

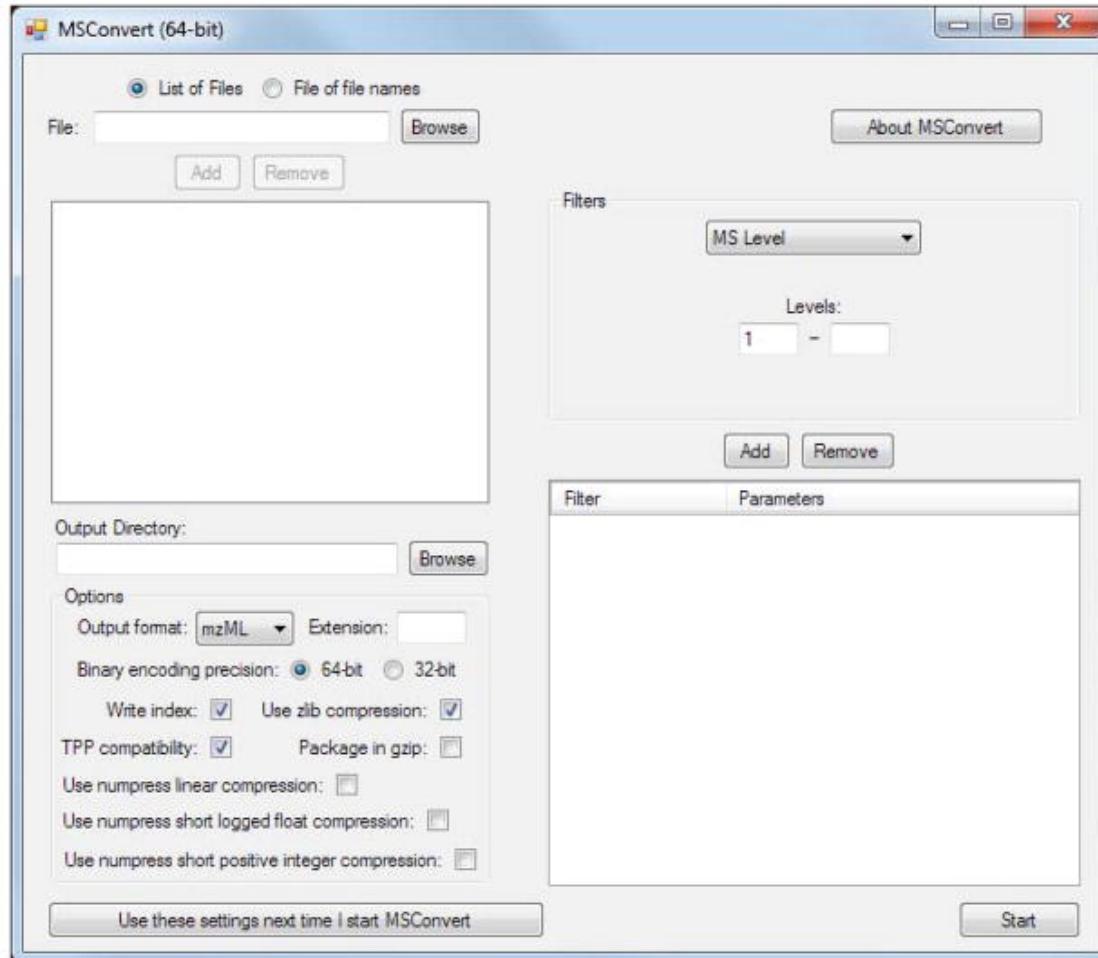
# Data transformation, Data analysis, and Metabolite identification

Mar Garcia-Aloy

# Data Transformation

# Data conversion

## ProteoWizard: MSConvert



Source: Holman JD, Tabb DL, Mallick P. Curr Protoc Bioinformatics. 2014;46:13.24.1-9.

