Supporting Information

Moving beyond the van Krevelen diagram: A new stoichiometric approach for compound classification in organisms.

Albert Rivas-Ubach[1†*], Yina Liu[1,2†], Thomas S. Bianchi[3], Nikola Tolic[1], Christer Jansson[1], Ljiljana Paša-Tolić[1]

1. Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, 99354, WA, USA.

2. Geochemical and Environmental Research Group, Texas A&M University, College Station, 77845, TX, USA.

3. Department of Geological Sciences, University of Florida, Gainesville, 32611-2120, FL, USA.

† Authors contributed equally to this manuscript.

*** Author of correspondence:**
Albert Rivas-Ubach
Environmental Molecular Sciences Division,
Pacific Northwest National Laboratory,
Richland, WA, USA, 99354
Tel: 971 319 5962
e-mail: albert.rivas.ubach@gmail.com

Table of Contents

Supplementary text

Considerations of the compound databases and for the determination of the MSCC.

The stoichiometry and elements for different compound categories were examined to establish the MSCC thresholds with a minimum overlap between compound categories (Figure S-2). Below we describe some aspects of the databases that had to be considered before determining the MSCC thresholds.

Carbohydrates (amino sugars excluded) is a group of compounds containing only C, O and H. Therefore, it was classified solely based on their H:C and O:C ratios, resembling the classic compound classification by vK diagram. Phosphorylated sugars, with higher O:C ratios, and polysaccharides also match the proposed MSCC defining Carbohydrates$_c$. Most polysaccharides (glycogen, cellulose, starch) are composed of 6-carbon monosaccharides with a molecular formula of $(C_6H_{10}O_5)_n$ and O:C and H:C ratios of 0.83 and 1.66, respectively.

According to the classic definition, **amino-sugars** are monosaccharides with one hydroxyl group (-OH) replaced by an amine group (-NH_2); however, a large variety of amino-sugar derivatives are still commonly considered amino-sugars. The complexity and diversity of amino-sugar biochemistry makes it challenging to accurately define amino-sugars' characteristics based on databases. The replacement of a hydroxyl group by an amine shifts the O:C ratios to lower values, and thus some amino sugars could be wrongly assigned as peptides or amino-lipids, when considering exclusively their O:C and H:C ratios. Hence, the inclusion of the N:C ratio is necessary to separate amino-sugars from peptides and high O:C amino-lipids. Furthermore, some amino-sugars can undergo multiple reactions to yield structural derivatives that are substantially different from their original parent sugar, and this can largely shift their original O:C, H:C and N:C molecular ratios. It is important to consider that some of these substantial modifications can thus result in molecules that no longer resemble a "typical" sugar found in organisms. For this reason, we excluded the amino-acid derivatives with O:C < 0.6 (typically highly dehydroxylated), N:C > 0.2 (e.g. high degree of replacement of hydroxyl group by amino group for relatively small molecules) and/or amino-acid derivatives containing long-carbon side chains. These excluded compounds represented 14.3% of the molecules considered amino-sugars in databases and mainly represent metabolites with antibiotic properties, specifically istamycins, fortimicins, and sannamycins. The MSCC for A-Sugars$_c$ is also suitable to properly classified poly-amino-sugars such as chitin and amino-sugar-phosphates.

Nucleotide O:C and H:C ratios showed considerable overlap with other compound categories (Figure 1 main text). However, nucleotides can be segregated from the other compound categories if

N:C, C:P, N:P, N, P and S, and the mass range are considered. Yet, we found some nucleotides that, within the classified Nucleotides$_c$, also fitted within the stoichiometric constraints of Protein$_c$ and A-Sugars$_c$ (Table 2 main text). Contrary, we did not find any peptide from the 93,245 or any amino-sugars from the 142 included in the databases matching the proposed constraints of the Nucleotide$_c$ indicating that the probability of including a no-nucleotide compound as nucleotide is practically zero. For this reason, any double match found in Nucleotide$_c$ should be considered exclusively as a nucleotide.

Lipids, **peptides,** and **phytochemical compounds** showed a large overlap in O:C and H:C ratios (Figure 1 main text). All peptides contain N, but most phytochemical compounds or lipids do not; therefore the N:C ratio is a crucial discriminant variable between Peptides$_c$ and Lipids$_c$ and Phytochemical$_c$ (Figure S-9). On the other hand, H:C ratio is the stoichiometric ratio used to discriminate between Lipids$_c$ and Phytochemical$_c$ (Figures S-1, S-2 and S-9). The overlap of O:C and H:C ratios between lipids and phytochemical compounds is largely due to the fact that several secondary metabolites, such as polyketides or prenol lipids, are lipid-related.[1] We also found that all glucosinolates (phytochemical compounds), except those derived from phenylalanine, tyrosine and tryptophan, matched in the A-Sugars$_c$ (Table 2 main text). Glucosinolates are N and S containing compounds derived from amino acids that cannot be differentiated from amino-sugars.

Alkaloids, found in plants but also isolated from animals, insect, microorganisms, and invertebrates, are a very diverse group of secondary metabolites with large C:H:O:N variability.[2] Contrary to flavonoids, most alkaloids are commonly species-specific,[2] with only a limited number of those compounds present in a single organism.[3] The presence of alkaloids in samples represents thus an insignificant fraction of the total detected compounds; hence, alkaloids were not considered for the determination of MSCC for the Phytochemical$_c$.

Isoprenoids (prenol lipids), also known as terpenoids or terpenes, belong to both lipids and phytochemical compound categories,[1] and showed large overlap with Lipids$_c$ and Phytochemical$_c$ in vK diagrams (Figure S-10). We found no stoichiometric variable that could efficiently discriminate isoprenoids from lipids and phytochemical compounds. Although most isoprenoids would match into the current vK stoichiometric constraints of Lipids$_c$ (~80%), we decided to exclude them for the determination of the MSCC for Lipids$_c$ and Phytochemical$_c$. Thus, any isoprenoid compound present in the analyzed samples would be ultimately "correctly" matched into Lipids$_c$ or Phytochemical$_c$.

Validation of MSCC threshold computation for Lipids$_c$, Phytochemical$_c$ and Protein$_c$ - the categories showing the largest overlapping across their O:H:C:N:P stoichiometry.

To validate the robustness of the MSCC threshold, the most determinant stoichiometric thresholds for Lipids$_c$, Phytochemical$_c$ and Protein$_c$ were determined with 50% of the data from the lipids, phytochemical compounds and peptide databases, respectively. The 50% of compound from each database were randomly selected. Due to the low number of carbohydrates (82), amino-sugars (142), and nucleotides (37) included in the databases for MSCC determination, the validation of the stoichiometric thresholds for those groups using only 50% of the data was not considered; because the compound diversity on those categories (carbohydrates, amino-sugars, nucleotides) is relatively low in databases and thus will not be able to generate robust test results. It should be noted, however, the determination of stoichiometric thresholds for compound classification will be more accurate with more compounds included considered to calculate it.

H:C ratio is the most determinant stoichiometric ratio to separate Lipids from phytochemical compounds with minimum overlapping (Figs. S-2, S-10). We found that, using the 50% of compounds, the H:C boundary with the minimum proportion of compounds from both databases (minimal relative overlapping) was 1.32 (see file S-1). Therefore, the lowest H:C boundary for Lipids$_c$ would be 1.32, coinciding with the largest H:C boundary for Phytochemical$_c$.

N:C ratio is the most determinant stoichiometric ratio for discriminating peptides from lipids, and peptides from phytochemical compounds. The N:C thresholds with the minimal relative overlapping using the 50% of compounds for each group, were 0.126 for Protein$_c$ vs. Lipids$_c$, and for Protein$_c$ vs. Phytochemical$_c$ (see file S-1). Tehrefore, 0.126 was the minimum value for N:C for Protein$_c$, and the maximum value for Lipids$_c$ and Phytochemical$_c$.

Those results prove that H:C and N:C ratios, the most determinant stoichiometry to discriminate Lipids$_c$, Phytochemical$_c$ and Protein$_c$ remained identical if using 50% or 100% of compounds from the databases (see Table 1 of the main manuscript).

Additional validation of the MSCC performance for Lipids$_c$, Phytochemical$_c$ and Protein$_c$ - the categories showing the largest overlapping across their O:H:C:N:P stoichiometry.

The performance of the established thresholds (Table 1 of the main manuscript) of Lipids$_c$, Phytochemical$_c$ and Protein$_c$, the compound categories with the largest number of compounds and the largest overlapping across their elemental stoichiometry, were additionally tested by using the elemental stoichiometry of compounds that were not included in the databases utilized for MSCC determination. The performance for the stoichiometric thresholds for Lipids$_c$ and Phytochemical$_c$ (oxy-aromatic compounds) were tested by matching 764 and 330 compounds, respectively, from the *HMDB* database (http://www.hmdb.ca/) that were not found in the databases used to determine the MSCC (see Table S-1). The performance for Protein$_c$ stoichiometric thresholds was tested by using 1,200 random peptides from *Swiss-Prot* database (http://www.uniprot.org/) that were not previously utilized for MSCC determination.

We found that 96.99% of the lipids from *HMDB* database matched within the stoichiometric constraints of Lipids$_c$ (Table 1 main text) (0.26% matched into A-Sugar$_c$, 0.13% matched into Carbohydrates$_c$, 0.79% matched into Phytochemical$_c$, 0.92% matched into Protein$_c$, and 0.92% did not match any category) (see file S-2). For oxy-aromatic compounds (Phytochemical compounds), 97.27% of the compounds from the *HMDB* database matched into the stoichiometric constraints of Phytochemical$_c$ of the MSCC (Table 1 main text) (1.8% matched into Lipids$_c$, 0.6% matched into Protein$_c$ and 0.3% did not match any category) (see file S-2). From the 1,200 random peptides *from Swiss-Prot* database that were not utilized for the determination of the MSCC, we found that only 1 peptide did not match to any of the compound categories making thus the 99.92% of peptides matching properly into the Protein$_c$ of the proposed MSCC (Table 1 main text) (see database S-2).

The performance of the Lipids$_c$, Phytochemical$_c$ and Protein$_c$ tested with the databases used for their MSCC calculation (Table 2 main manuscript) was, therefore, very similar when using compounds not included in the databases for MSCC determination (see file S-2):

· Lipids$_c$: **97.1%** (Table 2 main manuscript; 30,729 total compounds) vs. **96.99%** (with only compounds not included for MSCC determination; 764 total compounds).

· Phytochemical$_c$: **96.5%** (Table 2 main manuscript; 7,774 total compounds) vs. **97.27%** (with only compounds not included for MSCC determination; 330 total compounds).

· Protein$_c$: **99.9%** (Table 2 main manuscript; 93,245 total compounds) vs. **99.92%** (with only compounds not included for MSCC determination; 1,200 total compounds).

Formula assignment error determination of Compound Identification Algorithm (CIA).

