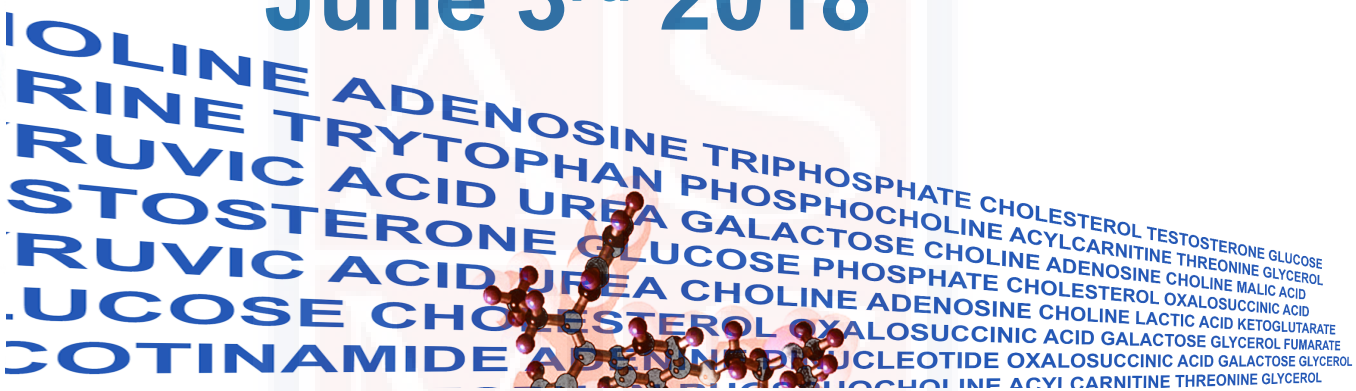
Advanced Metabolomics

June 3rd 2018

Advanced Metabolomics

June 3rd 2018

Gary Siuzdak



H. Paul Benton

Xavi Domingo

Erica Forsberg

Rafa Montenegro

Carlos Guijas

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

Fundamental Metabolomics

Pre-analytical

Sample amount

- biofluids (20-100µL)
- cell cultures (~ 1.10^6 cells)
- tissues (~ 10mg fresh weight)

Metabolism quenching

- snap freezing (liquid N_2)
- heat fixation

Sample storage

- freezing -80°C

Analytical

Metabolite extraction

- polar : MeOH : H_2O (4:1)
- lipid: CH_2Cl_2 : MeOH (1:1)
- option: SPE

Sample normalisation

- ~ creatinine
- ~ protein / DNA content
- ~ dry weight

Data acquisition

- targeted LC-MS/MS
- untargeted LC-MS/MS

Experimental Design

Fundamental Metabolomics Advanced

Pre-analytical

Sample amount

- biofluids (20-100µL)
- cell cultures (~ 1.10^6 cells)
- tissues (~ 10mg fresh weight)

Metabolism quenching

- snap freezing (liquid N_2)
- heat fixation

Sample storage

- freezing -80°C

Analytical

Metabolite extraction

- polar : MeOH : H_2O (4:1)
- lipid: CH_2Cl_2 : MeOH (1:1)
- option: SPE

Sample normalisation

- ~ creatinine
- ~ protein / DNA content
- ~ dry weight

Data acquisition

- targeted LC-MS/MS
- untargeted LC-MS/MS

Post-analytical

Raw data processing

- peak detection
- RT correction
- peak (re)grouping

XCMS Server

Data curation

- noise filtering
- metabolite ID

MetaboTools

METLIN
HMDB
MetaboData
...

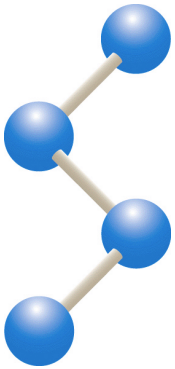
Statistical analyses

- multivariate
- univariate

SIMCA

R

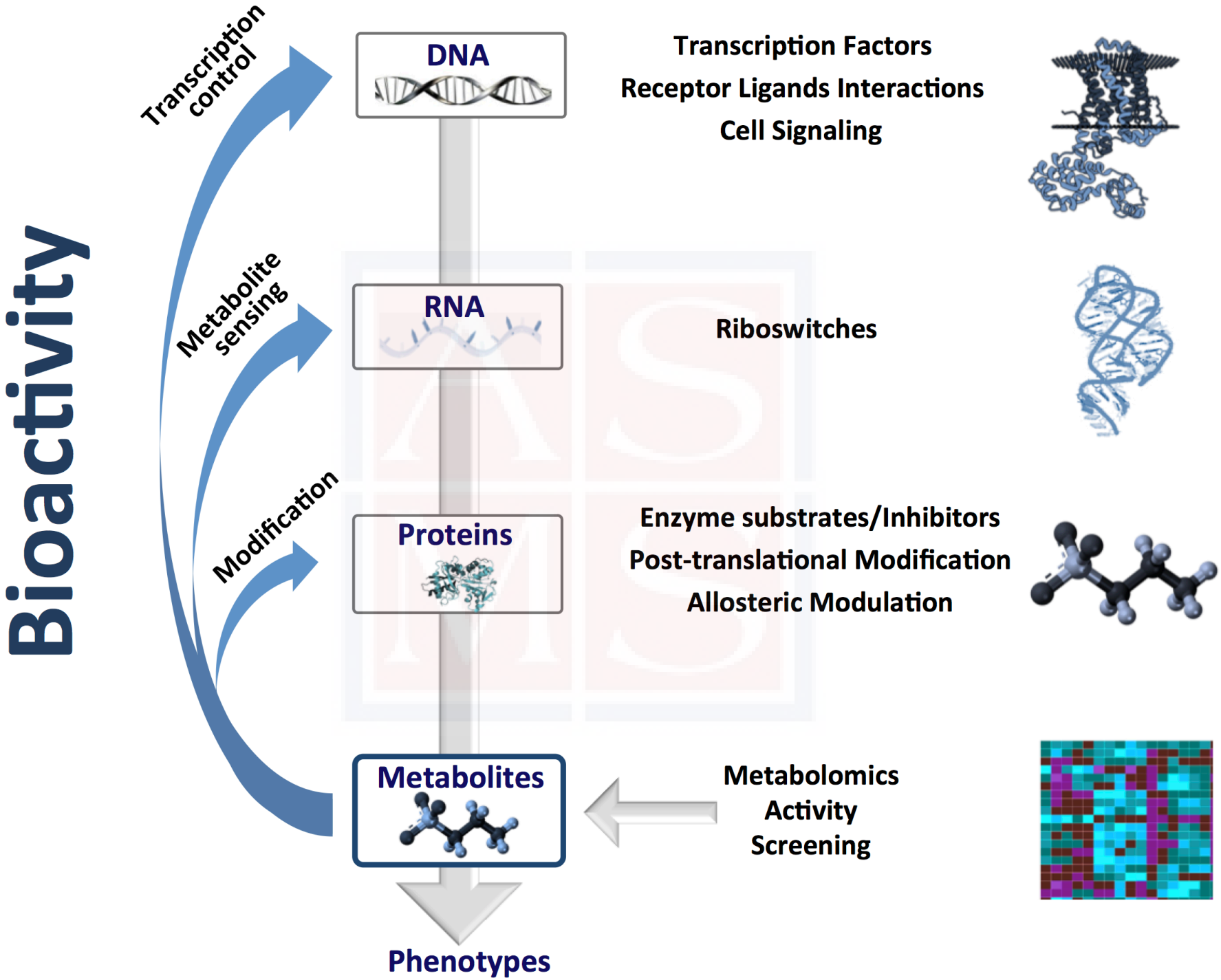
Experimental Design Biology



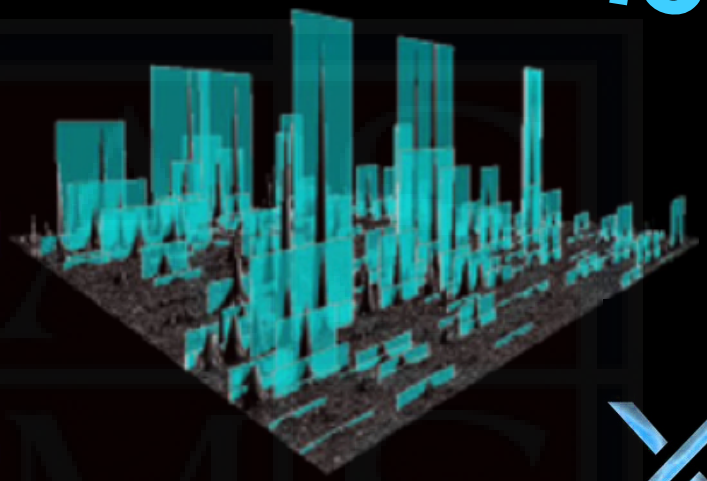
Fundamental Metabolomics

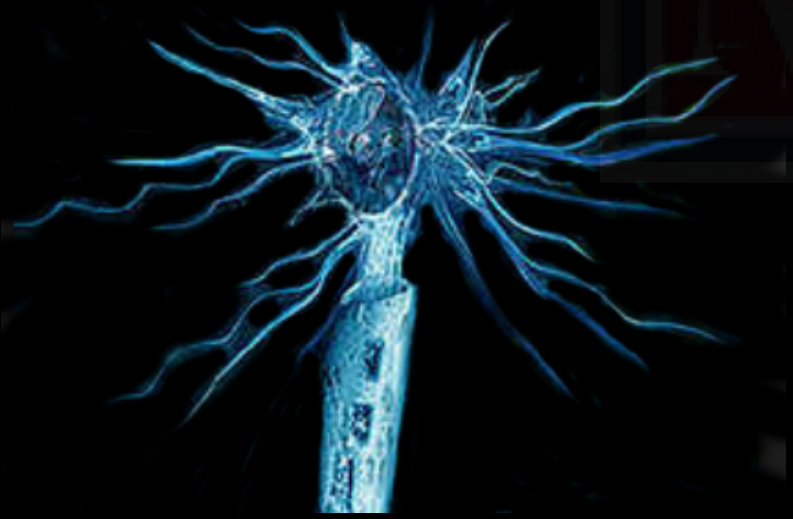
Genomics is a discipline in which the complete set of **DNA** within cells or an organism is analyzed.

Metabolomics is a discipline in which the complete set of **metabolites** within cells or an organism is analyzed.


Bioactivity
Transcription control
Metabolite sensing
Modification
DNA
RNA
Proteins
Metabolites
Phenotypes
Transcription Factors
Receptor Ligands Interactions
Cell Signaling
Riboswitches
Enzyme substrates/Inhibitors
Post-translational Modification
Allosteric Modulation
Metabolomics
Activity
Screening

Modulating Phenotype with Metabolites

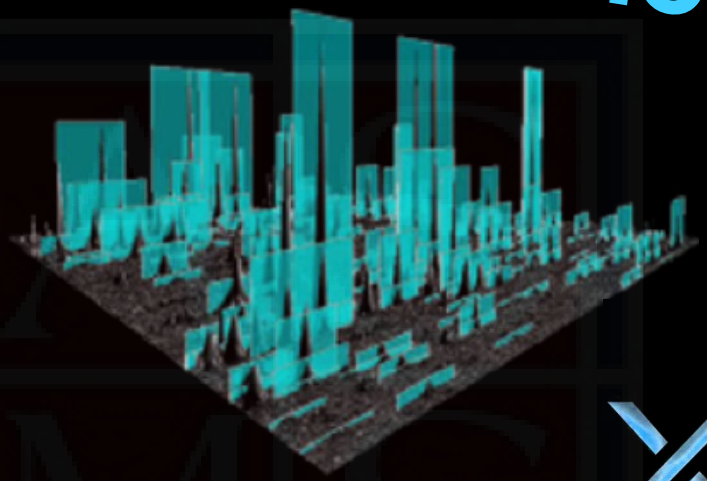


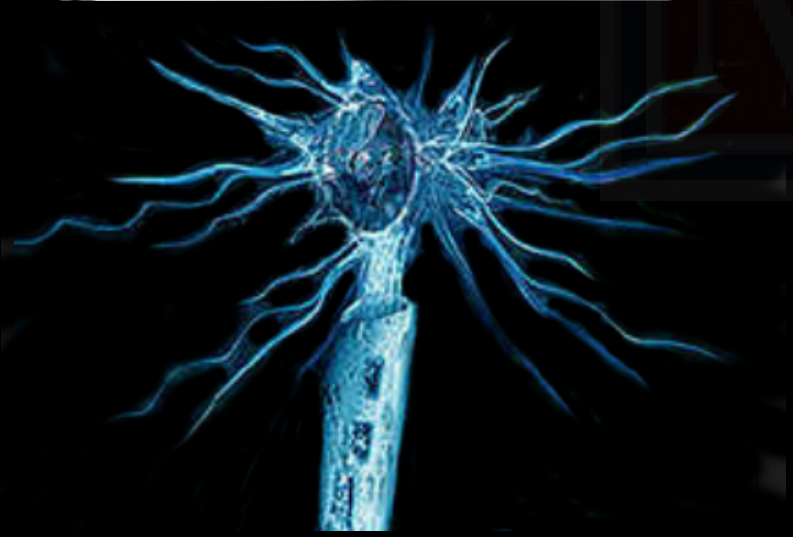


xcms Online

Nature Methods 2017
Cell Metabolism 2018
Nature Protocols 2018
Cell Chemical Biology 2018
Nature Biotechnology 2018
Nature Chemical Biology 2018

Modulating Phenotype with Metabolites





Modulating Phenotype with Metabolites

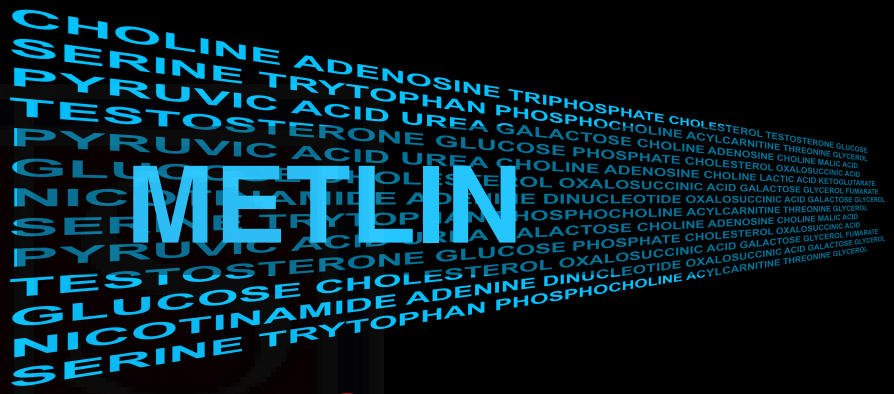
CHOLINE ADENOSINE TRIPHOSPHATE CHOLESTEROL TESTOSTERONE GLUCOSE
SERINE TRYTOPHAN PHOSPHOCHOLINE ACYLCARNITINE THREONINE GLYCEROL
PYRUVIC ACID UREA GALACTOSE CHOLINE ADENOSINE CHOLINE MALIC ACID
TESTOSTERONE GLUCOSE PHOSPHATE CHOLESTEROL OXALOSUCCINIC ACID
PYRUVIC ACID UREA CHOLINE ADENOSINE CHOLINE LACTIC ACID KETOGLUTARATE
GLUCOSE CHOLESTEROL OXALOSUCCINIC ACID GALACTOSE GLYCEROL FUMARATE
NICOTINAMIDE ADENINE DINUCLEOTIDE OXALOSUCCINIC ACID GALACTOSE GLYCEROL
SERINE TRYTOPHAN PHOSPHOCHOLINE ACYLCARNITINE THREONINE GLYCEROL
PYRUVIC ACID UREA GALACTOSE CHOLINE ADENOSINE CHOLINE MALIC ACID
TESTOSTERONE GLUCOSE PHOSPHATE CHOLESTEROL OXALOSUCCINIC ACID
OLESTEROL OXALOSUCCINIC ACID GALACTOSE GLYCEROL FUMARATE
E ADENINE DINUCLEOTIDE OXALOSUCCINIC ACID GALACTOSE GLYCEROL
OPHAN PHOSPHOCHOLINE ACYLCARNITINE THREONINE GLYCEROL

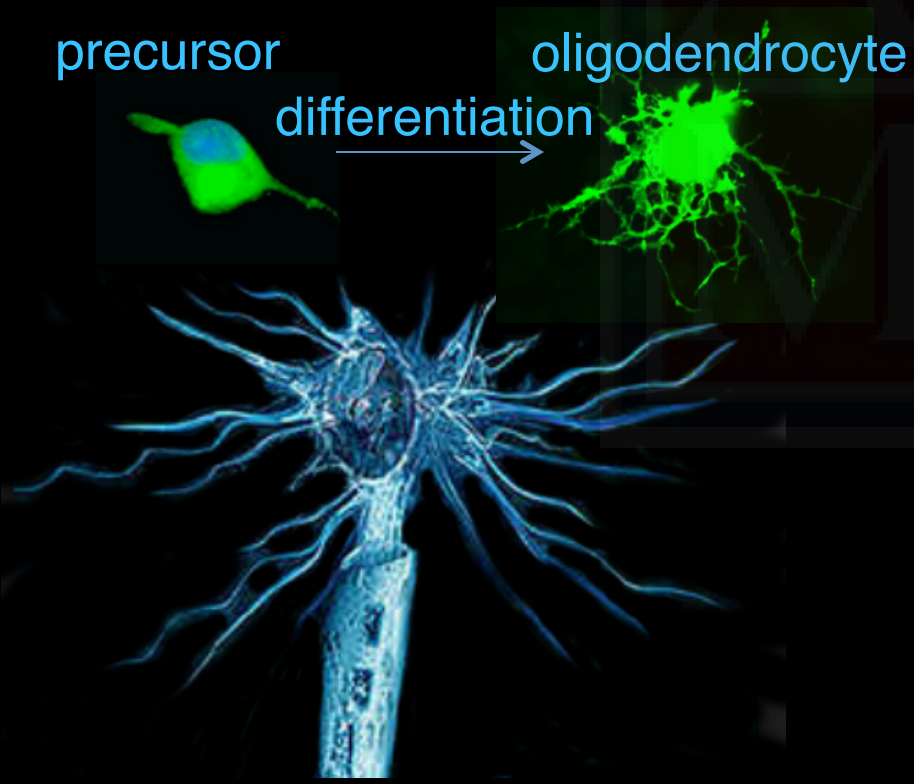
Multiple Sclerosis

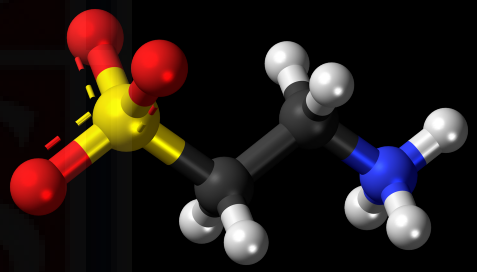
Nature Chem. Biol. 2018

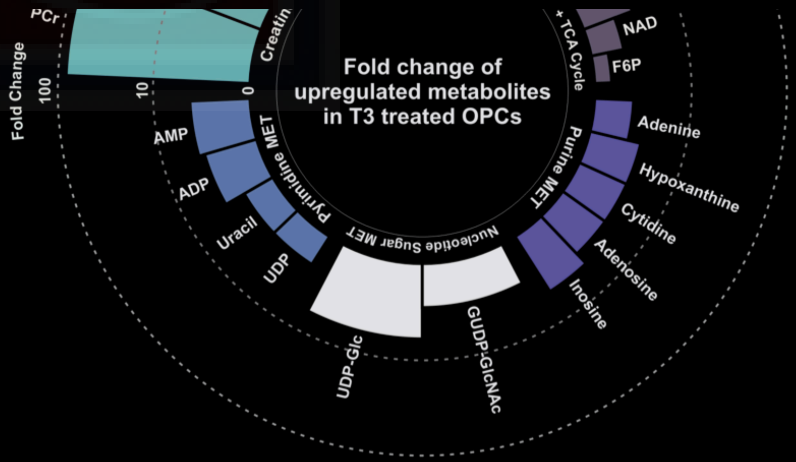
***Nature 2013* (Lairson)**

Multiple Sclerosis

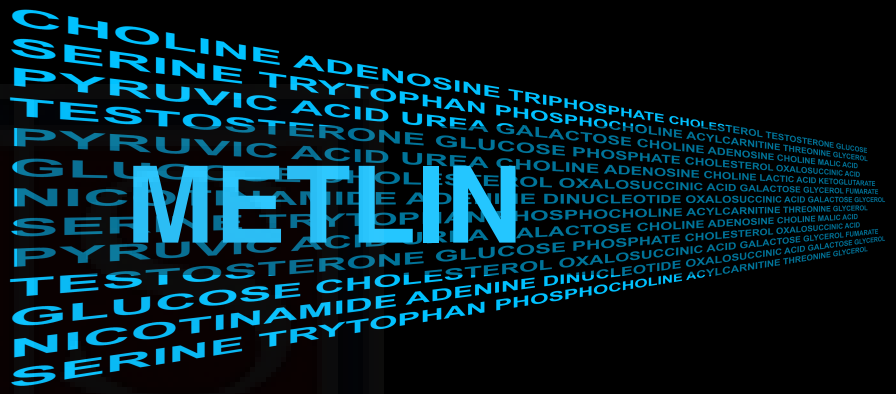


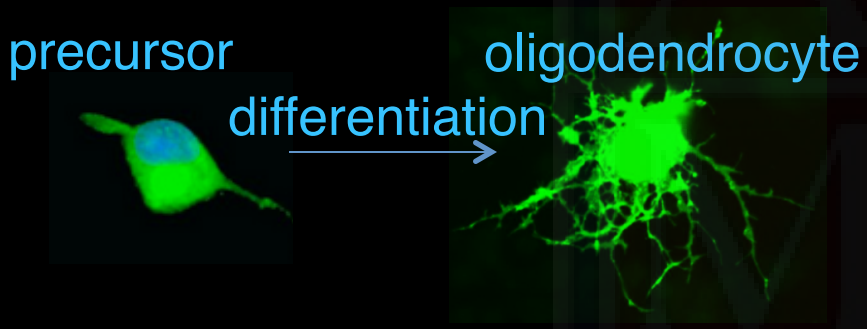


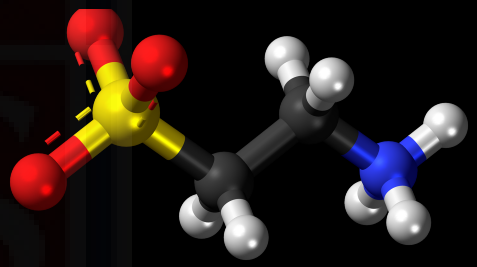


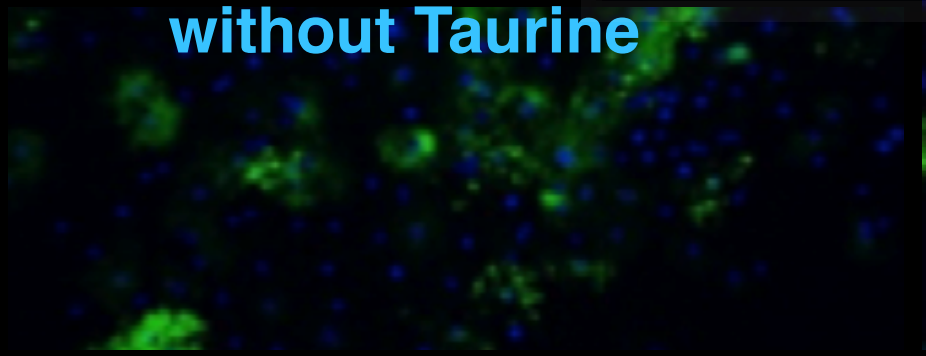


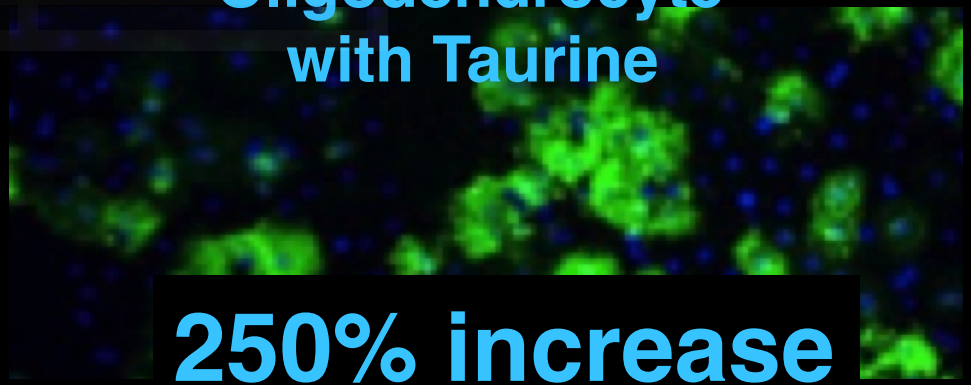
Multiple Sclerosis



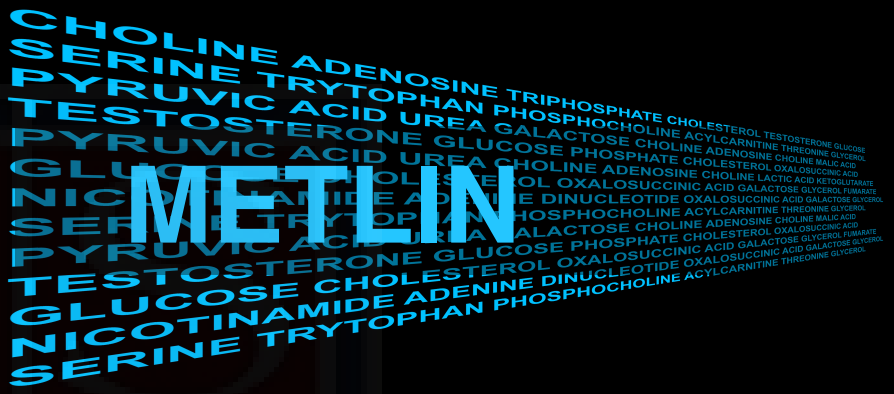


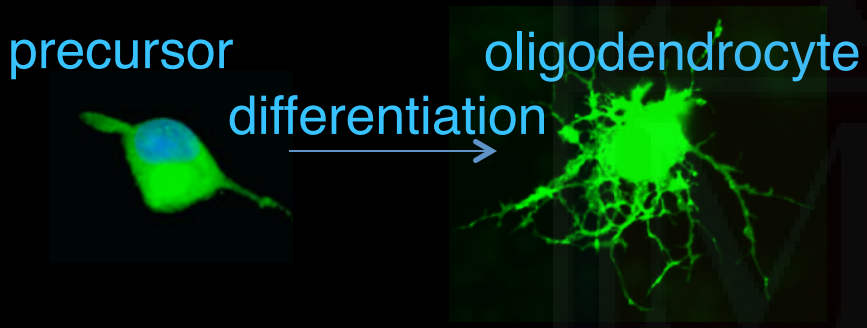


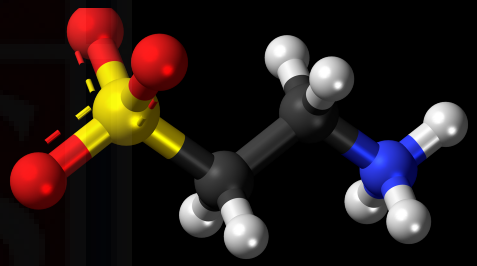


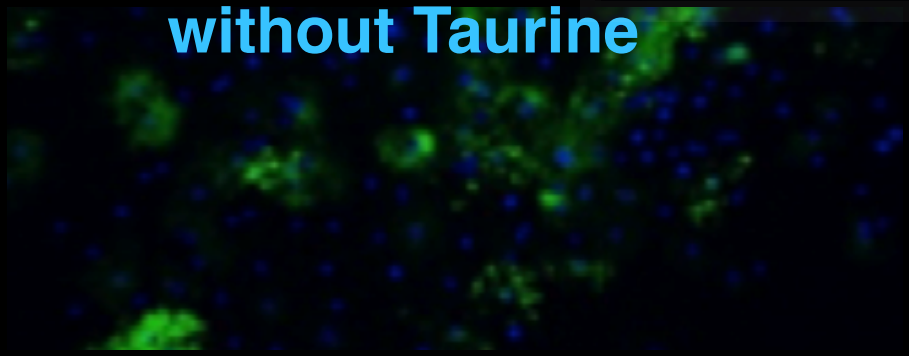


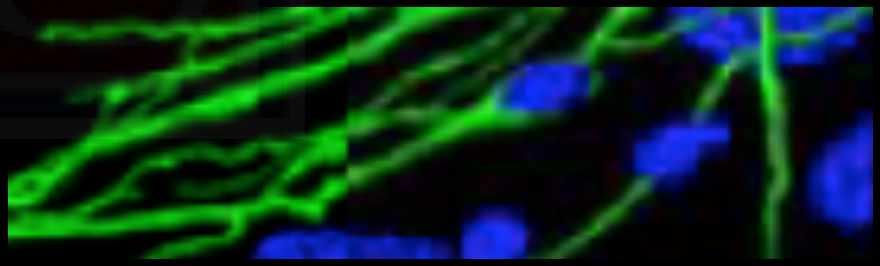
Multiple Sclerosis











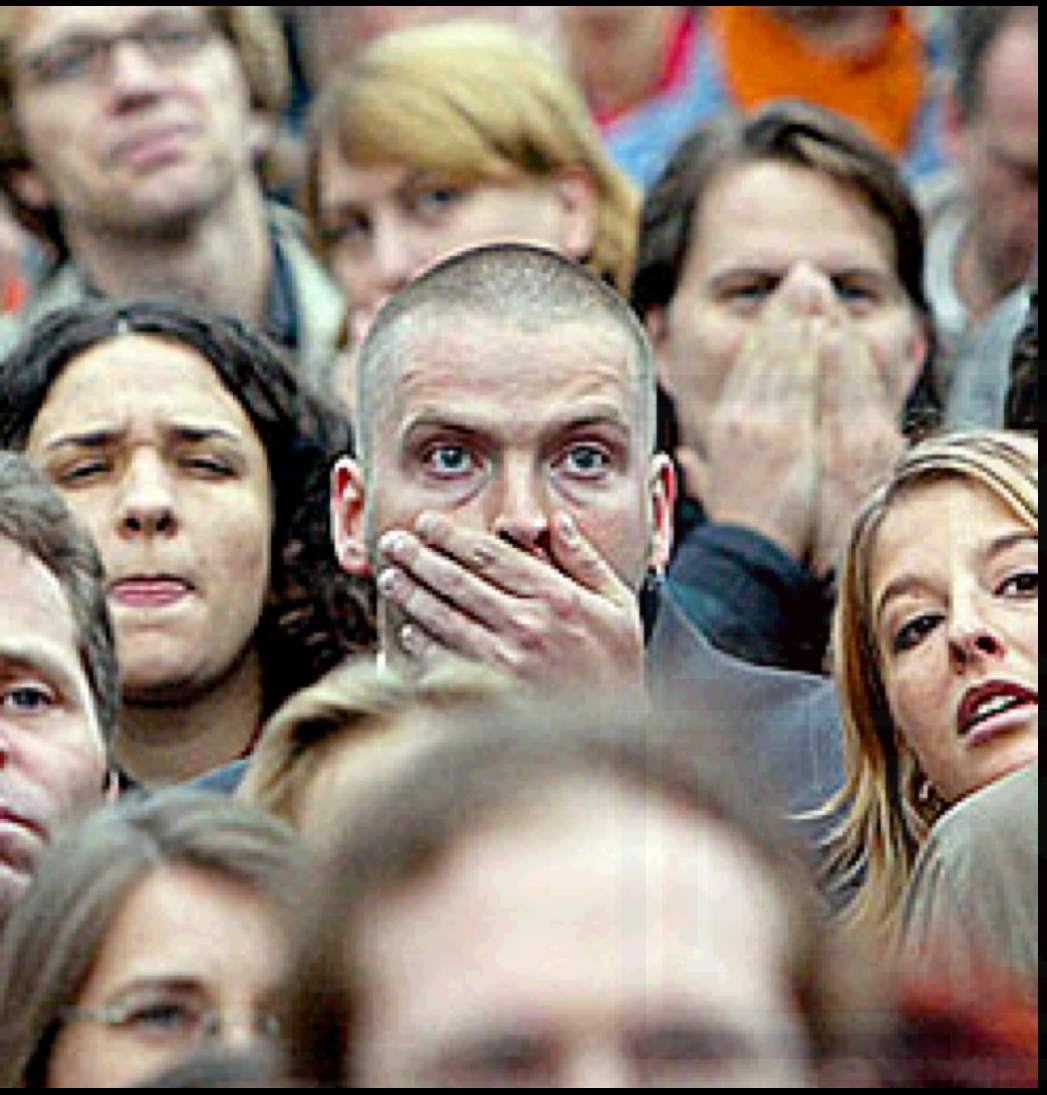
Myelin Sheath
Neuron regeneration

Precedence

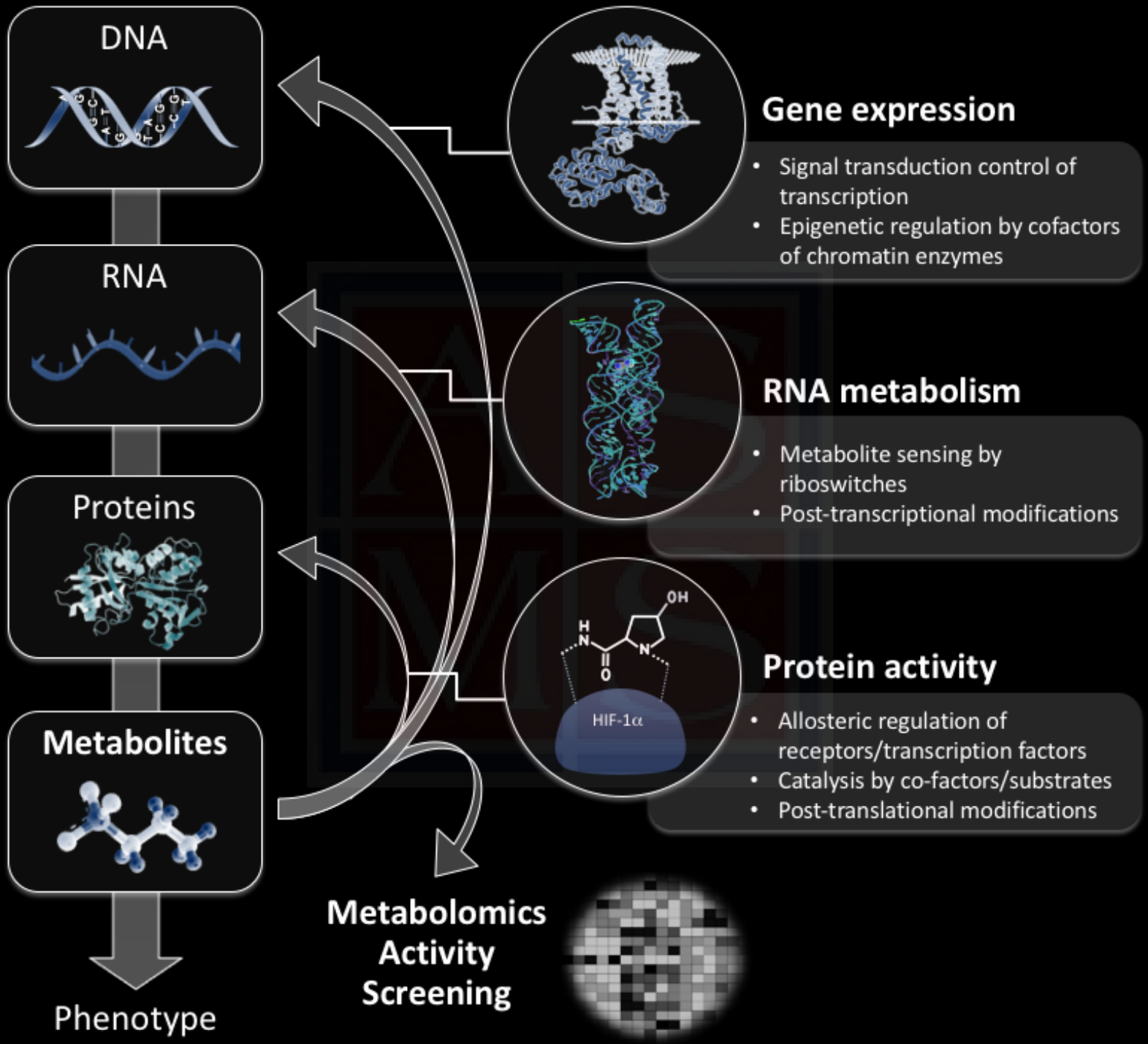
Metabolite	System	Journal	
Oleamide	Sleep	Science	1995
Neuroprotectin D1	Stem Cell Regulation	Nature Chem. Biology	2010
TMAO	Cardiac Disease	Nature	2011
Nicotinamide	Stem Cell Regulation	Nature Chem. Biology	2013
Dimethylsphingosine	Chronic Pain	Nature Chem. Biology	2013
PI (20:4/20:4)	Pathogen Killing	Journal of Immunology	2013
FAHFAs	Type 2 Diabetes	Cell	2014
CMP-furanpropan. Acid	Diabetes	Cell Metabolism	2014
TMAO	Cardiac Disease	Cell	2015
Polyamines/Lipids	Immuno-oncology	Cell Metabolism	2015
Hexadecenoic acid	Cardiovascular Disease	Cell Chemical Biology	2016
Taurine	Multiple Sclerosis	Nature Chem. Biology	2018
Itaconate	Anti-Inflammatory	Nature 3/2018	Nature 4/2018

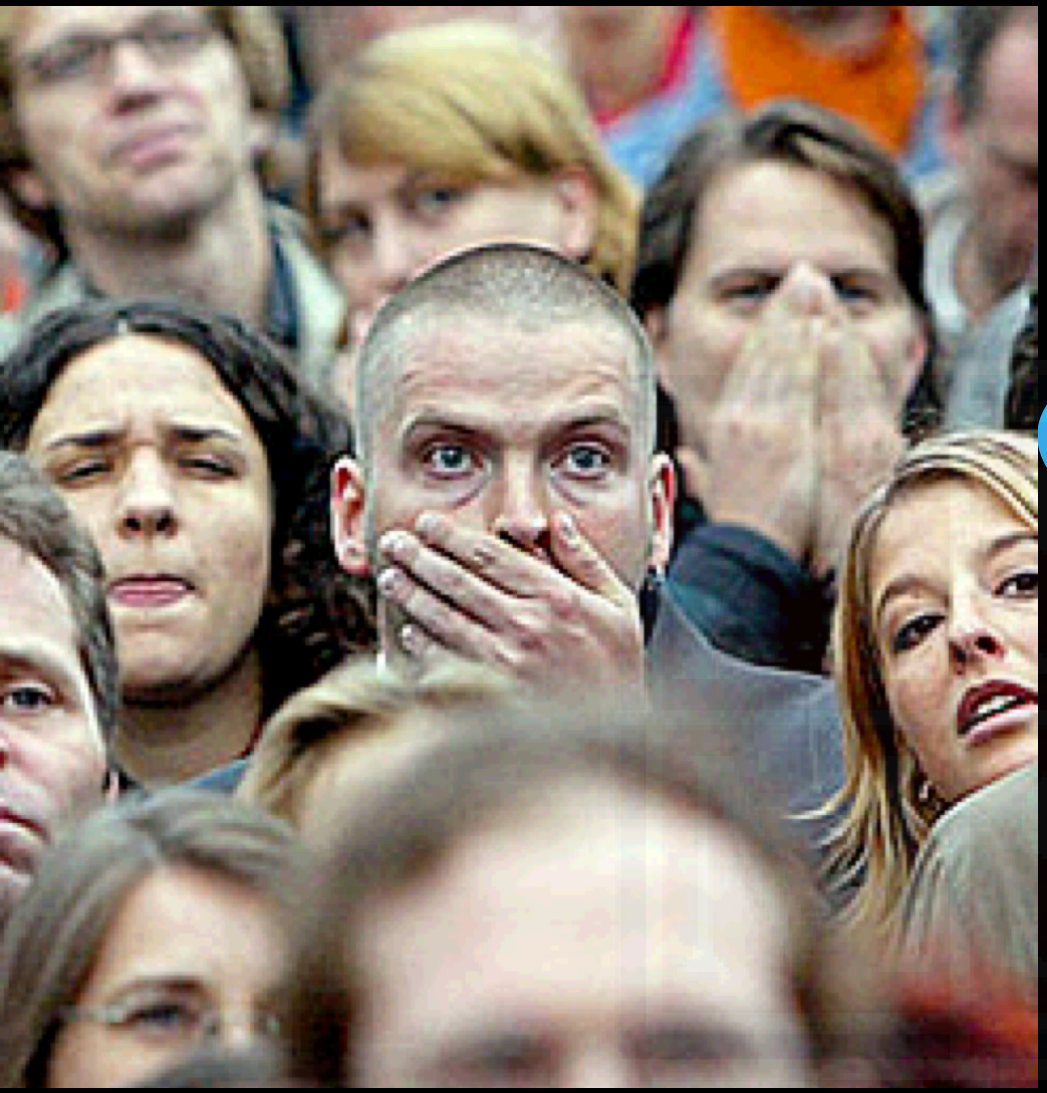
Precedence

Metabolite	System	Journal	
Oleamide	Sleep	Science	1995
Neuroprotectin D1	Stem Cell Regulation	Nature Chem. Biology	2010
TMAO	Cardiac Disease	Nature	2011
Nicotinamide	Stem Cell Regulation	Nature Chem. Biology	2013
Dimethylsphingosine	Chronic Pain	Nature Chem. Biology	2013
PI (20:4/20:4)	Pathogen Killing	Journal of Immunology	2013
FAHFAs	Type 2 Diabetes	Cell	2014
CMP-furanpropan. Acid	Diabetes	Cell Metabolism	2014
TMAO	Cardiac Disease	Cell	2015
Polyamines/Lipids	Immuno-oncology	Cell Metabolism	2015
Hexadecenoic acid	Cardiovascular Disease	Cell Chemical Biology	2016
Taurine	Multiple Sclerosis	Nature Chem. Biology	2018
Itaconate	Anti-Inflammatory	Nature 3/2018	Nature 4/2018



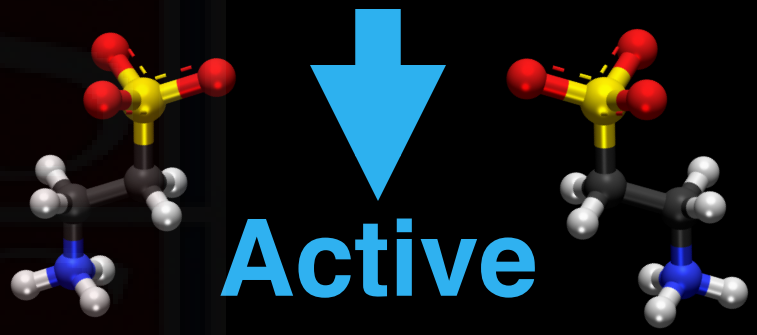
Primary Message?


DNA
RNA
Proteins
Metabolites
Phenotype
Gene expression
Signal transduction control of transcription
Epigenetic regulation by cofactors of chromatin enzymes
RNA metabolism
Metabolite sensing by riboswitches
Post-transcriptional modifications
Protein activity
Allosteric regulation of receptors/transcription factors
Catalysis by co-factors/substrates
Post-translational modifications
HIF-1α
OH
H
N
O
Metabolomics Activity Screening



Metabolomics

Biomarkers
Pathways
(Passive Observations)



Active Participants that can Modulate Phenotype

Primary Message?

Nature Biotechnology 2018

Advanced Metabolomics

June 3rd 2018

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

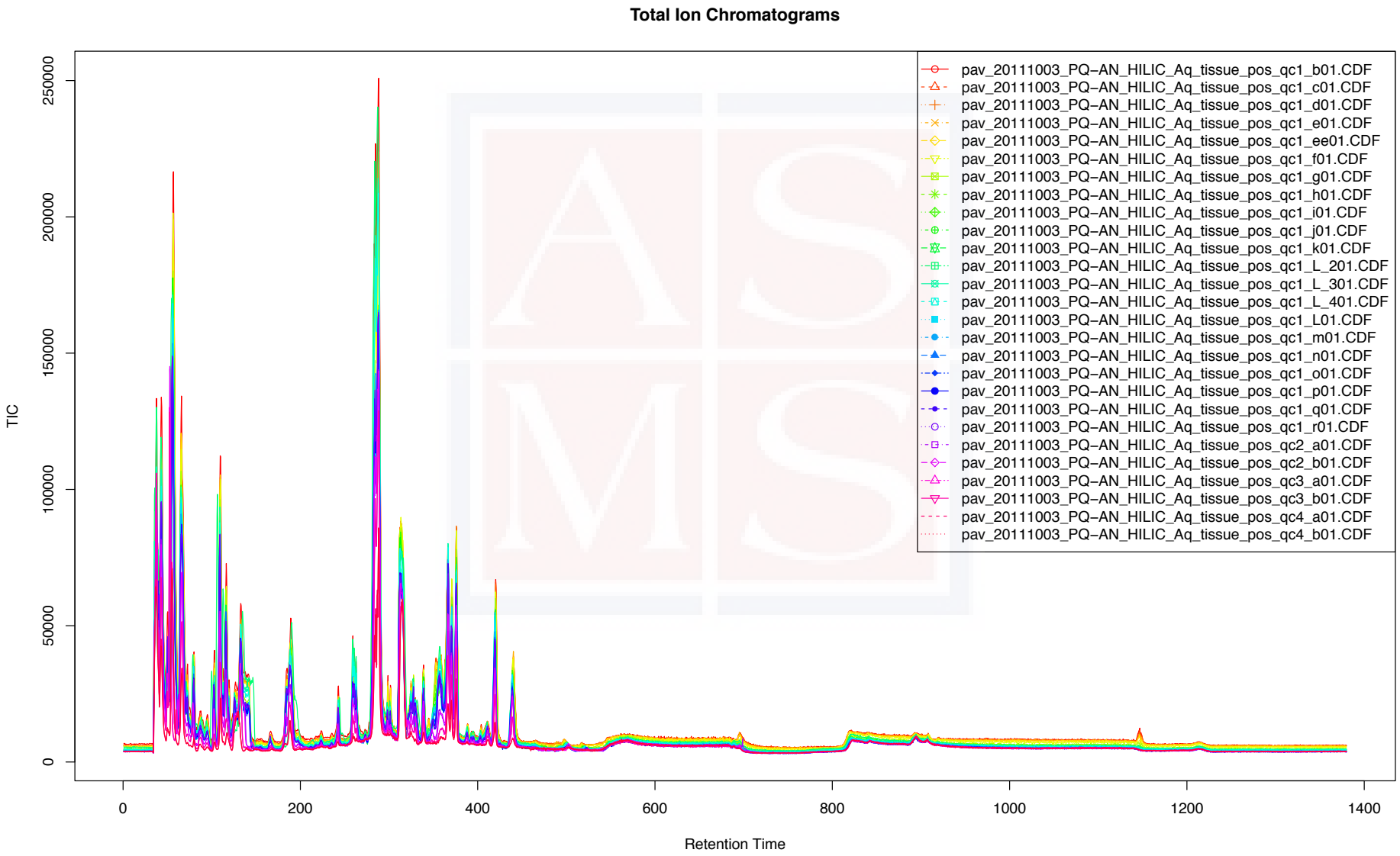
---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

What do we want to do ?

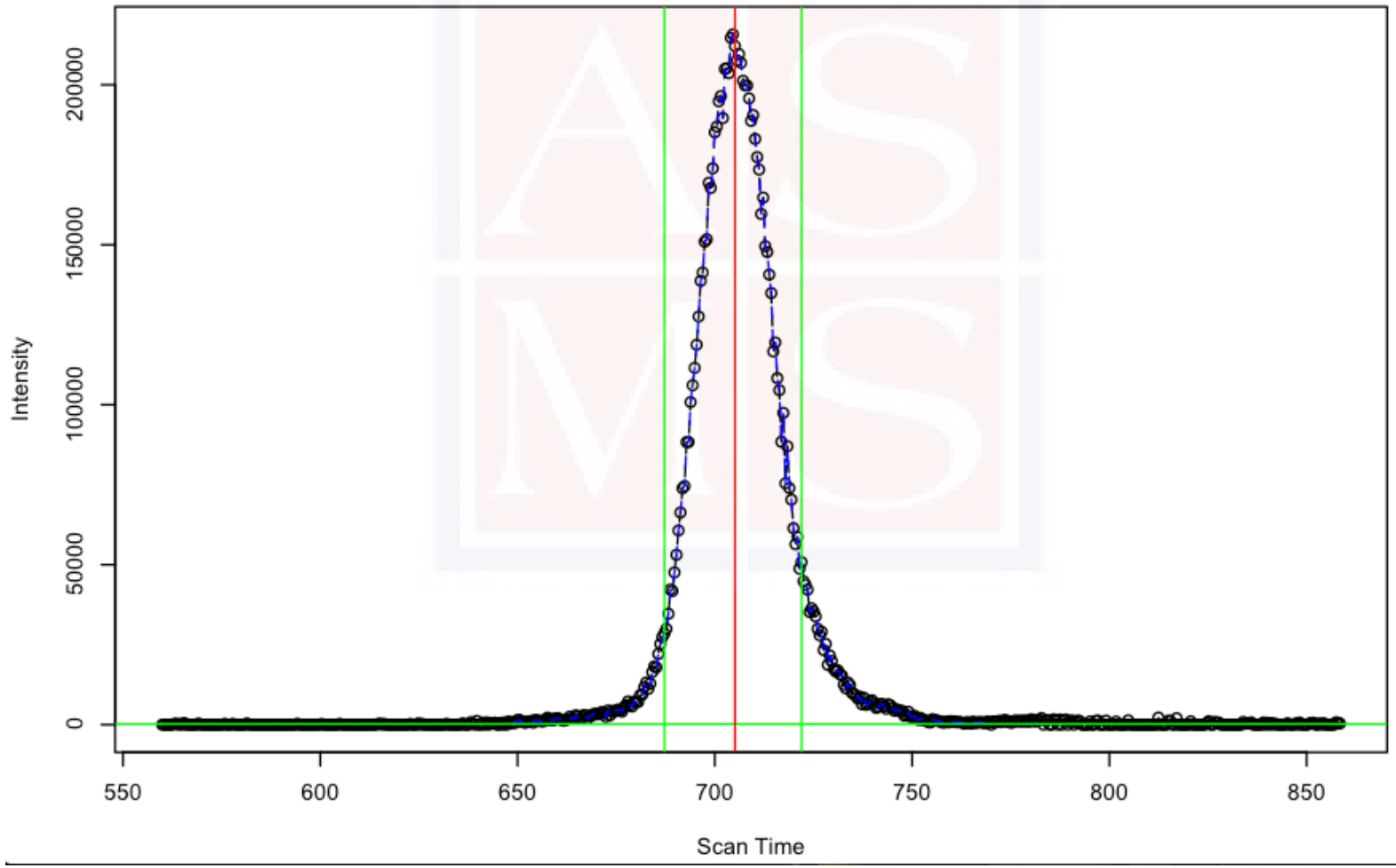
- Compare sample between different classes and analyse which metabolites are responsible/contributing to that difference
  - Run some statistics
  - Identify compounds
  - Understand biology

What are we dealing with



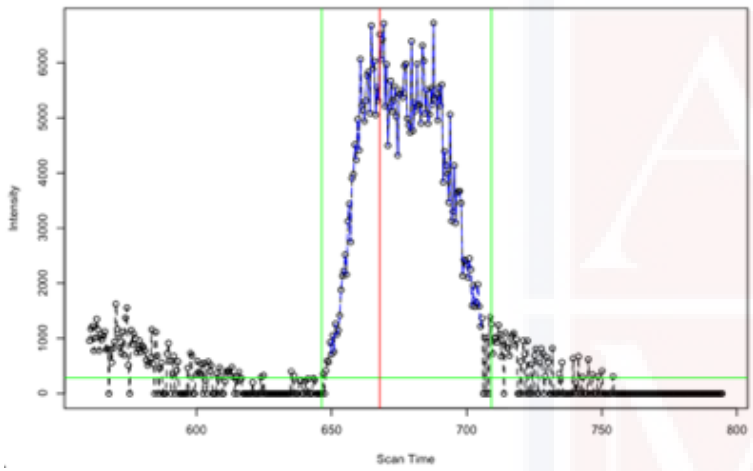
How to deal with it?

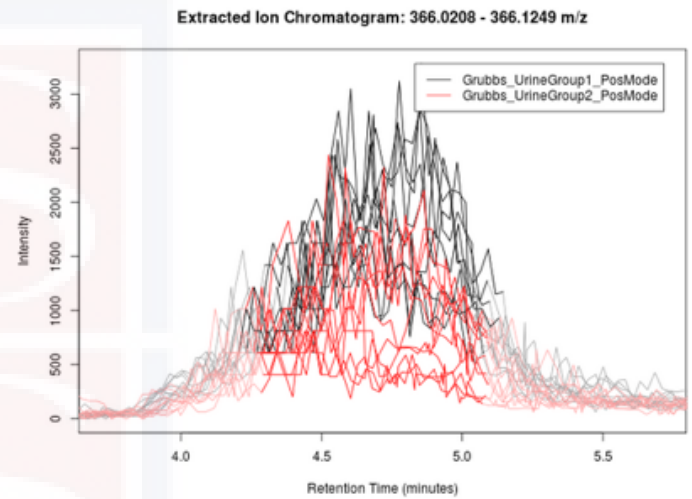
- Peak Detection… Easy

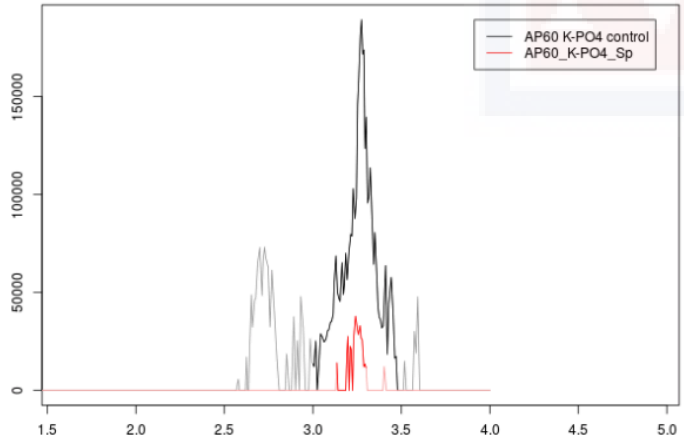


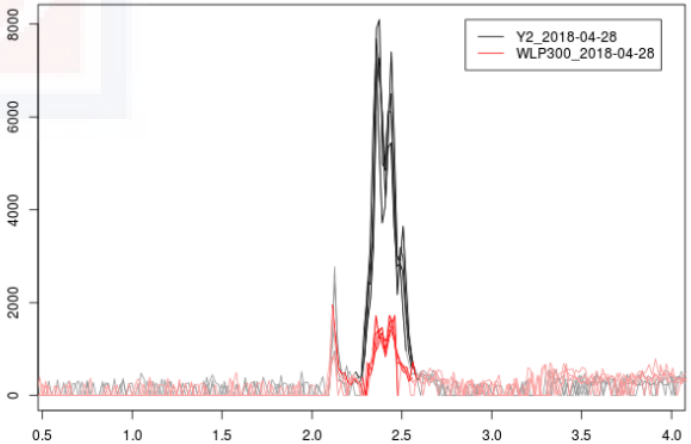
How to deal with it?

- Peak Detection… Easy …. Maybe not









What does current software do?

Analytica Chimica Acta

Available online 4 May 2018

In Press, Accepted Manuscript — Note to users

Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection

Zhucui Li[a, b, c], Yan Lu[a, b, d], Yufeng Guo[c], Haijie Cao[e], Qinhong Wang[c], Wenqing Shui[b, d]

Show more

https://doi.org/10.1016/j.aca.2018.05.001

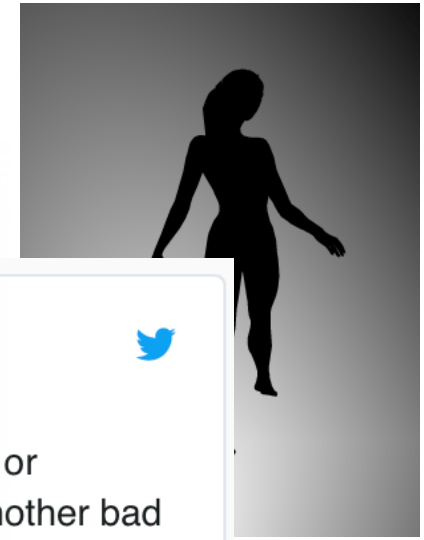
Get rights and content

Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395 (2010).

R Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. E. XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry* **84,** 5035–5039 (2012).

Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Meth* **12,** 523–526 (2015).

What does current software do?



Philadelphia Police
@PhillyPolice

Please don't call 911 to ask if we're hearing "Laurel" or "Yanny". The only thing we hear is the creation of another bad hashtag. (And Laurel. We're definitely hearing Laurel).

5:31 AM - May 16, 2018

4,119 1,513 people are talking about this

Hear Both Yanny and Laurel

Click to pause

Laurel Yanny

Click here to submit when you first hear the words change

General workflow

Peak Detection

Grouping
Clustering of Peaks
Across replicates

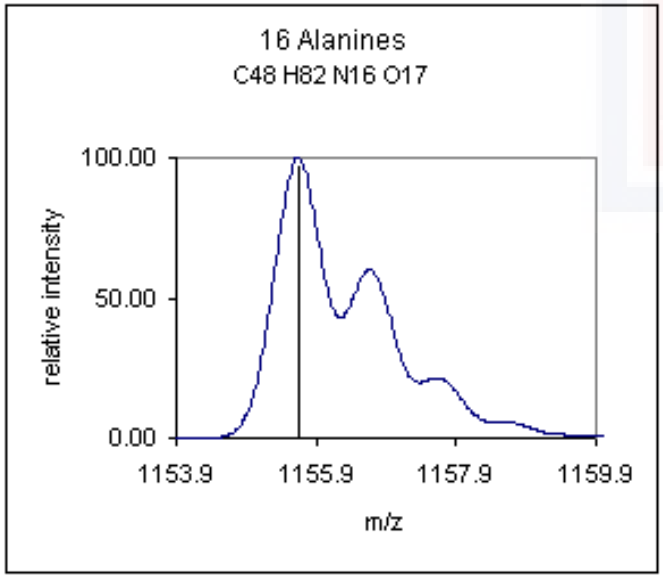
Retention time
Alignment

Statistical Analysis

mzMine 2 – peak detection

- Breaks the collection of peaks into 3 steps
  - Mass Detection
  - Chromatogram building
  - Peak Deconvolution

- Mass detection is centroiding of the data

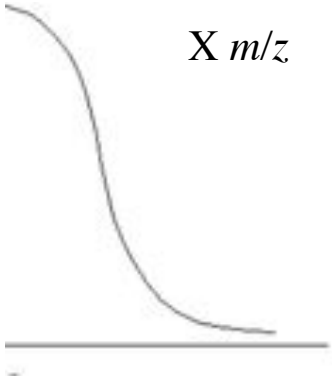


- MzMine has many algorithms –
  - Basics are to look for a peak and assign the centre as the centroid
  - NB m/z domain only

mzMine 2 – peak detection

- Breaks the collection of peaks into 3 steps
  - Mass Detection
  - Chromatogram building
  - Peak Deconvolution

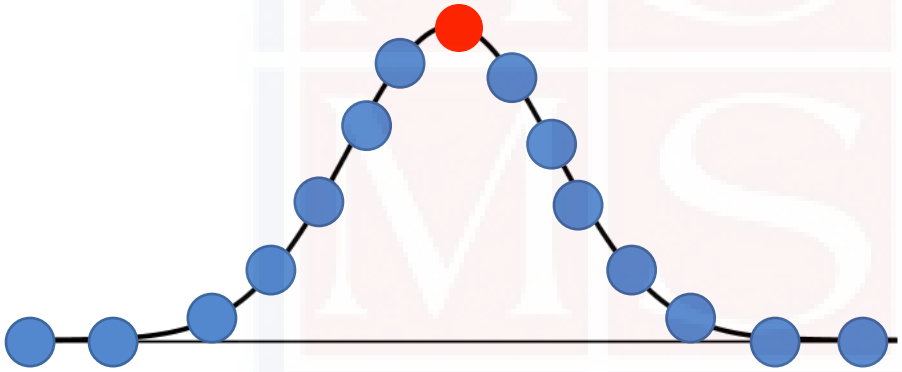
- Connect an m/z slice by intensity with the most intense ions first
- Look for distributions within a time window.



