# Figure legend

## Figure 1-7

**Fig. 1 | The overview of Mcnebula Visualization of serum metabolomics dataset.** The **Parent-nebula** was visualized in network layout algorithm of [layout\_with\_kk](https://cran.microsoft.com/web/packages/igraph/igraph.pdf" \o "https://cran.microsoft.com/web/packages/igraph/igraph.pdf). The nodes mapped all features collated by MCnebula. The color of nodes mapped the superclasses of highest posterior probability of classes prediction (PPCP). The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. For better visualization, the isolated nodes were removed from parent-nebula. The **Child-nebulae** were visualized in network layout algorithm of [layout\_with\_kk](https://cran.microsoft.com/web/packages/igraph/igraph.pdf). All the classified labels mapped nebula-index and the nebula-name of corresponding sub-network. The nodes，features and edges were seperated from parent-nebula according to classify of MCnebula. The color of nodes were different from that of parent-nebula.The color mapped the dominant or sub-structural classses of highest PPCP in priority (level 5 subclass class superclass). The edges were cut off to leave less than 5 with top spectral similarity.

**Fig. 2 | Workflow of MCnebula：End-to-end analysis from samples to multi-chemcial nebulae.** First, the data (.raw) of prepared sample were obtained via LC-MS instrument. Second，the .raw was then converted to get the data of m/z extensible markup language (mzML) , followed by feature detection of LC-MS processing. Feature table and MS/MS list (MGF file) were obtained. Third，Run SIRIUS software identification workflow, involving SIRIUS, ZODIAC, CSI:fingerID, CANOPUS. Run MCnebula in R. MCnebula conducted data collating and integration. Ultimately, the multi-chemical nebulae as well as other annotation file were achieved.

**Fig. 3 | Tracing top rank metabolites in child-nebulae to discover biomarkers of serum metabolomic dataset.** According to classify of TopMs in MCnebula, the nebula-index was rebuilt (the irrelevant classes were filtered out) and lead to neo-child-nebulae. The neo-child-nebulae was visualized in network layout algorithm of [layout\_with\_fr](https://cran.microsoft.com/web/packages/igraph/igraph.pdf). The TopMs were makred in neo-child-nebulae. **Other annotation:** The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.

**Fig. 4 | In-depth visualization of child-nebula of ‘Acyl carnitines’.** The nodes ofTopMs were marked with color. The nodes of features were annotated with structures, ring diagram and bar plot of posterior probability of classes prediction (PPCP). The top score **Structures** of features were mapped into nodes. The atomic coordinates of molecular structures were calculated by [Molconvert](https://docs.chemaxon.com/display/docs/molconvert.md), since its less structural overlap (the default setting for MCnebula, ChemmineOB was performed for calculation). The **Ring diagram** mapped relative summed peak area of per feature detected within each metadata group (NN: non-hospital, non-infected; HN: hospital, non-infected; HS: hospital, survival; HM: hospital, mortality). The statistic data of ring diagram was obtained from study of Wozniak et al. and aligned with our re-analyzed feature list (0.01 m/z tolerance and 0.3 min retention time tolerance). The nodes without ring diagram indicated the features were detected in re-analysis but not found in previous study. The **Bar plot** mapped PPCP of structural (sub-structural or dominant structural) classes for the feature. These structural classes were in line with the classes in nebula-index. **Other annotation:** The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.

**Fig. 5 | Evaluation of classified accuracy and noise tolerance of MCnebula algorithm.** For the **Intermediate horizontal bar plot**, three levels of assessment were assigned for evaluation of accuracy. The ‘true’ indicated the classified classes were in line with that of ClassyFire. The ‘latent’ indicated the classified classes were not in line with that of ClassyFire, but thier parent classes of ‘class’ level (illustrated by the legend of the **Left tile diagram**) were in line with that of ClassyFire. The ‘false’ indicated the classified classes were completely inconsistent with that of ClassyFire. Noise was added into original dataset to evaluate the stability of MCnebula algorithm. Both for ‘true’ and ‘false’ assessment, the arrow indicated the middle noise or high noise lead to accuracy shift（increasing or decreasing.）The accuracy evaluation were only performed with the classified feature number . If the noise lead to classified number , the class was excluded from assessment of noise tolerance. The **Right horizontal bar plot** indicated the classified amount of features.    **Fig. 6 | Evaluation of classified amount，identification accuracy and noise tolerance of benchmark method and MCnebula.** The **Left faceted lollipop diagram** illustrated a comparasion of classified amount and noise tolerance of MCnebula and benchmark method. When noise was added into original dataset, some classified feature amount was occurred . A cut-off (amount 50) was set to exclude these classes from assessment. The **Right lollipop diagram** illustrated the identified accuracy of MCnebula. A cut-off (Tanimoto similarity 0.5) was set to get structures of high matching score for evaluation.