The MSCC applies to elemental formulas which are commonly assigned to metabolic features acquired from the samples. Correct elemental formula assignment to metabolic features is thus a critical prerequisite for accurate compound classification and subsequent data interpretation. To assess the final error of MSCC, we used the metabolite database described above to examine the performance of the automated CIA[4] for assigning elemental formulas. All formulas from the database were converted into exact masses before applying the CIA for elemental formula assignment. The CIA results were then compared to the known formulas from the database to determine correct assignment.

Applying the automated compound assignment algorithm[4] to all exact masses of compounds from the databases we found that 96.94% of the masses were correctly assigned, with only 0.21% of compounds not assigned and 2.84% incorrectly assigned. All carbohydrate formulas were correctly assigned, followed by lipids (98.08%), peptides (96.6%; including phosphorylated peptides), and phytochemical compounds (93.67%). An estimated 70.27% and 57.25% of nucleotides and amino sugars were correctly assigned to molecular formulas, respectively, while 21.62% and 34.78% were incorrectly assigned, and 8.11% and 7.97% not assigned (Table S3).

R script for compound classification of stoichiometric ratios.

Copy and paste the script below to "R-Studio", "Tinn-R", "RKward" or your favourite R editor/interface:

```
#############
### MSCC ###
#############

# VARIABLES REQUIRED IN THE DATASET (make sure your dataset includes the following variables with the names as described below; variables need to be in columns and the detected features need to be in rows):
# O.C <- O:C ratio column
# H.C <- H:C ratio column
# N.C <- N:C ratio column
# P.C <- P:C ratio column
# N.P <- N:P ratio column
# O <- O column
# N <- N column
# P <- P column
# S <- S column
# Mass <- exact mass column



## THE FOLLOWING 3 SECTIONS HAVE TO BE USED BY THE USER ##
# In "R", directories Paths are written with two backslashes "\\".
# Example: C:\\DATA\\MSCC\\R\\MSCC_Test.csv

# Read the DATASET in CSV format containing all the required variables.
# Example: C:\\DATA\\MSCC\\R\\MSCC_Test.csv
DATASET <- read.csv("Directory_of_the_dataset_in CSV_File", sep=",", header=T)

# Specify the directory of the resulting matchin results summary
# Example: C:\\DATA\\MSCC\\R\\MSCC_Test_Summary_Table.csv
Destination.File.Dataset <- "Directory_of_the_generated_matching_results_in_CSV_Format"

# Specify the directory for generating a summary of the reults in proportions
# Example: C:\\DATA\\MSCC\\R\\MSCC_Test_Summary_Proportions_Table.csv
Destination.File.Proportions <- "Directory_of_the_summary_proportion_results_in_CSV_Format"



## RUN THE FULL CODE BELOW ##

## 1st. STEP – ASSIGNATION OF COMPOUNDS ##
# Create a list for each compound category to keep the compound matches
list() -> Matching.Lipids
list() -> Matching.Carbohydrates
list() -> Matching.AminoSugars
list() -> Matching.Phytochemical
list() -> Matching.Protein.1
```

```
list() -> Matching.Protein.2
list() -> Matching.Nucleotides


# Loops for each compound category (we perform a single loop for each category to facilitate double matching detection)
# LIPID CONSTRAINTS
for (i in 1:nrow(DATASET)){
 if((DATASET[i,]$O.C <= 0.6) &&
    (DATASET[i,]$H.C >= 1.32) &&
    (DATASET[i,]$N.C <= 0.126) &&
    (DATASET[i,]$P.C < 0.35) &&
    (DATASET[i,]$N.P <= 5)){
   paste0("Lipid") -> Matching.Lipids[i]
 } else {
   paste0("") -> Matching.Lipids[i]
 }
}

# CARBOHYDRATE CONSTRAINTS
for (i in 1:nrow(DATASET)){
 if((DATASET[i,]$O.C >= 0.8) &&
    (DATASET[i,]$H.C >= 1.65) &&
    (DATASET[i,]$H.C < 2.7) &&
    (DATASET[i,]$N == 0)){
   paste0("Carbohydrate") -> Matching.Carbohydrates[i]
 } else {
   paste0("") -> Matching.Carbohydrates[i]
 }
 }

# AMINO-SUGAR CONSTRAINTS
for (i in 1:nrow(DATASET)){
 if((DATASET[i,]$O.C >= 0.61) &&
    (DATASET[i,]$H.C >= 1.45) &&
    (DATASET[i,]$N.C <= 0.2) &&
    (DATASET[i,]$N.C > 0.07) &&
    (DATASET[i,]$P.C < 0.3) &&
    (DATASET[i,]$N.P <= 2) &&
    (DATASET[i,]$O >= 3) &&
    (DATASET[i,]$N >= 1)){
   paste0("Amino.Sugar") -> Matching.AminoSugars[i]
 } else {
   paste0("") -> Matching.AminoSugars[i]
 }
}

# PHYTOCHEMICAL/OXYAROMATIC COMPOUND CONSTRAINTS
for (i in 1:nrow(DATASET)){
 if((DATASET[i,]$O.C <=1.15) &&
    (DATASET[i,]$H.C < 1.32) &&
    (DATASET[i,]$N.C < 0.126) &&
    (DATASET[i,]$P.C <= 0.2) &&
```

```
    (DATASET[i,]$N.P <= 3)){
   paste0("Phytochemical.Oxyaromatic.Compound") -> Matching.Phytochemical[i]
  } else {
   paste0("") -> Matching.Phytochemical[i]
  }
 }

 # PROTEIN (1) CONSTRAINTS
 for (i in 1:nrow(DATASET)){
  if((DATASET[i,]$O.C > 0.12) &&
    (DATASET[i,]$O.C <= 0.6) &&
    (DATASET[i,]$H.C > 0.9) &&
    (DATASET[i,]$H.C < 2.5) &&
    (DATASET[i,]$N.C >= 0.126) &&
    (DATASET[i,]$N.C <= 0.7) &&
    (DATASET[i,]$P.C < 0.17) &&
    (DATASET[i,]$N >= 1)){
   paste0("Protein") -> Matching.Protein.1[i]
  } else {
   paste0("") -> Matching.Protein.1[i]
  }
 }

 # PROTEIN (2) CONSTRAINTS
 for (i in 1:nrow(DATASET)){
  if((DATASET[i,]$O.C > 0.6) &&
    (DATASET[i,]$O.C <= 1) &&
    (DATASET[i,]$H.C > 1.2) &&
    (DATASET[i,]$H.C < 2.5) &&
    (DATASET[i,]$N.C > 0.2) &&
    (DATASET[i,]$N.C <= 0.7) &&
    (DATASET[i,]$P.C < 0.17) &&
    (DATASET[i,]$N >= 1)){
   paste0("Protein") -> Matching.Protein.2[i]
  } else {
   paste0("") -> Matching.Protein.2[i]
  }
 }

 # NUCLEOTIDE CONSTRAINTS
 for (i in 1:nrow(DATASET)){
  if((DATASET[i,]$O.C >= 0.5) &&
    (DATASET[i,]$O.C < 1.7) &&
    (DATASET[i,]$H.C > 1) &&
    (DATASET[i,]$H.C < 1.8) &&
    (DATASET[i,]$N.C >= 0.2) &&
    (DATASET[i,]$N.C <= 0.5) &&
    (DATASET[i,]$P.C >= 0.1) &&
    (DATASET[i,]$P.C <= 0.35) &&
    (DATASET[i,]$N.P > 0.6) &&
    (DATASET[i,]$N.P <= 5) &&
    (DATASET[i,]$N >= 2) &&
    (DATASET[i,]$P >= 1) &&
```

```
    (DATASET[i,]$S == 0) &&
    (DATASET[i,]$Mass > 305) &&
    (DATASET[i,]$Mass < 523)){
   paste0("Nucleotide") -> Matching.Nucleotides[i]
  } else {
   paste0("") -> Matching.Nucleotides[i]
  }
 }

 # Concatenate all lists into a single one
 Matchings.pasted.01 <- as.list(paste(Matching.Nucleotides, Matching.Carbohydrates, Matching.Lipids,
 Matching.AminoSugars, Matching.Phytochemical, Matching.Protein.1, Matching.Protein.2))

 # Trim each row of the list (delete "spaces")
 Matchings.pasted.02 <- as.list(gsub(" ", "", Matchings.pasted.01, fixed =TRUE))

 # Add "Not.Matched" to those cells that were not matched to any compound category
 Matchings.pasted.02[Matchings.pasted.02==""] <- "Not.Matched"

 # Mark the potential Double Matches
 # Create a new List
 Matchings.list <- list()

 # Loop on the generated list (double matchings will be marked by "Double.Matched")
 for (i in 1:length(Matchings.pasted.02)){
  if (Matchings.pasted.02[i] == "Lipid"){
   paste0("Lipid") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Carbohydrate"){
   paste0("Carbohydrate") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Amino.Sugar"){
   paste0("Amino.Sugar") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Phytochemical.Oxyaromatic.Compound"){
   paste0("Phytochemical.Oxyaromatic.Compound") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Protein"){
   paste0("Protein") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Nucleotide"){
   paste0("Nucleotide") -> Matchings.list[i]
  } else if (Matchings.pasted.02[i] == "Not.Matched"){
   paste0("Not.Matched") -> Matchings.list[i]
  } else {
   paste0(paste("Double.Match_",Matchings.pasted.02[i])) -> Matchings.list[i]
  }
 }

 Matchings <- as.data.frame(do.call(rbind, Matchings.list))

 Matchings[Matchings == "Double.Match_NucleotideProtein"] <- "Nucleotide" # Double matches with
 nucleotides will be Nucleotides
 Matchings[Matchings == "Double.Match_NucleotideAmino.Sugar"] <- "Nucleotide" # Double matches with
 nucleotdies will be Nucleotides
 DATASET.MATCHED <- DATASET
 DATASET.MATCHED["Compound.Match"] <- Matchings # Add a new column called "Compound.Match" into the
 DATASET.
```

```
410
411  # SAVE DATASET INTO A CSV FILE
412  write.table(data.frame(DATASET.MATCHED), file= Destination.File.Dataset)
413
414
415
416  ## 2nd STEP - CALCULATE THE PROPORTIONS OF EACH COMPOUND CATEGORY ##
417  Protein.Proportion <- length(which(Matchings == "Protein"))/nrow(Matchings)*100
418  Phytochemical.Oxyaromatic.Compound.Proportion <- length(which(Matchings ==
419  "Phytochemical.Oxyaromatic.Compound"))/nrow(Matchings)*100
420  Lipid.Proportion <- length(which(Matchings == "Lipid"))/nrow(Matchings)*100
421  Carbohydrate.Proportion <- length(which(Matchings == "Carbohydrate"))/nrow(Matchings)*100
422  Amino.Sugar.Proportion <- length(which(Matchings == "Amino.Sugar"))/nrow(Matchings)*100
423  Nucleotide.Proportion <- length(which(Matchings == "Nucleotide"))/nrow(Matchings)*100
424  Not.Matched.Proportion <- length(which(Matchings == "Not.Matched"))/nrow(Matchings)*100
425  Double.Matched.Proportion <- length(which(Matchings != "Protein" & Matchings !=
426  "Phytochemical.Oxyaromatic.Compound" & Matchings != "Lipid" & Matchings != "Carbohydrate" & Matchings
427  != "Amino.Sugar" & Matchings != "Nucleotide" & Matchings != "Not.Matched"))/nrow(Matchings)*100 #
428  Including double matches
429
430  # Integrate all the proportions together into a single categorical vector
431  Compound.Proportions <- c(Carbohydrate.Proportion, Amino.Sugar.Proportion, Nucleotide.Proportion,
432  Lipid.Proportion, Protein.Proportion, Phytochemical.Oxyaromatic.Compound.Proportion,
433  Not.Matched.Proportion, Double.Matched.Proportion)
434
435  # Create a Data Frame with the proportions
436  Compound.Proportions.DF <- as.data.frame(Compound.Proportions)
437
438  # Create the Labels for each proportion (has to follow the same order as the integration of the proportions)
439  Labels <- c("Carbohydrates", "Amino.Sugars", "Nucleotides", "Lipids", "Proteins",
440  "Phytochemical.Oxyaromatic.Compounds", "Not.Matched", "Double.Matched")
441
442  # Add a new column into the Data Frame with the name of the compounds
443  Compound.Proportions.DF["Compound"] <- Labels
444
445  # SAVE DATASET INTO A CSV FILE
446  write.table(data.frame(Compound.Proportions.DF), file= Destination.File.Proportions)
447
448
449
450  ## 3rd STEP - PIE CHART OF THE COMPOUND PROPORTIONS ##
451  # Constrain the number of decimals to 2
452  Pie.Proportions <- list()
453  for (i in 1:length(Compound.Proportions.DF$Compound.Proportions)){
454    format(round(Compound.Proportions.DF$Compound.Proportions[i], 2), nsmall=2) -> Pie.Proportions[i]
455  }
456
457  # Create the labels for the Pie Chart
458  Labels.Plot <- paste (Labels, Pie.Proportions) # Add The percentage value to each label.
459  Labels.Plot.2 <- paste(Labels.Plot,"%", sep="") # Add "%" to each label.
460
461  # Plot the Pie Chart
462  pie(Compound.Proportions, labels = Labels.Plot.2, col= rainbow(length(Labels.Plot.2)))
```