## Supported data formats:

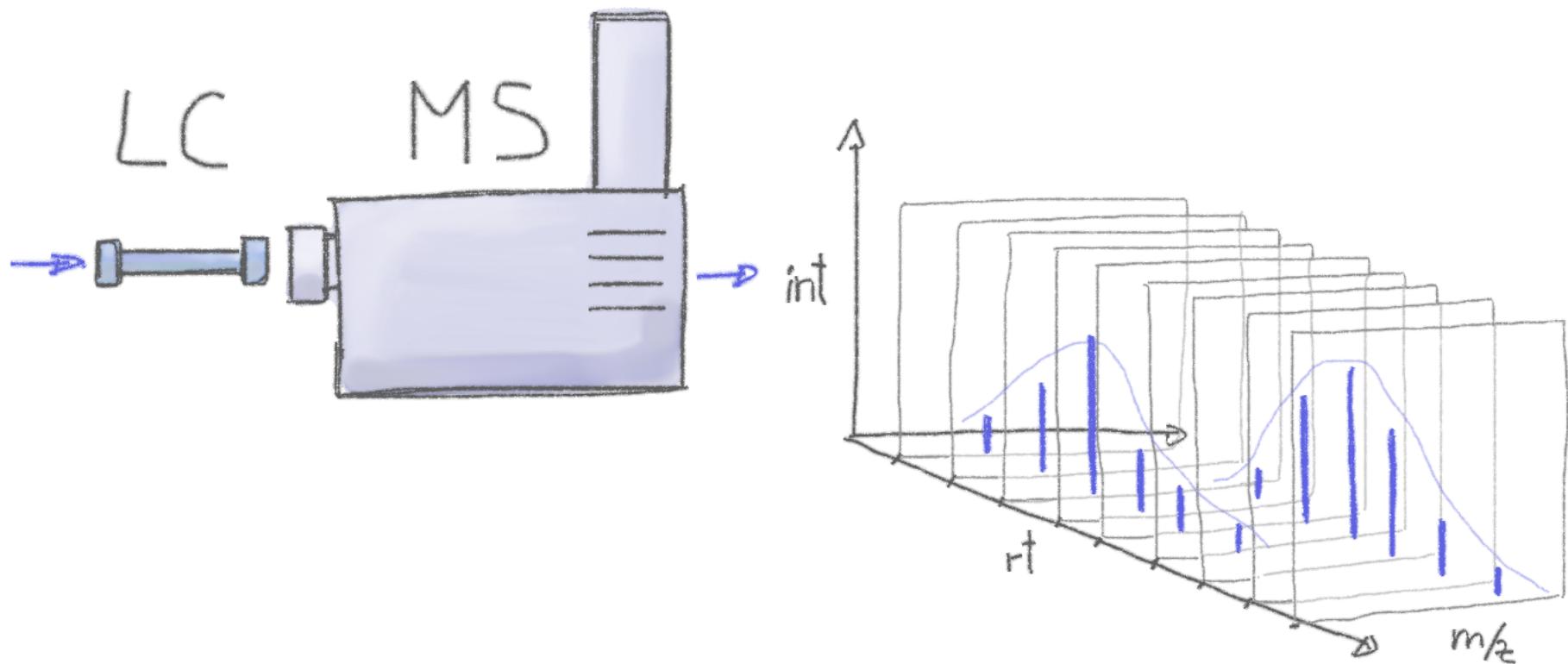
---

Vendor	Format
ABSciex	WIFF; T2D (with DataExplorer)
Agilent	MassHunter (.d directories)
Bruker	FID; .d directories; XMASS XML
Thermo	RAW
Waters	raw directories
<hr/>	
HUPO PSI	mzML
ISB Seattle Proteome Center	mzXML
Matrix Science	MGF
Yates/MacCoss Laboratories	MS2/CMs2/BMS2
Steen & Steen Laboratory	mz5

---

Source: Holman JD, Tabb DL, Mallick P. Curr Protoc Bioinformatics. 2014;46:13.24.1-9.

# Data processing



Source: Rainer J: Metabolomics data pre-processing using xcms (<https://jorainer.github.io/metabolomics2018/xcms-preprocessing.html>).

