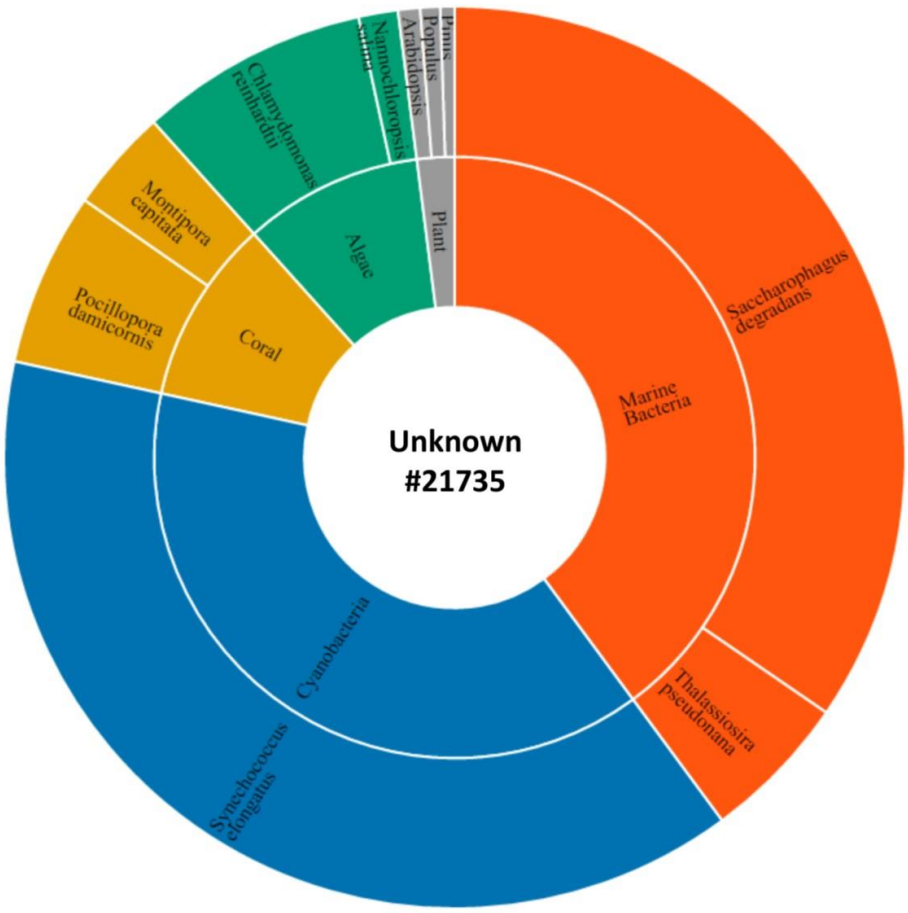
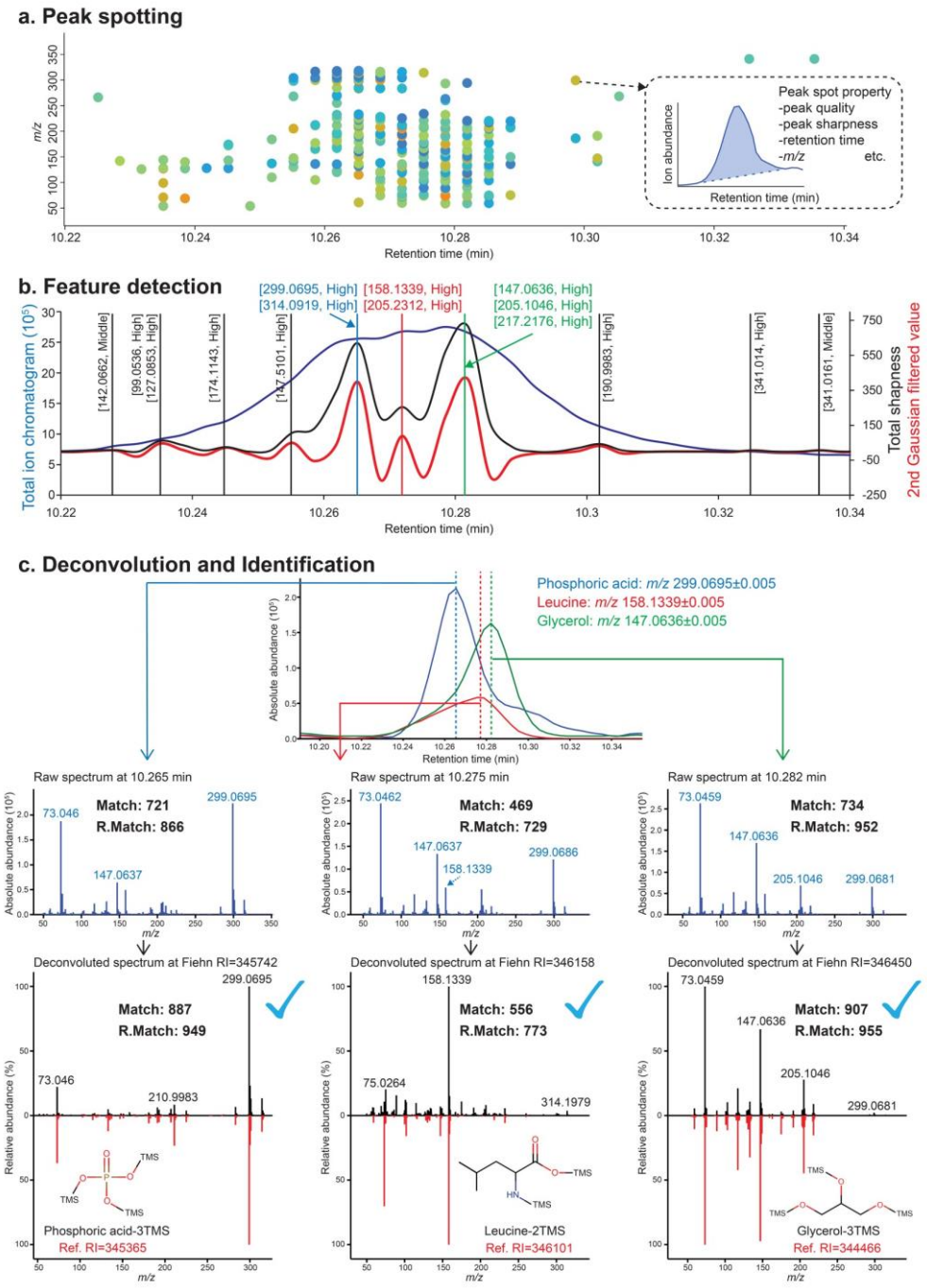