463 **Supplementary Tables**

464 **Table S-1.** Compounds included in each of the online examined databases (lipids, amino sugars,
465 phytochemical compounds, carbohydratres and nucleotides) and the corresponding source.

Lipids (Source: LipidMAP)	
Fatty Acyls [FA]	Docosanoids [FA04]
	Eicosanoids [FA03]
	Acyltrehaloses [SL03]
	Fatty Acids and Conjugates [FA01]
	Fatty acyl glycosides [FA13]
	Fatty alcohols [FA05]
	Fatty aldehydes [FA06]
	Fatty amides [FA08]
	Fatty esters [FA07]
	Hydrocarbons [FA11]
	Octadecanoids [FA02]
	Oxygenated hydrocarbons [FA12]
Glycerolipids [GL]	Diradylglycerols [GL02]
	Glycosyldiradylglycerols [GL05]
	Glycosylmonoradylglycerols [GL04]
	Monoradylglycerols [GL01]
	Triradylglycerols [GL03]
Glycerophospholipids [GP]	CDP-Glycerols [GP13]
	Glycerophosphates [GP10]
	Glycerophosphocholines [GP01]
	Glycerophosphoethanolamines [GP02]
	Glycerophosphoinositols [GP06]
	Glycerophosphoglycerols [GP04]
	Glycerophosphoglycerophosphoglycerols [GP12]
	Glycerophosphoinositol bisphosphates [GP08]
	Glycerophosphoinositol monophosphates [GP07]
	Glycerophosphoinositol trisphosphates [GP09]
	Glycerophosphoinositolglycans [GP15]
	Glycosylglycerophospholipids [GP14]
	Glycerophosphoserines [GP03]
	Glyceropyrophosphates [GP11]
	Oxidized glycerophospholipids [GP20]
Prenol Lipids [PR]	Quinones and hydroquinones [PR02]
	Polyprenols [PR03]
Saccharolipids [SL]	Acylaminosugar glycans [SL02]
	Acylaminosugars [SL01]
	Acyltrehaloses [SL03]
	Other acyl sugars [SL05]
Sphingolipids [SP]	Ceramides [SP02]
	Neutral glycosphingolipids [SP05]
	Phosphonosphingolipids [SP04]
	Phosphosphingolipids [SP03]
	Sphingoid bases [SP01]
Sterol Lipids [ST]	Bile acids and derivatives [ST04]

	Secosteroids [ST03]
	Steroid conjugates [ST05]
	Steroids [ST02]
	Sterols [ST01]
Phytochemical Compounds (Sources: LipidMAP and KEGG)	
From LipidMAP	
Polyketides [PK]	Linear polyketides [PK01]
	Halogenated acetogenins [PK02]
	Annonaceae acetogenins [PK03]
	Macrolides and lactone polyketides [PK04]
	Ansamycins and related polyketides [PK05]
	Polyenes [PK06]
	Linear tetracyclines [PK07]
	Polyether antibiotics [PK09]
	Aflatoxins and related substances [PK10]
	Cytochalasins [PK11]
	Flavonoids [PK12]
	Aromatic polyketides [PK13]
Prenol Lipids [PR]	Hopanoids [PR04]
From KEGG	
Flavonoids	Flavonoids
	Isoflavonoids
	Complex flavonoids
	Monolignols
	Lignans
	Coumarins
Skimate / acetate	malonate pathway derived compounds
Polyketides	Anthraquinones
	Pyrones
	Others
Fatty acids related compounds	Fatty acids
Amino acid related compounds	Betalains
	Cyanogenic glucosides
	Glucosinolates
	Others
Others	Naphthoquinones
	Tannins and galloyl derivatives
Amino-Sugars (Sources: KEGG and ChEBI)	
From KEGG	Amino sugars
From ChEBI	15993, 16062, 16173, 16702, 17122, 17274, 17316, 17411, 17446, 17911, 18207, 18232, 21615, 21977, 24108, 25505, 27438, 27459, 27465, 27503, 27625, 28000, 28132, 28207, 28255, 28401, 28761, 28879, 28944, 28945, 28999, 29006, 29025, 29711, 31747, 31748, 32570, 32571, 32572, 35418, 39610, 44230, 46991, 47966, 47968, 47987, 52079, 52426, 57832, 59239, 59277, 59732, 59986, 61033, 61437, 62169, 62325, 63120, 63153, 63287, 64888, 68682, 7125, 7203, 72626,

	72725, 73783, 79970, 79971, 81450, 83930, 84560, 84569, 84941, 85106, 87176, 87177, 87178, 87179, 87180, 87313, 88130, 95151
Carbohydrates (Source: KEGG)	
Monosaccharides	Aldoses
	Ketoses
	Deoxy sugars
	Sugar acids
	Sugar alcohols
Oligosaccharides	Disaccharides
	Tetrasaccharides
Nucleotides (Source: KEGG)	
Nucleotides	Ribonucleotides
	Deoxyribonucleotides
	Cyclic nucleotides

480 **Table S-2.** Proportions of compounds from databases that correctly matched (CM), not matched (NM), incorrectly matched (IM), and double
481 matched (DM) with lipids, protein, amino sugar and carbohydrate categories delimited by our constraints (Table 1 of main text) or the O:C and
482 H:C constraints proposed for other 21 studies. Each of the 21 bibliographical studies is referenced with a different number and the citations are
483 placed as a footnote. The proportions of IM considering DM as incorrect match (IM_{+DM}), the CM without consider the NM and DM ($CM_{-(NM+DM)}$)
484 and the CM/IM_{+DM} and $CM/(IM_{+DM} + NM)$ ratios are also shown. The total proportions and ratios considering all categories together are shown
485 and are based on the absolute number of compounds in databases and on the relative number of compounds.

	Study number (references as footnote)																				
	Present study	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**	**16.**	**17.**	**18.**	**19.**	**20.**	**21.**
Lipids																						
CM (%)	97.05	17.45	53.10	26.35	9.13	58.92	77.98	34.51	33.36	75.00	62.35	25.29	49.21	72.92	24.77	29.26	27.73	66.18	36.98	25.81	23.93	58.30
IM (%)	1.67	1.81	4.95	30.44	39.03	34.58	12.09	25.58	12.10	18.35	17.63	1.23	11.05	3.87	6.44	39.19	16.11	11.74	25.59	8.29	35.52	23.97
NM (%)	1.28	80.74	41.94	34.25	51.83	3.60	9.59	38.21	54.54	6.33	20.02	73.48	39.74	23.21	3.19	31.54	56.16	22.07	37.43	65.90	28.05	17.71
DM (%)	0.00	0.00	0.00	8.96	0.01	2.90	0.34	1.70	0.00	0.33	0.00	0.00	0.00	0.00	65.60	0.00	0.00	0.00	0.00	0.00	12.50	0.01
IM_{+DM} (%)	1.67	1.81	4.95	39.40	39.04	37.48	12.43	27.28	12.10	18.67	17.63	1.23	11.05	3.87	72.04	39.19	16.11	11.74	25.59	8.29	48.02	23.99
$CM_{-(NM+DM)}$ (%)	98.31	90.59	91.47	46.40	18.96	63.02	86.58	57.43	73.38	80.35	77.96	95.35	81.67	94.96	79.36	42.75	63.26	84.93	59.10	75.68	40.25	70.86
CM/IM_{+DM}	58.13	9.63	10.72	0.67	0.23	1.57	6.27	1.27	2.76	4.02	3.54	20.50	4.45	18.85	0.34	0.75	1.72	5.64	1.45	3.11	0.50	2.43
$CM/(IM_{+DM} + NM)$	32.92	0.21	1.13	0.36	0.10	1.43	3.54	0.53	0.50	3.00	1.66	0.34	0.97	2.69	0.33	0.41	0.38	1.96	0.59	0.35	0.31	1.40
Peptides																						
CM (%)	99.89	13.74	17.90	72.44	68.03	65.25	37.53	56.00	43.59	55.69	45.50	17.29	24.70	14.21	32.02	31.39	16.76	19.34	61.13	15.70	66.11	64.49
IM (%)	0.01	0.29	7.24	3.30	2.55	7.14	26.19	2.95	0.82	22.36	10.26	5.36	1.49	30.79	0.97	10.45	1.44	10.71	2.40	4.24	0.63	6.59
NM (%)	0.10	85.98	74.86	23.00	29.34	26.80	35.58	40.57	55.59	20.94	44.24	77.36	73.81	55.00	0.02	58.15	81.79	69.95	36.47	80.06	32.99	28.81
DM (%)	0.00	0.00	0.00	1.25	0.08	0.82	0.70	0.49	0.00	1.01	0.00	0.00	0.00	0.00	66.99	0.00	0.00	0.00	0.00	0.00	0.27	0.11
IM_{+DM} (%)	0.01	0.29	7.24	4.55	2.63	7.95	26.89	3.43	0.82	23.37	10.26	5.36	1.49	30.79	67.96	10.45	1.44	10.71	2.40	4.24	0.90	6.70
$CM_{-(NM+DM)}$ (%)	99.99	97.94	71.20	95.64	96.39	90.14	58.90	95.00	98.16	71.35	81.61	76.35	94.32	31.58	97.05	75.02	92.07	64.37	96.22	78.74	99.05	90.73
CM/IM_{+DM}	7761.83	47.61	2.47	15.91	25.89	8.20	1.40	16.31	53.20	2.38	4.44	3.23	16.62	0.46	0.47	3.00	11.61	1.81	25.42	3.70	73.21	9.62
$CM/(IM_{+DM} + NM)$	904.29	0.16	0.22	2.63	2.13	1.88	0.60	1.27	0.77	1.26	0.83	0.21	0.33	0.17	0.47	0.46	0.20	0.24	1.57	0.19	1.95	1.82
Amino Sugars																						
CM (%)	98.59			9.15	6.34	11.97	4.93	6.34		22.54	4.93							16.20	6.34			4.93
IM (%)	0.00			35.21	42.25	50.00	55.63	0.00		0.00	46.48							13.38	38.73			0.00