- Re-order the slice by time and apply a filter to integrate

MS-DIAL – peak detection

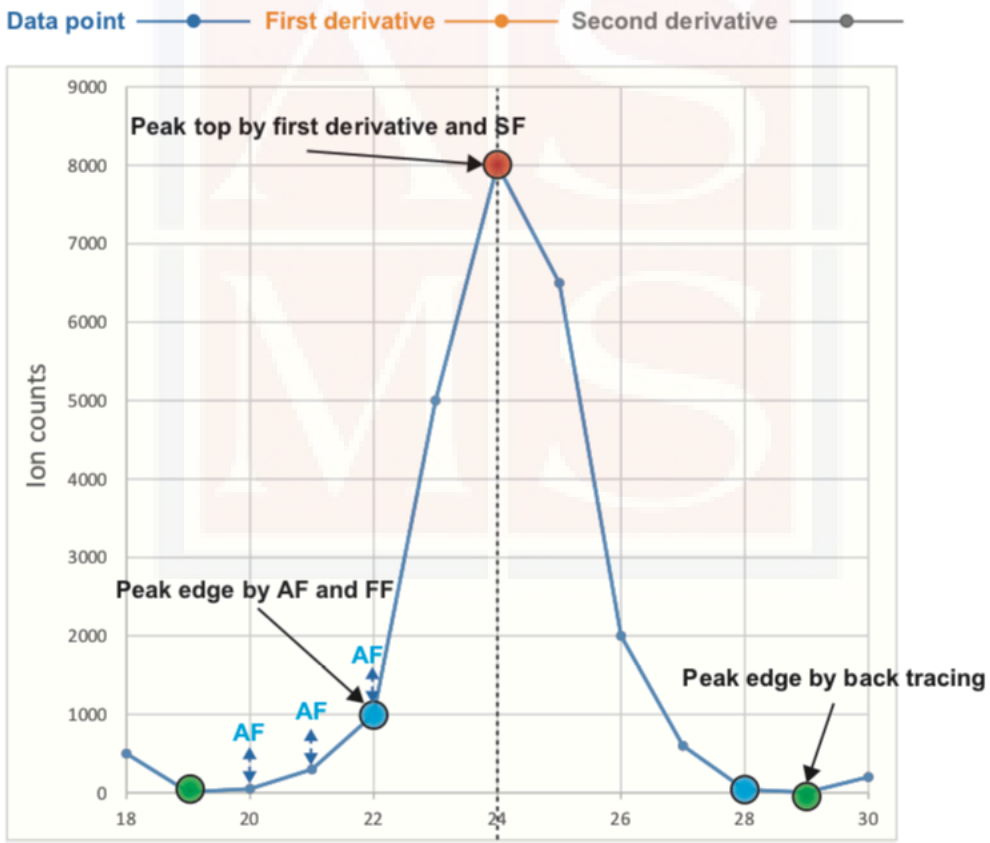
- Peak detection
- Peak spotting



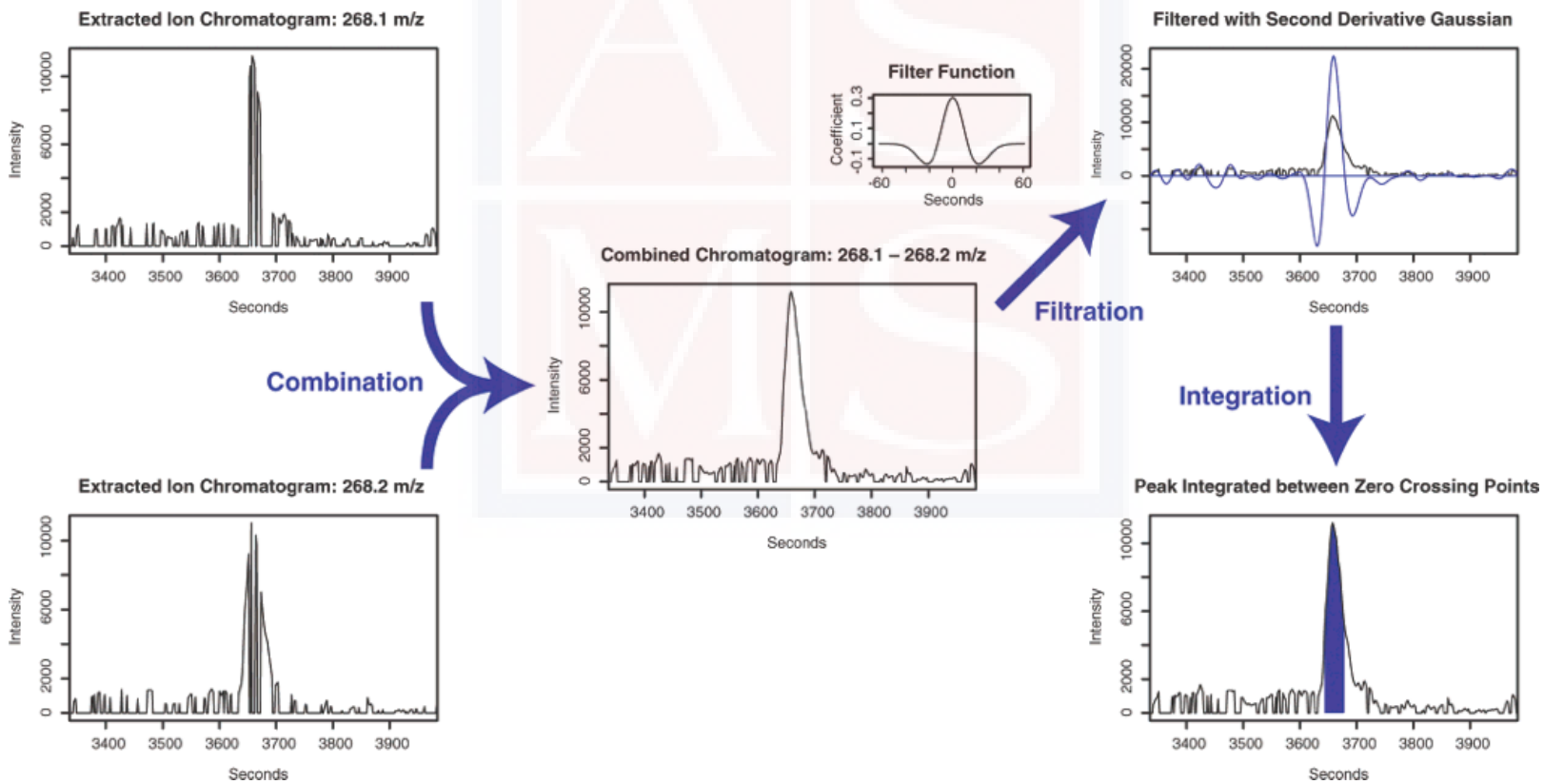
MS-DIAL

MS-DIAL – peak detection

- Peak detection
- Peak spotting

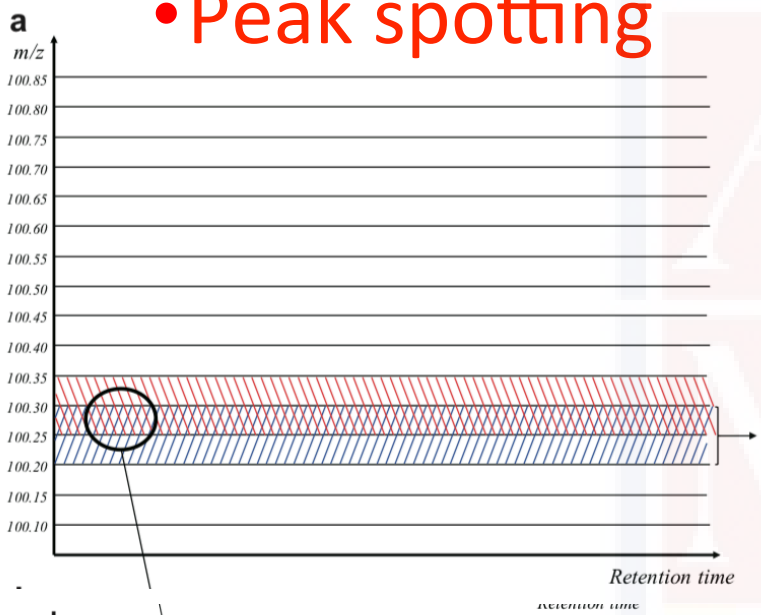


XCMS – matched Filter

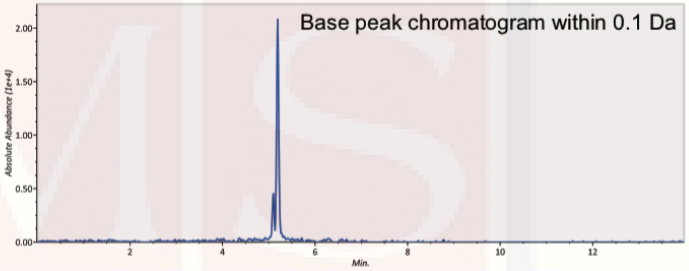


MS-DIAL – peak detection

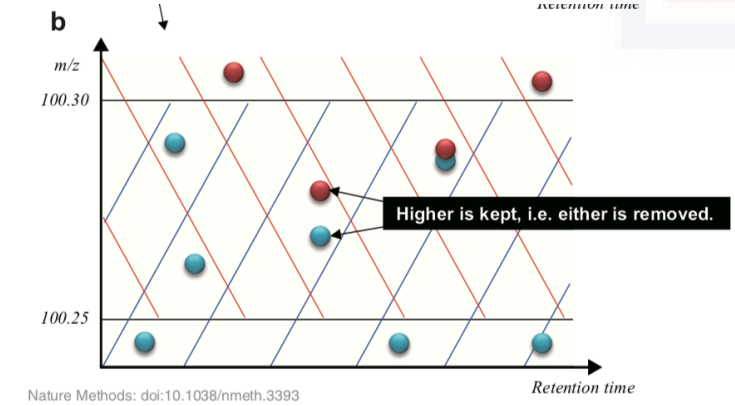
- Peak detection
- Peak spotting



Scan number	Retention time [min]	Base peak *m/z*	Base peak intensity
1	0.1	100.2054	1
2	0.12	100.2053	10
3	0.14	100.2053	5
4	0.16	100.2052	50
5	0.18	100.2051	200
6	0.2	100.2054	1500
7	0.22	100.2054	3000
8	0.24	100.2054	1700
9	0.26	100.2053	180
10	0.28	100.205	60



MS-DIAL



Nature Methods: doi:10.1038/nmeth.3393

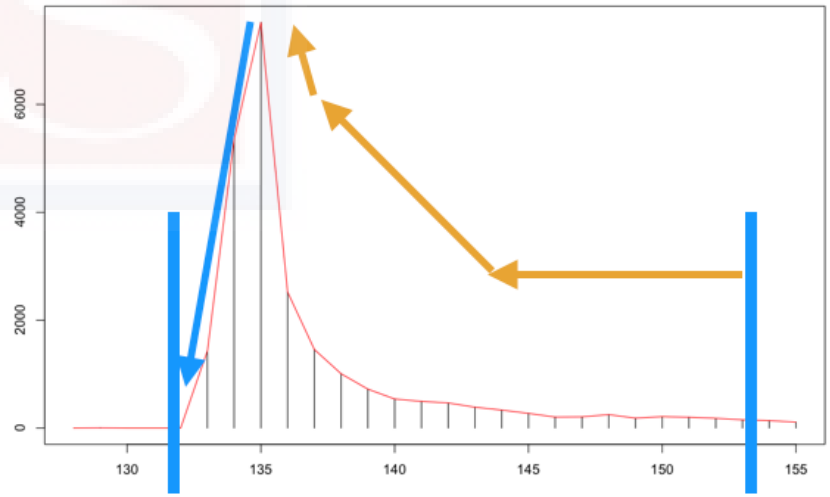
XCMS – centWave

- LC-MS traces are like Missiles!

XCMS – CentWave

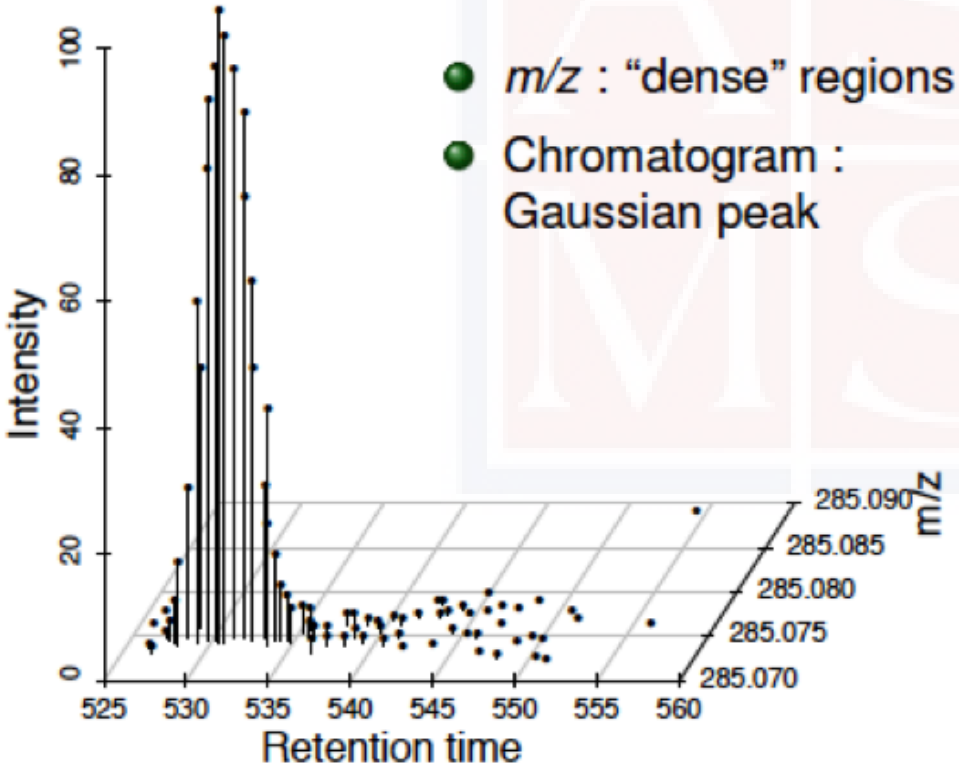
Tracking Missiles is
like tracking LC-MS traces

Trace backward along the trace
This will define the area of the 'bin'

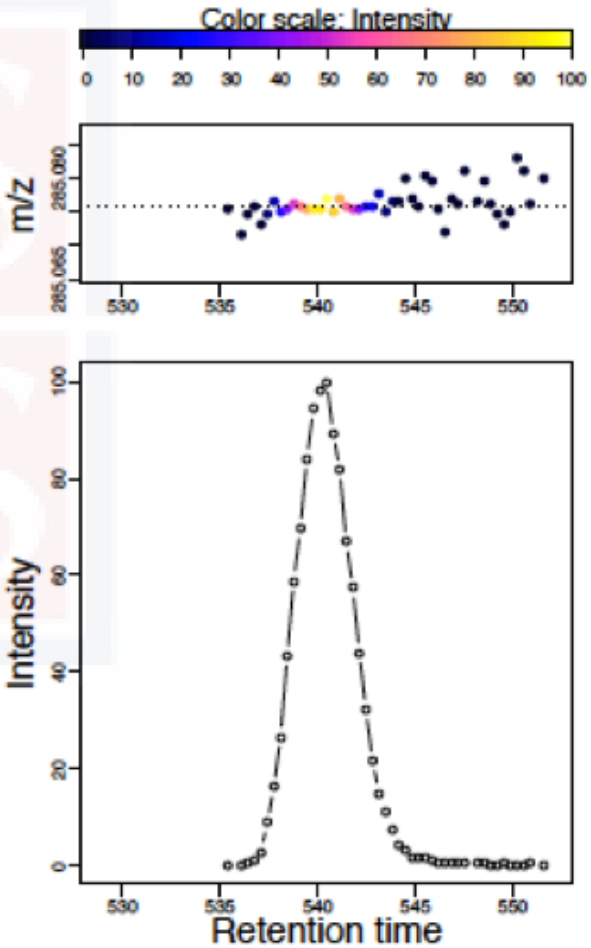


1. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9,** 504 (2008).

XCMS - CentWave



- *m/z* : “dense” regions
- Chromatogram : Gaussian peak



General workflow

Peak Detection

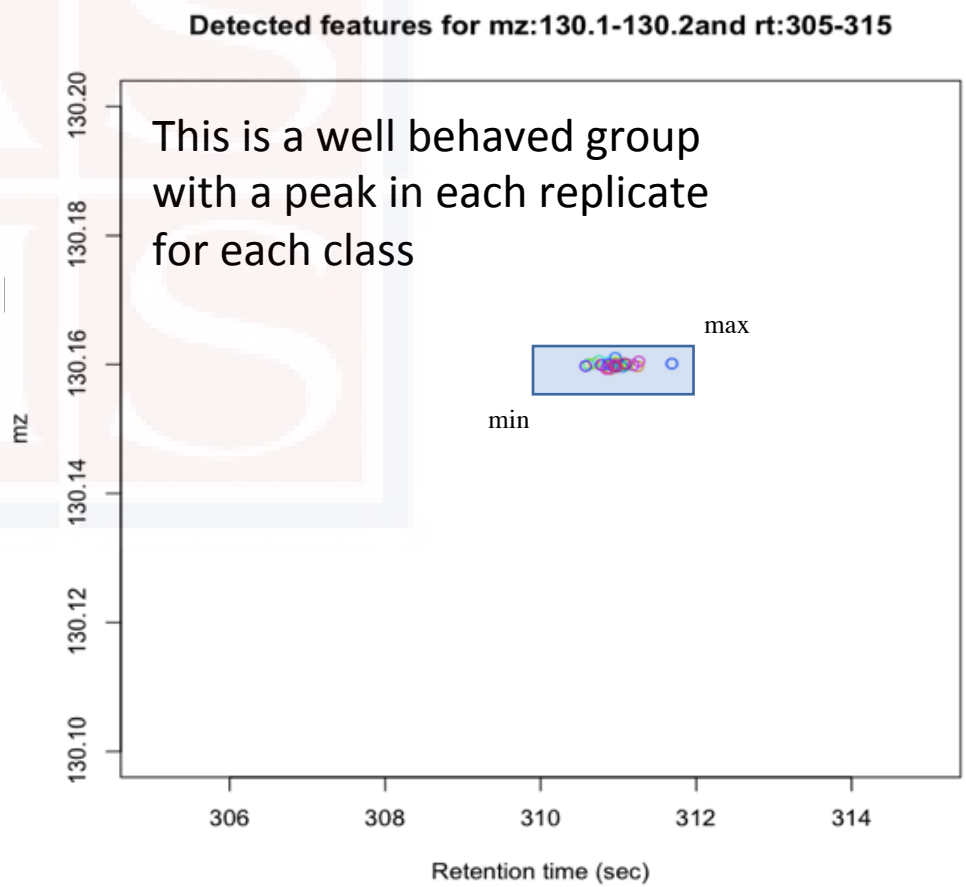
Grouping
Clustering of Peaks
Across replicates

Retention time
Alignment

Statistical Analysis

XCMS - Grouping

- <u>Density algorithm</u>
- First time using all files
- Looks for closly clustered/dense peaks across multiple files
  - Once grouped togethe in xcms terms they are a features



All - Grouping

- mzMine uses grouping to also align simtanously.
- This works on a nearest neighbor system
- MS-Dial reused this algorithm
- A reference spectra is setup by finding features that closely grouped together.
- Features that are further are scored to be in that group
- Live demo of algorithm
- NB also alignment for MS-Dial and mzMine

General workflow

Peak Detection

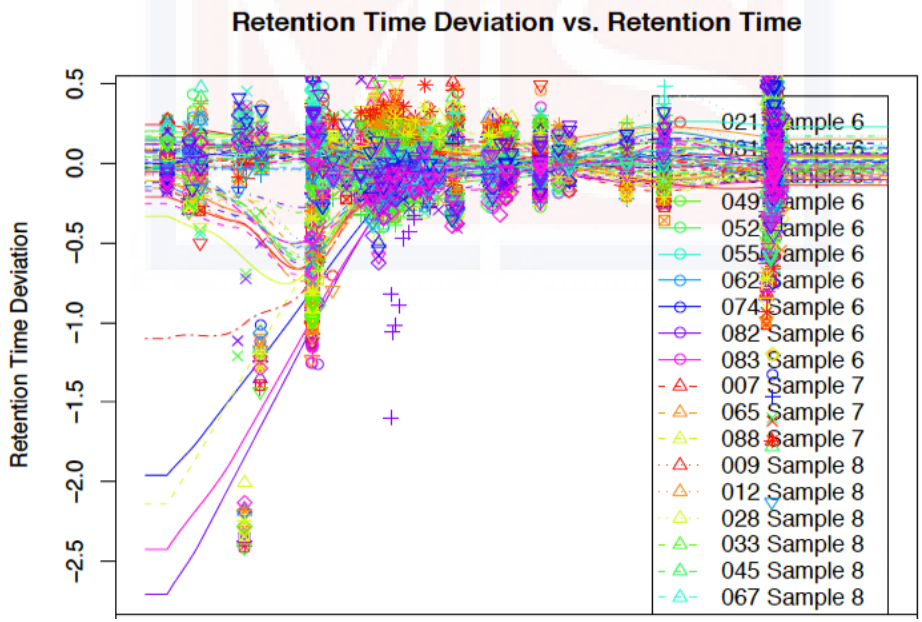
Grouping
Clustering of Peaks
Across replicates

Retention time
Alignment

Statistical Analysis

XCMS - Retention time alignment

- <u>Peak groups alignment</u>
- Particular to XCMS
- Uses internal features that are naturally well grouped as anchors
- Uses a local regression (loess) between these anchors to find deviation profile



XCMS - Retention time alignment

- <u>Obiwarp algorithm</u>
- Retention time correction based on spectra similarity
- No initial grouping needed
  - Re-reads raw files
- Warps the chromatogram to a median profile
  - Acts as a mold to which other chromatograms are warped
  - A Dynamic programming technique to find paths of greatest similarity between each.
  - The path is the deviation profile
  - Similar technique to blast transcript alignments

•John T. Prince and Edward M. Marcotte **Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping** Analytical Chemistry, 2006 78 (17), 6140-6152

General workflow

Peak Detection

Grouping
Clustering of Peaks
Across replicates

Retention time
Alignment

Statistical Analysis

Results …

MZmine 2

XCMS Online

MS-DIAL

Was the software able to find the compound?

			Total features	Consensus features	True features	True feature ID rate[a] (%)
QE HF dataset	Targeted		-	-	836	-
	Untargeted	Compound Discoverer	10,525	10,525	748	89.5
		MS-Dial	21,545	17,726	799	95.6
		MZmine 2	20,021	18,871	769	92.0
		XCMS	35,215	30,680	820	98.1

Results …

MZmine 2

XCMS Online

MS-DIAL

Was the software able to find quantify the compound?

			Accurately quantified true features	Quantification accuracy rate (%)	True discriminating markers	False discriminating markers
QE HF dataset	Targeted		836	100	50	0
	Untargeted	Compound Discoverer	482	64.4	41	111
		MS-Dial	654	81.9	42	42
		MZmine 2	761	99.0	48	3
		XCMS	731	89.2	45	51

Conclusions and questions

- Different software different results…
- Personal taste (to some extent)
- Some software's do more than what was discussed.
  - SWATH processing – MS-DIAL
  - ADAP algorithm – mzMine2
  - System biology – XCMS Online
  - Etc…

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

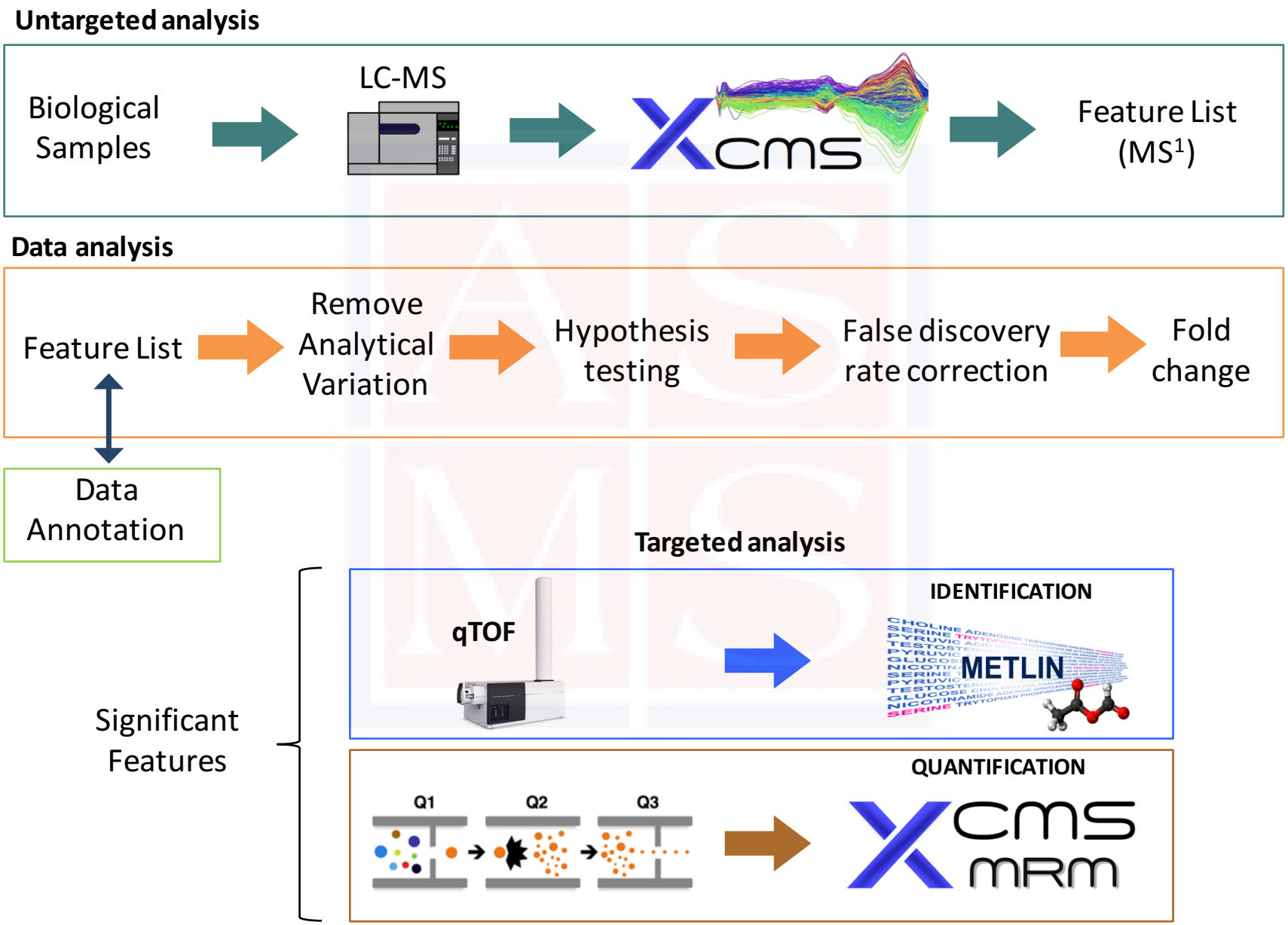
---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

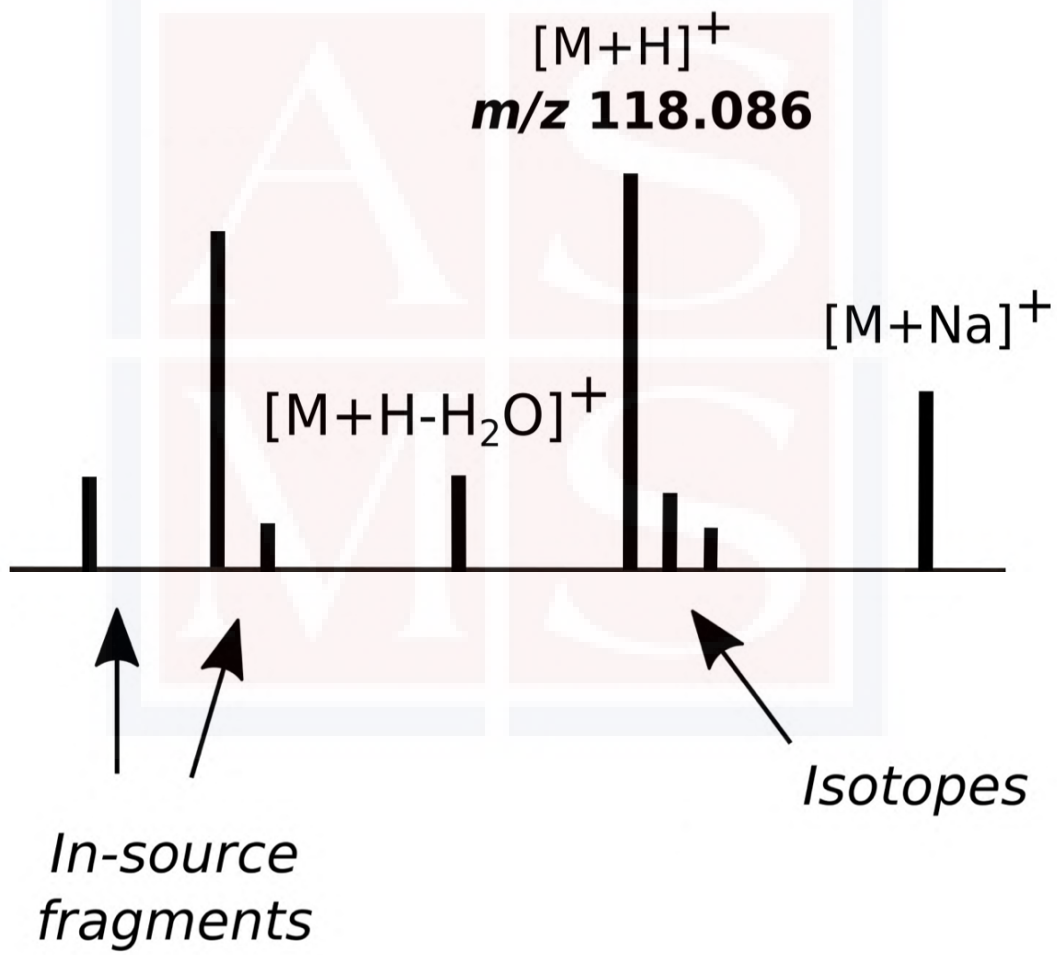
Metabolite Annotation

- **Overview**
- **Annotation strategies**
  1. MS^1 pseudo-spectra extraction
  2. Adduct mass rules
  3. Biochemical knowledge
  4. Use and integration of tandem MS data
  5. Retention time calibration
- **Annotation in practice**
  1. CAMERA
  2. xMSannotator
  3. Everest
  4. eRah (GC/MS)

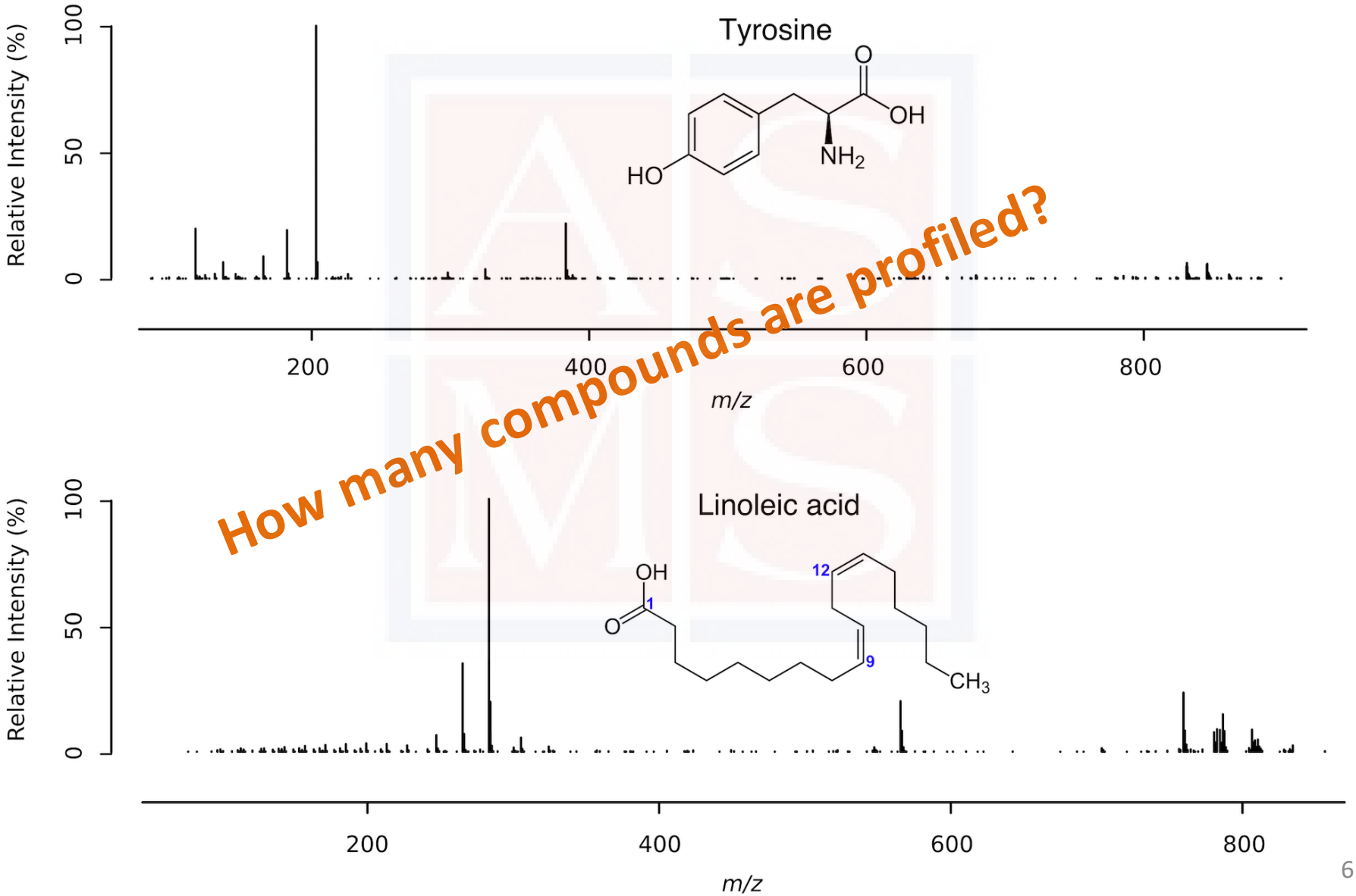
The Untargeted Metabolomics Workflow



LC-MS data: highly dimensional and redundant

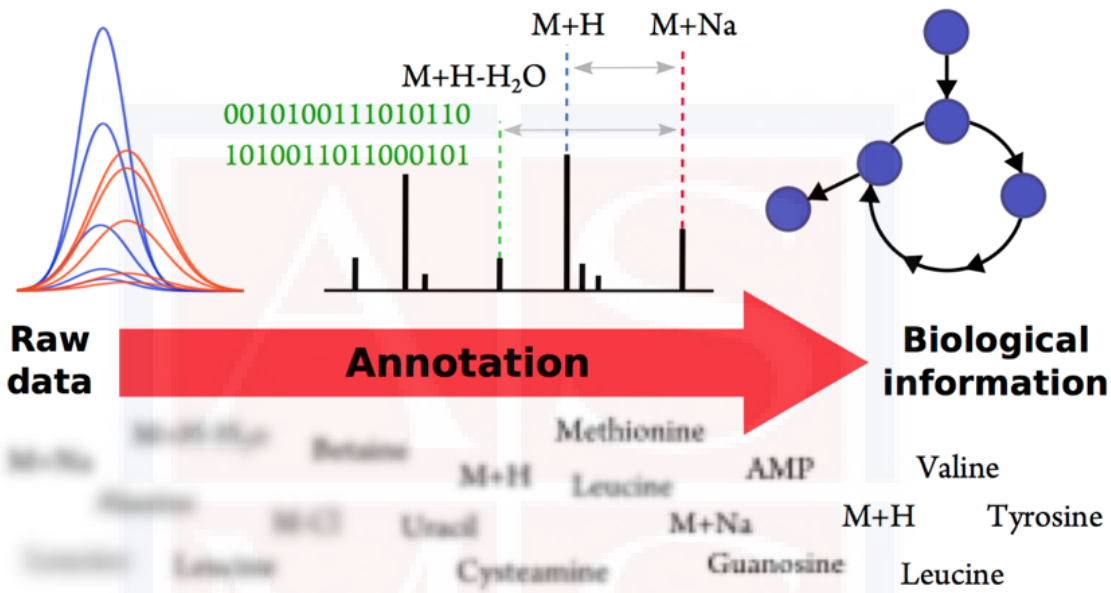


LC-MS data: highly dimensional and redundant



How many compounds are profiled?

Annotation

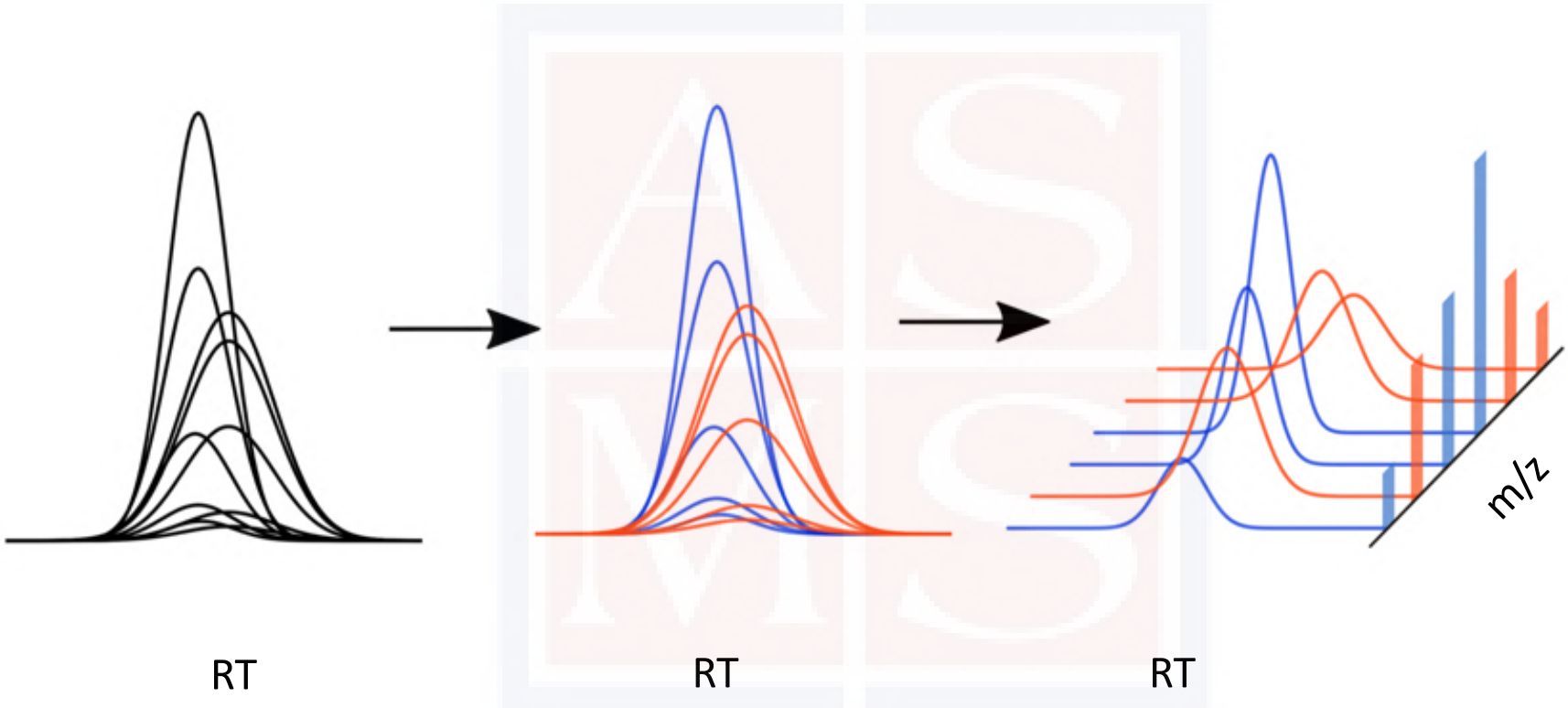


Annotation is defined as the process of "noting" and thus, assigning each observed feature with their identity.

Summary

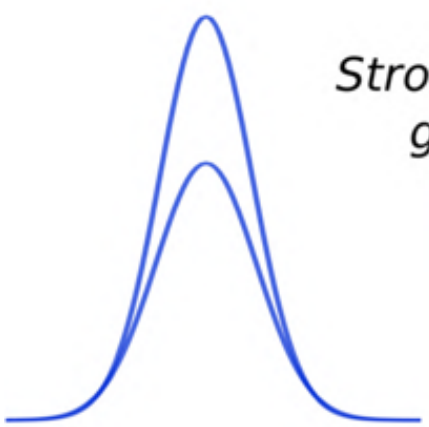
1. MS^1 pseudo-spectra extraction
2. Adduct mass rules
3. Biochemical knowledge
4. Use and integration of tandem MS data
5. Retention time calibration

MS^1 pseudo-spectra extraction



MS^1 pseudo-spectra extraction

Peak Shape Correlation



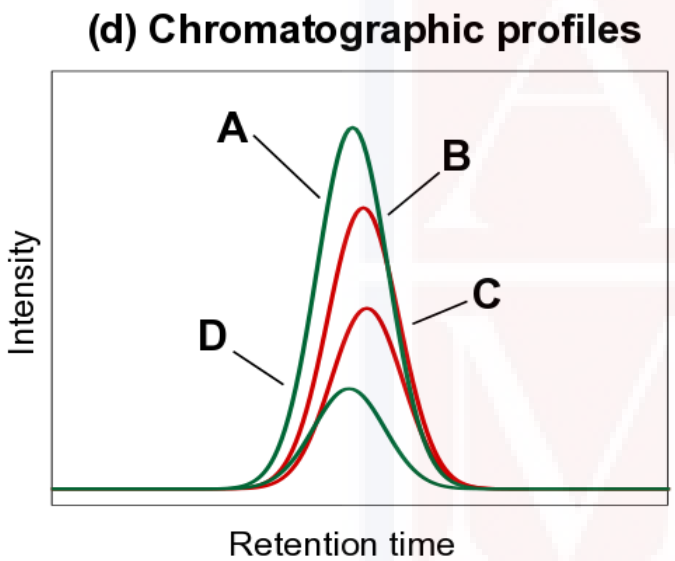
Strong correlation:
group peaks

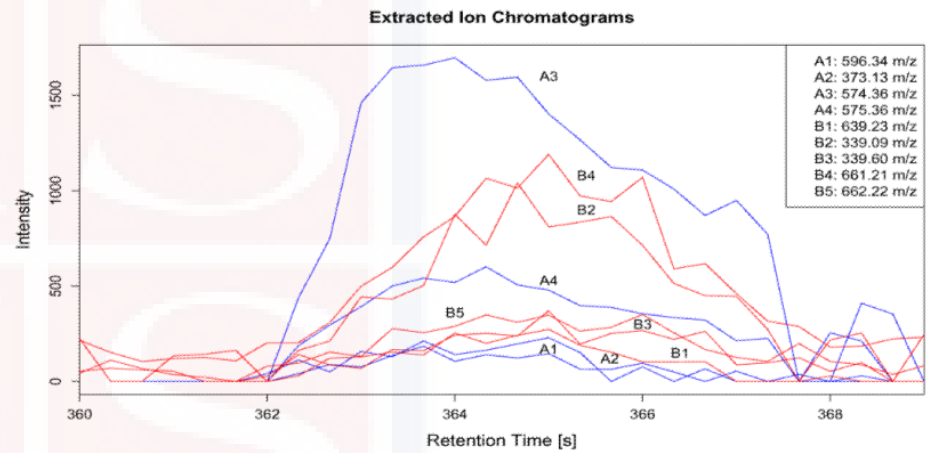
Weak correlation:
do not group peaks

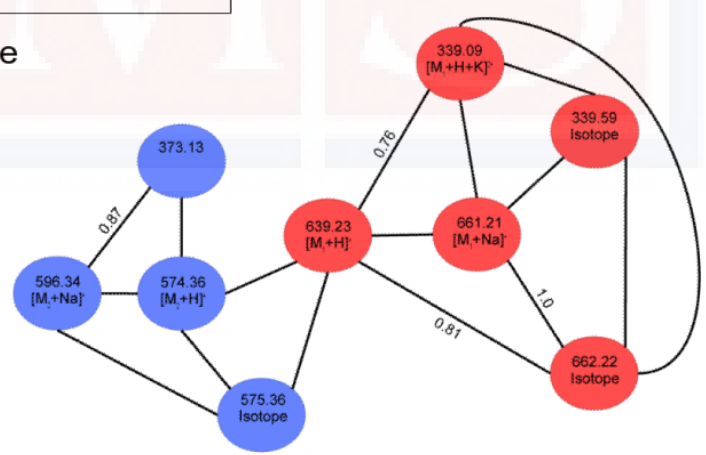
MS[1] pseudo-spectra extraction

Peak Shape Correlation

CAMERA



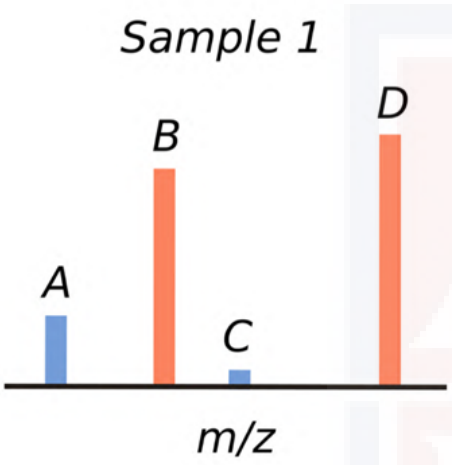




Picture from C. Kuhl et al., Anal. Chem., 84 (2012) 283-289

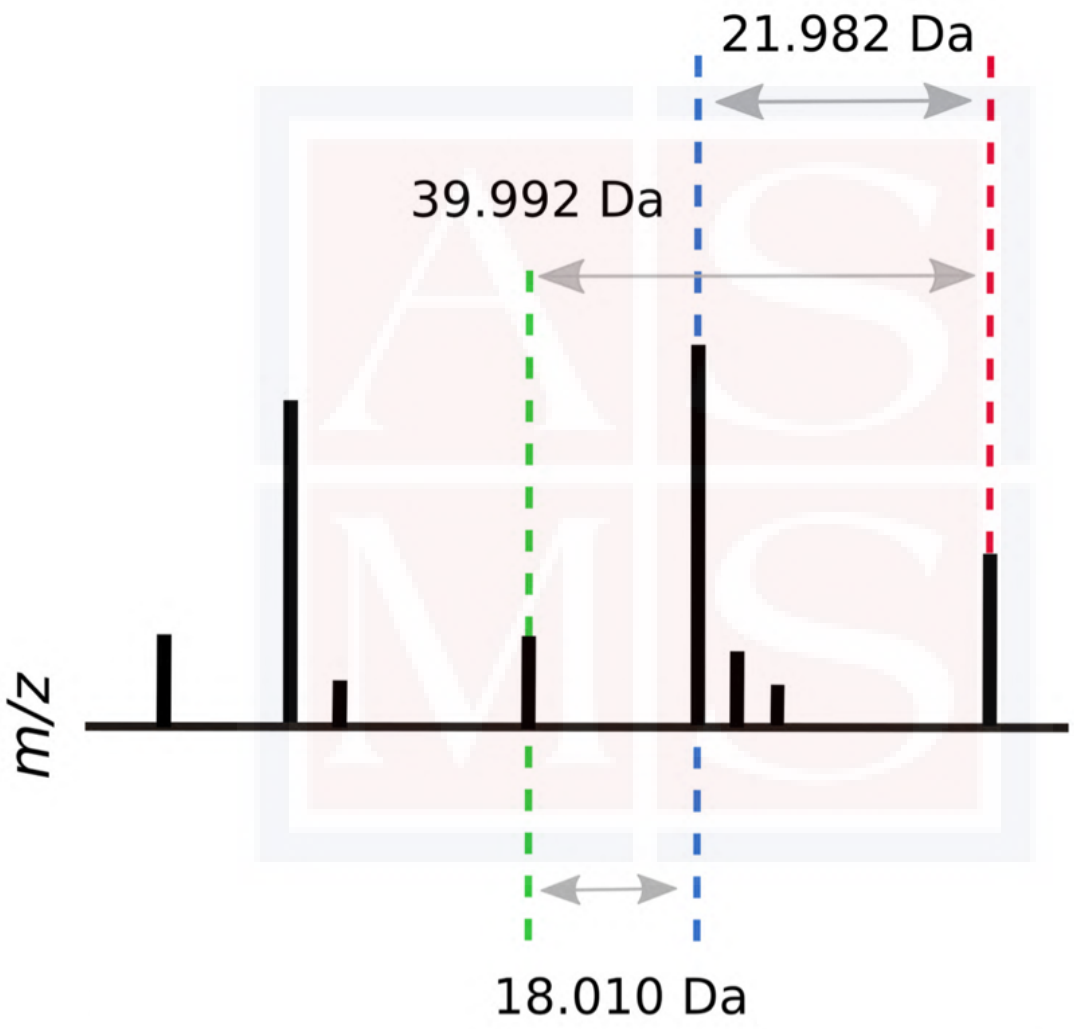
MS^1 pseudo-spectra extraction

Peak Abundance Correlation

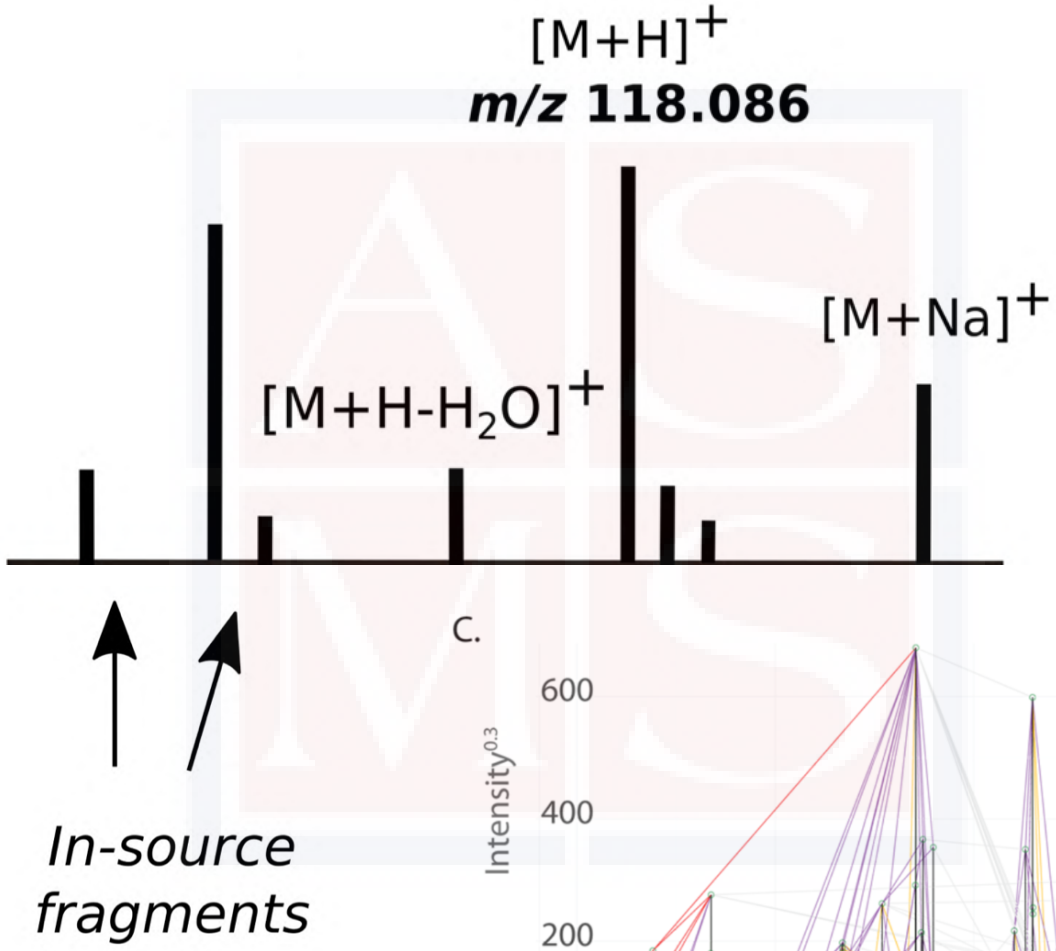


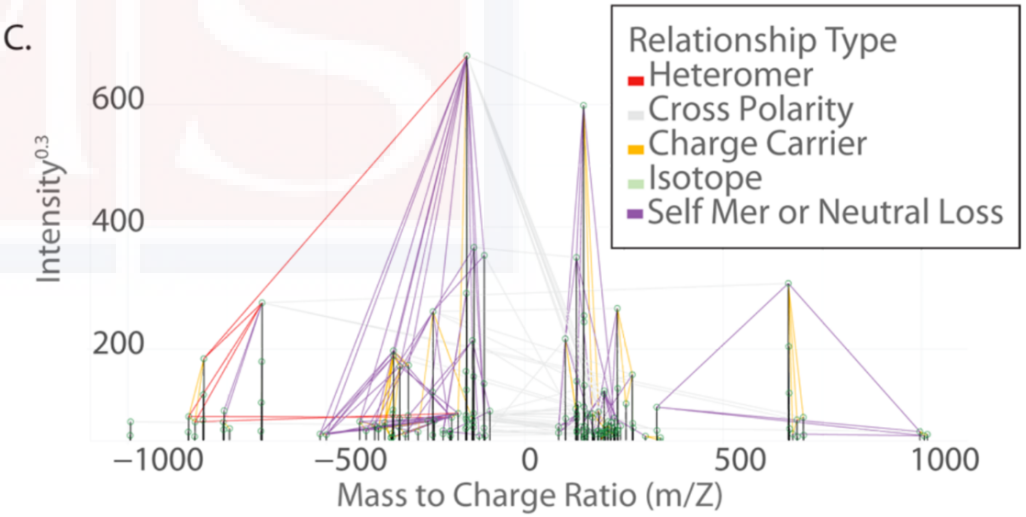
Picture from Domingo-Almenara et al., Anal. Chem., 90 (2018)

Adduct mass rules



Adduct mass rules

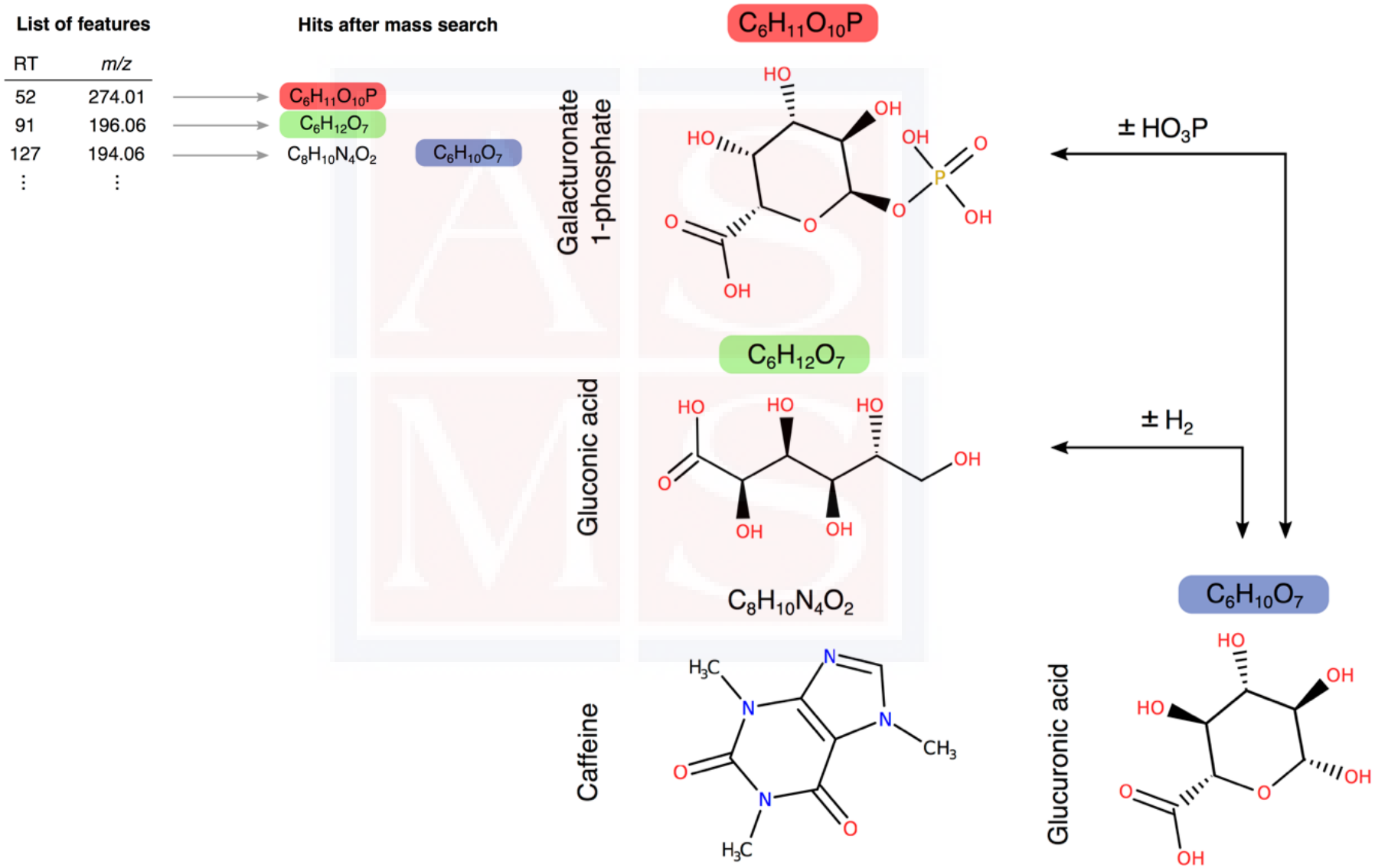




N. G. Mahieu *et. al.* **Defining and Detecting Complex Peak Relationships in Mass SpectralData: The Mz.unity Algorithm.** *Anal. Chem. 88 (2016) 9037-9046*

Biochemical knowledge

Biotransformations



Picture from Domingo-Almenara et al., Anal. Chem., 90 (2018)

Biochemical knowledge

Projection onto pathways

Assumption: if a list of putative identifications are “true”, these should reflect a biological activity and thus show an enrichment on local pathway regions

Li et al., PLoS Comput. Bio., 9 (2013) e1003123
Huan et al., Nat. Meth., 14 (2017) 461-462

Biochemical knowledge

Projection onto pathways

Feature List
10,000 features

Library search

Metabolite list
100,000 metabolites???