Supplementary Figure 1

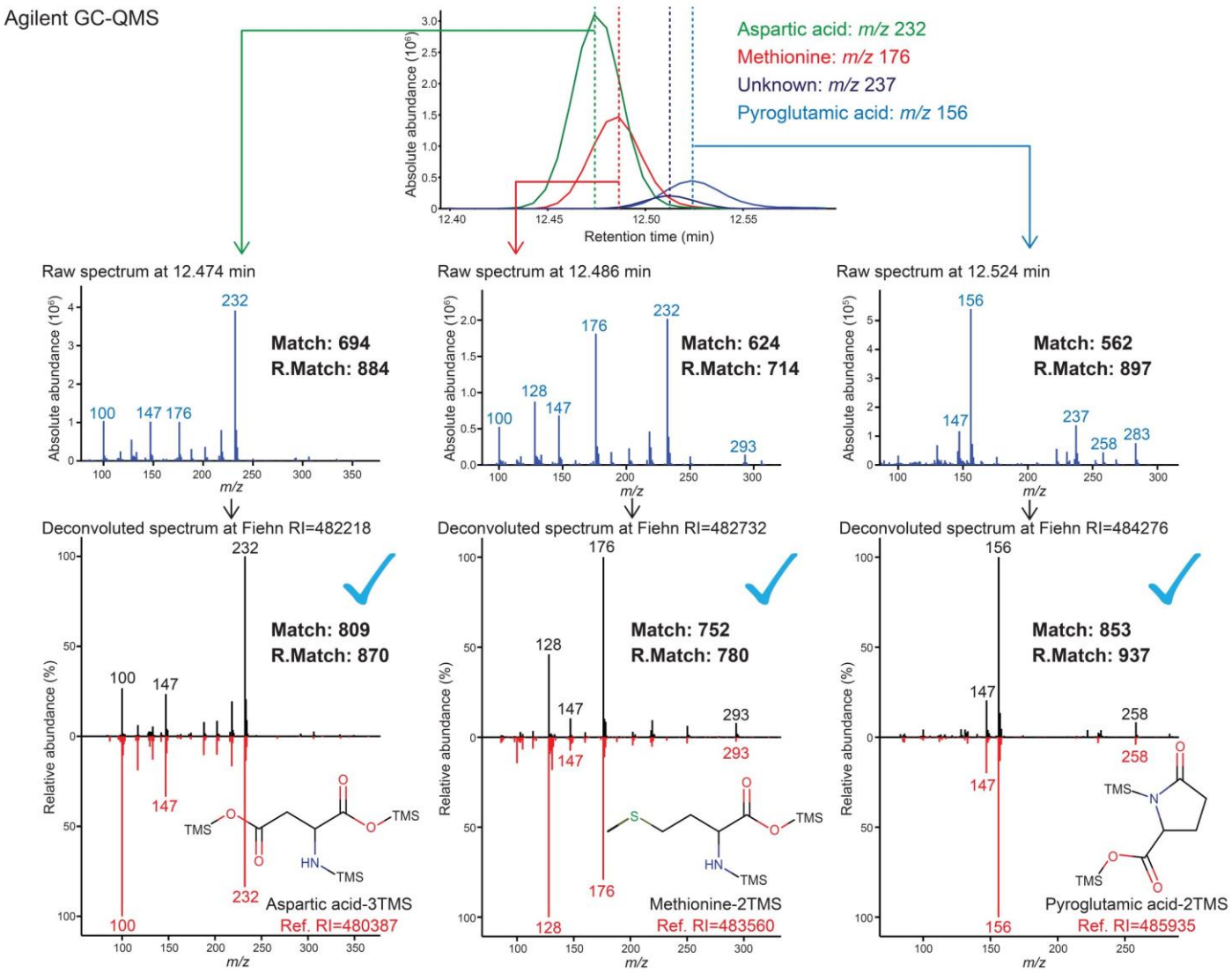
Cross-study specificity and relevance analysis of unknown BinBase ID 21735 with BinVestigate.



Supplementary Figure 2

Methodology and example for MS-DIAL 2.0 program.

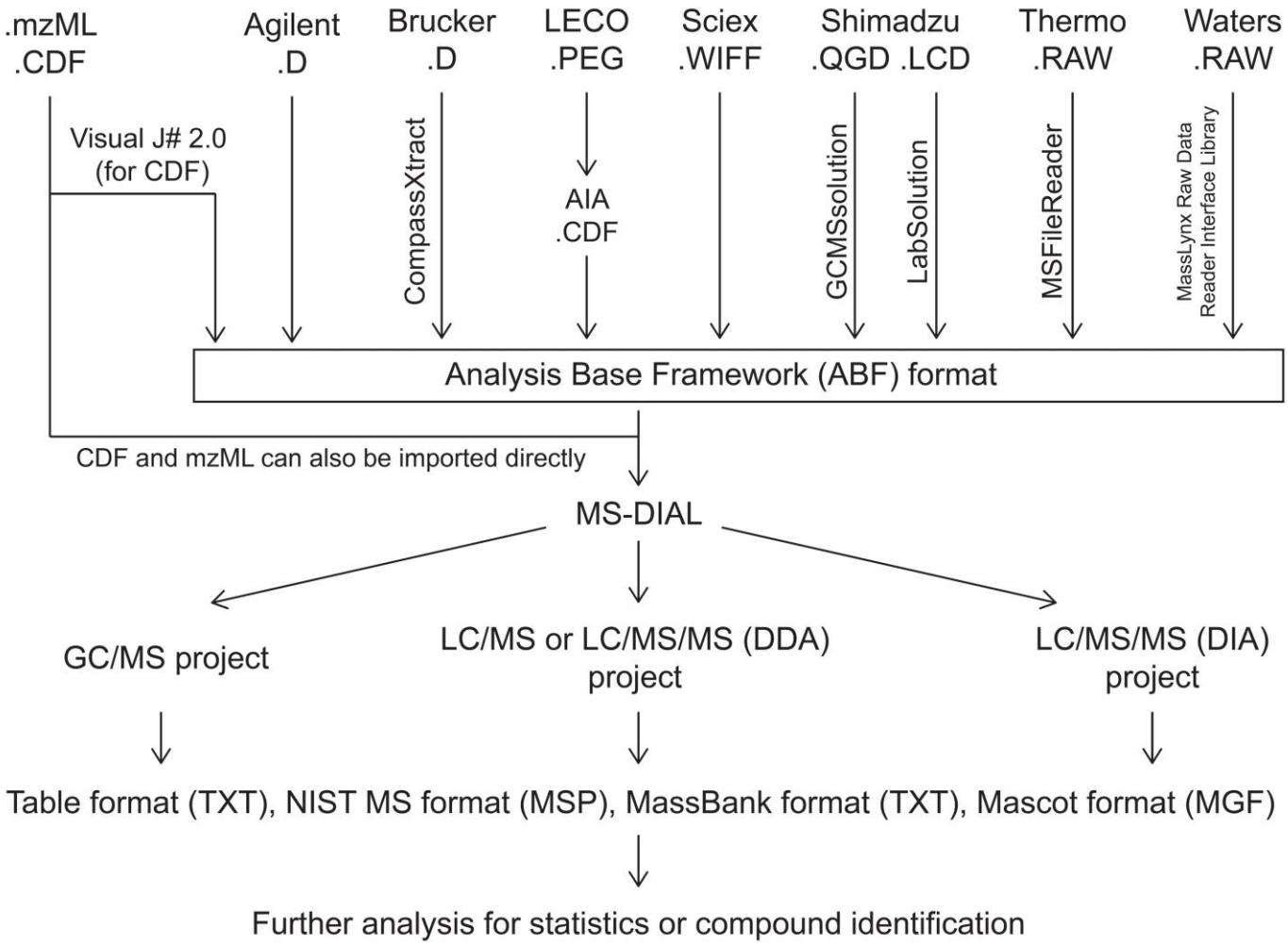
(a) Peak spotting: to determine fragment ions for GC-MS spectra, the detected m/z-RT features are termed as ‘peak spots’ with computed peak quality and peak sharpness values. (b) Feature detection: all peak spots with identical peak widths and peak top retention times are combined into single array. For each array, peak sharpness values are totaled and a second Gaussian derivative filter is applied to construct ‘peak groups’. (c) Deconvolution and identification: MS1Dec chromatogram deconvolution and open access MoNA mass spectral database are utilized to annotate the coeluting metabolites – phosphate, leucine, and glycerol – with 0.4-0.6 s peak top differences. The terms “Match” and “R.Match” mean dot-product and reverse dot-product values calculated in NIST MS search program, respectively.



