# MCnebula: Classified analysis methods for untargeted LC-MS/MS dataset

# Abstract

   Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed MCnebula workflow, a novel strategy proposed with multidimensional view of visualization, termed multi-chemical Nebulae, involving in scope of abundant classes, classification, structures, sub-structural characteristics and fragmentation similarity. 1Many state-of-the-art technologies and popular methods were incorporated in MCnebula workflow to boost chemical discovery. 2Notably, MCnebula workflow can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library. 3Researchers were able to export the analysis process into a report, enabling integrated analysis from initial input to final output. 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. The core methods of MCnebula workflow was integrated in R package MCnebula2 and is public available for custom R pipeline analysis.

# 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 the 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). Nevertheless, we found that taxonomy or ontology for chemistry is not a one-off method for pharmacological or biological researches. Numerous key metabolites or drugs in classes are distributed in diverse hierarchies, such as ‘Bile acids, alcohols and derivatives’ (subclass)[49](#ref-2020cr), ‘Indoles and derivatives’ (class)[50](#ref-2022am), ‘Acyl carnitines’ (level 5)[51](#ref-2020s). These classes represent a family of compounds with either similar biological functions or activity, however, function- or activity-independent scattered on different branches of the diverse ranks of taxonomy. It confuse us and which are potential biomarkers? Indeed, untargeted mass spectra dataset is more like a black box. For unknown metabolites, locating the appropriate classes as manifestation from a complicated list of chemical taxonomy (> 4800 different categories) is challenging. Previous study applied such analogous method while in binary comparison but not yet systematized[52](#ref-duhrkop_systematic_2021).

   For above comprehensive consideration, we proposed a classified analysis workflow, named MCnebula, for untargeted LC-MS/MS dataset analysis. Currently, the core tool for MCnebula workflow is MCnebula2, a R package written primarily in S4 system of object-oriented programming. MCnebula2 leverages the state-of-the-art *in silico* tools, SIRIUS workflow (SIRIUS, ZODIAC, CSI:fingerID, CANOPUS)[31](#ref-duhrkop_searching_2015),[36](#ref-duhrkop_sirius_2019),[52](#ref-duhrkop_systematic_2021)–[55](#ref-ludwig_database-independent_2020), for compounds formulae prediction, structures retrieve and classes prediction.1 For the first time, MCnebula2 integrates an abundance-based classes (ABC) selection algorithm for compounds annotation. The ABC selection algorithm allows researchers to screen chemical classes for downstream data analysis, including statistical analysis, screening for key metabolites, pathway enrichment analysis, etc. With the help of MCnebula2, we can switch from untargeted analysis to targeted analysis which focusing on our interesting compounds or classes precisely. 2In addition, we provide a number of methods featuring MCnebula workflow for downstream custom analysis, involving pathway enrichment analysis, heat map analysis focused on chemical classes, querying compounds for synonyms, assembling identification tables, etc. These tools are integrated in the ‘exMCnebula2’ package to provide users with further options. With the MCnebula workflow, researchers could also export the analysis process into a report, enabling integrated analysis from initial input to final output. 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

   MCnebula2 was an R package for the analysis of LC-MS/MS datasets and was part of the MCnebula workflow. 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 key metabolites and their structural characteristics, pathway enrichment, querying compound synonyms, etc. In a nutshell, chemical classes focused analysis was the core of MCnebula workflow, but it was not only focused on chemical classes.

   Currently, the R package MCnebula2 has integrated methods for implementing general templates for MCnebula workflow, involving: application programming interface (APIs) for obtaining data from the cutting-edge SIRIUS software; filtering candidates of chemical formula, structure, classes; creating Nebulae visualizations; forming analytical reports, etc. In addition, methods for customizing advanced analyses, such as statistical analysis, feature selection, pathway enrichment analysis, querying compound synonyms, assembling identification tables, etc., have been initially integrated into the ‘exMCnebula2’ package. Custom advanced analysis was known to be a complex analytical process that requires a priori knowledge and a thorough understanding of the data. The ‘exMCnebula2’ package could not fully undertake automated analysis, but offered a range of functions that has the characteristics of MCnebula workflow, providing users with further options. Next, we would illustrate the MCnebula workflow with demo data.

