# MCnebula: Select critical chemical classes to classify and boost identification by visualization for downstream analysis of untargeted LC-MS/MS dataset analysis

# Abstract

   Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed the MCnebula approach to address this challenge by focusing on critical chemical classes and visualizing them in multiple dimensions: via abundance-based classes (ABC) selection algorithm, critical chemical classes were selected to classify ‘features’ (compounds); visualizing these chemical classes as multiple Child-Nebulae (network graph) and annotated them in depth with chemical classification and structure. Notably, MCnebula can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library; it can also be used for pathway spotting and biomarker discovery. MCnebula was implemented in the R language. We provide 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, output analysis reports, etc. For illustration, we conducted investigation of human-derived dataset of serum metabolomics by tracing structural classes of biomarkers so that facilitating metabolic pathway spotlight; we also investigated a plant-derived dataset of herbal *E. ulmoides* to achieve a rapid identification and discovery of chemistry.

# Introduction

   Analyzing untargeted liquid chromatography/tandem mass spectrometry (LC-MS/MS) dataset is complicated, due to the massive of data volume, complexity of spectrogram profiles, and structural diversity of compounds. In the past decades, a great deal of research were anchored to address the issues. Many technical software or web server attempted to provide a one-stop bulk solution for data analysis [[1](#ref-2020p)–[4](#ref-2016a)]. These solution apply or suggest flexible mass spectrogram processing tools or analogous algorithms [[5](#ref-2012d)–[8](#ref-2010)]. To reduce false-positive and false-negtive results, more algorithms have been implemented for peak deconvolution, feature selection or statistical filtering [[9](#ref-2017f)–[13](#ref-2017i)]. Per feature corresponding to compound within sample or parallel samples, and it prevalently equipped with fragmentation spectra to identification. In this context, researchers have to be confronted with a barrier: how to identify so many compounds accurately?

   Until today, several strategies were developed for identifying with fragmentation spectra, including **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)](https://mona.fiehnlab.ucdavis.edu/), Global Natural Products Society molecular networking (GNPS) [[4](#ref-2016a),[14](#ref-2010p)–[17](#ref-2020cp)]. In counterpart, these fragmentation spectra are available via their web severs, third-party platform (e.g., [CompMass](http://prime.psc.riken.jp/compms/msdial/main.html#MSP%3E)) or specific tools (MASST) [[18](#ref-2020cm)]. However, comparing with structure database (PubChem harbours over 100 million records), spectra library is too small in size that limit the application of mass spectrometry. To cross this barrier, **2)** Matching with fragmentation spectra of *in silico* simulation. *In silico* tools have been increasingly developed for simulating fragmentation spectra [[19](#ref-2010c)–[27](#ref-2020cn)]. Some databases such as MoNA collated *in silico* fragmentation spectra for public available [[28](#ref-2013w),[29](#ref-2015aj)]. **3)** *In silico* prediction with matching learning. Prevalently, the algorithms make machine train from reference mass dataset or libraries, then ‘learned’ how to predict chemical fingerprints or principles so as to retrieve the correct structure within structure database [[30](#ref-2012ab)–[35](#ref-2021cy)].

*In silico* methods are developing quickly. Up to now, the cutting-edge technology, SIRIUS 4 [[36](#ref-duhrkop_sirius_2019)], integrated with many advanced algorithms of artificial intelligence, has been reported accuracy rate reached 70% while retrieving in structure libraries. This method help to identify metabolites beyond the scope of spectra library. While *in silico* tools boost chemical identification, it is still lack of an approach that incorporating and leveraging the state-of-the-art technology into biological research, i.e. biomarker discovery in untargeted mass spectral dataset. Identification and screening of biomarkers manually are time-consuming and the results are impressed by subjective factors. In terms of identification, molecular networking is increasingly popular due to its visualization and data transparency [[4](#ref-2016a),[37](#ref-2020b)]. Thus, we proposed a preliminary idea, could clustering features for visualization based on chemical classification contribute to biomarker discovery or metabolic pathway spotlight?

   The history of classification in chemistry dates back to at least the middle of the last century. The Chemical Fragmentation Coding system developed by Derwent World Patent Index (DWPI) in 1963. Chemical classification like Gene Ontology (GO) [[38](#ref-2000g)], has been proposed, more systematically, organizing with taxonomy and ontology in recent years [[39](#ref-2016)]. ClassyFire is increasingly used for compound annotation either in mass dataset analysis or not due to its computation available and systematicness [[40](#ref-2019bt)–[48](#ref-2022al)]. The taxonomy and ontology for chemistry is beneficial. For example, a hierarchical classification-based method, termed Qemistree, was proposed to address chemical relationships at a dataset-wide level [[45](#ref-2021b)].

   Untargeted metabolomics emerged to profile cellular and organismal metabolism without prior knowledge dependence [[49](#ref-2016aq),[50](#ref-2017at)]. Researchers with the help of statistical methodologies from thousands of features screen out biomarker, towards pharmaceutical, physiological or pathological mechanisms [[51](#ref-2016ar),[52](#ref-2016ao)]. These statistical approaches involved classical statistic and artificial intelligence (e.g., random forests) [[53](#ref-2019bv),[54](#ref-2021de)]. Both approaches were impossible to avoid specific biases, owing to the complexity of feature set or algorithmic stability [[13](#ref-2017i)]. Further, evaluation in per feature level seemed incapable of profiling systematic effects in metabolites [[55](#ref-duhrkop_systematic_2021)]. In this view, analyzing at chemical classified level may be a comprehensive settlement. However, we can not neglect the differences of metabolites at the same classified level. For example, small-molecules belonging to ‘Indoles and derivatives’ harbour structural dependent affection on aryl hydrocarbon receptor (AHR) [[56](#ref-2019c)]. Different structural characteristics may lead to diverse activities. The settlement for that is integrating either ‘per feature’ level statistic or classified level assessment.

   In addition to chemical classifying and statistical analysis, clustering visualization was another popular tool for untargeted mass spectrometry data analysis. For over a decade, Global Natural Products Social Molecular Networking (GNPS) developed as such a popular tool. Depending on fragmentation spectra similarity, structural annotations are propagated via network-based method [[57](#ref-2012a)–[61](#ref-2021d)]. Unfortunately, molecular networking is a highly spectral similarity dependent method instead of compounds structural or classified similarity. For example, Flavonoids were expected to be clustered together as its specific class and structural similarity. However, in previous research, it has been reported that some Flavonoids happened to be absent from the cluster of many Flavonoids compounds [[55](#ref-duhrkop_systematic_2021)]. In this context, visualization in a classified perspective is a better choice for untargeted mass spectra dataset. Earlier in 2012, molecular networking was proposed with visualization for mass data analysis for the first time [[57](#ref-2012a)]. At that time, *in silico* tools for predicting compound classification by fragmentation spectra were not available. Nowadays, with the development of automatic classified *in silico* tools [[39](#ref-2016),[55](#ref-duhrkop_systematic_2021)], it is time for a revolution of the visualization strategy. With higher confidence in classified level.

   For above consideration, we proposed a comprehensive analysis workflow, named MCnebula, for untargeted LC-MS/MS dataset analysis. For the first time, MCnebula integrated an abundance-based classes (ABC) selection algorithm for chemical classes selection. The ABC selection algorithm: first, applied an initial filtering to thousands of chemical classes based on the predicted probability; 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; 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 be serve to the subsequent analysis: visualized as Child-Nebulae and focus on these chemical classes (Nebulae) for biomarker finding or chemistry discovery. The top ‘features’ depend on statistic 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 the identification module of MCnebula utilized the cutting-edge technology of SIRIUS software [[31](#ref-duhrkop_searching_2015),[36](#ref-duhrkop_sirius_2019),[55](#ref-duhrkop_systematic_2021),[62](#ref-bocker_sirius_2009)–[64](#ref-ludwig_database-independent_2020)], which exceeds the limitations of spectral library matching. MCnebula has been implemented in the R language, which means that it can be easily integrated into the diverse biological analysis workflow of R. MCnebula (MCnebula2) was written primarily in S4 system of object-oriented programming. This allows all data for one analysis to be concentrated in one object from initial initiation to the end of the analysis, facilitating data management and reproducible analysis. In addition to the MCnebula (MCnebula2 package included tools such as ABC selection algorithm, Nebula visualization, statistical analysis, and output report etc.), we also provide an additional ‘exMCnebula2’ package for downstream analysis, which contains all the analysis tools used in this study, including: pathway enrichment analysis, heat map 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 to 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 processing of herbal medicine.