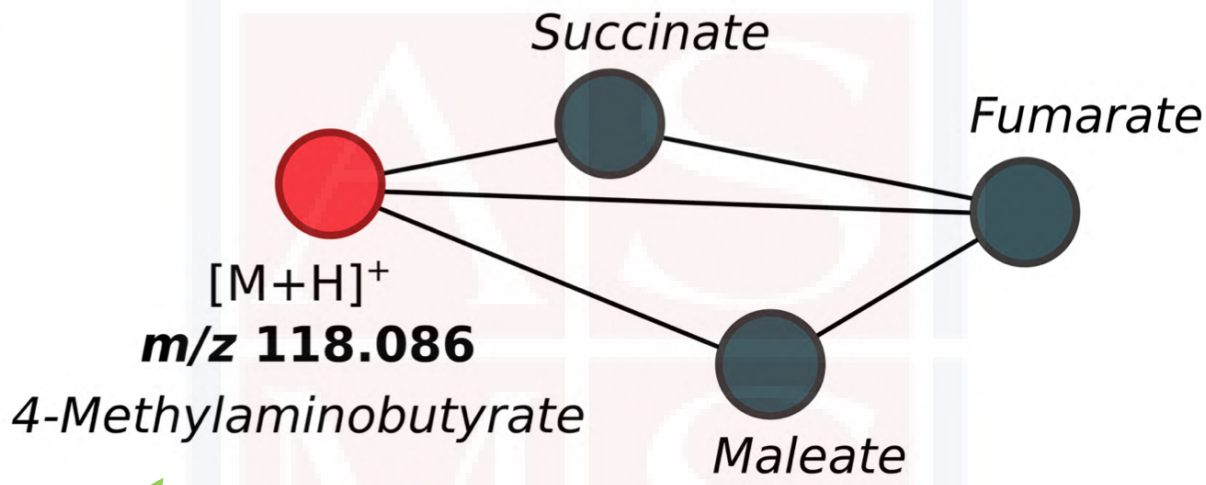
mzRT M118T56: 118.086

Betaine, *4-Methylamino-butyrate*, *Valine*, *Norvaline*,…

Hypothesis: *4-Methylamino-butyrate* ***?***

Biochemical knowledge

Projection onto pathways



mzRT M118T56

Biochemical knowledge Projection onto pathways

Feature List
10,000 features

Library search

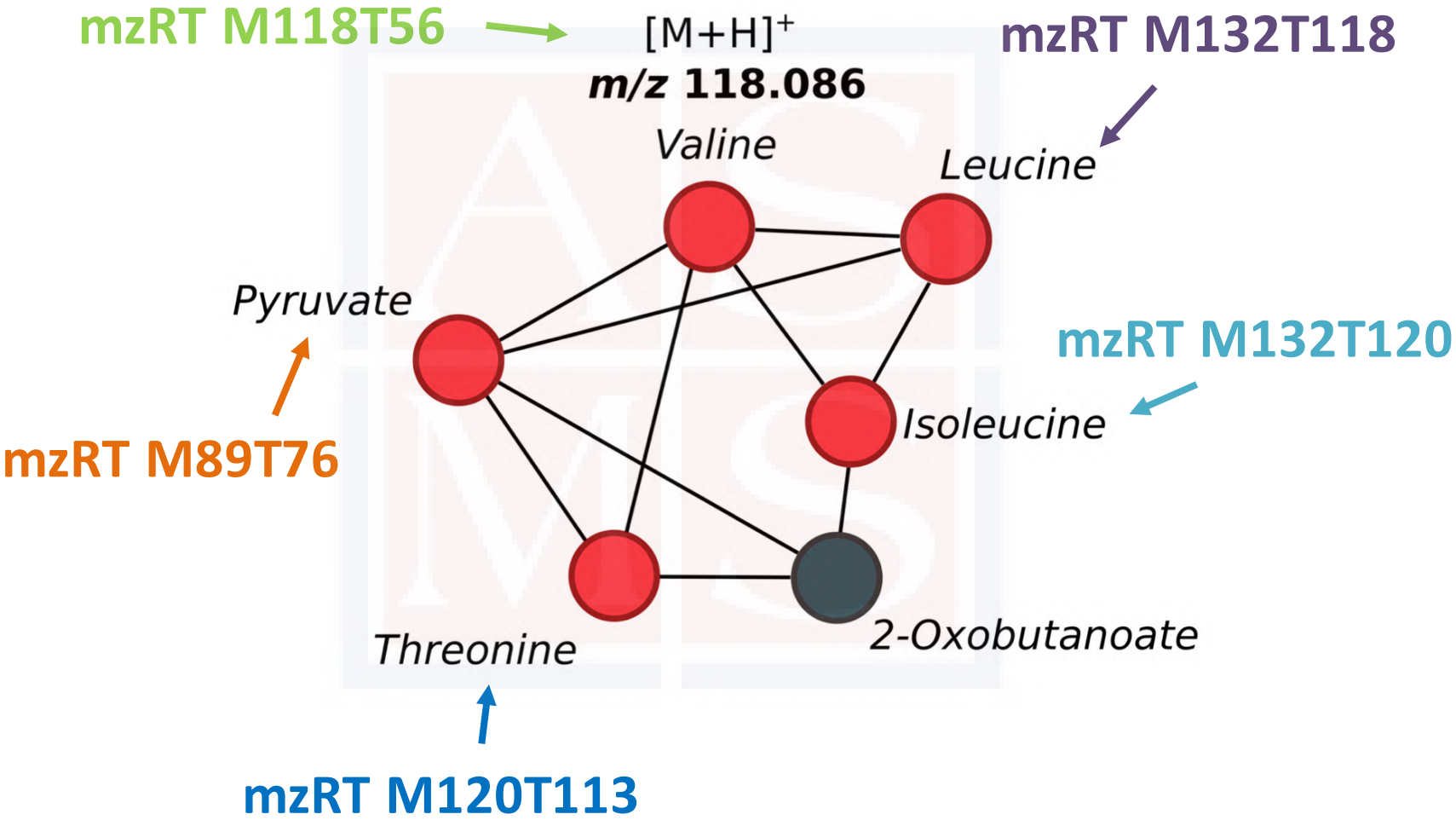
Metabolite list
100,000 metabolites???

mzRT M118T56: 118.086

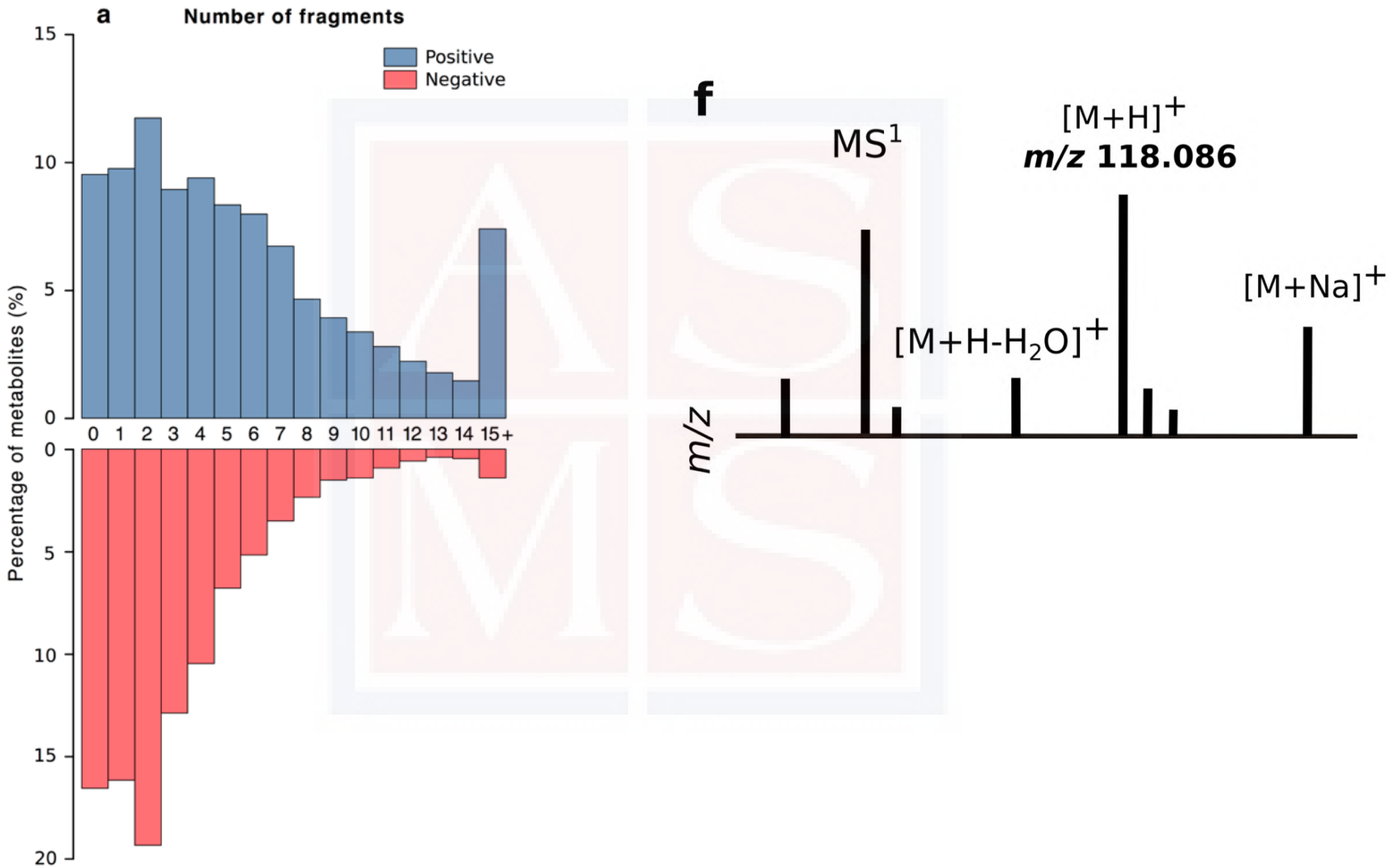
Betaine, *4-Methylamino-butyrate*, *Valine*, *Norvaline*,…

Hypothesis: Valine ?

Biochemical knowledge Projection onto pathways

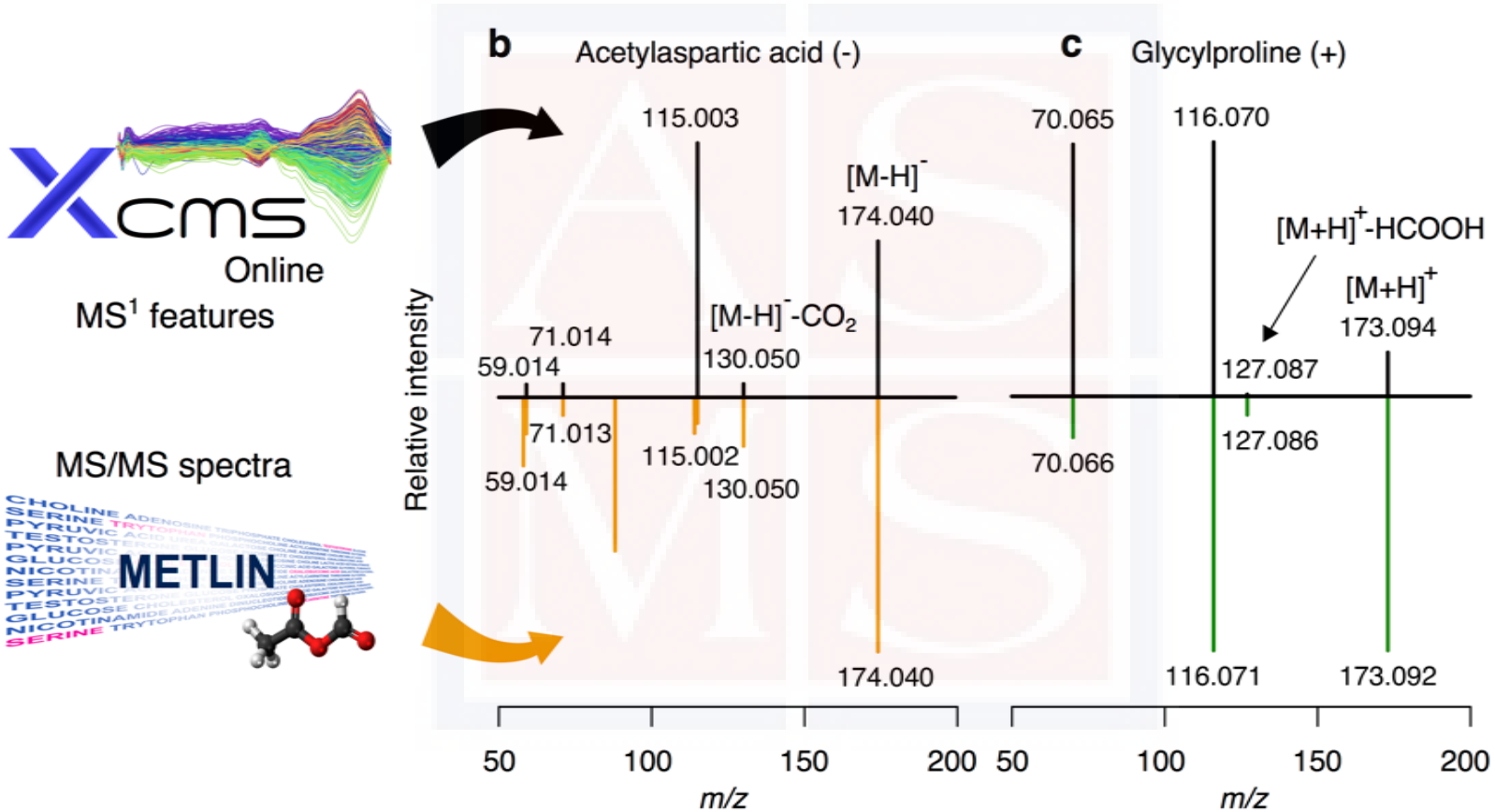


Use and integration of tandem MS data

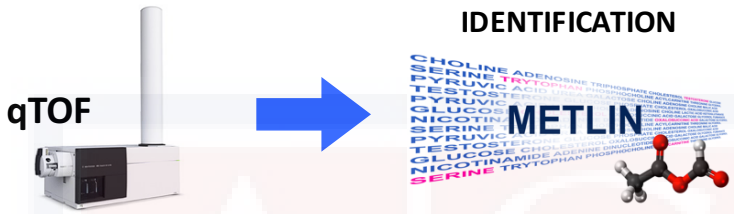


Picture from Domingo-Almenara et al., Anal. Chem., 90 (2018)

Use and integration of tandem MS data



Use and integration of tandem MS data

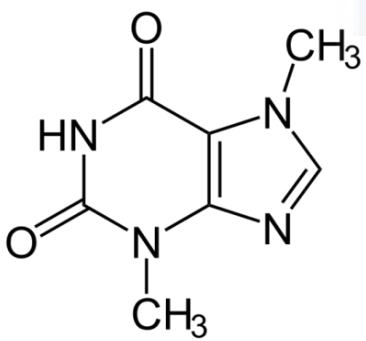


Experimental libraries… only 5% of MS/MS spectra

hmdb
The Human Metabolome Database

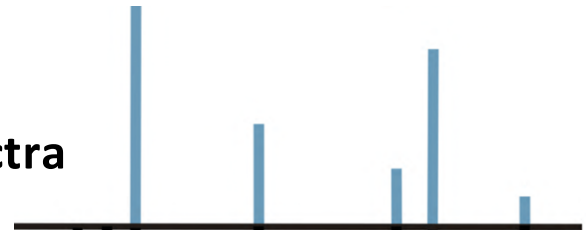
METLIN

Alternative… *in silico* prediction



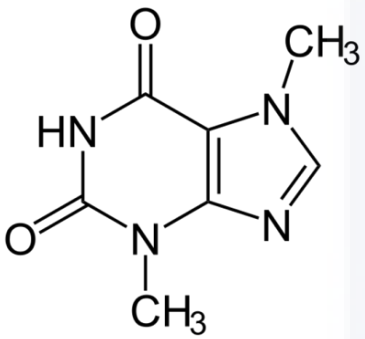
Structure

***In silico* MS/MS spectra**

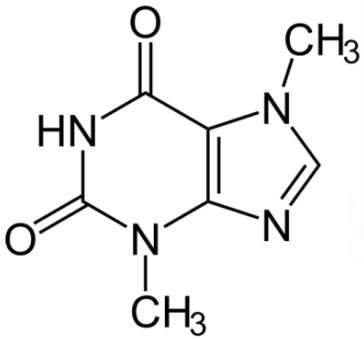


Use and integration of tandem MS data

Alternative… *in silico* prediction



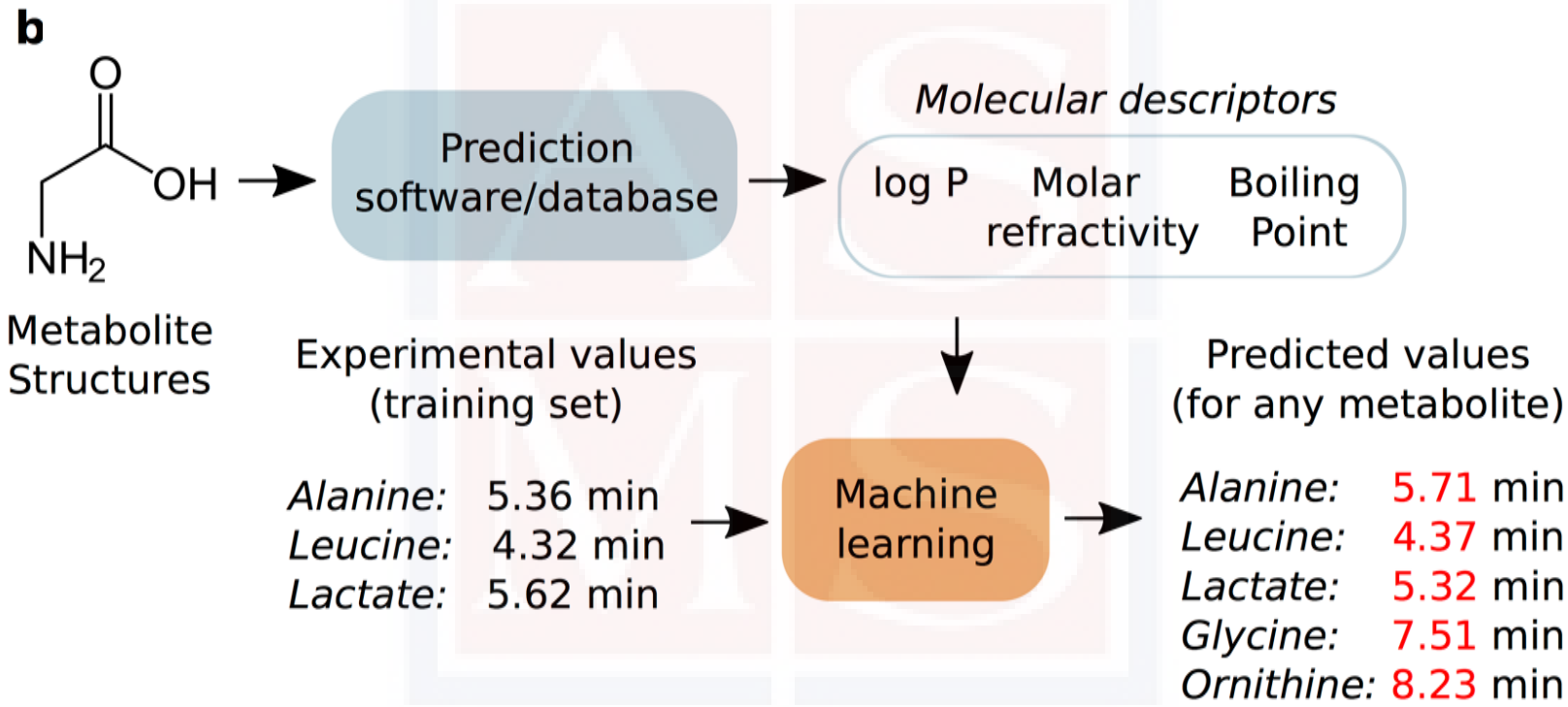
Alternative… spectral characterization or *de novo* identification



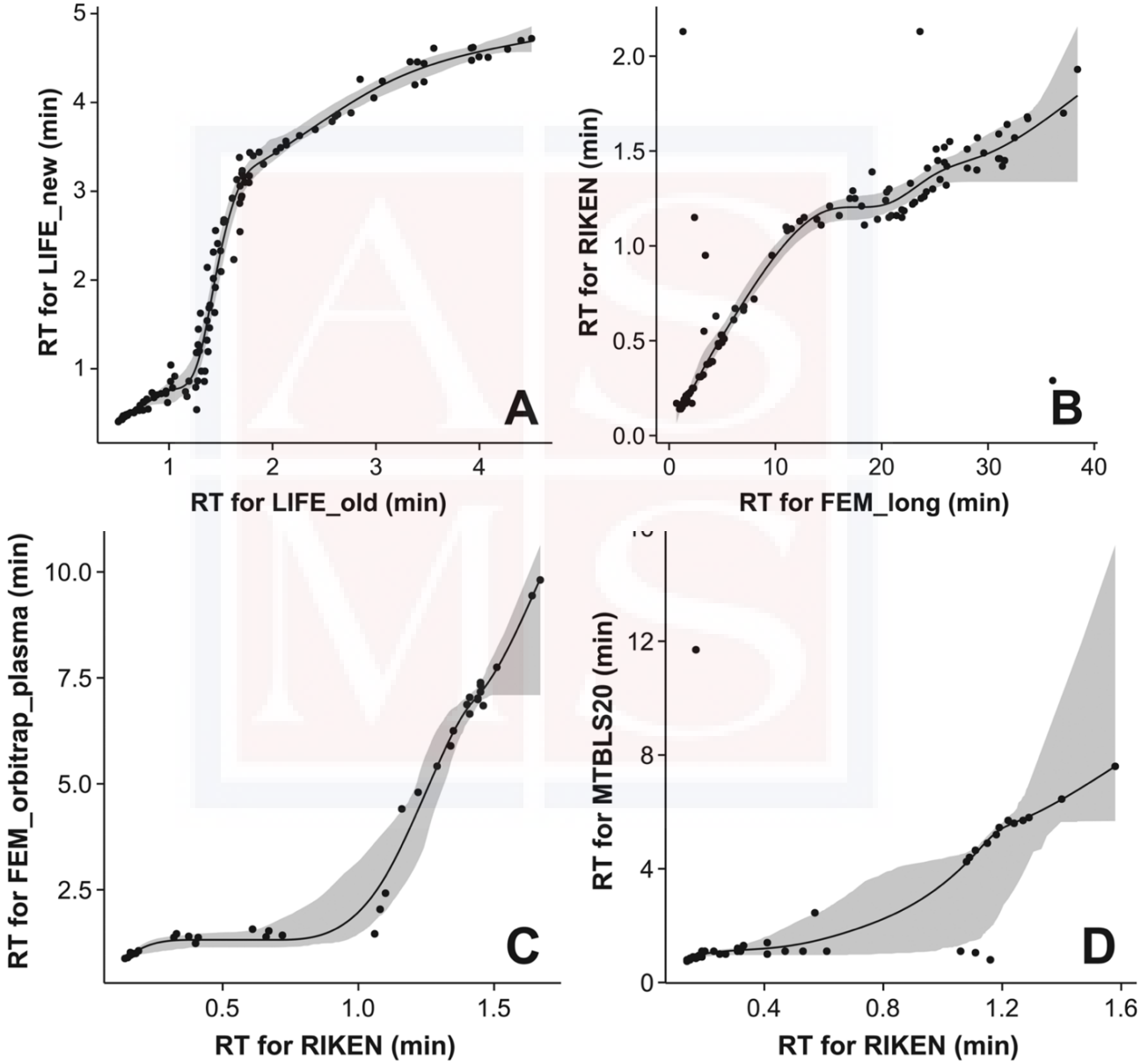
MS-FINDER, CSI:FinderID, iMet…

Tsugawa, et al. Anal. Chem. 2016, 88, 7946-7958
Dührkop, et al. Proc. Natl. Acad. Sci. U. S. A. 2015, 112, 12580–12585
Aguilar-Mogas, et al., Anal. Chem. 2017, 89, 3474-3482

Retention time calibration

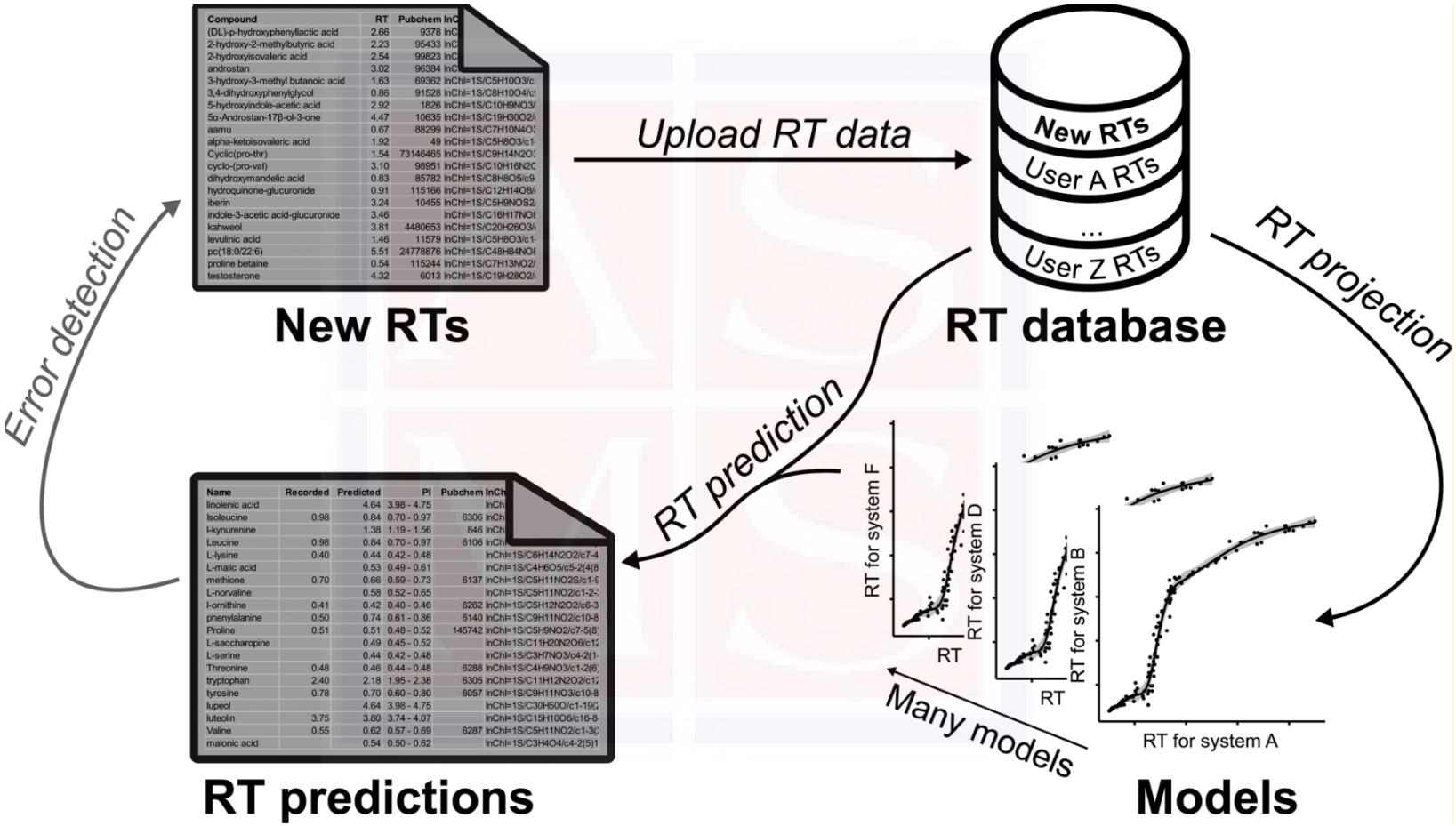


Retention time calibration



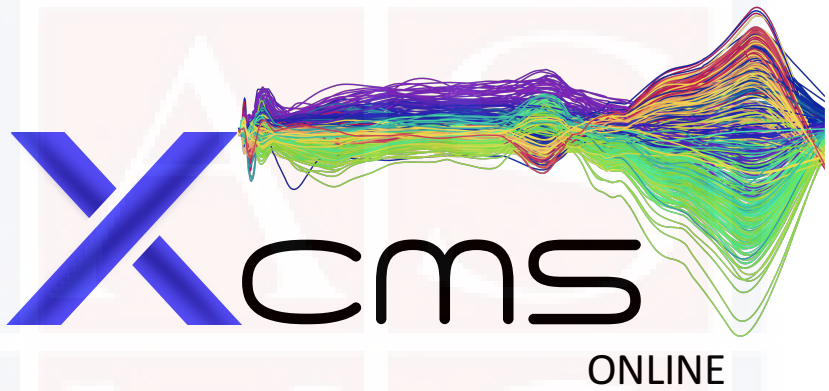
Picture from Stanstrup et al., Anal. Chem. 2015, 87, 9421–9428

Retention time calibration



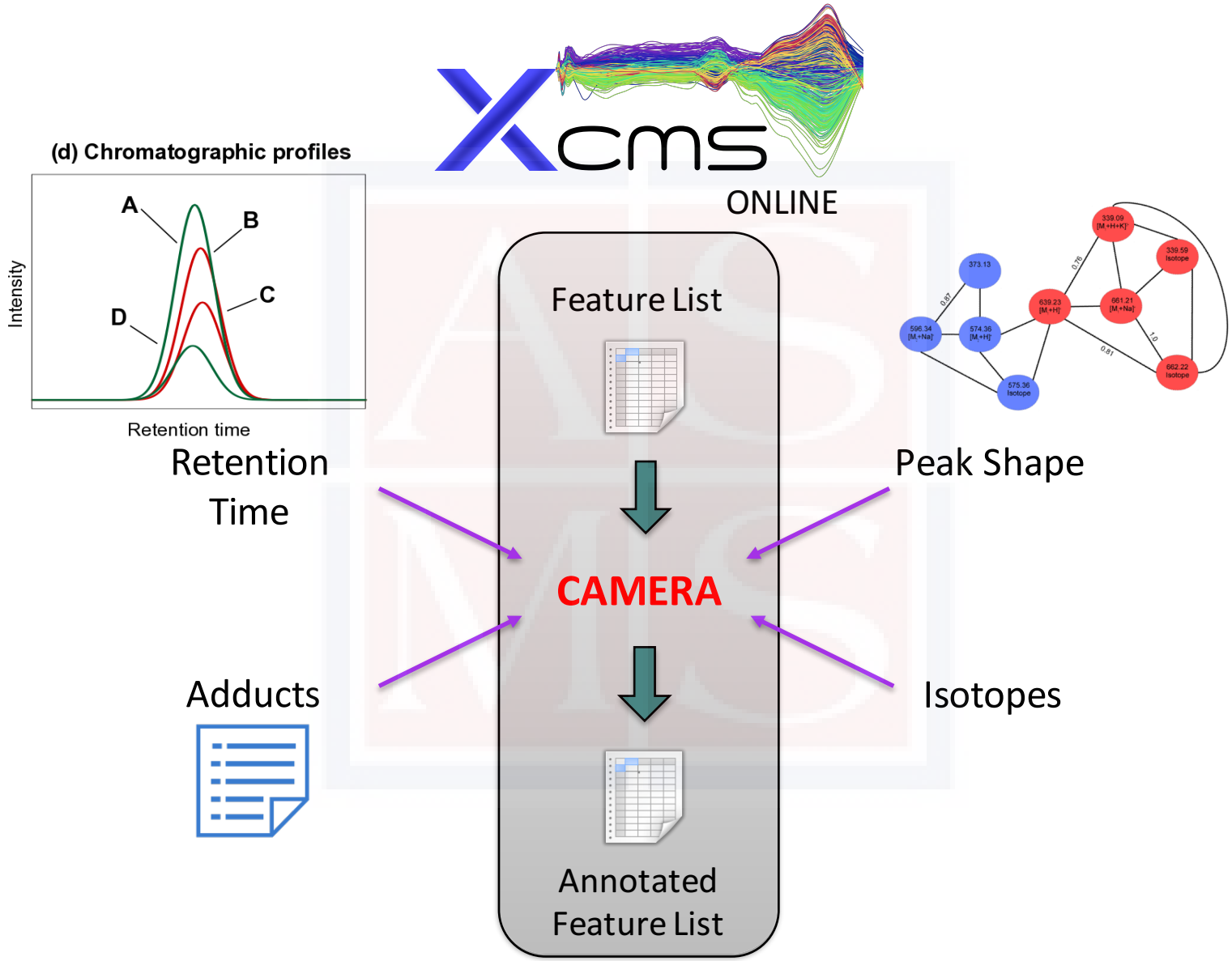
Picture from Stanstrup et al., Anal. Chem. 2015, 87, 9421–9428

Annotation with CAMERA



The R-package **CAMERA** is a **C**ollection of **A**lgorithms for **ME**tabolite p**R**ofile **A**nnotation

Annotation with XCMS: CAMERA



Results

JOB#1129809 : E COLI TEST 11

Columns Hide isotopic peaks Page 1 of 77 100 View 1 - 100 of 7,655

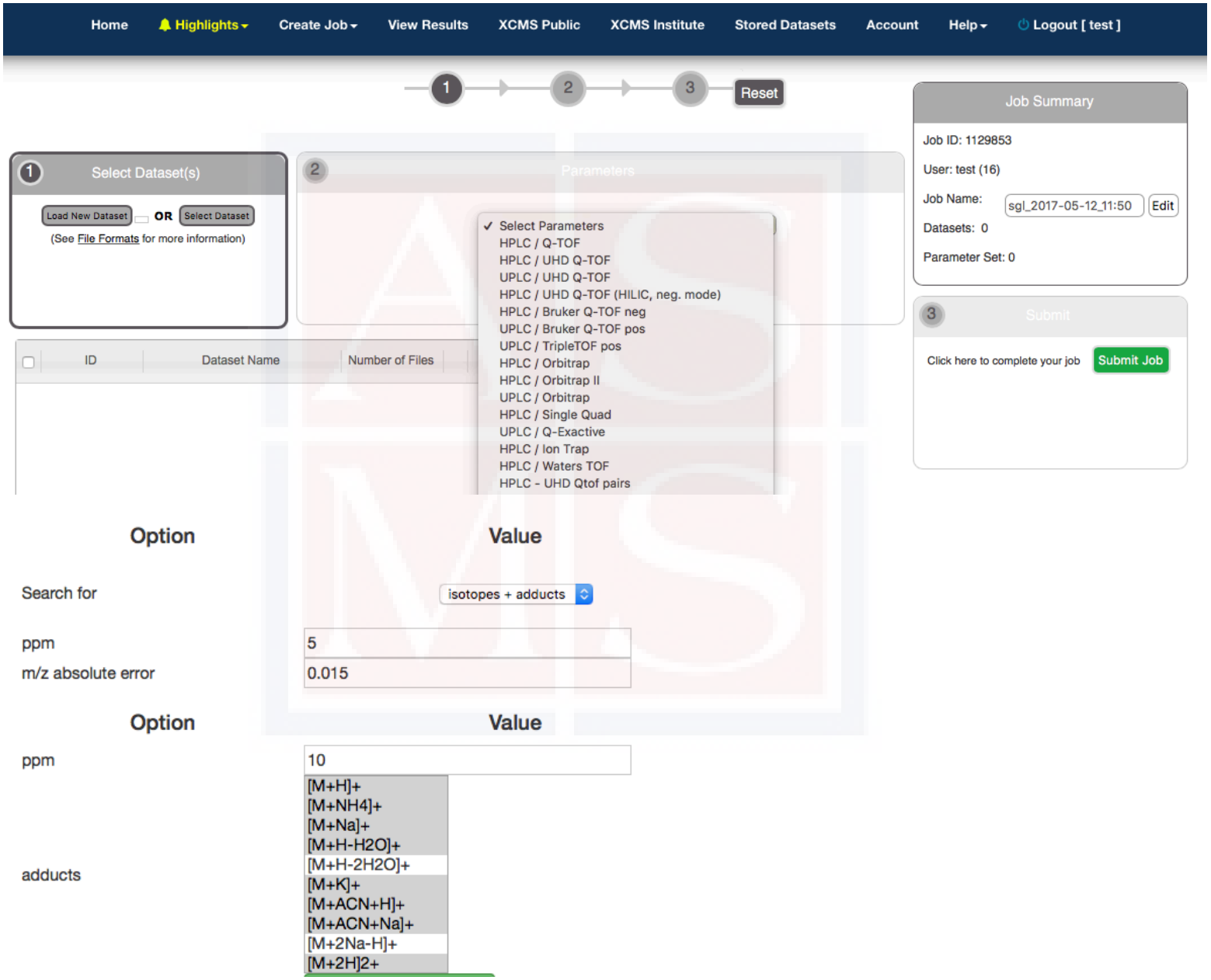
featureidx	fold	pvalue	updown	mzmed	rtmed	maxint	dataset1_mean	dataset2_mean	isotopes	adducts	peakgroup	usernotes
1	6.5	4.28435e-12	DOWN	274.1407	29.51	4,720	41,437	6,342			417	
2	4.3	5.08127e-12	DOWN	88.0403	22.31	2,586	29,222	6,871			45	
3	2.1	5.42721e-12	DOWN	587.0298	34.00	4,156	47,410	23,072	[513][M]2-		141	
4	4.4	9.90963e-12	DOWN	885.2532	32.88	4,562	71,713	16,228			91	
5	3.3	1.10874e-11	UP	135.0299	24.12	56,120	272,828	905,768	[34][M]-		154	
6	3.1	1.54532e-11	DOWN	628.0571	33.50	1,522	14,656	4,677		[M-2H+Na]- 607.0	35	
7	10.7	1.55786e-11	DOWN	145.0505	21.53	15,408	193,164	18,093	[46][M]-	[M-H]- 146.058 [M	24	
8	4.7	2.52691e-11	DOWN	885.7549	32.87	3,562	52,349	11,116			91	
9	4.6	3.40119e-11	DOWN	886.2566	32.87	2,166	28,494	6,186	[743][M]2-		91	
10	2.5	3.62674e-11	DOWN	965.2527	33.48	804	5,940	2,421			267	
11	2.4	4.41243e-11	DOWN	202.0722	32.62	6,326	81,181	33,351		[M-H]- 203.081	168	
12	3.6	6.10290e-11	DOWN	486.2681	26.35	3,580	45,341	12,541		[M+Cl]- 451.299 [M	286	
13	3.2	1.24508e-10	DOWN	607.0775	33.49	5,040	57,145	17,693	[522][M]-	[M+Cl]- 572.109 [M	267	
14	183.3	1.28438e-10	UP	135.0315	12.70	8,302,420	3,556,297	651,816,133	[35][M]-	[M-H]- 136.044	1	
15	6.2	1.39947e-10	DOWN	370.9553	21.23	1,140	10,011	1,623		[M-H-H20]- 389.97	28	
16	123.2	1.89734e-10	UP	135.0313	11.99	8,302,420	3,031,654	373,359,453	[38][M]-		2	
17	3.4	2.03586e-10	DOWN	608.0791	33.49	1,420	16,128	4,705	[522][M+1]-		267	
18	4.5	2.93537e-10	DOWN	967.7846	32.99	2,364	27,631	6,106			41	
19	2.6	3.98879e-10	DOWN	875.7411	33.41	2,242	19,723	7,707			145	
20	32.1	4.18445e-10	DOWN	210.0385	32.77	10,142	135,679	4,232		[M-2H+Na]- 189.0	21	
21	3.0	4.39714e-10	UP	232.1191	22.55	1,580	3,983	11,851			468	
22	5.5	4.72939e-10	UP	273.1204	29.08	1,282	1,777	9,738		[M-H]- 274.125	221	
23	7.9	5.26189e-10	DOWN	133.0506	20.84	44,504	604,836	76,256	[31][M]-		15	
24	14.2	5.76048e-10	DOWN	246.0028	17.88	3,258	44,658	3,156		[M-H]- 247.011	31	
25	4.1	6.23145e-10	DOWN	886.7567	32.88	1,054	11,046	2,688	[743][M+1]2-		91	
26	2.7	7.30410e-10	UP	106.0226	23.90	3,242	27,271	73,488	[12][M+1]-		174	
27	45.0	7.52775e-10	UP	136.0340	12.66	659,846	917,843	41,307,417	[35][M+1]-		1	
28	5.4	9.00549e-10	DOWN	526.2424	29.94	12,606	113,847	21,215	[462][M]-		275	
29	3.3	9.20177e-10	DOWN	606.0746	33.50	20,694	273,705	83,271		[M-H]- 607.082	35	
30	2.2	1.00253e-9	DOWN	541.3383	21.23	2,838	51,024	23,575			28	
31	3.5	1.06396e-9	UP	1,068.3862	41.51	2,198	5,016	17,787	[852][M+1]-		265	

Columns Export Page 1 of 77 100 View 1 - 100 of 7,655

Results

id	mz	rt	isotopes	adduct	pc
65	176.04	280.09			4
76	136.05	280.43	[14][M+1]1+		5
77	135.05	280.43	[14][M]1+		5
74	153.06	280.43		[M+H]+ 152.05437	5
75	175.04	280.43		[M+Na]+ 152.05437	5
73	197.02	280.76		[M+2Na-H]+ 152.05437	5
78	377.74	286.15			6
79	732.5	286.49			6
83	488.32	286.82		[M+Na]+ 465.33205	7
82	466.34	286.82		[M+H]+ 465.33205	7
...					

Parameters


Home
Highlights
Create Job
View Results
XCMS Public
XCMS Institute
Stored Datasets
Account
Help
Logout [test]
1
2
3
Reset
Job Summary
Job ID: 1129853
User: test (16)
Job Name:
sgl_2017-05-12_11:50
Edit
Datasets: 0
Parameter Set: 0
1
Select Dataset(s)
Load New Dataset
OR
Select Dataset
(See File Formats for more information)
2
Parameters
Select Parameters
HPLC / Q-TOF
HPLC / UHD Q-TOF
UPLC / UHD Q-TOF
HPLC / UHD Q-TOF (HILIC, neg. mode)
HPLC / Bruker Q-TOF neg
UPLC / Bruker Q-TOF pos
UPLC / TripleTOF pos
HPLC / Orbitrap
HPLC / Orbitrap II
UPLC / Orbitrap
HPLC / Single Quad
UPLC / Q-Exactive
HPLC / Ion Trap
HPLC / Waters TOF
HPLC - UHD Qtof pairs
ID
Dataset Name
Number of Files
3
Submit
Click here to complete your job
Submit Job
Option
Value
Search for
isotopes + adducts
ppm
5
m/z absolute error
0.015
Option
Value
ppm
10
adducts
[M+H]+
[M+NH4]+
[M+Na]+
[M+H-H2O]+
[M+H-2H2O]+
[M+K]+
[M+ACN+H]+
[M+ACN+Na]+
[M+2Na-H]+
[M+2H]2+

Annotation with xMSannotator

analytical chemistry

Technical Note

pubs.acs.org/ac

xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data

Karan Uppal,[†] Douglas I. Walker,[†,‡] and Dean P. Jones*[,†]

```
> library(xMSannotator)
> library(openxlsx)

> metdRaw <- read.xlsx('XCMS.diffreport.MultiClass.xlsx’)

> multilevelannotation(dataA, mode='pos', outloc=getwd(),
num_nodes=4)
```

Help can be accessed through: `?multilevelannotation`

Annotation with Everest

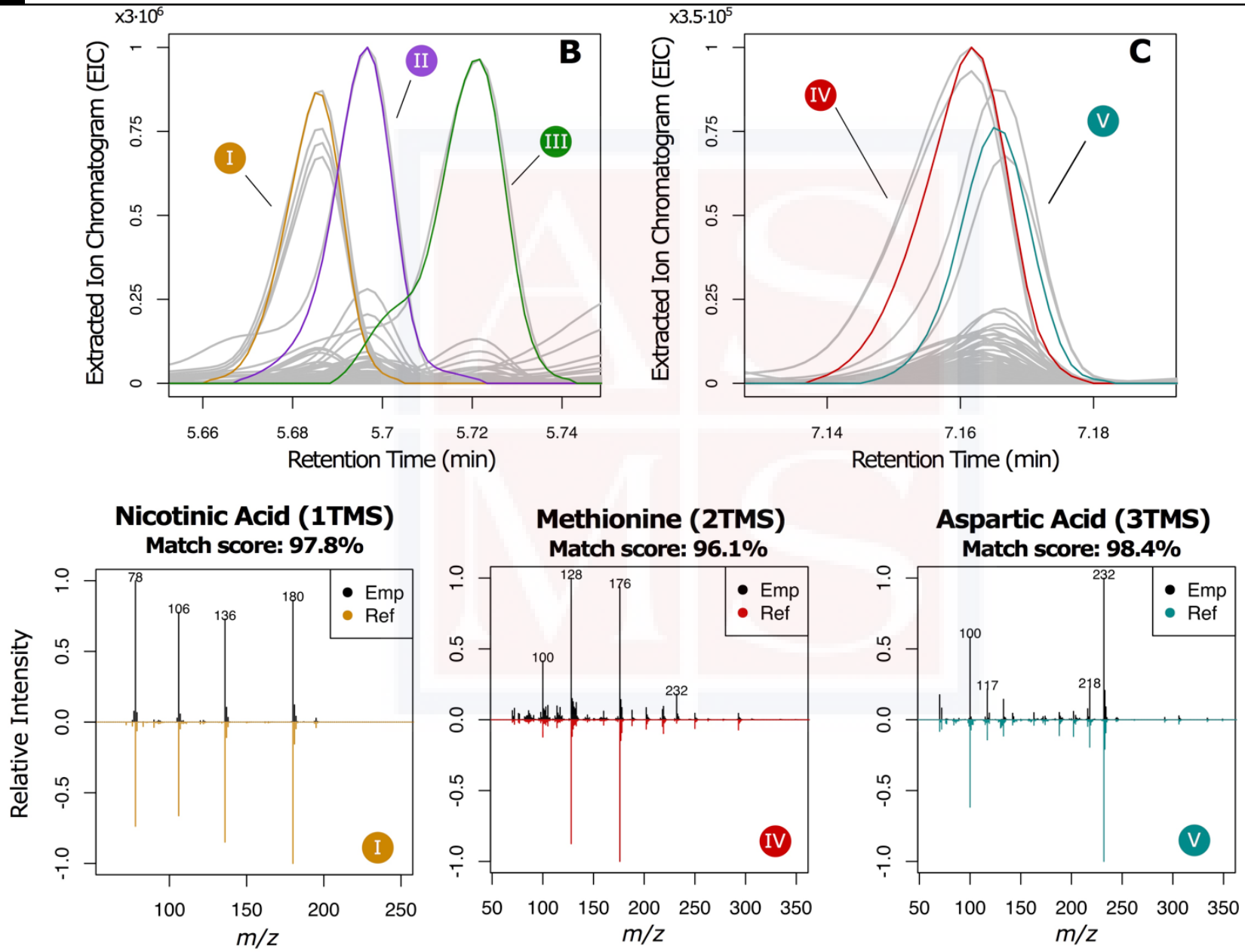
```
> library(everest)

> metdRaw <- read.xlsx('XCMS.diffreport.MultiClass.xlsx’)

> ex <- evAnnotate(xcmsSet=xset3, data.table=NULL,
ion.mode='pos', min.correlation=0.6, maz.time.dist=1)
> anTab <- annoTable(ex)
> write.xlsx(anTab, file='anResults.xlsx')
```

Help can be accessed through: `?evAnnotate`

`vignette("everestManual", package="everest”)`

row.names	mzmed	rtmed	Adduct	AnnID	AlignID	Isotope	toMSMS
M102T60	102.0904	60.158			144		
M162T60	162.1113	60.357	(M+H)+ 161.104076	1	144		yes
M163T60	163.1144	60.357	M+1 161.104076	1	144	yes	
M164T60	164.1164	60.354	M+2 161.104076	1	144	yes	
M184T60	184.0921	60.167	(M+Na)+ 161.104076	1	144		
M110T268	110.034	268.3875			413		
M128T268	128.0443	268.118			413		
M134T268	134.045	268.0125	(M+H)+[-H2O] 151.048623	1	413		
M135T268	135.0293	268.011	[M+H-NH3]+ 151.048623	1	413		
M136T268	136.0327	268.0125	M+1 151.048623	1	413	yes	
M137T269	137.045	268.5305			413		
M138T269	138.0476	268.5305			413		
M139T269	139.0492	268.516			413		
M152T268	152.0556	267.954	(M+H)+ 151.048623	1	413		yes
M153T268	153.0577	267.954	M+1 151.048623	1	413	yes	
M154T268	154.0207	268.437			413		
M193T269	193.0012	268.638			413		
M209T269	208.9737	268.5245			413		
M216T269	216.0172	268.6355			413		
M221T269	220.9965	268.5705			413		
M226T269	225.9767	268.504			413		
M232T269	231.9895	268.633			413		
M234T269	233.9895	268.65			413		
M237T268	236.968	268.49			413		
M238T269	237.9856	268.5705			413		
M250T269	249.9765	268.634			413		
M254T269	253.9708	268.6245			413		
M263T269	262.9841	268.5705			413		
M264T268	263.9885	268.438			413		
M265T268	264.9871	268.408			413		
M269T269	269.0862	268.504	(M+H)+ 268.078924; (M+H)+[-H2O] 286.090024	2; 4	413		
M270T269	270.0889	268.5705	M+1 268.078924; M+1 286.090024	2; 4	413	yes	
M271T269	271.0907	268.545	M+2 268.078924; M+2 286.090024	2; 4	413	yes	
M278T268	277.9943	268.011			413		
M282T268	281.9657	268.497			413		
M284T268	284.097	267.963	(M+H)+ 283.089759	11	413		
M285T268	285.0995	267.954	M+1 283.089759	11	413	yes	
M286T268	286.1014	268.011	M+2 283.089759	11	413	yes	
M291T269	291.0675	268.662	(M+Na)+ 268.078924; (M+Na)+[-H2O] 286.090024	2; 4	413		
M292T269	292.0706	268.665	M+1 268.078924; M+1 286.090024	2; 4	413	yes	
M306T268	306.0781	267.937	(M+Na)+ 283.089759	11	413		
M307T268	307.0412	268.497	(M+K)+ 268.078924; (M+K)+[-H2O] 286.090024	2; 4	413		

Annotation in GC/MS



Picture from Domingo-Almenara et al., Anal. Chem., 88 (2016)

Thank you for your attention!

Questions?

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Pathway Analysis and Multi-Omic Integration

1. Prerequisites
2. Biomarkers vs. Biological Relevance
3. Pathway Tools for Annotated Data
4. Pathway Tools for Unannotated Data
5. Multi-Omic Integration

Pathway Analysis and Multi-Omic Integration

1. **Prerequisites**
2. Biomarkers vs. Biological Relevance
3. Pathway Tools for Annotated Data
4. Pathway Tools for Unannotated Data
5. Multi-Omic Integration

Pathway Analysis Prerequisites

Feature detection
Retention time alignment
Statistical analysis

Pathway Analysis Prerequisites

Feature detection
Retention time alignment
Statistical analysis

Pathway Analysis Prerequisites

fold	log2fold	tstat	pvalue	qvalue	updown	mzmed
271.094972	8.08265455	169.812117	6.6221E-08	0.00054666	UP	544.456801
59.6958696	5.89955921	45.1775663	4.8957E-06	0.00721752	UP	559.43644
20.7924201	4.					493.350039
16.5527842	4.					495.356237
20.2312615	4.					494.353508
26.6099804	4.					510.3276
5.66123563	2.					573.491742
30.1386729	4.					511.32522
24.7943117	-4.6319373	-34.711263	6.1082E-05	0.02653869	DOWN	465.318744
2.44206375	1.28810086	17.5080138	0.00010347	0.03502456	UP	629.474016
158.373415	7.30718637	76.41158	0.00012401	0.03791842	UP	531.404284
335.205024	8.38889996	64.9274058	0.00018207	0.0443508	UP	500.306416
22.5169447	4.49293918	30.2821087	0.00021427	0.04713741	UP	572.487987
18.6838461	4.22371956	1[illegible]6569615	0.00023106	0.04838631	UP	490.400591
[illegible]	[illegible]	[illegible]552	0.00024929	0.04961304	UP	558.436677
[illegible]	[illegible]	[illegible]807	0.00024929	0.0		
[illegible]	[illegible]	[illegible]601	0.00030816	0.0		
[illegible]	[illegible]	[illegible]415	0.00031404	0.0		
[illegible]	[illegible]	[illegible]962	0.0003933	0.0		
[illegible]	[illegible]	[illegible]575	0.00052482	0.0		
[illegible]	[illegible]	[illegible]644	0.00062048	0.0		
35.1611492	5.13591032	23.5738056	0.00062429	0.06877211	UP	488.394625
348.642181	8.44560332	38.8693701	0.00063927	0.06922533	UP	593.417786
34.8260631	5.12209549	24.3373088	0.00071137	0.07120843	UP	550.41601
2.81883619	-1.4950996	-13.764555	0.00075704	0.07231659	DOWN	329.316413
391.572992	8.61313745	35.6765283	0.00078474	0.0729406	UP	768.571762

Feature detection
Retention time alignment
Statistical analysis

XCMS Online

MZmine

Pathway Analysis Prerequisites

fold	log2fold	tstat	pvalue	qvalue	updown	mzmed
271.094972	8.08265455	169.812117	6.6221E-08	0.00054666	UP	544.456801
59.6958696	5.89955921	45.1775663	4.8957E-06	0.00721752	UP	559.43644
20.7924201	4.37798578	33.1607101	5.2453E-06	0.00729893	UP	493.350039
16.5527842	4.04900199	33.3355079	5.4782E-06	0.00734826	UP	495.356237
20.2312615	4.33851437	29.9435847	7.4163E-06	0.00765272	UP	494.353508
26.6099804	4.73389554	25.473017	2.6633E-05	0.01738435	UP	510.3276
5.66123563	2.50111697	24.96156	2.6763E-05	0.01742602	UP	573.491742
30.1386729	4.91354399	22.2731985	3.8769E-05	0.02053971	UP	511.32522
24.7943117	-4.6319373	-34.711263	6.1082E-05	0.02653869	DOWN	465.318744
2.44206375	1.28810086	17.5080138	0.00010347	0.03502456	UP	629.474016
158.373415	7.30718637	76.41158	0.00012401	0.03791842	UP	531.404284
335.205024	8.38889996	64.9274058	0.00018207	0.0443508	UP	500.306416
22.5169447	4.49293918	30.2821087	0.00021427	0.04713741	UP	572.487987
18.6838461	4.22371956	15.6569615	0.00023106	0.04838631	UP	490.400591
63.9046097	5.9978481	26.040552	0.00024929	0.04961304	UP	558.436677
24.1278502	4.59262747	23.8607807	0.00024929	0.04961313	UP	516.425919
358.44269	8.48559866	54.7628601	0.00030816	0.05364511	UP	658.524971
1785.67658	10.8022551	56.4160415	0.00031404	0.05401676	UP	726.022636
36.9654994	5.2081075	24.0682962	0.0003933	0.05907364	UP	566.389999
4.91567968	2.29739091	22.1105575	0.00052482	0.06531191	UP	626.534683
201.583154	7.65523127	40.0821644	0.00062048	0.06865423	UP	578.447293
35.1611492	5.13591032	23.5738056	0.00062429	0.06877211	UP	488.394625
348.642181	8.44560332	38.8693701	0.00063927	0.06922533	UP	593.417786
34.8260631	5.12209549	24.3373088	0.00071137	0.07120843	UP	550.41601
2.81883619	-1.4950996	-13.764555	0.00075704	0.07231659	DOWN	329.316413
391.572992	8.61313745	35.6765283	0.00078474	0.0729406	UP	768.571762

Pathway Analysis Prerequisites

fold	log2fold	tstat	pvalue	qvalue	updown	mzmed
271.094972	8.08265455	169.812117	6.6221E-08	0.00054666	UP	544.456801
59.6958696	5.89955921	45.1775663	4.8957E-06	0.00721752	UP	559.43644
20.7924201	4.37798578	33.1607101	5.2453E-06	0.00729893	UP	493.350039
16.5527842	4.04900199	33.3355079	5.4782E-06	0.00734826	UP	495.356237
20.2312615	4.33851437	29.9435847	7.4163E-06	0.00765272	UP	494.353508
26.6099804	4.73389554	25.473017	2.6633E-05	0.01738435	UP	510.3276
5.66123563	2.50111697	24.96156	2.6763E-05	0.01742602	UP	573.491742
30.1386729	4.91354399	22.2731985	3.8769E-05	0.02053971	UP	511.32522
24.7943117	-4.6319373	-34.711263	6.1082E-05	0.02653869	DOWN	465.318744
2.44206375	1.28810086	17.5080138	0.00010347	0.03502456	UP	629.474016
158.373415	7.30718637	76.41158	0.00012401	0.03791842	UP	531.404284
335.205024	8.38889996	64.9274058	0.00018207	0.0443508	UP	500.306416
22.5169447	4.49293918	30.2821087	0.00021427	0.04713741	UP	572.487987
18.6838461	4.22371956	15.6569615	0.00023106	0.04838631	UP	490.400591
63.9046097	5.9978481	26.040552	0.00024929	0.04961304	UP	558.436677
24.1278502	4.59262747	23.8607807	0.00024929	0.04961313	UP	516.425919
358.44269	8.48559866	54.7628601	0.00030816	0.05364511	UP	658.524971
1785.67658	10.8022551	56.4160415	0.00031404	0.05401676	UP	726.022636
36.9654994	5.2081075	24.0682962	0.0003933	0.05907364	UP	566.389999
4.91567968	2.29739091	22.1105575	0.00052482	0.06531191	UP	626.534683
201.583154	7.65523127	40.0821644	0.00062048	0.06865423	UP	578.447293
35.1611492	5.13591032	23.5738056	0.00062429	0.06877211	UP	488.394625
348.642181	8.44560332	38.8693701	0.00063927	0.06922533	UP	593.417786
34.8260631	5.12209549	24.3373088	0.00071137	0.07120843	UP	550.41601
2.81883619	-1.4950996	-13.764555	0.00075704	0.07231659	DOWN	329.316413
391.572992	8.61313745	35.6765283	0.00078474	0.0729406	UP	768.571762

Pathway Analysis Prerequisites

fold	log2fold	tstat	pvalue	qvalue	updown	mzmed
271.094972	8.08265455	169.812117	6.6221E-08	0.00054666	UP	544.456801
59.6958696	5.89955921	45.1775663	4.8957E-06	0.00721752	UP	559.43644
20.7924201	4.37798578	33.1607101	5.2453E-06	0.00729893	UP	493.350039
16.5527842	4.04900199	33.3355079	5.4782E-06	0.00734826	UP	495.356237
20.2312615	4.33851437	29.9435847	7.4163E-06	0.00765272	UP	494.353508
26.6099804	4.73389554	25.473017	2.6633E-05	0.01738435	UP	510.3276
5.66123563	2.50111697	24.96156	2.6763E-05	0.01742602	UP	573.491742
30.1386729	4.91354399	22.2731985	3.8769E-05	0.02053971	UP	511.32522
24.7943117	-4.6319373	-34.711263	6.1082E-05	0.02653869	DOWN	465.318744
2.44206375	1.28810086	17.5080138	0.00010347	0.03502456	UP	629.474016
158.373415	7.30718637	76.41158	0.00012401	0.03791842	UP	531.404284
335.205024	8.38889996	64.9274058	0.00018207	0.0443508	UP	500.306416
22.5169447	4.49293918	30.2821087	0.00021427	0.04713741	UP	572.487987
18.6838461	4.22371956	15.6569615	0.00023106	0.04838631	UP	490.400591
63.9046097	5.9978481	26.040552	0.00024929	0.04961304	UP	558.436677
24.1278502	4.59262747	23.8607807	0.00024929	0.04961313	UP	516.425919
358.44269	8.48559866	54.7628601	0.00030816	0.05364511	UP	658.524971
1785.67658	10.8022551	56.4160415	0.00031404	0.05401676	UP	726.022636
36.9654994	5.2081075	24.0682962	0.0003933	0.05907364	UP	566.389999
4.91567968	2.29739091	22.1105575	0.00052482	0.06531191	UP	626.534683
201.583154	7.65523127	40.0821644	0.00062048	0.06865423	UP	578.447293
35.1611492	5.13591032	23.5738056	0.00062429	0.06877211	UP	488.394625
348.642181	8.44560332	38.8693701	0.00063927	0.06922533	UP	593.417786
34.8260631	5.12209549	24.3373088	0.00071137	0.07120843	UP	550.41601
2.81883619	-1.4950996	-13.764555	0.00075704	0.07231659	DOWN	329.316413
391.572992	8.61313745	35.6765283	0.00078474	0.0729406	UP	768.571762

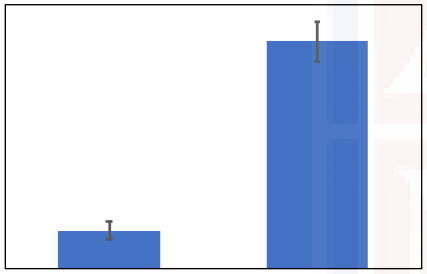
Pathway Analysis Prerequisites

fold	og2fold	tstat	pvalue	qvalue	updown	mzmed
271.094972	8.08265455	169.812117	6.6221E-08	0.00054666	UP	544.456801
59.6958696	5.89955921	45.1775663	4.8957E-06	0.00721752	UP	559.43644
20.7924201	4.37798578	33.1607101	5.2453E-06	0.00729893	UP	493.350039
16.5527842	4.04900199	33.3355079	5.4782E-06	0.00734826	UP	495.356237
20.2312615	4.33851437	29.9435847	7.4163E-06	0.00765272	UP	494.353508
26.6099804	4.73389554	25.473017	2.6633E-05	0.01738435	UP	510.3276
5.66123563	2.50111697	24.96156	2.6763E-05	0.01742602	UP	573.491742
30.1386729	4.91354399	22.2731985	3.8769E-05	0.02053971	UP	511.32522
24.7943117	-4.6319373	-34.711263	6.1082E-05	0.02653869	DOWN	465.318744
2.44206375	1.28810086	17.5080138	0.00010347	0.03502456	UP	629.474016
158.373415	7.30718637	76.41158	0.00012401	0.03791842	UP	531.404284
335.205024	8.38889996	64.9274058	0.00018207	0.0443508	UP	500.306416
22.5169447	4.49293918	30.2821087	0.00021427	0.04713741	UP	572.487987
18.6838461	4.22371956	15.6569615	0.00023106	0.04838631	UP	490.400591
63.9046097	5.9978481	26.040552	0.00024929	0.04961304	UP	558.436677
24.1278502	4.59262747	23.8607807	0.00024929	0.04961313	UP	516.425919
358.44269	8.48559866	54.7628601	0.00030816	0.05364511	UP	658.524971
1785.67658	10.8022551	56.4160415	0.00031404	0.05401676	UP	726.022636
36.9654994	5.2081075	24.0682962	0.0003933	0.05907364	UP	566.389999
4.91567968	2.29739091	22.1105575	0.00052482	0.06531191	UP	626.534683
201.583154	7.65523127	40.0821644	0.00062048	0.06865423	UP	578.447293
35.1611492	5.13591032	23.5738056	0.00062429	0.06877211	UP	488.394625
348.642181	8.44560332	38.8693701	0.00063927	0.06922533	UP	593.417786
34.8260631	5.12209549	24.3373088	0.00071137	0.07120843	UP	550.41601
2.81883619	-1.4950996	-13.764555	0.00075704	0.07231659	DOWN	329.316413
391.572992	8.61313745	35.6765283	0.00078474	0.0729406	UP	768.571762

Pathway Analysis and Multi-Omic Integration

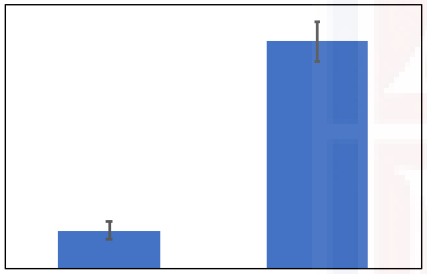
1. Prerequisites
2. **Biomarkers vs. Biological Relevance**
3. Pathway Tools for Annotated Data
4. Pathway Tools for Unannotated Data
5. Multi-Omic Integration

Biomarkers vs. Biological Relevance

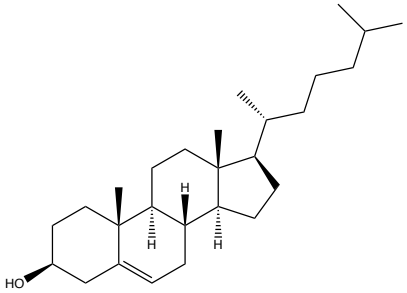


"A measurable substance in an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure."

