# MCnebula: Critical chemical classes for the classification 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, 4\***

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.

4 The Third Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, 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 a large amount of time on data analysis, especially for system biology. A framework called Multiple-Chemical nebula (MCnebula) was developed herein to facilitate mass spectrometry data analysis process by focusing on critical chemical classes and visualization in multiple dimensions. This framework consisted of the three vital steps as follows: (1) abundance-based classes (ABC) selection algorithm, (2) critical chemical classes to classify ‘features’ (compounds), and (3) visualization as multiple Child-Nebulae (network graph) with annotation, chemical classification, and structure. Notably, MCnebula can be used to explore the classification and structural characteristic of unknown compounds beyond the limit of spectral library. Moreover, it is intuitive and convenient for pathway analysis and biomarker discovery because of its function of ABC selection and visualization. MCnebula was implemented in the R language. A series of tools in the R packages was provided 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. The broad utility of MCnebula was illustrated, 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. A plant-derived dataset of herbal *Eucommia ulmoides* was investigated to achieve a rapid unknown compound annotation and discovery.

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

## Introduction

The analysis of untargeted liquid chromatography/tandem mass spectrometry (LC-MS/MS) dataset is complicated because of the massive of data volume, complexity of spectra, and structural diversity of compounds. In the past decades, many researchers have attempted to address the issues. Many technical software or web-based interfaces have been 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). False-positive and false-negative results were reduced through the implementation of algorithms 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 the problem of the quick and accurate identification of a large number of compounds.

Until today, several strategies have been developed for identifying compounds with fragmentation spectra. **1)** Spectral library matching. Several public available databases were built to settle this matter by 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). These fragmentation spectra are available via their web servers, third-party platform (e.g., [CompMass](http://prime.psc.riken.jp/compms/msdial/main.html" \l "MSP%3E)), or specific tools (MASST)[13](#ref-2020cm). However, compared with structure database (PubChem harbors over 100 million records), the spectral library has a very small size, thus limiting the application of mass spectrometry. This barrier was addressed via **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 are collated in silico fragmentation spectra and are publicly available [18](#ref-2013w). **3)** In silico prediction with matching learning. At present, the algorithms allow machine training from reference mass dataset or libraries and learn how to predict chemical fingerprints or principles 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 in identifying metabolites beyond the scope of spectral libraries. While in silico tools boost chemical identification, it still lacks a proper framework that could incorporate and leverage SIRIUS 4 into a user-friendly way for biological research, such as biomarker discovery and pathway analysis of mass spectral dataset. The manual compound annotation and screening of biomarkers are both time-consuming, and the results are impressed by subjective factors. Molecular networking is becoming increasingly popular because of its visualization and data transparency. Molecular networking is a spectral correlation and visualization method that can detect spectra from related molecules (so-called spectral networks), even when the spectra do not match with 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 can be traced back to the middle of the last century. The Chemical Fragmentation Coding system was first developed by Derwent World Patent Index in 1963. Thereafter, chemical classifications such as Gene Ontology (GO)[23](#ref-2000g), which was organized with taxonomy and ontology, was proposed systematically[24](#ref-2016). ClassyFire is popular for compound annotation in LC-MS dataset analysis because of its computational availability and systematicness[25](#ref-2019bt)–[28](#ref-2019bq). The taxonomy and ontology are robust and useful for chemistry. A hierarchical classification-based method called Qemistree has been 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 because of 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 effectively assess health outcomes[31](#ref-2016ar). These statistical approaches mainly involve classical statistical and artificial intelligence models (e.g., random forests)[32](#ref-2019bv). Both approaches introduce specific biases, owing to the complexity of feature set or algorithmic stability[33](#ref-2017i). Furthermore, the analysis at feature level is unable to profile systematic effects on metabolites unbiased[34](#ref-duhrkop_systematic_2021). Therefore, the analysis at chemical classified level may be a good settlement. However, the differences of metabolites at the same classified level need to be noted. For example, small-molecules belonging to ‘Indoles and derivatives’ have structure-dependent affection on aryl hydrocarbon receptor[35](#ref-2019c). Different structural characteristics will lead to diverse activities. This condition can be addressed by integrating both ‘feature’ level statistic and classified level assessment.