Supplementary Figure 3

MS-DIAL 2.0 deconvolution example for Agilent GC-Q(MS).

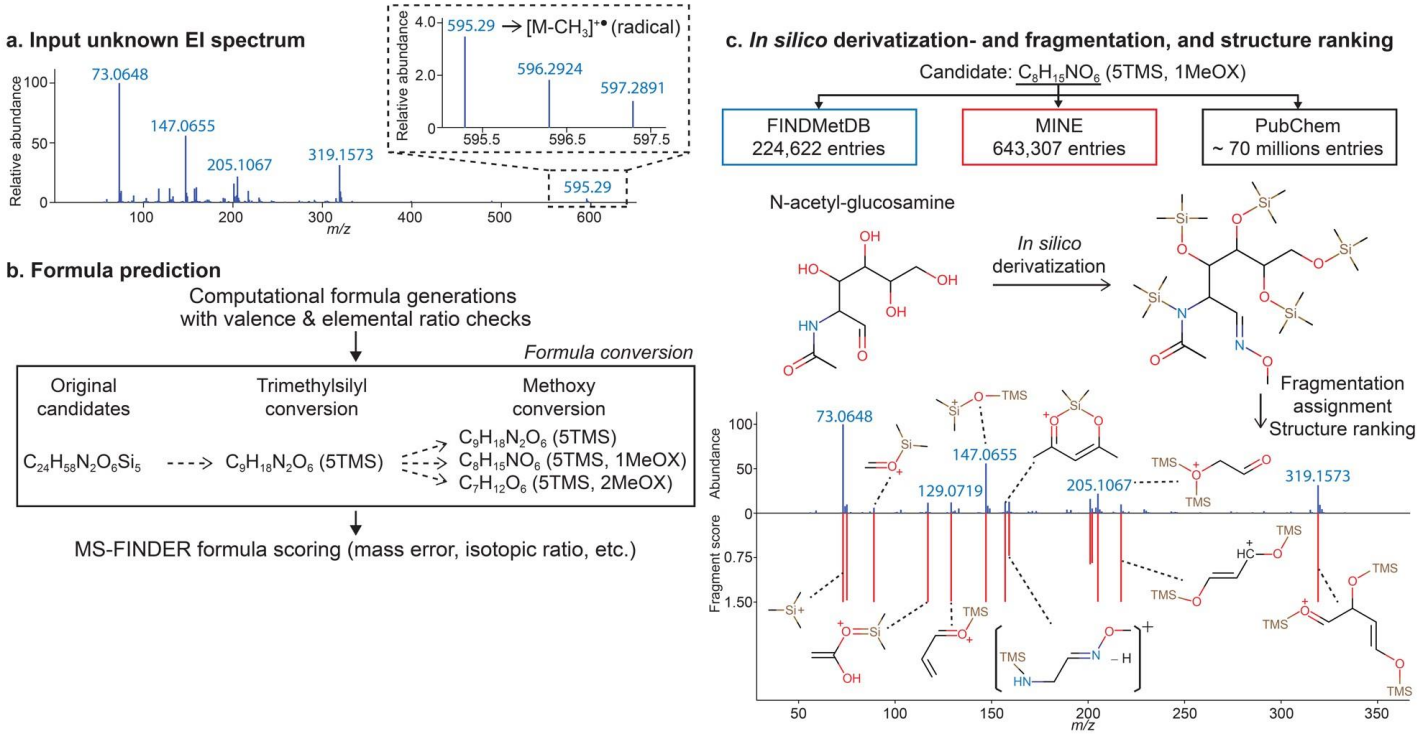
The accuracy of GC-MS chromatogram deconvolution is confirmed by analyzing a biological sample in Agilent GC-Q(MS) system. The other examples using LECO GC-TOF(MS), Shimadzu GC-Q(MS), Bruker GC-Q(MS), and Thermo GC-QExactive (MS) data are shown in Supplementary Data 1.



Supplementary Figure 4

Data stream of the MS-DIAL 2.0 program.