Biomarkers vs. Biological Relevance

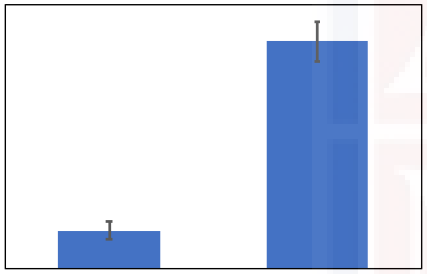


"A measurable substance in an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure."

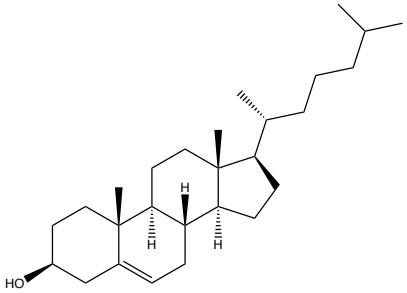


Cholesterol
→CVD

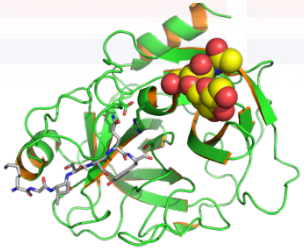
Biomarkers vs. Biological Relevance



"A measurable substance in an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure."

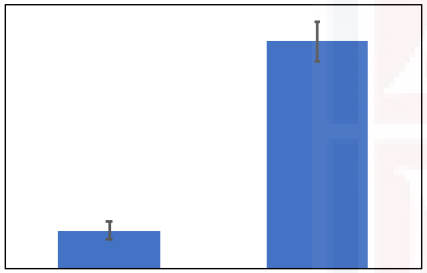


Cholesterol
→CVD



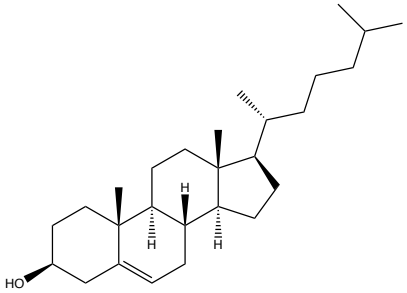
PSA
→Colon cancer

Biomarkers vs. Biological Relevance

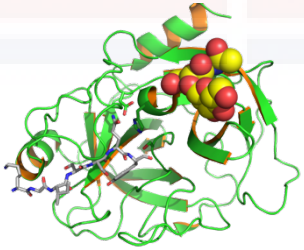


Healthy vs. Disease

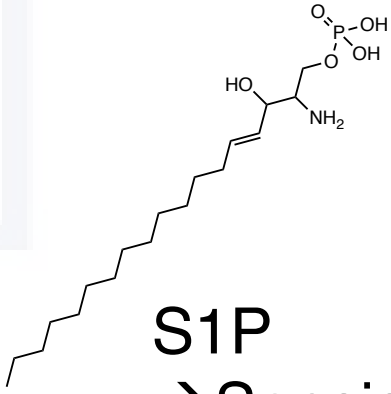
"A measurable substance in an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure."



Cholesterol
→CVD



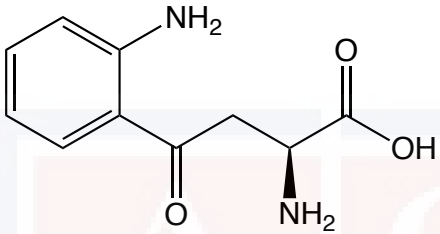
PSA
→Colon cancer

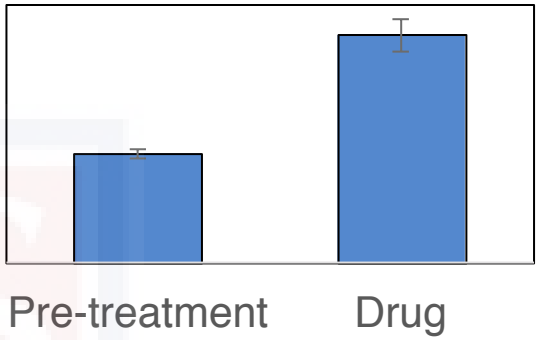


S1P
→Sepsis
→Alzheimer's
→Cancer…

Biomarkers vs. Biological Relevance

L-Kynurenine

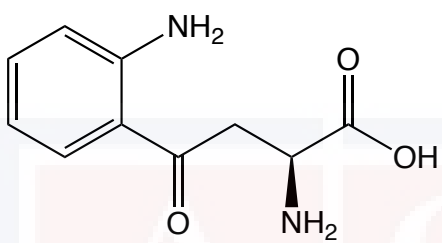


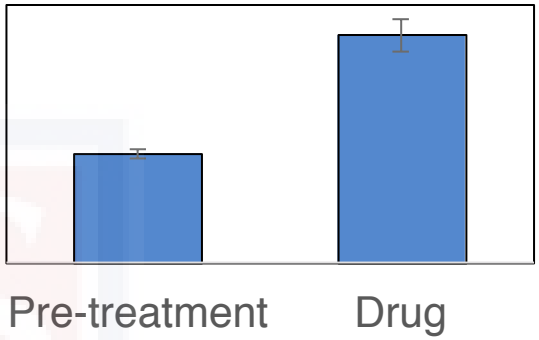


Gendelman et al NPJ Parkinson's Disease Journal 2017

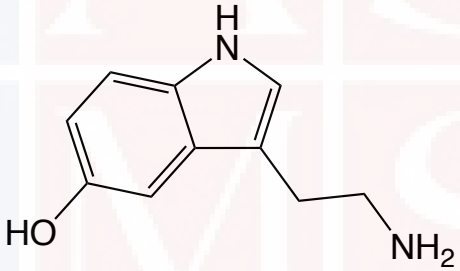
Biomarkers vs. Biological Relevance

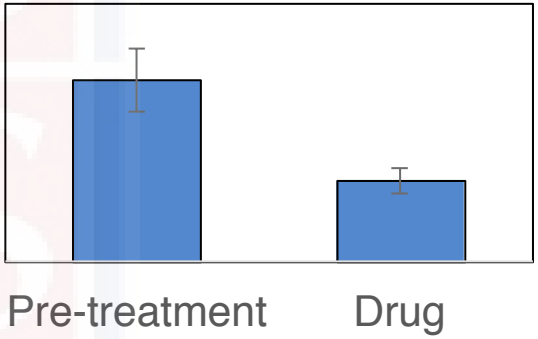
L-Kynurenine



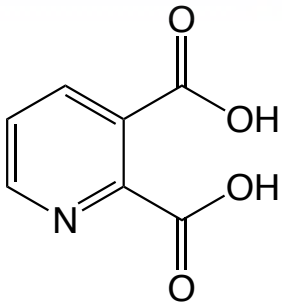


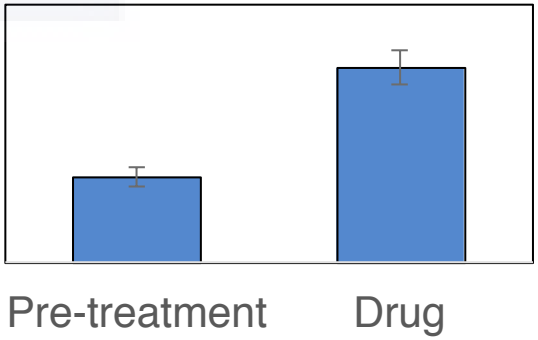
Serotonin





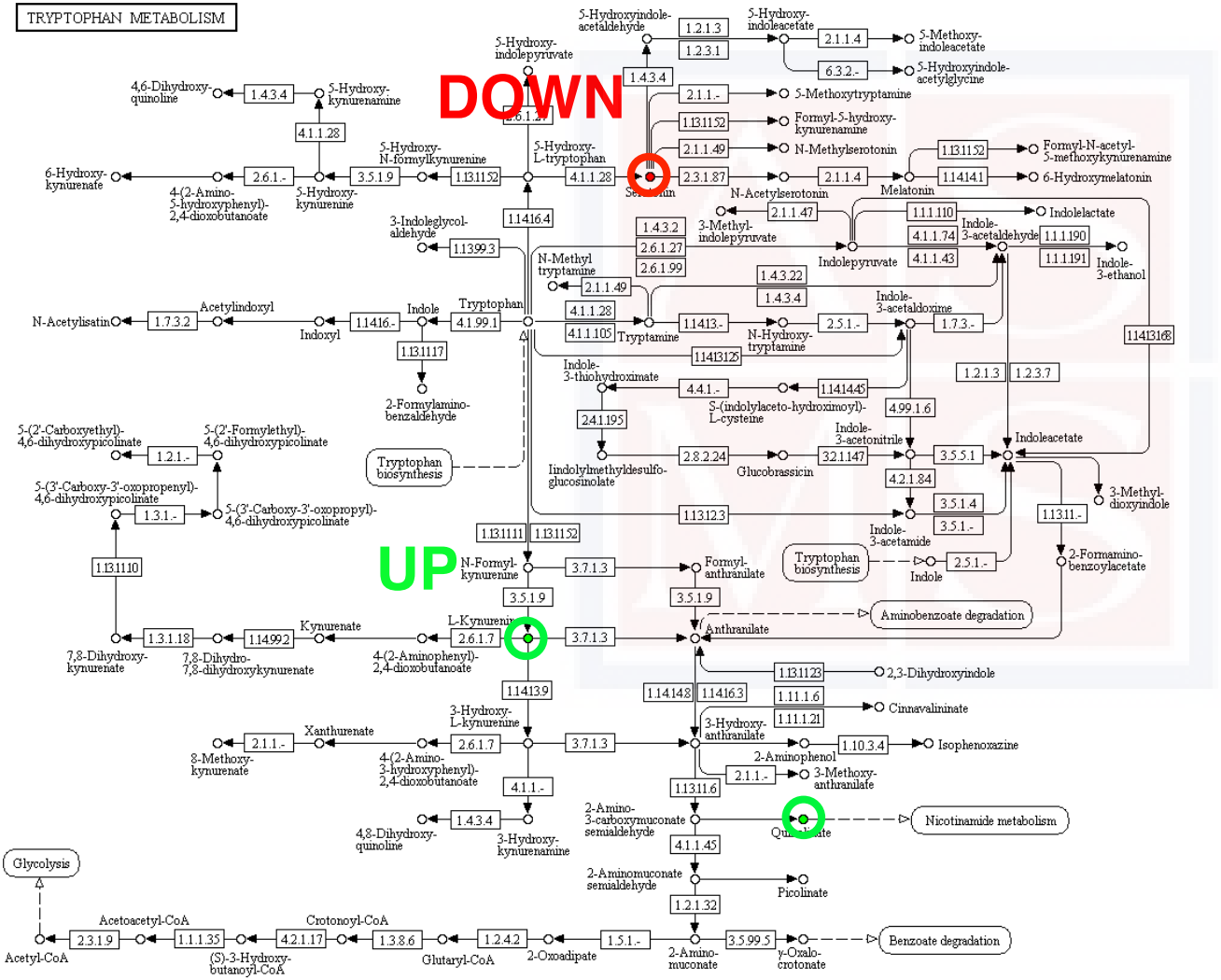
Quinolinic acid





Gendelman et al NPJ Parkinson's Disease Journal 2017

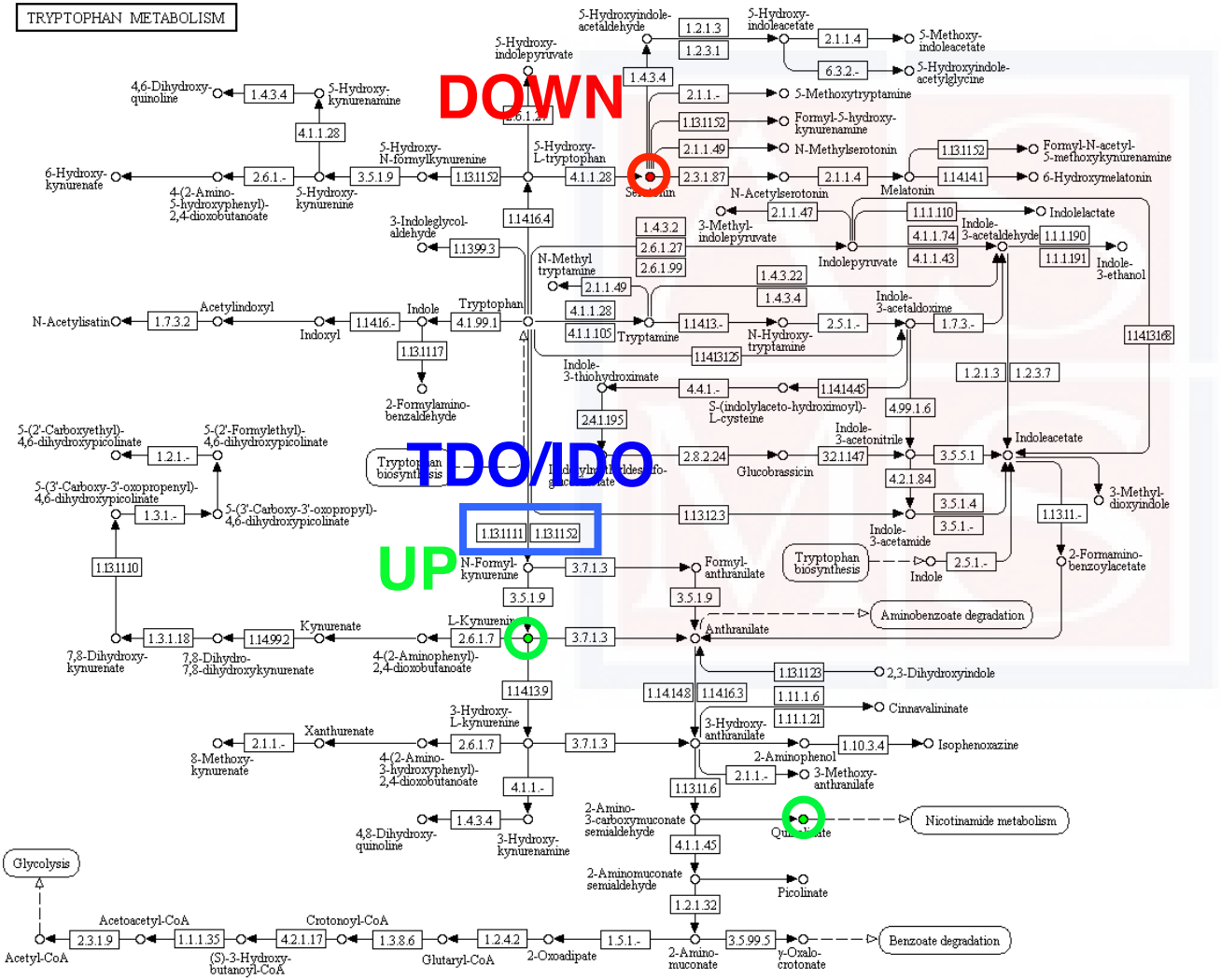
Biomarkers vs. Biological Relevance



Biomarkers

Kenihisa, Nucleic Acids Research 2017

Biomarkers vs. Biological Relevance

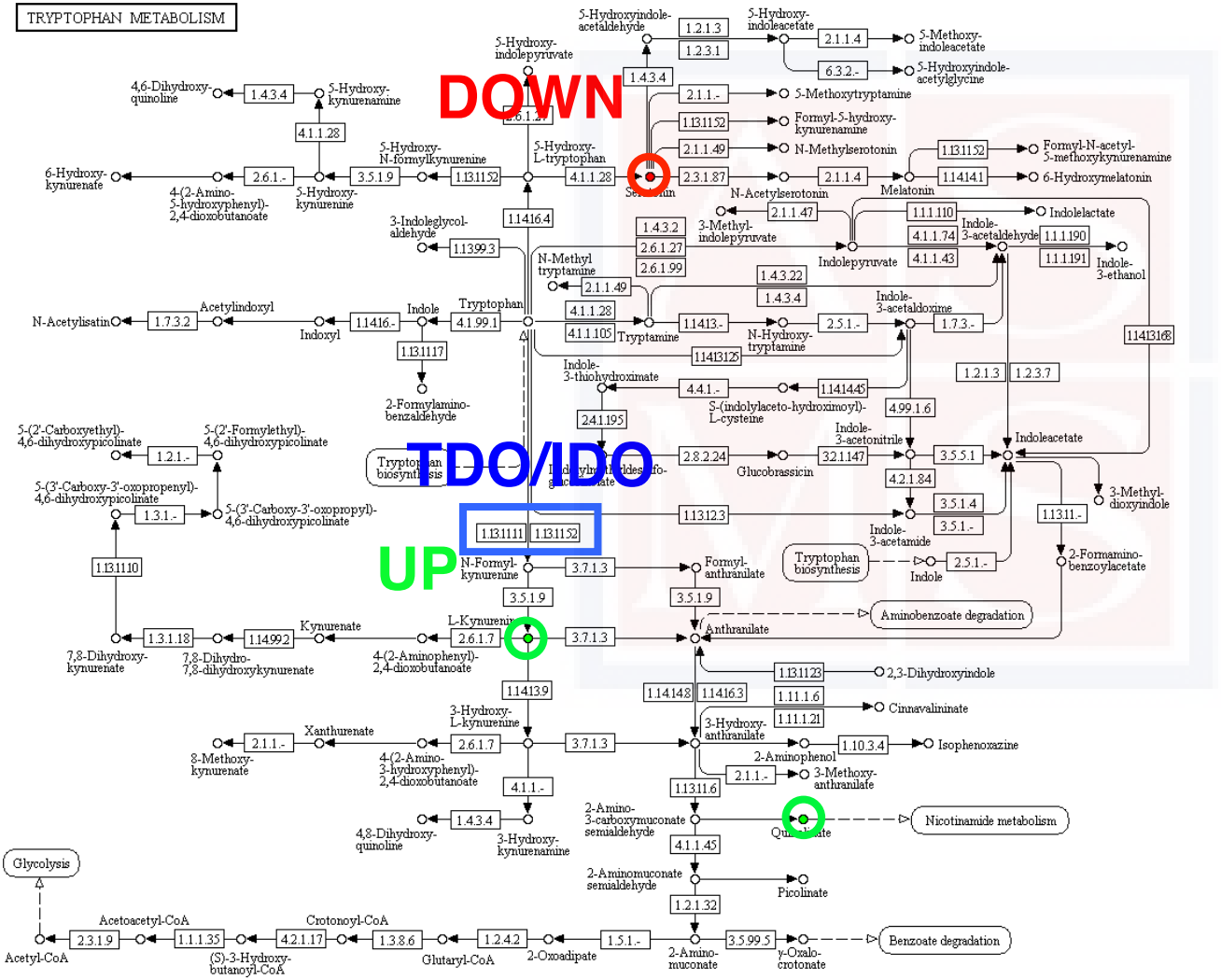


Biomarkers

Pathways

Kenihisa, Nucleic Acids Research 2017

Biomarkers vs. Biological Relevance



Biomarkers

Pathways

Potential Target

Kenihisa, Nucleic Acids Research 2017

Pathway Analysis and Multi-Omic Integration

1. Prerequisites
2. Biomarkers vs. Biological Relevance
3. **Pathway Tools for Annotated Data**
4. Pathway Tools for Unannotated Data
5. Multi-Omic Integration

Pathway Tools for Annotated Data

Pathway Tools for Annotated Data

1. Putative identification with accurate mass

Pathway Tools for Annotated Data

1. Putative identification with accurate mass
2. MS/MS matched with database spectra

Pathway Tools for Annotated Data

1. Putative identification with accurate mass
2. MS/MS matched with database spectra
3. MS/MS and RT matched with standard

Pathway Tools for Annotated Data

1. Putative identification with accurate mass
2. MS/MS matched with database spectra
3. **MS/MS and RT matched with standard**

Pathway Tools for Annotated Data

XCMS→~1,000,000 molecules
https://metlin.scripps.edu

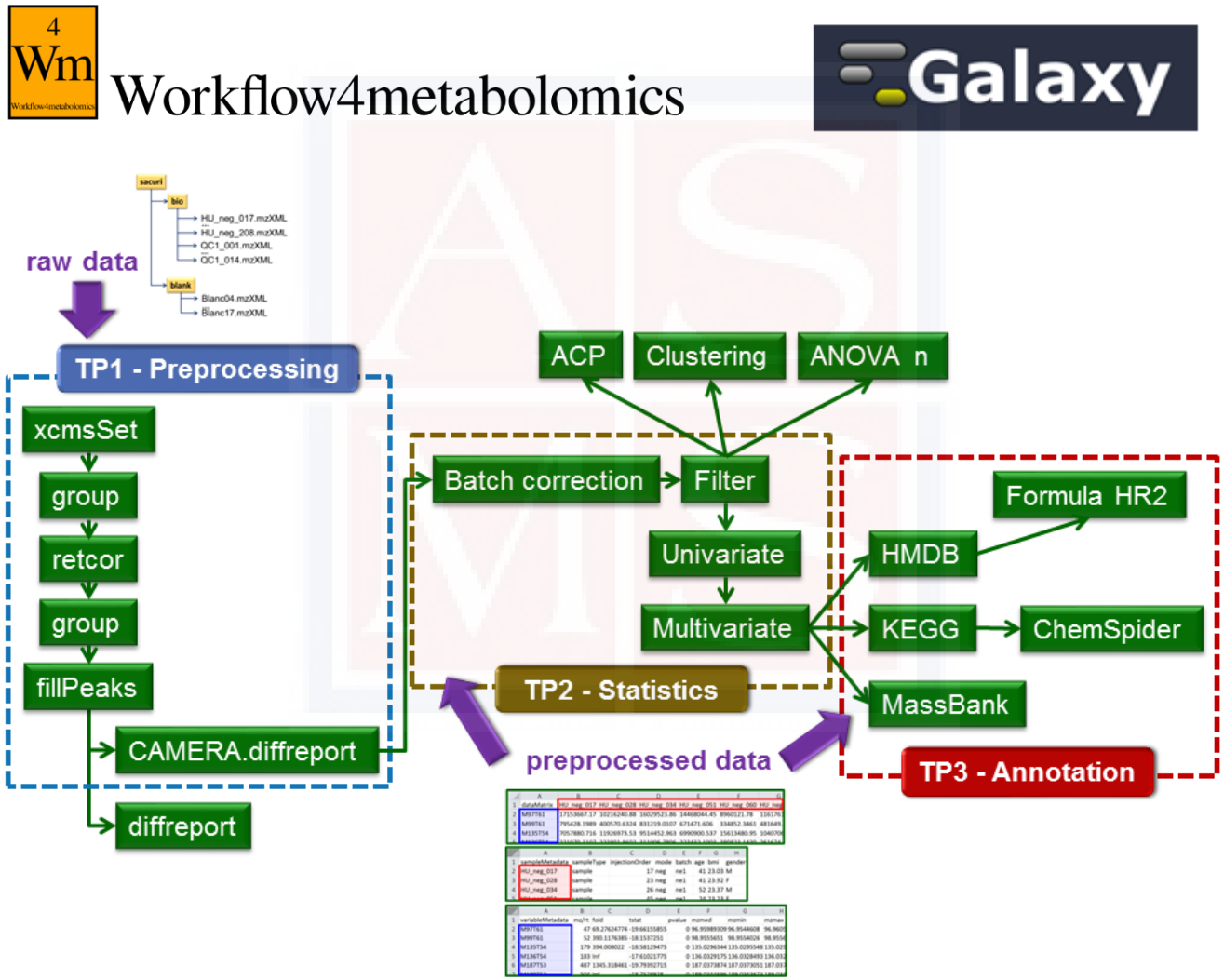
~115,000 molecules
http://www.hmdb.ca/

~40,000 lipids
http://www.lipidmaps.org/tools/

MINE

KEGG, EcoCyc, YMDB, SEED
http://minedatabase.mcs.anl.gov/

Pathway Tools for Annotated Data

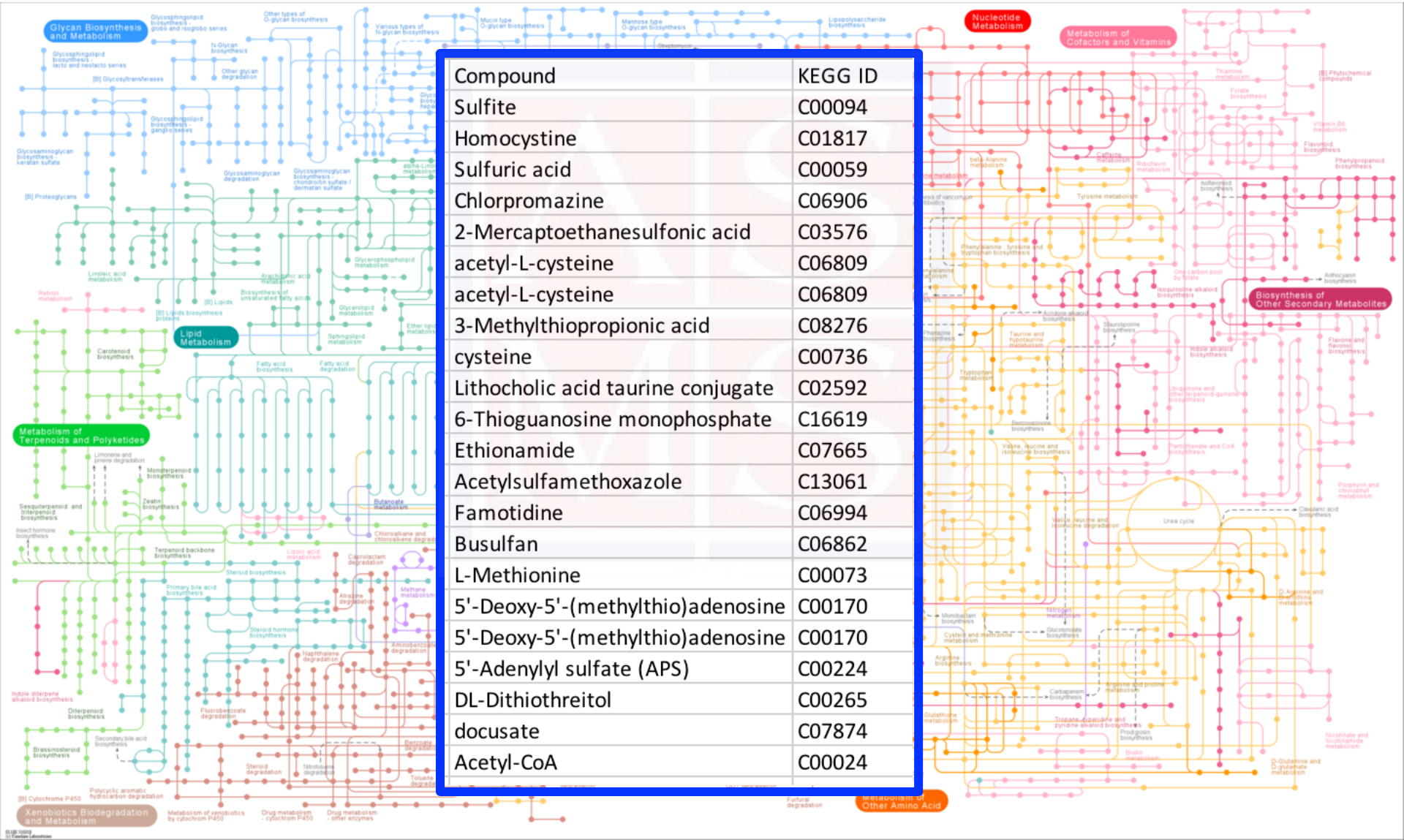


http://workflow4metabolomics.org/

Pathway Tools for Annotated Data

Compound	KEGG ID
Sulfite	C00094
Homocystine	C01817
Sulfuric acid	C00059
Chlorpromazine	C06906
2-Mercaptoethanesulfonic acid	C03576
acetyl-L-cysteine	C06809
acetyl-L-cysteine	C06809
3-Methylthiopropionic acid	C08276
cysteine	C00736
Lithocholic acid taurine conjugate	C02592
6-Thioguanosine monophosphate	C16619
Ethionamide	C07665
Acetylsulfamethoxazole	C13061
Famotidine	C06994
Busulfan	C06862
L-Methionine	C00073
5'-Deoxy-5'-(methylthio)adenosine	C00170
5'-Deoxy-5'-(methylthio)adenosine	C00170
5'-Adenylyl sulfate (APS)	C00224
DL-Dithiothreitol	C00265
docusate	C07874
Acetyl-CoA	C00024

Pathway Tools for Annotated Data



Compound	KEGG ID
Sulfite	C00094
Homocystine	C01817
Sulfuric acid	C00059
Chlorpromazine	C06906
2-Mercaptoethanesulfonic acid	C03576
acetyl-L-cysteine	C06809
acetyl-L-cysteine	C06809
3-Methylthiopropionic acid	C08276
cysteine	C00736
Lithocholic acid taurine conjugate	C02592
6-Thioguanosine monophosphate	C16619
Ethionamide	C07665
Acetylsulfamethoxazole	C13061
Famotidine	C06994
Busulfan	C06862
L-Methionine	C00073
5'-Deoxy-5'-(methylthio)adenosine	C00170
5'-Deoxy-5'-(methylthio)adenosine	C00170
5'-Adenylyl sulfate (APS)	C00224
DL-Dithiothreitol	C00265
docusate	C07874
Acetyl-CoA	C00024

Pathway Tools for Annotated Data

KEGG Mapper

442 Eukaryotes

4694 Bacteria

269 Archea

Kanehisa et al, Nucleic Acids Research, 2011

Pathway Tools for Annotated Data

KEGG Mapper

442 Eukaryotes

4694 Bacteria

269 Archea

Kanehisa et al, Nucleic Acids Research, 2011
Yamada et al, Nucleic Acids Research, 2011

Pathway Tools for Annotated Data

KEGG Mapper

442 Eukaryotes

4694 Bacteria

269 Archea

Kanehisa et al, Nucleic Acids Research, 2011
Yamada et al, Nucleic Acids Research, 2011
López-Ibáñez et al, Nucleic Acids Research, 2016

Pathway Tools for Annotated Data

KEGG Mapper

442 Eukaryotes

4694 Bacteria

269 Archea

iPATH

interactive Pathways Explorer

MBROLE

MetaboAnalyst 4.0

Statistics

Visualization

Integration

Biomarker

Pathway

Function

MSEA

MetPA

Kanehisa et al, Nucleic Acids Research, 2011

Yamada et al, Nucleic Acids Research, 2011

López-Ibáñez et al, Nucleic Acids Research, 2016

Xia and Wishart, Nucleic Acids Research, 2010

Pathway Tools for Annotated Data

KEGG Mapper

Pathway Tools for Annotated Data

KEGG Mapper

Search against: Enter: map, ko, ec, rn, hsadd, or org

Primary ID: KEGG identifiers (Outside IDs for organism-specific pathways only)

Enter objects one per line followed by bgcolor, fgcolor:

Pathway Tools for Annotated Data

KEGG Mapper

Select Organism

Search against: dvu Enter: map, ko, ec, rn, hsadd, or org

Primary ID: KEGG identifiers (Outside IDs for organism-specific pathways only)

Enter objects one per line followed by bgcolor, fgcolor:

```
C06809    red
C06809    red
C06862    green
C06906    red
C06994    green
C07665    red
C07874    green
C08276    red
C13061    red
C16619    red
```

Pathway Tools for Annotated Data

Annotated Data

KEGG ID
C00094
C01817
C00059
C06906
C03576
C06809
C06809
C08276
C00736
C02592
C16619
C07665
C13061
C06994
C06862
C00073
C00170
C00170
C00224
C00265
C07874
C00024

Compound IDs

KEGG Mapper

Select Organism

Search against: dvu Enter: map, ko, ec, rn, hsadd, or org

Primary ID: KEGG identifiers (Outside IDs for organism-specific pathways only)

Enter objects one per line followed by bgcolor, fgcolor:

```
C06809
C06809
C06862
C06906
C06994
C07665
C07874
C08276
C13061
C16619
```

Pathway Tools for Annotated Data

Annotated Data

KEGG ID
C00094
C01817
C00059
C06906
C03576
C06809
C06809
C08276
C00736
C02592
C16619
C07665
C13061
C06994
C06862
C00073
C00170
C00170
C00224
C00265
C07874
C00024

Compound IDs

KEGG Mapper

Select Organism

Search against: dvu Enter: map, ko, ec, rn, hsadd, or org

Primary ID: KEGG identifiers (Outside IDs for organism-specific pathways only)

Enter objects one per line followed by bgcolor, fgcolor:

```
C06809    red
C06809    red
C06862    green
C06906    red
C06994    green
C07665    red
C07874    green
C08276    red
C13061    red
C16619    red
```

UP

DOWN

Pathway Tools for Annotated Data

KEGG Mapper

Pathway Search Result

Following object(s) was/were not found cpd:C00265 cpd:C00736 cpd:C02592 cpd:C06809 cpd:C06862 cpd:C06906 cpd:C0

Sort by the pathway list

Show all objects

- dvu01100 Metabolic pathways - Desulfovibrio vulgaris Hildenborough (7)
- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)
- dvu01120 Microbial metabolism in diverse environments - Desulfovibrio vulgaris Hildenborough (5)
- dvu00920 Sulfur metabolism - Desulfovibrio vulgaris Hildenborough (4)
- dvu01130 Biosynthesis of antibiotics - Desulfovibrio vulgaris Hildenborough (3)
- dvu01200 Carbon metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00430 Taurine and hypotaurine metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00680 Methane metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00261 Monobactam biosynthesis - Desulfovibrio vulgaris Hildenborough (2)
- dvu01110 Biosynthesis of secondary metabolites - Desulfovibrio vulgaris Hildenborough (2)
- dvu00230 Purine metabolism - Desulfovibrio vulgaris Hildenborough (2)

Pathway Tools for Annotated Data

KEGG Mapper

Number of overlapping metabolites

Pathway Search Result

Following object(s) was/were not found cpd:C00265 cpd:C00736 cpd:C02592 cpd:C06809 cpd:C06862 cpd:C06906 cpd:C0

Sort by the pathway list

Show all objects

- dvu01100 Metabolic pathways - Desulfovibrio vulgaris Hildenborough (7)
- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)
- dvu01120 Microbial metabolism in diverse environments - Desulfovibrio vulgaris Hildenborough (5)
- dvu00920 Sulfur metabolism - Desulfovibrio vulgaris Hildenborough (4)
- dvu01130 Biosynthesis of antibiotics - Desulfovibrio vulgaris Hildenborough (3)
- dvu01200 Carbon metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00430 Taurine and hypotaurine metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00680 Methane metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00261 Monobactam biosynthesis - Desulfovibrio vulgaris Hildenborough (2)
- dvu01110 Biosynthesis of secondary metabolites - Desulfovibrio vulgaris Hildenborough (2)
- dvu00230 Purine metabolism - Desulfovibrio vulgaris Hildenborough (2)

Pathway Tools for Annotated Data

KEGG Mapper

Number of overlapping metabolites

Pathway Search Result

Following object(s) was/were not found cpd:C00265 cpd:C00736 cpd:C02592 cpd:C06809 cpd:C06862 cpd:C06906 cpd:C0

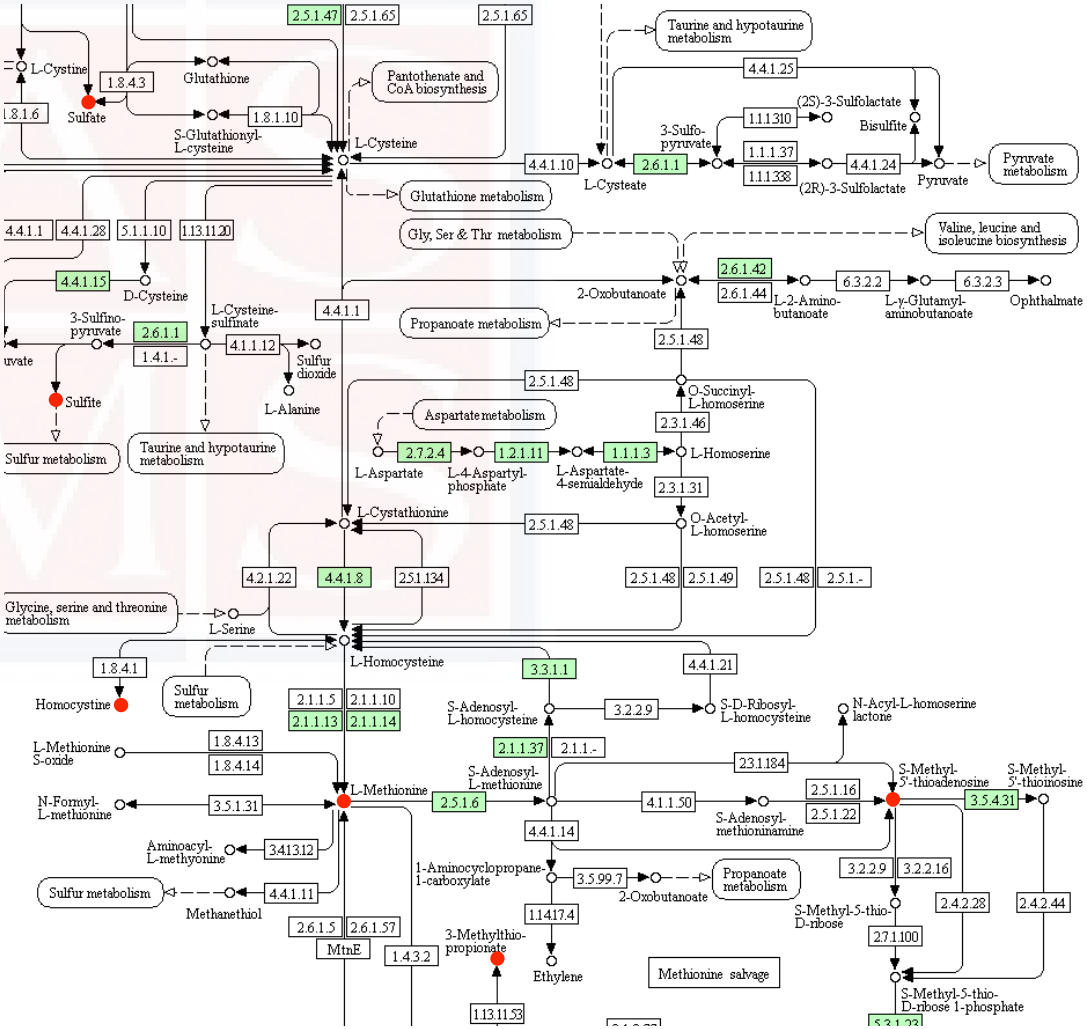
Sort by the pathway list

Show all objects

- dvu01100 Metabolic pathways - Desulfovibrio vulgaris Hildenborough (7)
- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)
- dvu01120 Microbial metabolism in diverse environments - Desulfovibrio vulgaris Hildenborough (5)
- dvu00920 Sulfur metabolism - Desulfovibrio vulgaris Hildenborough (4)
- dvu01130 Biosynthesis of antibiotics - Desulfovibrio vulgaris Hildenborough (3)
- dvu01200 Carbon metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00430 Taurine and hypotaurine metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00680 Methane metabolism - Desulfovibrio vulgaris Hildenborough (2)
- dvu00261 Monobactam biosynthesis - Desulfovibrio vulgaris Hildenborough (2)
- dvu01110 Biosynthesis of secondary metabolites - Desulfovibrio vulgaris Hildenborough (2)
- dvu00230 Purine metabolism - Desulfovibrio vulgaris Hildenborough (2)

Pathway Tools for Annotated Data

- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)



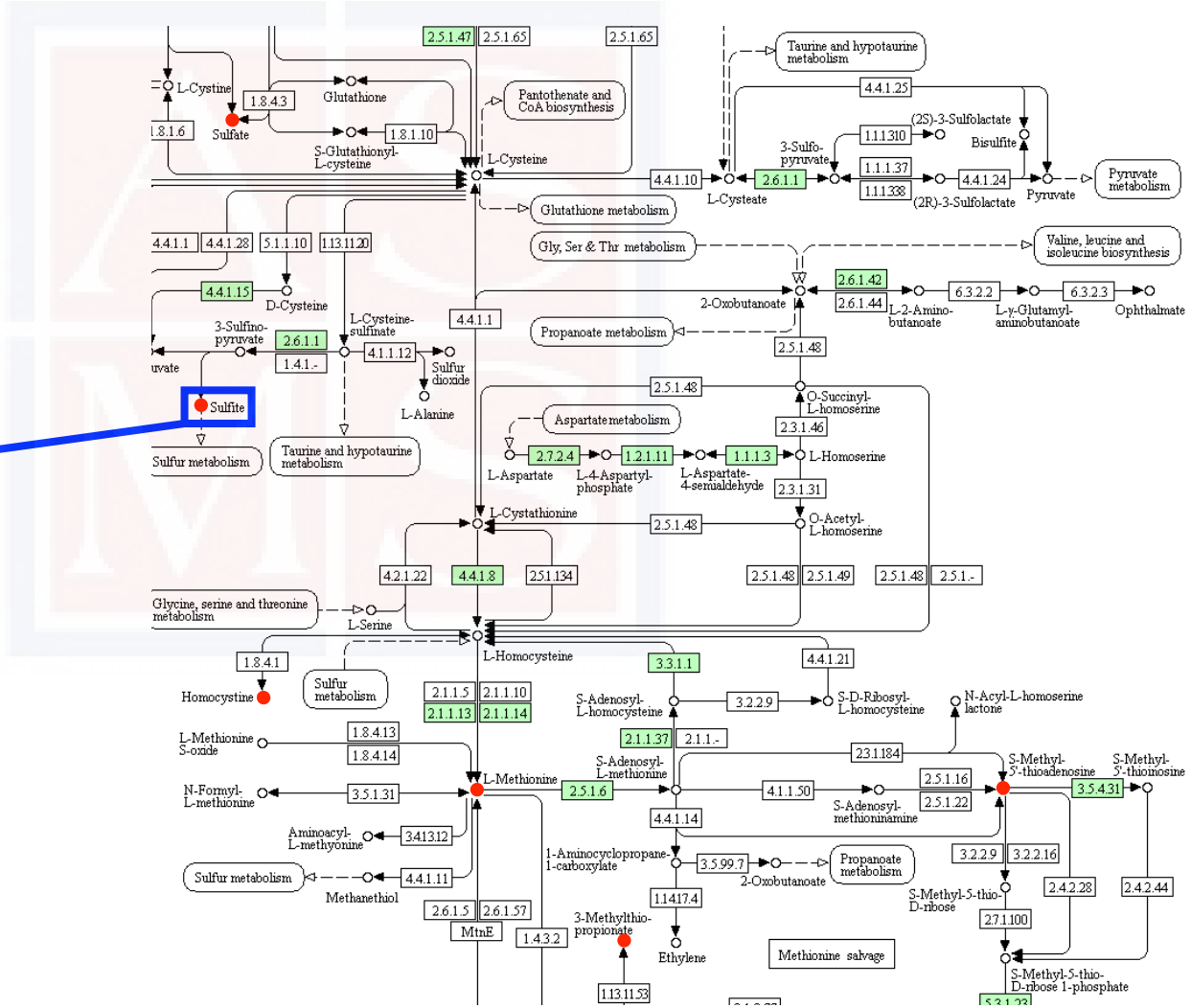
KEGG Mapper

Pathway Tools for Annotated Data

- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)

Overlapping metabolite

KEGG Mapper

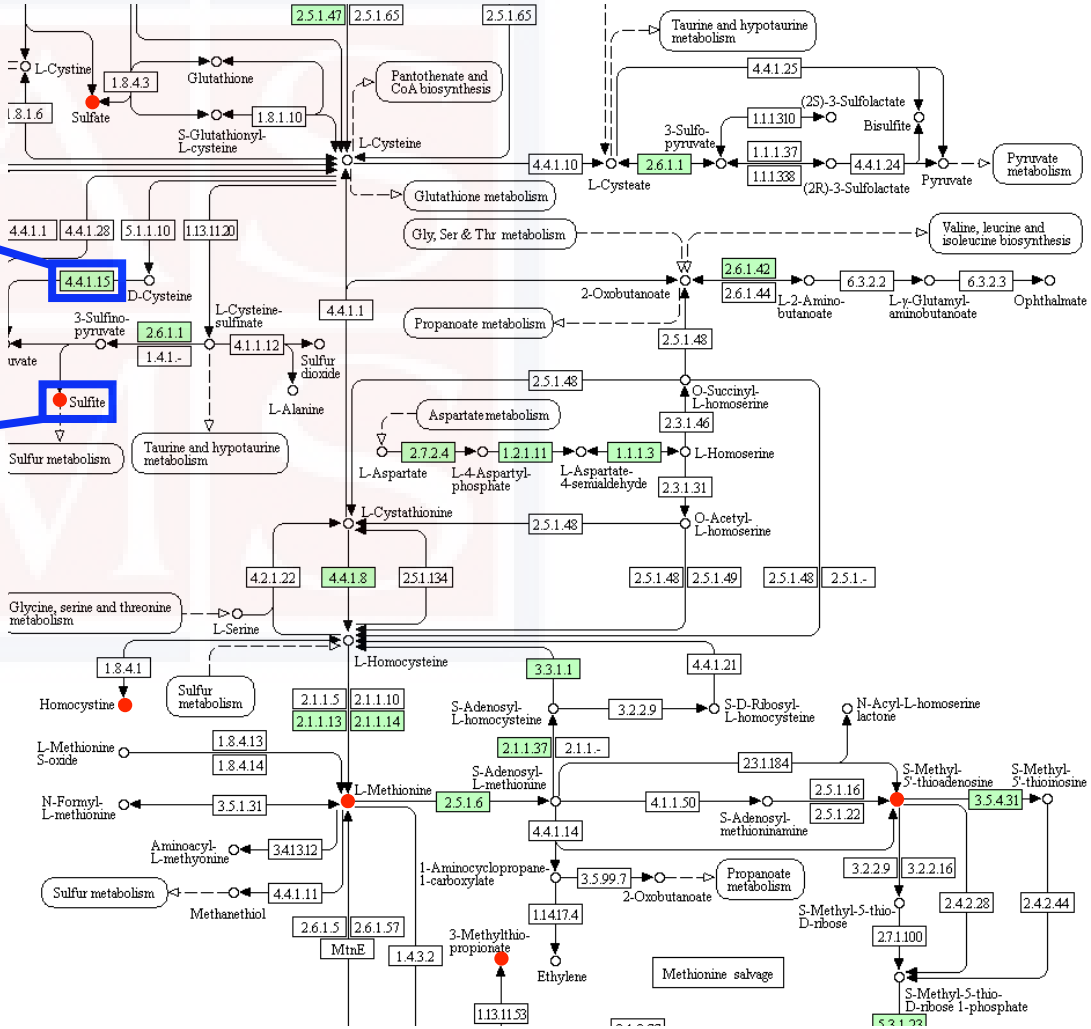


Pathway Tools for Annotated Data

- dvu00270 Cysteine and methionine metabolism - Desulfovibrio vulgaris Hildenborough (6)

Enzymes known in organism

Overlapping metabolite



KEGG Mapper

Pathway Tools for Annotated Data

iPATH

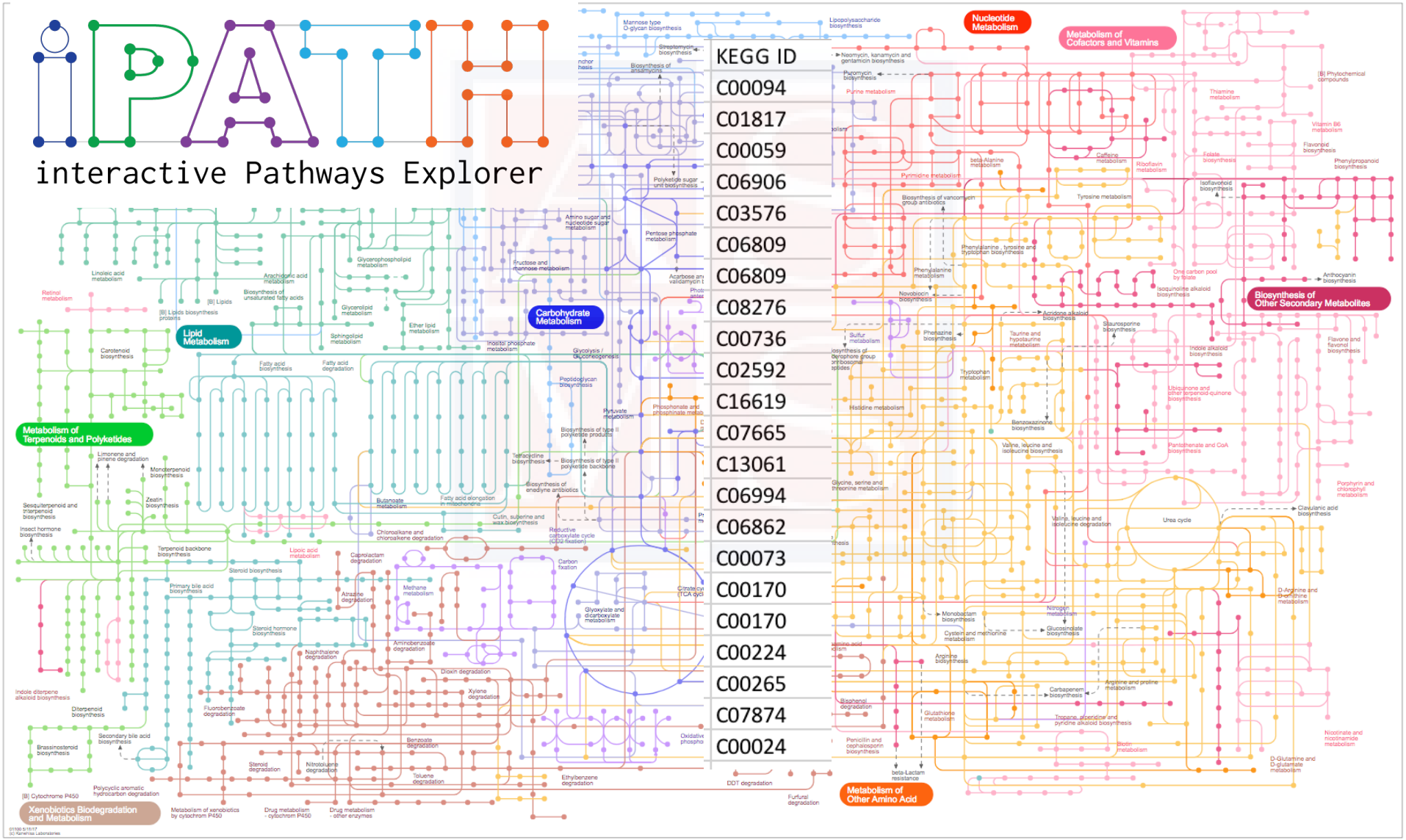
interactive Pathways Explorer

Pathway Tools for Annotated Data

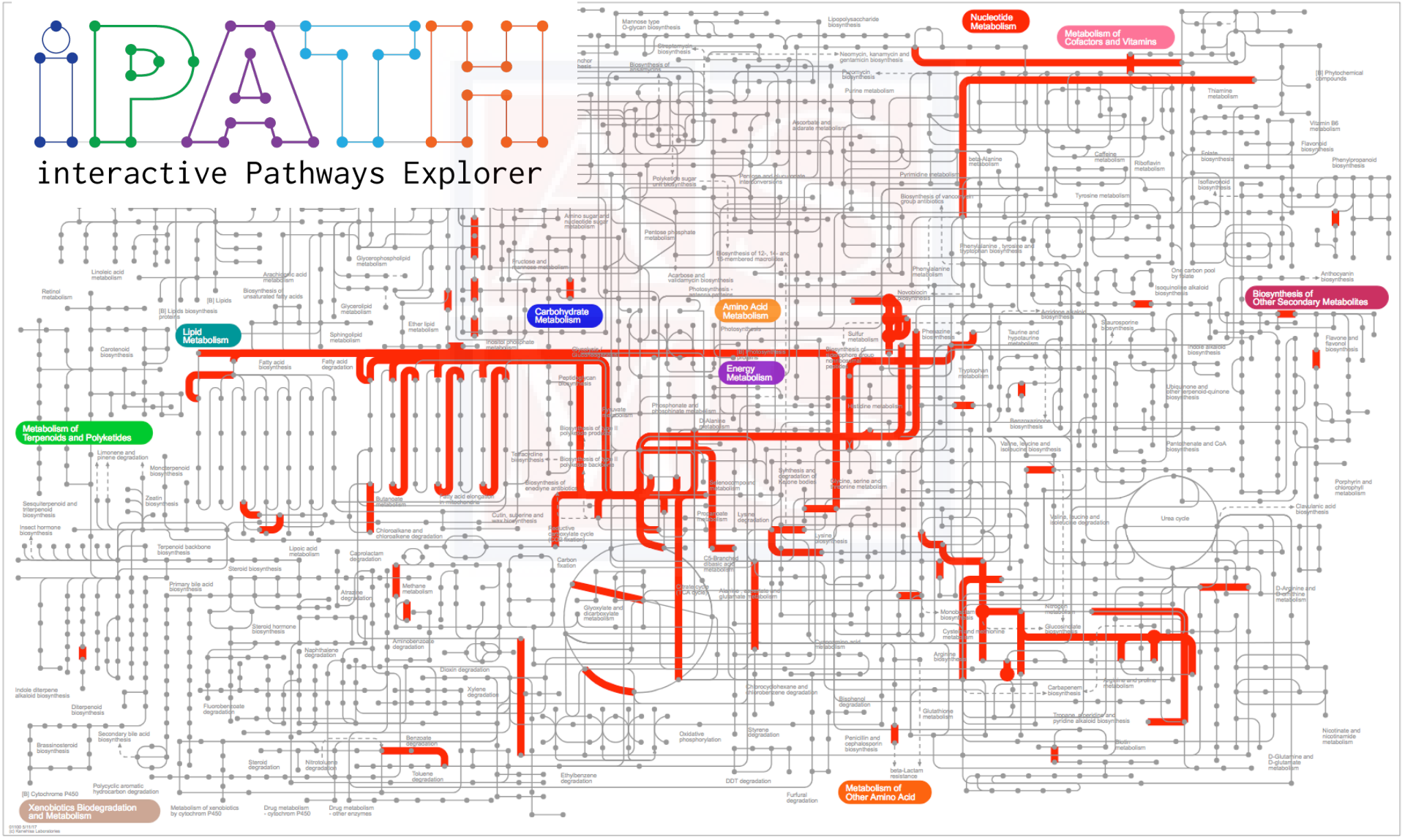
iPATH
interactive Pathways Explorer

KEGG ID
C00094
C01817
C00059
C06906
C03576
C06809
C06809
C08276
C00736
C02592
C16619
C07665
C13061
C06994
C06862
C00073
C00170
C00170
C00224
C00265
C07874
C00024

Pathway Tools for Annotated Data

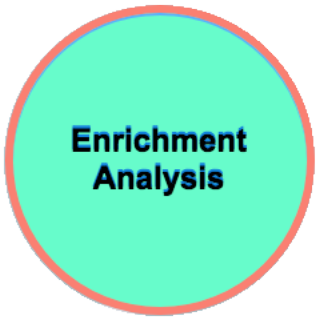


Pathway Tools for Annotated Data



Pathway Tools for Annotated Data

MSEA

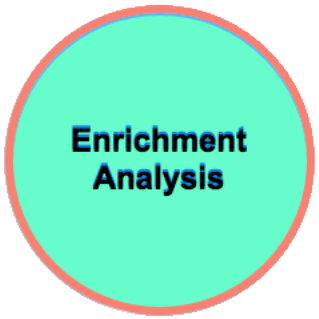


KEGG ID
C00094
C01817
C00059
C06906
C03576
C06809
C06809
C08276
C00736
C02592
C16619
C07665
C13061
C06994
C06862
C00073
C00170
C00170
C00224
C00265
C07874
C00024