# Data processing workflow

1. **Peak Picking:** Identify detected peaks in each sample
2. **Retention time correction:** Adjust the shifts in RT between measurement runs from samples within an experiment
3. **Peak Grouping:** Match the peak lists across the samples to find a consensus list of features
4. **Get the data matrix:** Select a measure of signal abundance (peak area, height...) and arrange everything in a data matrix

**Provide reliable ways to visualize all the previous steps and check their consistency**

Source: Franceschi P. Creating a data matrix in xcms from raw LC-MS data ([https://github.com/pietrofranceschi/Metabolomics\\_lectures](https://github.com/pietrofranceschi/Metabolomics_lectures))

# Data processing softwares

There are several software solutions available for the processing of metabolomics data:

1. Commercial suites
2. Open source solutions which can be run locally (e.g. **xcms**, **MzMine**, **MSDial**)
3. Freely available web services



Source: Franceschi P. Software Resources for preprocessing ([https://github.com/pietrofranceschi/Metabolomics\\_lectures](https://github.com/pietrofranceschi/Metabolomics_lectures))

# Data treatment

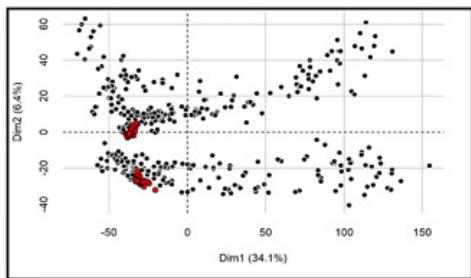
- Data transformation (*e.g. log or sqrt*)
- Variable scaling (*e.g. Pareto scalig or unit variance*)
- Sample normalization

# Data Analysis

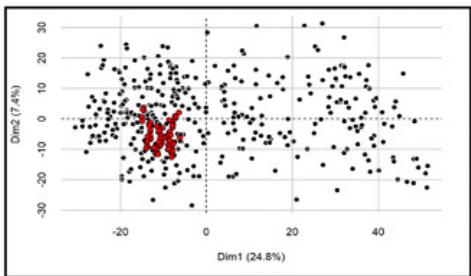
# Exploratory analysis

## VERIFICATION OF QUALITY CONTROL SAMPLES AND OUTLIERS WITH PCA

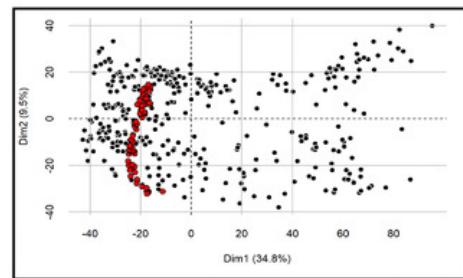
**Example of Batch Effect** – samples divided into two clouds due to a prompt problem during injections (RAW, not normalized data)



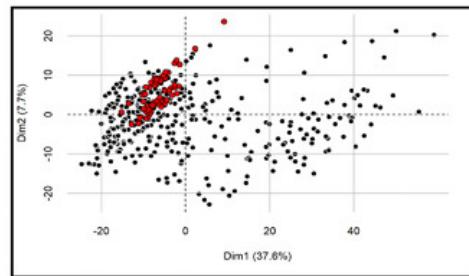
The same sample set normalized by:  
1) median of each feature/plate = 1  
2) Intensity of creatinine  $m/z$  feature



**Example of Drift due to loss of signal intensity** – Samples and QCs suffered from drop in the signal intensity during injections (RAW, not normalized data)



The same sample set normalized by:  
1) median of each feature/plate = 1  
2) Intensity of creatinine  $m/z$  feature



QC samples



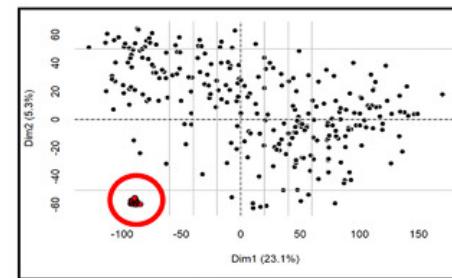
Study samples



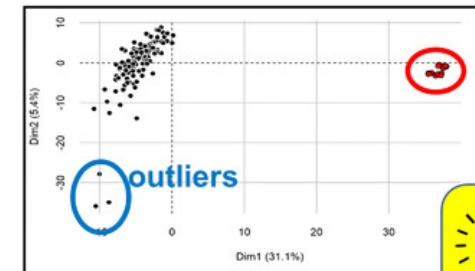
Outliers samples

**Example of QC prepared from study samples**

– a tight QCs cloud is located nearby a wide cloud of real samples (RAW, not normalized data)

