# **MCnebula: Classified visualization for spotlighting structural characteristics of underlying biomarkers and unknown compounds**

# **Abstract**

   Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed MCnebula, a novel visualization strategy proposed with multidimensional view, termed multi-chemical nebulae, involving in scope of abundant classes, classification, structures, sub-structural characteristics and fragmentation similarity. Many state-of-the-art technologies and popular methods were incorporated in MCnebula workflow to boost chemical discovery. Notably, MCnebula can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library. Reference spectral data was used for evaluation and MCnebula outperformed than popular benchmark method in classify with high noise tolerance. In virtue of MCnebula, 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 explore chemical transformation during drug processing. MCnebula was first integrated in R package and is now public available for custom R statistical pipline analysis.

# **Introduction**

   Analyzing untarget 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 [@2020p; @2018az; @2020co; @2016a]. These solution apply or suggest flexible mass spectrogram processing tools or analogous algorithms [@2012d; @2016e; @2006a; @2010]. To reduce false-positive and false-negtive results, more algorithms have been implemented for peak deconvolution, feature selection or statistical filtering [@2017f; @2022; @2017ao; @2022b; @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 acchieving reusability of reference fragmentation spectra, such as MassBank, MassBank of [North America (MoNA)](https://mona.fiehnlab.ucdavis.edu/), Global Natural Products Society molecular networking (GNPS) [@2010p; @2016a; @2012ac; @2015ak; @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) [@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 [@2010c; @2015c; @2016am; @2017aq; @2017ap; @2018ax; @2019bn; @2020; @2020cn]. Some databases such as MoNA collated *in silico* fragmentation spectra for public available [@2013w; @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 [@2012ab; @2015a; @2018ay; @2019bo; @2019bk; @2021cy].

*In silico* methods are developing quickly. Up to now, the cutting-edge technology, SIRIUS 4 [@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 untarget 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 [@2016a; @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) [@2000g], has been proposed, more systematically, organizing with taxonomy and ontology in recent years [@2016]. ClassyFire is increasingly used for compound annotation either in mass dataset analysis or not due to its computation available and systematicness [@2019bt; @2019br; @2019bs; @2019bq; @2020cq; @2021b; @2021da; @2021cz; @2022al]. The taxonomy and ontology for chemistry is benificial. For example, a hierarchical classification-based method, termed Qemistree, was proposed to address chemical relationships at a dataset-wide level [@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) [@2020cr], ‘Indoles and derivatives’ (class) [@2022am], ‘Acyl carnitines’ (level 5) [@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, untarget 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 [@2021a].

   For above comprehensive consideration, we proposed a classified visualization method, named MCnebula, for untarget LC-MS/MS dataset analysis. MCnebula leverages the state-of-the-art *in silico* tools, SIRIUS workflow (SIRIUS, ZODIAC, CSI:fingerID, CANOPUS) [@2009; @2019; @2015; @2015a; @2020a; @2021a], for compounds formulae prediction, structures retrieve and classes prediction. For the first time, MCnebula integrates an abundance-based classes selection algorithm for compounds annotation. MCnebula also incorporates the benefits of molecular networking, i.e., intuitive visualization and a great deal of information that can be conducted. In virtue of MCnebula, we can switch from untarget analysis to target analysis which focusing on our interesting compounds or classes precisely. MCnebula has massive potential functions, involving metabolites indentification, biomarker tracing in classes, drug discovery, chemical change exploration, etc. In this article, two datesets 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. we evaluated and validated MCnebula with several datasets, involving reference spectra library and published dataset.

# **Results**

## **Overview of MCnebula**

   MCnebula primarily performs an abundance-based class selection algorithm before visualization. MCnebula tend to filter out those classes with too large scope (e.g., possibly be ‘Lipids and lipid-like molecules’ but data dependent) or too sparse compounds (data depend). We termed these summarised classes as nebula-index. To begin with, like classical feature-based molecular networking (FBMN) pattern [@2020d], features make up the initial network, which we termed parent-nebula. Subsequently, according to nebula-index and the posterior probability of classes prediction (PPCP) of features [@2021a], nodes or edges from parent-nebula are divided into sub-networks. We termed these sub-networks as child-nebulae and their names, termed nebula-name, are in line with the classes name within nebulae-index (Fig. {@fig:eu.couple.nebula}{nolink=True}). The nebula-names contained the sub-structural or dominant-structural characteristics for features within child-nebulae. Collectively, all the network and sub-networks termed multi-chemical nebulae. In general, parent-nebula is too informative to show, so child-nebulae was used to dipict the abundant classes of metabolites in a grid panel, intuitively (Fig. {@fig:eu.couple.nebula}{nolink=True}). In a bird’s eye view of child-nebulae, we can obtain many characteristics of features, involving classes distribution, structure identified accuracy, as well as spectral similarity within classes. An end-to-end analysis using MCnebula is illustrated in Fig. {@fig:workflow}{nolink=True}.