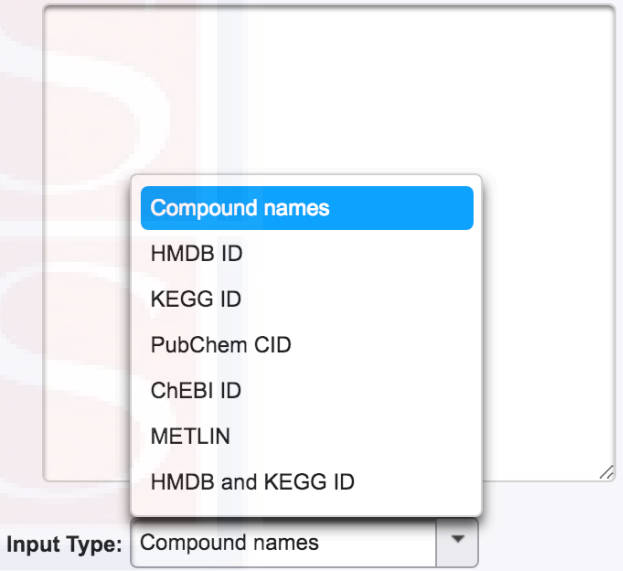
Xia and Wishart, Nucleic Acids Research, 2010

Pathway Tools for Annotated Data

MSEA



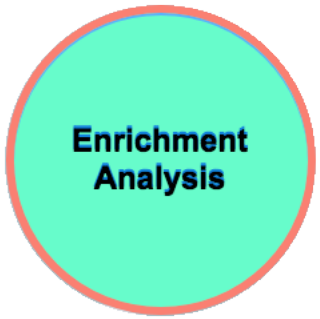
KEGG ID
C00094
C01817
C00059
C06906
C03576
C06809
C06809
C08276
C00736
C02592
C16619
C07665
C13061
C06994
C06862
C00073
C00170
C00170
C00224
C00265
C07874
C00024



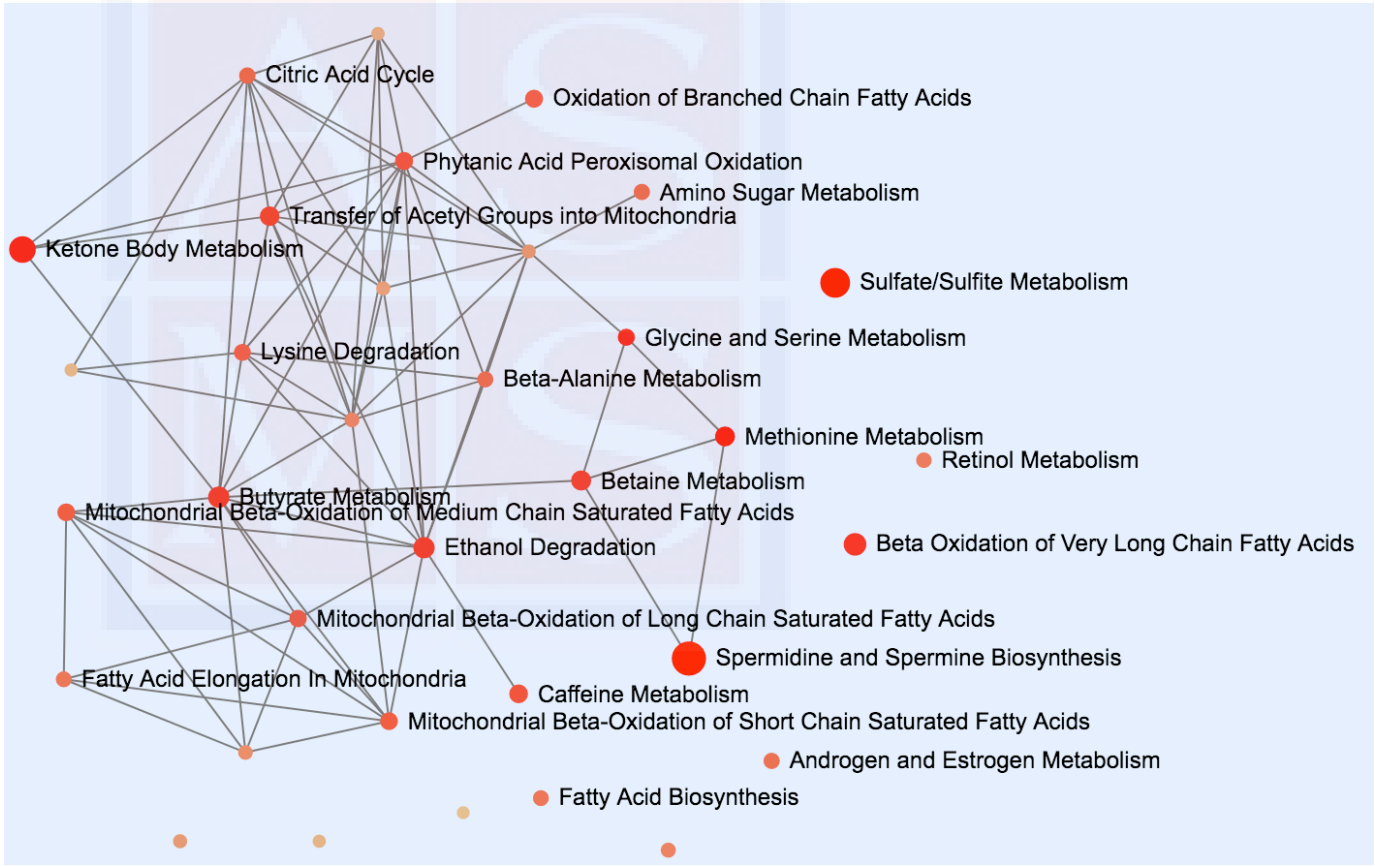
Xia and Wishart, Nucleic Acids Research, 2010

Pathway Tools for Annotated Data

MSEA



Metabolite Functional Enrichment

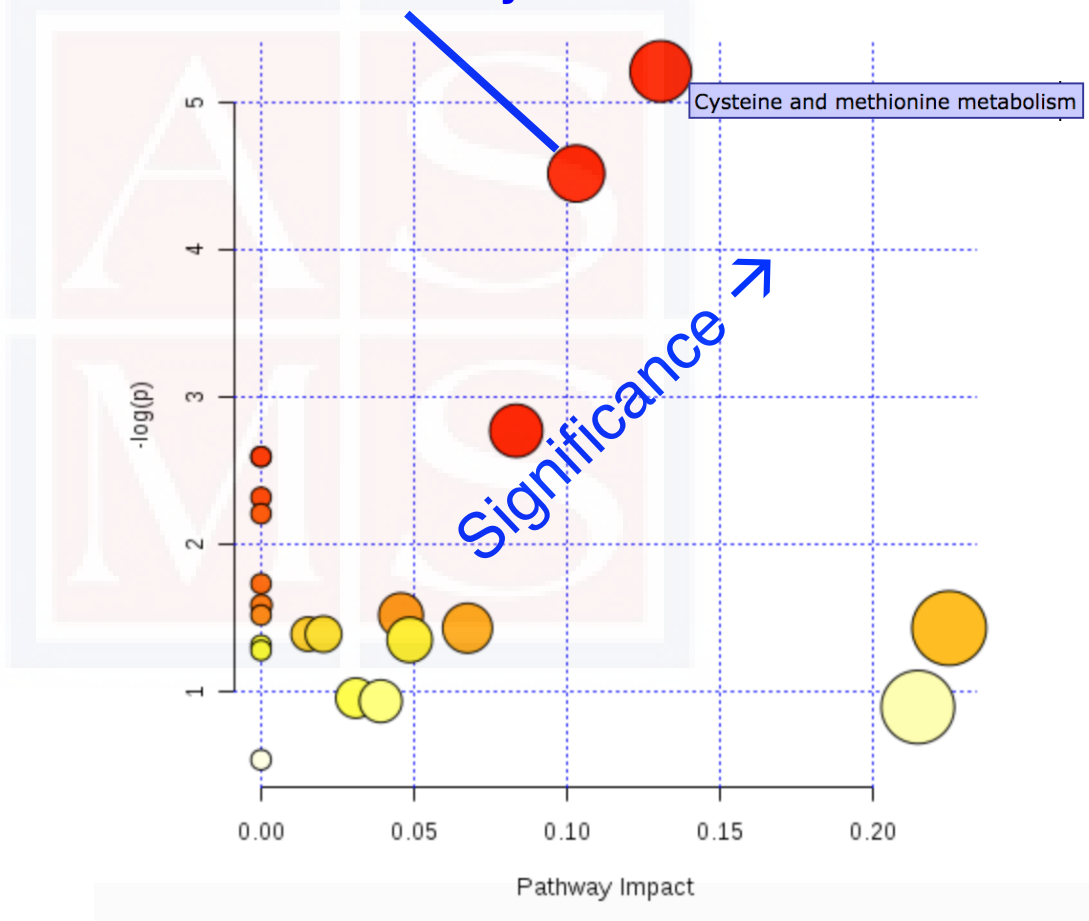


Xia and Wishart, Nucleic Acids Research, 2010

Pathway Tools for Annotated Data

MetPA

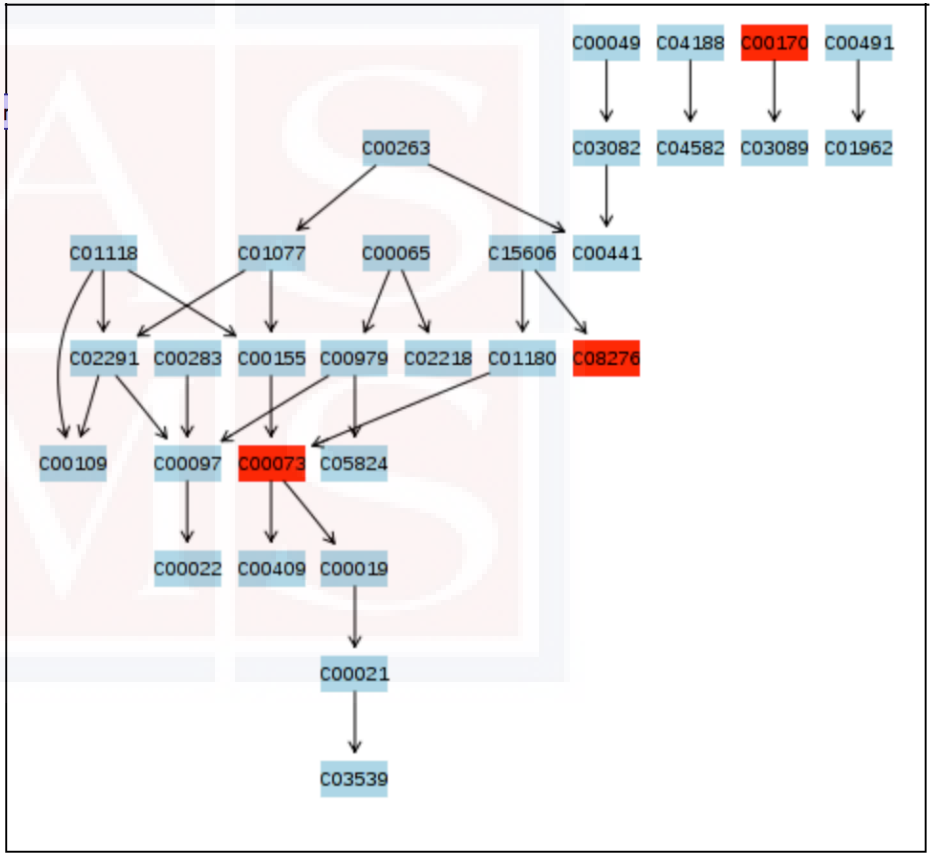
Metabolic Pathways



Xia and Wishart, Bioinformatics, 2010

Pathway Tools for Annotated Data

MetPA



Xia et al Bioinformatics 2010

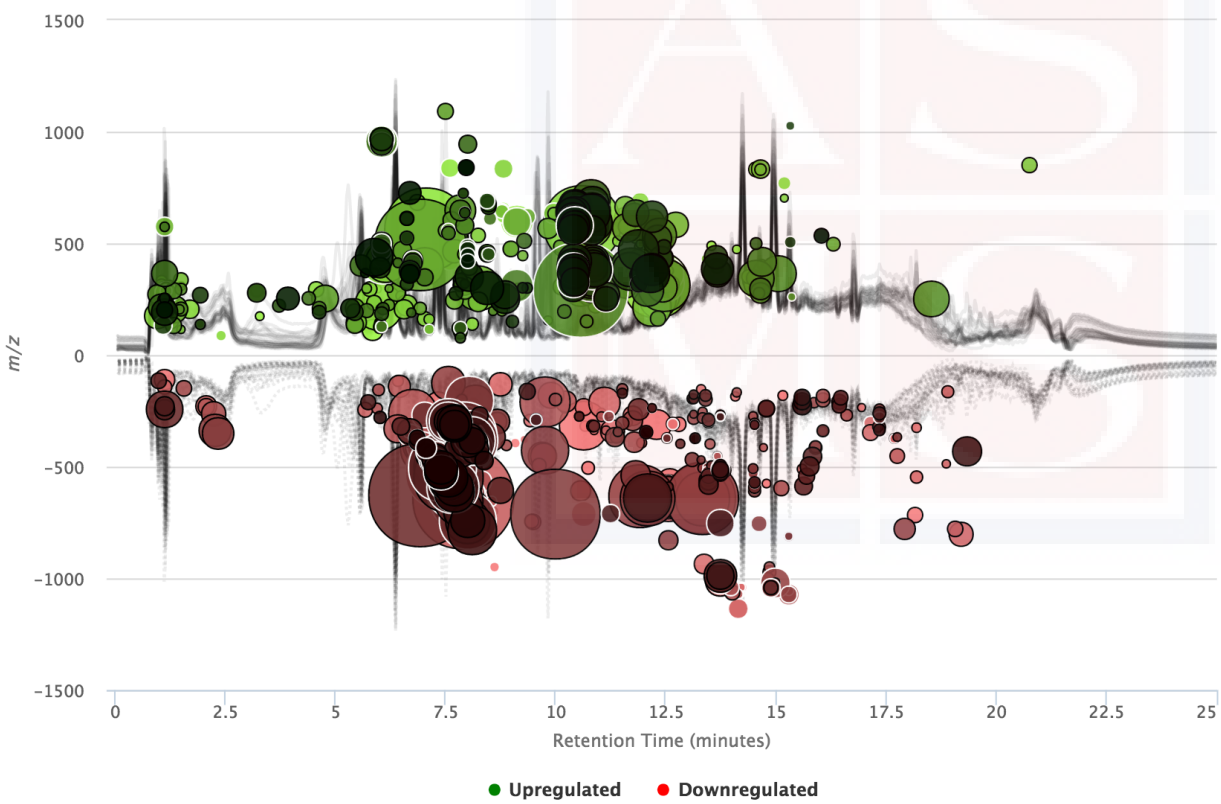
Pathway Analysis and Multi-Omic Integration

1. Prerequisites
2. Biomarkers vs. Biological Relevance
3. Pathway Tools for Annotated Data
4. **Pathway Tools for Unannotated Data**
5. Multi-Omic Integration

Pathway Tools for Unannotated Data

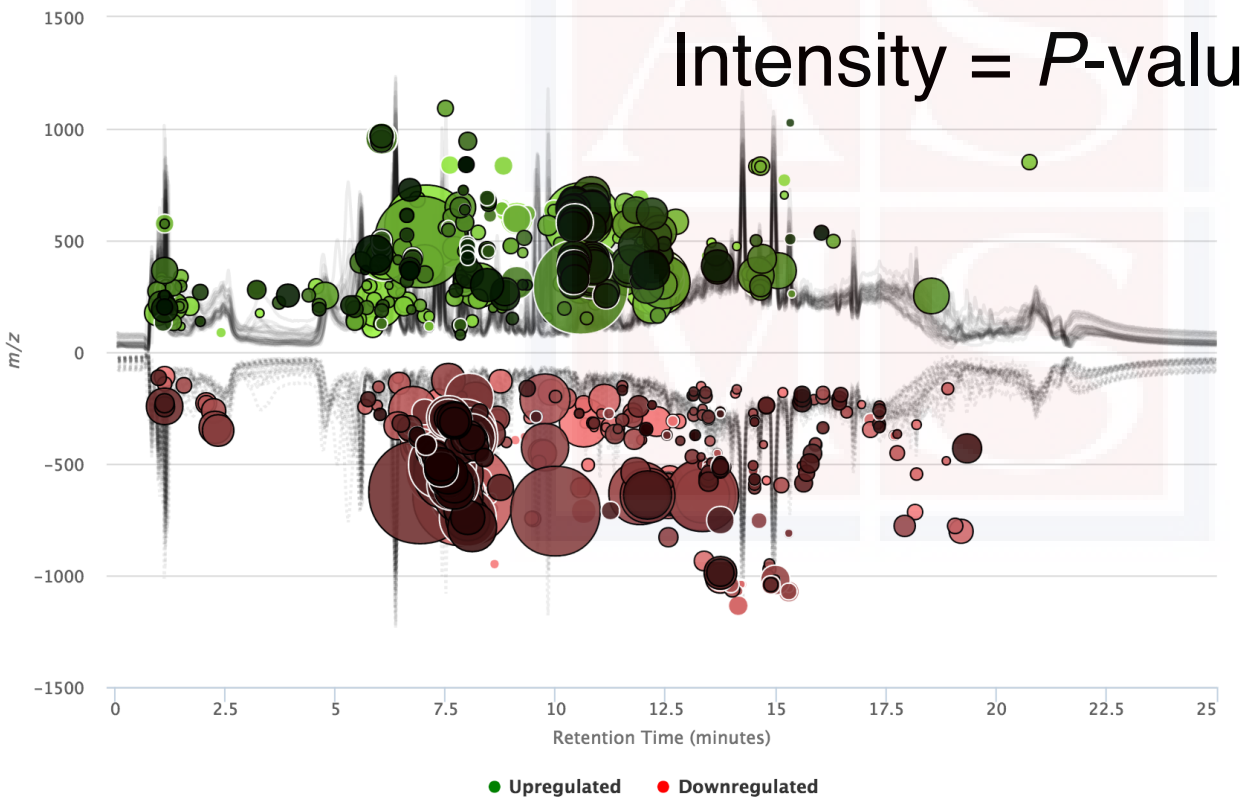
Feature Mapping

Feature Mapping



Feature Mapping

Each circle = 1 metabolite feature
Radius = fold change
Intensity = *P*-value



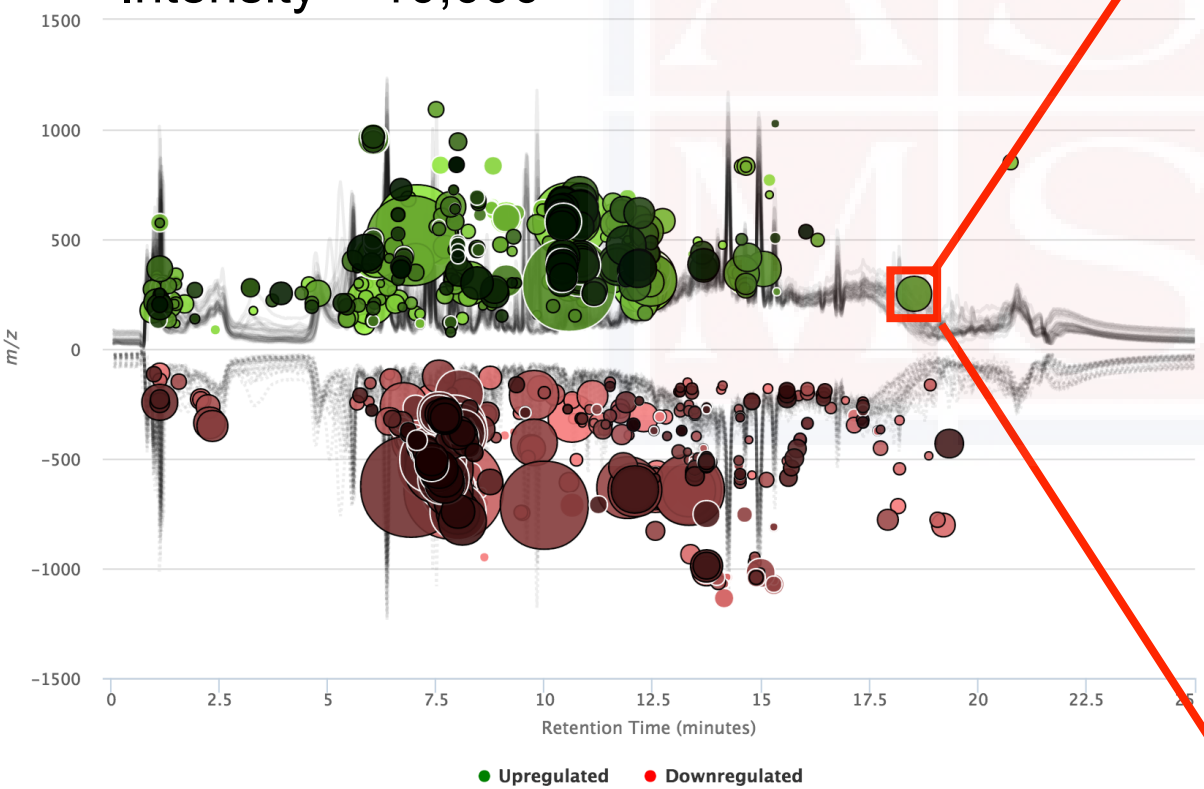
Feature Mapping

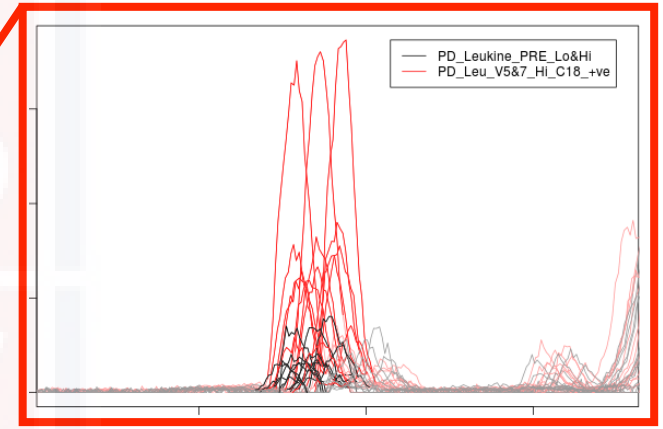
893 Features

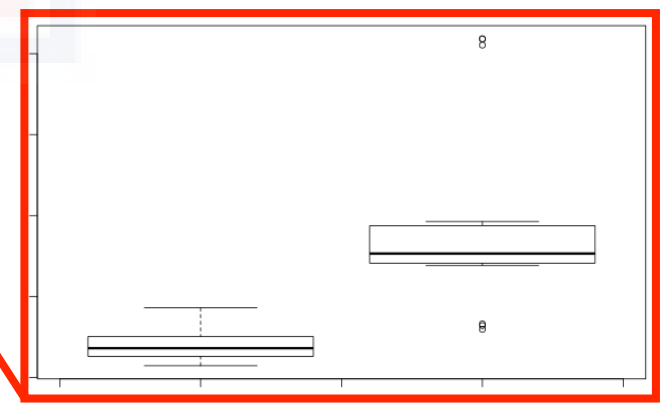
Fold change > 1.5

p-value < 0.01

Intensity > 10,000

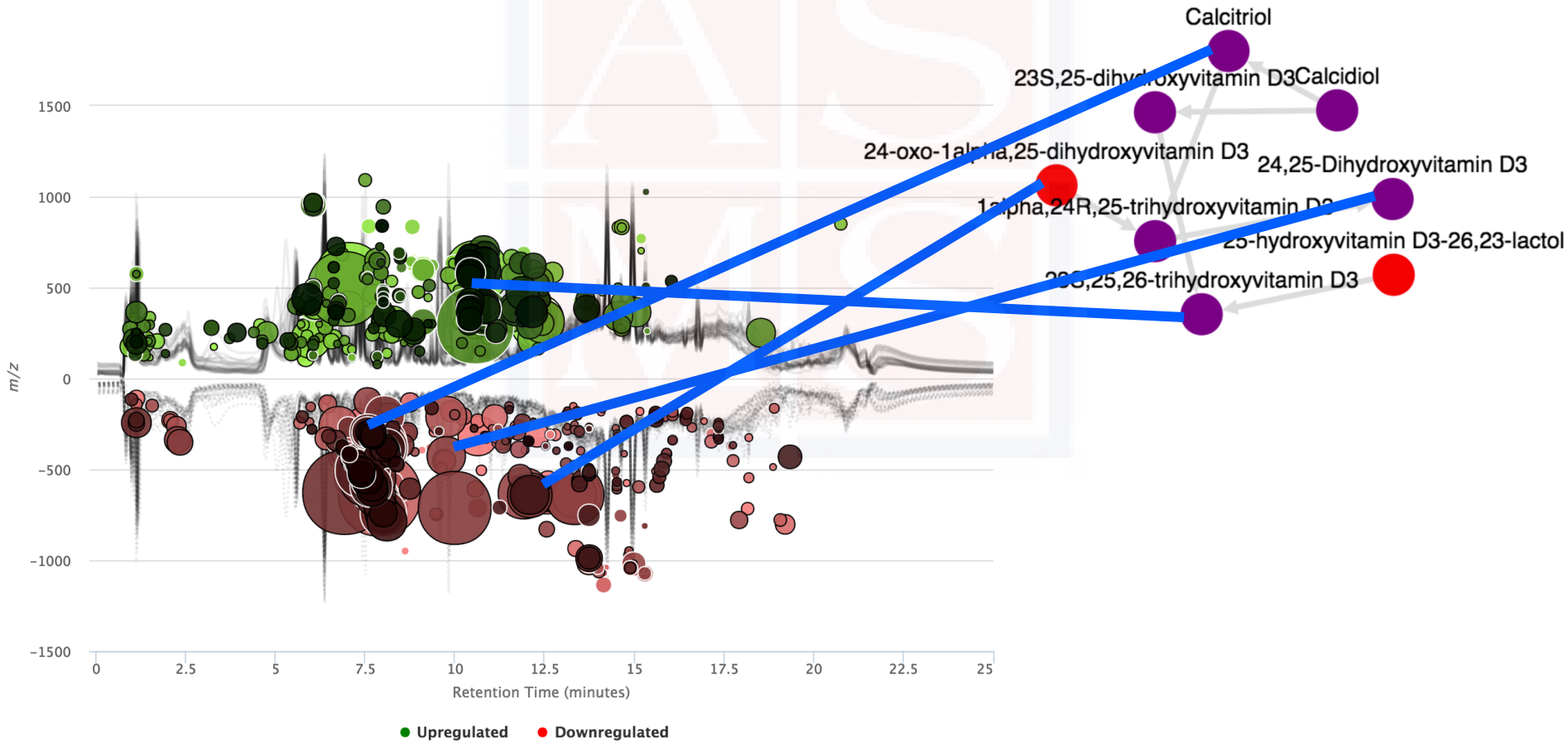






Feature Mapping

Features directly onto metabolic pathways



Feature Mapping

Automated Workflows using Mummichog:

Feature Mapping

Automated Workflows using Mummichog:

S. Li et al, *PLoS Comput. Biol.*, **2013**
Shuzao Li
Emory University

Mummichog

m/z

207.0759
228.0515

182.0833
164.0722
203.0579

152.0722
212.0925
193.0995
198.0767
134.0607

m/z features

Li et al, PLoS Computational Biology, 2013

Mummichog

m/z

L-kynurenine Neutral mass: 208.0848

207.0759

228.0515

adrenaline Neutral mass: 183.0895

182.0833

164.0722

203.0579

dopamine Neutral mass: 153.0790

152.0722

212.0925

193.0995

198.0767

134.0607

m/z features

Monoisotopic mass

Li et al, PLoS Computational Biology, 2013

Mummichog

m/z		
L-kynurenine	Neutral mass:	208.0848
207.0759	M-H[-]	-0.0002
228.0515	M+Na-2H[-]	-0.0007
adrenaline	Neutral mass:	183.0895
182.0833	M-H[-]	0.0010
164.0722	M-H2O-H[-]	0.0005
203.0579	M+Na-2H[-]	0.0009
dopamine	Neutral mass:	153.0790
152.0722	M-H[-]	0.0005
212.0925	M+CH3COO[-]	0.0002
193.0995	M+ACN-H[-]	0.0013
198.0767	M+HCOO[-]	0.0001
134.0607	M-H2O-H[-]	-0.0004

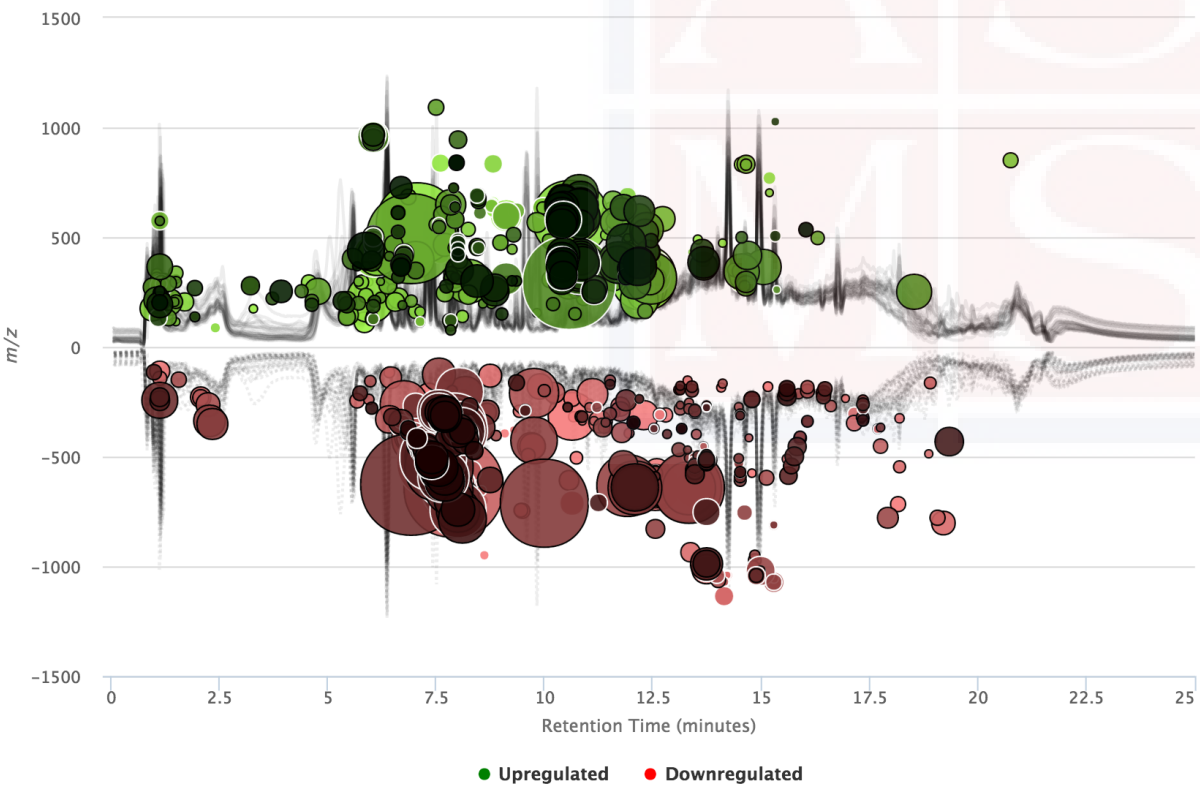
m/z features

Monoisotopic mass

Mass difference (Da)

Li et al, PLoS Computational Biology, 2013

Mummichog

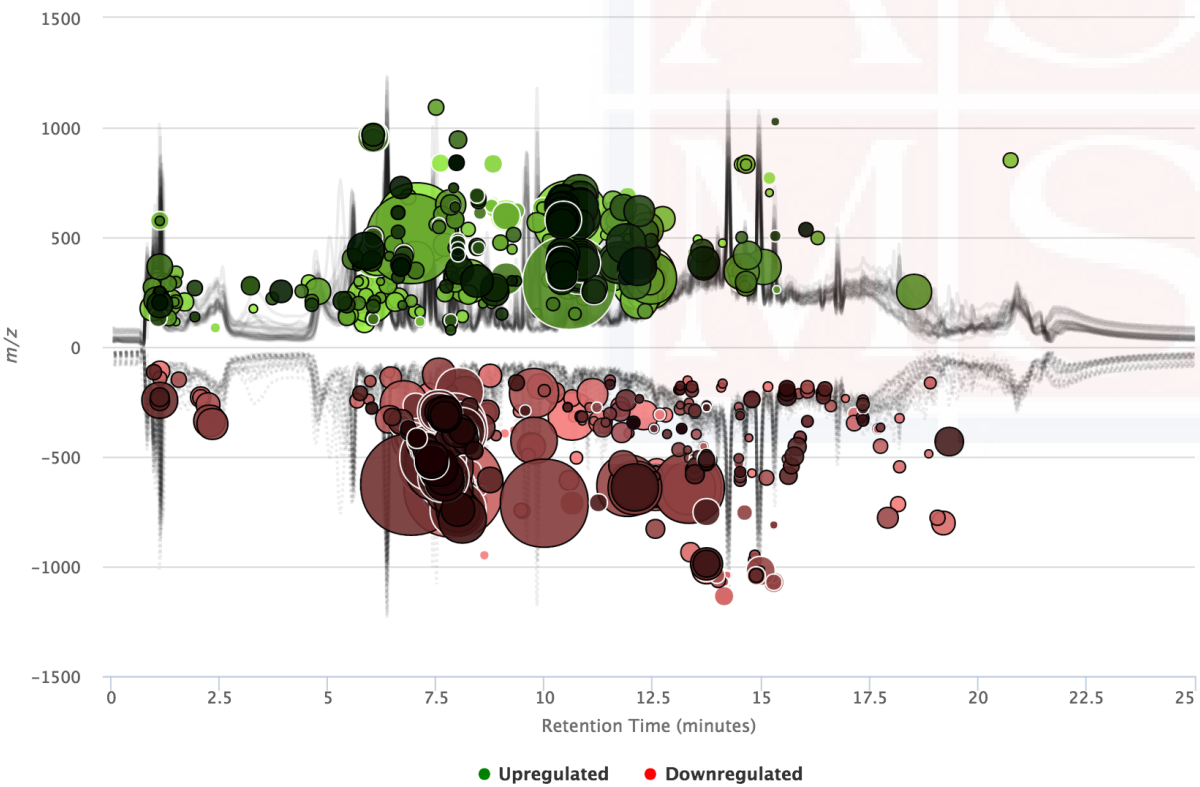


Significant

439.0639
207.0759
583.2114
616.1190
182.0833
300.0444
314.0600
260.5147
495.2696
152.0722
518.1004
437.0628
140.0100
615.1152
139.0398
184.0011
384.9432
338.2470
507.1602
268.1587

Mummichog

Fisher's Exact Test



Significant	Reference
439.0639	387.2271
207.0759	152.0722
583.2114	699.2394
616.1190	543.3510
182.0833	338.2470
300.0444	240.9840
314.0600	158.1189
260.5147	357.3067
495.2696	563.3888
152.0722	182.0833
518.1004	240.1712
437.0628	203.0691
140.0100	261.6342
615.1152	253.0341
139.0398	283.0455
184.0011	572.2922
384.9432	207.0759
338.2470	484.1861
507.1602	138.0226

Mummichog

FET:

Statistical significance of matched significant list vs. nonsignificant list compared with a random permutation of the signifcant list vs. nonsignificant list

Mummichog

FET:

Statistical significance of matched significant list vs. nonsignificant list compared with a random permutation of the signifcant list vs. nonsignificant list

Pathway	Overlapping putative metabolites[1]	All metabolites[2*]	p-values
pyrimidine deoxyribonucleotides *de novo* biosynthesis I	12	15	5.4e-3
glycolysis I (from glucose 6-phosphate)	11	14	7.8e-3
glycolysis II (from fructose 6-phosphate)	11	14	7.8e-3
purine deoxyribonucleosides degradation I	7	8	1.0e-2
UDP-*N*-acetyl-D-glucosamine biosynthesis I	8	10	1.5e-2

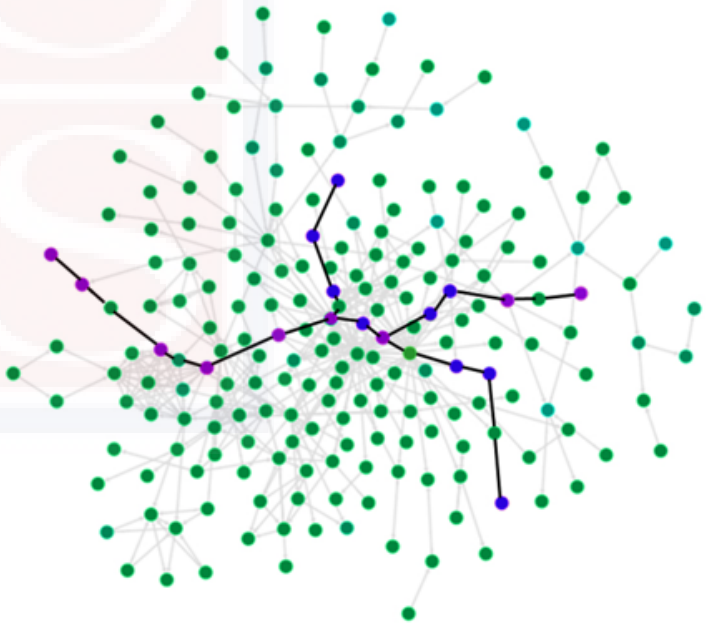
Mummichog

FET:

Statistical significance of matched significant list vs. nonsignificant list compared with a random permutation of the signifcant list vs. nonsignificant list

Significant

439.0639
207.0759
583.2114
616.1190
182.0833
300.0444
314.0600
260.5147
495.2696
152.0722
518.1004
437.0628



Feature Mapping

Automated Workflows using Mummichog:

Raw MS data

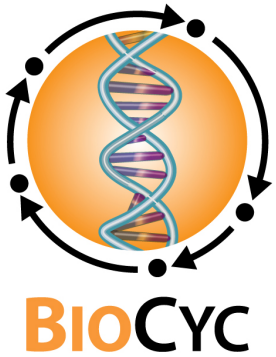
Processed data

Huan et al Nature Methods, 2017
Forsberg et al Nature Protocol 2018

Chong et al Nucleic Acids Research 2017

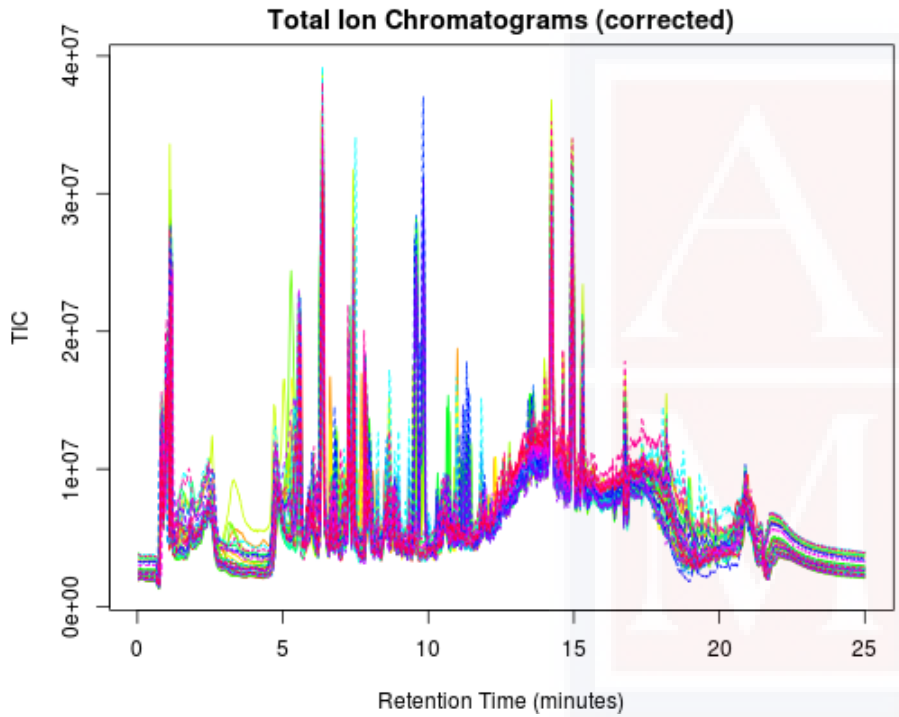
User uploads raw MS data

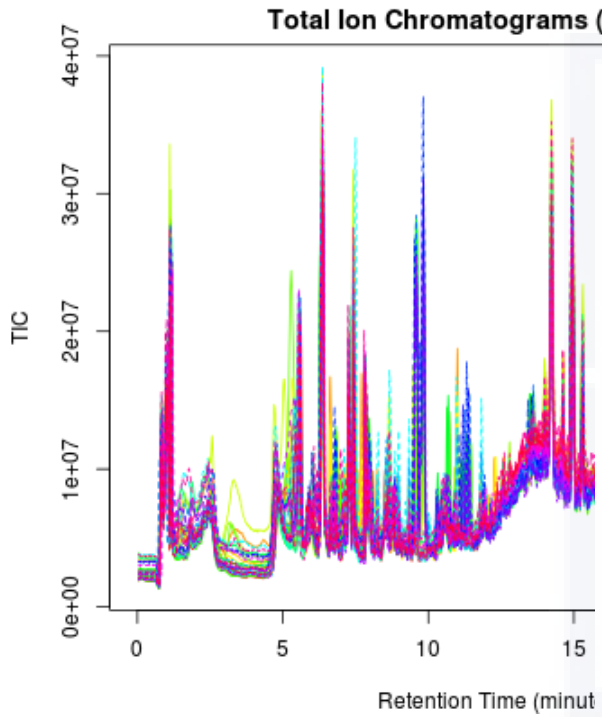
Mummichog



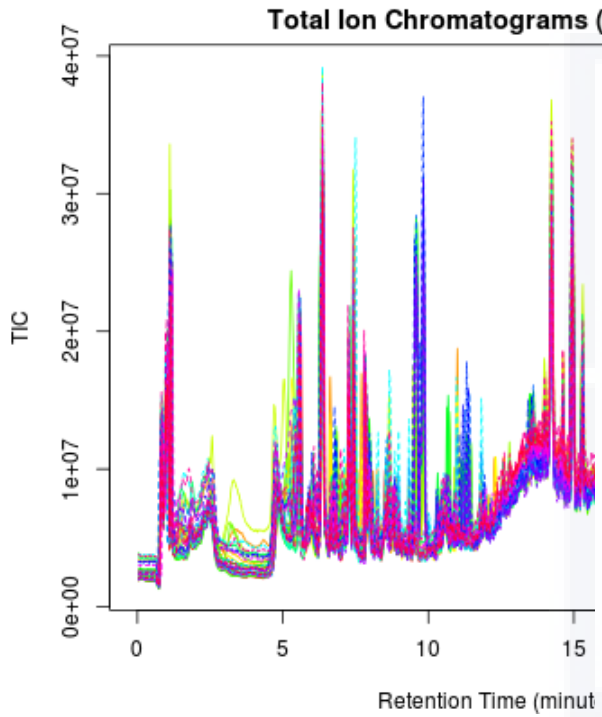
7600 Biosources – model organisms

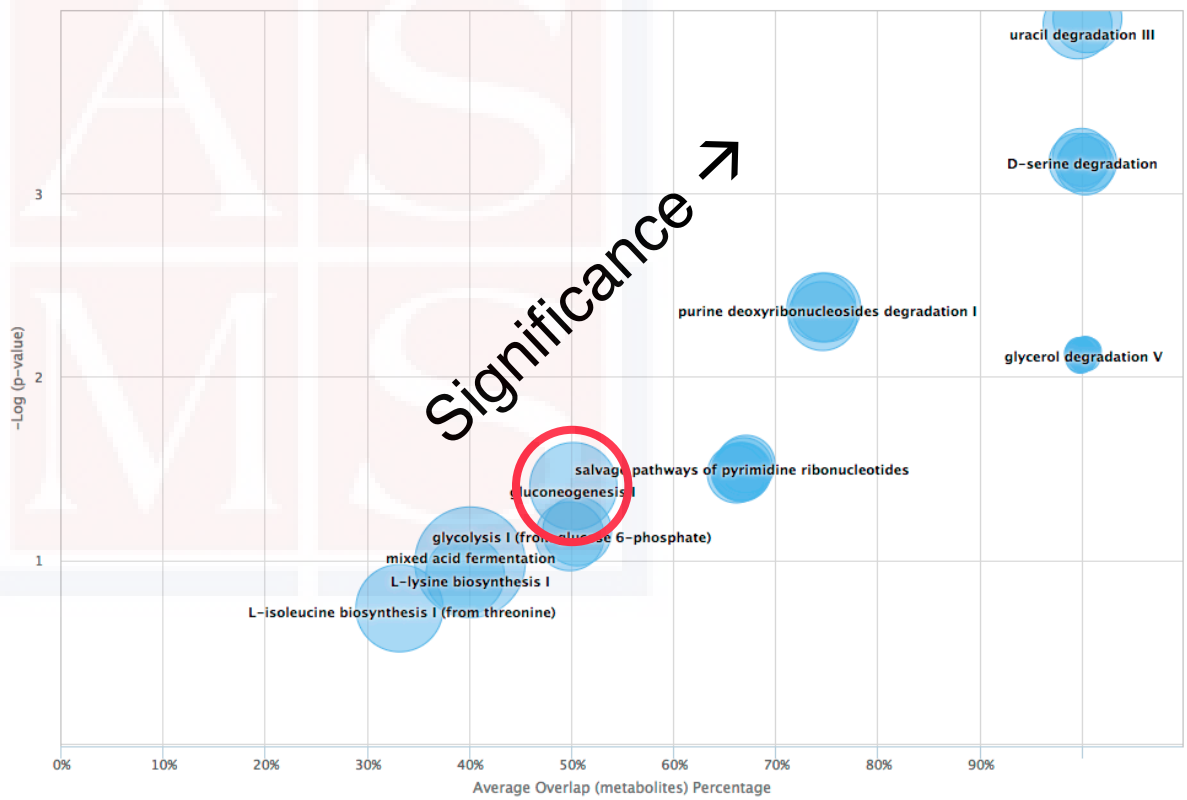
Li et al, PLoS Computational Biology, 2013


Total Ion Chromatograms (corrected)
TIC
0e+00
1e+07
2e+07
3e+07
4e+07
0
5
10
15
20
25
Retention Time (minutes)

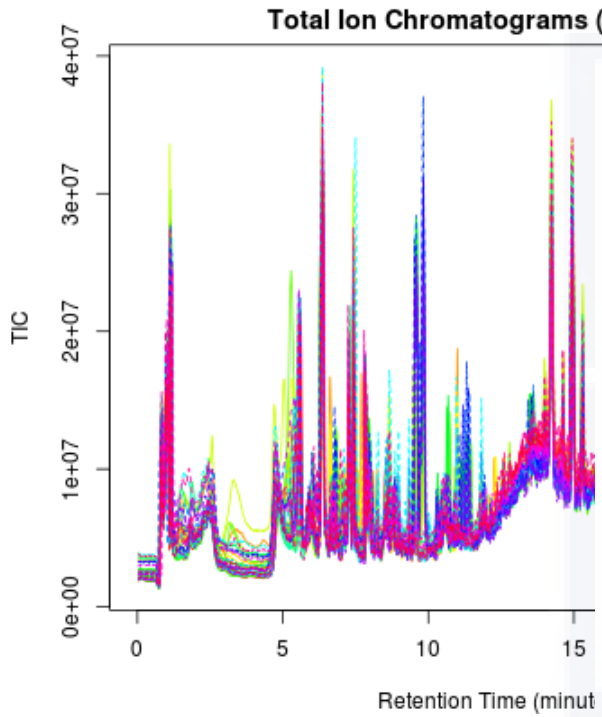


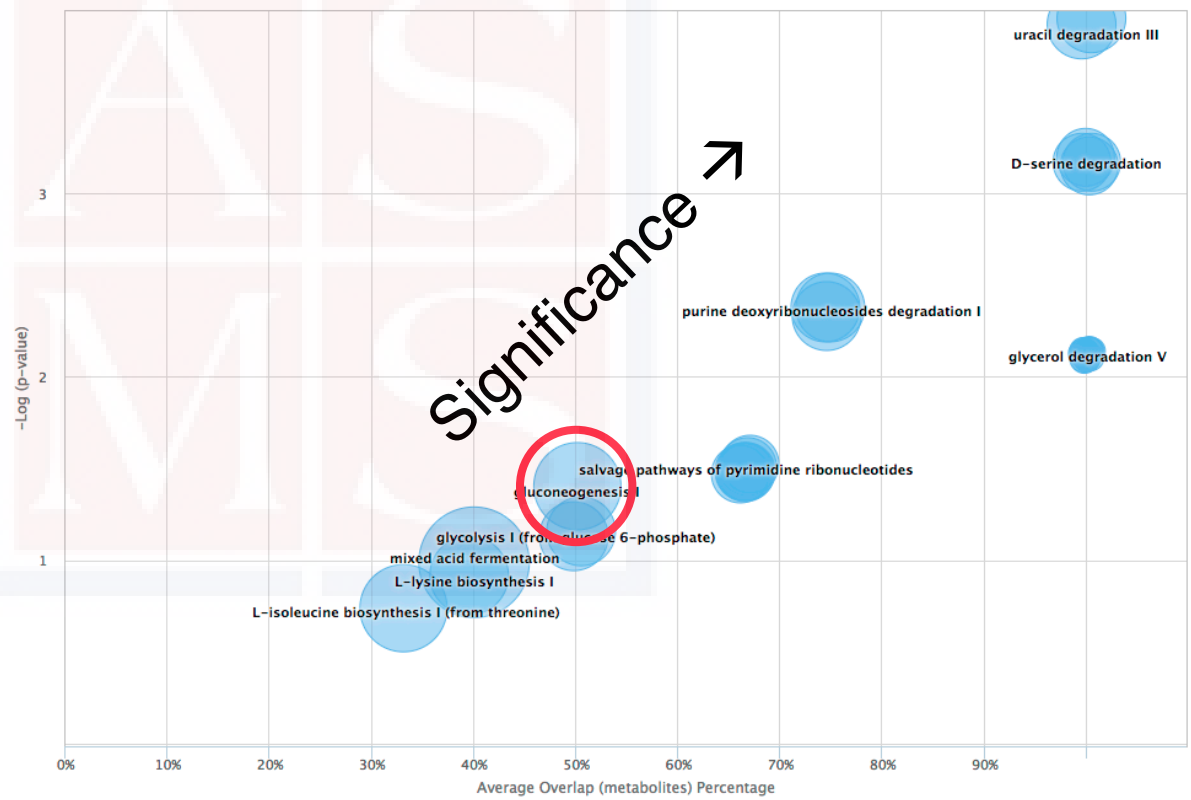
Huan et al Nature Methods, 2017
Forsberg et al Nature Protocol 2018





Huan et al Nature Methods, 2017
Forsberg et al Nature Protocol 2018





Pathway Name	Overlapping Genes	Overlapping Proteins	Overlapping Metabolites *	p-value
gluconeogenesis I	10	8	3	0.042
glycolysis I (from glucose 6-phosphate)	12	7	2	0.075
glycolysis II (from fructose 6-phosphate)	11	6	2	0.075

Huan et al Nature Methods, 2017
Forsberg et al Nature Protocol 2018

Metabolite Overlap with Pathway: gluconeogenesis I

Search:

Metabolites	METLIN ID	KEGG ID	Dysregulation	Fold Change	p-value	*m/z*	Retention Time	Adduct Form	Feature Details
(*S*)-malate									
	118	C00149	"DOWN"	2.7	1.1e-6	115.0037	33.69	M-H2O-H[-]	352
	118	C00149	"DOWN"	3.7	6.4e-7	133.0143	33.68	M-H[-]	307
	118	C00149	"DOWN"	15.8	1.4e-4	168.9911	46.97	M+Cl[-]	1279
2-phospho-D-glycerate									
	151	C00631	"DOWN"	3.1	1.6e-4	184.9856	42.83	M-H[-]	1299
	151	C00631	"DOWN"	2.5	6.7e-5	184.9855	43.49	M-H[-]	1092
3-phospho-D-glycerate									
	150	C00197	"DOWN"	3.1	1.6e-4	184.9856	42.83	M-H[-]	1299
	150	C00197	"DOWN"	2.5	6.7e-5	184.9855	43.49	M-H[-]	1092
D-glyceraldehyde 3-phosphate									
	3294	C00118	"DOWN"	15.8	1.4e-4	168.9911	46.97	M-H[-]	1279
fructose 1,6-bisphosphate									
	147	C00354	"DOWN"	31.3	9.8e-5	338.9892	47.04	M-H[-]	1196
	147	C00354	"DOWN"	15.8	1.4e-4	168.9911	46.97	M-2H[2-]	1279
glycerone phosphate									
	148	C00111	"DOWN"	15.8	1.4e-4	168.9911	46.97	M-H[-]	1279

Metabolite Overlap with Pathway: gluconeogenesis I

Search:

Metabolites	METLIN ID	KEGG ID	Dysregulation	Fold Change	p-value	*m/z*	Retention Time	Adduct Form	Feature Details
(*S*)-malate									
	118	C00149	"DOWN"	2.7	1.1e-6	115.003	33.69	M-H2O-H[-]	352
	118	C00149	"DOWN"	3.7	6.4e-7	133.014	33.68	M-H[-]	307
	118	C00149	"DOWN"	15.8	1.4e-4	168.991	46.97	M+Cl[-]	1279
2-phospho-D-glycerate									
	151	C00631	"DOWN"	3.1	1.6e-4	184.985	42.83	M-H[-]	1299
	151	C00631	"DOWN"	2.5	6.7e-5	184.985	43.49	M-H[-]	1092
3-phospho-D-glycerate									
	150	C00197	"DOWN"	3.1	1.6e-4	184.985	42.83	M-H[-]	1299
	150	C00197	"DOWN"	2.5	6.7e-5	184.985	43.49	M-H[-]	1092
D-glyceraldehyde 3-phosphate									
	3294	C00118	"DOWN"	15.8	1.4e-4	168.991	46.97	M-H[-]	1279
fructose 1,6-bisphosphate									
	147	C00354	"DOWN"	31.3	9.8e-5	338.989	47.04	M-H[-]	1196
	147	C00354	"DOWN"	15.8	1.4e-4	168.991	46.97	M-2H[2-]	1279
glycerone phosphate									
	148	C00111	"DOWN"	15.8	1.4e-4	168.991	46.97	M-H[-]	1279

Feature Mapping

Chong et al Nucleic Acids Research 2018

Feature Mapping

21 Model Systems:
Mammals
Plant

Chong et al Nucleic Acids Research 2018

Feature Mapping

21 Model Systems:
Mammals
Plant

Upload a peak list profile

Mass Accuracy (ppm): 0.1 (editable)

Analytical Mode: Positive Mode Negative Mode

P-value Cutoff: 1.0E-4 (editable)

Choose Data File: Choose File export.txt

Chong et al Nucleic Acids Research 2018

Feature Mapping

21 Model Systems:
Mammals
Plant

m.z	p.value	t.score
135.029882	1.133E-11	24.02133688
145.0504748	1.59743E-11	-16.75295439
526.2424105	7.55667E-11	-23.752828
135.0314479	1.33987E-10	41.06575322
135.03129	1.89734E-10	39.10495648
210.0384291	4.33849E-10	-34.49820828
606.074636	9.31523E-10	-21.90779617
89.02423101	1.18565E-09	-31.22736519
88.04032549	1.57119E-09	-27.93033212
105.0192214	2.11853E-09	-14.67692019
170.0459139	3.1135E-09	-27.69435678
165.0401703	4.61572E-09	21.37085582
191.0560661	4.66254E-09	-25.21491821
98.02469887	6.56477E-09	-25.06503599
966.2797467	7.35356E-09	-20.12428665
243.0986141	7.39342E-09	-24.55919426
245.114304	9.86668E-09	23.57684187
160.0613397	1.62841E-08	22.46107258
187.0012806	1.95217E-08	-21.83924786
147.0297863	2.00281E-08	-15.53239635
128.0330011	2.06351E-08	-10.39857714
155.0089072	2.1046E-08	16.17600945

Chong et al Nucleic Acids Research 2018

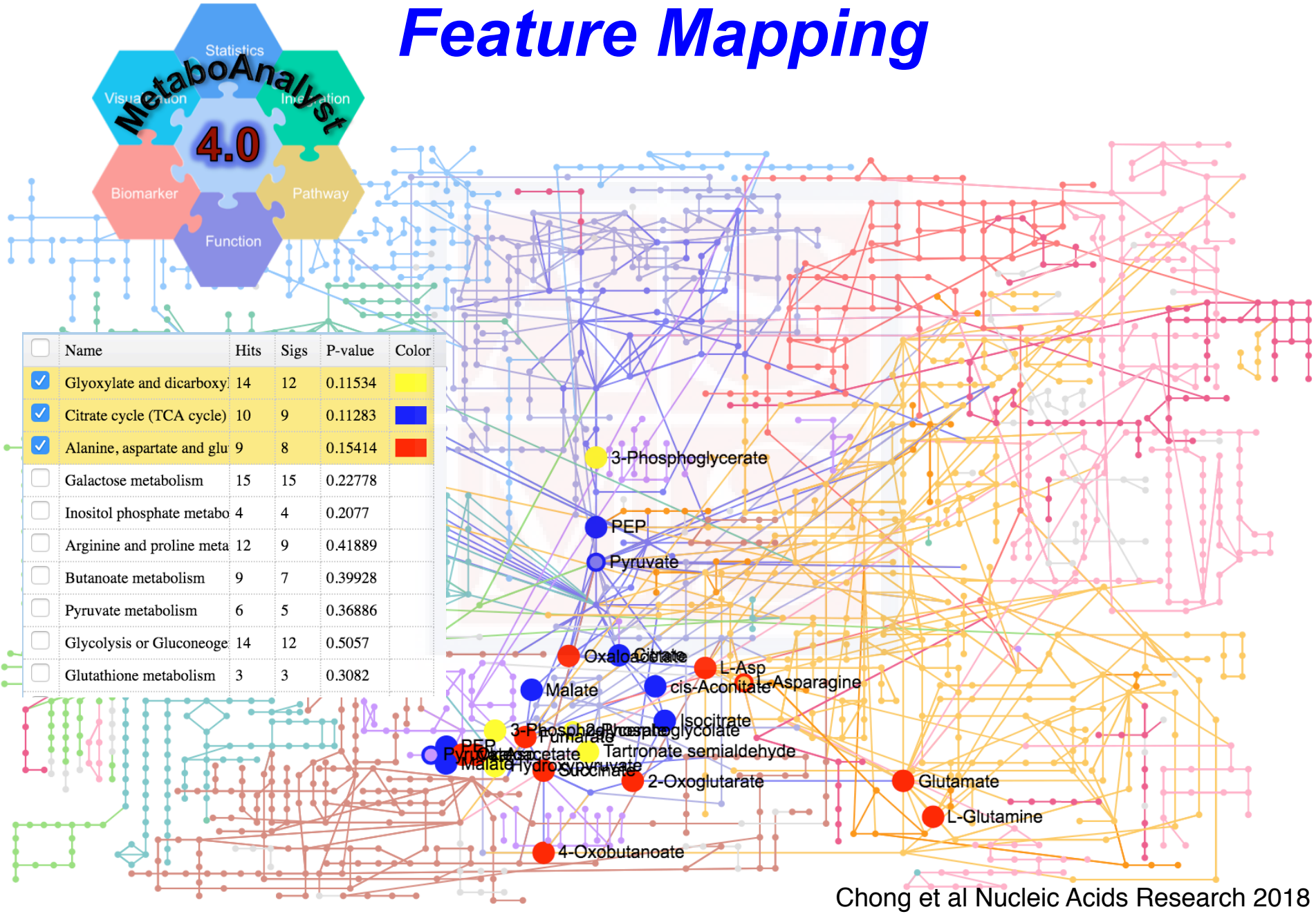
Feature Mapping

21 Model Systems:
Mammals
Plant

Pathway Name	Total	Hits (all)	Hits (sig.)	Fisher's P	EASE Score	Gamma P
Glyoxylate and dicarboxylate metabolism	29	14	12	0.11534	0.27603	0.0016258
Citrate cycle (TCA cycle)	20	10	9	0.11283	0.31351	0.0019478
Alanine, aspartate and glutamate metabolism	18	9	8	0.15414	0.39319	0.0028784
Galactose metabolism	37	15	12	0.22778	0.42515	0.003376
Inositol phosphate metabolism	8	4	4	0.2077	0.60543	0.0086507
Arginine and proline metabolism	41	12	9	0.41889	0.65342	0.011297
Butanoate metabolism	18	9	7	0.39928	0.66905	0.012349
Pyruvate metabolism	26	6	5	0.36886	0.69539	0.014386
Glycolysis or Gluconeogenesis	29	14	10	0.5057	0.71303	0.015969
Glutathione metabolism	21	3	3	0.3082	0.75083	0.020105
Glycerolipid metabolism	14	3	3	0.3082	0.75083	0.020105

Chong et al Nucleic Acids Research 2018

Feature Mapping



	Name	Hits	Sigs	P-value	Color
☑	Glyoxylate and dicarboxyl	14	12	0.11534	
☑	Citrate cycle (TCA cycle)	10	9	0.11283	
☑	Alanine, aspartate and glu	9	8	0.15414	
☐	Galactose metabolism	15	15	0.22778	
☐	Inositol phosphate metabo	4	4	0.2077	
☐	Arginine and proline meta	12	9	0.41889	
☐	Butanoate metabolism	9	7	0.39928	
☐	Pyruvate metabolism	6	5	0.36886	
☐	Glycolysis or Gluconeoge	14	12	0.5057	
☐	Glutathione metabolism	3	3	0.3082	

Chong et al Nucleic Acids Research 2018

Networking

PIUMet

Online: Uses untargeted data to perform pathway overlap and gain network information; takes both pos and neg MS data

http://fraenkel-nsf.csbi.mit.edu/piumet2/
Pirhaji et al, Nature Methods, 2016

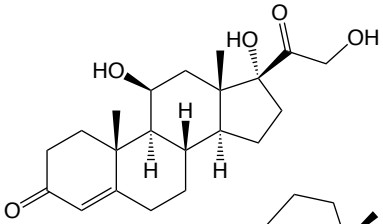
Metscape

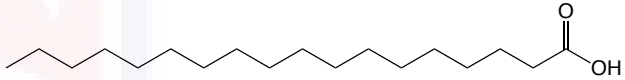
Cytoscape
Network Visualization and Analysis

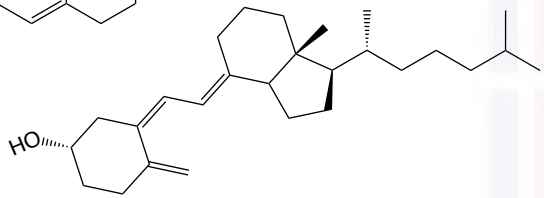
http://fraenkel-nsf.csbi.mit.edu/piumet2/
Pirhaji et al, Nature Methods, 2016

Metabolite Coverage

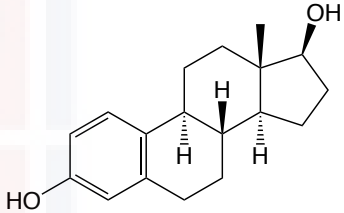
Untargeted = Unbiased?