## Data analysis

**Serum metabolic analysis.** To illustrate the application of MCnebula2 in metabolism, we re-analyzed the serum data from Wozniak et al.[51](#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 using MZmine2 on the LC-MS/MS dataset. After predicting the compounds by MS/MS spectra with SIRIUS software, subsequent analysis was performed using MCnebula2. Of these, 6501 ‘features’ were annotated with molecular formula and further, 3449 ‘features’ were annotated with chemical structure formulae. With the basic workflow of MCnebula2, Parent-Nebula and Child-Nebulae were obtained (Fig. S1, Fig. S2). By interrogating Child-Nebulae, we have 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’; the top 50 ‘features’ were set as ‘tracers’ to mark them in Child-Nebulae (Fig. 2). By combining the features selection algorithm, the chemical classes exhibited in Child-Nebulae were reduced. The log\_2\_(Fold Change) (log\_2\_(FC)) quantification for the HM versus HS groups was visualized in Child-Nebulae (Fig. 3). In Fig. 3, the overall level of ‘Bile acids, alcohols and derivatives’ (BAs) class and ‘Acyl carnitines’ (ACs) 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[51](#ref-2020s),[56](#ref-2021db),[57](#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. 4, Fig. S3). 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. S4b). 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. S4c). 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[58](#ref-2018bi). Their function implemented via bi-directional transport of acyl moieties Between cytosol and mitochondria (Fig. S4a).

   In research of Wozniak et al[51](#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 MCnebula2 workflow, sulfur element is detectable for SIRIUS in isotopes pattern with high mass accuracy[53](#ref-bocker_sirius_2009). However, for the MS/MS spectra of ‘HEPC’, there was no candidate formula that harbouring sulfur element. In addition, the high quality identified compounds were collated in Tab. S1

**Herbal medicine analysis.** *Eucommia ulmoides Oliv.* (*E. ulmoides*)[59](#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. MCnebula2 was successfully applied to analyze plant-derived chemical composition. Fig. S5 shown the focused classes with top ‘features’. It was noticed there were several characteristic classes ‘lignan’,

‘iridoid’ or ‘terpenoid’. In *E. ulmoides*, numerous literatures have reported the medicinal value of compounds in these classes[60](#ref-2021cq)–[66](#ref-2015q). Fig. S6 shown the MS/MS spectra and chemical structure of the identified top ‘features’. All the high quality identified compounds were collated in Tab. S2.

# Discussion

   MCnebula offers a unique analytical perspective, termed multi-chemical Nebulae to achieve unknown compound identification and classes focus. The visualization is equipped with a more precise, flexible and perceptive capturing ability of chemical classes which is different from classical molecular networking pattern. Meanwhile, it draws on the superiority of the classical pattern. Recently, molecular networking is a popular method for visualization and annotation of mass spectra. Depending on fragmentation spectra similarity, structural annotations are propagated via network-based method[67](#ref-2012a)–[71](#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[52](#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[67](#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),[52](#ref-duhrkop_systematic_2021), it is time for a revolution of the visualization strategy.

   Untargeted metabolomics emerged to profile cellular and organismal metabolism without prior knowledge dependence[72](#ref-2016aq),[73](#ref-2017at). Researchers with the help of statistical methodologies from thousands of features screen out biomarker, towards pharmaceutical, physiological or pathological mechanisms[74](#ref-2016ar),[75](#ref-2016ao). These statistical approaches involved classical statistic and artificial intelligence (e.g., random forests)[76](#ref-2019bv),[77](#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[52](#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 denpendent affection on aryl hydrocarbon receptor (AHR)[78](#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. Therewith, MCnebula is proposed to screen and trace biomarkers with higher confidence in classified level.