**Fig. 7 | Marking features with fold change in child-nebulae to explore chemical transformation during processing of *E. ulmoides*.** The **Left horizontal bar plot** illustrated the rank of variation relative abundance (VRA) of classified classes in *E. ulmoides* dataset. The VRA was calculated as: feature amount of divided by feature sum in classified nebula. Those classes of VRA in nebula-index were extracted to generated neo-nebula-index. The **Right neo-child-nebulae** were visualized according to neo-nebula-index and illustrated the features of in classified nebula. **Other annotation:** The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.

## Supplementary figure

**Fig. S1 | Evaluation of classified accuracy and noise tolerance of benchmark algorithm.** TFor the **Intermediate horizontal bar plot**, three levels of assessment were assigned for evaluation of accuracy. The ‘true’ indicated the classified classes were in line with that of ClassyFire. The ‘latent’ indicated the classified classes were not in line with that of ClassyFire, but thier parent classes of ‘class’ level (illustrated by the legend of the **Left tile diagram**) were in line with that of ClassyFire. The ‘false’ indicated the classified classes were completely inconsistent with that of ClassyFire. Noise was added into original dataset to evaluate the stability of MCnebula algorithm. Both for ‘true’ and ‘false’ assessment, the arrow indicated the middle noise or high noise lead to accuracy shift of increasing or decreasing. The accuracy evaluation were only performed with the classified feature number . If the noise lead to classified number , the class was excluded from assessment of noise tolerance. The **Right horizontal bar plot** indicated the classified amount of features. When noise was added, the classified amount was decreased.

**Fig. S3 | In-depth visualization of child-nebula of ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs) facilitated compound identification and biomarker discovery in serum metabolomic dataset.** The nodes of top metabolites (TopMs) were marked with color. The **Ring diagram** mapped relative summed peak area of per feature detected within each metadata group (NN: non-hospital, non-infected; HN: hospital, non-infected; HS: hospital, survival; HM: hospital, mortality). The statistic data of ring diagram was obtained from study of Wozniak et al. and aligned with our re-analyzed feature list (0.01 m/z tolerance and 0.3 min retention time tolerance). The nodes without ring diagram indicated the features were detected in re-analysis but not found in previous study. **Other annotation:** The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.

**Fig. S5 | In-depth visualization of child-nebula of ‘Pyranones and derivatives’ (PDs) and ‘Iridoid O-glycosides’ (IOGs) facilitated discovery of chemical transformation in *E. ulmoides* dataset.** The child-nebulae of PDs and IOGs were visualized according to the neo-child-nebulae (Fig. S7). The **Ring diagram** mapped relative summed peak area of per feature detected within each metadata group The top score **Structures** of features were mapped into nodes. In particular, herein the atomic coordinates of molecular structures were calculated by [Molconvert](https://docs.chemaxon.com/display/docs/molconvert.md), since its less structural overlap (the default setting for MCnebula, ChemmineOB was performed for calculation). **Other annotation:** The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.

**Fig. S6 | Mass spectrometry inspection for remarkable features of lignans and iridoids in *E. ulmoides* dataset.** The features were picked with , Tanimoto similarity , and fine peak shape. (**a**) The extracted ion chromatogram (EIC) plot illustrated the peak shape detected via Automated Data Analysis Pipeline (ADAP) algorithm. (**b**) The mirrored MS/MS spectra plots illustrated the raw MS/MS spectra (back bar) and the noise filtered MS/MS spectra (red bar) via building fragmentation tree in SIRIUS workflow. The dot above the bar implied a corresponding relation. The top score structures were mapped into mirrored MS/MS spectra.    **Fig. S8 | Interrogation of classes distribution of raw spectral library collection in child-nebulae.** The Child-nebulae were visualized in network layout algorithm of [layout\_with\_fr](https://cran.microsoft.com/web/packages/igraph/igraph.pdf). All the classified labels mapped nebula-index and the nebula-name of corresponding sub-network. The size of nodes mapped the Tanimoto similarity of structure match. The edges mapped the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and implied the identification quality (Tanimoto similarity ) of features. The edges were cut off to leave less than 5 with top spectral similarity.