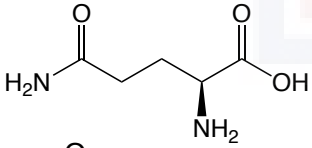


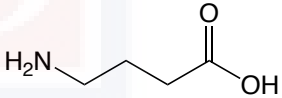


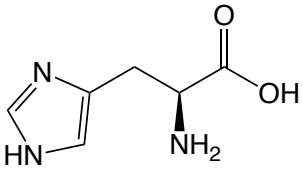
Hydrophobic

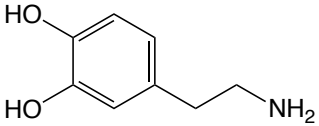


Hydrophilic

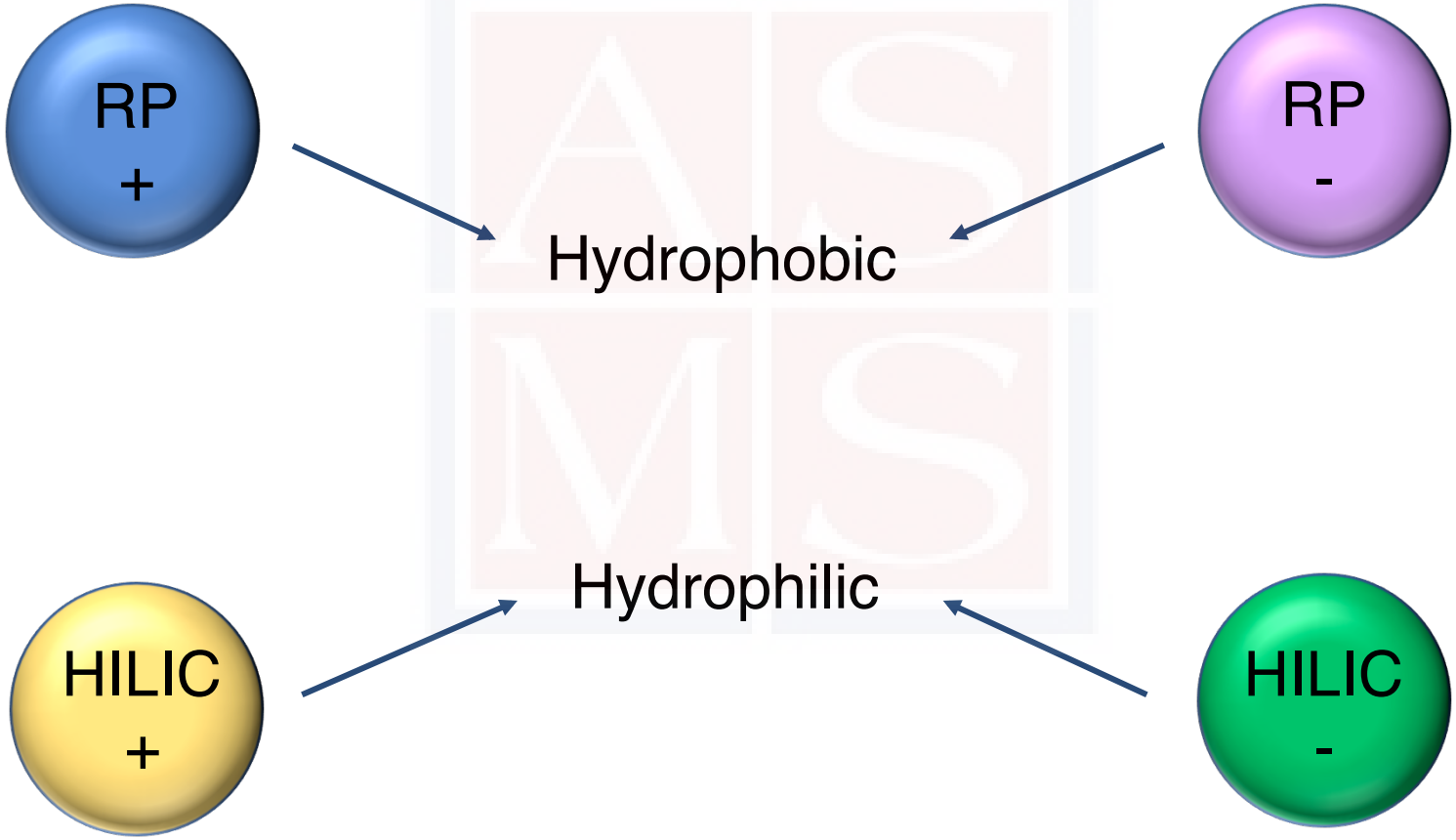




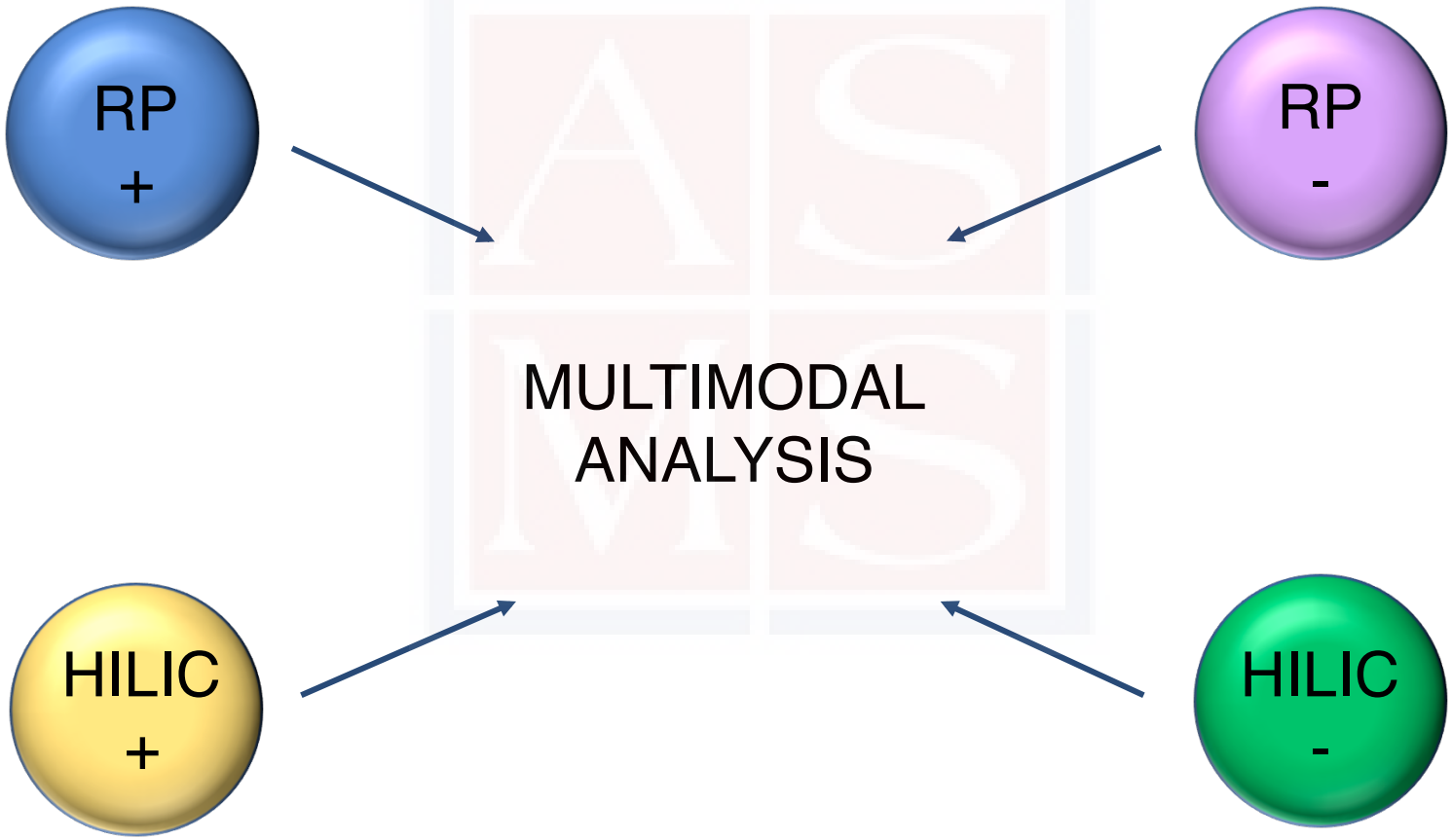




Mummichog



Mummichog

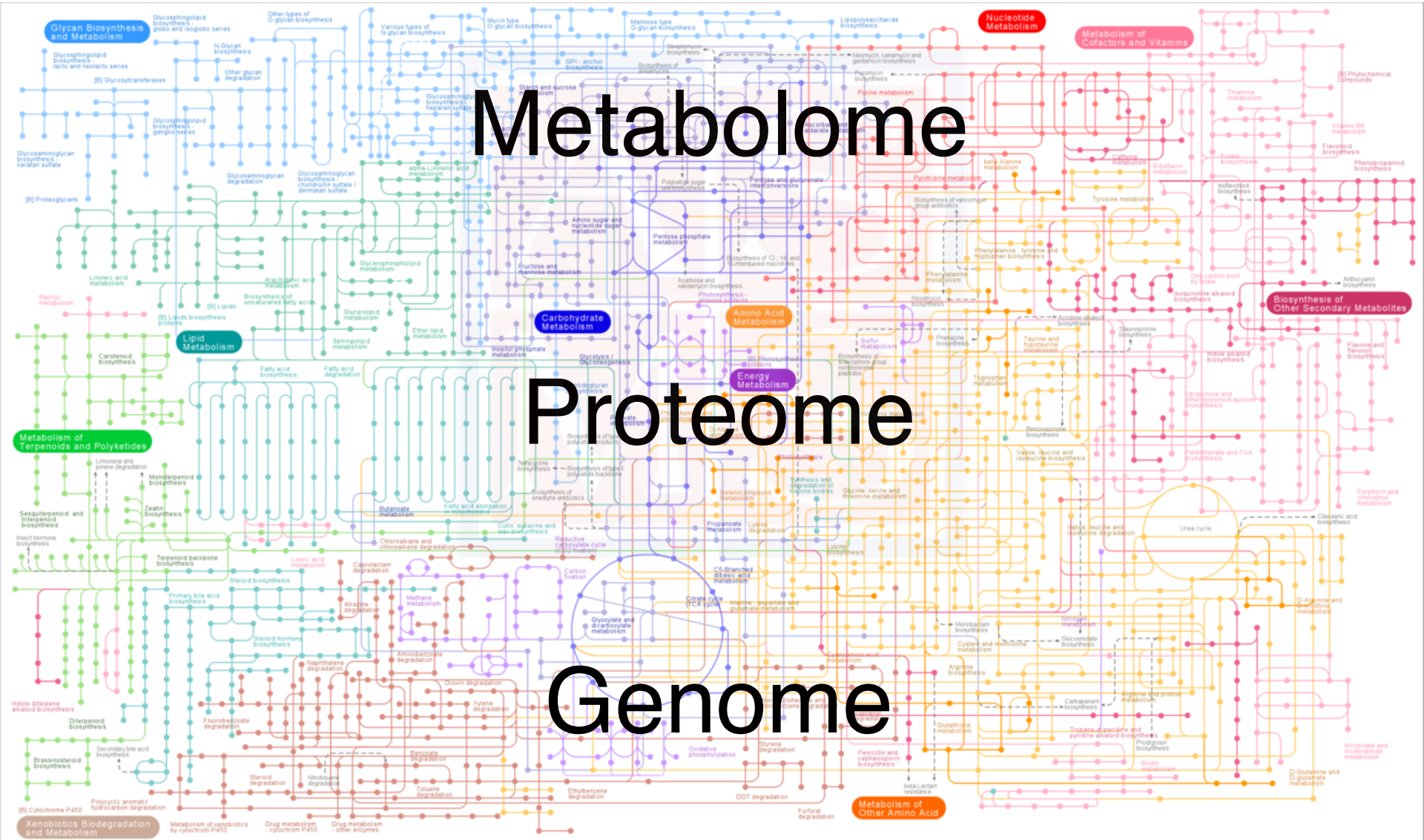


Input both +ve and –ve mode data

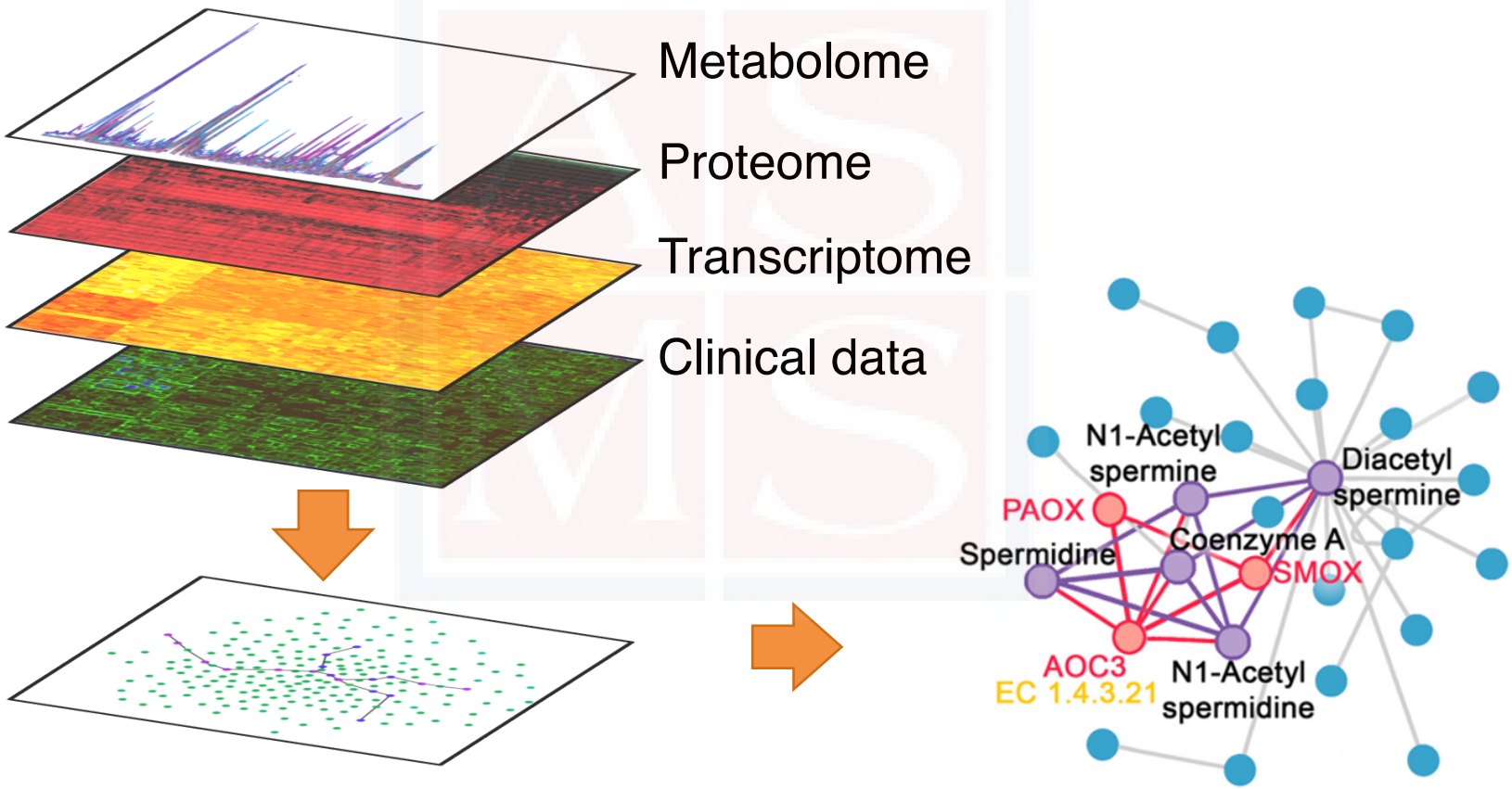
Pathway Analysis and Multi-Omic Integration

1. Prerequisites
2. Biomarkers vs. Biological Relevance
3. Pathway Tools for Annotated Data
4. Pathway Tools for Unannotated Data
5. **Multi-Omic Integration**

Multi-Omic Integration

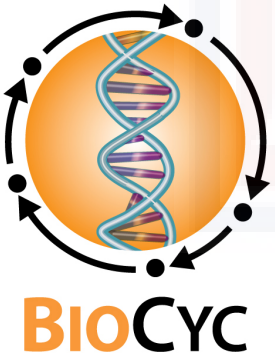


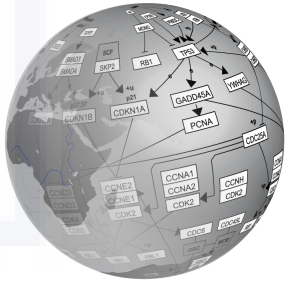
Mummichog


Metabolome
Proteome
Transcriptome
Clinical data
N1-Acetyl spermine
Diacetyl spermine
PAOX
Spermidine
Coenzyme A
SMOX
AOC3
EC 1.4.3.21
N1-Acetyl spermidine

Mummichog

Need access to more than just pathways





IMPaLA

Integrated Molecular Pathway Level Analysis

Gene Data

MDHM_HUMAN	-1.16	2.51
MDHC_HUMAN	0.25	1.14
DLDH_HUMAN	-0.82	3.20
DHSA_HUMAN	-0.64	2.59
DHSB_HUMAN	1.69	2.46
C560_HUMAN	-0.75	0.03
DHSD_HUMAN	-0.05	1.81
ODO2_HUMAN	-1.27	2.79
ODO1_HUMAN	-0.62	2.80
CISY_HUMAN	-0.44	3.13
ACON_HUMAN	-1.82	2.03
IDH3A_HUMAN	-1.37	2.24
IDH3B_HUMAN	-0.78	0.64

C00002	0.25	3.61
C00011	1.39	1.23
C00001	-1.14	0.84
C00004	0.43	1.05
C00080	0.41	3.77
C00003	0.94	-0.63
C00008	-0.13	0.04
C00009	-0.71	1.49
C00024	-0.93	1.08
C00010	-1.33	0.56
C00122	0.65	1.75
C00026	-0.27	3.33
C00042	-0.38	0.84

Metabolite Data

Kamurov et al, Bioinformatics 2011

IMPaLA

Integrated Molecular Pathway Level Analysis

Gene Data

MDHM_HUMAN	-1.16	2.51
MDHC_HUMAN	0.25	1.14
DLDH_HUMAN	-0.82	3.20
DHSA_HUMAN	-0.64	2.59
DHSB_HUMAN	1.69	2.46
C560_HUMAN	-0.75	0.03
DHSD_HUMAN	-0.05	1.81
ODO2_HUMAN	-1.27	2.79
ODO1_HUMAN	-0.62	2.80
CISY_HUMAN	-0.44	3.13
ACON_HUMAN	-1.82	2.03
IDH3A_HUMAN	-1.37	2.24
IDH3B_HUMAN	-0.78	0.64

Metabolite Data

C00002	0.25	3.61
C00011	1.39	1.23
C00001	-1.14	0.84
C00004	0.43	1.05
C00080	0.41	3.77
C00003	0.94	-0.63
C00008	-0.13	0.04
C00009	-0.71	1.49
C00024	-0.93	1.08
C00010	-1.33	0.56
C00122	0.65	1.75
C00026	-0.27	3.33
C00042	-0.38	0.84

pathway name	pathway source	overlapping genes	all genes	P_{genes}	Q_{genes}	overlapping metabolites	all metabolites	$P_{metabolites}$	$Q_{metabolites}$	P_{joint}	Q_{joint}
TCA cycle	HumanCyc	15	18 (18)	3.12e-44	4.71e-41	18	22 (23)	2.36e-45	9.94e-42	1.5e-86	3.84e-83
TCA cycle	EHMN	17	30 (30)	1.77e-46	4e-43	17	36 (36)	6.04e-36	2.12e-33	2e-79	2.56e-76
superpathway of conversion of glucose to acetyl CoA and entry into the TCA cycle	HumanCyc	16	47 (48)	1.38e-38	1.25e-35	18	34 (36)	7.09e-40	1.49e-36	1.74e-75	1.48e-72
Citric acid cycle (TCA cycle)	Reactome	14	22 (22)	3.81e-38	2.87e-35	17	30 (30)	8.44e-38	1.18e-34	5.54e-73	3.54e-70
Citrate cycle	INOH	16	32 (32)	5.52e-42	6.25e-39	16	35 (35)	4.16e-33	8.34e-31	3.91e-72	2e-69
TCA Cycle	Wikipathways	13	17 (17)	7.08e-37	4.58e-34	16	23 (24)	2.53e-37	1.53e-34	3.02e-71	1.29e-68
Pyruvate dehydrogenase deficiency (E3)	SMPDB	13	21 (21)	6.04e-35	1.44e-32	17	32 (33)	3.98e-37	1.53e-34	3.94e-69	7.75e-67

Gene List

Gene list with optional fold changes

```
#Entrez  logFC
1737  -1.277784317
83440  -1.034136439
3939  -2.231729728
10911  -1.045657875
10690  -0.968308832
10010  -0.861541301
11224  1.187399591
63826  -1.405238611
11031  0.785011172
4190  -1.778774832
10782  -2.140715987
10993  -0.925083829
10455  1.732172706
10963  1.177511121
10282  -1.20754269
```

ID Type: (Human) Entrez ID

Metabolite List

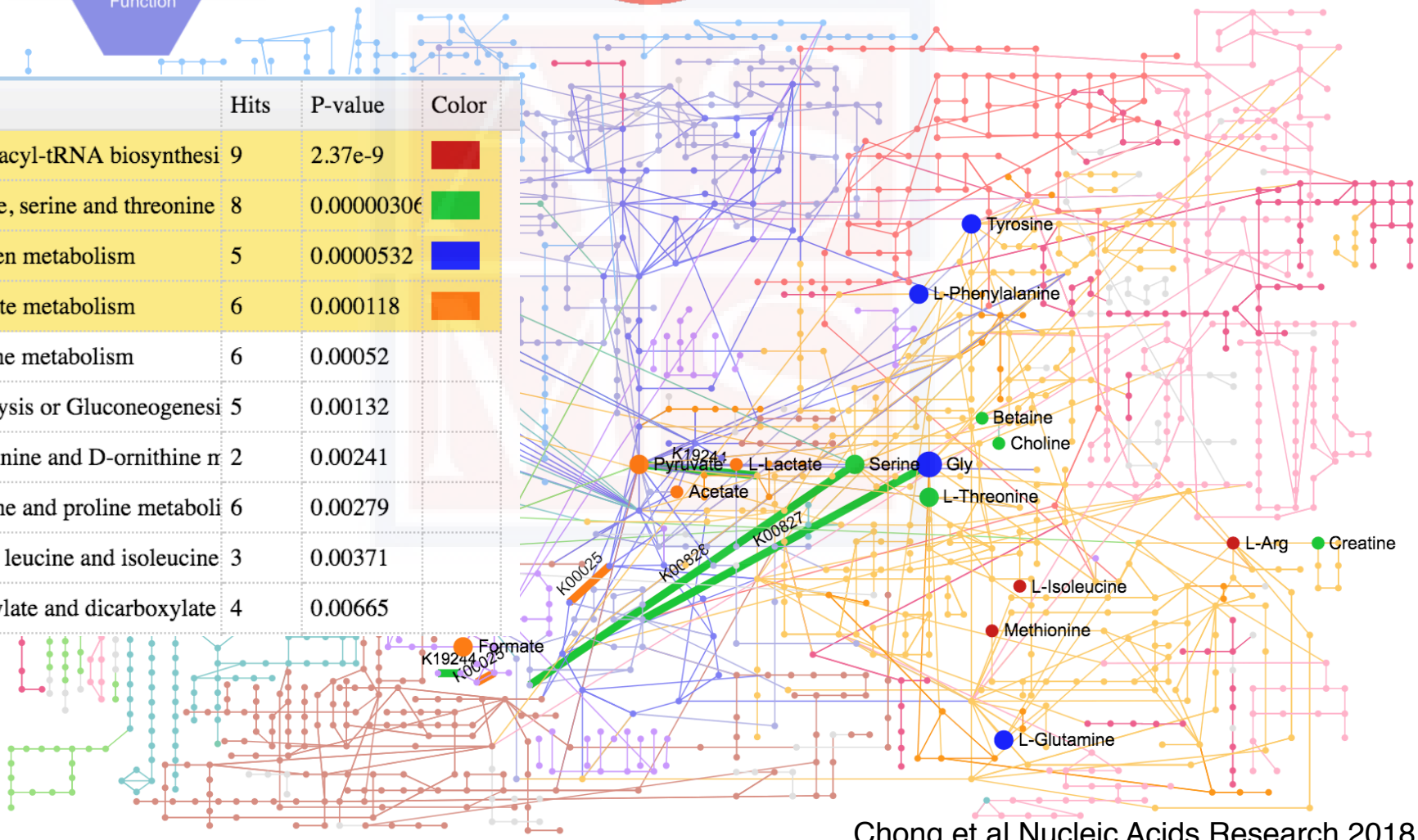
Compound list with optional fold changes

```
#KEGG  logFC
C00116  1.010972619
C00565  -0.714283001
C00033  0.822193121
C00583  -1.005192252
C00022  -0.623838569
C00719  -0.406052491
C05984  -0.390152174
C00207  -0.932835099
C00065  0.903658797
C00031  0.548035915
C00079  0.416744818
C02632  -0.515041676
C00064  -0.497216411
C00114  1.102078837
C00073  0.516193785
```

ID Type: KEGG ID

Chong et al Nucleic Acids Research 2018

	Name	Hits	P-value	Color
☑	Aminoacyl-tRNA biosynthesi	9	2.37e-9	
☑	Glycine, serine and threonine	8	0.00000306	
☑	Nitrogen metabolism	5	0.0000532	
☑	Pyruvate metabolism	6	0.000118	
☐	Methane metabolism	6	0.00052	
☐	Glycolysis or Gluconeogenesi	5	0.00132	
☐	D-Arginine and D-ornithine m	2	0.00241	
☐	Arginine and proline metaboli	6	0.00279	
☐	Valine, leucine and isoleucine	3	0.00371	
☐	Glyoxylate and dicarboxylate	4	0.00665	



Chong et al Nucleic Acids Research 2018

Data Upload

Data Upload

FileID	Filename	Upload Date	List Type	Accession ID	Metabolic Matches	Remove
160564	Ecoli_gene	2017-05-29 18:39:00	Genes	Gene symbol	View	✕
160565	Ecoli_prot	2017-05-29 18:39:11	Proteins	✓ UNIPROT Gene symbol	View	✕

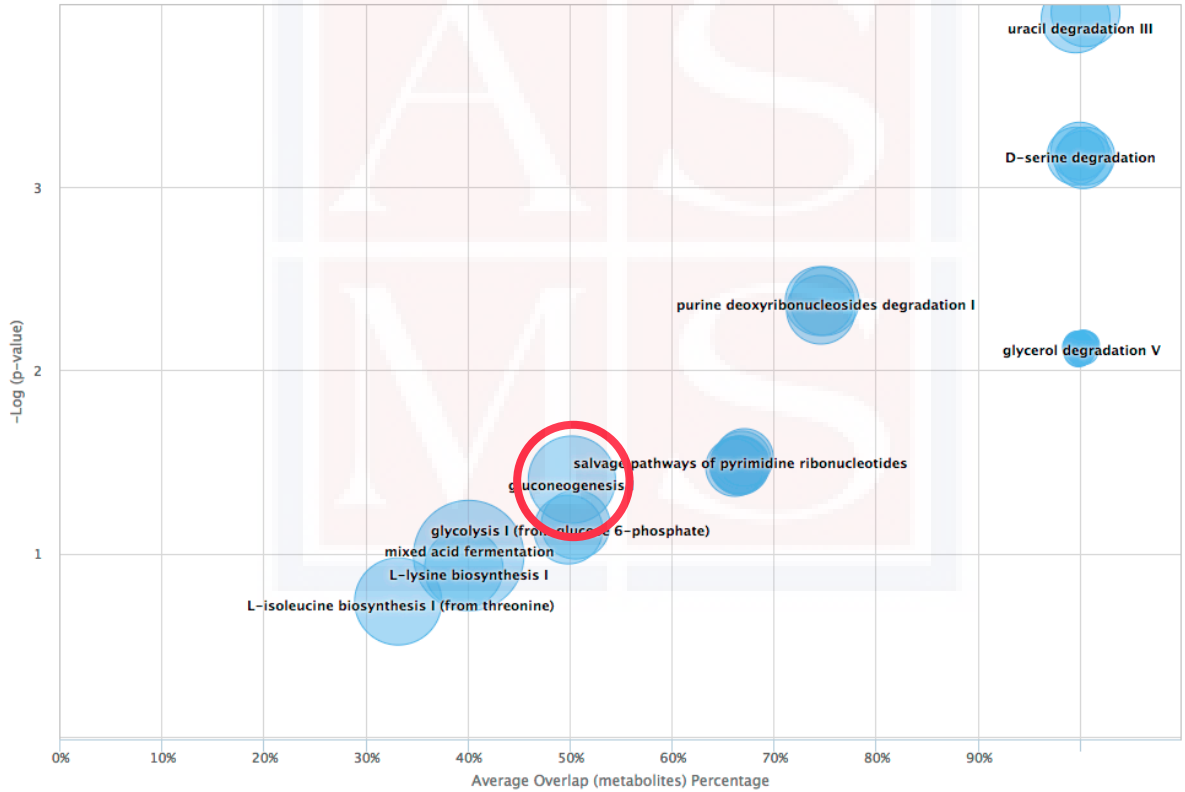
Run matching subjobs

Huan et al Nature Methods, 2017
Forsberg et al Nature Protocol 2018

Overlap Table

Pathway	Overlapping genes	All genes*	Overlapping proteins	All proteins*	Overlapping putative metabolites[1]	All metabolites[2]*	p-values
gluconeogenesis I	10	17	8	17	3	6	4.2e-2
glycolysis I (from glucose 6-phosphate)	12	18	7	18	2	4	7.5e-2
glycolysis II (from fructose 6-phosphate)	11	18	6	18	2	4	7.5e-2
methylglyoxal degradation II	3	3	1	2	2	3	3.2e-2
mixed acid fermentation	17	29	1	20	4	10	9.6e-2

Multi-Omic Cloud Plot



Pathway Name	Overlapping Genes	Overlapping Proteins	Overlapping Metabolites *	p-value
gluconeogenesis I	10	8	3	0.042
glycolysis I (from glucose 6-phosphate)	12	7	2	0.075
glycolysis II (from fructose 6-phosphate)	11	6	2	0.075

Advanced Metabolomics

Thank you!

Questions?

Erica Forsberg, PhD
Dept. Chemistry & Biochemistry
eforsberg@sdsu.edu

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

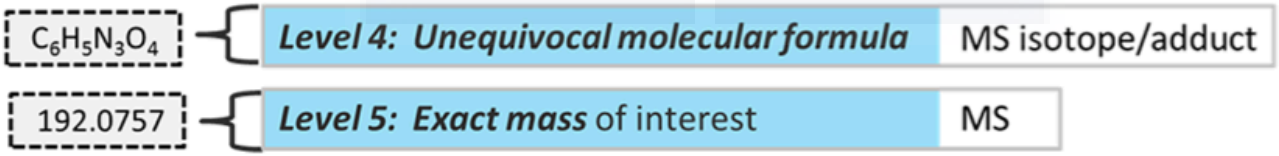
---- 02:15 pm Break ----

Identifying Metabolites: The Big Obstacle

Identifying Metabolites: The Big Obstacle
Levels of identification

Schymanski, E. et. al. *Environ. Sci. Technol.*, **2014**, *48* (4), 2097–2098
Sumner, L. W. et. al. *Metabolomics,* **2007,** 3: 211

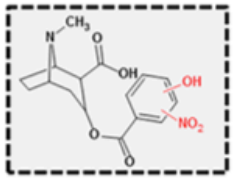
Identifying Metabolites: The Big Obstacle
Levels of identification



4

Schymanski, E. et. al. *Environ. Sci. Technol.*, **2014**, *48* (4), 2097–2098
Sumner, L. W. et. al. *Metabolomics*, **2007,** 3: 211

Identifying Metabolites: The Big Obstacle
Levels of identification



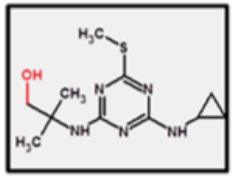
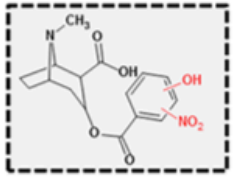
	Level 3: Tentative candidate(s) structure, substituent, class	MS, MS^2, Exp. data
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct
192.0757	***Level 5: Exact mass*** of interest	MS

3

4

Schymanski, E. et. al. *Environ. Sci. Technol.*, **2014**, *48* (4), 2097–2098
Sumner, L. W. et. al. *Metabolomics*, **2007,** 3: 211

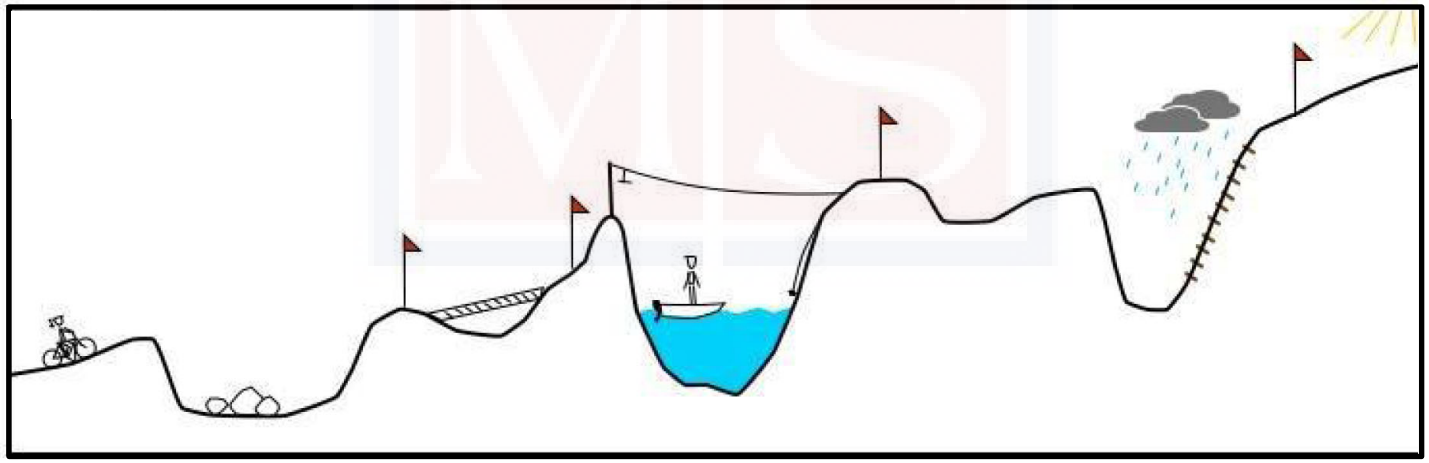
Identifying Metabolites: The Big Obstacle
Levels of identification

Example	Identification confidence	Minimum data requirements	
	***Level 1: Confirmed structure*** by reference standard	MS, MS^2, RT, Reference Std.	**1**
	Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MS^2, Library MS^2 MS, MS^2, Exp. data	**2**
	***Level 3: Tentative candidate(s)*** structure, substituent, class	MS, MS^2, Exp. data	**3**
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct	**4**
192.0757	***Level 5: Exact mass*** of interest	MS	

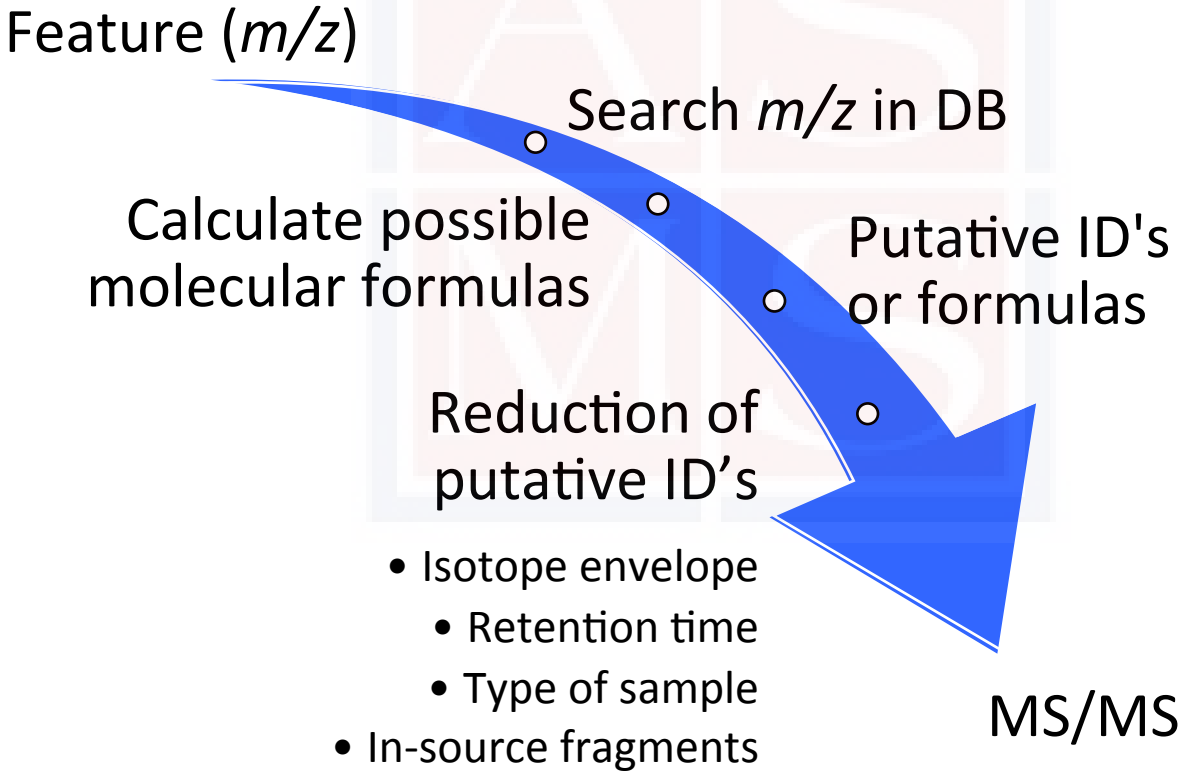
Schymanski, E. et. al. *Environ. Sci. Technol.*, **2014**, *48* (4), 2097–2098
Sumner, L. W. et. al. *Metabolomics*, **2007,** 3: 211

Identifying Metabolites: The Big Obstacle

Identifying Metabolites: The Big Obstacle



Identifying Metabolites: The Big Obstacle Identification prior to MS/MS



Identifying Metabolites: The Big Obstacle Identification with MS/MS

MS/MS (CID)

Compare MS/MS of putative ID's in DB

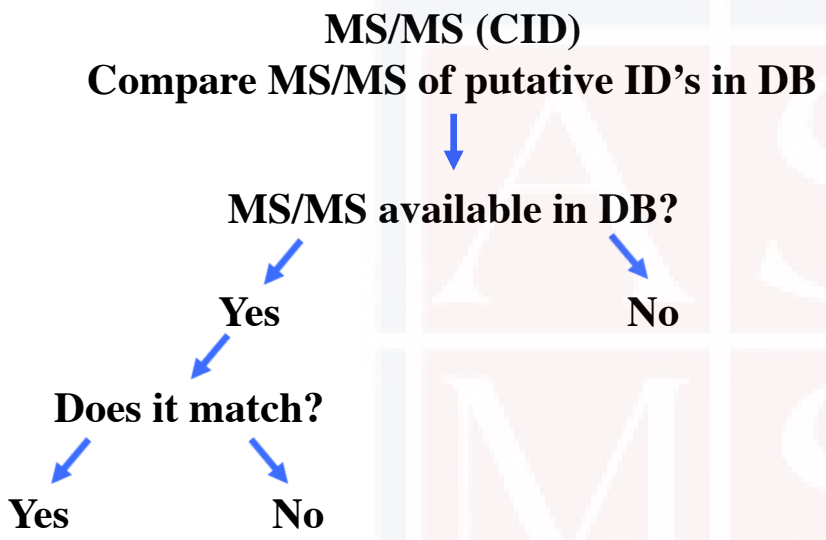
↓

MS/MS available in DB?

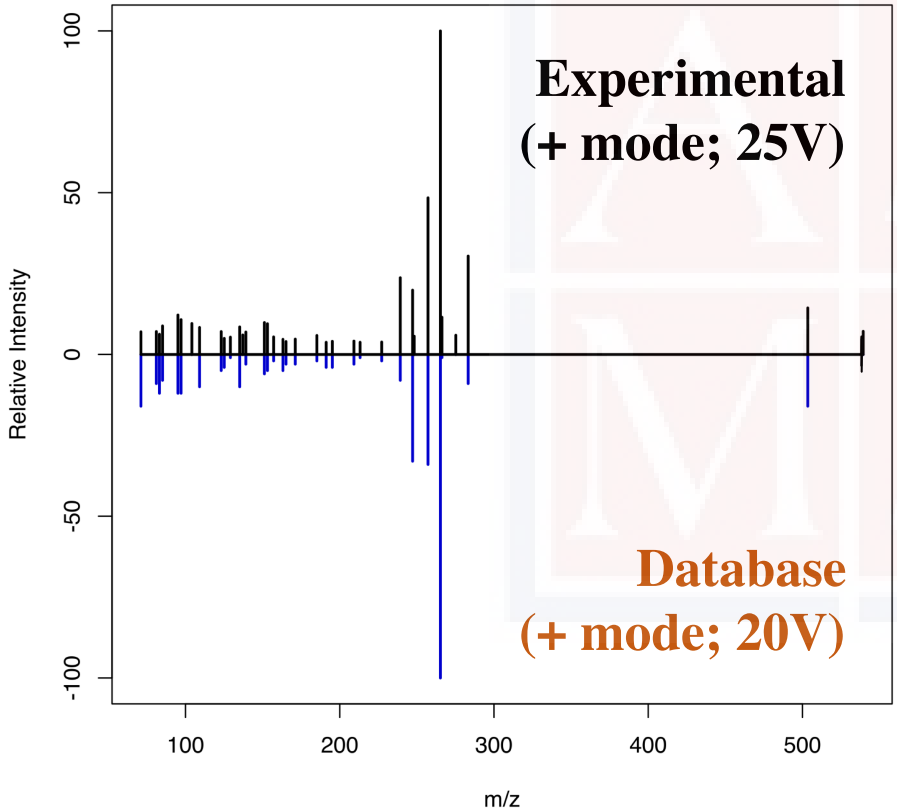
↙ ↘

Yes **No**

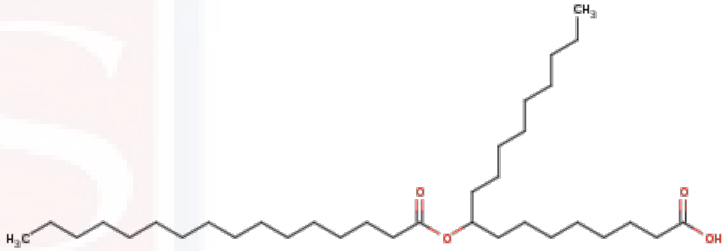
Identifying Metabolites: The Big Obstacle Identification with MS/MS



Example 1: 9-PAHSA

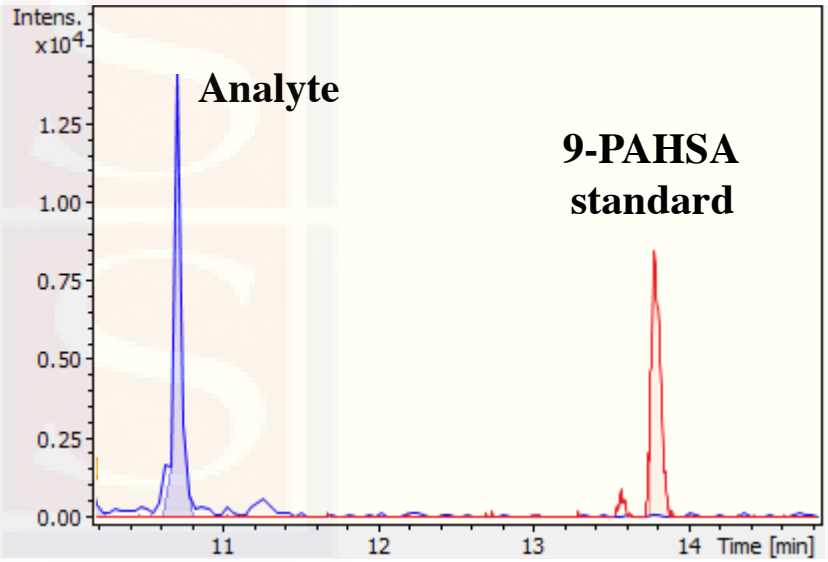


Is this a match?

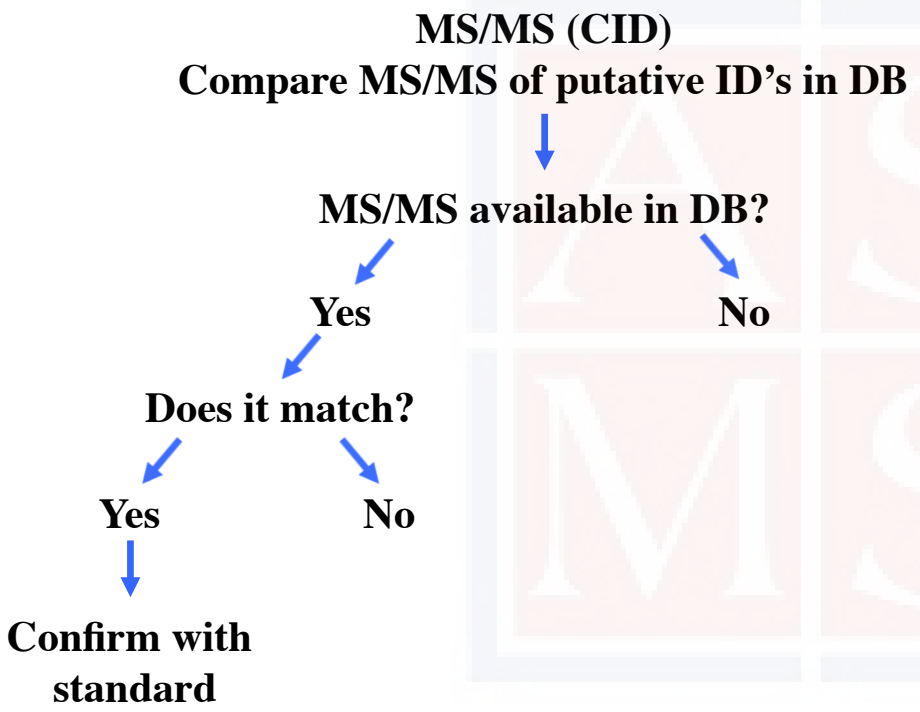


Example 1: 9-PAHSA

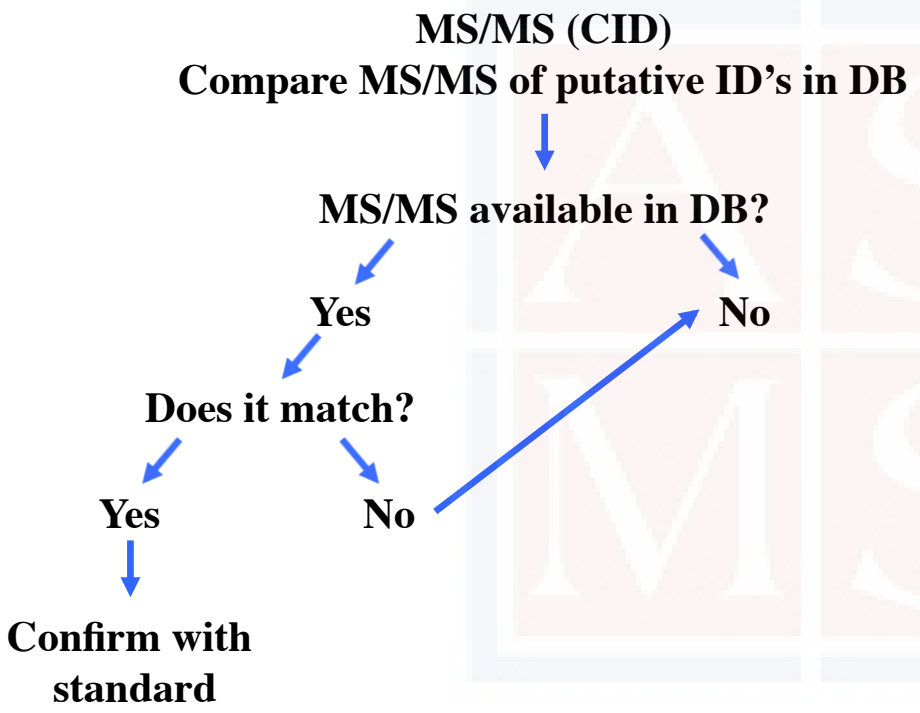
- At first glance it looks like a match
- RT of the standard did not match
- Fragmentation in negative mode did not match
- Ion ID corresponds to $[2M+H]^+$ of stearic and palmitic acid



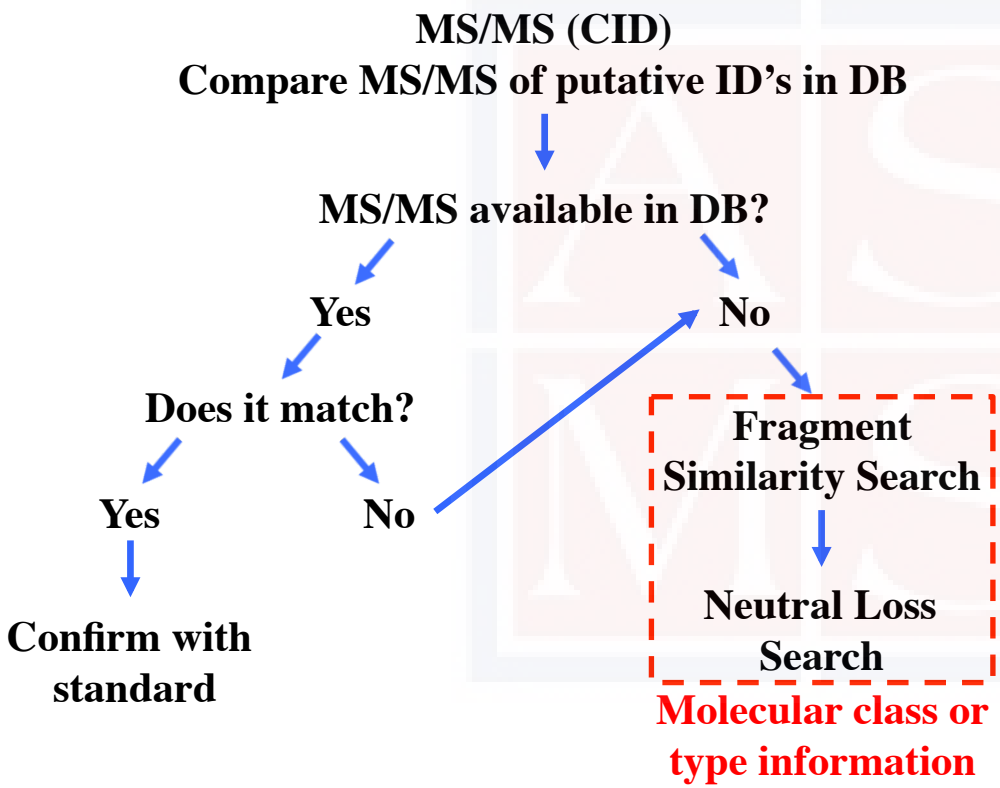
Identifying Metabolites: The Big Obstacle Identification with MS/MS



Identifying Metabolites: The Big Obstacle Identification with MS/MS



Identifying Metabolites: The Big Obstacle Identification with MS/MS



Identifying Metabolites: The Big Obstacle Fragment Similarity Search

Home | isoMETLIN | Simple Search | Advanced Search | Batch Search | Fragment Similarity Search | Neutral Loss Search | MS/MS Spectrum Match Search | MRM | Logout [rmont]

Fragment Similarity Search

Fragment M/Z (Maximum Number of M/Z is 5, separated by comma): 355.1511, 299.0892, 235.0942, 179.0341,

Tolerance: 30 PPM

Mode: Positive

Filter Out Fragments with Intensity Less than: 5 %

Order By: ΔPPM / Intensity

Fragments with Structure Only

Precursor M/Z (optional)

Search | Clear

Metabolite(s) containing 3 fragment(s)

METLIN ID: 52097 NAME: **Xanthohumol** MASS: **354.1467** View MS/MS STRUCTURE:

Show 10 entries Search:

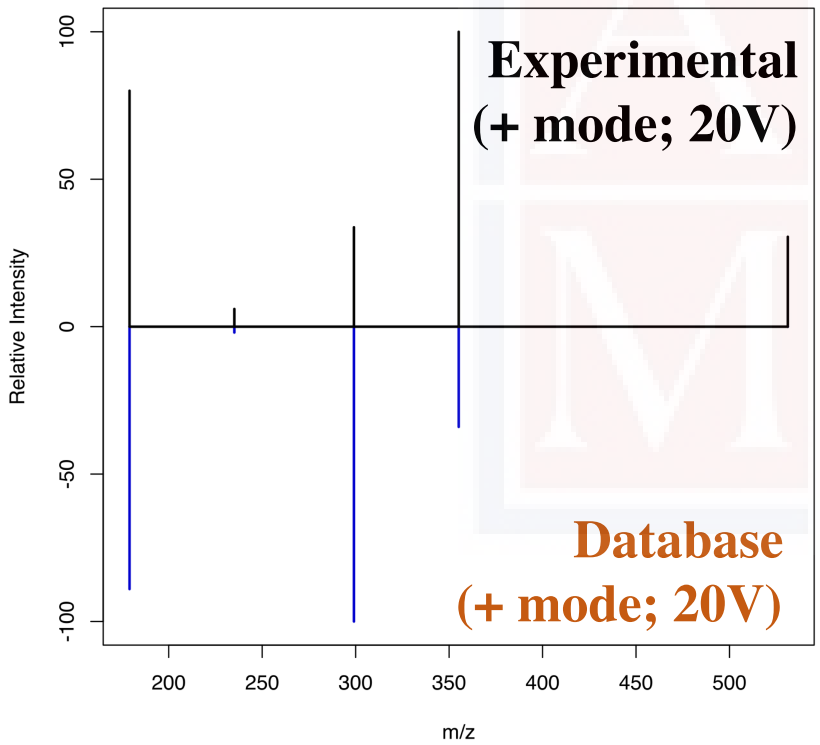
Frag. m/z	Δppm	Intensity	CE	Predicted Ion Type	Predicted Fragment Structure
179.0340	0	100.0	10, 20, 40	[M]+	
299.0890	0	100.0	10, 20	[M-H+2H]+	
355.1510	0	34.8	10	-	No Structure Information is available

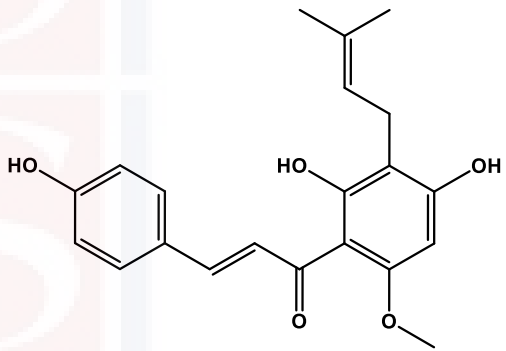
Showing 1 to 3 of 3 entries Previous 1 Next

METLIN ID: 18901 NAME: **Asn Ala Glu** MASS: **332.1332** View MS/MS STRUCTURE:

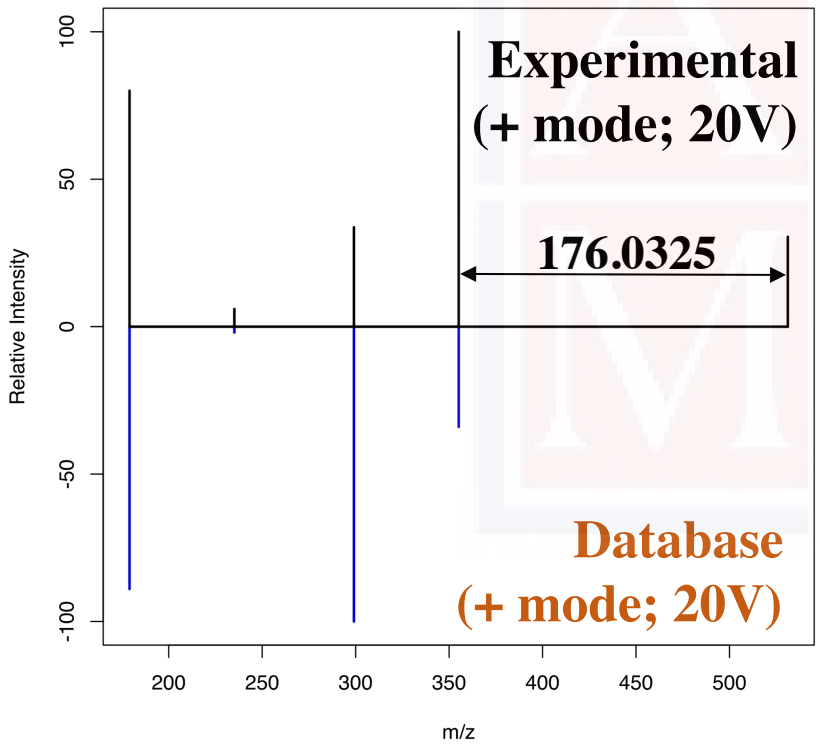
Metabolite(s) containing 2 fragment(s)