NM (%)	1.41			48.59	51.41	38.03	39.44	93.66		77.46	48.59							70.42	54.93			95.07
DM (%)	0.00			7.04	0.00	0.00	0.00	0.00		0.00	0.00							0.00	0.00			0.00
IM_{+DM} (%)	0.00			42.25	42.25	50.00	55.63	0.00		0.00	46.48							13.38	38.73			0.00
$CM_{-(NM+DM)}$ (%)	100.00			20.63	13.04	19.32	8.14	100.00		100.00	9.59							54.76	14.06			100.00
CM/IM_{+DM}	∞			0.22	0.15	0.24	0.09	∞		∞	0.11							1.21	0.16			∞
$CM/(IM_{+DM} + NM)$	70.00			0.10	0.07	0.14	0.05	0.07		0.29	0.05							0.19	0.07			0.05
Carbohydrates																						
CM (%)	98.78	6.10	6.10	82.93	82.93	86.59	93.90		4.88		82.93	28.05	39.02	4.88	97.56	1.22	35.37	34.15	82.93	39.02	37.80	
IM (%)	0.00	0.00	0.00	0.00	1.22	1.22	0.00		0.00		0.00	0.00	0.00	0.00	1.22	0.00	0.00	1.22	0.00	0.00	0.00	
NM (%)	1.22	93.90	93.90	17.07	15.85	12.20	6.10		95.12		17.07	71.95	60.98	95.12	1.22	98.78	64.63	64.63	17.07	60.98	62.20	
DM (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
IM_{+DM} (%)	0.00	0.00	0.00	0.00	1.22	1.22	0.00		0.00		0.00	0.00	0.00	0.00	1.22	0.00	0.00	1.22	0.00	0.00	0.00	
$CM_{-(NM+DM)}$ (%)	100.00	100.00	100.00	100.00	98.55	98.61	100.00		100.00		100.00	100.00	100.00	100.00	98.77	100.00	100.00	96.55	100.00	100.00	100.00	
CM/IM_{+DM}	∞	∞	∞	∞	68.00	71.00	∞		∞		∞	∞	∞	∞	80.00	∞	∞	28.00	∞	∞	∞	
$CM/(IM_{+DM} + NM)$	81.00	0.06	0.06	4.86	4.86	6.45	15.40		0.05		4.86	0.39	0.64	0.05	40.00	0.01	0.55	0.52	4.86	0.64	0.61	
Total Absolute (according to the total absolute number of metabolites) **For those studies with no amino sugar or carbohydrate category; those clases were not considered for computation of totals.**																						
CM (%)	99.19	14.63	26.58	60.98	53.40	63.64	47.54	50.62	40.98	60.43	49.65	19.25	30.75	28.71	30.23	30.81	19.47	30.94	55.11	18.20	55.58	62.89
IM (%)	0.42	0.70	6.68	10.05	11.62	13.97	22.72	8.55	3.63	21.34	12.12	4.36	3.91	24.09	2.40	17.57	5.10	10.96	8.18	5.28	9.31	10.89
NM (%)	0.39	84.67	66.74	25.81	34.92	21.06	29.13	40.04	55.39	17.38	38.23	76.39	65.35	47.20	0.84	51.62	75.43	58.10	36.71	76.52	31.82	26.14
DM (%)	0.00	0.00	0.00	3.16	0.06	1.33	0.61	0.79	0.00	0.84	0.00	0.00	0.00	0.00	66.52	0.00	0.00	0.00	0.00	0.00	3.30	0.08
IM_{+DM} (%)	0.42	0.70	6.68	13.21	11.68	15.30	23.33	9.33	3.63	22.18	12.12	4.36	3.91	24.09	68.93	17.57	5.10	10.96	8.18	5.28	12.60	10.97
$CM_{-(NM+DM)}$ (%)	99.58	95.46	79.92	85.85	82.13	82.00	67.66	85.46	91.86	73.84	80.38	81.54	88.73	54.38	92.64	63.68	79.26	73.84	87.07	77.52	85.66	85.16
CM/IM_{+DM}	234.64	21.01	3.98	4.61	4.57	4.16	2.04	5.42	11.29	2.72	4.10	4.42	7.87	1.19	0.44	1.75	3.82	2.82	6.74	3.45	4.41	5.73
$CM/(IM_{+DM} + NM)$	121.73	0.17	0.36	1.56	1.15	1.75	0.91	1.03	0.69	1.53	0.99	0.24	0.44	0.40	0.43	0.45	0.24	0.45	1.23	0.22	1.25	1.69
Total Relative (giving the same weight to each database independently of the number of metabolites included in each one) **For those studies with no amino sugar or carbohydrate category; those clases were not considered for computation of totals.**																						
CM (%)	98.58	12.43	25.70	47.72	41.61	55.68	53.59	32.28	27.28	51.08	48.93	23.54	37.65	30.67	51.45	20.62	26.62	33.97	46.84	26.84	42.61	42.57
IM (%)	0.42	0.70	4.06	17.24	21.26	23.23	23.48	9.51	4.31	13.57	18.59	2.20	4.18	11.55	2.88	16.55	5.85	9.26	16.68	4.18	12.05	10.19
NM (%)	0.88	86.87	70.23	30.73	37.11	20.15	22.68	57.48	68.42	34.91	32.48	74.26	58.18	57.78	1.48	62.83	67.53	56.77	36.48	68.98	41.08	47.20
DM (%)	0.00	0.00	0.00	4.31	0.02	0.93	0.26	0.73	0.00	0.45	0.00	0.00	0.00	0.00	44.20	0.00	0.00	0.00	0.00	0.00	4.26	0.04
IM_{+DM} (%)	0.42	0.70	4.06	21.55	21.29	24.16	23.74	10.24	4.31	14.01	18.59	2.20	4.18	11.55	47.07	16.55	5.85	9.26	16.68	4.18	16.31	10.23
$CM_{-(NM+DM)}$ (%)	99.57	96.18	87.56	65.67	56.74	67.77	63.40	84.14	90.51	83.90	67.29	90.57	92.00	75.51	91.73	72.59	85.11	75.15	67.35	84.81	79.77	87.20

CM/IM_{+DM}	234.39	17.74	6.32	2.21	1.95	2.30	2.26	3.15	6.33	3.64	2.63	10.72	9.01	2.65	1.09	1.25	4.55	3.67	2.81	6.43	2.61	4.16
$CM/(IM_{+DM} + NM)$	272.05	0.15	0.47	1.99	1.79	2.48	4.90	0.62	0.44	1.52	1.85	0.31	0.65	0.97	13.60	0.29	0.38	0.73	1.77	0.39	0.96	1.09

1. (Kim, Kramer, & Hatcher, 2003)[5]
2. (Mopper, Stubbins, Ritchie, Bialk, & Hatcher, 2007)[6]
3. (Podgorski et al., 2012)[7]
4. (D'Andrilli, Foreman, Marshall, & McKnight, 2013)[8]
5. (Minor, Swenson, Mattson, & Oyler, 2014)[9]
6. (Tfaily et al., 2015)[10]
7. (Schmidt, Elvert, Koch, Witt, & Hinrichs, 2009)[11]
8. (Bhatia, Das, Longnecker, Charette, & Kujawinski, 2010)[12]
9. (Lusk & Toor, 2016)[13]
10. (Xu et al., 2013)[14]
11. (Saenger, Cécillon, Sebag, & Brun, 2013)[15]
12. (Liu, Sleighter, Zhong, & Hatcher, 2011)[16]
13. (Wang, Goual, & Colberg, 2012)[17]
14. (Hockaday, Purcell, Marshall, Baldock, & Hatcher, 2009)[18]
15. (Nebbioso & Piccolo, 2013)[19]
16. (Thevenot et al., 2013)[20]
17. (Grannas, Hockaday, Hatcher, Thompson, & Mosley-Thompson, 2006)[21]
18. (Mann et al., 2015)[22]
19. (Stubbins et al., 2010)[23]
20. (Osborne et al., 2013)[24]
21. (Hodgkins et al., 2014)[25]

Table S-3. Percentage of database compound exact masses that were correctly, incorrectly and not assigned to the corresponding molecular formula by applying compound identification algorithm (CIA)[4]. The absolute number of is shown in brackets. Correctly assigned formulas excluding the not assigned, and the ratios correctly-assigned/incorrectly-assigned and the correctly-assigned/(incorrectly-assigned + not-assigned) are also shown. The total proportions are shown on the calculations based on the absolute number of compounds in databases and on the relative number of compounds.

	Correctly Assigned	Incorrectly Assigned	Not Assigned	Correctly Assigned excluding Not-Assigned	Correctly Assigned / Incorrectly Assigned ratio	Correctly Assigned / (Incorrectly Assigned + Not Assigned) ratio
Lipids	98.08%	1.14%	0.78%	98.08%	86.03	51.08
Peptides	96.6%	3.01%	0.03%	96.6%	32.09	31.78
Non-phosphorilated Peptides	98.39%	1.58%	0.03%	98.39%	62.27	61.11
Phosphopeptides	89.4%	10.59%	0.01%	89.4%	8.44	8.43
Amino sugars	57.25%	34.78%	7.97%	57.25%	1.64	1.34
Carbohydrates	100%	0%	0%	100%	∞	∞
Nucleotides	70.27%	21.62%	8.11%	70.27%	3.25	2.36
Phytochemical compounds	93.67%	5.98%	0.35%	93.67%	15.66	14.8
TOTAL	96.94%	2.84%	0.21%	96.94%	34.13	31.78

Supplementary Figures

Figure S-1. Correlation plots of stoichiometric variables (O:C, H:C, N:C, P:C and N:P) for all compound databases (Amino sugars, yellow; Carbohydrates, orange; Lipids, blue; Nucleotides, cyan; Peptides, red; Phytochemical compounds, green;). Box plots showing the distribution of compounds of each database for each variable are shown, extreme values are shown by dots. Left panels represent the distribution of each of the compounds of each database along the specified stoichiometric variable.