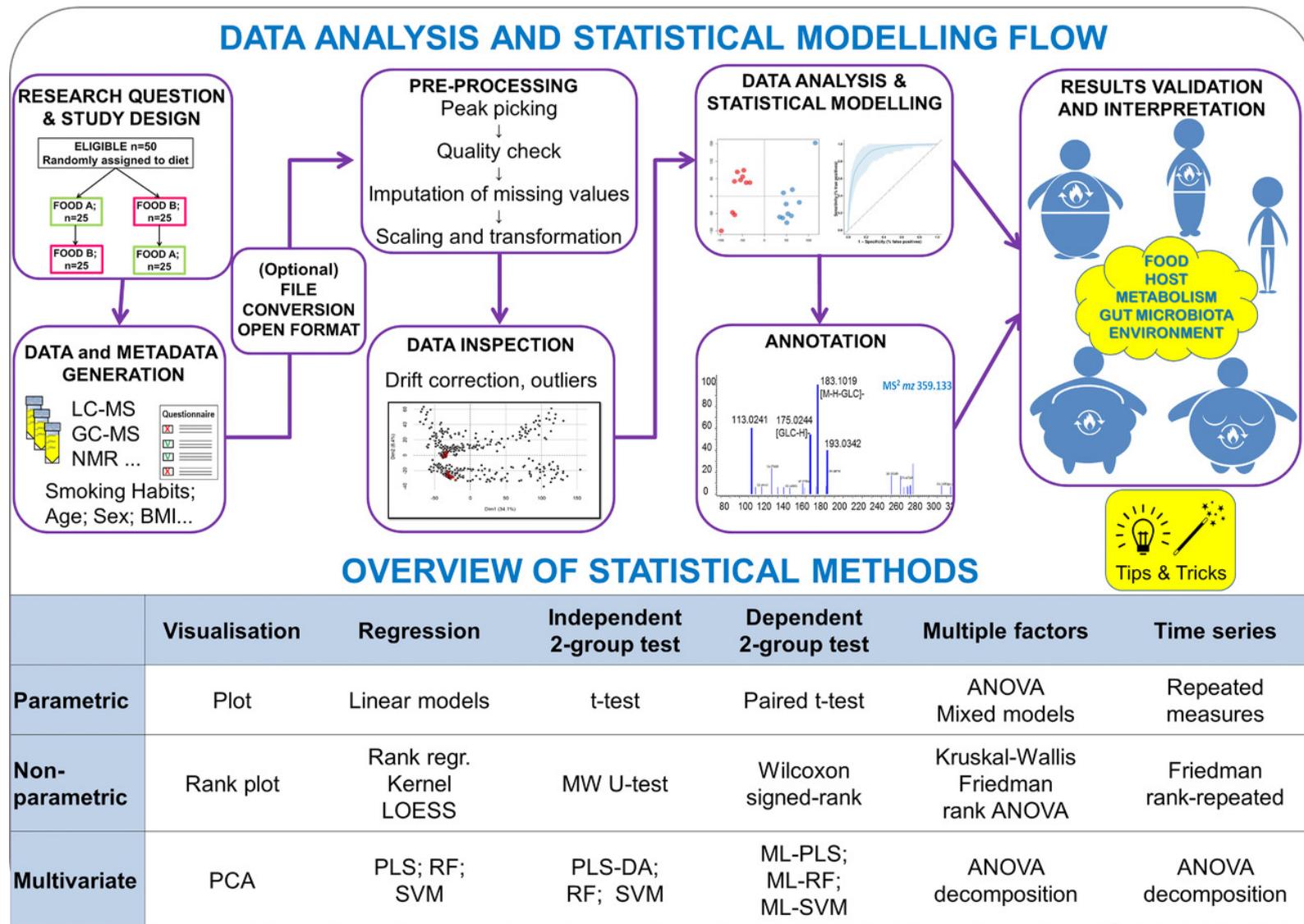


**Example of QC prepared from commercial biological fluid (i.e. plasma)** – a tight QCs cloud is located far from a wider cloud of real samples (RAW, not normalized data). Three samples (outliers) are separated from the rest of samples (blue circle)