In addition to chemical classifying and statistical analysis, clustering visualization is a popular tool for untargeted mass spectrometry data analysis. Over the last decade, GNPS is becoming an increasingly popular clustering visualization tool based on MS dataset. GNPS applies molecular networking connecting mass spectra of molecules based on the similarity of their fragmentation patterns[36](#ref-2012a). Unfortunately, the molecular networking of GNPS mainly depend on spectral similarity instead of structural or classified similarity of the compounds. For example, flavonoids consist of an aromatic ring joined to an oxygenated heterocyclic ring linked to a phenyl group, which are expected to be clustered together because of its specific class and structural similarity. However, some compounds belonging to flavonoids are absent from the cluster of other flavonoids compounds [34](#ref-duhrkop_systematic_2021). Thus, clustering visualization in a classified level is a good 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. Nowadays, with the development of automatic classified in silico tools[24](#ref-2016), the visualization strategy can be revolutionized with higher confidence in classified level.

Accordingly, a comprehensive framework, named MCnebula, was proposed herein for untargeted LC-MS/MS dataset analysis. MCnebula integrates a new abundance-based classes (ABC) selection algorithm for chemical classes selection. The ABC selection algorithm involves the following principles: (1) applies an initial filtering to thousands of chemical classes based on the predicted probability, (2) regards all ‘features’ as a whole, examines the number and abundance of ‘features’ of each chemical classification (classification at different levels, classification of sub-structure and dominant structure), and then selects representative classes, and (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 are important for the subsequent analysis. They can be visualized as Child-Nebulae, and these chemical classes/Nebulae can be used 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 R language and can be easily integrated into the diverse biological analysis workflow of R. MCnebula (updated to MCnebula2, which includes more tools such as ABC selection algorithm, Nebula visualization, statistical analysis, and output report) was written primarily in the S4 system of object-oriented programming. It allowed all data for one-button analysis from the beginning to the end, thus facilitating data processing. In addition to the basic function of MCnebula, an additional ‘exMCnebula2’ package was provided for downstream analysis. This package contains all the analysis tools used in this study such as pathway enrichment analysis, heatmap clustering analysis, spectral visualization analysis, and chemical information query. The downstream analysis of untargeted LC/MS-MS is complex and varies across different 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 to demonstrate the broad utility of our method. These dataset include a human-derived serum dataset that is correlated with mortality risk profiling of *Staphylococcus aureus* Bacteremia (SaB) and a plant-derived herbal dataset that is 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. The annotation of these features is time-consuming, and extraction of useful information for further biological research is challenging. In response to these issues, MCnebula, which has an abundance-based class (ABC) selection algorithm, was presented 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. The molecular formula was determined based on the molecular weight of elements and possible fragmentation pattern of the MS/MS spectrum by SIRIUS. The exact chemical structure was then searched from compound structure databases. However, this process is often uncertain, because 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, candidates for molecular formula, structure, and chemical classes were obtained. Some candidates are correct, while others are not. The correct predictions for molecular formula and chemical structure are unique, and the chemical classification may have multiple correct predictions belonging to different classes of hierarchy. The scoring method depends on personalized research purpose. Scores can be used based on various criteria such as isotopes, mass error, structural similarity, or chemical classes to rank and filter the candidates. Given the availability of numerous score systems, low-scoring candidates were filtered out, and those with higher scores, which are more likely to be the correct compound, were focused on. However, in most cases, the top candidates from the three scoring systems are not always consistent. Hence, a ‘specific candidate’ was chosen as a reference in the user-selected scoring system, and its chemical information was retrieved from the other systems for data integration. Unique molecular formulae and chemical structure were obtained for reference by scoring and ranking, while chemical classes require more work.

**Chemical classification**. Compounds with MS/MS spectrum can be classified based on their overall structure or local structure, and they can be referred to as the dominant structure and substructure, respectively. Therefore, in the chemical classification system, compounds can be classified based on their dominant structure, as well as their substructure. When the dominant structure is unavailable, or the MS/MS spectral fragment is insufficient, the compounds can be classified using substructure information to gain knowledge about the compound. Note: The classification of compounds based on their dominant structure is straightforward. For example, Taxifolin is classified as a Flavone rather than a phenol, although its local structure contains a phenol substructure. Compounds can be classified based on their dominant structure, because it is more concise and provides more information. However, during the MS/MS spectral analysis, compounds can only be classified based on their substructure sometimes.

**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 (Fig. 6).

* Create Stardust Classes (Inner filter). By using the posterior probability of classification prediction data for each feature, simple threshold, or absolute conditions were set to filter the chemical classes through a method called ‘inner’ filtering.
* Cross-filter Stardust Classes. This method involves the combination of 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 for the comparison of 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, in which the Parent-Nebula represents all features combined, and the Child-Nebulae represents individual classes with their ‘features’.