   MCnebula can be applied to discover biomarkers. We demonstrated the application of MCnebula by re-analyzing serum metabolic dataset. 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[57](#ref-2016at),[79](#ref-2020cv)–[81](#ref-2003n). In a recent review[79](#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[57](#ref-2016at). The investigation of spesis indicated LPCs concentrations in blood were established correlation with severe sepsis or septic shock[80](#ref-2014ao); In addition, LPCs was reported inversely correlate with mortality in sepsis patients[81](#ref-2003n). BAs’ disorder implied a liver dysfunction and imbalance of intestinal microphylactic homeostasis[82](#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[83](#ref-2018bd). LXRs modulate membrane phospholipid composition through activation of lysophosphatidylcholine acyltransferase 3 (LPCAT3), which directly related to LPCs[84](#ref-2021di). In addition, the above classes showed correlation with LXRs[83](#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 in Child-Nebulae with a scope of chemical classes. MCnebula is favorable to compound identification even for unknown compound. 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 MCnebula2. In the future, its methods and application will be extensively expanded.

# Methods

## MCnebula2 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 MCnebula2, was used for addressing these difficulty.

The MCnebula2 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, MCnebula2 performed downstream analysis by extracting the prediction data from SIRIUS project. The core of MCnebula2 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, MCnebula2 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 MCnebula2 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 MCnebula2 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’*. 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’*. 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’*. 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 MCnebula2 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** MCnebula2 used a number of existing R packages to integrate and reformat data[85](#ref-2021z)–[92](#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.

**Data structure** MCnebula2 was written mainly in R S4 system of object-oriented programming. When analyzing data with MCnebula2, all data (whether ‘features’ annotation data or visualization data) was stored in a one object (class ‘mcnebula’). This reduced the difficulty of using the MCnebula2 package, and made the data easy to manage and the analysis easy to repeat.

**Report system** MCnebula2 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 MCnebula2 package. The reporting system was associated with the ‘rmarkdown’ R package[93](#ref-allaire_rmarkdown_2022)–[95](#ref-xie_r_2018).

**Code Compatibility** MCnebula2 performs downstream analysis by extracting the data from the already computed SIRIUS project. The SIRIUS project is the main source of data for MCnebula2 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 MCnebula2 is not invalidated due to versioning issues, its application program interface (API) for the SIRIUS project has been designed to be flexible. MCnebula2 is able to perform data extraction for different SIRIUS versions.

## 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)[51](#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),[53](#ref-bocker_sirius_2009), ZODIAC[55](#ref-ludwig_database-independent_2020), CSI:fingerID[31](#ref-duhrkop_searching_2015), CANOPUS[52](#ref-duhrkop_systematic_2021). In particular, SIRIUS was customized set to detect Iodine element while predicting formula. MCnebula2 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[96](#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[97](#ref-2018bj),[98](#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*[99](#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),[100](#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). 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 MCnebula2 is available at <https://github.com/Cao-lab-zcmu/MCnebula2>. 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 MCnebula2 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 MCnebula2 package and exMCnebula2 package, all the analyses of this study can be reproduced by executing a few lines of 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](https://doi.org/10.1038/s41587-020-0531-2). Nature Biotechnology. 2020 Oct;38(10):1159–63.

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](https://doi.org/10.1093/nar/gky310). Nucleic Acids Research. 2018 Jul;46(W1):W486–94.

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

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](https://doi.org/10.1038/nbt.3597). Nature Biotechnology. 2016 Aug;34(8):828–37.

5. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. [A cross-platform toolkit for mass spectrometry and proteomics](https://doi.org/ghh626). Nature Biotechnology. 2012 Oct;30(10):918–20.

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](https://doi.org/f82r32). Nature Methods. 2016 Sep;13(9):741–8.