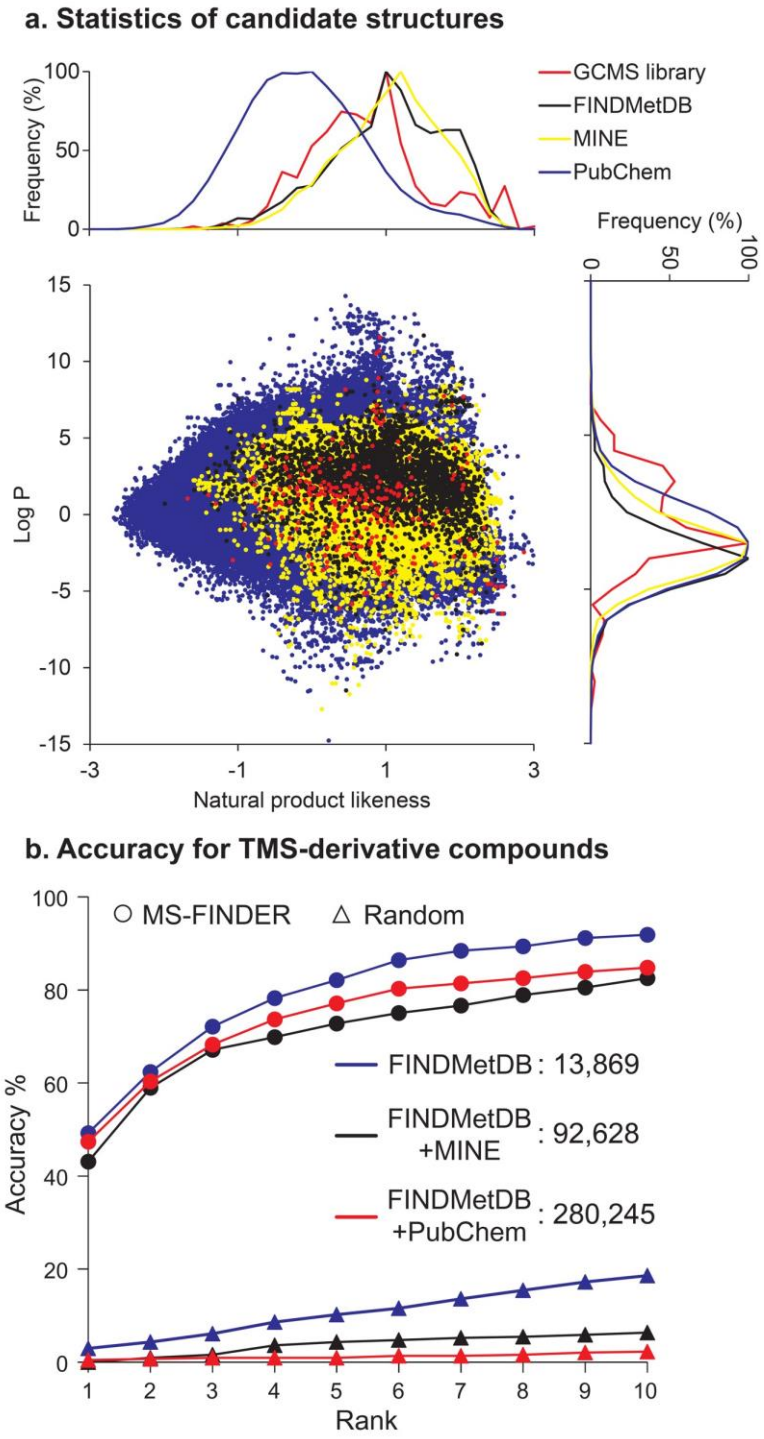
MS-DIAL 2.0 is designed as a universal software for MS data processing. First, MS vendor format or common format (mzML/CDF) data are converted to the ABF binary format for rapid data retrieval while the common formats, while mzML and netCDF can be directly imported. Then MS-DIAL 2.0 performs chromatogram deconvolution with support for any MS analytical platform, ranging from low and high resolution GC-MS (MS/MS) or LC-MS (MS/MS) to data dependent or data independent acquisition method. Finally, the program achieves compound annotation by matching against mass spectral library and further statistical analysis.



Supplementary Figure 5

Workflow for the MS-FINDER 2.0 program.

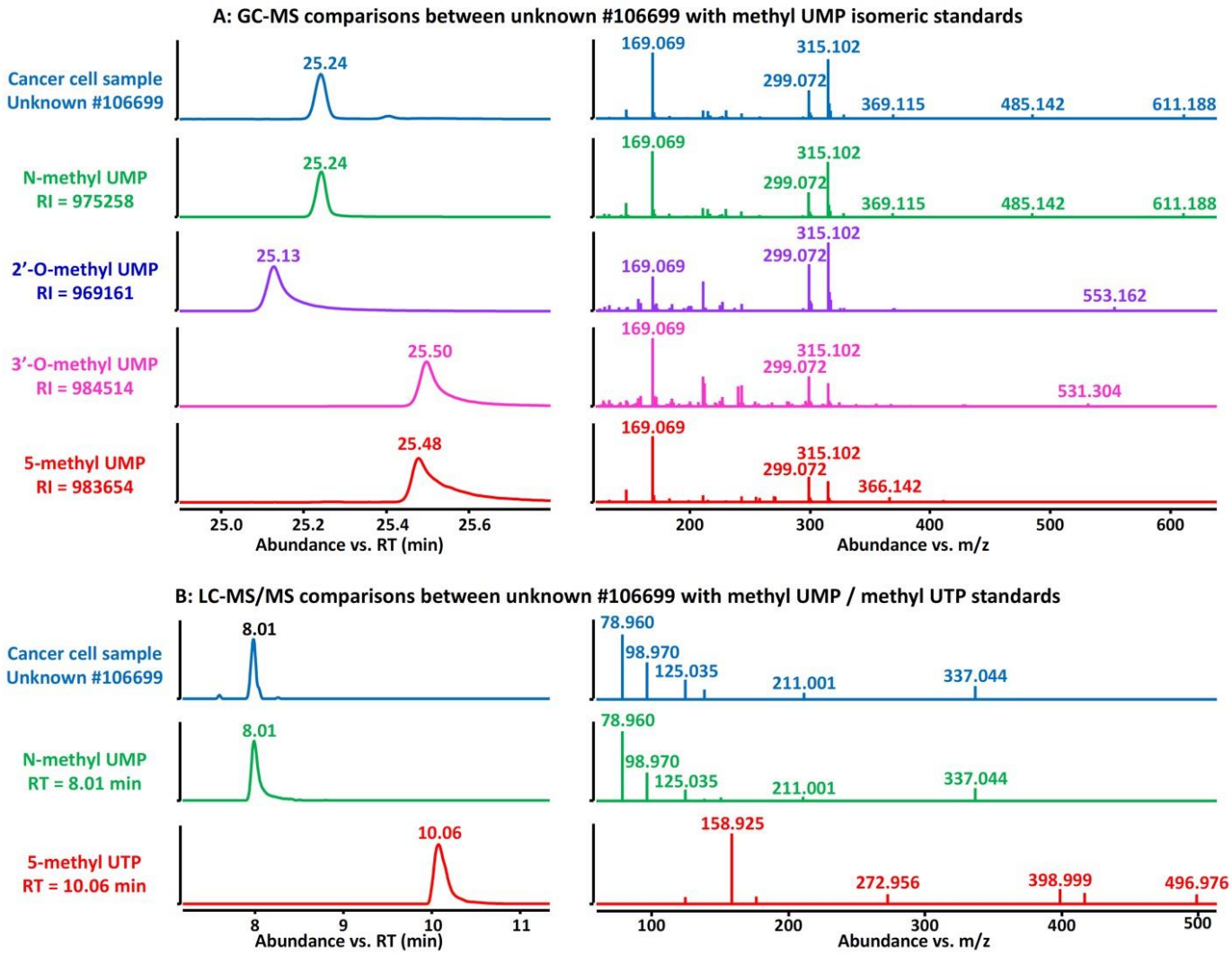
(a) Accurate mass GC-EI-MS data is utilized as input with defined molecular ion and its adduct type. (b) Derivatized formulas are computationally generated and ranked based on valence and elemental ratio check in combination with the original MS-FINDER formula scoring algorithm. (c) Structure candidates are retrieved from multiple databases. After the candidate is computationally derivatized, the candidates are ranked by the result of substructure assignments from computational mass fragmentations.