   Child-nebulae equipped with feature selection algorithm to trace biomarker in classes (Fig. {@fig:trace.bio}{nolink=True}) [@2017i]. It assisted to focus on the targeted metabolites or compounds that we were interested in from complicated untarget analysis. Researchers can focus on specific class based on priori knowledge. Additionaly, calling nebula-name of interest in R with function, and a focused visualization with in deep annotation is available (Fig. {@fig:ac.zoom}{nolink=True}).

## **Evaluation of MCnebula**

**Classified accurary.** For evaluation, we assigned GNPS molecular networking as benchmark method, since its prestige, popularity and as well emerged as a visualized strategy [@2016a; @2020d]. Considering parallelity and repeatability, we employed FBMN and equipped it with MolnetEnhancer for assessment. In recent years, MolnetEnhancer has extensively adopted to boost molecular networking function with annotated classification [@2021cr; @2021cs; @2021cu; @2021ct; @2021cx; @2021cv; @2021cw; @2021cl]. Although we attempted to compare both methods in completely parallel way, there were several demarcation points for them: **1)** MCnebula conducted abundance-based classes selection and filtering, whereas benchmark method assigned all positive supperclass, class and subclass annotation; **2)** MCnebula gather features into classified index as child-nebulae, whereas in benchmark method, these possibly be scattered across network or as isolated nodes. **3)** MCnebula either performed dominant structural clustering or sub-structural clustering, whereas for benchmark method, features were annotated by dominant structural class.

   Both MCnebula and benchmark methods were run with a collection of GNPS spectral library (positive ion mode). ‘Noise’ was added into spectra to evaluate the stability of both algorithm. For MCnebula, as the figure shows (Fig. {@fig:mc\_noise\_tolerance\_bar}{nolink=True}), the classified accuracy is around 80% overall (‘true’ combined with ‘latent’, average 81.2%, 80.6%, 78.4% in original dataset, middle noise dataset and high noise dataset respectively). The annotated amount exhibited high stability even with high noise.

   For benchmark method, we collated all the annotated classes (superclass, class and subclass) and their harboured features. Some classes were too sparse with features hence we set a cut off to filter out those (features number 50). The stats of three datasets were gathered (Fig. [S{@s.fig](mailto:S%7B@s.fig):molnet\_noise\_tolerance\_bar}{nolink=True}). With original dataset, the benchmark method exhibited high classified accuracy and large annotated amount. For example, the figure (Fig. [S{@s.fig](mailto:S%7B@s.fig):molnet\_noise\_tolerance\_bar}{nolink=True}) showed nearly 500 ‘Flavonoides’ classified and the accuracy was 87.0%. Nevertheless, the benchmark method has low tolerance to noise. When it assessed with middle noise dataset and high noise dataset, the annotated amount is reduced to 30% and 60%, respectively. Taking the same class ‘Flavonoids’ as an example, the annotated amount was decreased from 500 to nearly 200. In particular with high noise, the rest amount was only around 100. The comparable classified amount of the in line classes are showed in lollipop diagram (Fig. {@fig:comp.accu\_identi}{nolink=True}a). The annotated amount of MCnebula exhibited better than benchmark method in middle noise, high noise and even in original dataset for some classes.

**Identified accurary.** MCnebula leverages structures retrieved by CSI:fingerID to annotate features within child-nebulae. Foremost, CSI:fingerID retrieves through public available structure databases that beyond the limitation of spectral library match. It facilitated discovering of novel compounds. In current, the high identified accuracy and outperforming of CSI:fingerID within SIRIUS workflow has been reported [@2019; @2021]. Herein, we evalueated the identified accuracy of features within child-nebulae. The original dataset was employed for evaluation. Overfitting issues did not exist since there was no spectral matching in CSI:fingerID algorithm. Without any filter or exclusion, there were 8782 fragmentation spectra in original dataset. Few compounds were in specific precursor adduct type, such as ‘[2M+H]+’, ‘[2M+Na]+’ and even ‘[M+H-99]+’ (totally 30 compounds). It is worth noting that several compounds consists of Iodine element (totally 7 compounds) which was sparse in current metabolite libraries. Considering untarget LC-MS/MS dataset as a complex ensembles of wild range of compounds, we did no filter out. After preprocessing and collated by MCnebula, a total of 8058 compounds were identified with formulae. Among them, a total of 6610 compounds were identified with chemical structures.