# 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. 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 of 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 compounds 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.

   In order to evaluate MCnebula more objectively, we chose an existing method as a benchmark method for parallel comparison. We finally chose the molecular networking provided by GNPS (Global Natural Products Social Molecular Networking) as the benchmark method because we considered that it is also a visual analysis method that has been popular among chemists for a long time. In GNPS, the workflow of Feature-based molecular networking (FBMN) combined with MolNetEnhancer provides a way to classify compounds present in spectral data. This is a traditional classified method, which first identifies the exact chemical structure based on spectral data, and then determines the chemical class based on the 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, and relative false rate. For the classified number, MCnebula outperformed GNPS in three dataset (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: 10.5%, 19.8%; GNPS: 41.7%, 50.1%) (Fig. 2a). For the last indicator, 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. 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 relative false rate in three dataset (MCnebula: 30.2%, 32.9%, 32.6%; GNPS: 51.9%, 48.8%, 47.6%) (Fig. 2a). 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. These 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 [[36](#ref-duhrkop_sirius_2019)]). 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 [[36](#ref-duhrkop_sirius_2019)].

## Data analysis

**Serum metabolic analysis.** To illustrate the application of MCnebula in metabolism, we re-analyzed the serum data from Wozniak et al. [[65](#ref-2020s)]. The serum samples were collected from patients 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 molecular formula and further, 3449 ‘features’ were annotated with chemical structure. Using ABC selection algorithm: more than one thousand chemical classes were filtered out in module of ‘inner filter’ (see method section of ABC selection algorithm); 508 chemical classes were filtered out while performing ‘cross filter’; for the remaining 41 chemical classes, 19 chemical classes were manually filtered out; the final 22 chemical classes make up the Nebula-Index, and these chemical classes would be 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 [[65](#ref-2020s)–[67](#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, βGCS 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 [[68](#ref-2018bi)]. Their function implemented via bi-directional transport of acyl moieties Between cytosol and mitochondria (Fig. S5a).

   In research of Wozniak et al [[65](#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 belong to ‘Cholines’ (level 5) from ‘Organic nitrogen compounds’ (superclass) family, whereas LPC15:0 belong 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 [[62](#ref-bocker_sirius_2009)]. However, for the MS/MS spectra of ‘HEPC’, there was no candidate formula that harbouring 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 [[65](#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.** *Eucommia ulmoides Oliv.* (*E. ulmoides*) [[69](#ref-2021n)], as a traditional Chinese medicine (TCM), after being processed with saline water, was applied to the treatment of renal diseases for a long time in China. Due to its complex composition, discovering chemical changes during processing (such as processed with saline water) is challenging. MCnebula was instructive for addressing this challenge. With the help of ABC selection algorithm, the 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 MS/MS 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 after the processing. Their chemical structures were 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, 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 well been 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 increasing in the level of the compound of ID 1642 (its spectra were shown in Fig. S7 and S8) may also be associated with the reduction of the compound with 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 untargeted LC-MS/MS data is challenging because of its large data volume and the unknown nature of the potential information as well as the limited size of the current reference spectral library. Researchers often need to map out critical compounds from this black box to advance to the next step in biology, but this process was time consuming. We developed MCnebula to advance this process. MCnebula assists researchers in focusing on potential critical compounds 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 that the overall direction of the study is not biased; it is an effective guarantee for statistical analysis to produce top features for tracing analysis. Because the results of statistical analysis may be biased, filtering on the basis of chemical classes can prevent bias. The Child-Nebula mapped on the basis of the chemical classes obtained by the ABC selection algorithm makes it visualizing the huge untargeted dataset as a single graph. It is important to stress that the parameters of the ABC selection algorithm are 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 in the dataset and filter out those that are too large or too small in conceptual scope.

   For identification, spectral library matching is still the mainstay of MS/MS data analysis because of its high accuracy. The general classifying of compounds is also based on it, i.e., the chemical structure is first identified by spectral matching, and then its chemical class is evaluated based on the chemical structure. However, the classifying technique (CANOPUS [[55](#ref-duhrkop_systematic_2021)]) incorporated in MCnebula bypassed the step of first identifying the chemical structure and 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 visualization of Child-Nebulae to make it possible to explore unknown compounds. We have compared the classifying method of MCnebula with GNPS of which method relies on chemical structure identification. Since we choose the reference spectral data, which was included in the matching library of GNPS, for the origin dataset, the classifying method of MCnebula was not as good as GNPS in terms of accuracy and the classified number. However, when different levels of noise were added, the number of classified compounds of GNPS decreased remarkably compared to the stable performance of MCnebula. For the actual acquired MS/MS spectra, they were often not as good as the reference spectra and contained some noise. This means that the conditions we evaluate are actually present. Using MCnebula can avoid failure to classify due to noise. At the end of the evaluation, we examined the accuracy of the MCnebula identification, although its identification technologies belongs to SIRIUS [[36](#ref-duhrkop_sirius_2019)], and we confirmed that the accuracy of its identification fluctuated around 70%.

   We used serum metabolomics data to illustrate that MCnebula can be used for pathway spotting and Biomarker discovery. Most of our identification were in line with that of Wozniak et al., except 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 consistent in our re-identification, leaving one controversial. ACs compounds mentioned by Wozniak et al. for their correlation with SaB disease 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 [[65](#ref-2020s)]. It should be added that 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., we found a total of 37 overlaps, including the key metabolite of L-Thyroxine in the study of W et al. Top ‘features’ can differ depending on the feature selection algorithm. The reliability of the ‘binary comparison’ method is indirectly illustrated by our ranked results with those of Wozniak et al. In addition to the consistent parts, with MCnebula, we have more findings about other chemical classes associated with SaB disease. We discovered additional classes, i.e. ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs), that were not concerned in previous study. Previously, LPCs have been extensively investigated in the context of inflammation and atherosclerosis development [[67](#ref-2016at),[70](#ref-2020cv)–[72](#ref-2003n)]. In a recent review [[70](#ref-2020cv)], the complex roles of LPCs in vascular inflammation have been 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 [[67](#ref-2016at)]. The investigation of spesis indicated LPCs concentrations in blood were established correlation with severe sepsis or septic shock [[71](#ref-2014ao)]. LPCs was reported inversely correlate with mortality in sepsis patients [[72](#ref-2003n)]. BAs’ disorder implied a liver dysfunction and imbalance of intestinal microphylactic homeostasis [[73](#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) harboured pivotal roles in the transcriptional control of lipid metabolism [[74](#ref-2018bd)]. LXRs modulate membrane phospholipid composition through activation of lysophosphatidylcholine acyltransferase 3 (LPCAT3), which directly related to LPCs [[75](#ref-2021di)]. The above classes showed correlation with LXRs [[74](#ref-2018bd)]. Unfortunately, LXRs’s specific role in SaB infection or mortality has not been documented and beyond the scope of this research.