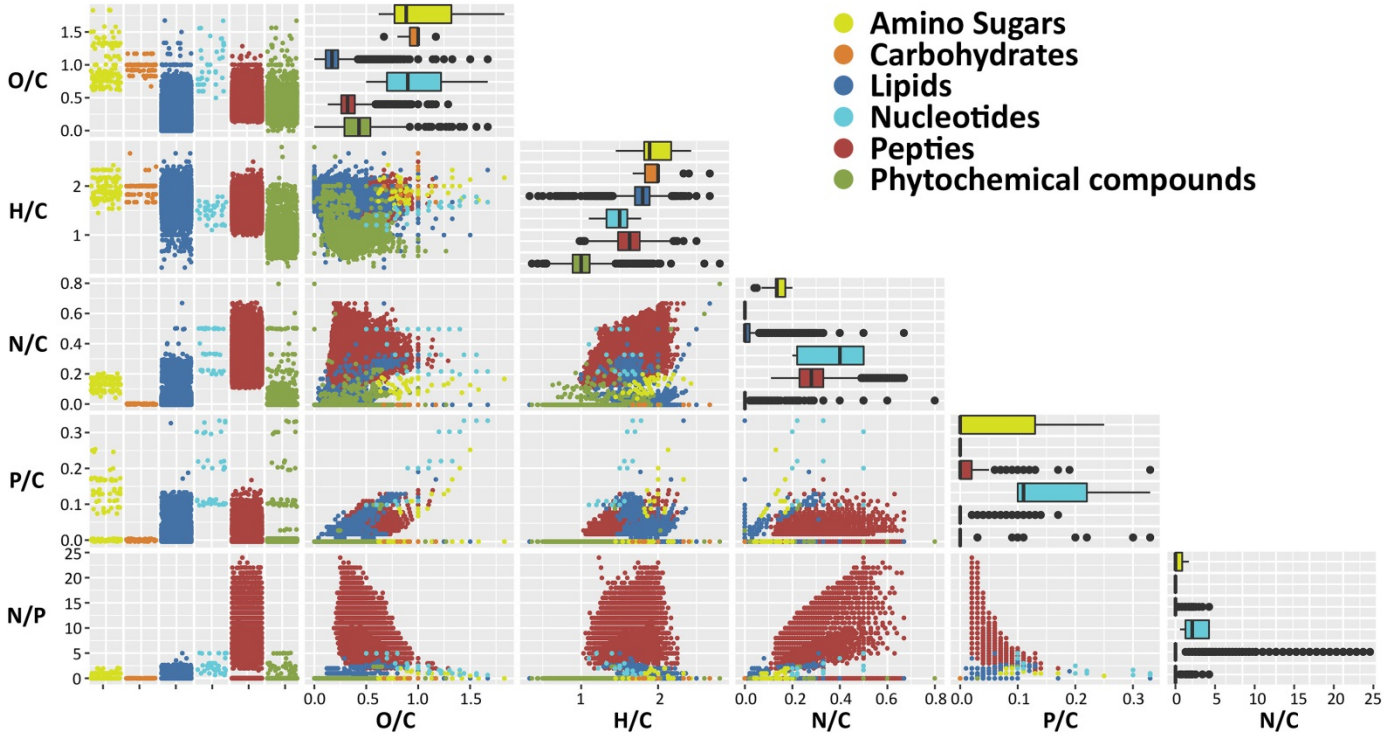


Figure S-2. Figure example showing the implemented criteria to determine the threshold value to separate two categories (Lipids and Phytochemical compounds in this example) that showed overlapping in all stoichiometric variables. The stoichiometric variable that showed better separation between the two compound categories was the one considered to discriminate them; H:C ratio in this case. First, a normal distribution fitting was created for each compound category along the selected variable (**a**). The intersection value between both distribution fittings was considered as a reference threshold (**b**). We created 2000 numbers at 0.0001 step value (threshold candidates) above and below the reference threshold value. Each threshold candidate value determines thus a distribution range for each compound category along the variable (H:C); in this example, the variable range below the candidate value corresponds to phytochemical compounds and above corresponds to lipids. For each of the 4,000 threshold candidate values we calculated the proportion of features of each compound category outside their alleged distribution range. Total overlapping distribution along the 4000 threshold candidates for H:C (**c**). The candidate value that separated the two categories with the minimum proportional number of total overlapped compounds (Lipids + Phytochemical compounds) was considered as the cut-off for those compound categories and variable (H:C): 1.32 in this example.

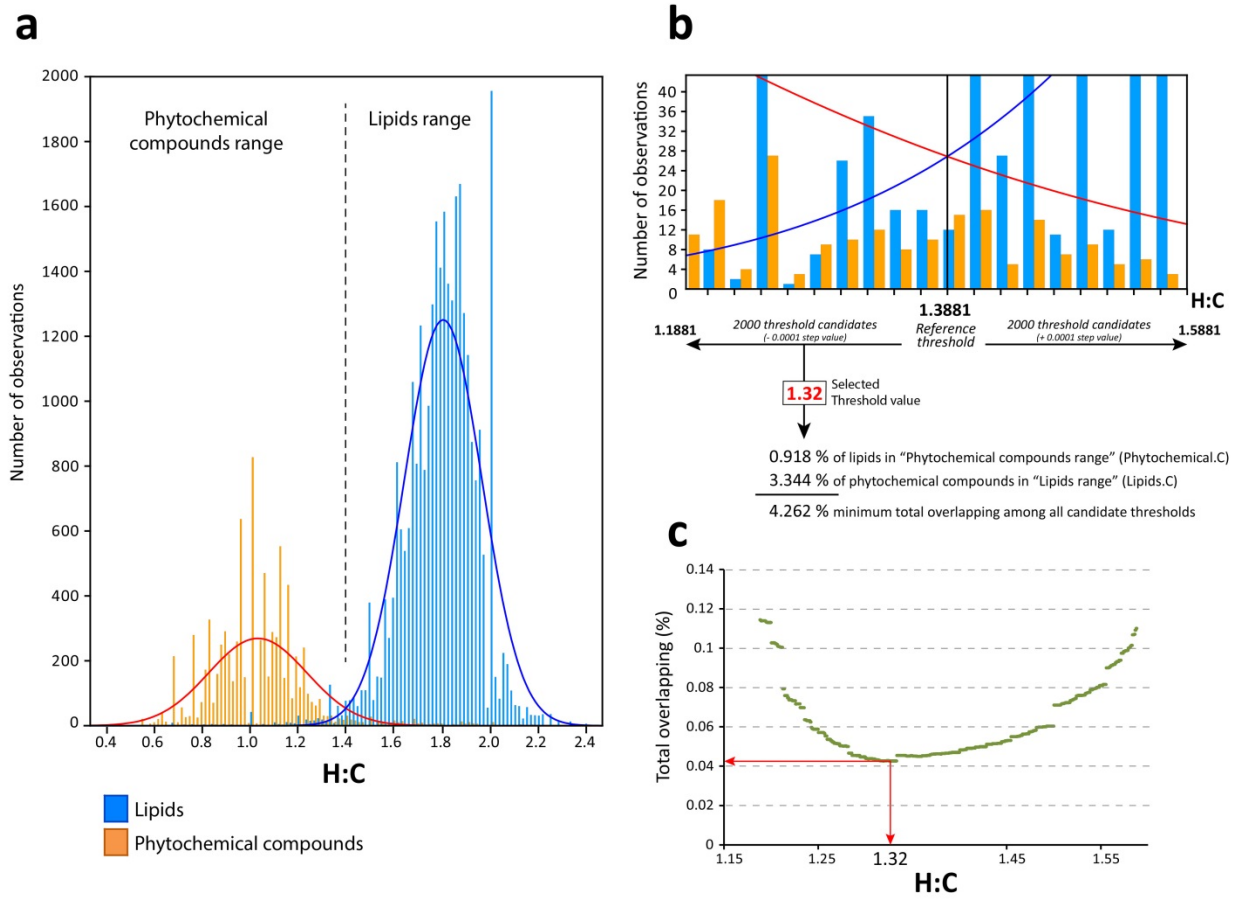
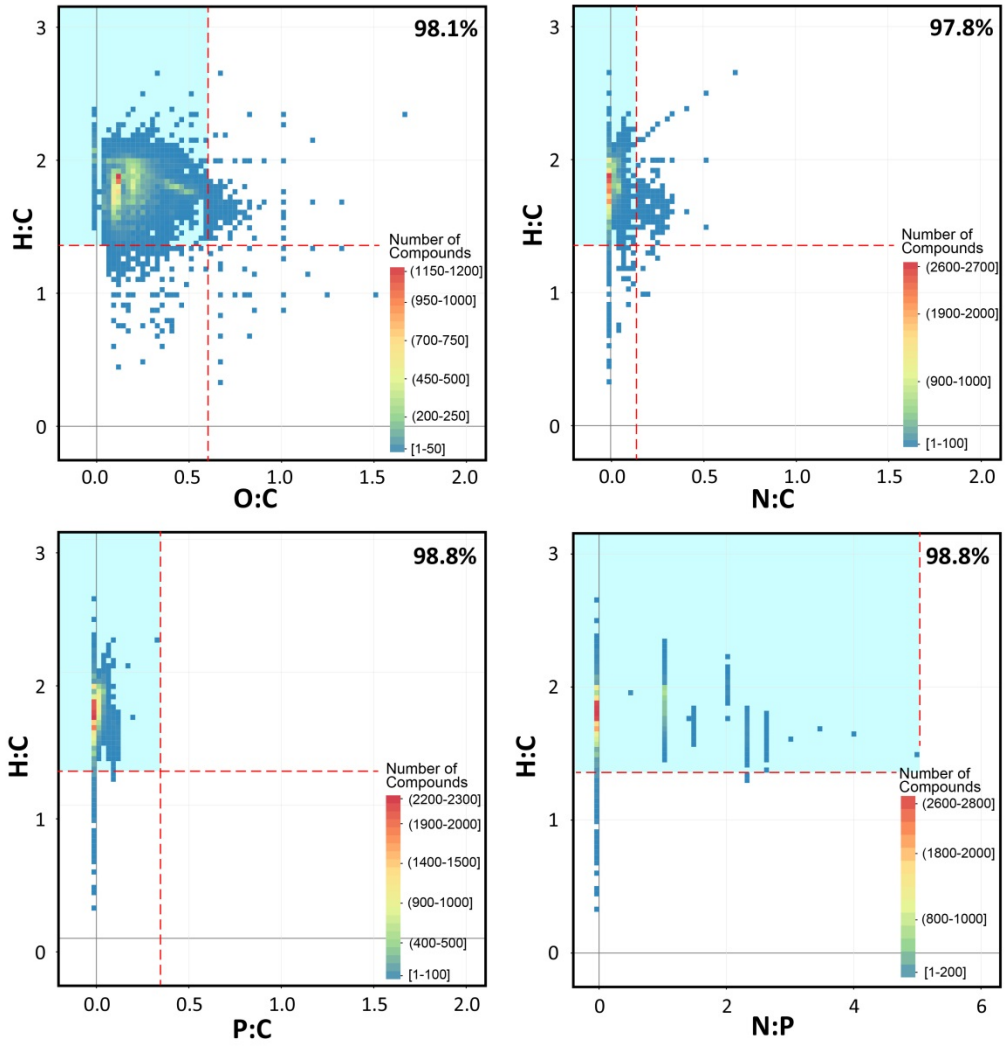
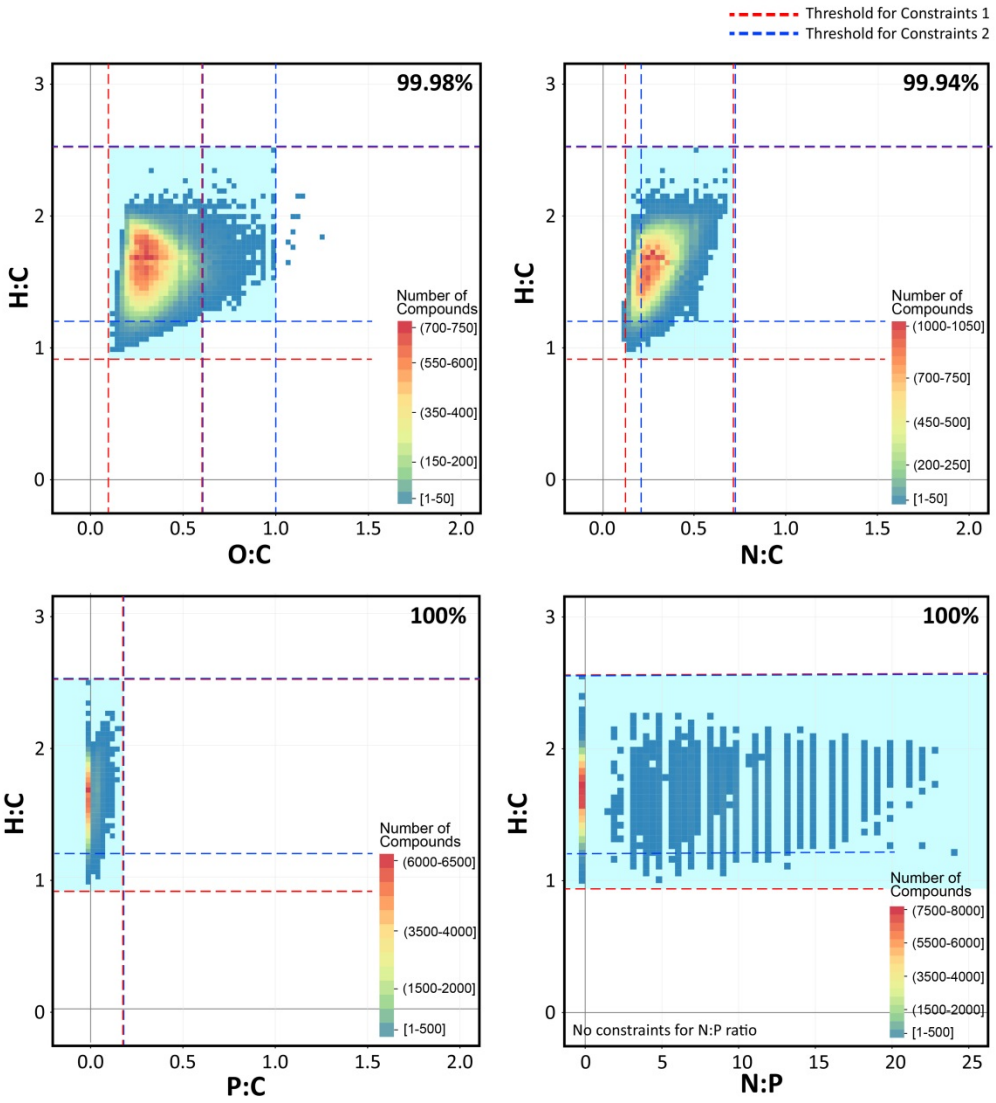