   For each feature in child-nebulae, we collated top score structure for assessment. In line with classified evaluation, those dominant-structral classes were picked. As figure shown (Fig. {@fig:comp.accu\_identi}{nolink=True}b), most of the identified accuracy were between 60% to 70%. However, some classes were quite low, such as ‘Long-chain fatty acids’ (LCFA) or ‘Lignans, neolignans and related compounds’ (LN-RC). Actually, researchers usually have no confidence for those structure with low matching score. Tanimoto similarity provides the matching degree of chemical fingerprints with structures [@2019]. In the assessment, we set 0.5 as cut-off value with Tanimoto similarity to filter those with low score.

## **Data analysis with MCnebula**

**Serum metabolomic analysis.** To illustrate the application of MCnebula in metabolome, we re-analyzed the serum data from Wozniak et al. [@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).

   In previous research, a total of four top metabolites (TopMs) were identified as 2-Hexadecanoylthio-1-Ethylphosphorylcholine (HEPC) (original ID: 103 or 2385), sphingosine-1-phosphate (S1P) (original ID: 114), T4 (original ID: 1110) and decanoyl-carnitine (original ID: 119). Except for HEPC, others were all identified in our re-analysis (Tab. [S{@s.tbl](mailto:S%7B@s.tbl):serum.bio}{nolink=True}). Intriguingly, ‘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 [@2009]. However, for feature of original ID: 103 or 2385, there was no candidate formula that harbouring sulfur element. In addition, the match of LPC15:0 was in high COSMIC confidence score (0.82) [@2021].

   In addition to four known metabolites, there were 11 new identified metabolites and some metabolites were with high COSMIC confidence (> 0.7) (Tab. [S{@s.tbl](mailto:S%7B@s.tbl):serum.bio}{nolink=True}). Child-nebulae were employed to trace those biomarker (TopMs) in abundant classes. We set the child-nebula min possess as 10 features (), max possess percentage of all features as 0.1 (). To reduce hitted classes, we post-modified the max possess as 300 features (Fig. {@fig:trace.bio}{nolink=True}). Overall, the visualized child-nebulae covered 13 of 16 TopMs whereas the rest were filtered out algorithmically. In-depth analysis of nebula-index, many prominent classes were noteworthy for exploration.

* ‘Acyl carnitines’ (ACs) were a sepsis related indicators [@2018bc], which as well agreed with Wozniak et al. (Fig. [S{@fig](mailto:S%7B@fig):ac.zoom}{nolink=True}). We verified 5 identified ACs metabolites presented in previous article. Furthermore, more metabolites of ACs were identified in the child-nebula. As the ring diagrams show (the statistic data were merged from Wozniak et al.), most of the ACs were increasing in HM group. Comparing with the ACs in TopMs or other previous identified ACs [@2020s], there were 4 ACs with a more remarkble increasing (ID of 8795, 3286, 3203, 14196). Of note, these 4 ACs are out of the main large cluster of ACs, because their functional group locating at carbon chain end were all Carboxyl groups. MCnebula captured the same sub-structure precisely. Incidently, the attractive speculation was that the carboxy-modified ACs were more indicative while referring to spesis and its liver dysfunction. The heatmap of levels of ACs showed a correlation of their level with SaB infection (Fig. {@fig:path}{nolink=True}a). A fundamental role of ACs in tuning the switch between the glucose and fatty acid metabolism was reviewed [@2018bi]. Their function implemented via bi-directional transport of acyl moieties between cytosol and mitochondria (Fig. {@fig:path}{nolink=True}b).
* ‘Lysophosphatidylcholines’ (LPCs) were a group of bioactive lipids, which were not refered by Wozniak et al. In our re-analysis, three of TopMs were enriched in this class, involving LPC15:0 (ID: 1819, original ID: 103 or 2385). Indeed, LPCs were associated with septic mortality [@2003n; @2014ao]. Here, we found a correlation between LPCs level with SaB infection and mortality which implied a pathogenesis of spesis. Focused on child-nebula of LPCs (Fig. [S{@s.fig](mailto:S%7B@s.fig):lpc\_ba}{nolink=True}a), comparing with control groups, the level of some LPCs in infection groups was remarkably lower. The heatmap of level of LPCs suggested a mortality risk of SaB infection, as the HM group was remarkably clustered (Fig. {@fig:path}{nolink=True}c). The significant LPCs (HS versus HM, ) were performed with Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway enrichment analysis. A kind of compounds termed ‘1-Acyl-sn-glycero-3-phosphocholine’ (KEGG ID: C04230) were hitted. The compounds of C04230 were characterized by its sub-structura (Fig. {@fig:path}{nolink=True}d). Almost all features classifed in child-nebula of LPCs were belonging to C04230 (Fig. [S{@s.fig](mailto:S%7B@s.fig):lpc\_ba}{nolink=True}a). As the figure {@fig:path}{nolink=True}d showed, C04230 affected multiple downstream pathway and most of which were correlated with lipids metabolism.
* ‘Bile acids, alcohols and derivatives’ (BAs) act as an important signaling molecule associated with liver function and intestinal microbial homeostasis. Diverse BAs structures were discovered in the chiled-nebula (Fig. [S{@s.fig](mailto:S%7B@s.fig):lpc\_ba}{nolink=True}b), and most were increasing in infection group. The heatmap of level of BAs implied a high correlation of SaB infection. The significant BAs (control group versus infection group, ) were performed with KEGG metabolic pathway enrichment analysis. The results foremost implied a correlation of SaB infection with bile secretion, cholesterol metabolism and primary bile acid biosynthesis.