   In herbal dataset analysis, we showed a quick identification 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 [[76](#ref-2021bi)]. In our study, the chemical classes obtained by ABC selection algorithm include ‘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) [[39](#ref-2016)] and phenolics may 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 we identified in chemical classes of LNARC and IAD (Tab. S2) have been reported in previous research about LC-MS/MS analysis of *E. ulmoides* [[77](#ref-2014w)–[82](#ref-2021bt)]. 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 into the Child-Nebulae. We then hypothesized that it was related to two structurally similar compounds by transformation. The application of MCnebula in the analysis of plant-derived compounds was illustrated with this example of quick identification and exploration of chemical changes. In contrast to reference spectral library for human-derived metabolites, which are more likely to be commercially valuable, the spectra for plant-derived compounds was more scarce. Although some specific database of plant-derived compounds have been constructed [[15](#ref-2012ac),[16](#ref-2015ak)], there were lack of enough fragmentation spectra for comprehensive library match. With the help of MCnebula, mostly via retrieving structural libraries, a rapid and reliable resolution of complex compositions of plant-derived can be achieved.

   In this article, due to limited space, few examples demonstrated MCnebula application. Indeed, MCnebula has a great potential in the field of chemistry, pharmacy and medicine. The latter, beyond this article, e.g., fields of application include natural products, foodomics, environmental research etc. As an integrated visualization method, MCnebula possibly be more popular with biologists or chemists. MCnebula has been proposed and implemented in the R language with package MCnebula. In the future, its methods and application will be extensively expanded.

# Methods

## MCnebula algorithm

**Overall consideration** We knew that the analysis of untargeted LC-MS/MS dataset generally begin with feature detection. It detected ‘peaks’ as features in MS1 (MASS level 1) data. Each feature may represents a compound, and assigned with MS2 (MASS level 2) spectra. The MS2 spectra was used to find out the compound identity. The difficulty lied in annotating these features to discover their compound identity, mining out meaningful information, so as to serve further biological research. In addition, the untargeted LC-MS/MS dataset was general a huge dataset, which leads to time-consuming analysis of the whole process. Herein, a classified visualization method, called MCnebula, was used for addressing these difficulty.

The MCnebula package itself did not contain any part of molecular formula prediction, structure prediction and chemical prediction of compounds, so the accuracy of these parts was not involved. Currently, MCnebula performed downstream analysis by extracting the prediction data from SIRIUS project. The core of MCnebula was its chemical classes filtering algorithm, called abundance-based classes (ABC) selection algorithm. To explain the ABC selection algorithm in details, we need to begin with MS/MS spectral analysis and identification of compounds.

**Chemical structure and formula.** The analysis of MS/MS spectrum was a process of inference and prediction. For example, we speculated the composition of elements based on the molecular weight of MS1; combined with the possible fragmentation pattern of MS2 spectrum, we speculated the potential molecular formula of a compound; finally, we look for the exact compound from the compound structure database. Sometimes, this process is full of uncertainty, because there are too many factors that affect the reliability of MS/MS data and the correctness of inference. It can be assumed that there are complex candidates for the potential chemical molecular formula, chemical structure and chemical class behind MS/MS spectrum. Suppose we had these data of candidates now, MCnebula extracted these candidates and obtained the unique molecular formula and chemical structure for each MS/MS spectrum based on the highest score of chemical structure prediction; in this process, as most algorithms do, we make a choice based on the score, and only select the result of highest score.

**Establish reference upon top candidate** We predicted a potential compound represented by LC-MS/MS spectrum, and obtained the candidates of chemical molecular formula, structure and chemical class. These candidates include both positive and negative results: for chemical molecular formula and chemical structure, the positive prediction was unique; for chemical class, multiple positive predictions that belong to various classification were involved. We did not know the exact negative and positive. Normally, we ranked and filtered these according to the scores. There were numerous scores, for isotopes, for mass error, for structural similarity, for chemical classes… Which score selected to rank candidates depends on the purpose of research. Such as:

* To find out the chemical structure mostly be positive, ranking the candidates by structural score.
* To determine whether the potential compound may be of a certain chemical classes, ranking the candidates by the classified score.

Ether by functions in MCnebula of ‘filter\_formula()’, ‘filter\_structure()’ or ‘filter\_ppcp()’, the candidate with top score can be obtained. However, for the three module (formula, structure, classes), sometimes their top score candidates were not in line with each other. That was, their top score towards different chemical molecular formulas. To find out the corresponding data in other modules, ‘create\_reference()’ should be performed to establish the ‘specific candidate’ as reference for subsequent data integrating.

Above, we talked about chemical molecular formula, chemical structural formula and chemical classes. We obtained the unique chemical molecular formula and chemical structure formula for reference by scoring and ranking. But for chemical classes, we can’t adopt such a simple way to get things done.

**Chemical classification.** Chemical classification is a complex system. Here, we only discuss the structure based chemo taxonomy system [[39](#ref-2016)], because the MS/MS spectrum is more indicative of the structure of compounds than biological activity and other information.

According to the division of the overall structure and local structure of compounds, we can call the structural characteristics as the dominant structure and substructure [[39](#ref-2016)]. Correspondingly, in the chemical classification system, we can not only classify according to the dominant structure, but also classify according to the substructure. The chemical classification based on the dominant structure of compounds is easy to understand, because we generally define it in this way. For example, we will classify Taxifolin as ‘flavones’, not ‘phenols’, although its local structure has a substructure of ‘phenol’. We hope to classify a compound by its dominant structure rather than substructure, because such classify is more concise and contains more information. However, in the process of MS/MS spectral analysis, we sometimes can only make chemical classification based on the substructure of compounds, which may be due to: uncertainty in the process of structural analysis; it may be an unknown compound; MS/MS spectral fragment information is insufficient. In this case, it was necessary for us to classify the compounds with the aid of substructure information, otherwise we had no knowledge of the compounds for which we cannot obtain dominant structure information.

We should also be clear about the complexity of another aspect of chemo taxonomy, i.e., the hierarchy of classification. This was easy to understand. For example, ‘Flavones’ belongs to its superior, ‘Flavonoids’; its next higher level, ‘Phynylpropanoids and polyketides’; the further upward classification is ‘organic compounds’.

**ABC selection.** The above section discusses the inferential prediction of individual MS/MS spectrum. In the untargeted LC-MS/MS dataset, each feature has a corresponding MS/MS spectrum, and there are thousands of features in total. The ABC selection algorithm regards all features as a whole, examined the number and abundance of features of each chemical classification (classification at different levels, classification of substructure and dominant structure), and then selected representative classes (mainly screening the classes according to the number or abundance range of features) to serve the subsequent analysis (Fig. 6).

* Create Stardust Classes (Inner filter). The posterior probability of classification prediction (PPCP) data belonged to each ‘feature’. When performing the filtering, only simple threshold conditions or absolute conditions were set to filter the chemical classes; there was no crossover between the different attributes and no crossover between the ‘features’. Therefore, we considered this as ‘inner’ filtering.
* Cross filter Stardust Classes. The data of the chemical classes and their classified ‘features’, i.e. Stardust Classes data, were combined and then grouped upon the chemical classes. After grouping, each chemical class had a certain quantity of ‘features’. When filtering, statistics may be performed on ‘features’ data within a group; statistics may be performed on these data in conjunction with ‘features annotation’ data; and statistics may be performed to compare groups with each other. As its crossover of attributes for filtering, we consider this as ‘cross’ filtering. (See details in next subsection about Cross filter Stardust Classes.)

Whether it is all filtered by the algorithm provided by MCnebula function or custom filtered for some chemical classes, we now have a data called Nebula-Index. This data records a number of chemical classes and the ‘features’ attributed to them. The subsequent analysis process or visualization will be based on it. Each chemical class is considered as a ‘nebula’ and its classified ‘features’ are the components of these ‘nebulae’. In the visualization, these ‘nebulae’ will be visualized as networks. Formally, we call these ‘nebulae’ formed on the basis of Nebula-Index data as Child-Nebulae. In comparison, when we put all the ‘features’ together to form a large network, then this ‘nebula’ is called Parent-Nebula.