Supplementary Figure 6

Performance validation of the MS-FINDER 2.0 program.

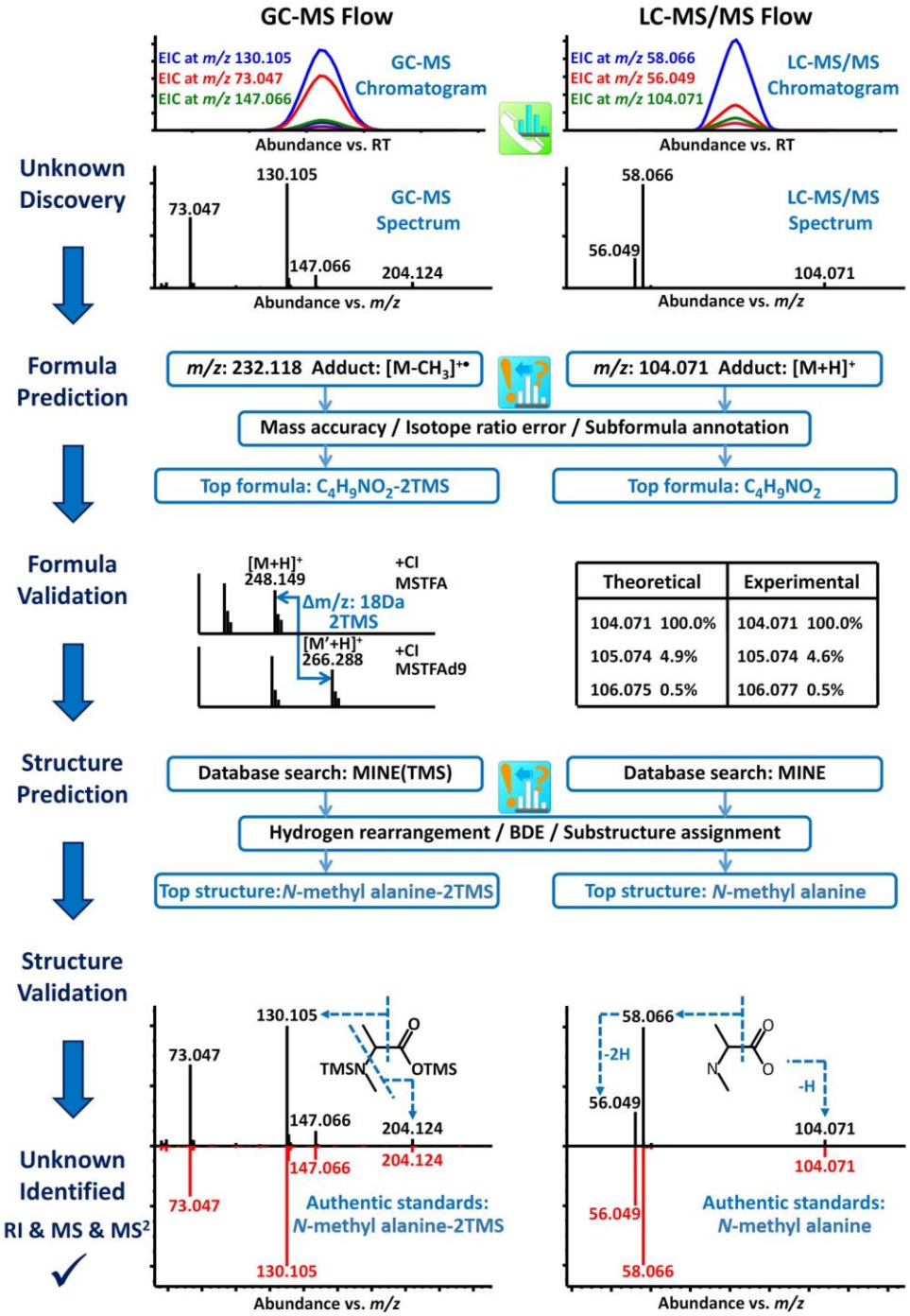
(a) The compound logP and natural product likeness comparison between the metabolite dataset for accuracy test (denoted as 'GCMS library') with the databases in MS-FINDER 2.0 (FINDMetDB, MINE, and PubChem). (b) The performance test results of MS-FINDER 2.0 and random sampling method with three structure resource sets.



Supplementary Figure 7

Authentic standard validation for the identification of *N*-methyl-UMP.

Mass spectra and retention times in GC-MS (a) and LC-MS/MS (b) were compared between BinBase ID 106699 in cancer cell sample with chemically synthesized *N*-methyl-UMP standard, as well as other isomeric compounds including 2'-*O*-methyl, 3'-*O*-methyl, and 5-methyl-UMP(UTP).