   The above classes, together with steroids and fatty acids related classes, all suggest a central role of liver in SaB induced infection or mortality [@2017au]. In addition, the specific compound T4 in study of Wozniak et al. were clustered into mainly substructural classes, ‘Phenylpropanoic acids’ and ‘Phenoxy compounds’. Unfortunately, as its uncommon element constitution (Iodine), the classes failed to show the correlation with other features. All the high quality predictions (Tanimoto similarity > 0.5) of compounds in our re-analysis were collated according to ClassyFire classification (Tab. [S{@s.tbl](mailto:S%7B@s.tbl):serum.compound}{nolink=True}).

**Herbal medicine analysis.** *Eucommia ulmoides Oliv.* (*E. ulmoides*) [@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 successfully applied to analyze plant-derived chemical composition. Two kinds of proccessed *E. ulmoides* dataset were aligned in feature lists (Raw-Eucommia and Pro-Eucommia, before processed with saline water or not), and run with MCnebula workflow. Focused on abundant chemical classes, we set the nebula and . Figure {@fig:eu.couple.nebula}{nolink=True} shows these focused classes. 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 [@2021cq; @2021bt; @2021ch; @2021aw; @2018u; @2016aj; @2015q].

   We filtered the features in child-nebulae by || > 1 and ranked the classes by variation relative abundance (Fig. {@fig:eu.rank}{nolink=True}). The class of ‘Pyranones and derivatives’ (PDs) was in top one rank. Of note, many Flavonoides were discovered in its nebula (Fig. [S{@s.fig](mailto:S%7B@s.fig):pyran\_iri}{nolink=True}a), as its sub-structural Pyranone. Previous studies have reported pharmacological functions of Flavonoides in *E. ulmoides* [@2021cp; @2021cg; @2019ai; @2019x]. Within the annotated nebula, the feature of ID:980 possess a remarkable . However, the structure was matched with low Tanimoto similarity (0.38). Recur to overview child-nebulae, the level of some compounds belonging ‘Lignan glycosides’ (LG) and ‘Iridoid O-glycosides’ (IOGs) were changed with || > 1. Strikingly, we found most of the IOGs were increasing in level (peak area) after processing. Focused on IOGs, we visualized the annonated child-nebula (Fig. [S{@s.fig](mailto:S%7B@s.fig):pyran\_iri}{nolink=True}b). Structures of IOGs are similar in molecular skeleton, which contain a sub-structural nucleus formed by a five-membered ring combined with a six-membered ring. Among them, the feature with ID of 3918 is remarkably increasing in Pro-Eucommia. Retrieving PubChem database via InChIKey planar (first hash block of InChIKey), we found the briefest synonyms termed ‘8,10-Didehydroargylioside’, a compound without literature but structural record. We showed extracted ion chromatography (EIC) and fragmentation spectra of ‘8,10-Didehydroargylioside’, together with some related compounds of IOGs, LG or their parent classes (picked with || > 1, Tanimoto similarity > 0.5, and better peak shape) (Fig. [S{@s.fig](mailto:S%7B@s.fig):eu.iso}{nolink=True}). These compounds all showed remarkable alteration after processing. Among them, ‘8,10-Didehydroargylioside’ was a new generated compound after processing. Via MCnebula and other chemical tools (ClassyFire, PubChem database etc.), we identified a total of 582 compounds (Tab. [S{@s.tbl](mailto:S%7B@s.tbl):eu.compound}{nolink=True}. Some compounds were not reported before.

# **Discussion**

   MCnebula is a novel visualization strategy that leverages state-of-the-art *in silico* technology and orients to overall compounds within dataset. It 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 [@2012a; @2017g; @2018d; @2020e; @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 [@2021a]. In this context, visualization in a classified perspective is a better choice for untarget mass spectra dataset. Earlier in 2012, molecular networking was proposed with visualization for mass data analysis for the first time [@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 [@2021a; @2016], it is time for a revolution of the visualization strategy.

