# Figure legend

## Figure 1-8

**Fig. 1 | The overview of MCnebula visualization of serum metabolomic dataset.** In **Parent-nebula**, the nodes map all features collated by MCnebula. The color of nodes maps the superclasses of highest posterior probability of classes prediction (PPCP). The size of nodes maps the Tanimoto similarity of structure match. The edges map the spectural similarity of noise filtered between features (cosine ; ZODIAC score ) and imply the identification quality (Tanimoto similarity ) of features. In **Child-nebulae**, all the classified labels map nebula-index and the nebula-name of corresponding sub-network. The color of nodes maps the dominant or sub-structural classses of highest PPCP in priority (level 5 > subclass > class > superclass).

**Fig. 2 | Workflow of MCnebula: End-to-end analysis from samples to multi-chemcial nebulae.** First, the data (.raw) of prepared sample are obtained via LC-MS instrument. Second, the .raw is 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) are obtained. Third, run SIRIUS software identification workflow, involving SIRIUS, ZODIAC, CSI:fingerID, CANOPUS. Run MCnebula in R. MCnebula conduct data collating and integration. Ultimately, the multi-chemical nebulae as well as other annotation file are achieved.

**Fig. 3 | Tracing top rank metabolites in child-nebulae to discover biomarker of serum metabolomic dataset.** According to classify of TopMs in MCnebula, the nebula-index is rebuilt (the irrelevant classes were filtered out) and lead to neo-child-nebulae. The TopMs are maked in neo-child-nebulae. Other mapping annotation refer to figure 1.

**Fig. 4 | In-depth visualization of child-nebula of ‘Acyl carnitines’** The nodes of TopMs are marked with color. The nodes of features are annotated with structures, ring diagram and bar plot of posterior probability of classes prediction (PPCP). The top score **Structures** of features are 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 is performed for calculation). The **Ring diagram** map 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** map PPCP of structural (sub-structural or dominant structural) classes for the feature. These structural classes are in line with the classes in nebula-index. Other mapping annotation refer to figure 1.

**Fig. 5 | Evaluation of classified accuracy and noise tolerance of MCnebula algorithm.** For the **Intermediate horizontal bar plot**, three levels of assessment are assigned for evaluation of accuracy. The ‘true’ indicates the classified classes are in line with that of ClassyFire. The ‘latent’ indicates the classified classes are not in line with that of ClassyFire, but thier parent classes of ‘class’ level (illustrated by the legend of the **Left tile diagram**) are in line with that of ClassyFire. The ‘false’ indicates the classified classes are completely inconsistent with that of ClassyFire. Noise is added into original dataset to evaluate the stability of MCnebula algorithm. Both for ‘true’ and ‘false’ assessment, the arrow indicates the middle noise or high noise lead to accuracy shift (increasing or decreasing). The accuracy evaluation are only performed with the classified feature number . If the noise lead to classified number < 50, the class is excluded from assessment of noise tolerance. The **Right horizontal bar plot** indicates the classified amount of features.

**Fig. 6 | Evaluation of classified amount，identification accuracy and noise tolerance of benchmark method and MCnebula**. **a)** Figure 6a illustrates a comparison of classified amount and noise tolerance of MCnebula and benchmark method. When noise is added into original dataset, some classified feature amount is occurred < 50. A cut-off (amount 50) is set to exclude these classes from assessment. **b)** Figure 6b illustrates the identified accuracy of MCnebula. A cut-off (Tanimoto similarity 0.5) is set to get structures of high matching score for evaluation.

**Fig. 7 | Heatmap and metabolomic pathway analysis of ‘Acyl carnitines’ (ACs), ‘Lysophosphatidylcholines’ (LPCs), ‘Bile acids, alcohols and derivatives’ (BAs) in serum metabolomic dataset.** Figure 7**a**, **c** and **e** show heatmap of level of ACs, LPCs and BAs. **b)** The carnitine system in mitochondria. Abbreviation: CPT1, carnitine-palmitoyltransferase-1; CACT, carnitine-acylcarnitine translocase; CrAT, carnitine acetyltransferase; CPT2, carnitine-palmitoyltransferase-2. **d**) Enrichment analysis of LPCs in pagerank algorithm with Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolomic pathway. Abbreviation: P A2, phospholipase A2; PC-Sterol O-AT, phosphatidylcholine-sterol O-acyltransferase; LP, lysophospholipase; 1-AGPC O-AT, 1-acylglycerophosphocholine O-acyltransferase; **f**) Enrichment analysis of BAs in pagerank algorithm with KEGG metabolomic pathway. Abbreviation: βGC, beta-glucuronidase; βGCS, beta-D-Glucuronoside; GT, glucuronosyltransferase; TCDC 6α-H, taurochenodeoxycholate 6alpha-hydroxylase; TCDC, taurochenodeoxycholate; GCC, Glycocholate; GCCDC, Glycochenodeoxycholate; Conju. BAs syn., ‘Conjugated bile acid biosynthesis, cholate’; BA-CoA, bile acid-CoA:amino acid N-acyltransferase.

**Fig. 8 | Marking features with fold change in child-nebulae to explore chemical transformation during processing of *E. ulmoides*.** Figure 8a illustrates the rank of variation relative abundance (VRA) of classified classes in *E. ulmoides* dataset. The VRA is calculated as: feature amount of || > 1 divided by feature sum in classified nebula. Those classes of VRA > 0 in nebula-index are extracted to generated neo-nebula-index. Figure 8b, the neo-child-nebulae are visualized according to neo-nebula-index and illustrates the features of || > 1 in classified nebula. Other mapping annotation refer to figure 1.

## Supplementation

**Fig. S1 | Evaluation of classified accuracy and noise tolerance of benchmark algorithm.** The illustration of this figure refer to figure 5.

**Fig. S2 | In-depth visualization of child-nebula of ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs).** The illustration of this figure refer to figure 4.

**Fig. S3 | 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 illustration of this figure refer to figure 4.

**Fig. S4 | Mass spectrometry inspection for remarkable features of lignans and iridoids in *E. ulmoides* dataset.** The features are picked with || > 1, Tanimoto similarity > 0.5, and fine peak shape. (**a**) The extracted ion chromatogram (EIC) plot illustrates 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) obtained 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. S5 | Interrogation of classes distribution of raw spectral library collection in child-nebulae.** The illustration of this figure refer to figure 1.