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

*Cross-filtering by ‘quantity’* (abundance selection): The ‘features’ quantity limitation was 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 were used as thresholds. Chemical classes with too many or too few ‘features’ were 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 was 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 () was calculated using the equation , 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’, namely, the expected Goodness value of ‘tolerance’ and the actual Goodness, which are related to the parameter ‘cutoff’.

Goodness assessment can be assigned to multiple target attributes. The chemical class is retained only if it passes the Goodness assessment of all target attributes. This step mainly aims 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 was set for chemical classification, and groups within this range were 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) and the value of the parameter ‘identical\_factor’ () were assessed as follows:

: 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. This step aims 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 the class in which the potential compound belongs to from the LC-MS/MS spectra.

**Network graph presentation**. The features and their annotations were 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 was 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. For the ease of visualization among users, the ‘ggset’ data class, which stores pre-defined ggplot2 plotting functions and parameters for visualizing Nebulae, was developed. 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 the differential expression analysis of RNA-sequence and microarray data[40]. Then, they are stored as package for the differential analysis of metabolomics data. The gene expression and feature quantification matrices of LC-MS are similar, and both have phenotypic variables (sample information) and dependent variables (gene expression or feature quantification values). The use of our method is appropriate for the statistical analysis of the feature quantification of experimental designs, in which explanatory variables are factorial variables, and the design matrix is without an intercept [41].

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

**Reporting system**. MCnebula includes a reporting system that enables the analysis process to be exported in PDF 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 requirements. Furthermore, the ‘rmarkdown’ R package [43] was 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. The effect of other version problems on MCnebula was eliminated by designing its application programming interface (API) for various SIRIUS versions.

### MCnebula evaluation

**Spectra dataset for evaluation**. The performance of MCnebula was evaluated using the spectra from the GNPS MS/MS library (http://prime.psc.riken.jp/compms/msdial/main.html#MSP). Overfitting during library match evaluation was prevented by adding ‘noise’ to the MS/MS spectra[44]. Two models of noise were simulated, including medium 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[45]. 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 (version 4.9.12) was applied for computation, and MS/MS spectra with empty fragmentation peaks were filtered out. In total, 7,524 out of 8,782 compounds were left for evaluation. ClassyFire was used to assess the classification accuracy[24]. After in silico annotation, structure annotation, International Chemical Identifier Key (InChIKey), and other metadata of these compounds were obtained. Considering that ClassyFire only supports chemical identities with structures that have been classified in its server previously, the first hash block of InChIKeys (InChIKey planar, which represents the molecular skeleton) was used to query the PubChem API (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest) [46]. This process provided us with all the possible InChIKeys of isomerism (stereo, isotopic substitution) [47]. The classification of small molecules depends on their molecular skeleton. Hence, chemical identities that share the same InChIKey planar are identical in classification. The InChIKey list was imported into ClassyFire to obtain chemical classification. In the R script, once any InChIKey of isomerism matched the classified data in the database, the acquisition status for this molecular skeleton was turned off. Finally, all these chemical annotations were collated, integrated, and assigned as standard reference.

Considering the differences in algorithms and classified results, the MCnebula and benchmark methods were evaluated separately. Given that sub-structural classification was not available for the benchmark method, these classes were excluded during the evaluation analysis. Nevertheless, some compounds within the remaining classes may still be classified into sub-structural classes. Three levels were assigned for evaluation, including ‘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’, while the false positive (FP) was assigned as ‘False’.

The identification of classes or structures was evaluated by merging the results with a standard reference by InChIKey planar. For the evaluation of chemical structure identification, a structure was considered as ‘True’ if it matched the chemical structure identified by InChIKey planar. However, this evaluation neglected stereochemistry because of 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 for the analysis of LC-MS/MS datasets starting from the raw data obtained from the sample and going through the various stages of analysis to obtain a complete analysis report (Fig. 2). The analysis process followed the general MCnebula analysis templates from filtering the candidates of chemical formula, structural formulae, and chemical classes until the creation of visual Nebulae. It also allowed for custom advanced analysis with the help of chemical class focused Nebulae to perform statistical analysis, feature selection, focus on critical metabolites (compounds) and their structural characteristics, pathway enrichment, and querying compound synonyms. The workflow of MCnebula in function with other popular public available methods was also evaluated. In terms of the evaluation of our chosen indicators, which cover identification and classification, MCnebula has a wider scope of applicability (Tab. S1).