**Details of Cross filter Stardust Classes**. This method were integration of the following three module (Fig. 6):

*Cross filter by ‘quantity’* (abundance selection). Set ‘features’ quantity limitation for each group (each group, i.e. a chemical class with its classified ‘features’). The groups with too many ‘features’ or too few ‘features’ would be filtered out. This means the chemical class would be filtered out. These thresholds are about:

* Minimum quantity: the ‘features’ within the group.
* Maximum proportion: the ‘features’ quantity within the group versus all ‘features’ (unique) quantity of all groups.

The purpose of this step is to filter out chemical classes that have too large or too subtle a conceptual scope. For example, ‘Organic compounds’, which covers almost all compounds that can be detected in metabolomics data, is too large in scope to be of any help to our biological research. The setting of parameters is not absolute, and there is no optimal solution. Users can draw up thresholds according to the necessity of the study.

*Cross filter by ‘score’* (Goodness assessment). This step associate Stardust Classes data with ‘features’ annotation data. For each group, the Goodness assessment is performed for each target attribute (continuous attribute, generally be a scoring attribute of compound identification, such as ‘Tanimoto similarity’). If the group met all the expected Goodness, the chemical class would be retained; otherwise, the chemical class would be filtered out. The Goodness () related with the ‘features’ within the group:

* : the quantity of ‘features’ of which target attributes satisfied with the cut-off.
* : the quantity of all ‘features’.

The Goodness: .

The assessment of Goodness is related to the parameters of ‘tolerance’ and ‘cutoff’:

* Expected Goodness, i.e. value of ‘tolerance’.
* Actual Goodness, related to parameter ‘cutoff’. .

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

*Cross filter by ‘identical’* (identicality assessment). A similarity assessment of chemical classes. Set a hierarchical range for chemical classification and let groups within this range be compared for similarity to each other. For two groups, if the classified ‘features’ almost identical to each other, the chemical class represented by one of the groups would be discarded. The assessment of identical degree of two groups (A and B):

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

**Network graph presentation** As mentioned above, ‘features’ and their annotations were integrated as Nebulae according to the Nebula-Index. These Nebulae were network-type graph data. The ‘features’ annotation data contains top candidate of chemical formula and structure (obtained with the MCnebula function ’filter\_\*()‘). The MS/MS spectral similarity (dotproduct) of the ’features’ was calculated and used to form the edge data for the network graph.

**Visualization system** MCnebula used a number of existing R packages to integrate and reformat data [[83](#ref-2021z)–[90](#ref-2020u)]. In particular, the network graph data was equipped with ‘ggplot2’ package for visualization. The ggplot2 package is known for its elaborate and aesthetic mapping characteristics. We designed the ‘ggset’ data class to store pre-defined ggplot2 plotting functions and parameters for visualizing Nebulae. This allows users to completely customize the visualization to suit their needs or the needs of the publisher. Its flexibility depends on the user’s knowledge of the ‘ggplot2’ package. If users were not experienced in ‘ggplot2’, then just follow the preset to visualization.

**Statistic analysis** MCnebula integrated the functions (mainly for: ‘limma::makeContrasts’, ‘limma::lmFit’, ‘limma::lmFit’, ‘limma::eBayes’) of the ‘limma’ package for differential expression analysis (analysis of RNA-sequence and microarray) and packaged them for differential analysis of metabolomics data [[91](#ref-gentleman_limma_2005-1),[92](#ref-ritchie_limma_2015)]. The gene expression matrix and the feature quantification matrix of LC-MS are similar in that both have corresponding explanatory variables (sample information) and dependent variables (gene expression values or feature quantification values), except that one represents the level of gene expression and the other the level of metabolite. We normalized the peak area levels of the features and transformed (log2) them, used the metadata information of the samples to create the design matrix and the contrast matrix [[91](#ref-gentleman_limma_2005-1),[92](#ref-ritchie_limma_2015)]; thus even if the data itself is not gene related, it can be performed differential analysis using the tools of the ‘limma’ package. The ‘limma’ is a powerful package for differential analysis using linear models that can handle not only simple experimental designs in which the explanatory variables are factors (e.g., Control group versus Model group), but also complex experimental designs in which the explanatory variables are covariates (e.g., groups containing time series). However, our packaged method was only appropriate to experimental designs in which explanatory variables are factorial variables and the design matrix without intercept (code: ‘model.matrix(~ 0 + group)’) [[93](#ref-law_guide_2020)]. Because of its simple applicability, we called it as ‘binary comparison’. Our evaluation section does not cover its evaluation (as it is not the main part of this study), but we used it in two demo datasets and validated them to some extent: in the serum dataset, we compared our obtained top ‘features’ with that of Wozniak et al. [[65](#ref-2020s)]; in the herbal dataset, we traced the obtained top ‘features’ to the EIC plot.

**Data structure** MCnebula was written mainly in R S4 system of object-oriented programming. When analyzing data with MCnebula, all data (whether ‘features’ annotation data or visualization data) was stored in a one object (class ‘mcnebula’). This reduced the difficulty of using the MCnebula package, and made the data easy to manage and the analysis easy to repeat.

**Report system** MCnebula integrated a reporting system that allows the analysis process to be output as a PDF document or in other formats. The reporting system was based on the data class ‘report’, which could stores each step of the analysis as a section and could be flexibly modified according to the user’s needs. In addition, the reporting system can be used to generate reports even if the analysis is completely irrelevant to MCnebula package. The reporting system was associated with the ‘rmarkdown’ R package [[94](#ref-allaire_rmarkdown_2022)–[96](#ref-xie_r_2018)].

**Code Compatibility** MCnebula performs downstream analysis by extracting the data from the already computed SIRIUS project. The SIRIUS project is the main source of data for MCnebula 2. The SIRIUS software is still being updated and improved. In fact, from SIRIUS version 4 to version 5 (<https://github.com/boecker-lab/sirius>), the data structure and attributes name in the project directory have changed. In order that the functionality of MCnebula is not invalidated due to versioning issues, its application program interface (API) for the SIRIUS project has been designed to be flexible. MCnebula is able to perform data extraction for different SIRIUS versions.

## MCnebula evaluation

**Spectra dataset for evaluation.** The spectra collection (in positive ion mode, for more spectra data) of GNPS MS/MS library were used for evaluation (.msp file) (<http://prime.psc.riken.jp/compms/msdial/main.html#MSP>). As Fragmentation spectra in reference library generally possess high quality, and while used for evaluation of library match, it may caused overfitting. To address the issue, refer to ref. [[97](#ref-2021)], we added ‘noise’ into these MS/MS spectra. In brief, the ‘noise’ involves mass shift, intensity shift, and the inserted noise peak; of note, the magnitude factors for these shift were drawn randomly from function of normal distribution. Overall, we simulated two model of ‘noise’ (medium noise and high noise). The ‘noise’ simulation was achieved in custom R script. The algorithm and parameters were parallel to the reference [[97](#ref-2021)]. We assign these datasets as origin dataset, medium noise dataset and high noise dataset. The details of simulation of noise were as following:

* A global mass shift was simulated by drawing a random number from (Normal distribution) and then shifting every peak mass by . The standard deviation was chosen as (medium noise) or (high noise), so that the interval represents a 10-ppm shift for medium noise and a 15-ppm shift for high noise.
* Individual mass deviations was simulated by drawing, for each peak with mass individually, a random number from and shifting the peak by . The standard deviation was chosen so that the interval represents a 10-ppm shift for medium noise and a 20-ppm shift for high noise.
* Intensity variations were simulated in the spectrum. Each peak intensity was multiplied by an individual random number drawn from . Variance was chosen as for medium noise and for high noise. 0.03 times the maximum peak intensity of the spectrum was subtracted from each peak intensity. If a peak intensity fell below the threshold of one thousands of the maximum intensity in the spectrum, the peak was discarded.
* Additional ‘noise peaks’ were added to the spectrum. In processing of origin dataset, a pool of ‘noise peaks’ was gathered from the fragmentation spectra, using all peaks that did not have a molecular sub-formula decomposition of the known molecular formula of the precursor. For each spectrum, of these ‘noise peaks’ were added to the spectrum, where is the number of peaks in the spectrum and for medium noise and for high noise. Intensities of ‘noise peaks’ were adjusted for maximum peak intensities in the contributing and receiving spectrum. ‘Noise peaks’ were randomly drew from the pool of ‘noise peaks’ and added to the spectrum.