Figure S-3. Bidimensional (2D) density plots of H:C vs. O:C, N:C, P:C, and N:P ratios for lipids database (including 30,729 elemental formulas). Color gradient indicates distinct number of features included in each squared area (red squares indicate the areas with higher density of lipids; blue squares indicate the areas with lower density of lipids). Stoichiometric thresholds for each variable (H:C, O:C, N:C, P:C, and N:P) are represented by red dashed lines (see Table 1 of the main text for exact stoichiometric thresholds). Light-blue area indicates the area included in the stoichiometric constraints. The percentage on the top-right corner of the plots indicate the proportion of compounds within the light-blue area (within the MSCC thresholds).



573 **Figure S-4.** Bidimensional (2D) density plots of H:C vs. O:C, N:C, P:C, and N:P ratios for peptide database
574 (including 93,245 elemental formulas). Color gradient indicates distinct number of features included in
575 each squared area (red squares indicate the areas with higher density of peptides; blue squares indicate
576 the areas with lower density of peptides). Stoichiometric thresholds for each variable (H:C, O:C, N:C, P:C,
577 and N:P) are represented by red dashed lines (constraints 1) and blue dashed lines (constraints 2) (see
578 Table 1 of the main text for exact stoichiometric thresholds). Stoichiometric constraints for Protein
579 category (Protein$_C$) is composed by constraints 1 and constraints 2 together. Light-blue area indicates
580 the area included in the stoichiometric constraints. The percentages on the top-right corner of the plots
581 indicate the proportion of compounds within the light-blue area (within the MSCC thresholds).

582



583

584

585

Figure S-5. Bidimensional (2D) density plots of H:C vs. O:C, N:C, P:C, and N:P ratios for the phytochemical compounds database (including 7,774 elemental formulas). Color gradient indicates distinct number of features included in each squared area (red squares indicate the areas with higher density of phytochemical compounds; blue squares indicate the areas with lower density of phytochemical compounds). Stoichiometric thresholds for each variable (H:C, O:C, N:C, P:C, and N:P) are represented by red dashed lines (see Table 1 of the main text for exact stoichiometric thresholds). Light-blue area indicates the area included in the stoichiometric constraints. The percentages on the top-right corner of the plots indicate the proportion of compounds within the light-blue area (within the MSCC thresholds).

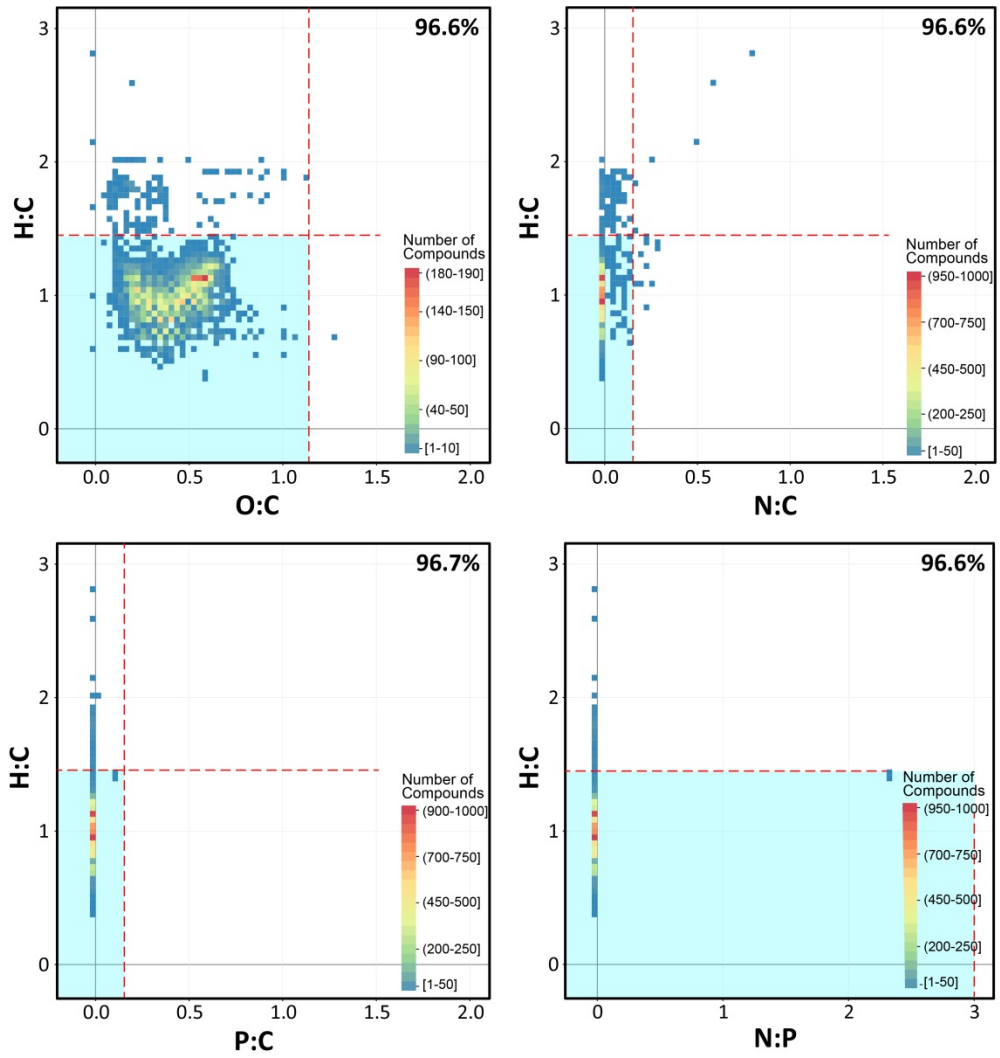


Figure S-6. Bidimensional (2D) density plots of H:C vs. O:C, N:C, P:C, and N:P ratios for amino-sugar database (including 142 elemental formulas). Color gradient indicates distinct number of features included in each squared area (red squares indicate the areas with higher density of amino-sugar; blue squares indicate the areas with lower density of amino-sugar). Stoichiometric thresholds for each variable (H:C, O:C, N:C, P:C, and N:P) are represented by red dashed lines (see Table 1 of the main text for exact stoichiometric thresholds). Light-blue area indicates the area included in the stoichiometric constraints. The percentages on the top-right corner of the plots indicate the proportion of compounds within the light-blue area (within the MSCC thresholds).