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](https://doi.org/b58xrd). Analytical Chemistry. 2006 Feb;78(3):779–87.

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](https://doi.org/bxbwnj). BMC Bioinformatics. 2010 Dec;11(1):395.

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](https://doi.org/gbrjtm). Analytical Chemistry. 2017 Sep;89(17):8696–703.

10. Fu J, Zhang Y, Wang Y, Zhang H, Liu J, Tang J, et al. [Optimization of metabolomic data processing using NOREVA](https://doi.org/10.1038/s41596-021-00636-9). Nature Protocols. 2022 Jan;17(1):129–51.

11. Mahieu NG, Patti GJ. [Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites](https://doi.org/10.1021/acs.analchem.7b02380). Analytical Chemistry. 2017 Oct;89(19):10397–406.

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

13. Neumann U, Genze N, Heider D. [EFS: An ensemble feature selection tool implemented as R-package and web-application](https://doi.org/10.1186/s13040-017-0142-8). BioData Mining. 2017 Dec;10(1):21.

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](https://doi.org/10.1002/jms.1777). Journal of mass spectrometry: JMS. 2010 Jul;45(7):703–14.

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](https://doi.org/10.1016/j.phytochem.2012.07.007). Phytochemistry. 2012 Oct;82:38–45.

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](https://doi.org/10.1021/acs.analchem.5b01559). Analytical Chemistry. 2015 Jul;87(14):7373–81.

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](https://doi.org/10.1016/j.phytochem.2020.112427). Phytochemistry. 2020 Sep;177:112427.

18. Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, Weldon K, et al. [Mass spectrometry searches using MASST](https://doi.org/10.1038/s41587-019-0375-9). Nature Biotechnology. 2020 Jan;38(1):23–6.

19. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. [In silico fragmentation for computer assisted identification of metabolite mass spectra](https://doi.org/10.1186/1471-2105-11-148). BMC Bioinformatics. 2010 Dec;11(1):148.

20. Allen F, Greiner R, Wishart D. [Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification](https://doi.org/10.1007/s11306-014-0676-4). Metabolomics. 2015 Feb;11(1):98–110.

21. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. [MetFrag relaunched: Incorporating strategies beyond in silico fragmentation](https://doi.org/10.1186/s13321-016-0115-9). Journal of Cheminformatics. 2016;8:3.

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](https://doi.org/10.1186/s13321-017-0219-x). Journal of Cheminformatics. 2017 May;9(1):32.

23. Wang Y, Wang X, Zeng X. [MIDAS-G: A computational platform for investigating fragmentation rules of tandem mass spectrometry in metabolomics](https://doi.org/10.1007/s11306-017-1258-z). Metabolomics. 2017 Oct;13(10):116.

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](https://doi.org/10.1371/journal.pcbi.1006089). PLoS computational biology. 2018 Apr;14(4):e1006089.

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](https://doi.org/10.3390/metabo9040072). Metabolites. 2019 Apr;9(4):72.

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

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](https://doi.org/10.1007/s00216-019-02351-7). Analytical and Bioanalytical Chemistry. 2020 Feb;412(6):1303–15.

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

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](https://doi.org/10.1186/s13321-015-0087-1). Journal of Cheminformatics. 2015;7:44.

30. Heinonen M, Shen H, Zamboni N, Rousu J. [Metabolite identification and molecular fingerprint prediction through machine learning](https://doi.org/10.1093/bioinformatics/bts437). Bioinformatics (Oxford, England). 2012 Sep;28(18):2333–41.

31. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. [Searching molecular structure databases with tandem mass spectra using CSI:FingerID](https://doi.org/10.1073/pnas.1509788112). Proceedings of the National Academy of Sciences. 2015 Oct;112(41):12580–5.

32. Ludwig M, Dührkop K, Böcker S. [Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints](https://doi.org/10.1093/bioinformatics/bty245). Bioinformatics (Oxford, England). 2018 Jul;34(13):i333–40.