#### Method evaluation

**Classified accuracy.** A publicly available reference spectral library was used to assess the accuracy of classification by MCnebula. The direct use of such a reference spectral library may lead to over-fitting during the evaluation. Noise was simulated to eliminate this consequence. Simulation of noise, which involves the addition of invalid noisy data to the reference spectrum or numerically shifting the existing data, also simulates data acquisition similar to a real scenario. Considering the different acquisition conditions, the spectral data in the real case will be noisier compared with the reference spectrum. By adding noise to the reference spectral library, three datasets become available for evaluation (origin, medium noise, and high noise dataset; commonly 7,524 compounds [spectra]). All three datasets were analyzed using MCnebula. Considering the richness of the compounds in the reference spectra, for the origin dataset, 152 chemical classes (each with a corresponding compound to be evaluated) were obtained using the 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. Comparison with other methods was facilitated by selecting only chemical classes that are likely to be dominant structures for evaluation. A total of 37 of such chemical classes were selected for evaluation. MCnebula was evaluated objectively by choosing the molecular networking provided by GNPS, 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. By 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.

The three datasets were uploaded to the GNPS server, and the results obtained were evaluated. For origin dataset, GNPS resulted in 44 chemical classes (parallel to MCnebula, with at least 50 compounds per chemical class). A total of 19 common classes were obtained. 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. 3a). 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. 3a). For the indicators of precision, the performance of classification was assessed by combining the level of the stability to calculate the relative false rate, rather than the absolute false rate. Then, this parameter was used to estimate precision. The relative false rate effectively simulated the actual application to LC-MS/MS analysis, because the actual spectral data contains noise and many unknown compounds that cannot be identified by spectral matching. In this context, MCnebula outperformed the GNPS in terms of the evaluation of the precision in three datasets (MCnebula: 69.8%, 67.1%, 67.4%; GNPS: 48.1%, 51.2%, 52.4%) (Fig. 3a). The recall was estimated using relative false rate as well. As shown in Fig. 3a, MCnebula (82.2%, 81.6%, and 81.6%) outperformed GNPS (51.2%, 52.7%, and 53.3%). In addition to the three indicators mentioned above, MCnebula and GNPS were compared at the individual level for the 19 chemical classes (Fig. 3b). Remarkably, MCnebula was more stable to noise than GNPS.

**Identified accuracy.** By using the MCnebula workflow, for the origin dataset containing 8,057 compounds (precursor ions m/z < 800), all of these compounds were predicted with chemical molecular formulae, in which 6,610 compounds were predicted with chemical structure. These chemical structures were evaluated for accuracy in a classified context. For the 37 chemical classes (Fig. 3c), the average false rate of identification was 37%, 156 compounds were identified in average. Among them, most of the identified false rate were between 30% to 40%, but 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 by providing 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%, 139 compounds were identified in average (Fig. 3c). The accuracy of the identification of compounds in each chemical class obtained by MCnebula. Notably, MCnebula itself does not contain any module for identification, and it only utilizes the top scoring candidate from the SIRIUS predicted results for annotation. For an extensive evaluation on identification, more details can be found in the publication and our previous related work[22](#ref-duhrkop_sirius_2019),[50](#ref-lai_deep_2022).

#### Data analysis

**Serum metabolic analysis.** The application of MCnebula in metabolism was illustrated by re-analyzing the serum data from Wozniak et al.[42](#ref-2020s). The serum samples were collected from patients in-hospital infected with SaB or not and healthy volunteers. Overall, the samples were divided into control groups, which include NN (non-hospital, non-infected) and HN (hospital, non-infected), and infection groups, which include HS (hospital, survival), HM (hospital, mortality).

A total of 7,680 ‘features’ were detected while running with LC-MS pre-processing on the serum dataset. After predicting the compounds by MS/MS spectra (with SIRIUS software), subsequent analysis was performed using MCnebula. Among these, 6501 ‘features’ were annotated with predicted molecular formula, and 3,449 ‘features’ were annotated with the predicted chemical structure. By using the ABC selection algorithm, more than 1,000 chemical classes were filtered out by applying the ‘inner filter’ module (see method section of ABC selection algorithm). A total of 508 chemical classes were further filtered out while performing ‘cross filter’. For the 41 remaining chemical classes, 19 chemical classes were manually filtered out, leaving 22 chemical classes to make up the Nebula-Index, which were further visualized as Child-Nebulae. Notably, the 527 filtered out (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). The analysis of Child-Nebulae provided insight into the chemical classes contained in the serum dataset. More information was obtained from Child-Nebulae by performing 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. 4). By combining the features selection algorithm about the *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). Th figure shows that the overall level of ‘Bile acids, alcohols and derivatives’ (BAs) class and ‘Acyl carnitines’ (ACs) (Fig. 5a and b) class increased remarkably, whereas the overall level of ‘Lysophosphatidylcholines’ (LPCs) class decreased remarkably. BAs, ACs, and LPCs are associated with liver dysfunction, imbalance of intestinal microphylactic homeostasis, and mortality risk[42](#ref-2020s),[51](#ref-2021db),[52](#ref-2016at).