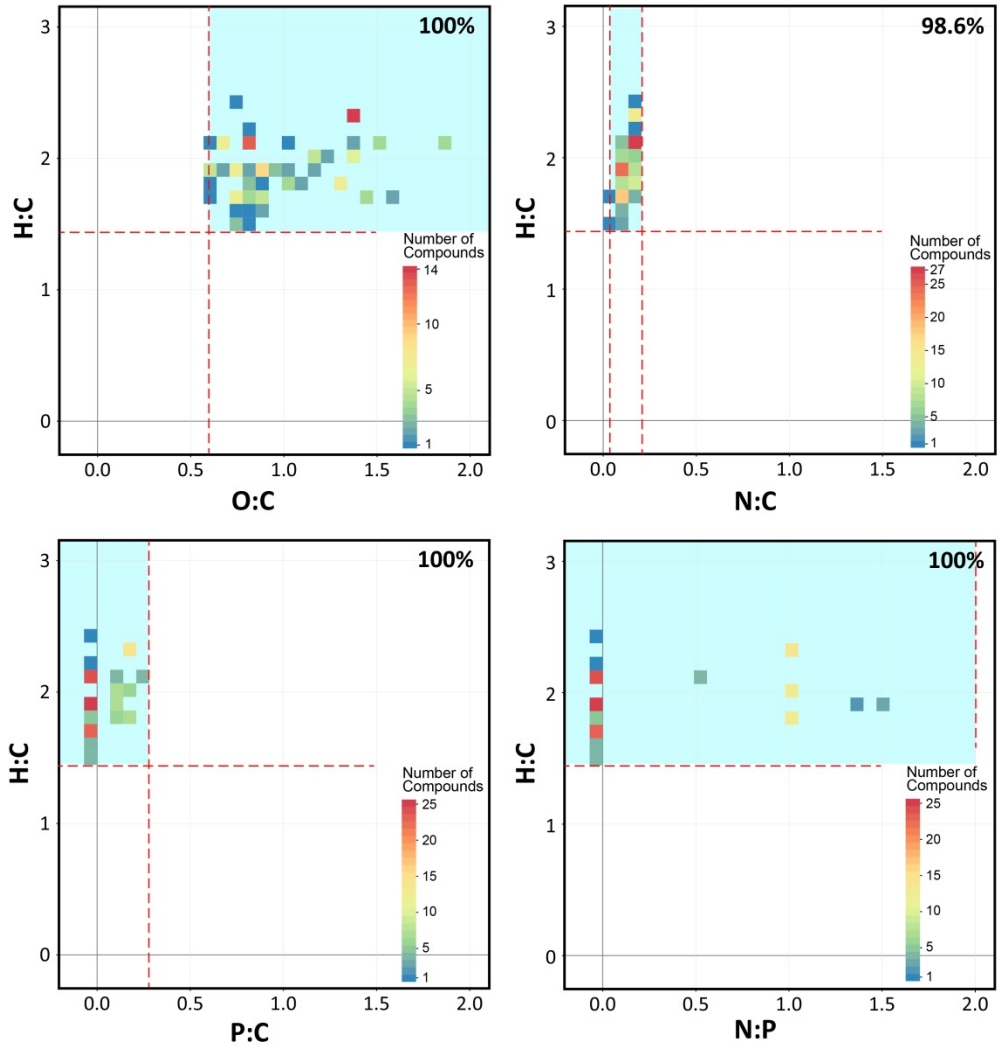


Figure S-7. Bidimensional (2D) density plots of H:C vs. O:C ratio for carbohydrate database (including 82 elemental formulas). Color gradient indicates distinct number of features included in each squared area (red squares indicate the areas with higher density of carbohydrates; blue squares indicate the areas with lower density of carbohydrates). Stoichiometric thresholds for each variable (H:C and O:C) are represented by red dashed lines (see Table 1 of the main text for exact stoichiometric thresholds). Light-blue area indicates the area included in the stoichiometric constraints. The percentage on the top-right corner of the plot indicates the proportion of compounds within the light-blue area (within the MSCC thresholds).

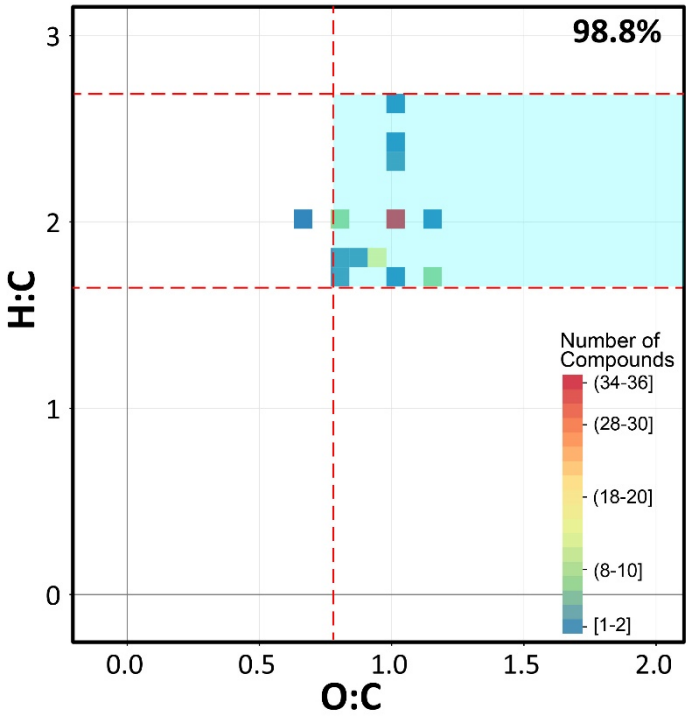
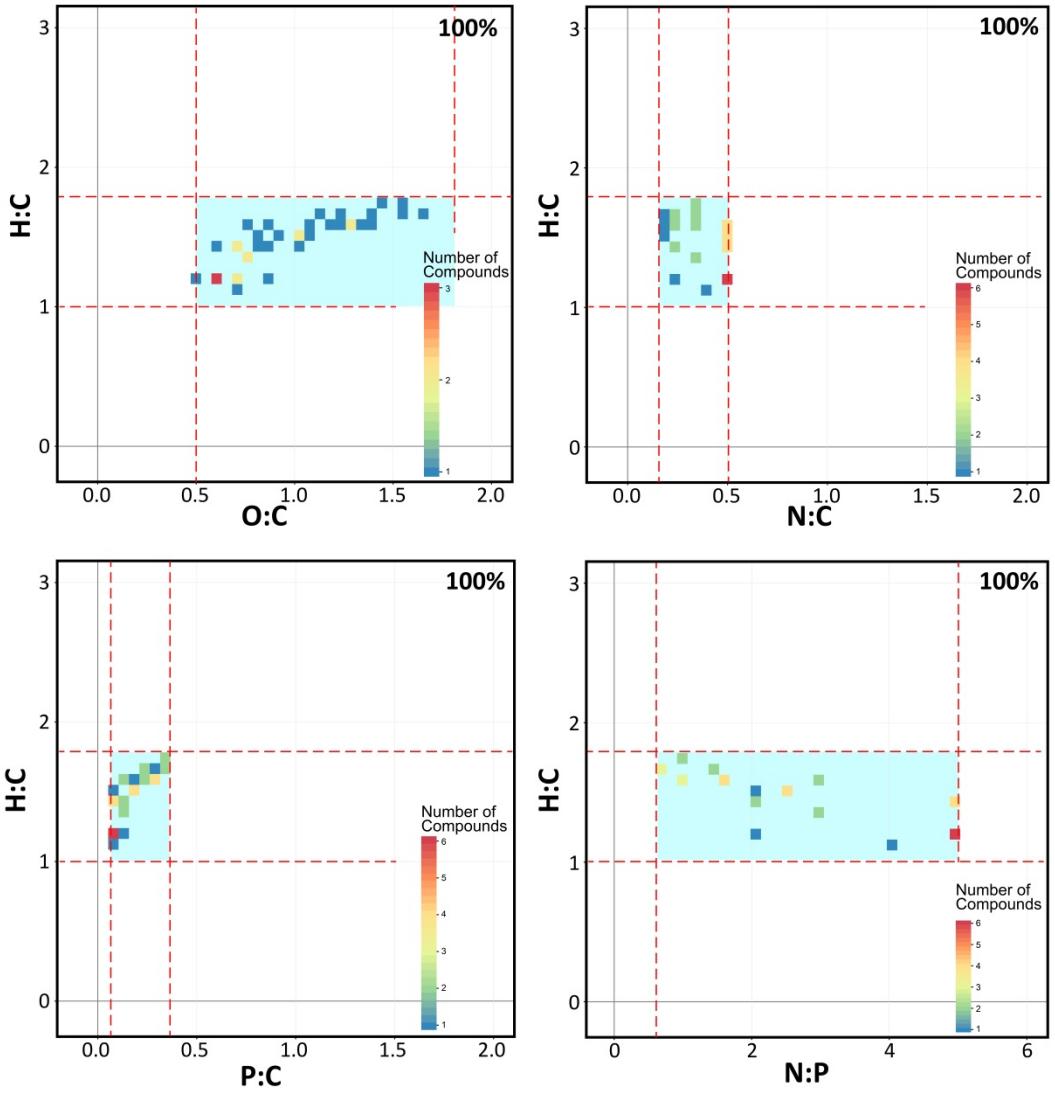
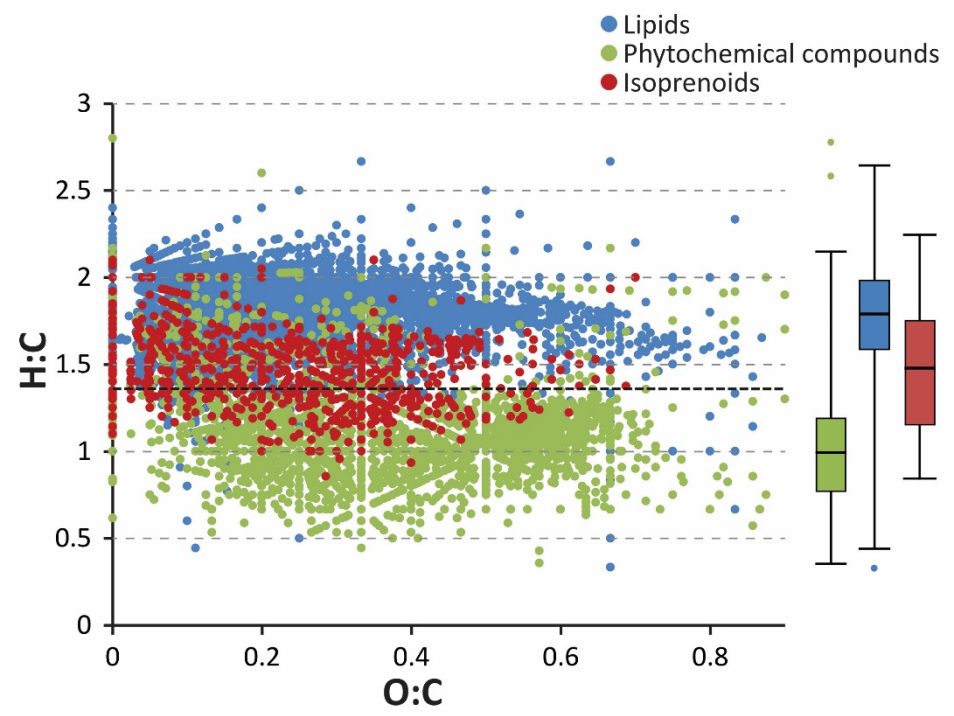


Figure S-8. Bidimensional (2D) density plots of H:C vs. O:C, N:C, P:C, and N:P ratios for nucleotide database (including 37 elemental formulas). Color gradient indicates distinct number of features included in each squared area (red squares indicate the areas with higher density of nucleotides; blue squares indicate the areas with lower density of nucleotides). Stoichiometric thresholds for each variable (H:C, O:C, N:C, P:C, and N:P) are represented by red dashed lines (see Table 1 of the main text for exact stoichiometric thresholds). Light-blue area indicates the area included in the stoichiometric constraints. The percentages on the top-right corner of the plots indicate the proportion of compounds within the light-blue area (within the MSCC thresholds).