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](https://doi.org/10.1093/bioinformatics/bty577). Bioinformatics (Oxford, England). 2019 Jan;35(2):274–83.

34. Nguyen DH, Nguyen CH, Mamitsuka H. [ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra](https://doi.org/10.1093/bioinformatics/btz319). Bioinformatics. 2019 Jul;35(14):i164–72.

35. Cao L, Guler M, Tagirdzhanov A, Lee Y-Y, Gurevich A, Mohimani H. [MolDiscovery: Learning mass spectrometry fragmentation of small molecules](https://doi.org/10.1038/s41467-021-23986-0). Nature Communications. 2021 Dec;12(1):3718.

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](https://doi.org/10.1038/s41592-019-0344-8). Nature Methods. 2019 Apr;16(4):299–302.

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](https://doi.org/10.1038/s41596-020-0317-5). Nature Protocols. 2020 Jun;15(6):1954–91.

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](https://doi.org/10.1038/75556). Nature Genetics. 2000 May;25(1):25–9.

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](https://doi.org/10.1186/s13321-016-0174-y). Journal of Cheminformatics. 2016 Dec;8(1):61.

40. Blaženović I, Kind T, Sa MR, Ji J, Vaniya A, Wancewicz B, et al. [Structure Annotation of All Mass Spectra in Untargeted Metabolomics.](https://doi.org/10.1021/acs.analchem.8b04698) Analytical chemistry. 2019 Feb;91(3):2155–62.

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.](https://doi.org/10.3390/metabo9070144) Metabolites. 2019 Jul;9(7).

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.](https://doi.org/10.1016/j.foodchem.2019.05.099) Food chemistry. 2019 Oct;295:368–76.

43. Sha B, Schymanski EL, Ruttkies C, Cousins IT, Wang Z. [Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs).](https://doi.org/10.1039/c9em00321e) Environmental science Processes & impacts. 2019 Nov;21(11):1835–51.

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.](https://doi.org/10.1128/mSystems.00824-20) mSystems. 2020 Nov;5(6).

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](https://doi.org/10.1038/s41589-020-00677-3). Nature Chemical Biology. 2021 Feb;17(2):146–51.

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.](https://doi.org/10.1007/s00216-021-03201-1) Analytical and bioanalytical chemistry. 2021 May;413(11):2879–91.

47. Neto FC, Raftery D. [Expanding Urinary Metabolite Annotation through Integrated Mass Spectral Similarity Networking.](https://doi.org/10.1021/acs.analchem.1c02041) Analytical chemistry. 2021 Sep;93(35):12001–10.

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.](https://doi.org/10.1093/nar/gkab941) Nucleic acids research. 2022 Jan;50(D1):D1317–23.

49. Quinn RA, Melnik AV, Vrbanac A, Fu T, Patras KA, Christy MP, et al. [Global chemical effects of the microbiome include new bile-acid conjugations](https://doi.org/10.1038/s41586-020-2047-9). Nature. 2020 Mar;579(7797):123–9.

50. Hezaveh K, Shinde RS, Klötgen A, Halaby MJ, Lamorte S, Ciudad MT, et al. [Tryptophan-derived microbial metabolites activate the aryl hydrocarbon receptor in tumor-associated macrophages to suppress anti-tumor immunity](https://doi.org/10.1016/j.immuni.2022.01.006). Immunity. 2022 Feb;55(2):324–340.e8.

51. 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](https://doi.org/10.1016/j.cell.2020.07.040). Cell. 2020 Sep;182(5):1311–1327.e14.

52. 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](https://doi.org/10.1038/s41587-020-0740-8). Nature Biotechnology. 2021 Apr;39(4):462–71.

53. Böcker S, Letzel MC, Lipták Z, Pervukhin A. [SIRIUS: Decomposing isotope patterns for metabolite identification](https://doi.org/10.1093/bioinformatics/btn603). Bioinformatics. 2009 Jan;25(2):218–24.