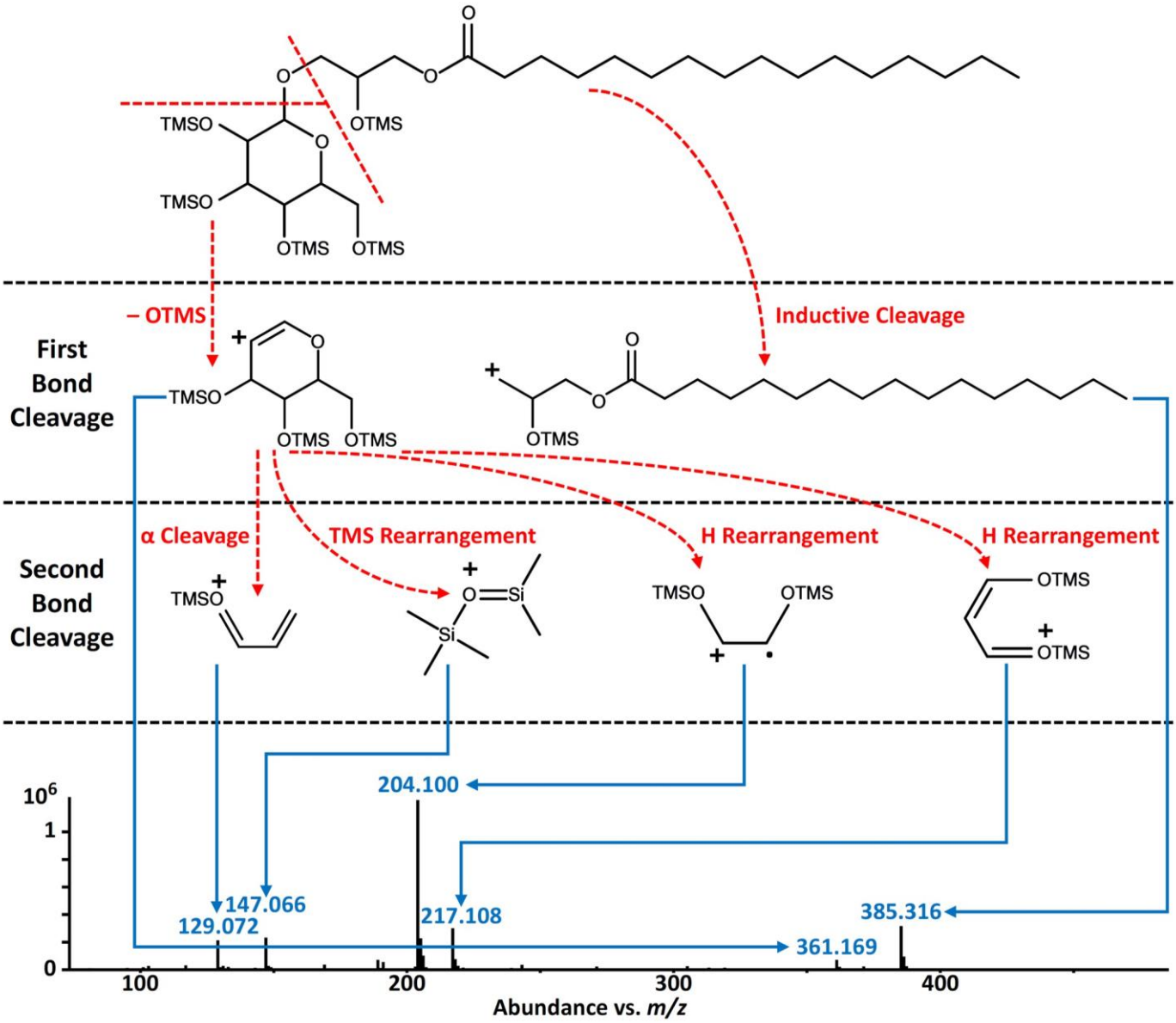


Supplementary Figure 8

A workflow for GC-MS and LC-MS/MS identification of *N*-methylalanine in MS-FINDER 2.0.

The workflow is the same as shown in Figure 3. High resolution GC-MS analytics was used for structure elucidation (left), then LC-MS/MS was applied as additional evidence line (right). Unknown discovery: fragment ions and molecular adduct ions of BinBase ID 160842 were deconvoluted by MS-DIAL 2.0. Formula prediction: C4H9NO2 was scored and ranked at 1st in MS-FINDER 2.0 based on mass errors, isotope ratio errors, and subformula assignments. Formula validation: for GC-MS flow, chemical ionization data with different derivatization methods (MSTFA vs. MSTFAd9) were obtained to verify the formula as well as to yield the number of acidic protons; for LC-MS flow, between theoretical values and experimental values, the mass errors were only 1 mDa, and the isotopic ratio

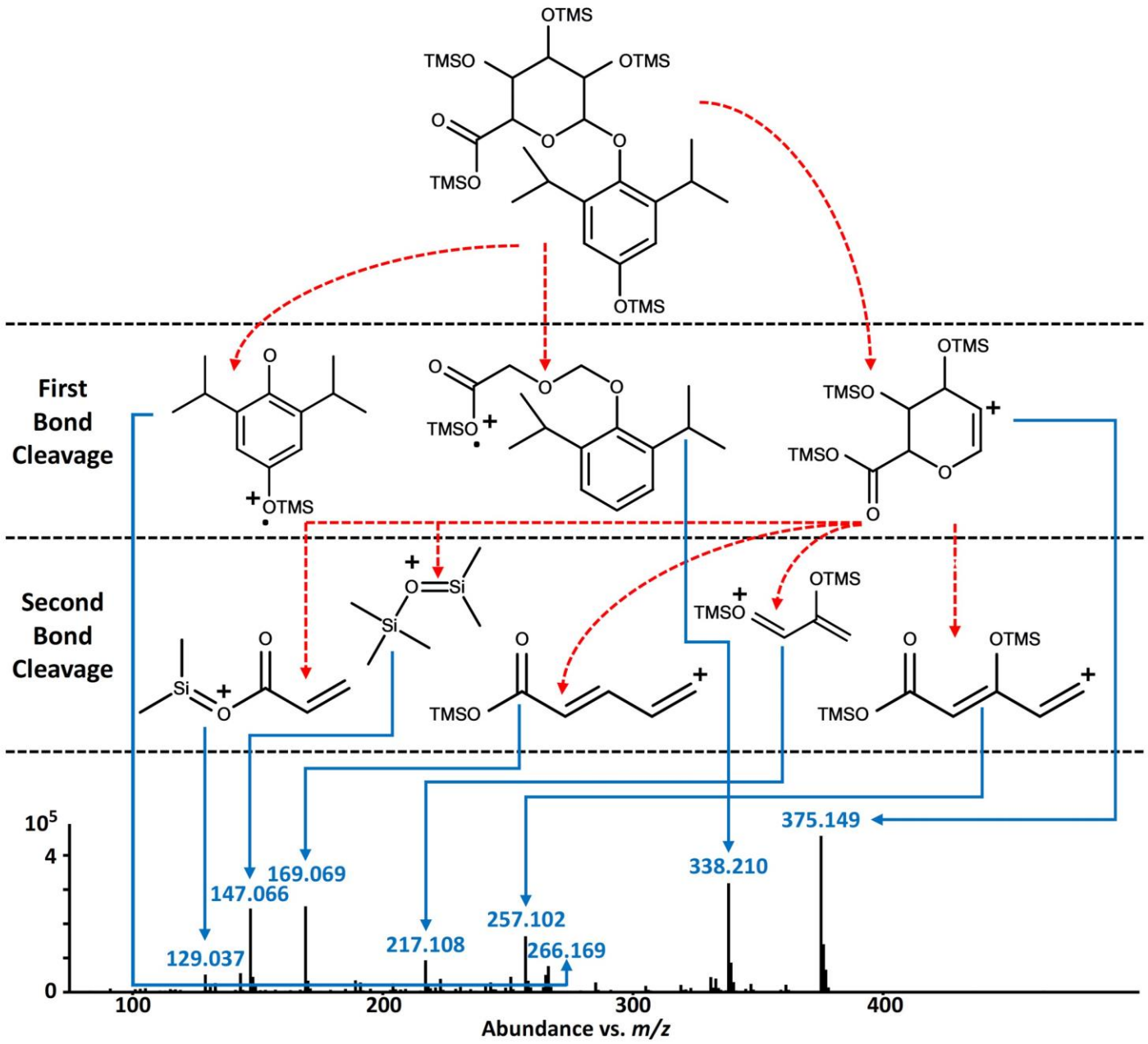
errors were within 1%. Structure prediction: structure candidates were retrieved from MINE DB in addition to internal metabolome database, and *in silico* fragmented based on hydrogen rearrangement rules, bond dissociation energy, and comprehensive fragmentation rule library (including GC-EI-MS and LC-ESI-MS/MS). *N*-methyl-alanine was ranked at the most likely structure in MS-FINDER 2.0 with computational assigned substructures. Structure validation: the mass spectra and retention times in GC-MS (left) and LC-MS/MS (right) were matched with chemically synthesized *N*-methyl-alanine standard.