   Herein, we evaluated MCnebula with its accuracy of both classify and identification, it outperformed than feature-based molecular networking (FBMN) equipped with MolnetEnhancer in classify. Particularly, MCnebula exhibited higher tolerance of noise, since its workflow passed through building fragmentation tree [@2015]. MCnebula is more robust while dealing with metabolites of those without spectral library. MCnebula leverages dataset of posterior probability of classes prediction (PPCP) computated by CANOPUS to classify features and facilitate annotation of even unknown metabolites.

   Untarget metabolomics emerged to profile cellular and organismal metabolism without prior knowledge dependence [@2016aq; @2017at]. Researchers in virtue of statistical methodologies from thousands of features screen out biomarker, towards pharmaceutical, physiological or pathological mechanisms [@2016ar; @2016ao]. These statistical approaches involved classical statistic and artificial intelligence (e.g., random forests) [@2019bv; @2021de]. Both approaches were impossible to avoid specific biases, owing to the complexity of feature set or algorithmic stability [@2017i]. Further, evaluation in per feature level seemed incapable of profiling systematic effects in metabolites [@2021a]. 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) [@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 metabolomic dataset. The accuracy of MCnebula for metabolite identification and its contribution to the discovery of biomarkers in classified level was confirmed. We found more ‘Acyl carnitines’ (ACs) than previous study. Intriguingly, 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 [@2020cv; @2016at; @2014ao; @2003n]. In a recent review [@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 desease of increasing mortality risk [@2016at]. The investigation of spesis indicated LPCs concentrations in blood were established correlation with severe sepsis or septic shock [@2014ao]; In addition, LPCs was reported inversely correlate with mortality in sepsis patients [@2003n]. BAs’ disorder implied a liver dysfunction and imbalance of intestinal microphylactic homeostasis [@2021dg]. The chemical multiformity 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 [@2018bd]. LXRs modulate membrane phospholipid composition through activation of lysophosphatidylcholine acyltransferase 3 (LPCAT3), which directly related to LPCs [@2021di]. In addition, the above classes showed correlation with LXRs [@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 flexible exploration in child-nebulae with a scope of classification. The instance was enumerated with abundant classes. Child-nebulae could be set to trace sparse classes according to manual defination. MCnebula is favorable to compound identification even for unknown compound. For discovery of novel compound from complex herbal medicine, the visualization of child-nebulae is robust since it involved in scope of all abundant classes, classification, structures and even sub-structural characteristics. Although some specific database of plant-derived compounds have been constructed [@2015ak; @2012ac], there were lack of enougth fragmentation spectra for comprehensive library match. In virtue 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’s application. Indeed, MCnebula has a great potential in the field of chemistry, pharmacy and medcine. The latter, beyond this article, e.g., fields of application include natural products, foodomics, environmental research etc. In addition, as an integrated visualization method, MCnebula possibly be more popular with biologists or chemists. Currently, MCnebula was first proposed and implemented in the R language. In the future, its function and application will be extensively expanded.

# **Methods**

## **MCnebula algorithm**

**Data preprocessing.** MCnebula algorithm builds on the feature detection and SIRIUS compound identification workflow of untarget LC-MS/MS data. In brief, after feature detection of the untargeted LC-MS/MS data (via MZmine2 or other mass data processing tools [@2010a; @2006a; @2016e; @2020p]), feature table and MS/MS list (.mgf format file) were exported; SIRIUS 4 soft, used .mgf file as input, performed SIRIUS (predict molecular formula), ZODIAC (re-rank molecular formula), CSI:fingerID (retrieve structure library), CANOPUS (predict compound classification) in sequence. All results of SIRIUS workflow written down into SIRIUS soft project space (a directory). The overview preprocessing step were as follow:

* Convert raw mass spectrometry data (.RAW) to m/z extensible markup language (mzML) via MSconvert Proteowizard [@2011b; @2012d].
* Perform feature detection via MZmine2 (version 2.53).
* Perform SIRIUS soft compound identification workflow, involving SIRIUS, ZODIAC, CSI:fingerID, CANOPUS.

**MCnebula processing** MCnebula processing workflow was implemented into a R package. In R console or studio, through loading MCnebula package and employing several functions, MCnebula targeted at SIRIUS soft project space, accomplished data collating, integrating and visualization. The algorithms in detail were discriped:

* Collate molecular formulae. For each feature, as computation results, multiple molecular formula candidates may exist. MCnebula took comprehensive consideration of ZODIAC and CSI:fingerID scores to get top excellent formula. If CSI:fingerID retrieved any structure candidate, in default setting, MCnebula took formula of top score structure preferentially. While there was no structure candidates, MCnebula took the formula with top ZODIAC scores. Prioritization of picking top formula with either ZODIAC score or CSI:fingerID score can be reversed manually. Note that picking a correct molecular formula is the most essential step before structure identification as well as MCnebula workflow. Subsequently, The picked top formula for all features were gathered as MCnebula formula set (.MCn.formula\_set).
* Collate structures. According to .MCn.formula\_set, for each feature, only considering top formula, MCnebula took top CSI:fingerID score structure within the candidates. Then the picked structures were gathered as MCnebula structure set (.MCn.structure\_set).
* Collate PPCP. Analogously, according to .MCn.formula\_set, for each feature, only considering top formula, MCnebula took PPCP data of all classification. Theres data were gathered as MCnebula PPCP dataset (.MCn.ppcp\_dataset).
* Summarise nebula-class. Within .MCn.ppcp\_dataset, for each feature, posterior probability of thousands of classes prediction exist. A thereshold ( in default) was set to filter these data. Further, a paramater of classes hierarchy priority ( in default, which are equivalent to level 5, subclass, class, superclass of ClassyFire) was set to filter and sort these classes. The raw .MCn.ppcp\_dataset contained a large amount of sub-sturcture or dominant structure classes prediction data. This step aimed to get those classes favorable for identification. After filtering, the dataset was gathered as MCnebula nebula class (.MCn.nebula\_class).
* Summarise nebula-index. Although the raw .MCn.ppcp\_dataset was filtered via the previous step, all these classes were still too redundancy to perform an overview visualization of untarget LC-MS dataset in classification. In this step, several measures were adopted to implement an auto-filtering.
  1. Those subtile classes which represented a location of chemical function were removed. In deed, MS/MS spectra is not proficient at distinguishing position isomerism. Due as characteristics of International uion AppliedChemistry Rules (IUPAC rules), this measures was achieved simply via removing those class names involving Arabic numerals while matching in pattern. For example, as sub-catalogues of ‘Hydroxyflavonoids’, ‘6-hydroxyflavonoids’ and ‘7-hydroxyflavonoids’ were romoved.
  2. Filter via features of max possess and min possess setting of a class. Using previous filtered classes to traverse .MCn.ppcp\_dataset. For those any class, while the PPCP of a feature reached , the feature would be collated into the index of this class. After that, feature numbers in all classes index were stated respectively, and determined whether this class would be filtered out. The thereshold of min possess () was difined in absolute number whereas the thereshold of max possess () was difined in relative number (e.g., 20% of all feature number). The former paramater aim at filtering out the class possess sparse features, and the later aim at filtering out the class which covering too large scope of compounds. For example, while overview E. ulmoides, we appreciated abundant class such as ‘Lignan glycosides’ instead of ‘Nitrobenzenes’ etc.; for compounds belong to ‘Lipids and lipid-like molecules’ (superclass), the better choices is representing them in that sub-catalogues, such as ‘Steroids and steroid derivatives’, ‘Prenol lipids’ etc.
  3. Fiter out the classes that containing almost the same features. The thereshold of top hierarchy ( in default. i.e., level of class in ClassyFire) and the thereshold of identical factor ( in default.) were difined. All classes lower than were compared in binary. While each other possess more than of the same features, the class which possess less features would be filtered out. In deed, only few classes were filtered out in this algorithm. However, if a lower value of is set, some sub-catalogues may be removed (e.g., ‘Hydroxyflavonoids’ removed but ‘Flavonoids’ kept).
  4. Filter out the classes of features with low degree of structural identification. In most of cases, incorrect molecular formula lead to failed fingerprint which predicted from corresponding fragmentation tree. Both structure and PPCP were matched or computated depending on fingerprint. The false positive molecular formula would cause both wrong in structure identification and classes prediction. Reflecting in class, some classes harboured abundant features, however, almost no structures were matched or all matched structures with low similarity score. To filter out those classes, a similarity score-based algorithm was defined. First, the evaluation of similarity score type was set ( = ‘Tanimoto similarity’ in default). Then a cut-off of similarity score ( in default) was set. All classes harboured less than min reached ratio ( in default) of eligible features were filtered out. Ultimately, the rest classes and affiliated features were gathered as MCnebula nebula index (.MCn.nebula\_index).
* Generate parent-nebula. Analogous with molecular networking, parent-nebula consists of nodes (vehicles of feature information or annotation) and edges (annotation of fragmentation spectra similarity) data. To get edges and nodes data and merged as parent-nebula, MCnebula implement:
  1. Evaluation of MS/MS specturm similarity among features. MCnebula integrated ‘compareSpectra’ function of MSnbase R package to calculated cosine similarity (dotproduct) among MS/MS spectra [@2020v]. Unlike popularly spectral comparing method [@2012a; @2020d], instead of using raw MS/MS spectra, MCnebula collated all noise filtered MS/MS spectra for comparation. The noise filtered spectra were aquired from SIRIUS project space [@2015; @2017]. Different melecular formulae candidates of one feature, the corresponding MS/MS spectra may assign with diverse ‘valid’ or ‘noise’ peak pattern. To in line with above algorithm, all spectra picking based on molecular formulae within .MCn.formula\_set. In addition, to reduce time-consuming computation, spectra similarities were calculated only within the same nebula-index (); only classes hierarchy equal to or lower than thereshold (, in default, i.e., subclass of ClassyFire) were considered. Furthermore, if total feature number was more than 2000 (in default), ZODIAC scores ( in default) and Tanimoto similarity scores ( in default) were utilized to exclude features from computation. After that, a threshold of edges ( in default) was set to filter out low similarity. The results were formatted as edges data (.MCn.parent\_edges).
  2. Merging of multiple dataset. MCnebula merged .MCn.formula\_set with .MCn.structure\_set as nodes data (.MCn.parent\_nodes).
  3. Integation of .MCn.parent\_nodes and .MCn.parent\_edges as ‘graph’ project of igraph R package (.MCn.parent\_graph) [@2006b]. In addition, .grahml format file of parent-nebula was exported for surporting interactive exploration within Cytoscape [@2003].
* Generate child-nebulae. Depending on .MCn.nebula\_index, .MCn.parent\_nodes and .MCn.parent\_edges were accordingly divided and gathered as a variety of ‘graph’ project. At the meantime, for one nodes, a thereshold of max possessing ( in default) was defined to reduce edges for better visualization of child-nebulae. The edges imply lower similarity were preferentially be cut off. In the end, all child-nebulae ‘graph’ were saved into .MCn.child\_graph\_list and as well exported as .graphml format file, respectively. Note that a feature may exists in multiple child-nebulae, since compounds could be defined to diverse classes attributes to its sub-structures or dominant-structure.
* Visualize parent-nebula and child-nebula. In this step, ‘graph’ object obtained previously were visualized as individual or grid-based network via multiple R packages [@2016g; @2021y; @2018h; @2017j; @2019l; @2021ac; @2022f; @2022d; @2006b; @2020u]. In addition, Other R packages were used pass through all data processing if neccessary [@2021x; @2022c; @2021aa; @2007a; @2019k; @2022e].