Through the 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. 5c, Fig. S4). Subsequently, cluster heat map analysis and pathway enrichment analysis were performed on the compounds of these three classes. As shown in the clustering heat map (Fig. 6), the control group of ACs and BAs were remarkably separated from the infection group, indicating the infection relevance of ACs and BAs to SaB. By contrast, LPCs did not show remarkable SaB infection relevance or mortality relevance possibly because of the general consistency of this class of compounds for SaB disease. Pathway enrichment analysis was carried out 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’ (Fig. S5b). Among them, GCS is a class of compounds with the same parent nucleus. The results for LPCs suggest 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. However, a fundamental role of ACs in tuning the switch between the glucose and fatty acid metabolism was reviewed[53](#ref-2018bi). Their function implemented via bi-directional transport of acyl moieties between cytosol and mitochondria (Fig. S5a).

In the research of Wozniak et al.[42](#ref-2020s), five ACs compounds were identified. In addition, four top metabolites such as 2-hexadecanoylthio-1-ethylphosphorylcholine (HEPC), sphingosine-1-phosphate (S1P), decanoyl-carnitine, and L-Thyroxine (T4) were also identified. In our reanalysis, all identifications were consistent, except for HEPC (see ‘Data and code availability’ section for the report of serum dataset analysis). Based on the 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 terms of structure, but they have distinct in elemental composition (corresponding to C23H48NO5PS and C23H48NO7P respectively). HEPC belongs to ‘Cholines’ (level 5) from ‘Organic nitrogen compounds’ (superclass) family, while LPC15:0 belongs to ‘Lysophosphatidylcholines’ (level 5) from ‘Lipids and lipid-like molecules’ family. As a part of the MCnebula workflow, sulfur is detectable for SIRIUS in isotopes pattern with high mass accuracy[37](#ref-bocker_sirius_2009). However, for the MS/MS spectra of ‘HEPC’, no candidate formula contains sulfur element. Overall, more compounds were identified with the MCnebula workflow, and many of the results were in line with the analysis of Wozniak et al[42](#ref-2020s). All the identified compounds were collated in Table S2 (filtered with Tanimoto similarity > 0.5 and de-duplicated with the first hash block of InChIKey [molecular skeleton]). The compounds (top 50 of ensemble feature selection [EFS] and Mann-Whitney U [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.** MCnebula was used to interpretate structure diversity and chemical transformation during the traditional processing of a representative herbal medicine, *Eucommiae Cortex,* the peel of *Eucommia ulmoides Oliv. (E. ulmoides)*[54](#ref-2021n). After processing with saline water, it has long been used for the treatment of renal diseases in China, but the chemical basis still remained to be explored. With the help of ABC selection algorithm in MCnebula, 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). MCnebula was used 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, the ‘features’ of ID 1642, 1785, and 2321 were newly generated compounds, because the peak area levels before the processing were almost zero compared with those after processing. Their chemical structures are shown 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, the ‘feature’ of ID 1642 belongs to ‘Iridoids and derivatives’ (IAD), while the others belong to ‘Dialkyl ethers’ (DE; ID 1785) and ‘Phenylpropanoids and polyketides’ (PAP; ID 2321). The Child-Nebulae of IAD, DE, and PAP were annotated. 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; Fig. S9d, e, and f, respectively). Their levels of peak area decreased and then increased after the processing. Based on the chemical structures shown in Figs. S9d and e, 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 increase 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 compound ID 2110.

The methods of MCnebula that we have demonstrated for discovering significant compounds and discovering chemical changes can be applied to explore more compounds in Tab. S4. However, a detailed description was not provided here.