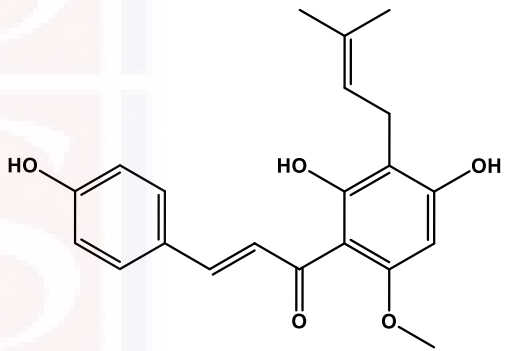
Example 2: Xanthohumol scaffold





Example 2: Xanthohumol scaffold





Identifying Metabolites: The Big Obstacle Neutral Loss Search

Home | isoMETLIN | Simple Search | Advanced Search | Batch Search | Fragment Similarity Search | Neutral Loss Search | MS/MS Spectrum Match Search | MRM | Logout [rmont]

Neutral Loss Search

Neutral Loss: 176.0325

Tolerance: 5 PPM

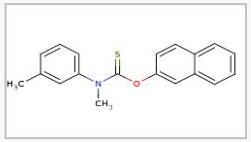
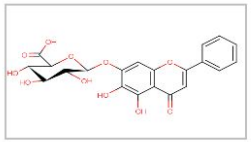
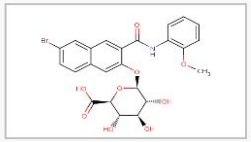
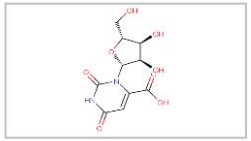
Mode: Positive

Neutral Loss with Structure Only: ☐

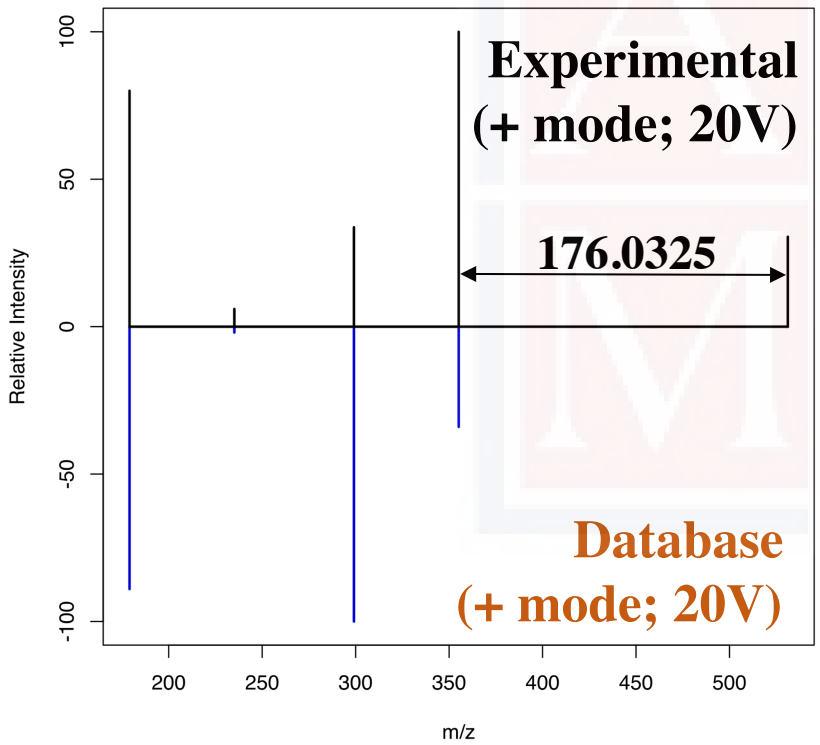
Search | Clear

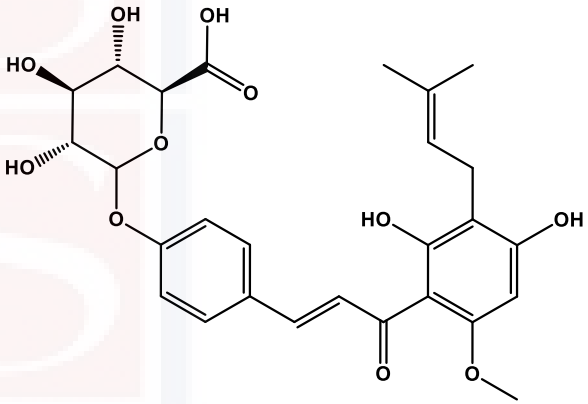
Show 10 entries

Search:

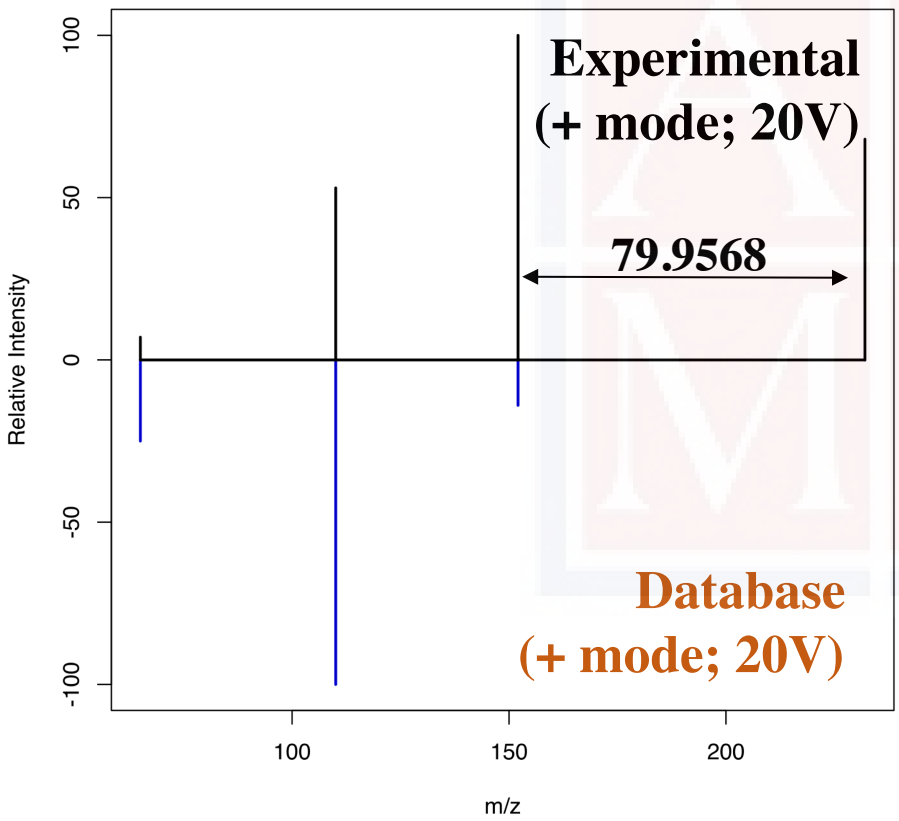
METLIN ID	Name	Neutral Loss [Fragment m/z]	Compound Mass	ΔPPM	Intensity	CE	Predicted Neutral Loss Structure	Compound Structure
43356	Tolnaftate	176.0323 [132.0781]	307.1031	0	1.9	40	No Structure Information is available	
49507	Baicalin	176.0328 [271.0594]	446.0849	1	100.0	10, 20, 40	No Structure Information is available	
4092	Naphthol AS-BI α-D-glucuronide	176.0322 [372.0229]	547.0478	1	100.0	10, 20	No Structure Information is available	
5754	Orotidine	176.0322 [135.0164]	288.0594	1	10.7	10, 20	No Structure Information is available	
4090	4-Aminophenyl 1-thio-β-D-glucuronide	176.0323 [126.0370]	301.0620	1	100.0	10, 20, 40	No Structure Information is available	

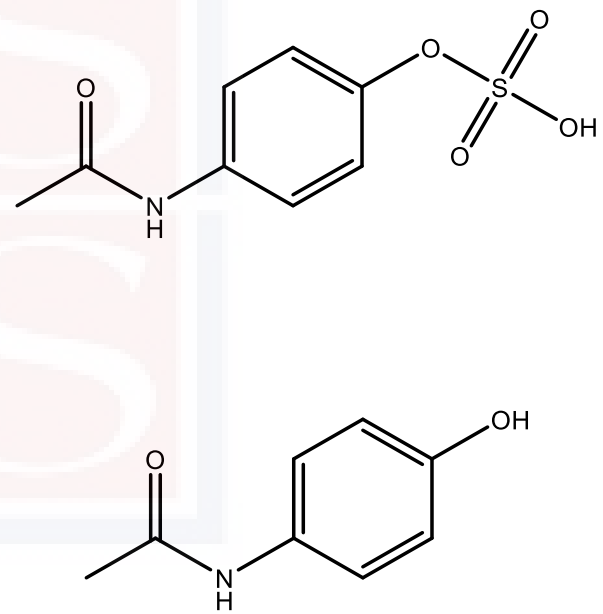
Example 2: Glucuronide loss



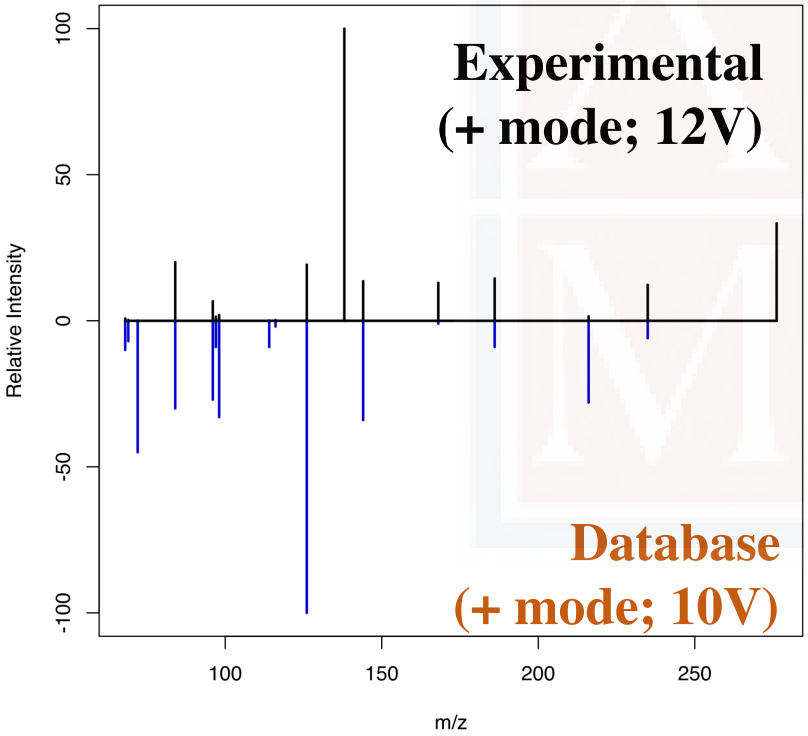


Example 3: Acetaminophen-sulfate

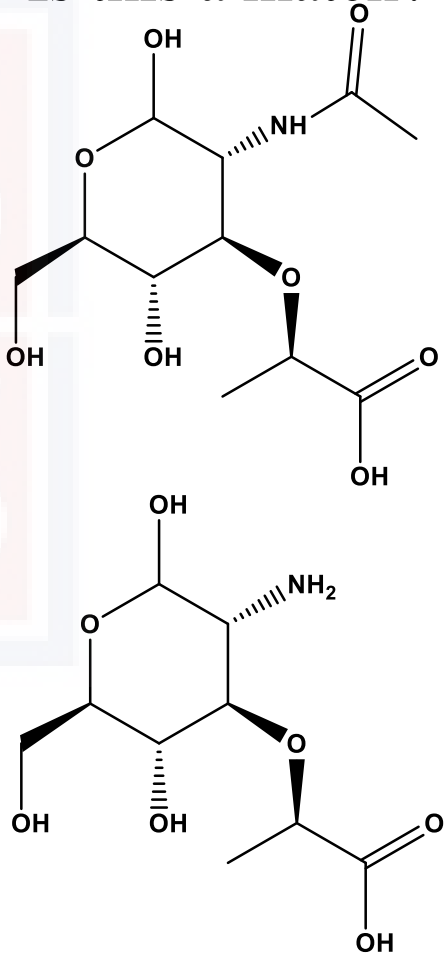




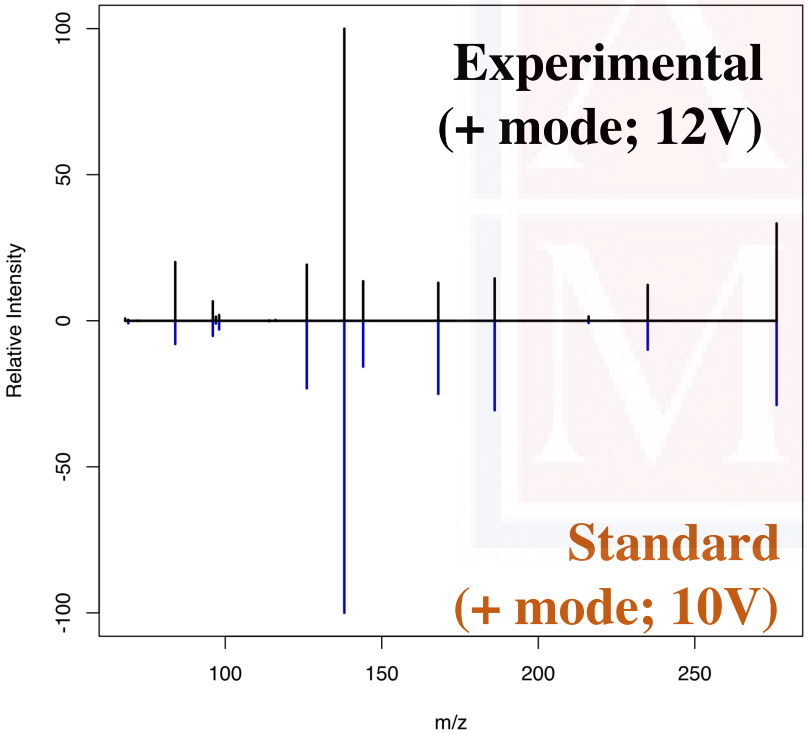
Example 4: N-acetylmuramic acid

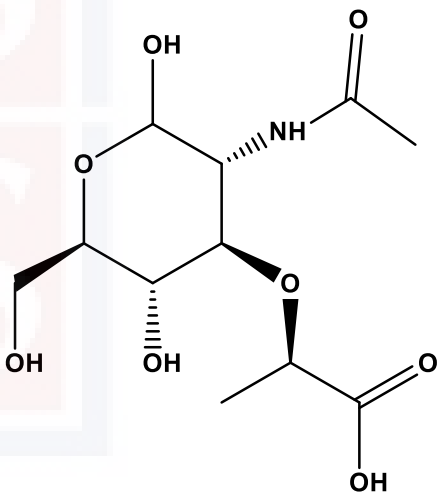


Is this a match?

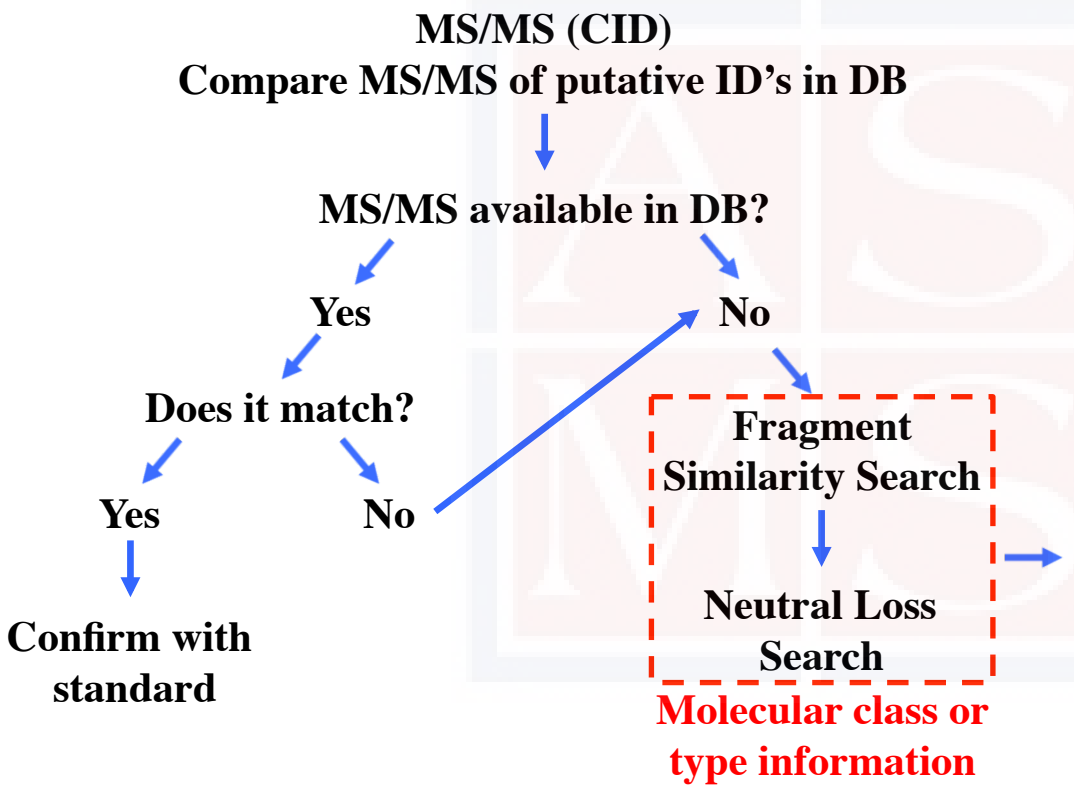


Example 4: N-acetylmuramic acid





Identifying Metabolites: The Big Obstacle Identification with MS/MS



1. Informatic tools: Metfrag, CFM-ID, CSI-Finger ID
2. Literature (peer-reviewed publications and dissertations)
3. Fragmentation rules

Identifying Metabolites: The Big Obstacle
Informatic tools for MS/MS prediction

Metfrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Sebastian Wolf[1*], Stephan Schmidt[1], Matthias Müller-Hannemann[2], Steffen Neumann[1]

CFM-ID

CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra

Felicity Allen[*], Allison Pon, Michael Wilson, Russ Greiner and David Wishart

CSI:FingerID

Searching molecular structure databases with tandem mass spectra using CSI:FingerID

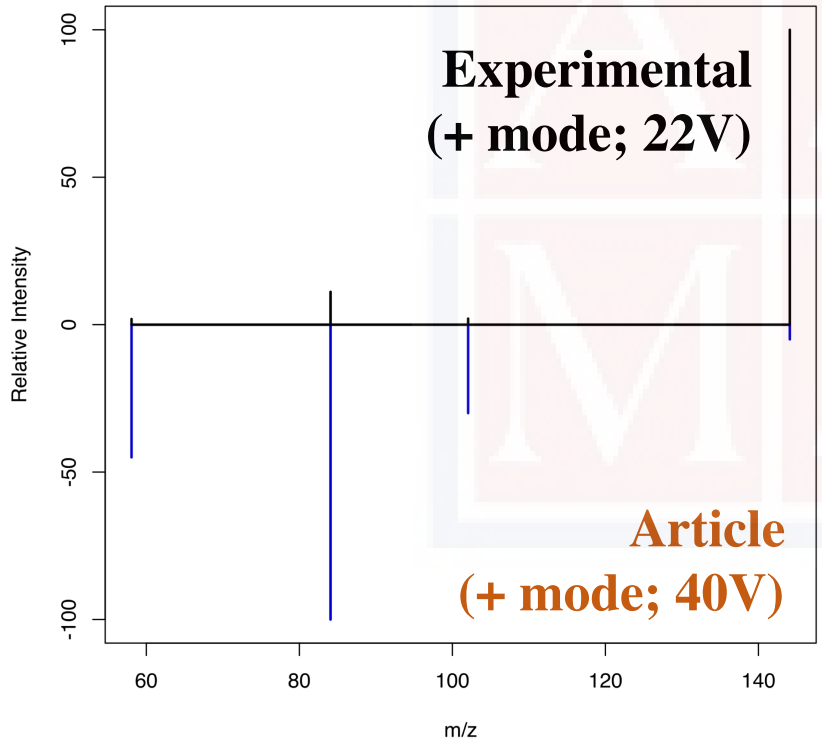
Kai Dührkop[a], Huibin Shen[b], Marvin Meusel[a], Juho Rousu[b], and Sebastian Böcker[a,1]

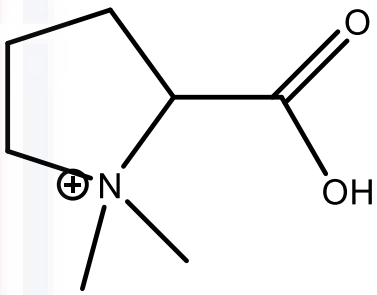
Wolf, S. et. al. *BMC Bioinformatics*, **2010**, 11:148
Allen, F. et. al. *Nucl. Acids Res*. **2014**, 42 (1), 94–99
Dührkop, K. et. al. *PNAS*, **2015** 112 (41), 12580-12585

Identifying Metabolites: The Big Obstacle Literature search for MS/MS prediction

- Large MS/MS spectra available in peer review articles (over 24,000 in Pubmed)
  - MS/MS not always HR but fragments can reduce number of putative identifications

Example 5: Proline betaine



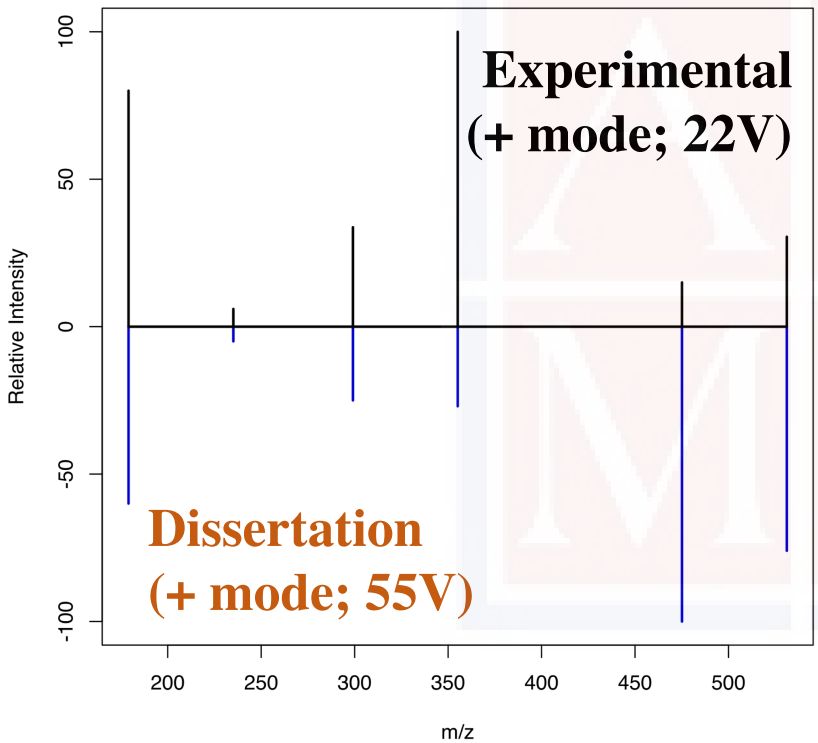


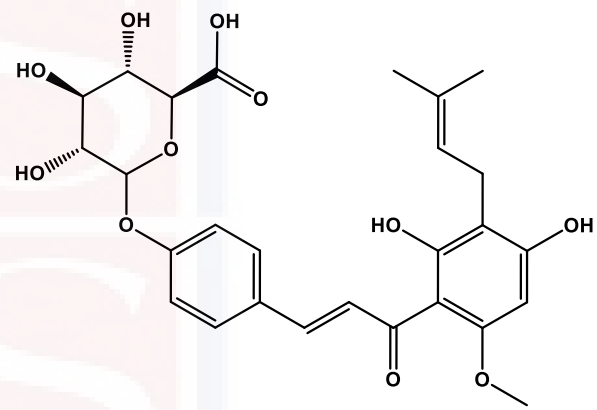
Lloyd, A. J. et. al. *Br. J. Nutr.* **2011**, 106 (6), 812-824
Yang, Q. et. al. *J. Sep. Sci.* **2010**, 33, 1495–1503

Identifying Metabolites: The Big Obstacle Literature search for MS/MS prediction

- Large MS/MS spectra available in peer review articles (over 24,000 in Pubmed)
  - MS/MS not always HR but fragments can reduce number of putative identifications
- Other literature works (e.g. dissertations) can contain useful MS/MS spectra

Example 5: Xanthohumol-glucuronide





Yilmazer, M. (**2001**) *Xanthohumol, a flavonoid from hops: in vitro and in vivo metabolism, antioxidant properties of metabolites and risk assessment in humans* (Doctoral dissertation). Retrieved from https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/fx719q33m

Identifying Metabolites: The Big Obstacle Literature search for MS/MS prediction

- Large MS/MS spectra available in peer review articles (over 24,000 in Pubmed)
  - MS/MS not always HR but fragments can reduce number of putative identifications
- Other literature works (e.g. dissertations) can contain useful MS/MS spectra
- Caution is recommended with these resources as errors can be found

Identifying Metabolites: The Big Obstacle Fragmentation Rules

1. Fragmentation results in charged and neutral species

$$A\text{-}B^{+} \longrightarrow A + B^{+}$$

Identifying Metabolites: The Big Obstacle Fragmentation Rules

1. Fragmentation results in charged and neutral species

2. Product ions mainly depend on number and strength of the bonds

Aliphatic < aromatic < conjugated
C-heteroatom < C-C

Identifying Metabolites: The Big Obstacle Fragmentation Rules

3. Fragments from $+H^+$ or $-H^+$ mainly have even number of e^-

Nitrogen rule:
Fragment with even *m/z* -> odd number of nitrogen atoms
(exception: loss of halogen from aromatic compounds)

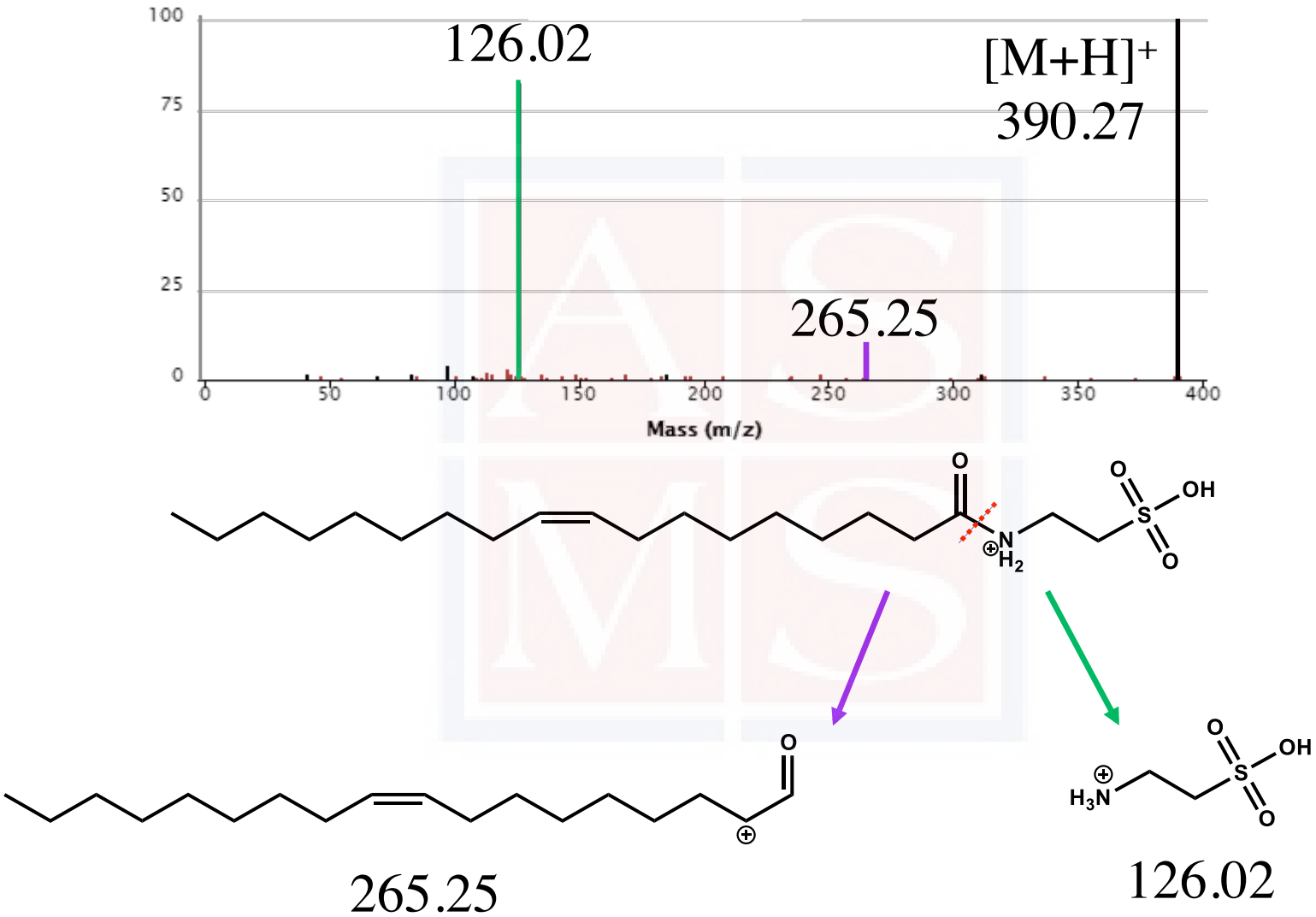
Identifying Metabolites: The Big Obstacle Fragmentation Rules

3. Fragments from $+H^+$ or $-H^+$ mainly have even number of e^-

4. Cleavage of C-(N, O and S) results in charge migration to α C or charge retention in (N, O and S) by H^+ rearrangement (N, O and S)

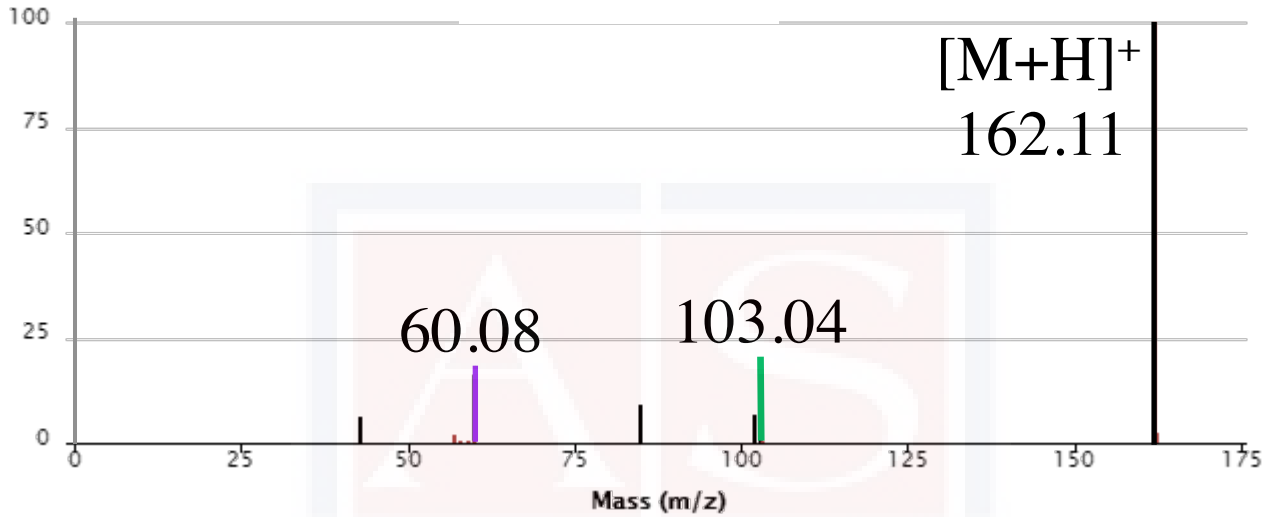
In some cases both fragments can be detected and sum of nominal masses equals nominal value of $[M+H]^+$+1 or $[M-H]^-$-1

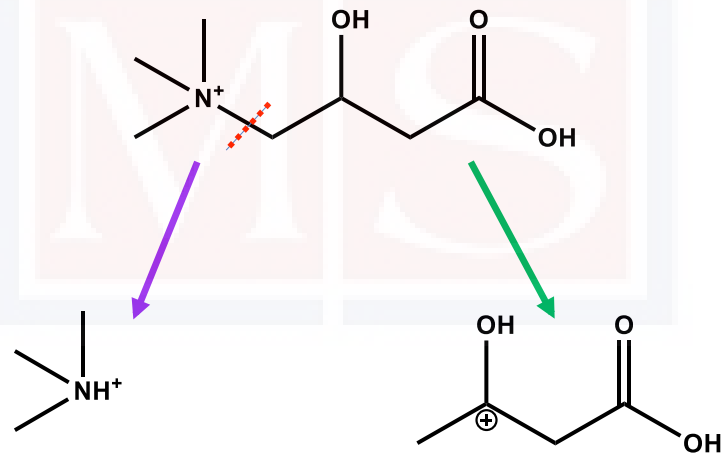
Example 5: N-oleyl taurine



$[M+H]^+ + 1$: 265 + 126 = 390 + 1

Example 6: Carnitine

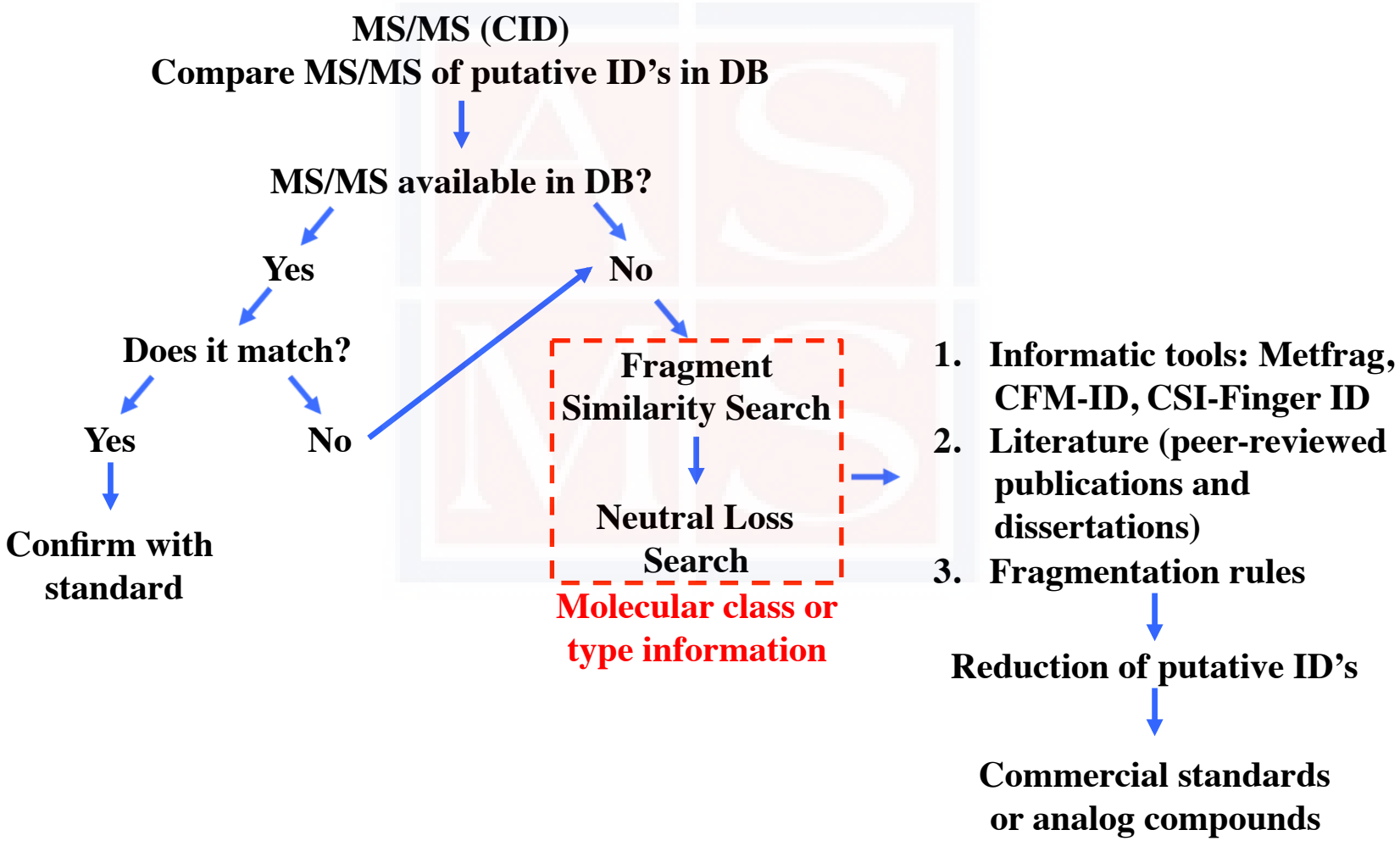




60.08 103.04

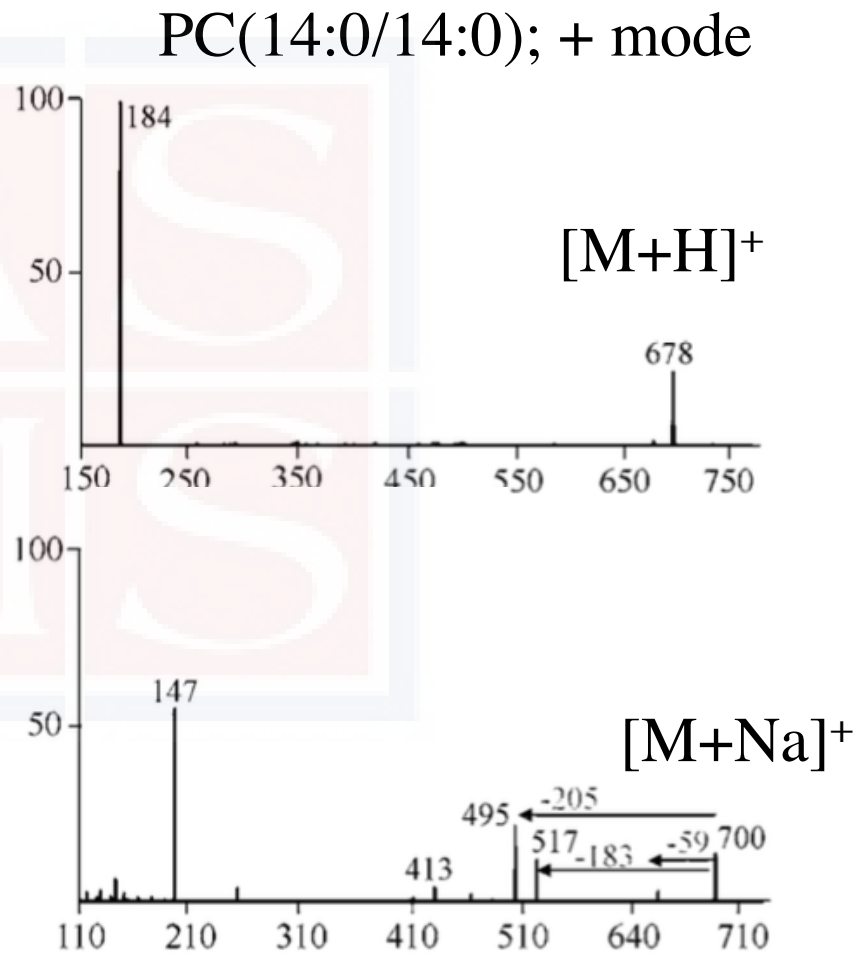
$[M+H]^+ + 1$: 60 + 103 = 162 + 1

Identifying Metabolites: The Big Obstacle Identification with MS/MS



Identifying Metabolites: The Big Obstacle
Other considerations and pitfalls

- H^+ vs. Na^+



Busik, J. V. et. al. *Lipidomics, Methods in Molecular Biology* **2009**, 579

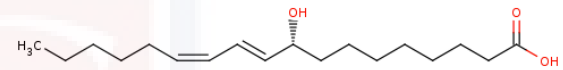
Identifying Metabolites: The Big Obstacle
Other considerations and pitfalls

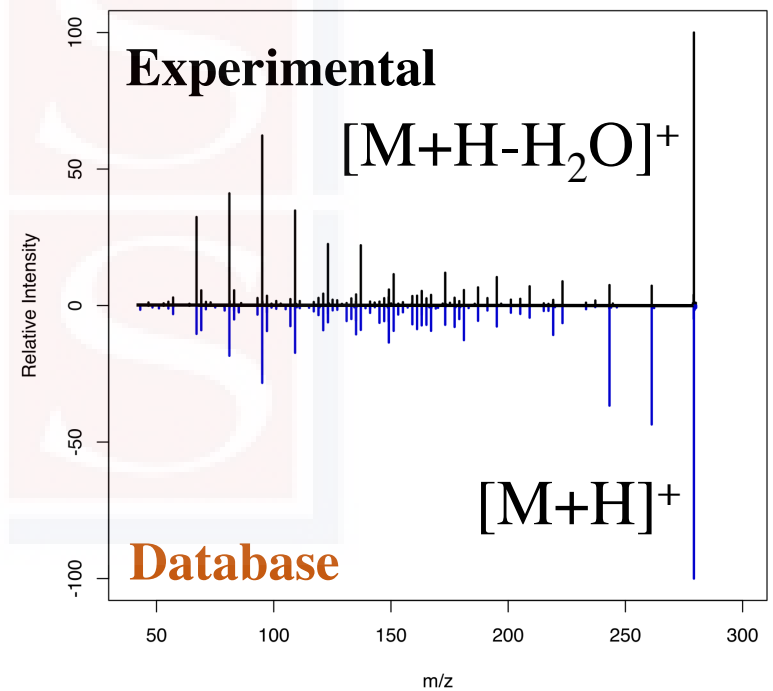
- H^+ vs. Na^+
- Isolation window and noise

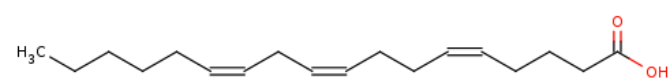
Identifying Metabolites: The Big Obstacle
Other considerations and pitfalls

- H^+ vs. Na^+
- Isolation window and noise
- In source fragments

9-HODE





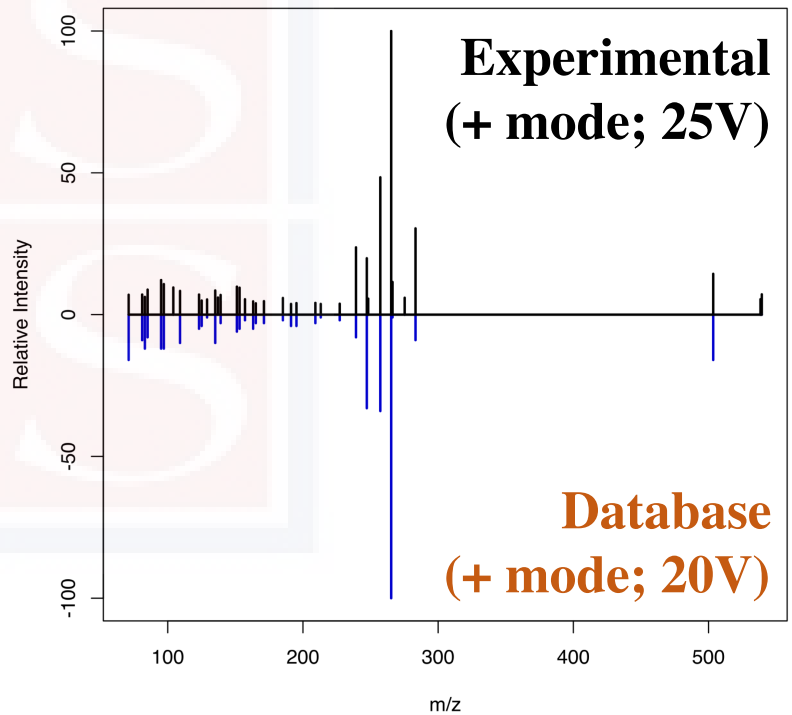


Pinolenic acid

Identifying Metabolites: The Big Obstacle
Other considerations and pitfalls

- H^+ vs. Na^+
- Isolation window and noise
- In source fragments
- Isobars, similar structures and multimers

9-PAHSA and stearic acid dimer



Identifying Metabolites: The Big Obstacle Conclusions

- Confirm ID with standards when possible
- Indicate level of identification
- Resources for MS/MS spectra interpretation can reduce number of putative ID's

Advanced Metabolomics

June 3rd 2018

Gary Siuzdak

H. Paul Benton

Xavi Domingo

Erica Forsberg

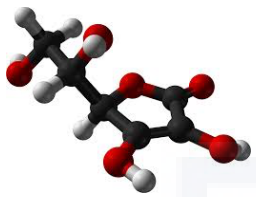
Carlos Guijas

Rafa Montenegro

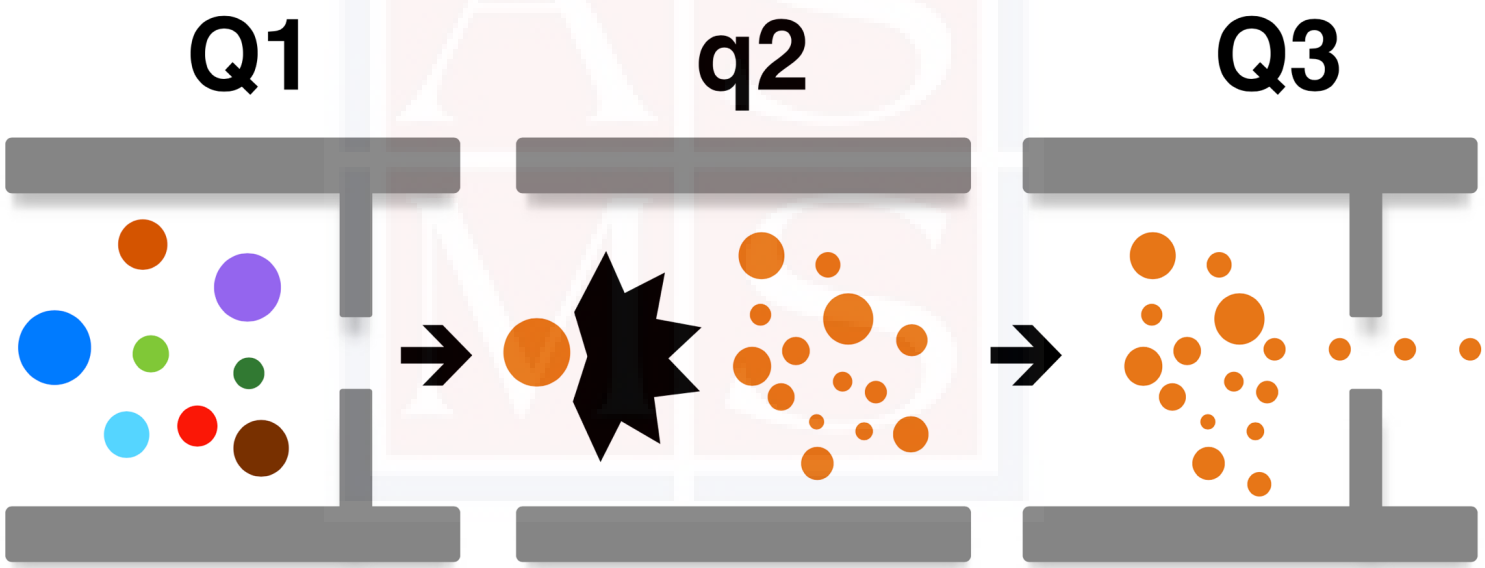
Advanced MRM

Xavi Domingo

Multiple Reaction Monitoring (MRM)



MRM transitions
(precursor -> product)

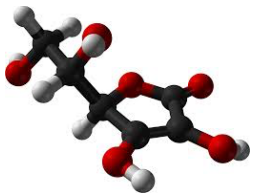


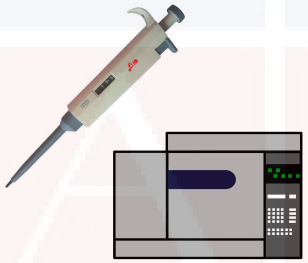
Specific precursor ions

CID: fragmentation

Specific fragments

Transition optimization via pure materials

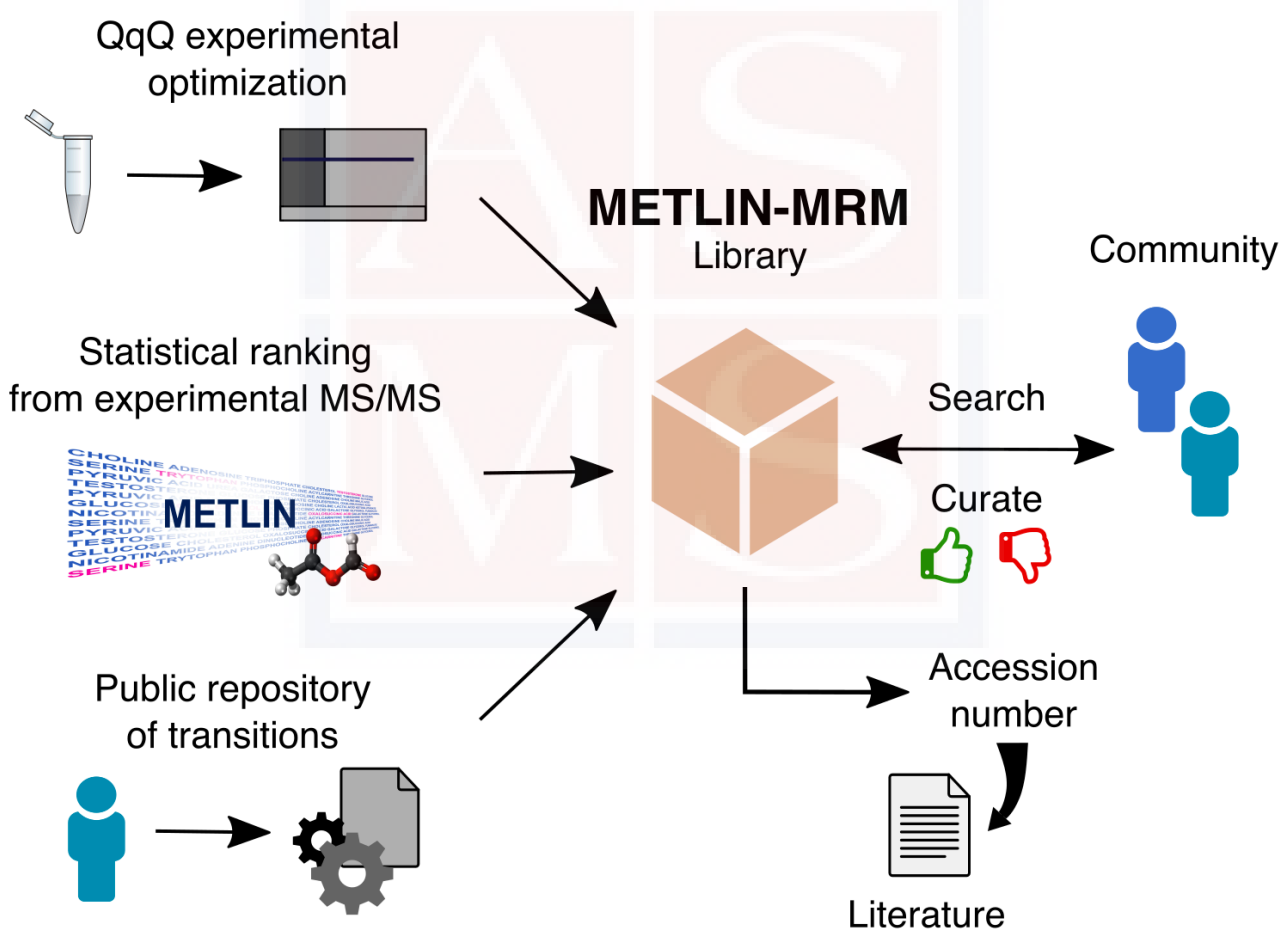




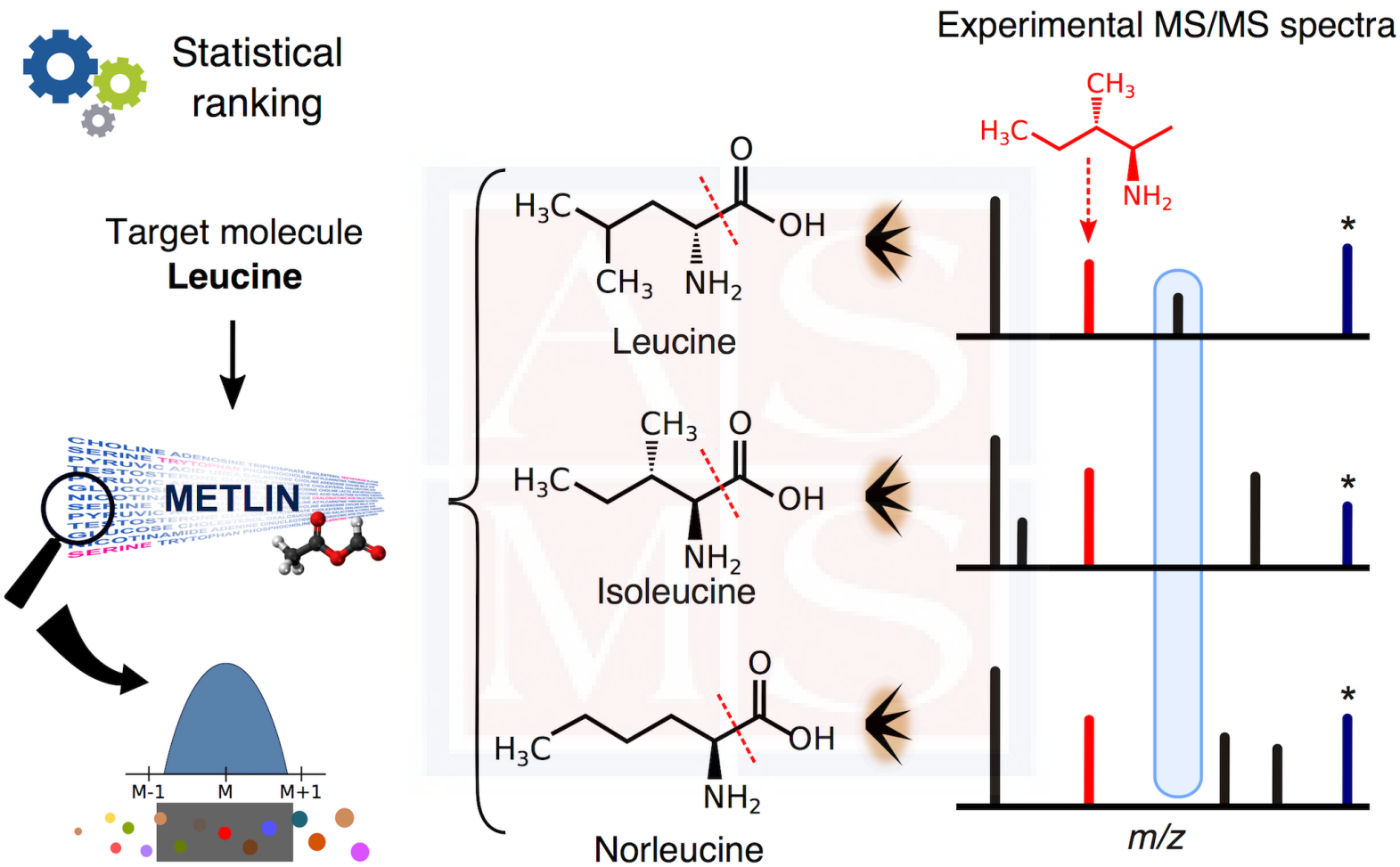
MRM transitions
(precursor -> product)

- Auto-calibration via vendor software
- MRM transitions selected based on chromatographic properties (S/N), ionization efficiency.