   For another issue, the spectra collection did not possess isotopes pattern. In real LC-MS processing (feature detection), isotope peaks were grouped and merged, which favorable for SIRIUS to detect some specific element [[62](#ref-bocker_sirius_2009)]. To simulate isotopes pattern, we used function of ‘get.isotopes.pattern’ within ‘rcdk’ R package to get isotope mass and its abundance [[98](#ref-2007j)]. Further, these mass were considered for the adduct type to increase or decrease exact mass. For the ‘intensity’ of these isotopes pattern, we simulated as relative intensity, i.e., the abundance of isotopes multiply by 100 as the value. These ‘isotope peaks’ were merged into MS1 list of its compounds. All the spectra collections were formatted to fit with input of MCnebula workflow or benchmark method (.mgf file and feature quantified table .csv).

**Evaluation method.** The three simulated data were all run with MCnebula workflow and benchmark method. While these data were put into SIRIUS 4 command-line interface (CLI) (version 4.9.12) for computation, the MS/MS spectra with empty fragmentation peak were auto-filtered. In addition, to reduce computation time, the compounds with over 800 m/z precursor were filtered out manually. These filtered out compounds were excluded from ultimate accuracy assessment. There were 8782 MS/MS spectra within the raw collection, and after filtered or excluded, totally 7524 compounds for ultimate evaluation.

   The assessment of classified accuracy was conducted with the help of ClassyFire [[39](#ref-2016)]. In detail, we traversed the raw .msp spectra file to collate metadata of these compounds, involving structure annotation. The International Chemical Identifier Key (InChIKey) of these compounds were available for ClassyFire database retrieve. However, since ClassyFire only support for those chemical identity of which structure have been classified in its server previously, we noticed all the InChIKeys were vetoed. To address that, we employed first hash block of these InChIKeys (InChIKey planar, represent molecular skeleton) to touch PubChem application programming interface (API) (<https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest>) [[99](#ref-2022ak)]. Accordingly, we got all the possibly InChIKeys of isomerism (stereo, isotopic substitution) [[100](#ref-2012e)]. The classification of small molecules are depending on its molecular skeleton hence these chemical identities possess the same InChIKey planar are identical in classification. We pushed the obtained InChIKey list to ClassyFire to obtain classification. In R script, once any InChIKey of isomerism meet the classified data in database, the acquisition status for this molecular skeleton was turn off. In the end, all these chemical annotation were collated, integrated and assigned as standard reference.

   The discrepancies between the MCnebula and benchmark methods in terms of algorithm and classified results disallow them to be evaluated at completely the same level. We evaluated both methods respectively. For MCnebula, before assess the accuracy, we interrogated the obtained classes generated from pre-analysis of the origin dataset. The classes over in half were classified based on sub-structural class, such as ‘Organic carbonic acids and derivatives’, ‘Hydroxy acids and derivatives’. These classes were small in structural size and were chemical function group within compounds. The principle of ClassyFire is selecting the most dominant structural class of compounds to substitute [[39](#ref-2016)]. But, in perspective of drug discovery, structure determines potency; many pharmacological action possibly depends on these sub-structure. To locate more universality among features, in algorithm, we reserved these classes in results. Sub-structural classify for benchmark method not available, hence we neglected these classes in evaluation. The rest classes, nevertheless, still possibly be sub-structural class while meet some compounds. We assigned three levels for evaluation, i.e., ‘True’, ‘Latent’, ‘False’. The ‘True’ indicated the classified classes were in line with that of ClassyFire. The ‘Latent’ indicated the classified classes were not in line with that of ClassyFire, but their parent classes of ‘class’ level were in line with that of ClassyFire. The ‘False’ indicated the classified classes were completely inconsistent with that of ClassyFire.

   To assess the identification of classes or structures, the results were merge with standard reference by InChIKey planar. For assessment of identification of chemical structure, once the identified chemical structure was in line with or chemical structure (match by InChIKey planar), we assigned it as ‘true’. Indeed, such evaluation neglect stereochemistry.

## Serum dataset

   We re-analyzed 245 LC-MS/MS data (.mzML) from MASSIVE (id no. [MSV000083593](https://massive.ucsd.edu/ProteoSAFe/QueryMSV?id=MSV000079949)) (blanks, controls and samples)[[65](#ref-2020s)]. MZmine2 (version 2.53) was performed for feature detection. The detection workflow mainly involves **1)** Automated Data Analysis Pipeline (ADAP) for peak detection and deconvolution [[9](#ref-2017f)], **2)** isotopes peak finder, **4)** parallel samples join alignment, **5)** gap filling algorithm. While exporting MS/MS spectra (.mgf) for SIRIUS 4 software computation, spectra were merged across samples into one fragmentation list with 30% Peak Count threshold filtering. The feature detection workflow was refer to [FBMN preprocessing](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-mzmine2/) and [SIRIUS computational prerequisites](https://boecker-lab.github.io/docs.sirius.github.io/prerequisites/). The output .mgf was run with SIRIUS 4 software (version 4.9.12) for computation with SIRIUS [[36](#ref-duhrkop_sirius_2019),[62](#ref-bocker_sirius_2009)], ZODIAC [[64](#ref-ludwig_database-independent_2020)], CSI:fingerID [[31](#ref-duhrkop_searching_2015)], CANOPUS [[55](#ref-duhrkop_systematic_2021)]. In particular, SIRIUS was customized set to detect Iodine element while predicting formula. MCnebula package were used for subsequent data analysis. All subsequent analysis have been organized into concise code and exported as reports (see section of Data and code availability).

   Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway enrichment analysis was performed with ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs), respectively. We used the identified InChIKey planar of structures to hit compounds in metabolic pathway. In detail, firstly, in order to avoid the identified structural deviations due to stereoisomerism, the InChIKey planar were used to obtain all possible InChIKeys via PubChem API. In this step, PubChem CID of those compounds were also obtained. The R package of MetaboAnalystR was used for converting PubChem CID to KEGG ID [[101](#ref-2020cx)]. Many compounds were not related to metabolic pathway so those were filtered out. The R package of FELLA was used for KEGG enrichment with ‘pagerank’ algorithm [[102](#ref-2018bj),[103](#ref-ilprints422)]. The above methods have been integrated as functions to interface with the MCnebula workflow. These functions are available in the ‘exMCnebula2’ package.

## Herbal dataset

**Material and processing.** *E. ulmoides* dried bark was obtained from company of ZheJiang ZuoLi Chinese Medical Pieces LTD. Raw-Eucommia and Pro-Eucommia were prepared as following: (1) Raw-Eucommia: The shreds or blocks of *E. ulmoides* dried bark were took, powdered and passed through 80-mesh sieves for further process. (2) Pro-Eucommia: The shreds or blocks of *E. ulmoides* dried bark were took, fried with saline water (the amount of salt is 2% of *E. ulmoides*, add 10 fold of water to dissolve), and smothered in airtight for 30 min. Then, the barks were dried in oven at 60 °C, followed by baking at 140 °C for 60 min. Finally, the baked barks were powdered and passed through 80-mesh sieves for further process. The processing method was based on previous studies of *E. ulmoides* [[104](#ref-2010b)].