Tips & Tricks

# Biomarker discovery



Source: Ułaszczyk et al. Mol Nutr Food Res. 2019;63(1):e1800384

**Univariate** analyses consider each variable separately and it applies “standard” statistical tools to spot the more *interesting* variables:

1. t-Tests, ANOVA
  2. Linear modeling (`lm`, `glm`, ...)
- 

## PRO

1. Statistical modeling
2. Simple output interpretation

## CONS

1. Multiple testing
2. Redundancy of data structure

Source: Franceschi P. Data analysis in metabolomics ([https://github.com/pietrofranceschi/Metabolomics\\_lectures](https://github.com/pietrofranceschi/Metabolomics_lectures))

**Multivariate** approaches consider more than 1 variable simultaneously and able to exploit the correlation between variables:

1. PCA
  2. PLS-DA
- 

## PRO

1. Potentially more powerful
2. Use of variable correlation

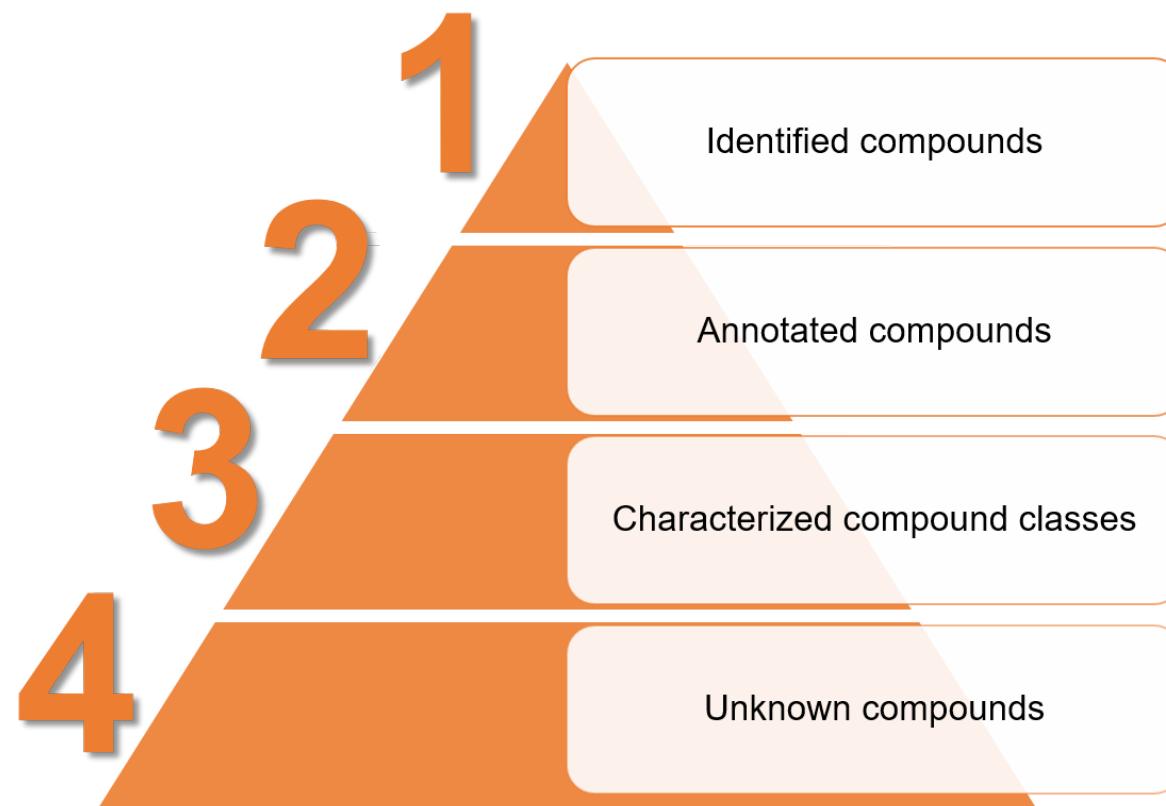
## CONS

1. Interpretation is more complex
2. Supervised methods: risk of getting overfitted models

Source: Franceschi P. Data analysis in metabolomics ([https://github.com/pietrofranceschi/Metabolomics\\_lectures](https://github.com/pietrofranceschi/Metabolomics_lectures))