   R presents a variety of flexible scientific stating and graphing tools. MCnebula provided a chanel to harbour massive of auto-annotated data of SIRIUS workflow into R analysis pipline. Users were encouraged to leverage R tools facilitating data integration and parsing. The visualization format of multi-chemical nebulae facilitate data perspective, which is favorable both for compound identification and for discovering biomarkers. In addition, GNPS FBMN could be incorporated into MCnebula for analysis. MCnebula could take the ‘edge’ file (a table file) generated by FBMN and performs classified grid as multi-chemical nebulae.

## **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 generaly possess high quality, and while used for evaluation of library match, it may caused overfitting. To address the issue, refer to 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 ref. [@2021]. We assign these dataset as original data, middle noise data and high noise data.

   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 [@2009]. To simulate isotopes pattern, we used function of ‘get.isotopes.pattern’ within rcdk R package to get isotope mass and its abundance [@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).

**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 stat. In this context, several compounds harboured Iodine element were excluded from stating either (totally 7), as it cost more time for SIRIUS to detect that (it is sparse in metabolites, SIRIUS do not detect that in default setting). There were 8782 MS/MS spectra within the raw collection, and after filtered or excluded, totally 7829 compounds for ultimate evaluation.

   The assessment of classification was in virtue of ClassyFire [@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, we noticed all the InChIKeys were vetoed. To address that, we emplyed 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>) [@2022ak]. Accordingly, we got all the possibly InChIKeys of isomerism (stereo, isotopic substitution) [@2012e]. The classification of small molecules are depending on its molecular skeletion hence these chimical identities possess the same InChIKey planar are identical in classification. We pushed the obtained InChIKey list to ClassyFire to catch 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. First, we evaluated both methods respectively. For MCnebula, before stat the accuracy, we interrogated the child-nebulae generated from the raw spectral collection (Fig. [S{@s.fig](mailto:S%7B@s.fig):collection.raw.child}{nolink=True}). The child-nebulae at least 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 [@2016]. However, 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 nebula-index. Sub-structural classify for benchmark method not available, hence we neglected these classes in evaluation. The rest classes, nevertheless, still possibley be sub-structural class while meet some compounds. We assigned three levels for evaluation, i.e., ‘true’, ‘latent’, ‘false’ (Fig. {@fig:mc\_noise\_tolerance\_bar}{nolink=True}).