**Sample preparation.** 2 g of Raw-Eucommia powder and Pro-Eucommia powder were weighed, respectively, added 50 ml of methanol/water (1:1, v/v) followed by ultrasonic (20 kHz for 40 min). After ultrasonic, the mixture was filtered to obtain filtrate and residue. The residue was added with 50 ml of methanol/water (1:1, v/v) and extracted with ultrasonic (40 kHz, 250 W for 20 min) again. The mixture was filtered. Then, the filtrate of the two extracts was combined, the solvent was evaporated. Methanol/water (1:1, v/v) was added to redissolve the extract and the volume was fixed to 5 ml. Finally, the supernatant was obtained by centrifugation (12,000 r.p.m. for 10 min) for further LC−MS analysis.

**LC–MS experiments.** LC−MS analysis was performed using a Dionex Ultimate 3000 UHPLC system (Dionex, Germany) coupled with a high-resolution Fourier-transform mass spectrometer (Orbitrap Elite, Thermo Fisher Scientific, Germany) using a Waters Acquity HSS T3 column (1.8 μm, 100 mm 2.1 mm, Waters Corporation, Milford, MA, USA). Solvent A, formic acid/water (0.1:99, v/v), and solvent B, formic acid/acetonitrile (0.1:99, v/v), were used as the mobile phase. The gradient profile for separation was as follows: 2% of solvent B at 0min, 5% of solvent B at 2 min, 15% of solvent B at 10 min, 25% of solvent B at 15 min, 50% of solvent B at 18 min, 100% of solvent B at 23 min, 2% of solvent B at 25 min, and 2% of solvent B at 30 min. The flow rate was 0.3 ml/min. The column temperature was set at 40°C. Mass spectrometric analysis was performed using an Orbitrap Elite instrument equipped with an ESI source (Thermo FisherScientific, Germany) that operated in the negtive ionization mode. The ESI source was operated at 50 °C with a capillary temperature of 275 °C, an ionization voltage of 3.5 kV, and a sheath gas flow rate of 35 L/min. The survey scans were obtained in the Orbitrap mass analyzer operating at a 120,000 (full width at half-maximum) resolving power. A mass range of 100−1500 m/z and a normalized collision energy of 30 eV were used for survey scans. The analysis method was set to analyze the top 10 most intense ions from the survey scan, and a dynamic exclusion was enabled for 15 s.

**MCnebula Workflow.** E.ulmoides dataset were preprocessed with MZmine2 for feature detection, followed by SIRIUS software computation. The subsequent analysis was similar to serum metabolic dataset and also been organized as a report (see section of Data and code availability).

## Data processing

   Raw data (.raw) were converted to m/z extensible markup language (mzML, i.e., .mzml format data) in centroid mode using MSConvert ProteoWizard [[5](#ref-2012d),[105](#ref-2011b)]. The .mzml files were processed with MZmine2 (v.2.53) and followed by SIRIUS 4 in Pop!-OS (Ubuntu) 22.04 LTS 64-bits workstation (Intel Core i9-10900X, 3.70GHz 20, 125.5 Gb of RAM) [[8](#ref-2010),[36](#ref-duhrkop_sirius_2019)]. MCnebula (MCnebula2) and other R packages were executed in Pop!\_OS (Ubuntu) 22.04 LTS 64-bits PC (Intel Core i7-1065G7, 1.3 GHz 8, 16 Gb of RAM).

# Data and code availability

   The serum dataset were available at MassIVE web service (id no. [MSV000083593](https://massive.ucsd.edu/ProteoSAFe/QueryMSV?id=MSV000079949)). The source code of MCnebula is available at <https://github.com/Cao-lab-zcmu/MCnebula>. The source code of exMCnebula2 is available at <https://github.com/Cao-lab-zcmu/exMCnebula2>. The code for all the analysis in this study can be found in the internal data directory (‘inst/extdata/’) of the ‘exMCnebula2’ package. In addition, .mgf files (msms spectra) and .csv files (feature quantification) and SIRIUS output files (use MCnebula function to filter and compress tens of GB of data to just a few tens of MB) and analysis report of serum and herbal dataset were compressed and stored in the exMCnebula2 package. By downloading and installing MCnebula package and exMCnebula2 package, all the analyses of this study can be reproduced by executing R codes (while using tools beyond R, such as MSconvert software, SIRIUS 4 software and MZmine2 are excluded).

# Reference

[1] Tsugawa H, Ikeda K, Takahashi M, Satoh A, Mori Y, Uchino H, et al. A lipidome atlas in MS-DIAL 4. Nature Biotechnology 2020;38:1159–63. <https://doi.org/10.1038/s41587-020-0531-2>.

[2] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Research 2018;46:W486–94. <https://doi.org/10.1093/nar/gky310>.

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

[4] Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology 2016;34:828–37. <https://doi.org/10.1038/nbt.3597>.

[5] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 2012;30:918–20. <https://doi.org/ghh626>.

[6] Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. Nature Methods 2016;13:741–8. <https://doi.org/f82r32>.

[7] Smith CA, Want EJ, 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;78:779–87. <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;11:395. <https://doi.org/bxbwnj>.

[9] Myers OD, Sumner SJ, 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;89:8696–703. <https://doi.org/gbrjtm>.

[10] Fu J, Zhang Y, Wang Y, Zhang H, Liu J, Tang J, et al. Optimization of metabolomic data processing using NOREVA. Nature Protocols 2022;17:129–51. <https://doi.org/10.1038/s41596-021-00636-9>.

[11] Mahieu NG, Patti GJ. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. Analytical Chemistry 2017;89:10397–406. <https://doi.org/10.1021/acs.analchem.7b02380>.

[12] Gloaguen Y, Kirwan JA, Beule D. Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics. Analytical Chemistry 2022;94:4930–7. <https://doi.org/10.1021/acs.analchem.1c02220>.

[13] Neumann U, Genze N, Heider D. EFS: An ensemble feature selection tool implemented as R-package and web-application. BioData Mining 2017;10:21. <https://doi.org/10.1186/s13040-017-0142-8>.

[14] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry: JMS 2010;45:703–14. <https://doi.org/10.1002/jms.1777>.

[15] Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, et al. 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>.

[16] Lei Z, Jing L, Qiu F, Zhang H, Huhman D, Zhou Z, et al. Construction of an Ultrahigh Pressure Liquid Chromatography-Tandem Mass Spectral Library of Plant Natural Products and Comparative Spectral Analyses. Analytical Chemistry 2015;87:7373–81. <https://doi.org/10.1021/acs.analchem.5b01559>.

[17] Lee S, Hwang S, Seo M, Shin KB, Kim KH, Park GW, et al. BMDMS-NP: A comprehensive ESI-MS/MS spectral library of natural compounds. Phytochemistry 2020;177:112427. <https://doi.org/10.1016/j.phytochem.2020.112427>.

[18] Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, Weldon K, et al. Mass spectrometry searches using MASST. Nature Biotechnology 2020;38:23–6. <https://doi.org/10.1038/s41587-019-0375-9>.

[19] Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 2010;11:148. <https://doi.org/10.1186/1471-2105-11-148>.

[20] Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 2015;11:98–110. <https://doi.org/10.1007/s11306-014-0676-4>.

[21] Ruttkies C, Schymanski EL, 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>.

[22] Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. 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;9:32. <https://doi.org/10.1186/s13321-017-0219-x>.

[23] Wang Y, Wang X, Zeng X. MIDAS-G: A computational platform for investigating fragmentation rules of tandem mass spectrometry in metabolomics. Metabolomics 2017;13:116. <https://doi.org/10.1007/s11306-017-1258-z>.

[24] da Silva RR, Wang M, Nothias L-F, van der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, et al. Propagating annotations of molecular networks using in silico fragmentation. PLoS Computational Biology 2018;14:e1006089. <https://doi.org/10.1371/journal.pcbi.1006089>.

[25] Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, Arndt D, et al. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. Metabolites 2019;9:72. <https://doi.org/10.3390/metabo9040072>.