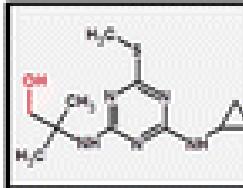
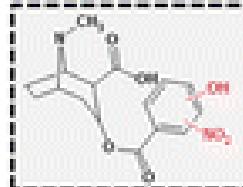
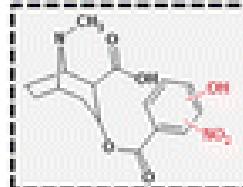
# Metabolite Identification

# 4 levels of annotation confidence



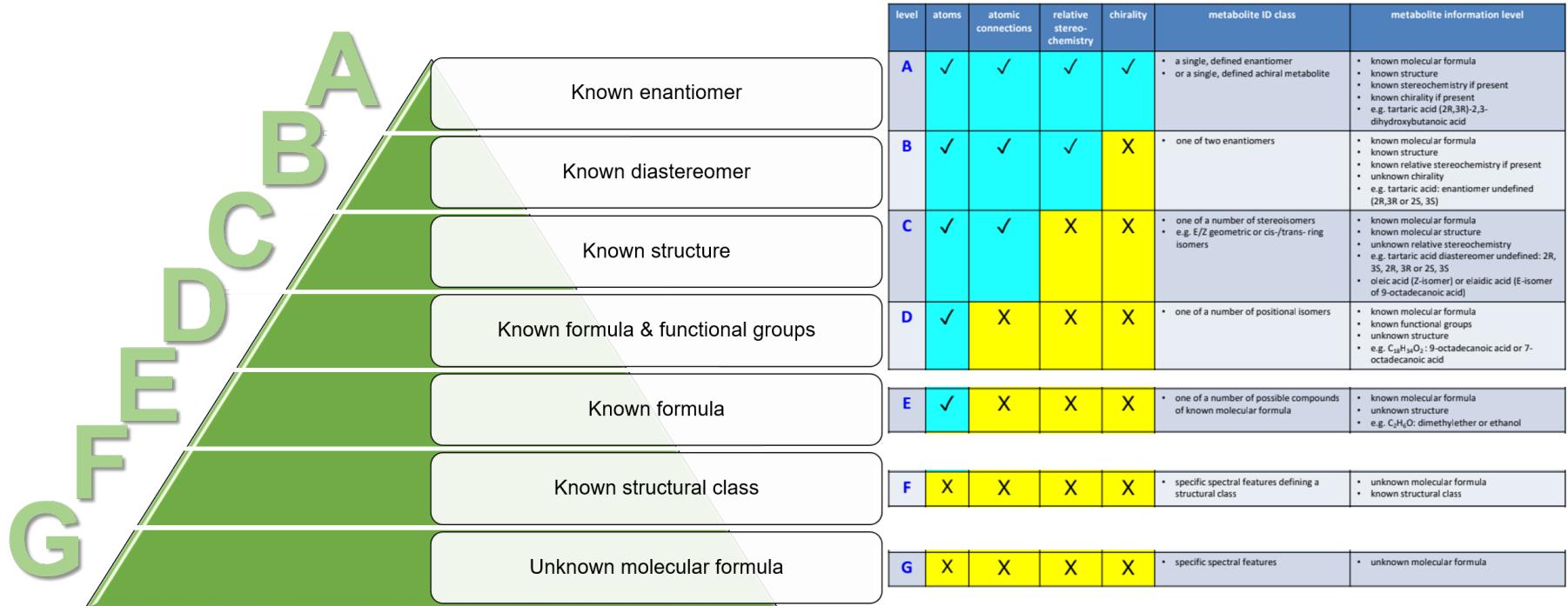
Source: Sumner et al. *Metabolomics*. 2007;3:211–21