   To assess the identification of classes or structures, the workflow results were merge with standard reference by InChIKey panar. Once the indentification results are in line with standard reference, we assigned it as ‘true’. For assessment of structure identification, indeed, such strategy 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)[@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 [@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 threshould 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 [@2019; @2009], ZODIAC [@2020a], CSI:fingerID [@2015a], CANOPUS [@2021a]. In particular, SIRIUS was customized set to detect Iodine element while predicting formula. The COSMIC confidence scores within SIRIUS 4 software output were used for assessment of identification [@2021c]. MCnebula package were used for collating data from SIRIUS 4 output file and visualizing child-nebulae or individual child-nebula.

   In research of Wozniak et al., an ensemble feature selection (EFS) approach and Mann-Whitney U (MWU) tests were used for feature selection (survival versus mortality) [@2017i]. We collated the feature annotation (m/z and retention time) of the top 25 EFS metabolites and the top MWU metabolites which were identified before, i.e. thyroxine (T4) and decanoyl-carnitine. We named these metabolites as TopMs. We aligned TopMs with our re-analyzed results by m/z (0.01 tolerance) and retention time (0.3 min tolerance) (Tab. [S{@s.tbl](mailto:S%7B@s.tbl):serum.bio}{nolink=True}). A total of 16 TopMs were aligned. Due to the algorithmic difference in feature detection, the re-analyzed feature list was not exactly in line with that of before. It should be noted that the new identity number (ID) was generated in this article, whereas original ID was from previous study [@2020s]. More preferable augments for MCnebula workflow was set (e.g., at least 10 ppm or 0.002 m/z tolerance for Automated Data Analysis Pipeline (ADAP) [@2017f]). The inconsistency were neglected, since the algorithmic evaluation of feature detection is out of the scope. For statistic data, we obtained [Metabolomics Data Resource](https://www.cell.com/cms/10.1016/j.cell.2020.07.040/attachment/f13178d1-d1ee-4179-9d33-227a02e604f1/mmc3.xlsx) of Wozniak et al. and aligned with our re-analyzed feature list (.csv file export from MZmine2). All statistic data (peak area) for serum dataset was used the previous [@2020s].

   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 metabolomic 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 [@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 [@2018bj; @ilprints422].

## ***E. ulmoides* 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, powedered 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 powedered and passed through 80-mesh sieves for further process. The processing method was based on previous studies of *E. ulmoides* [@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 at23 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 sheathgas 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. The algorithmic workflow was similar to serum metabolomic dataset. Here, some peaks in bad shape were filtered out manually. While export MS/MS spectra (.mgf) for SIRIUS computation, spectra were merged across samples into one fragmentation list without filtering. Similar to processing in serum dataset, MCnebula was used for collating data from SIRIUS 4 output file and visualize child-nebulae or individual child-nebula. Statistical analysis of nebula-index (classes) with peak area data were conducted. Those classes with features of which || > 1 (: peak area of Pro-Eucommia / Raw-Eucommia) were extracted as neo-nebula-index. The variation relative abundance (for each classes, number of features || > 1 divided by sum) were stated and sorted as rank. Based on neo-nebula-index, we re-visualized the overview child-nebulae and mask features with value of (Fig. {@fig:eu.rank}{nolink=True}).

## **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 [@2011b;@2012d]. For *E. ulmoides* dataset, the .mzml files were processed with MZmine2 (v.2.53) on Windows 10 1909 64-bits PC (intel Core i5-8300H, 2.30 GHz, 16Gb of RAM) [@2011b]; SIRIUS 4 and MCnebula were executed in Pop!\_OS (Ubuntu) 20.10 LTS 64-bits PC (intel Core i7-1065G7, 1.3 GHz 8, 16 Gb of RAM) [@2019b]. For the evaluation dataset (noise simulation dataset and serum dataset), all MCnebula workflow were implemented on Pop!-OS (Ubuntu) 20.04 LTS 64-bits workstation (intel Core i9-10900X, 3.70GHz 20, 125.5 Gb of RAM). R packages and custom R script were extensively used for data processing and scientific mapping.

# **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 submission job in GNPS of evaluation dataset are available: 1) original dataset: FBMN: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=05f492249df5413ba72a1def76ca973d>. MolnetEnhancer: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9d9c7f83fa2046c2bf615a3dbe35ca62>; 2) middle noise dataset: FBMN: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c65abe76cd9846c99f1ae47ddbd34927>; MolnetEnhancer: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7cc8b5a2476f4d4e90256ec0a0f94ca7>; 3) high noise dataset: FBMN: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=62b25cf2dcf041d3a8b5593fdbf5ac5e>; MolnetEnhancer: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6d08a335e814c5eac7c97598b26fb80>.

   The source code of MCnebula integrated in R is available at <https://github.com/Cao-lab-zcmu/MCnebula>. Other R scripts used for analysis or graphic mapping in this manuscript are available upon request.