[26] Ludwig M, Fleischauer M, Dührkop K, Hoffmann MA, Böcker S. De Novo Molecular Formula Annotation and Structure Elucidation Using SIRIUS 4. In: Li S, editor. Computational Methods and Data Analysis for Metabolomics, vol. 2104, New York, NY: Springer US; 2020, p. 185–207. <https://doi.org/10.1007/978-1-0716-0239-3_11>.

[27] Chao A, Al-Ghoul H, McEachran AD, Balabin I, Transue T, Cathey T, et al. In silico MS/MS spectra for identifying unknowns: A critical examination using CFM-ID algorithms and ENTACT mixture samples. Analytical and Bioanalytical Chemistry 2020;412:1303–15. <https://doi.org/10.1007/s00216-019-02351-7>.

[28] Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nature Methods 2013;10:755–8. <https://doi.org/10.1038/nmeth.2551>.

[29] Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of Cheminformatics 2015;7:44. <https://doi.org/10.1186/s13321-015-0087-1>.

[30] Heinonen M, Shen H, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics (Oxford, England) 2012;28:2333–41. <https://doi.org/10.1093/bioinformatics/bts437>.

[31] 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;112:12580–5. <https://doi.org/10.1073/pnas.1509788112>.

[32] Ludwig M, Dührkop K, Böcker S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. Bioinformatics (Oxford, England) 2018;34:i333–40. <https://doi.org/10.1093/bioinformatics/bty245>.

[33] Frainay C, Aros S, Chazalviel M, Garcia T, Vinson F, Weiss N, et al. MetaboRank: Network-based recommendation system to interpret and enrich metabolomics results. Bioinformatics (Oxford, England) 2019;35:274–83. <https://doi.org/10.1093/bioinformatics/bty577>.

[34] Nguyen DH, Nguyen CH, Mamitsuka H. ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. Bioinformatics 2019;35:i164–72. <https://doi.org/10.1093/bioinformatics/btz319>.

[35] Cao L, Guler M, Tagirdzhanov A, Lee Y-Y, Gurevich A, Mohimani H. MolDiscovery: Learning mass spectrometry fragmentation of small molecules. Nature Communications 2021;12:3718. <https://doi.org/10.1038/s41467-021-23986-0>.

[36] Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. Nature Methods 2019;16:299–302. <https://doi.org/10.1038/s41592-019-0344-8>.

[37] Aron AT, Gentry EC, McPhail KL, Nothias L-F, Nothias-Esposito M, Bouslimani A, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nature Protocols 2020;15:1954–91. <https://doi.org/10.1038/s41596-020-0317-5>.

[38] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 2000;25:25–9. <https://doi.org/10.1038/75556>.

[39] Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. Journal of Cheminformatics 2016;8:61. <https://doi.org/10.1186/s13321-016-0174-y>.

[40] Blaženović I, Kind T, Sa MR, Ji J, Vaniya A, Wancewicz B, et al. Structure Annotation of All Mass Spectra in Untargeted Metabolomics. Analytical Chemistry 2019;91:2155–62. <https://doi.org/10.1021/acs.analchem.8b04698>.

[41] Ernst M, Kang KB, Caraballo-Rodríguez AM, Nothias L-F, Wandy J, Chen C, et al. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites 2019;9. <https://doi.org/10.3390/metabo9070144>.

[42] Lee J, da Silva RR, Jang HS, Kim HW, Kwon YS, Kim J-H, et al. In silico annotation of discriminative markers of three Zanthoxylum species using molecular network derived annotation propagation. Food Chemistry 2019;295:368–76. <https://doi.org/10.1016/j.foodchem.2019.05.099>.

[43] Sha B, Schymanski EL, Ruttkies C, Cousins IT, Wang Z. Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs). Environmental Science Processes & Impacts 2019;21:1835–51. <https://doi.org/10.1039/c9em00321e>.

[44] Moreno-Ulloa A, Sicairos Diaz V, Tejeda-Mora JA, Macias Contreras MI, Castillo FD, Guerrero A, et al. Chemical Profiling Provides Insights into the Metabolic Machinery of Hydrocarbon-Degrading Deep-Sea Microbes. mSystems 2020;5. <https://doi.org/10.1128/mSystems.00824-20>.

[45] Tripathi A, Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, Nothias-Esposito M, et al. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. Nature Chemical Biology 2021;17:146–51. <https://doi.org/10.1038/s41589-020-00677-3>.

[46] Wang Y-K, Xiao X-R, Zhou Z-M, Xiao Y, Zhu W-F, Liu H-N, et al. A strategy combining solid-phase extraction, multiple mass defect filtering and molecular networking for rapid structural classification and annotation of natural products: Characterization of chemical diversity in Citrus aurantium as a case study. Analytical and Bioanalytical Chemistry 2021;413:2879–91. <https://doi.org/10.1007/s00216-021-03201-1>.

[47] Neto FC, Raftery D. Expanding Urinary Metabolite Annotation through Integrated Mass Spectral Similarity Networking. Analytical Chemistry 2021;93:12001–10. <https://doi.org/10.1021/acs.analchem.1c02041>.

[48] van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Research 2022;50:D1317–23. <https://doi.org/10.1093/nar/gkab941>.

[49] Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery 2016;15:473–84. <https://doi.org/10.1038/nrd.2016.32>.

[50] Liu X, Locasale JW. Metabolomics: A Primer. Trends in Biochemical Sciences 2017;42:274–84. <https://doi.org/10.1016/j.tibs.2017.01.004>.

[51] Guma M, Tiziani S, Firestein GS. Metabolomics in rheumatic diseases: Desperately seeking biomarkers. Nature Reviews Rheumatology 2016;12:269–81. <https://doi.org/10.1038/nrrheum.2016.1>.

[52] Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology 2016;17:451–9. <https://doi.org/10.1038/nrm.2016.25>.

[53] Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics 2019;20:492–503. <https://doi.org/10.1093/bib/bbx124>.

[54] Sharma A, Lysenko A, Boroevich KA, Vans E, Tsunoda T. DeepFeature: Feature selection in nonimage data using convolutional neural network. Briefings in Bioinformatics 2021;22:bbab297. <https://doi.org/10.1093/bib/bbab297>.

[55] Dührkop K, Nothias L-F, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nature Biotechnology 2021;39:462–71. <https://doi.org/10.1038/s41587-020-0740-8>.

[56] Platten M, Nollen EAA, Röhrig UF, Fallarino F, Opitz CA. Tryptophan metabolism as a common therapeutic target in cancer, neurodegeneration and beyond. Nature Reviews Drug Discovery 2019;18:379–401. <https://doi.org/gfvk74>.

[57] Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proceedings of the National Academy of Sciences 2012;109:E1743–52. <https://doi.org/10.1073/pnas.1203689109>.

[58] Quinn RA, Nothias L-F, Vining O, Meehan M, Esquenazi E, Dorrestein PC. Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. Trends in Pharmacological Sciences 2017;38:143–54. <https://doi.org/f9qzvt>.

[59] da Silva RR, Wang M, Nothias L-F, van der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, et al. Propagating annotations of molecular networks using in silico fragmentation. PLOS Computational Biology 2018;14:e1006089. <https://doi.org/gdc9cj>.

[60] Xie H-F, Kong Y-S, Li R-Z, Nothias L-F, Melnik AV, Zhang H, et al. Feature-Based Molecular Networking Analysis of the Metabolites Produced by *In Vitro* Solid-State Fermentation Reveals Pathways for the Bioconversion of Epigallocatechin Gallate. Journal of Agricultural and Food Chemistry 2020;68:7995–8007. <https://doi.org/10.1021/acs.jafc.0c02983>.

[61] Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nature Communications 2021;12:3832. <https://doi.org/10.1038/s41467-021-23953-9>.

[62] Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: Decomposing isotope patterns for metabolite identification. Bioinformatics 2009;25:218–24. <https://doi.org/10.1093/bioinformatics/btn603>.