# 5 levels of annotation confidence

Example	Identification confidence	Minimum data requirements
	<b>Level 1: Confirmed structure</b> by reference standard	MS, MS <sup>2</sup> , RT, Reference Std.
	<b>Level 2: Probable structure</b> a) by library spectrum match b) by diagnostic evidence	MS, MS <sup>2</sup> , Library MS <sup>2</sup> MS, MS <sup>2</sup> , Exp. data
	<b>Level 3: Tentative candidate(s)</b> structure, substituent, class	MS, MS <sup>2</sup> , Exp. data
$C_8H_5N_3O_4$	<b>Level 4: Unequivocal molecular formula</b>	MS isotope/adduct
192.0757	<b>Level 5: Exact mass of interest</b>	MS

Source: Schymanski et al. Environ Sci Technol. 2014;48(4):2097–8

# Metabolomics Society's Metabolite Identification Task Group



<https://drive.google.com/file/d/1PJLdPCkz8ymX8SgZ4WI5Sw4ZG-dlyWWU/view>

# Metabolite ID



**Supplementary Table S10.** Tentative markers for the ESI+ mode

Measured m/z	RT (min)	Annotation level <sup>g</sup>	Annotation <sup>1-8</sup>	HMDB ID	CHEBI ID	ChemSpider ID
205.098	8.61	1	tryptophan	HMDB0030396	CHEBI:27897	1116
214.090	11.11	1	1,2,3,4-Tetrahydroharmane-3-carboxylic acid	HMDB32102		133583
465.103	11.55	1	delphinidin 3-glucoside	HMDB37997	CHEBI:31463	391783
449.108	13.01	1	cyanidin 3-glucoside	HMDB0030684	CHEBI:28426	390284
479.118	13.66	1	petunidin 3-glucoside	HMDB38097	CHEBI:31985	391784
463.124	14.85	1	peonidin 3-glucoside	HMDB0013689	CHEBI:74793	391786
493.134	15.04	1	malvidin 3-glucoside	HMDB30777	CHEBI:31799	391785
441.084	15.49	1	epigallocatechin gallate	HMDB0003153	CHEBI:4806	58875
507.113	17.18	1	delphinidin 3-(6"-acetyl)-glucoside	HMDB38004	CHEBI:75678	30777226
517.138	18.16	1	pyrano malvidin glucoside			58191428
491.118	18.56	1	cyanidin 3-(6"-acetyl)-glucoside	HMDB0037971	CHEBI:131449	30780060
521.129	18.69	1	petunidin 3-(6"-acetyl)-glucoside			30779241
158.098	18.76	4	(correlation with 561,13 and 399,07)			
535.145	19.71	1	malvidin 3-(6"-acetyl)-glucoside	HMDB38008	CHEBI:75689	30779236
505.134	19.77	1	peonidin 3-(6"-acetyl)-glucoside		CHEBI:75697	30779239
655.166	20.08	1	malvidin caffeoyl glucoside	HMDB30099	CHEBI:75677	26559505
595.145	20.25	1	cyanidin 3-(6"-p-coumaroyl)-glucoside	HMDB37982	CHEBI:29560	4445294
625.155	20.25	1	petunidin 3-(6"-p-coumaroyl)-glucoside		CHEBI:75709	4445294
639.171	20.37	3	malvidin 3-(6"-p-coumaroyl)-glucoside	HMDB0038012	CHEBI:75693	30779237
609.160	20.40	3	peonidin 3-(6"-p-coumaroyl)-glucoside		CHEBI:75707	30779238
303.053	20.80	1	quercetin	HMDB0005794	CHEBI:16243	15139441
287.057	21.18	1	kaempferol	HMDB0005801	CHEBI:28499	4444395
170.098	21.31	4				

Table adapted from Ontañón et al. J Agric Food Chem. 2020;68(47):13367-79

**Thank you for your attention!**