642 **Figure S-9**. vK diagram (O:C vs H:C) representing the lipid, phytochemical compound and isoprenoid
643 databases. The threshold value separating lipids and phytochemical compounds along H:C is shown by a
644 dashed black line at H:C =1.32. Box plots for each category compound is shown for H:C variable. First
645 and third percentiles of box plots represent the 10% and 90% of the databases. Dots outside percentiles
646 are considered as outliers.

647



648

649

650

651

652

653

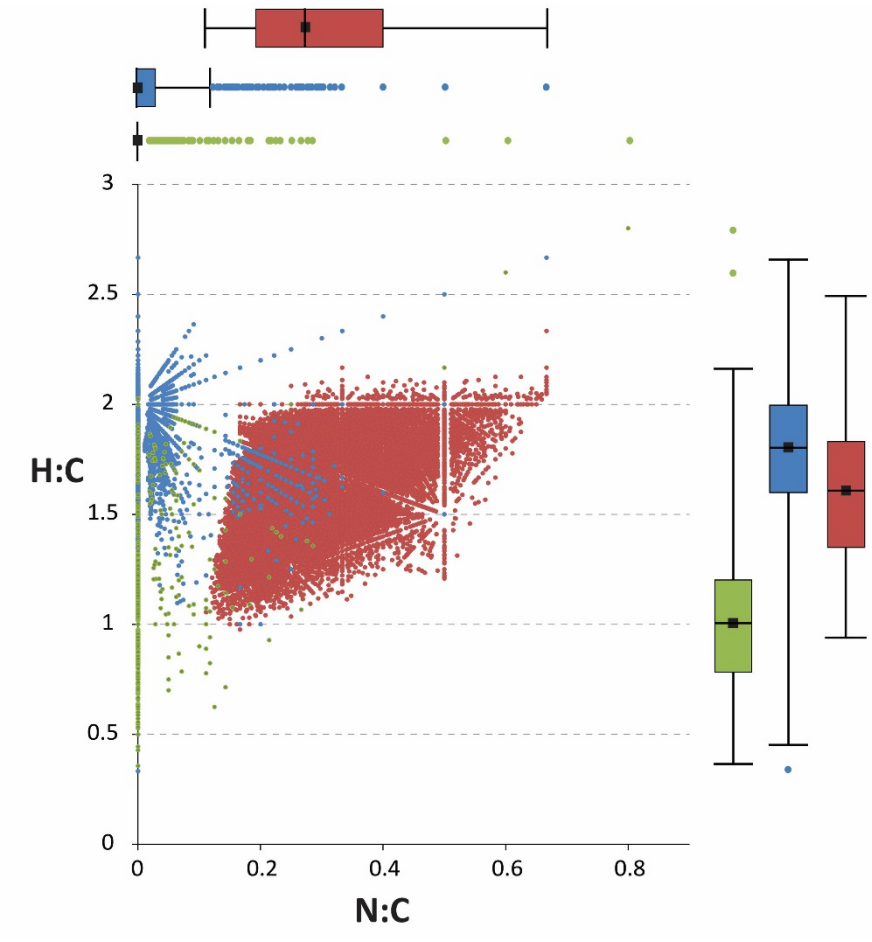
654

655

656

657

Figure S-10. H:C vs. N:C molecular ratios of all peptide (red), lipid (blue) and phytochemical compound (green) databases. Box plots for of each category compound is shown for the stoichiometric variables. First and third percentiles of box plots represent the 10% and 90% of the data. Squares represent the median values and the dots outside the quartiles are outlier compounds for each of the axis. Outliers were determined as the compounds presenting threefold higher values than the third quartile or threefold values lower than the first quartile. In this case we used H:C ratio as the discriminant variable discern lipids from phytochemical compounds while N:C ratio was the discriminant between peptides and the two other categories, especially lipids.



673 **References Supporting Information**

674

675 (1) Fahy, E.; Subramaniam, S.; Murphy, R. C.; Nishijima, M.; Raetz, C. R. H.; Shimizu, T.; Spener, F.;
676 van Meer, G.; Wakelam, M. J. O.; Dennis, E. A. *J. Lipid Res.* **2009**, *50 Suppl* (Supplement), S9-14.

677 (2) Roberts, M. F.; Wink, M. *Alkaloids: Biochemistry, Ecology, and Medicinal Applications*, 1st ed.;
678 Roberts, M. F., Wink, M., Eds.; Springer Science+Business Media, LLC: New Yor, 1998.

679 (3) Cimanga, K.; De Bruyne, T.; Pieters, L.; Claeys, M.; Vlietinck, A. *Tetrahedron Lett.* **1996**, *37* (10),
680 1703–1706.

681 (4) Kujawinski, E. B.; Behn, M. D. *Anal. Chem.* **2006**, *78* (13), 4363–4373.

682 (5) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75* (20), 5336–5344.

683 (6) Mopper, K.; Stubbins, A.; Ritchie, J. D.; Bialk, H. M.; Hatcher, P. G. *Chem. Rev.* **2007**, *107* (2), 419–
684 442.

685 (7) Podgorski, D. C.; Hamdan, R.; McKenna, A. M.; Nyadong, L.; Rodgers, R. P.; Marshall, A. G.;
686 Cooper, W. T. *Anal. Chem.* **2012**, *84* (3), 1281–1287.

687 (8) D'Andrilli, J.; Foreman, C. M.; Marshall, A. G.; McKnight, D. M. *Org. Geochem.* **2013**, *65*, 19–28.

688 (9) Minor, E. C.; Swenson, M. M.; Mattson, B. M.; Oyler, A. R. *Environ. Sci. Process. Impacts* **2014**, *16*
689 (9), 2064–2079.

690 (10) Tfaily, M. M.; Chu, R. K.; Tolić, N.; Roscioli, K. M.; Anderton, C. R.; Paša-Tolić, L.; Robinson, E. W.;
691 Hess, N. J. *Anal. Chem.* **2015**, *87* (10), 5206–5215.

692 (11) Schmidt, F.; Elvert, M.; Koch, B. P.; Witt, M.; Hinrichs, K.-U. *Geochim. Cosmochim. Acta* **2009**, *73*,
693 3337–3358.

694 (12) Bhatia, M. P.; Das, S. B.; Longnecker, K.; Charette, M. A.; Kujawinski, E. B. *Geochim. Cosmochim.*
695 *Acta* **2010**, *74*, 3768–3784.

696 (13) Lusk, M. G.; Toor, G. S. *Water Res.* **2016**, *96*, 225–235.

697 (14) Xu, C.; Chen, H.; Sugiyama, Y.; Zhang, S.; Li, H.-P.; Ho, Y.-F.; Chuang, C.-Y.; Schwehr, K. A.; Kaplan,
698 D. I.; Yeager, C.; Roberts, K. A.; Hatcher, P. G.; Santschi, P. H. *Sci. Total Environ.* **2013**, *449*, 244–
699 252.

700 (15) Saenger, A.; Cécillon, L.; Sebag, D.; Brun, J.-J. *Org. Geochem.* **2013**, *54*, 101–114.

701 (16) Liu, Z.; Sleighter, R. L.; Zhong, J.; Hatcher, P. G. *Estuar. Coast. Shelf Sci.* **2011**, *92*, 205–216.

702 (17) Wang, X.; Goual, L.; Colberg, P. J. S. *J. Hazard. Mater.* **2012**, *217–218* (218), 164–170.

703 (18) Hockaday, W. C.; Purcell, J. M.; Marshall, A. G.; Baldock, J. A.; Hatcher, P. G. *Limnol. Oceanogr.*
704 *Methods* **2009**, *7* (1), 81–95.

705 (19) Nebbioso, A.; Piccolo, A. *Anal. Bioanal. Chem.* **2013**, *405* (1), 109–124.

(20) Thevenot, M.; Dignac, M.-F.; Mendez-Millan, M.; Bahri, H.; Hatté, C.; Bardoux, G.; Rumpel, C. *Biol. Fertil. Soils* **2013**, *49* (5), 517–526.

(21) Grannas, A. M.; Hockaday, W. C.; Hatcher, P. G.; Thompson, L. G.; Mosley-Thompson, E. *J. Geophys. Res.* **2006**, *111* (D4), D04304.

(22) Mann, B. F.; Chen, H.; Herndon, E. M.; Chu, R. K.; Tolic, N.; Portier, E. F.; Roy Chowdhury, T.; Robinson, E. W.; Callister, S. J.; Wullschleger, S. D.; Graham, D. E.; Liang, L.; Gu, B. *PLoS One* **2015**, *10* (6), e0130557.

(23) Stubbins, A.; Spencer, R. G. M.; Chen, H.; Hatcher, P. G.; Mopper, K.; Hernes, P. J.; Mwamba, V. L.; Mangangu, A. M.; Wabakanghanzi, J. N.; Six, J. *Limnol. Oceanogr.* **2010**, *55* (4), 1467–1477.

(24) Osborne, D. M.; Podgorski, D. C.; Bronk, D. A.; Roberts, Q.; Sipler, R. E.; Austin, D.; Bays, J. S.; Cooper, W. T. *Rapid Commun. Mass Spectrom.* **2013**, *27* (8), 851–858.

(25) Hodgkins, S. B.; Tfaily, M. M.; McCalley, C. K.; Logan, T. A.; Crill, P. M.; Saleska, S. R.; Rich, V. I.; Chanton, J. P. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (16), 5819–5824.