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

55. 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](https://doi.org/10.1038/s42256-020-00234-6). Nature Machine Intelligence. 2020 Oct;2(10):629–41.

56. 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](https://doi.org/10.1016/j.chroma.2021.462307). Journal of Chromatography A. 2021 Aug;1651:462307.

57. Krautbauer S, Eisinger K, Wiest R, Liebisch G, Buechler C. [Systemic saturated lysophosphatidylcholine is associated with hepatic function in patients with liver cirrhosis](https://doi.org/10.1016/j.prostaglandins.2016.06.001). Prostaglandins & Other Lipid Mediators. 2016 Jul;124:27–33.

58. Melone MAB, Valentino A, Margarucci S, Galderisi U, Giordano A, Peluso G. [The carnitine system and cancer metabolic plasticity](https://doi.org/10.1038/s41419-018-0313-7). Cell Death & Disease. 2018 Feb;9(2):228.

59. Huang L, Lyu Q, Zheng W, Yang Q, Cao G. [Traditional application and modern pharmacological research of Eucommia ulmoides Oliv.](https://doi.org/gnmwxx) Chinese Medicine. 2021 Dec;16(1):73.

60. Huang Q, Tan J-B, Zeng X-C, Wang Y-Q, Zou Z-X, Ouyang D-S. [Lignans and phenolic constituents from Eucommia ulmoides Oliver.](https://doi.org/10.1080/14786419.2019.1700250) Natural product research. 2021 Oct;35(20):3376–83.

61. 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.](https://doi.org/10.1016/j.biopha.2021.111735) Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie. 2021 Aug;140:111735.

62. Chen Y, Pan R, Zhang J, Liang T, Guo J, Sun T, et al. [Pinoresinol diglucoside (PDG) attenuates cardiac hypertrophy via AKT/mTOR/NF-B signaling in pressure overload-induced rats.](https://doi.org/10.1016/j.jep.2021.113920) Journal of ethnopharmacology. 2021 May;272:113920.

63. Han R, Yuan T, Yang Z, Zhang Q, Wang W-W, Lin L-B, et al. [Ulmoidol, an unusual nortriterpenoid from Eucommia ulmoides Oliv. Leaves prevents neuroinflammation by targeting the PU.1 transcriptional signaling pathway.](https://doi.org/10.1016/j.bioorg.2021.105345) Bioorganic chemistry. 2021 Nov;116:105345.

64. Zhu Y-L, Sun M-F, Jia X-B, Zhang P-H, Xu Y-D, Zhou Z-L, et al. [Aucubin alleviates glial cell activation and preserves dopaminergic neurons in 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-induced parkinsonian mice.](https://doi.org/10.1097/WNR.0000000000001075) Neuroreport. 2018 Sep;29(13):1075–83.

65. 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.](https://doi.org/10.1007/s13318-015-0282-5) European journal of drug metabolism and pharmacokinetics. 2016 Oct;41(5):595–603.

66. Jing X, Huang W-H, Tang Y-J, Wang Y-Q, Li H, Tian Y-Y, et al. [Eucommia ulmoides Oliv. (Du-Zhong) Lignans Inhibit Angiotensin II-Stimulated Proliferation by Affecting P21, P27, and Bax Expression in Rat Mesangial Cells.](https://doi.org/10.1155/2015/987973) Evidence-based complementary and alternative medicine : eCAM. 2015;2015:987973.

67. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. [Mass spectral molecular networking of living microbial colonies](https://doi.org/10.1073/pnas.1203689109). Proceedings of the National Academy of Sciences. 2012 Jun;109(26):E1743–52.

68. 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](https://doi.org/f9qzvt). Trends in Pharmacological Sciences. 2017 Feb;38(2):143–54.

69. 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](https://doi.org/gdc9cj). Schlessinger A, editor. PLOS Computational Biology. 2018 Apr;14(4):e1006089.