Supplementary Figure 9

GC-MS identification of lyso-monogalactosyl-monopalmitin with *in silico* fragmentation and substructure assignments in MS-FINDER 2.0.

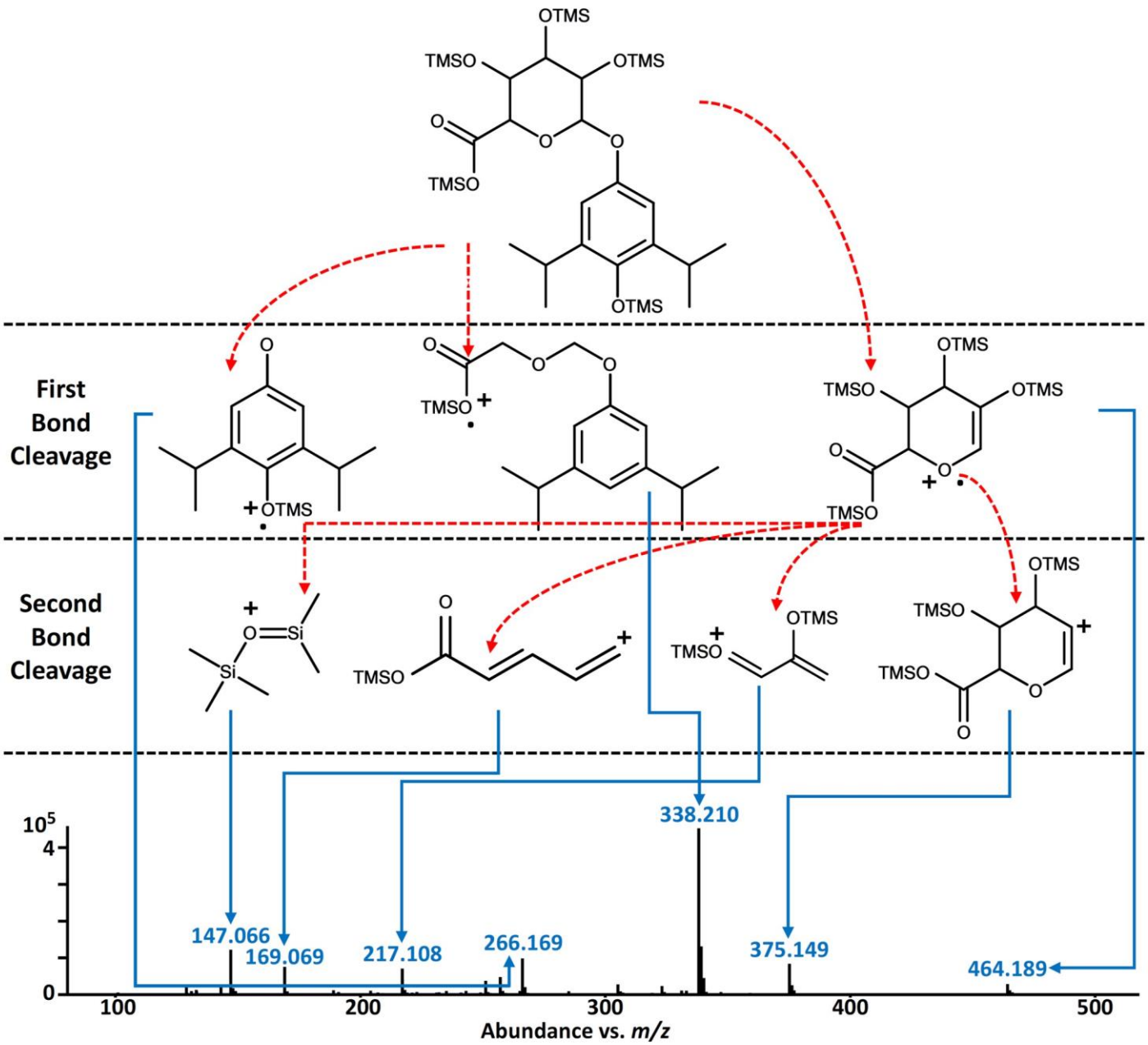
After the structure candidates were suggested by MS-FINDER 2.0, the molecular skeleton was confirmed by the result of substructure assignments with manual inspection.



Supplementary Figure 10

GC-MS identification of 4-hydroxypropofol-1-glucuronide with *in silico* fragmentation and substructure assignments in MS-FINDER 2.0.

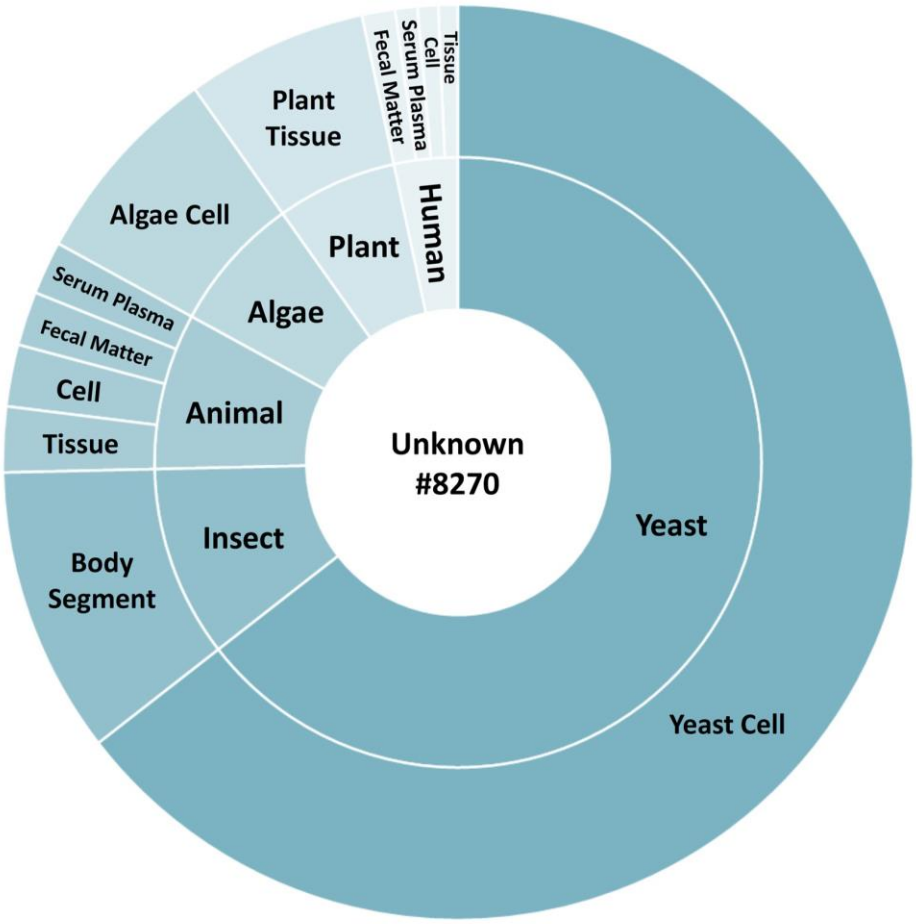
After the structure candidates were suggested by MS-FINDER 2.0, the molecular skeleton was confirmed by the result of substructure assignments with manual inspection. Finally, the structure was identified by the authentic standard compound.