[63] Dührkop K, Böcker S. Fragmentation Trees Reloaded. In: Przytycka TM, editor. Research in Computational Molecular Biology, vol. 9029, Cham: Springer International Publishing; 2015, p. 65–79. <https://doi.org/10.1007/978-3-319-16706-0_10>.

[64] Ludwig M, Nothias L-F, Dührkop K, Koester I, Fleischauer M, Hoffmann MA, et al. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. Nature Machine Intelligence 2020;2:629–41. <https://doi.org/10.1038/s42256-020-00234-6>.

[65] Wozniak JM, Mills RH, Olson J, Caldera JR, Sepich-Poore GD, Carrillo-Terrazas M, et al. Mortality Risk Profiling of Staphylococcus aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures. Cell 2020;182:1311–1327.e14. <https://doi.org/10.1016/j.cell.2020.07.040>.

[66] Chen Y-H, Bi J-H, Xie M, Zhang H, Shi Z-Q, Guo H, et al. 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>.

[67] 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>.

[68] Melone MAB, Valentino A, Margarucci S, Galderisi U, Giordano A, Peluso G. The carnitine system and cancer metabolic plasticity. Cell Death & Disease 2018;9:228. <https://doi.org/10.1038/s41419-018-0313-7>.

[69] Huang L, Lyu Q, Zheng W, Yang Q, Cao G. Traditional application and modern pharmacological research of Eucommia ulmoides Oliv. Chinese Medicine 2021;16:73. <https://doi.org/gnmwxx>.

[70] 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:E4501. <https://doi.org/10.3390/ijms21124501>.

[71] Park DW, Kwak DS, Park YY, Chang Y, Huh JW, Lim C-M, et al. 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;29:882.e5–11. <https://doi.org/10.1016/j.jcrc.2014.05.003>.

[72] Drobnik W, Liebisch G, Audebert F-X, Fröhlich D, Glück T, Vogel P, et al. Plasma ceramide and lysophosphatidylcholine inversely correlate with mortality in sepsis patients. Journal of Lipid Research 2003;44:754–61. <https://doi.org/10.1194/jlr.M200401-JLR200>.

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

[74] Wang B, Tontonoz P. Liver X receptors in lipid signalling and membrane homeostasis. Nature Reviews Endocrinology 2018;14:452–63. <https://doi.org/10.1038/s41574-018-0037-x>.

[75] Zhang Q, Yao D, Rao B, Jian L, Chen Y, Hu K, et al. The structural basis for the phospholipid remodeling by lysophosphatidylcholine acyltransferase 3. Nature Communications 2021;12:6869. <https://doi.org/10.1038/s41467-021-27244-1>.

[76] Huang L, Lyu Q, Zheng W, Yang Q, Cao G. Traditional application and modern pharmacological research of Eucommia ulmoides Oliv. Chinese Medicine 2021;16:73. <https://doi.org/10.1186/s13020-021-00482-7>.

[77] Huang Y-X, Liu E-W, Wang L, Huo Y, Wang Q, Olaleye O, et al. 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;12:469–76. <https://doi.org/10.1016/S1875-5364(14)60073-X>.

[78] Hu F, An J, Li W, Zhang Z, Chen W, Wang C, et al. 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;169:145–55. <https://doi.org/10.1016/j.jep.2015.04.007>.

[79] Li Y, Gong Z, Cao X, Wang Y, Wang A, Zheng L, et al. A UPLC-MS Method for Simultaneous Determination of Geniposidic Acid, Two Lignans and Phenolics in Rat Plasma and its Application to Pharmacokinetic Studies of Eucommia ulmoides Extract in Rats. European Journal of Drug Metabolism and Pharmacokinetics 2016;41:595–603. <https://doi.org/10.1007/s13318-015-0282-5>.

[80] Yan Y, Zhao H, Chen C, Zou L, Liu X, Chai C, et al. Comparison of Multiple Bioactive Constituents in Different Parts of Eucommia ulmoides Based on UFLC-QTRAP-MS/MS Combined with PCA. Molecules (Basel, Switzerland) 2018;23. <https://doi.org/10.3390/molecules23030643>.

[81] Jia J, Liu M, Wen Q, He M, Ouyang H, Chen L, et al. Screening of anti-complement active ingredients from Eucommia ulmoides Oliv. Branches and their metabolism in vivo based on UHPLC-Q-TOF/MS/MS. Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences 2019;1124:26–36. <https://doi.org/10.1016/j.jchromb.2019.05.029>.

[82] Huang Q, Zhang F, Liu S, Jiang Y, Ouyang D. Systematic investigation of the pharmacological mechanism for renal protection by the leaves of Eucommia ulmoides Oliver using UPLC-Q-TOF/MS combined with network pharmacology analysis. Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie 2021;140:111735. <https://doi.org/10.1016/j.biopha.2021.111735>.

[83] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

[84] Wickham H. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York; 2016.

[85] Horan K, Girke T. ChemmineOB: R interface to a subset of OpenBabel functionalities. 2021.

[86] Pedersen TL. Ggraph: An implementation of grammar of graphics for graphs and networks. 2021.

[87] Xiao N. Ggsci: Scientific journal and sci-fi themed color palettes for ’Ggplot2’. 2018.

[88] Auguie B. gridExtra: Miscellaneous functions for "Grid" graphics. 2017.

[89] Potter S, Murrell P. grImport2: Importing ’SVG’ graphics. 2019.

[90] Pedersen TL. Tidygraph: A tidy API for graph manipulation. 2020.

[91] Smyth GK. Limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, New York: Springer-Verlag; 2005, p. 397–420. <https://doi.org/10.1007/0-387-29362-0_23>.

[92] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015;43:e47. <https://doi.org/10.1093/nar/gkv007>.

[93] Law CW, Zeglinski K, Dong X, Alhamdoosh M, Smyth GK, Ritchie ME. A guide to creating design matrices for gene expression experiments. F1000Research 2020;9:1444. <https://doi.org/10.12688/f1000research.27893.1>.

[94] Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, et al. Rmarkdown: Dynamic documents for R. 2022.

[95] Xie Y, Dervieux C, Riederer E. R markdown cookbook. Boca Raton, Florida: Chapman and Hall/CRC; 2020.

[96] Xie Y, Allaire JJ, Grolemund G. R markdown: The definitive guide. Boca Raton, Florida: Chapman and Hall/CRC; 2018.

[97] Hoffmann MA, Nothias L-F, Ludwig M, Fleischauer M, Gentry EC, Witting M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. Nature Biotechnology 2021. <https://doi.org/10.1038/s41587-021-01045-9>.

[98] Guha R. Chemical informatics functionality in R. Journal of Statistical Software 2007;18.

[99] Temple Lang D. RCurl: General network (HTTP/FTP/...) Client interface for R. 2022.

[100] Pletnev I, Erin A, McNaught A, Blinov K, Tchekhovskoi D, Heller S. InChIKey collision resistance: An experimental testing. Journal of Cheminformatics 2012;4:39. <https://doi.org/10.1186/1758-2946-4-39>.

[101] Pang Z, Chong J, Li S, Xia J. MetaboAnalystR 3.0: Toward an optimized workflow for global metabolomics. Metabolites 2020. <https://doi.org/10.3390/metabo10050186>.

[102] Picart-Armada S, Fernandez-Albert F, Vinaixa M, Yanes O, Perera-Lluna A. FELLA: An R package to enrich metabolomics data. BMC Bioinformatics 2018;19:538. <https://doi.org/10.1186/s12859-018-2487-5>.

[103] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab / Stanford InfoLab; 1999.

[104] Cao Y. The history of *E. Ulmoides* and the progress of research in the past 20 years 2010.

[105] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzMLa Community Standard for Mass Spectrometry Data. Molecular & Cellular Proteomics 2011;10:R110.000133. <https://doi.org/dxkg99>.