70. 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](https://doi.org/10.1021/acs.jafc.0c02983). Journal of Agricultural and Food Chemistry. 2020 Jul;68(30):7995–8007.

71. 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](https://doi.org/10.1038/s41467-021-23953-9). Nature Communications. 2021 Dec;12(1):3832.

72. Wishart DS. [Emerging applications of metabolomics in drug discovery and precision medicine](https://doi.org/10.1038/nrd.2016.32). Nature Reviews Drug Discovery. 2016 Jul;15(7):473–84.

73. Liu X, Locasale JW. [Metabolomics: A Primer](https://doi.org/10.1016/j.tibs.2017.01.004). Trends in Biochemical Sciences. 2017 Apr;42(4):274–84.

74. Guma M, Tiziani S, Firestein GS. [Metabolomics in rheumatic diseases: Desperately seeking biomarkers](https://doi.org/10.1038/nrrheum.2016.1). Nature Reviews Rheumatology. 2016 May;12(5):269–81.

75. Johnson CH, Ivanisevic J, Siuzdak G. [Metabolomics: Beyond biomarkers and towards mechanisms](https://doi.org/10.1038/nrm.2016.25). Nature Reviews Molecular Cell Biology. 2016 Jul;17(7):451–9.

76. Degenhardt F, Seifert S, Szymczak S. [Evaluation of variable selection methods for random forests and omics data sets](https://doi.org/10.1093/bib/bbx124). Briefings in Bioinformatics. 2019 Mar;20(2):492–503.

77. Sharma A, Lysenko A, Boroevich KA, Vans E, Tsunoda T. [DeepFeature: Feature selection in nonimage data using convolutional neural network](https://doi.org/10.1093/bib/bbab297). Briefings in Bioinformatics. 2021 Nov;22(6):bbab297.

78. Platten M, Nollen EAA, Röhrig UF, Fallarino F, Opitz CA. [Tryptophan metabolism as a common therapeutic target in cancer, neurodegeneration and beyond](https://doi.org/gfvk74). Nature Reviews Drug Discovery. 2019 May;18(5):379–401.

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

80. 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](https://doi.org/10.1016/j.jcrc.2014.05.003). Journal of Critical Care. 2014 Oct;29(5):882.e5–11.

81. 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](https://doi.org/10.1194/jlr.M200401-JLR200). Journal of Lipid Research. 2003 Apr;44(4):754–61.

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

83. Wang B, Tontonoz P. [Liver X receptors in lipid signalling and membrane homeostasis](https://doi.org/10.1038/s41574-018-0037-x). Nature Reviews Endocrinology. 2018 Aug;14(8):452–63.

84. 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](https://doi.org/10.1038/s41467-021-27244-1). Nature Communications. 2021 Nov;12(1):6869.

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

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

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

88. Pedersen TL. Ggraph: An implementation of grammar of graphics for graphs and networks. 2021.

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

90. Auguie B. gridExtra: Miscellaneous functions for "Grid" graphics. 2017.

91. Potter S, Murrell P. grImport2: Importing ’SVG’ graphics. 2019.

92. Pedersen TL. Tidygraph: A tidy API for graph manipulation. 2020.

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

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

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

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

97. Picart-Armada S, Fernandez-Albert F, Vinaixa M, Yanes O, Perera-Lluna A. [FELLA: An R package to enrich metabolomics data](https://doi.org/10.1186/s12859-018-2487-5). BMC Bioinformatics. 2018;19(1):538.

98. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab / Stanford InfoLab; 1999 Nov. Report No.: 1999-66.

99. Cao Y. The history of *E. Ulmoides* and the progress of research in the past 20 years. 2010;

100. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. [mzMLa Community Standard for Mass Spectrometry Data](https://doi.org/dxkg99). Molecular & Cellular Proteomics. 2011 Jan;10(1):R110.000133.