.2021.462307>.

(50) Krautbauer, S.; Eisinger, K.; Wiest, R.; Liebisch, G.; Buechler, C. Systemic Saturated Lysophosphatidylcholine Is Associated with Hepatic Function in Patients with Liver Cirrhosis. *Prostaglandins & Other Lipid Mediators* **2016**, *124*, 27–33. <https://doi.org/10.1016/j.prostaglandins.2016.06.001>.

(51) Melone, M. A. B.; Valentino, A.; Margarucci, S.; Galderisi, U.; Giordano, A.; Peluso, G. The Carnitine System and Cancer Metabolic Plasticity. *Cell Death & Disease* **2018**, *9* (2), 228. <https://doi.org/10.1038/s41419-018-0313-7>.

(52) Huang, L.; Lyu, Q.; Zheng, W.; Yang, Q.; Cao, G. Traditional Application and Modern Pharmacological Research of Eucommia Ulmoides Oliv. *Chinese Medicine* **2021**, *16* (1), 73. <https://doi.org/gnmwxx>.

(53) Krettler, C. A.; Thallinger, G. G. A Map of Mass Spectrometry-Based *in Silico* Fragmentation Prediction and Compound Identification in Metabolomics. *Briefings in Bioinformatics* **2021**, *22* (6), bbab073. <https://doi.org/10.1093/bib/bbab073>.

(54) Knuplez, E.; Marsche, G. An Updated Review of Pro- and Anti-Inflammatory Properties of Plasma Lysophosphatidylcholines in the Vascular System. *International Journal of Molecular Sciences* **2020**, *21* (12), E4501. <https://doi.org/10.3390/ijms21124501>.

(55) Park, D. W.; Kwak, D. S.; Park, Y. Y.; Chang, Y.; Huh, J. W.; Lim, C.-M.; Koh, Y.; Song, D.-K.; Hong, S.-B. Impact of Serial Measurements of Lysophosphatidylcholine on 28-Day Mortality Prediction in Patients Admitted to the Intensive Care Unit with Severe Sepsis or Septic Shock. *Journal of Critical Care* **2014**, *29* (5), 882.e5–11. <https://doi.org/10.1016/j.jcrc.2014.05.003>.

(56) Drobnik, W.; Liebisch, G.; Audebert, F.-X.; Fröhlich, D.; Glück, T.; Vogel, P.; Rothe, G.; Schmitz, G. Plasma Ceramide and Lysophosphatidylcholine Inversely Correlate with Mortality in Sepsis Patients. *Journal of Lipid Research* **2003**, *44* (4), 754–761. <https://doi.org/10.1194/jlr.M200401-JLR200>.

(57) Perino, A.; Demagny, H.; Velazquez-Villegas, L.; Schoonjans, K. Molecular Physiology of Bile Acid Signaling in Health, Disease, and Aging. *Physiological Reviews* **2021**, *101* (2), 683–731. <https://doi.org/10.1152/physrev.00049.2019>.

(58) Wang, B.; Tontonoz, P. Liver X Receptors in Lipid Signalling and Membrane Homeostasis. *Nature Reviews. Endocrinology* **2018**, *14* (8), 452–463. <https://doi.org/10.1038/s41574-018-0037-x>.

(59) Zhang, Q.; Yao, D.; Rao, B.; Jian, L.; Chen, Y.; Hu, K.; Xia, Y.; Li, S.; Shen, Y.; Qin, A.; Zhao, J.; Zhou, L.; Lei, M.; Jiang, X.-C.; Cao, Y. The Structural Basis for the Phospholipid Remodeling by Lysophosphatidylcholine Acyltransferase 3. *Nature Communications* **2021**, *12* (1), 6869. <https://doi.org/10.1038/s41467-021-27244-1>.

(60) Huang, L.; Lyu, Q.; Zheng, W.; Yang, Q.; Cao, G. Traditional Application and Modern Pharmacological Research of Eucommia Ulmoides Oliv. *Chinese medicine* **2021**, *16* (1), 73. <https://doi.org/10.1186/s13020-021-00482-7>.

(61) Huang, Y.-X.; Liu, E.-W.; Wang, L.; Huo, Y.; Wang, Q.; Olaleye, O.; Wang, T.; Gao, X.-M. LC/MS/MS Determination and Pharmacokinetic Studies of Six Compounds in Rat Plasma Following Oral Administration of the Single and Combined Extracts of Eucommia Ulmoides and Dipsacus Asperoides. *Chinese journal of natural medicines* **2014**, *12* (6), 469–476. <https://doi.org/10.1016/S1875-5364(14)60073-X>.

(62) Hu, F.; An, J.; Li, W.; Zhang, Z.; Chen, W.; Wang, C.; Wang, Z. UPLC-MS/MS Determination and Gender-Related Pharmacokinetic Study of Five Active Ingredients in Rat Plasma After Oral Administration of Eucommia Cortex Extract. *Journal of ethnopharmacology* **2015**, *169*, 145–155. <https://doi.org/10.1016/j.jep.2015.04.007>.

(63) Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T.; Hirai, M. Y.; Saito, K. RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-based Data Resource and Database. *Phytochemistry* **2012**, *82*, 38–45. <https://doi.org/10.1016/j.phytochem.2012.07.007>.