# Figure legend

## Figure 1-6

**Fig. 1 | Data stream of MCnebula workflow.** The MCnebula workflow can be divided into two parts, depending on the platform on which the data presents. The first is the part beyond R (before MCnebula2): from the **Samples** to **LC-MS/MS** to obtain the raw data; the stage of **Convert raw data** is implemented using the popular MSconvert tool derived from Proteowizard; for **Feature detection**, users can implement it with any LC-MS processing tool, such as MZmine, XCMS, OpenMS, etc.; then .mgf or other file format of MS/MS spectra is imported into SIRIUS for computations. The part inside R, MCnebula2 implements integrating data and creating Nebulae within ‘mcnebula’ object (see section of MCnebula2 algorithm in article for details).

**Fig. 2 | Tracing top ‘features’ in Child-Nebulae of serum metabolomics dataset.** According to the rankings of ‘features’ by statistic analysis, the top ‘features’ are marked with different colors in Child-Nebulae.

**Fig. 3 | Showing log2(Fold change) of groups in Child-Nebulae of serum metabolomics dataset.** The log2(Fold change) value of HM versus HS group is shown in Child-Nebulae as gradient color. The nodes with white color indicate ‘features’ with missing quantification value (these ‘features’ were detected in our re-analysis, but not in Wozniak et al.)

**Fig. 4 | In-depth visualization of Child-Nebulae of ‘Acyl carnitines’** **a)** The nodes of top ‘features’ 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 candidate of Chemical structure of ‘features’ are mapped into nodes. 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 nodes without ring diagram indicate ‘features’ with missing quantification value (these ‘features’ were detected in our re-analysis, but not in Wozniak et al.) The Bar plot map PPCP of structural (sub-structural or dominant structural) classes for the ‘feature’. **b)** The zoom in node of ‘feature’ 2068 (ID) and its legend.

**Fig. 5 | Heat maps of ‘Acyl carnitines’ (ACs), ‘Lysophosphatidylcholines’ (LPCs), ‘Bile acids, alcohols and derivatives’ (BAs) in serum metabolomics dataset.** Figure 5**a**, **c** and **e** show heat maps of level of ACs, LPCs and BAs.

**Fig. 6 | Mechanism for filtering chemical classes of MCnebula2.** This figure illustrates how MCnebula2 filters chemical classes of prediction from ‘features’ to form a Nebula-Index to create Child-Nebulae. The **Inner filter** filter out the chemical classes by Regex match of names (names without without Arabic numerals) and set threshold for value of posterior probability. To create **Stardust Classes**, the previous filtered data is re-grouped according to chemical classes instead of ‘features’ ID. The **Cross filter** conduct further filtering of chemical classes via combining Stardust Classes and ‘features’ annotation data. The details of Cross filter was described in MCnebula2 Algorithm section in article.

## Supplementation

**Fig. S1 | Evaluation of classified accuracy of MCnebula** 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 their 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. Medium and high 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). In general, for each evaluated chemical class , there are two pairs of arrows, with the left pair indicating the ‘T’ offset and the right pair indicating the ‘F’ offset. Each pair of arrows: first starts at the position indicated by the yellow arrow and ends at the position indicated by the purple arrow; then starts at the position indicated by the purple arrow and ends at the purple bar. 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 number of features.

**Fig. S2 | Evaluation of classified accuracy of benchmark method (GNPS).** The illustration of this figure refer to Fig. S1.

**Fig. S3 | Comparison of classifying of MCnebula with benchmark method and Evaluation of identification accuracy of MCnebula**. **a)** Fig. S3a illustrates a comparison of classified number of MCnebula and benchmark method. When noise is added into original dataset, some number of classified features are occurred < 50, a cut-off (number 50) is set to exclude these classes from assessment. **b)** Fig. S3b illustrates the identified accuracy of MCnebula. A cut-off (Tanimoto similarity 0.5) is set to get chemical structures of high matching score for evaluation.

**Fig. S4 | Parent-Nebula of serum metabolomics dataset.** In **Parent-nebula**, ‘features’ are mapped as nodes in network graph. The edges illustrated the spectral similarity of adjacent ‘features’. Not all ‘features’ are shown in the Parent-Nebula, as the isolated nodes are removed.

**Fig. S5 | Child-Nebulae of serum metabolomics dataset.** The Child-Nebulae are created according to chemical classes in Nebula-Index. The classified ‘features’ of chemical classes are mapped into corresponding Child-Nebulae.

**Fig. S6 | In-depth visualization of Child-Nebulae of ‘Lysophosphatidylcholines’ and ‘Bile acids, alcohols and derivatives’.** See Fig. 4 for description.

**Fig. S7 | Pathway analysis of ‘Acyl carnitines’ (ACs), ‘Lysophosphatidylcholines’ (LPCs), ‘Bile acids, alcohols and derivatives’ (BAs) in serum metabolomics dataset.** **a)** The carnitine system in mitochondria. Abbreviation: CPT1, carnitine-palmitoyltransferase-1; CACT, carnitine-acylcarnitine translocase; CrAT, carnitine acetyltransferase; CPT2, carnitine-palmitoyltransferase-2. **b)** 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; **c)** 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. S8 | Tracing top ‘features’ in Child-Nebulae of herbal medicine dataset** According to the rankings of ‘features’ by statistic analysis, the top ‘features’ are marked with different colors in Child-Nebulae.

**Fig. S9 | MS/MS spectra of top ‘features’ of herbal medicine dataset.** For top ‘features’, the mirrored MS/MS spectra plots illustrated the raw MS/MS spectra (back bar) and the noise filtered MS/MS spectra (red bar) predicted by SIRIUS. The dot above the bar implied a corresponding relation. The top candidate of chemical structure of ‘features’ are mapped into the plot.

**Fig. S10 | Extracted ions chromatography (EIC) of top ‘features’ of herbal medicine dataset.** The EIC plot illustrates the peak shape of the top ‘features’ (drew via MCnebula; detected via Automated Data Analysis Pipeline (ADAP) algorithm in MZmine2).

**Fig. S11 | Focus on location of top ‘features’ in annotated Child-Nebulae.** Fig. S11 **a**, **b** and **c** illustrate local view of the annotated Child-Nebula of ‘Iridoids and derivatives’, ‘Dialkyl ethers’ or ‘Phenylpropanoids and polyketides’, respectively. Fig. S11 **d** and **e** show the chemical structure of ‘feature’ (compound) of ID 2110 and ID 854, respectively. Fig. S11 **f** and **g** show the mirrored MS/MS spectra (refer to Fig. S9 for description) and extracted ions chromatography (EIC) of the ‘features’.