### Discussion

The analysis of LC-MS/MS data is challenging because of its large dataset and the potential information of the unknown compounds and the limited of reference spectral library. Researchers often require much time to map 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 to make the overall direction of the study unbiased. Moreover, 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 because of the loss of information, and filtering on the basis of chemical classes level can prevent the bias to some extent. The Child-Nebula, which was mapped on the basis of the chemical classes obtained by the ABC selection algorithm, achieved the visualization of the huge untargeted dataset as a single graph. 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 mainly used for LC-MS/MS data because of its high accuracy. The general classification of compounds is also based on it, that is, the chemical structure is first identified by spectral matching. Then, its chemical class is evaluated based on the chemical structure. Considering the limit of reference spectral library, classification techniques such as CANOPUS[34](#ref-duhrkop_systematic_2021), which wasincorporated 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 achieved the visualization of Child-Nebulae, thus allowing the exploration of unknown compounds beyond spectral library. The classification method of MCnebula was compared 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. The reality of MS/MS spectra is much closer to the condition with noise. Therefore, MCnebula can resist noise interference to some extent. After the evaluation, the accuracy of the identification was examined by MCnebula. The results confirmed that the accuracy of identification fluctuated around 70%, which was the same as SIRIUS[22](#ref-duhrkop_sirius_2019).

Serum metabolomics data was applied to illustrate that MCnebula can be used for pathway analysis and biomarker discovery. Most of the results were consistent with those of reference[42](#ref-2020s). Notably, more metabolites beyond the scope of spectral library matching were identified. 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 that AC compounds had correlation with SaB disease, and AC compounds were re-identified in our study. Wozniak et al. used a joint approach of EFS and MWU tests to screen top metabolites[42](#ref-2020s). When 50 top ‘features’ were 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., 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 compared 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. Additional classes such as ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs) were not concerned in previous study. LPCs have been extensively investigated in the context of inflammation and atherosclerosis development[52](#ref-2016at),[55](#ref-2020cv),[56](#ref-2014ao). In a recent review[55](#ref-2020cv), the complex roles of LPCs in vascular inflammation were well described, involving the context-dependent pro- or anti-inflammatory action, as well as the effect in innate immune cells and adaptive immune system. The decrease in LPCs was associated with wild range of diseases of increasing mortality risk[52](#ref-2016at). The investigation of sepsis indicated LPC concentrations in blood were correlated with severe sepsis or septic shock[56](#ref-2014ao). LPCs was inversely correlated with mortality in patients with sepsis[57](#ref-2003n). BAs’ disorder implied a liver dysfunction and imbalance of intestinal microphylactic homeostasis[58](#ref-2021dg). The chemical multiversity of BAs, which were discovered in the BAs’ child-nebula, were determined by the intestinal microbiome and allowed the complex regulation of adaptive responses in host. In the present 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, steroid-related classes, ‘Lineolic acids and derivatives’, and other fatty acid-related classes showed that liver plays a central role in SaB infection and mortality. Liver X receptors (LXRs) play pivotal roles in the transcriptional control of lipid metabolism[59](#ref-2018bd). LXRs modulate membrane phospholipid composition through the activation of lysophosphatidylcholine acyltransferase 3 (LPCAT3), which is directly related to LPCs[60](#ref-2021di). The above classes are correlated with LXRs[59](#ref-2018bd). However, LXRs’ specific role in SaB infection or mortality has not been reported and is 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* include lignans, iridoids, phenolics, flavonoids, steroid, and terpenoids[61](#ref-huang_traditional_2021). In the present study, the chemical classes obtained by ABC selection algorithm included ‘Lignans, neolignans and related compounds’ (LNARC), ‘Iridoids and derivatives’ (IAD), ‘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 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 were identified in chemical classes of LNARC and IAD (Tab. S1) have been reported in previous research about LC-MS/MS analysis of *E. ulmoides*[62](#ref-2014w),[63](#ref-2015v). Top features have been obtained 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. Therefore, it was related to two structurally similar compounds by transformation. The application of MCnebula in the analysis of plant-derived compounds was well illustrated by this example, particularly for the 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 have been developed[64](#ref-2012ac), the fragmentation spectra for comprehensive library match remain insufficient. 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, voluminous information of the unknown compounds, and the limited of reference spectral library. Thus, a framework called MCnebula was established 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 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 reached 70%. The broad utility of MCnebula was illustrated by 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. A plant-derived dataset of herbal *E. ulmoides* was also investigated 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, and medicine.

**Acknowledgements**

This work was financially supported by the National Natural Science Foundation of China (Nos.81973481 and 82274101), Zhejiang Province Traditional Chinese Medicine Science and Technology Project (Nos. LZ22H280001 and 2022ZQ033).