Supplementary Figure 11

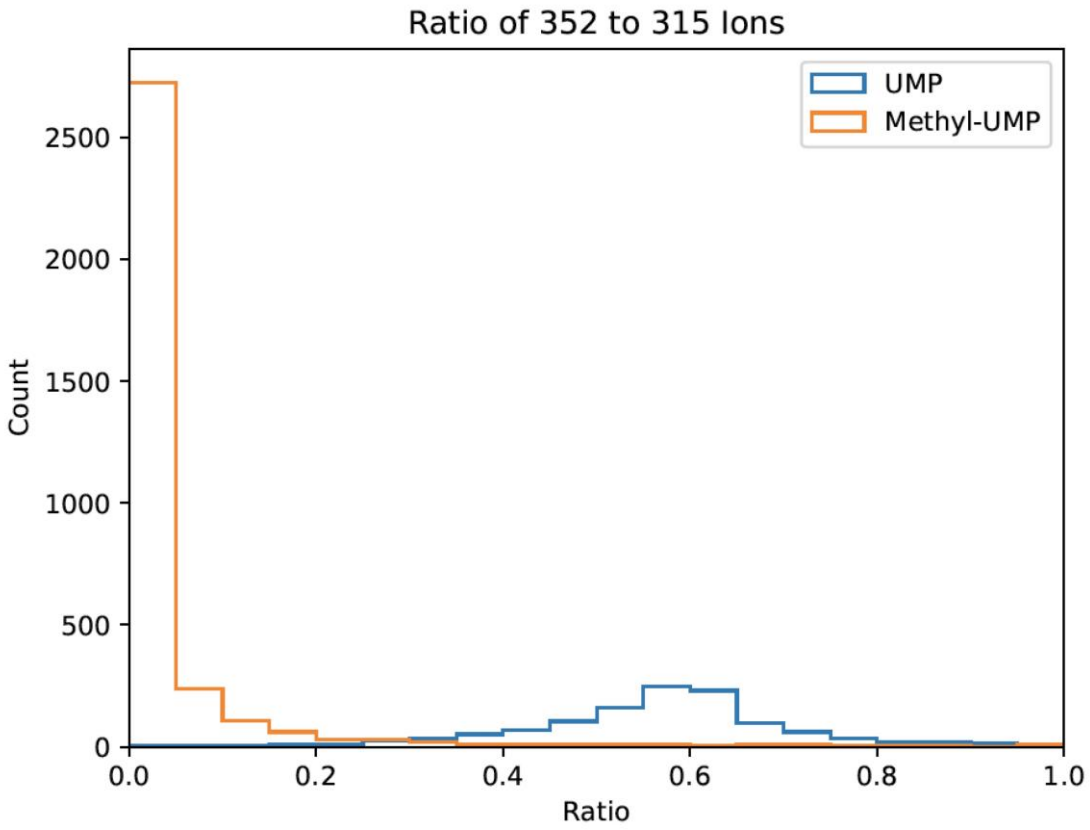
GC-MS identification of 4-hydroxypropofol-4-glucuronide with *in silico* fragmentation and substructure assignments in MS-FINDER 2.0.

After the structure candidates were suggested by MS-FINDER 2.0, the molecular skeleton was confirmed by the result of substructure assignments with manual inspection. Finally, the structure was identified by the authentic standard compound.



Supplementary Figure 12

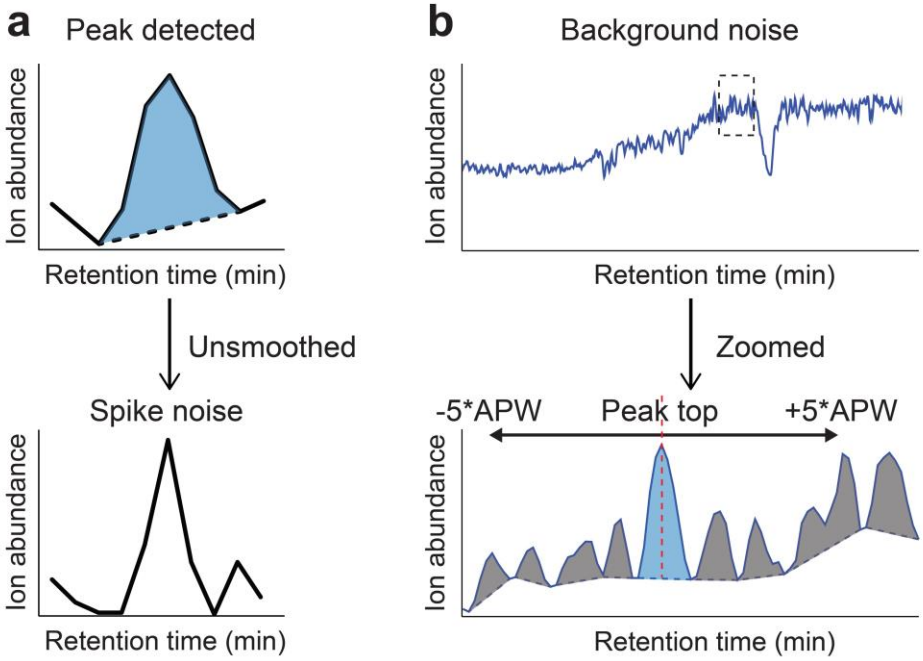
Cross-study specificity and relevance analysis of unknown BinBase ID 8270 with BinVestigate.



Supplementary Figure 13

Investigation for the unique mass ratio of *m/z* 352 to *m/z* 315 among the EI-MS spectra of UMP and *N*-methyl-UMP in BinBase.

The x- and y-axes show the ratio of *m/z* 352 to *m/z* 315 and the count of EI-MS records, respectively.



Supplementary Figure 14

MS-DIAL 2.0 background subtraction in peak detection.

(a) Peaks were excluded as spike noise if the ion abundance of one neighbor point from the peak top is zero in unsmoothed raw chromatogram. (b) Peaks were excluded as baseline noise if 4 spike noise signals were programmatically detected within a ± 5 APW region of the peak top.