METLIN-MRM

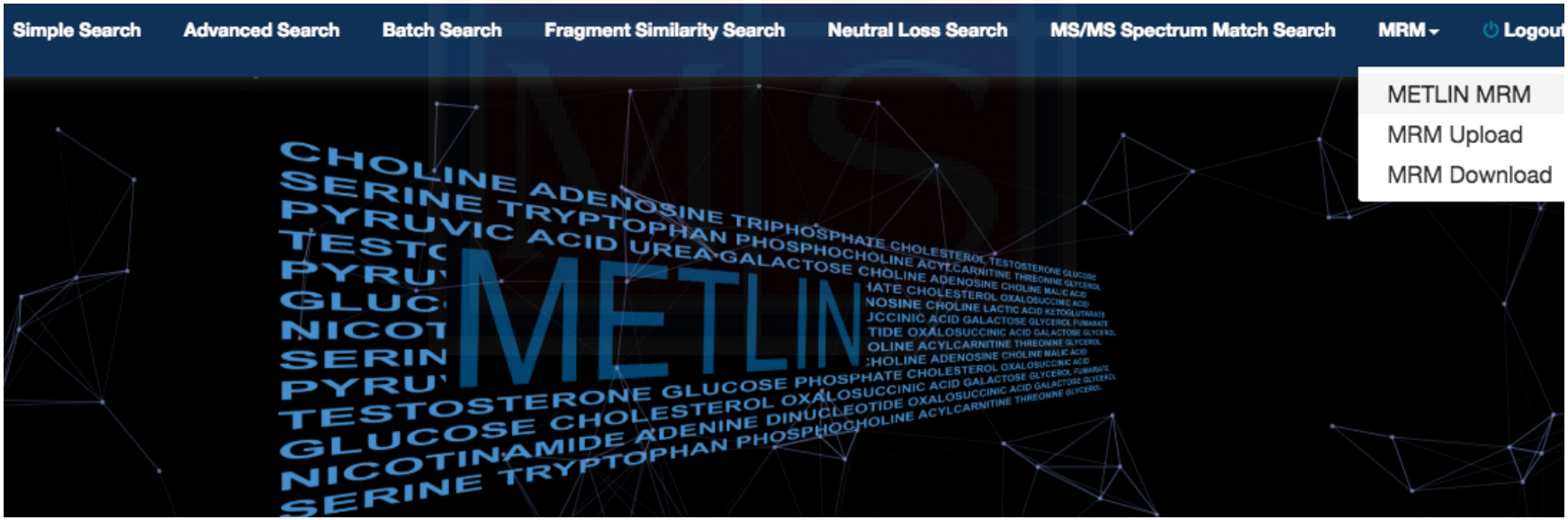

QqQ experimental optimization
METLIN-MRM
Library
Community
Statistical ranking from experimental MS/MS
METLIN
Search
Curate
Public repository of transitions
Accession number
Literature

Computational optimization



METLIN-MRM

1) Go to http://metlin.scripps.edu
2) Click on MRM/METLIN-MRM



Statisticaly Optimized Experimental Transitions

View Selected Fragment(s)

Name: **L-Tryptophan**, MID: **33**

Show 10 entries

Search:

Precursor	Adduct	Mode	Col. E.	MZ	Rating
203.1	M-H	-	20	116.1	(0) (0)
203.1	M-H	-	40	142.1	(0) (0)
203.1	M-H	-	10	159.1	(0) (0)
205.1	M+H	+	10	188.1	(0) (0)
205.1	M+H	+	10	146.1	(0) (0)
Precursor	**Adduct**	**Mode**	**Col. E.**	**MZ**	**Rating**

Showing 1 to 5 of 5 entries

Previous 1 Next

Agilent

View Selected Fragment(s)

Name: **L-Tryptophan**, MID: **33**

Show 10 entries

Search:

Precursor	Adduct	Mode	Col. E.	MZ	Rating
203.08208	M-H	-	20	116	(0) (0)
205.09768	M+H	+	8	188	(0) (0)
205.09768	M+H	+	20	146	(0) (0)
Precursor	**Adduct**	**Mode**	**Col. E.**	**MZ**	**Rating**

Showing 1 to 3 of 3 entries

Previous 1 Next

Sciex

View Selected Fragment(s)

Name: **L-Tryptophan**, MID: **33**

Show 10 entries

Search:

Precursor	Adduct	Mode	Col. E.	MZ	Rating
203.1	M-H	-	20	116	(0) (0)
205.09768	M+H	+	20	146	(0) (0)
205.09768	M+H	+	15	188	(0) (0)
Precursor	**Adduct**	**Mode**	**Col. E.**	**MZ**	**Rating**

Showing 1 to 3 of 3 entries

Previous 1 Next

User Supplied MRM Data

Name: **L-Tryptophan**, MID: **33**

Show 10 entries

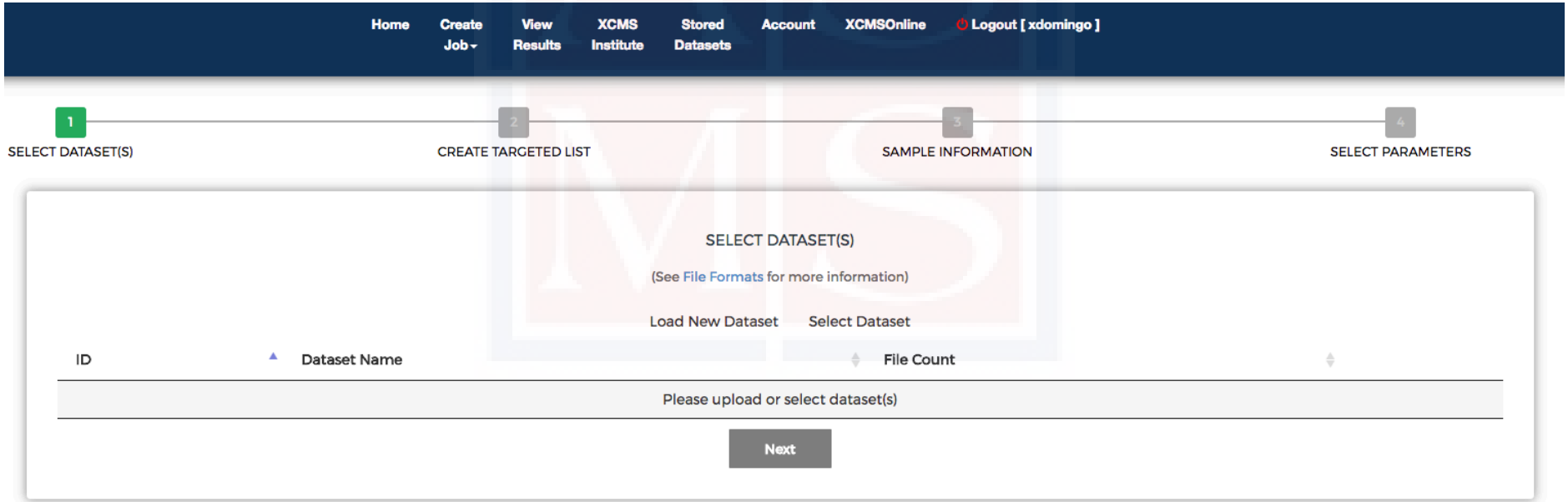
Search:

MRM ID	Name	Adduct	Precursor	Product	Col. E.	Mode	DOI	HMDB	Formula	PubChem
2	L-Tryptophan	M+H	205	146	21	+	10.1007/s11306-017-1264-1	HMDB00929	C11H12N2O2	
2	L-Tryptophan-13C	M+H	216	155	21	+	10.1007/s11306-017-1264-1	HMDB00929	*C11H12N2O2	
2	5-Hydroxy-L-tryptophan	M-H	219	144	22	-	10.1007/s11306-017-1264-1	HMDB00472	C11H12N2O3	
8	5-Hydroxy-L-tryptophan (5-HTP)	[M-H]-	219	157	24	-	10.1016/j.aca.2015.08.056	HMDB00472	C11H12N2O3	
9	5-hydroxy-L-tryptophan	[M+H]+	221.2	162.1	19	+	10.1177/1535370217694098			
9	5-hydroxy-L-tryptophan	[M+H]+	221.2	204	11	+	10.1177/1535370217694098			
MRM ID	**Name**	**Adduct**	**Precursor**	**Product**	**Col. E.**	**Mode**	**DOI**	**HMDB**	**Formula**	**PubChem**

Showing 1 to 6 of 6 entries

Previous 1 Next

1) Go to http://xcmsonline-mrm.scripps.edu
2) Click on Create Job/XCMS-MRM



XCMS-MRM

1 SELECT DATASET(S) — 2 CREATE TARGETED LIST — 3 SAMPLE INFORMATION — 4 SELECT PARAMETERS

TARGETED LIST

Choose File Target_list.csv

(List Example)

Show 10 entries

Name	Precursor	Product	RT.min	RT.max	Prec.Labeled	Prod.Labeled
Leucine	132.1	43.096				
Leucine	132.1	44.096				
Isoleucine	132.1	44.096				
Isoleucine	132.1	69.066				
Leucine	132.1	86.086				
Isoleucine	132.1	86.086				
Phenylalanine	166.08	103.096				
Phenylalanine	166.08	120.076				
Phenylalanine	166.08	130.996				
Tyrosine	182.08	136.096				

Previous 1 2 3 Next

Previous Next

XCMS-MRM

1 SELECT DATASET(S) — 2 CREATE TARGETED LIST — 3 SAMPLE INFORMATION — 4 SELECT PARAMETERS

SAMPLE INFORMATION

Auto-generate Targeted List

Previous Next

SAMPLE INFORMATION

Auto-generate Targeted List

	Dataset ID	File ID	File Name	Sample Type	Sample Grou	Leucine	Isoleucine	Phenylalanir	Tyrosine	Caffeine
1	214178	1518191	Plasma_0.mzML	Sample-Calib		0				
2	214178	1518192	Plasma_50.mzML	Sample-Calib		50				
3	214178	1518193	Plasma_1a.mzML	Sample-Calib		1				
4	214178	1518194	Plasma_5.mzML	Sample-Calib		5				
5	214178	1518195	Plasma_10.mzML	Sample-Calib		10				
6	214178	1518196	Plasma_100.mzML	Sample-Calib		100				

Previous Next

XCMS-MRM

SAMPLE INF

Auto-generate

	Dataset ID	File ID	File Name	Sample Type	Sample Grou	Leucine	Isoleuc
1	214178	1518191	Plasma_0.mzML	✓ Sample			
2	214178	1518192	Plasma_50.mzML	Blank			
3	214178	1518193	Plasma_1a.mzML	Calibration			
4	214178	1518194	Plasma_5.mzML	Sample-Calibration			
5	214178	1518195	Plasma_10.mzML	Sample			
6	214178	1518196	Plasma_100.mzML	Sample			

QC

Previous

XCMS-MRM

1 SELECT DATASET(S) — 2 CREATE TARGETED LIST — 3 SAMPLE INFORMATION — 4 SELECT PARAMETERS

SELECT PARAMETERS

Average peak width (seconds)

10

Scans per second (virtual)

5

Detection limit

3

Quantification limit

10

Calibration Curve

linear

Previous Submit

XCMS-MRM

Home | Create Job | View Results | XCMS Institute | Stored Datasets | Account | XCMSOnline | Logout [xdomingo]

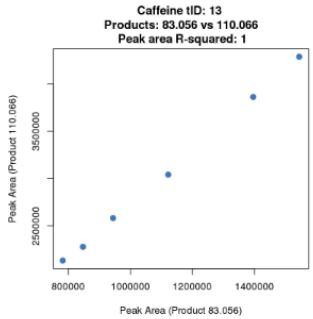
Submit Date	Finish Date	Parameter ID#	Log	Shared	Download Results
2017-12-18 15:08:07	2017-12-18 15:11:56	name (24427)	View Log	NOT SHARED	hash: dfd21f0cfb26d06028d8139defef30fc

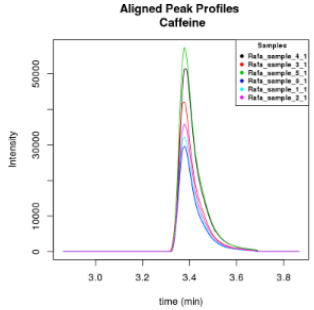
Show 25 rows | TSV | Print | Image Viewer | Profile Adjustment

Search:

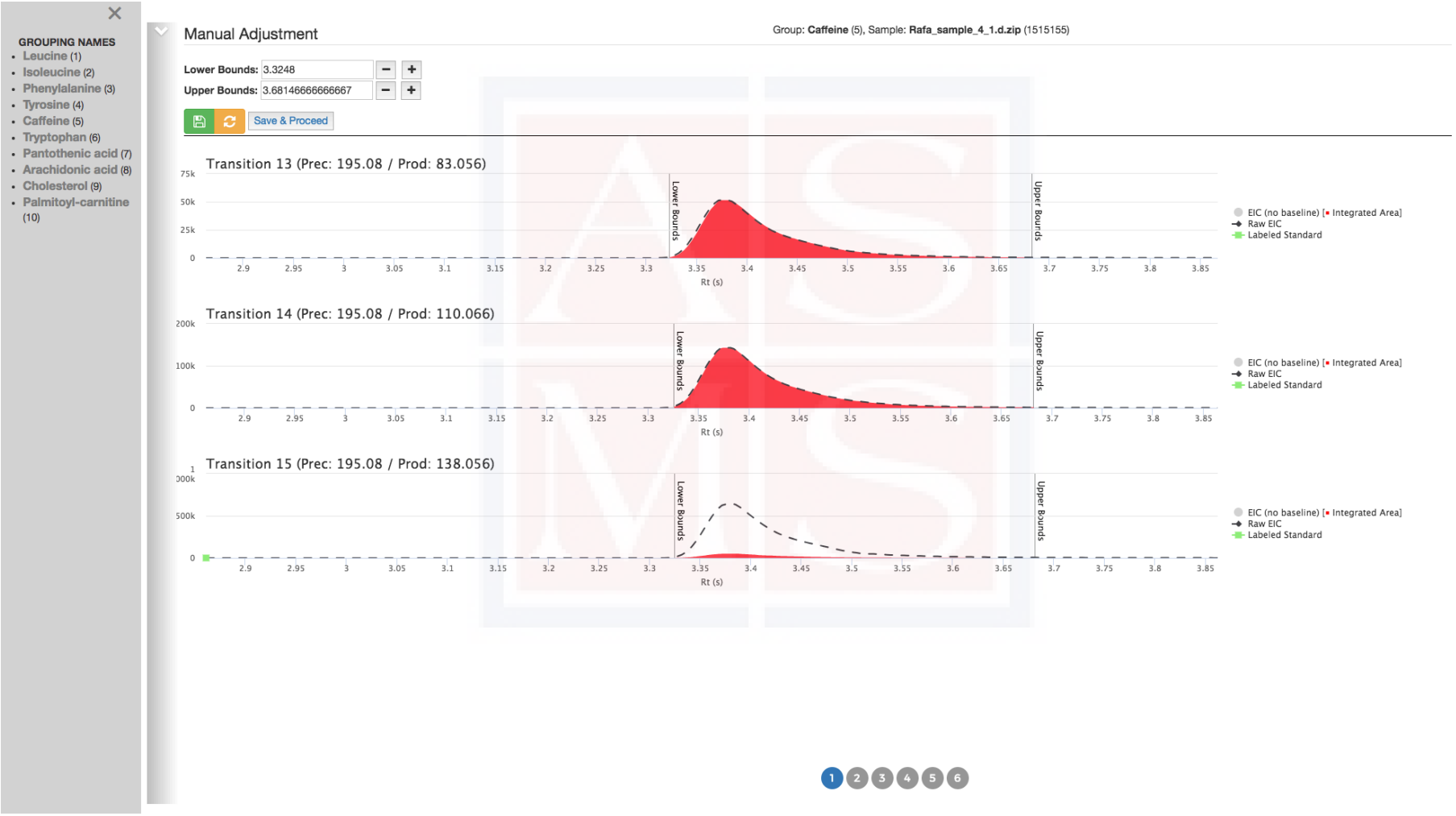
tID	Name	Precursor	Product	RT(min)	RT(max)	RT(mean)	LOD	LOQ	R2	p-value	FC	CV
1	Leucine	132.1	43.096	0.709	1.7	1.48			1.00			
2	Leucine	132.1	44.096	0.707	1.7	1.29			1.00			
3	Leucine	132.1	86.086	0.704	1.7	1.29			1.00			
4	Isoleucine	132.1	44.096	0.708	1.705	1.29			0.98			
5	Isoleucine	132.1	69.066	0.706	1.703	1.29			0.99			
6	Isoleucine	132.1	86.086	0.705	1.701	1.29			1.00			
7	Phenylalanine	166.08	103.096	2.012	3.004	2.55			1.00			
8	Phenylalanine	166.08	120.076	2.01	3.003	2.55			0.99			
9	Phenylalanine	166.08	130.996	2.007	3.003	2.55			0.99			
10	Tyrosine	182.08	136.096	0.607	1.599	1.22			1.00			
11	Tyrosine	182.08	146.996	0.604	1.598	1.23			1.00			
12	Tyrosine	182.08	165.096	0.601	1.597	1.23			1.00			
13	Caffeine	195.08	83.056	2.863	3.867	3.38			1.00			
14	Caffeine	195.08	110.066	2.862	3.864	3.38			1.00			
15	Caffeine	195.08	138.056	2.861	3.862	3.38			1.00			
16	Tryptophan	205.09	146.046	2.512	3.501	3.07			1.00			
17	Tryptophan	205.09	188.066	2.511	3.499	3.07			1.00			
18	Pantothenic acid	220.11	69.996	2.204	3.196	2.78			1.00			
19	Pantothenic acid	220.11	90.096	2.202	3.195	2.78			1.00			
20	Pantothenic acid	220.11	184.096	2.201	3.194	2.78			1.00			
21	Arachidonic acid	305.25	67.096	8.911	10.903	10.14			1.00			

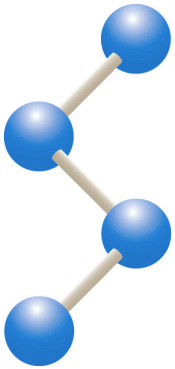
Caffeine (13)
Precursor: 195.08





XCMS-MRM


GROUPING NAMES
Leucine (1)
Isoleucine (2)
Phenylalanine (3)
Tyrosine (4)
Caffeine (5)
Tryptophan (6)
Pantothenic acid (7)
Arachidonic acid (8)
Cholesterol (9)
Palmitoyl-carnitine (10)
Manual Adjustment
Group: Caffeine (5), Sample: Rafa_sample_4_1.d.zip (1515155)
Lower Bounds: 3.3248
Upper Bounds: 3.68146666666667
Save & Proceed
Transition 13 (Prec: 195.08 / Prod: 83.056)
Transition 14 (Prec: 195.08 / Prod: 110.066)
Transition 15 (Prec: 195.08 / Prod: 138.056)
Lower Bounds
Upper Bounds
Rt (s)
EIC (no baseline) [Integrated Area]
Raw EIC
Labeled Standard
1 2 3 4 5 6



Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

June 3rd

---- 09:00 am Begin ----

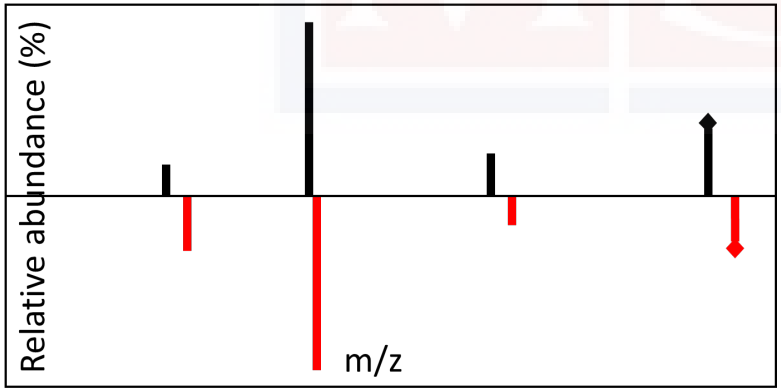
---- 10:15 am Break ----

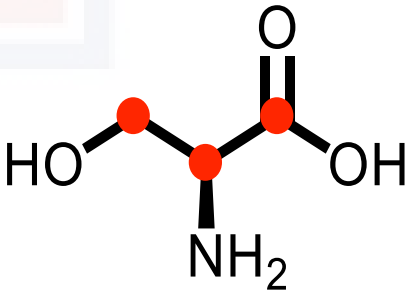
---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

IDENTIFYING METABOLITES USING ISOTOPES





CARLOS GUIJAS

TABLE OF CONTENTS

- ENDOGENOUS ISOTOPIC DISTRIBUTION OF A FEATURE.
- GENERATION OF ISOTOPIC DATA USING UNIFORMLY-LABELED MICROORGANISMS.
- IDENTIFICATION OF UNKNOWNS USING ISOTOPES.

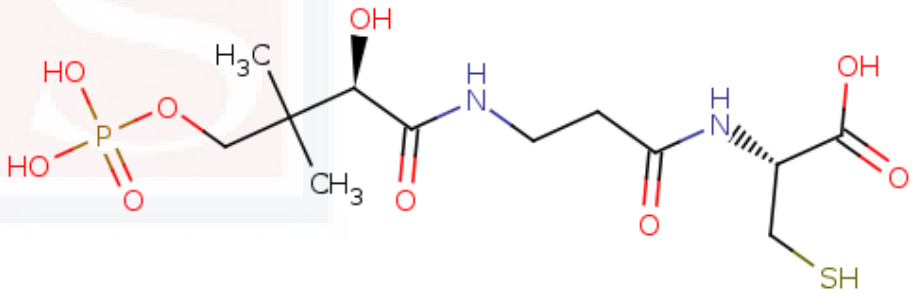
- **Isotopes**: Atoms of the same element with different mass due to the presence of neutrons in the nucleus.

Isotope	Mass (a.m.u.)	Abundance (%)
^{1}H	1.0078	99.985
^{2}H	2.0141	0.015
^{12}C	12.0000	98.89
^{13}C	13.0034	1.11
^{14}N	14.0031	99.64
^{15}N	15.0001	0.36
^{16}O	15.9949	99.76
^{17}O	16.9991	0.04
^{18}O	17.9992	0.20
^{31}P	30.9738	100
^{32}S	31.9721	94.93
^{33}S	32.9715	0.76
^{34}S	33.9679	4.29
^{36}S	35.9671	0.02

* Only stable isotopes

Biomolecules:

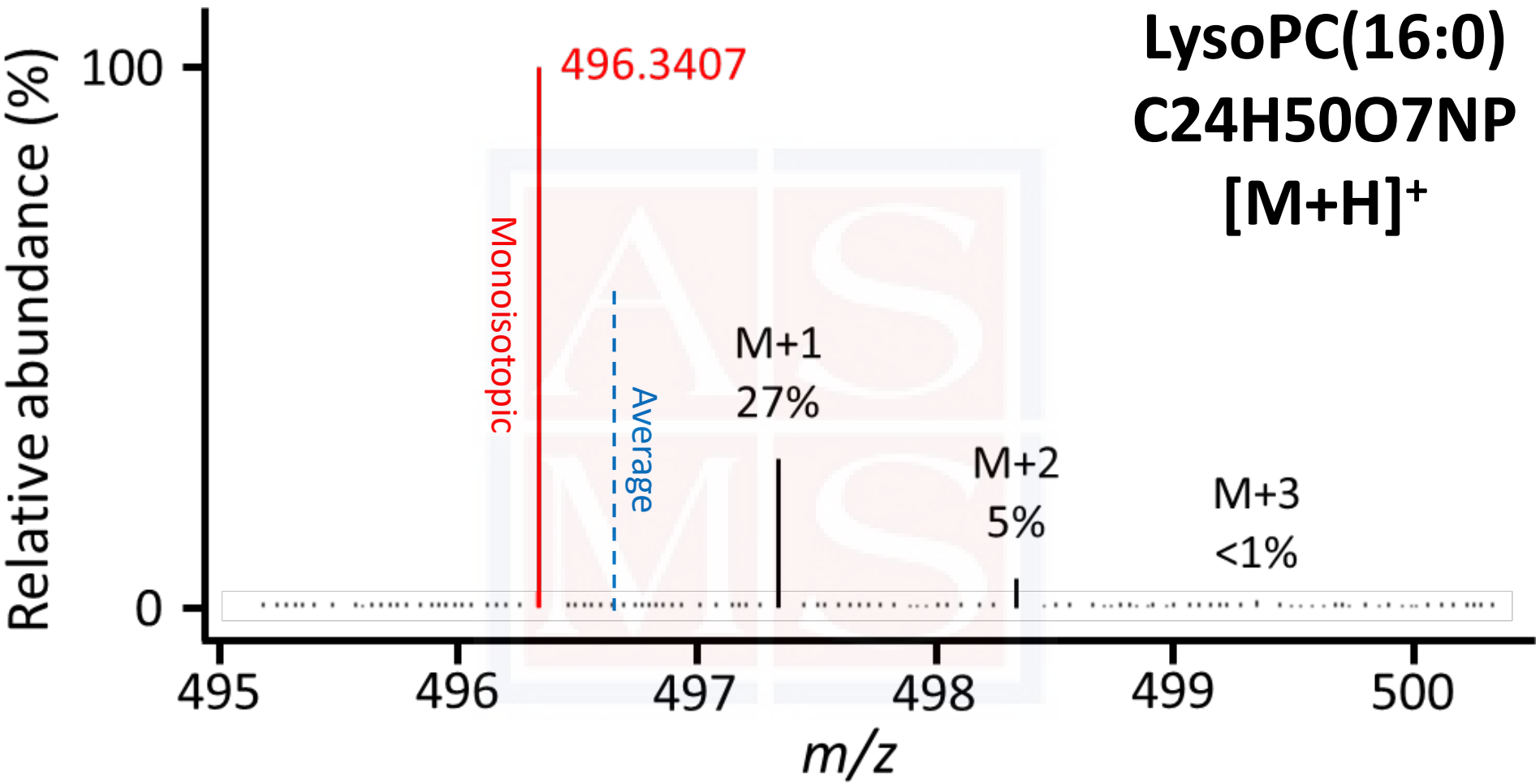
C-H-N-O-P-S



4-Phosphopantothenoylcysteine

ISOTOPIC DISTRIBUTION

Information extracted from the isotopic distribution of a feature

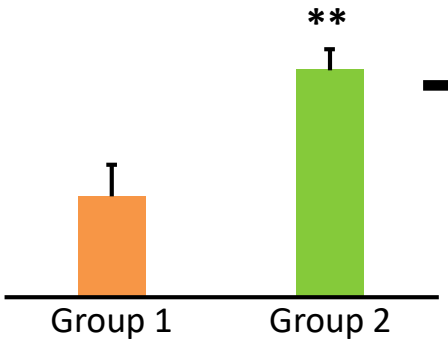


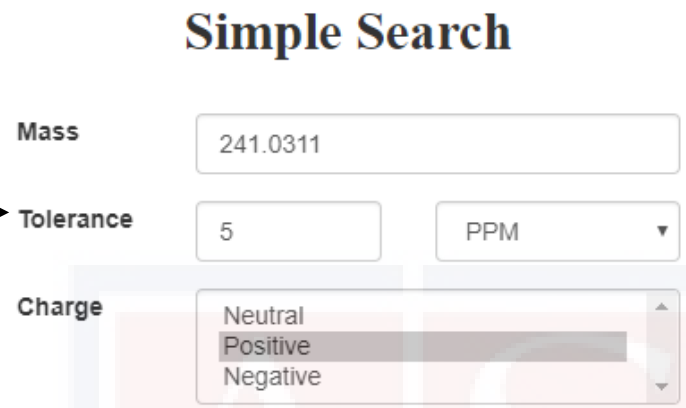
Data acquired with a Bruker Impact II Q-ToF (Resolution~ 34.000)

ISOTOPIC DISTRIBUTION

Example I. Isotopic distribution

Unknown feature
m/z=241.0311

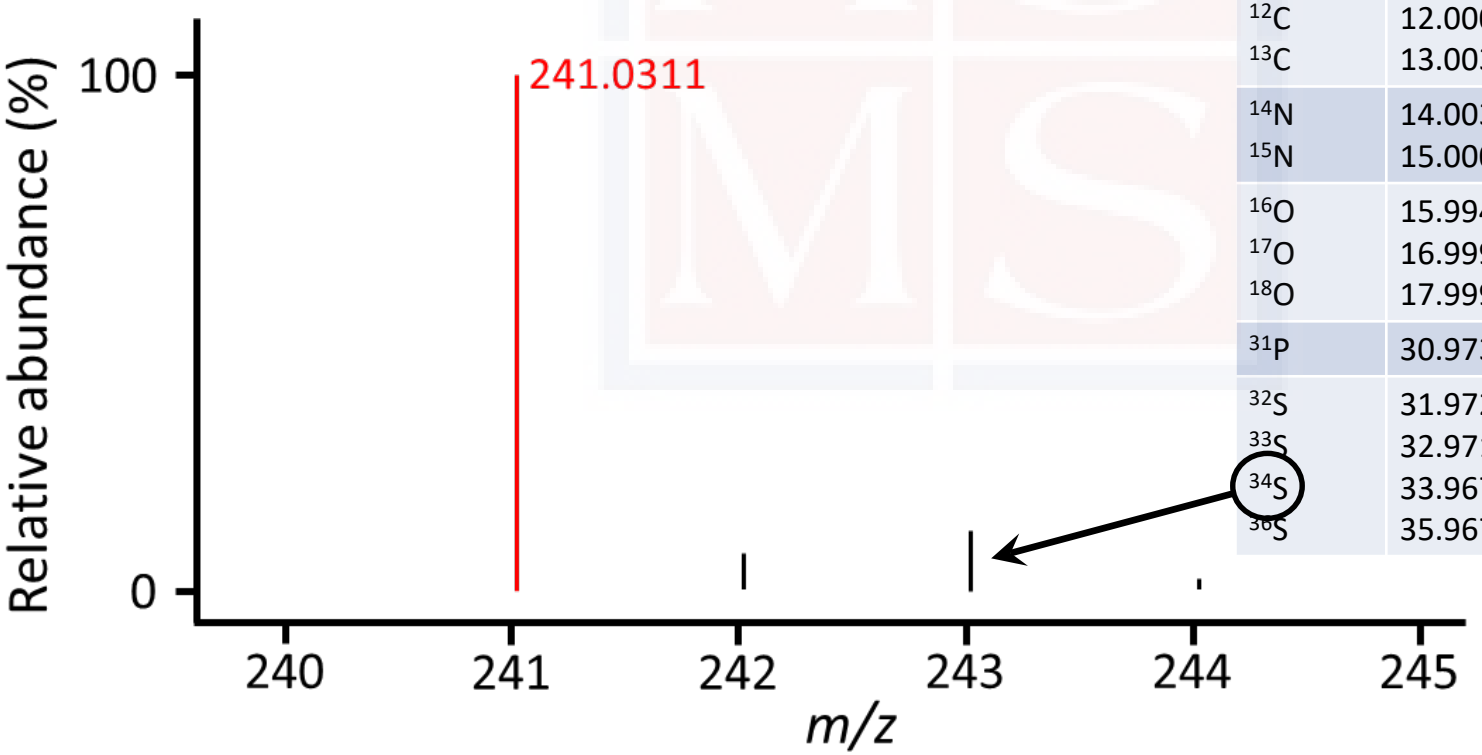




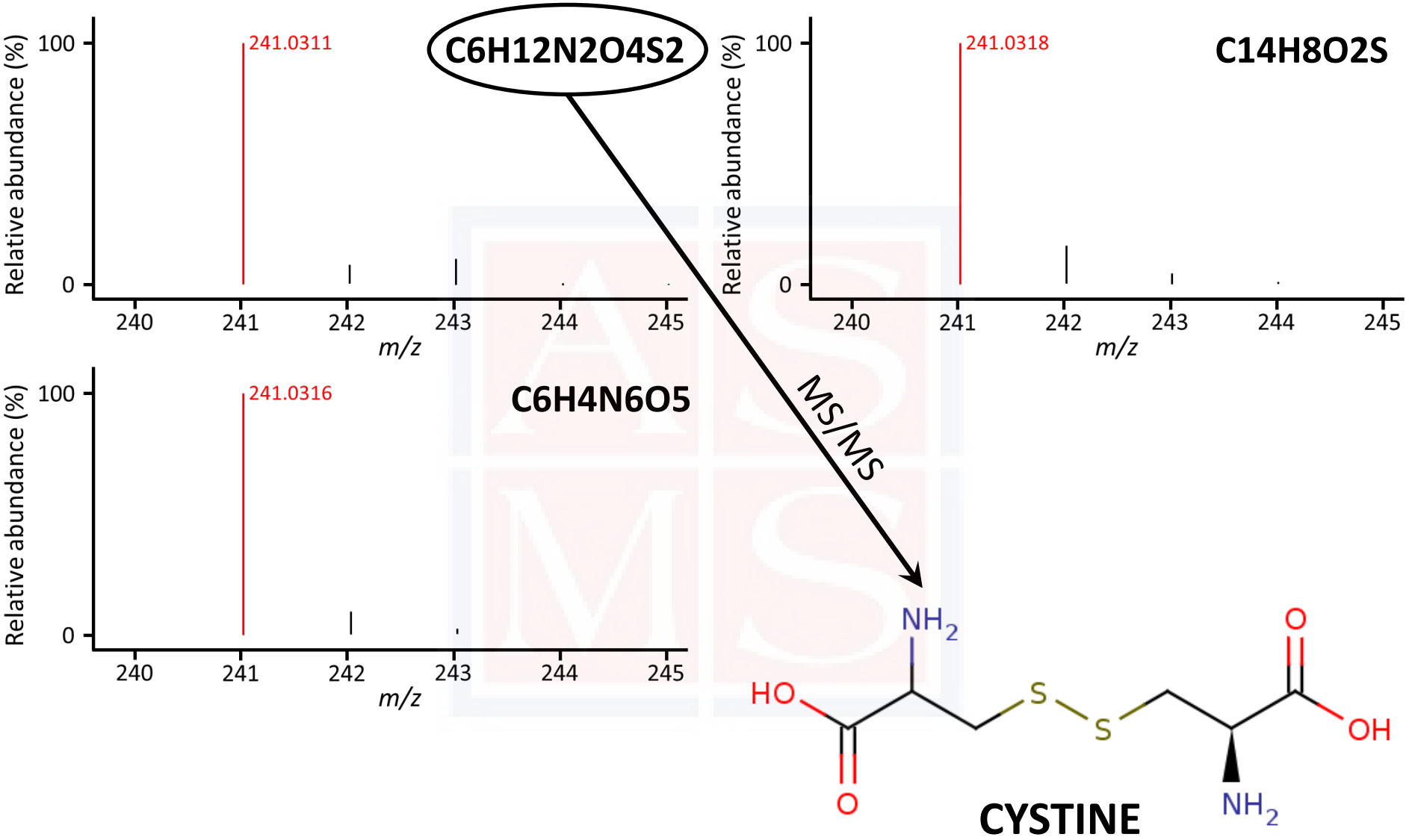
List of putative formulas:

- C6H12N2O4S2
- C14H8O2S
- C6H4N6O5

Isotope	Mass (a.m.u.)	Abundance (%)
^{1}H	1.0078	99.985
^{2}H	2.0141	0.015
^{12}C	12.0000	98.89
^{13}C	13.0034	1.11
^{14}N	14.0031	99.64
^{15}N	15.0001	0.36
^{16}O	15.9949	99.76
^{17}O	16.9991	0.04
^{18}O	17.9992	0.20
^{31}P	30.9738	100
^{32}S	31.9721	94.93
^{33}S	32.9715	0.76
^{34}S	33.9679	4.29
^{36}S	35.9671	0.02



Example I. Isotopic distribution



Simulation of isotopic distributions: https://www.envipat.eawag.ch/

Mass defect of a biomolecule

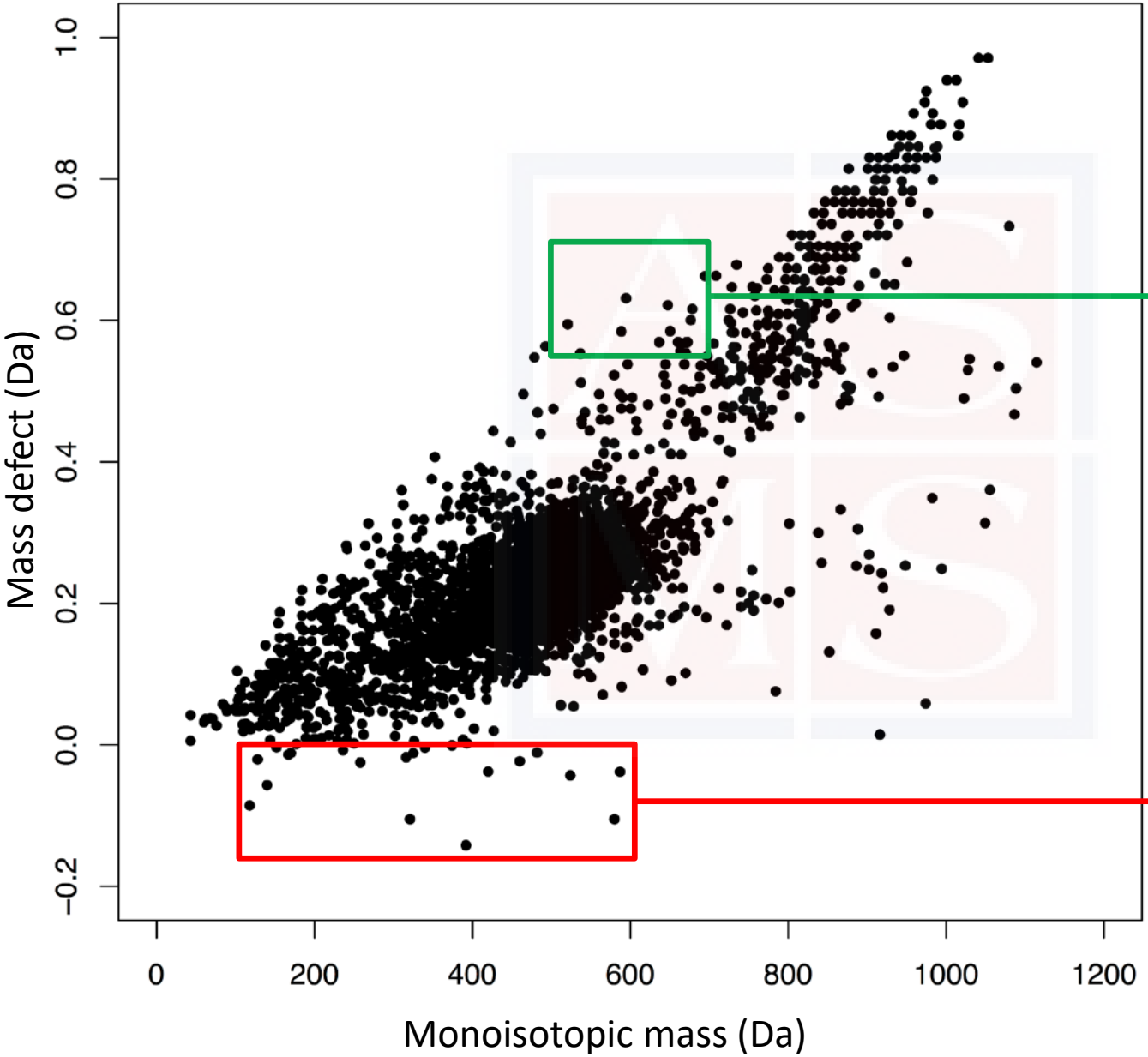
- **Mass defect**: Difference between an element exact mass and its nominal mass.

Isotope	Exact mass (Da)	Nominal mass (Da)	Mass defect (Da)
^{1}H	1.0078	1	0.0078
^{12}C	12.0000	12	0.0000
^{14}N	14.0031	14	0.0031
^{16}O	15.9949	16	-0.0051
^{31}P	30.9738	31	-0.0262
^{32}S	31.9721	32	-0.0279

* Only most abundant isotopes

Discrimination based on mass defect

Search of 5000 random molecules in METLIN

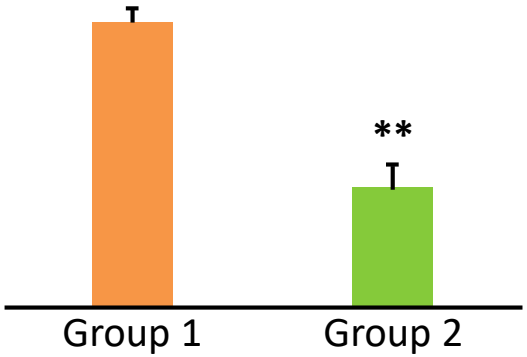


- CE (20:5) (**C47H74**O2)
- DAG(40:5) (**C43H74**O5)
- Cer(d18:1/24:1) (**C42H81**NO3)
- Heptatriacontane (**C37H76**)

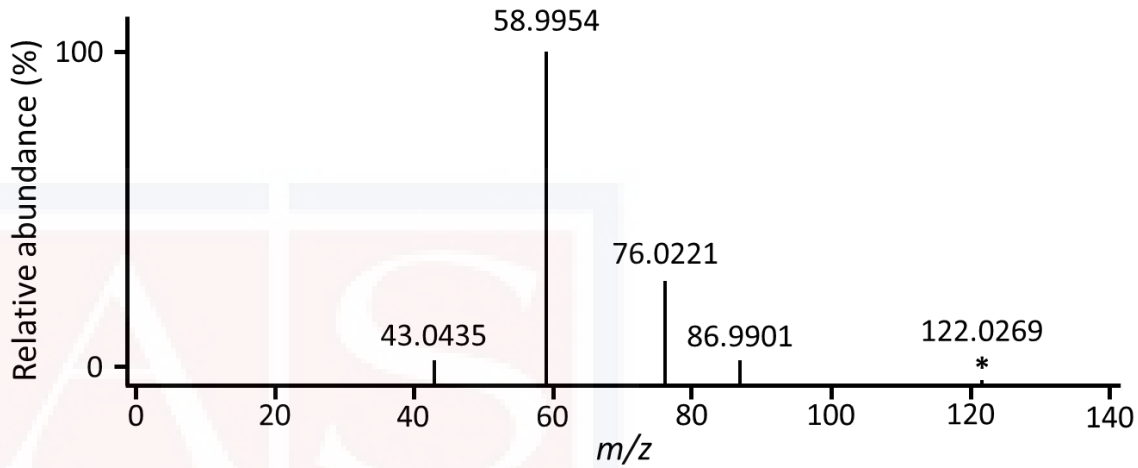
- Inositol trisphosphate (C6H15**O15P3**)
- Glucose bisphosphate (C6H14**O12P2**)
- Sulfolactate (C3H6**O6S**)
- Triazolidinonethione (C12H12N3**OSBr**)

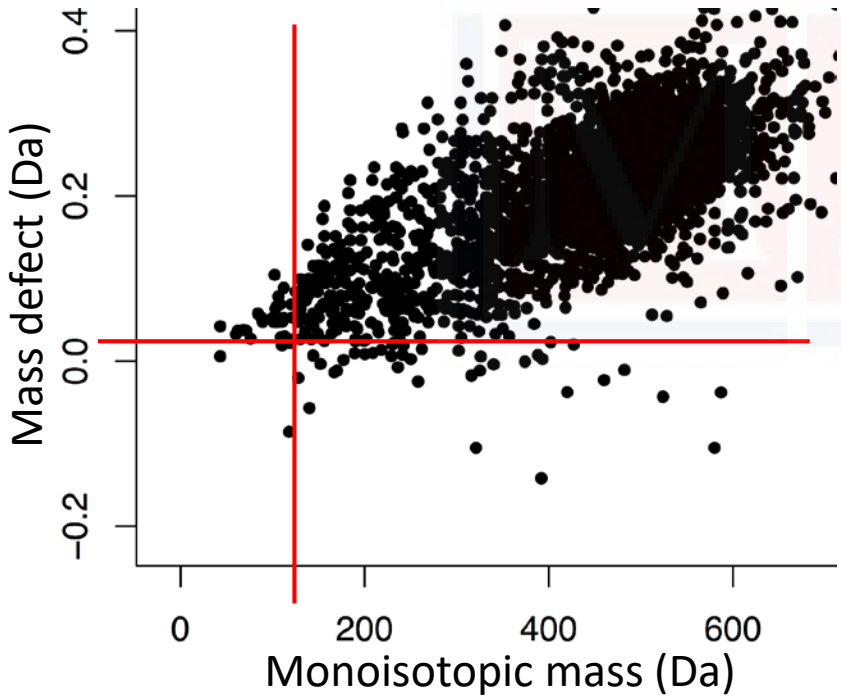
Example II. Mass defect

Unknown feature
m/z=122.0270

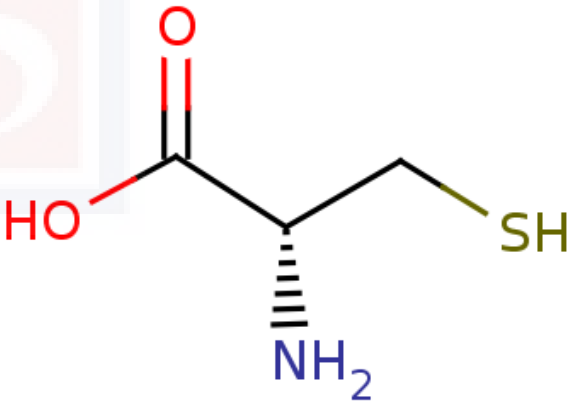


Experimental MS/MS of 122.0270 in positive mode



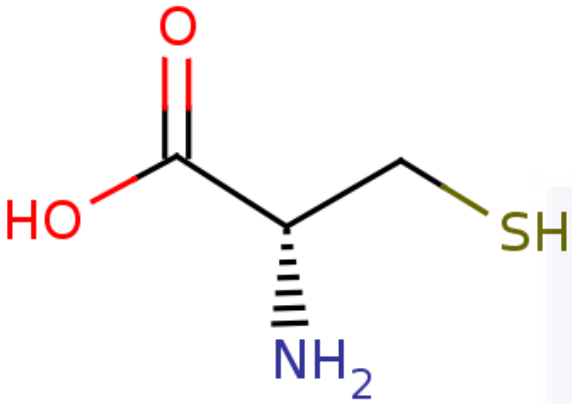


O, S, P

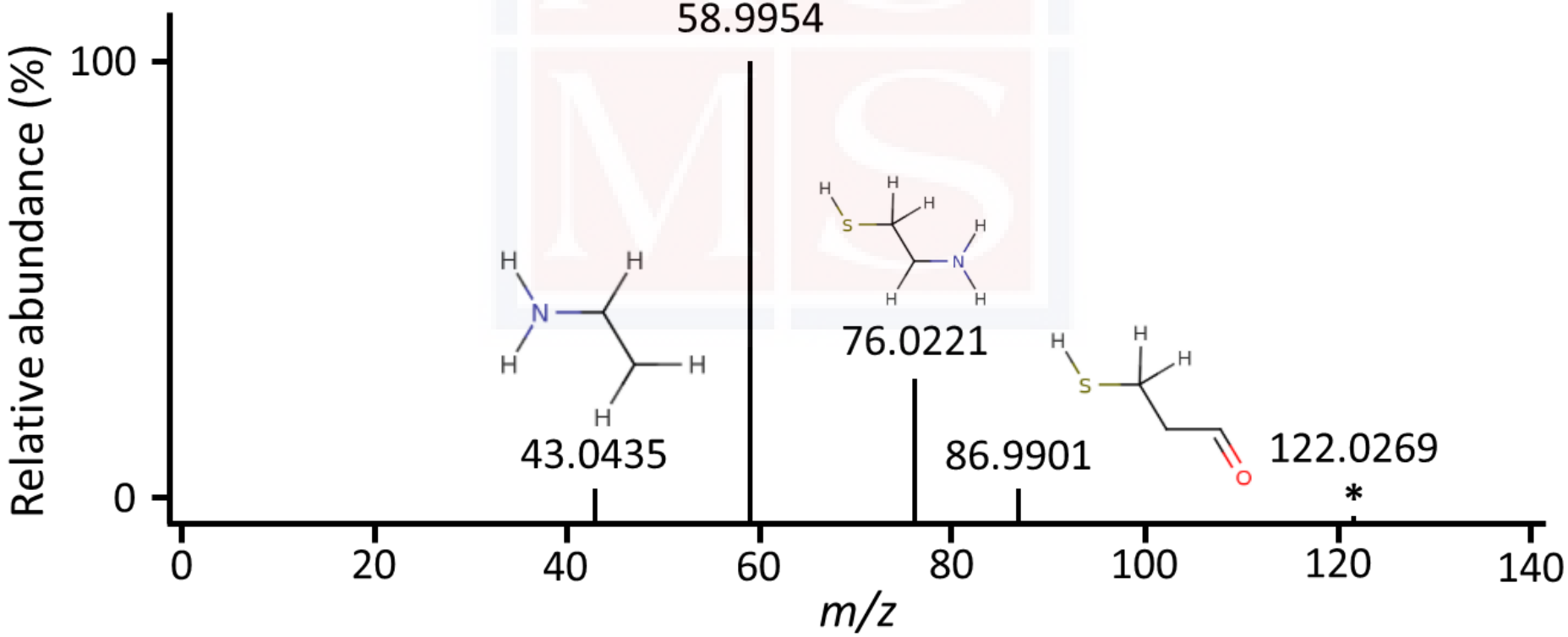


Cysteine (C3H7NO2S)

Putative metabolite



→ **Cysteine (C3H7NO2S)**



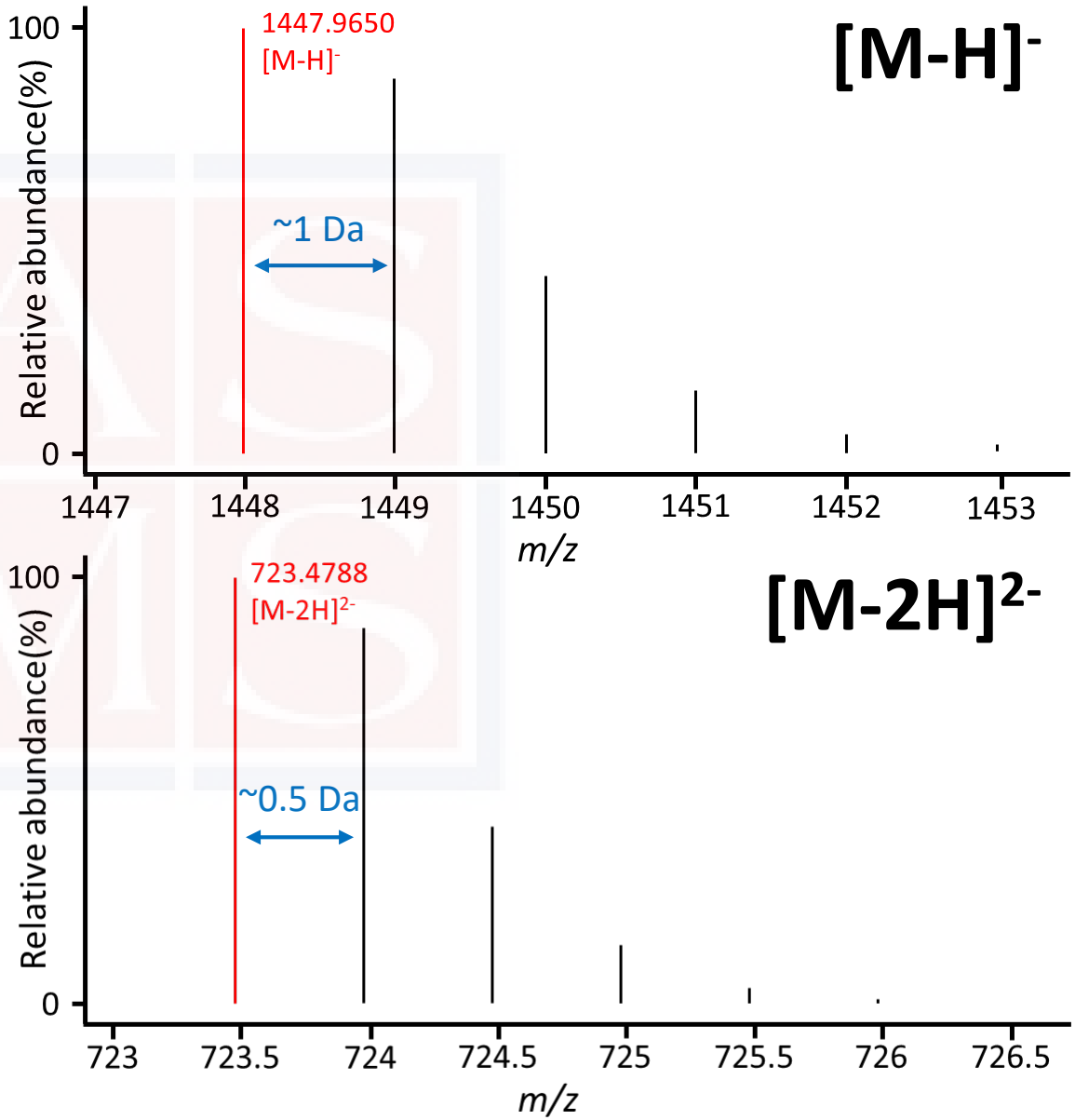
Multiple charged species

- Very common phenomenon in proteomics. Rarely observed in metabolomics.

$$\frac{M+1-H}{1} - \frac{M-H}{1} = 1$$

CARDIOLIPIN(72:8)
C81H142O17P2

$$\frac{M+1-2H}{2} - \frac{M-2H}{2} = 0.5$$



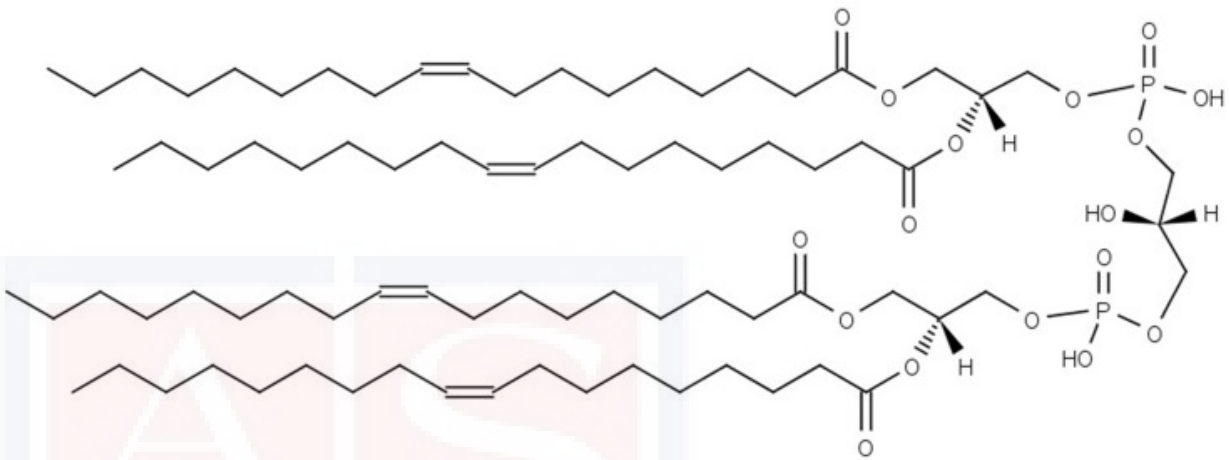
Example III. Multiple charged species

Which information can be extracted from the presence of a multiple charged feature?

1. Annotation. Incorporated into XCMS Online.
2. Presence of two or more highly ionizable functional groups: phosphates in negative and amines in positive.
3. Discrimination from features coming from single charged molecules.

Example III. Multiple charged species

CARDIOLIPIN(72:6)
C81H142O17P2
Mass=1448.9722



- Abundance of $[M-H]^-$ and $[M-2H]^{2-}$ ions is similar for this family of molecules (depending on the instrument).

1. Typical profiling experiments in untargeted lipidomics go up to *m/z*=1000-1200. The $[M-H]^-$ ions are overlooked. 👎
2. Use of $[M-2H]^{2-}$ ions (range 670-760) to study these molecules. 👍
3. This mass range is the same for the major phospholipid class PE. 👎
4. The isotopic pattern can be used to differentiate between PE isotopes and cardiolipin double charged parent ions. 👍

TABLE OF CONTENTS

- ENDOGENOUS ISOTOPIC DISTRIBUTION OF A FEATURE.
- GENERATION OF ISOTOPIC DATA USING UNIFORMLY-LABELED MICROORGANISMS.
- IDENTIFICATION OF UNKNOWNS USING ISOTOPES.

Uniformly-labeled metabolites

Information

Complexity

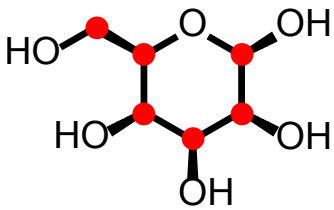
Use of intrinsic stable isotopes of metabolites to gain insight about features and help in annotation and identification

Growth of uniformly-labeled microorganisms

Generation of molecules where all atoms are stable isotopes: uniformly-labeled metabolites

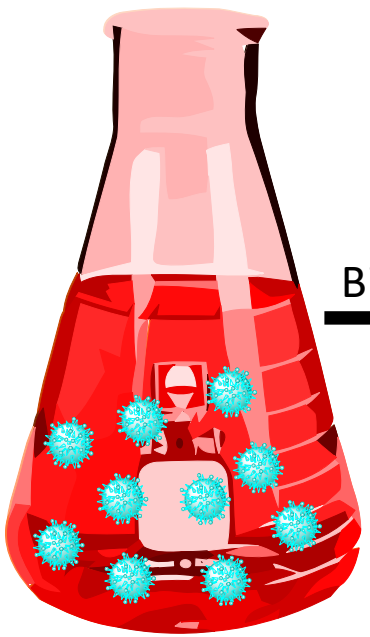
Growth of uniformly-labeled microorganisms

^{13}C-Glucose

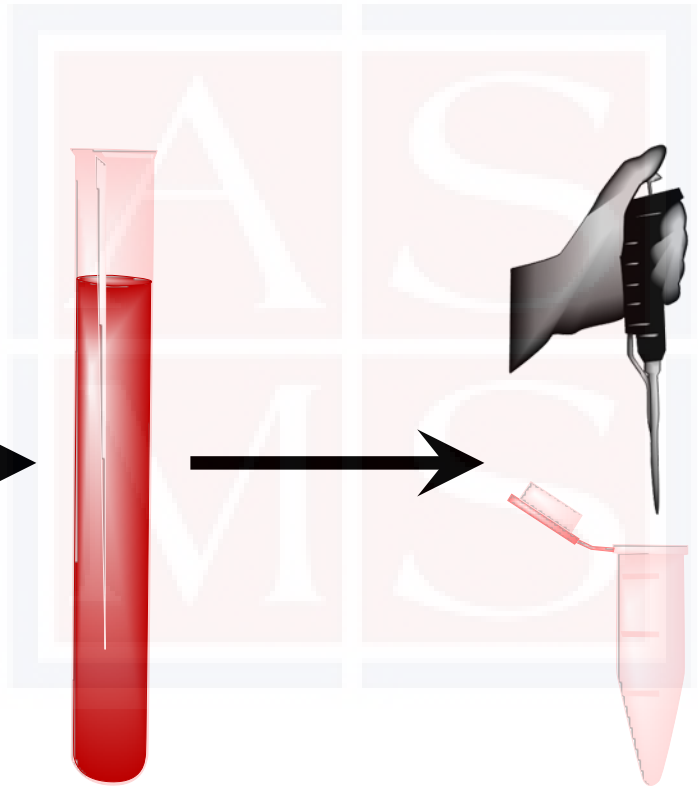


^{13}C-Labeled Microorganisms

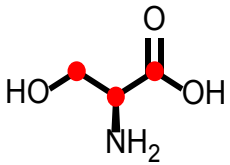
Metabolite extraction



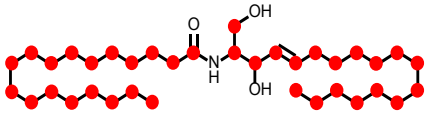
Bioreactor



Polar metabolites



Lipids



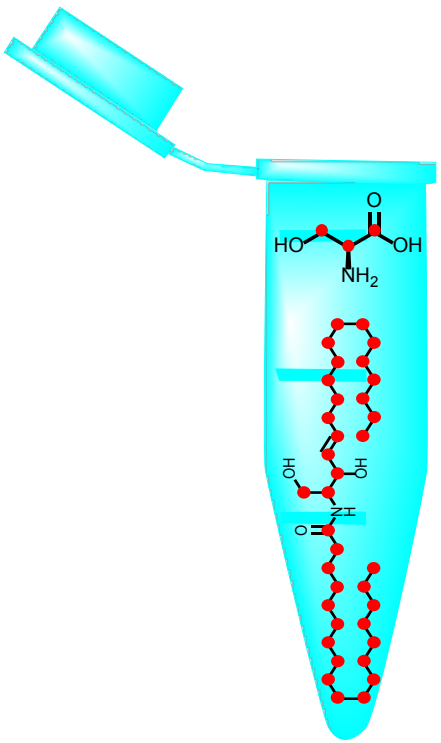
- Bacteria (Weiner, 2015)
- Yeast (Rampler, 2017)
- Plants (Bueschl, 2014)

>99% labeling efficiency

Uniformly-labeled extracts as internal standards

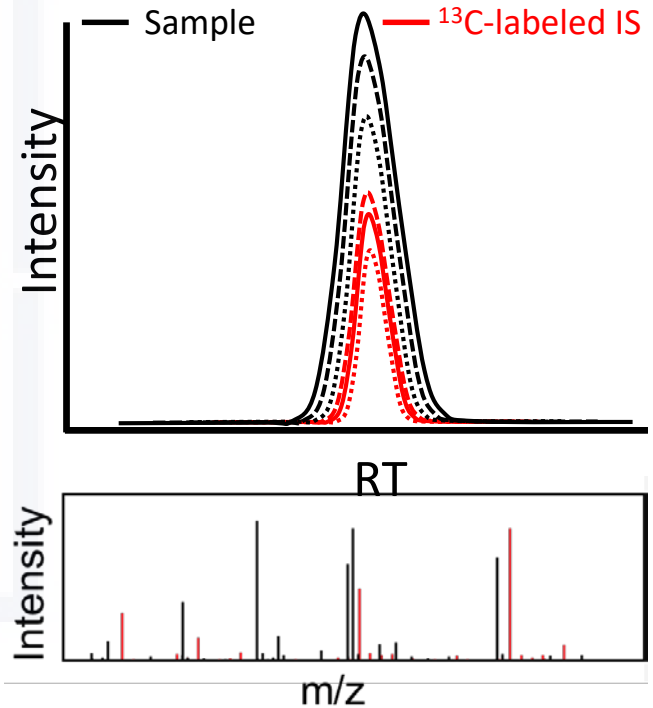
Add to samples as internal standards

Quantification of many compounds

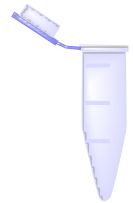


Untargeted generation of accurate MS/MS spectra of the ^{13}C-labeled molecules

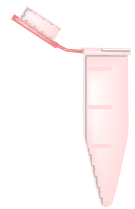
We envisioned the opportunity to generate a vast isotopic MS/MS data to help identifying compounds



Generation of isotopic data



^{12}C-labeled



^{13}C-labeled

Lipid extraction → Lipidomics

Polar metabolites extraction → Reversed-phase (hydrophobic) / HILIC (hydrophilic)

Peak detection alignment and integration

Sorting of isotopologues

→

1. Samples ***2. Sample prep*** ***3. LC/MS*** ***4. Data analysis***

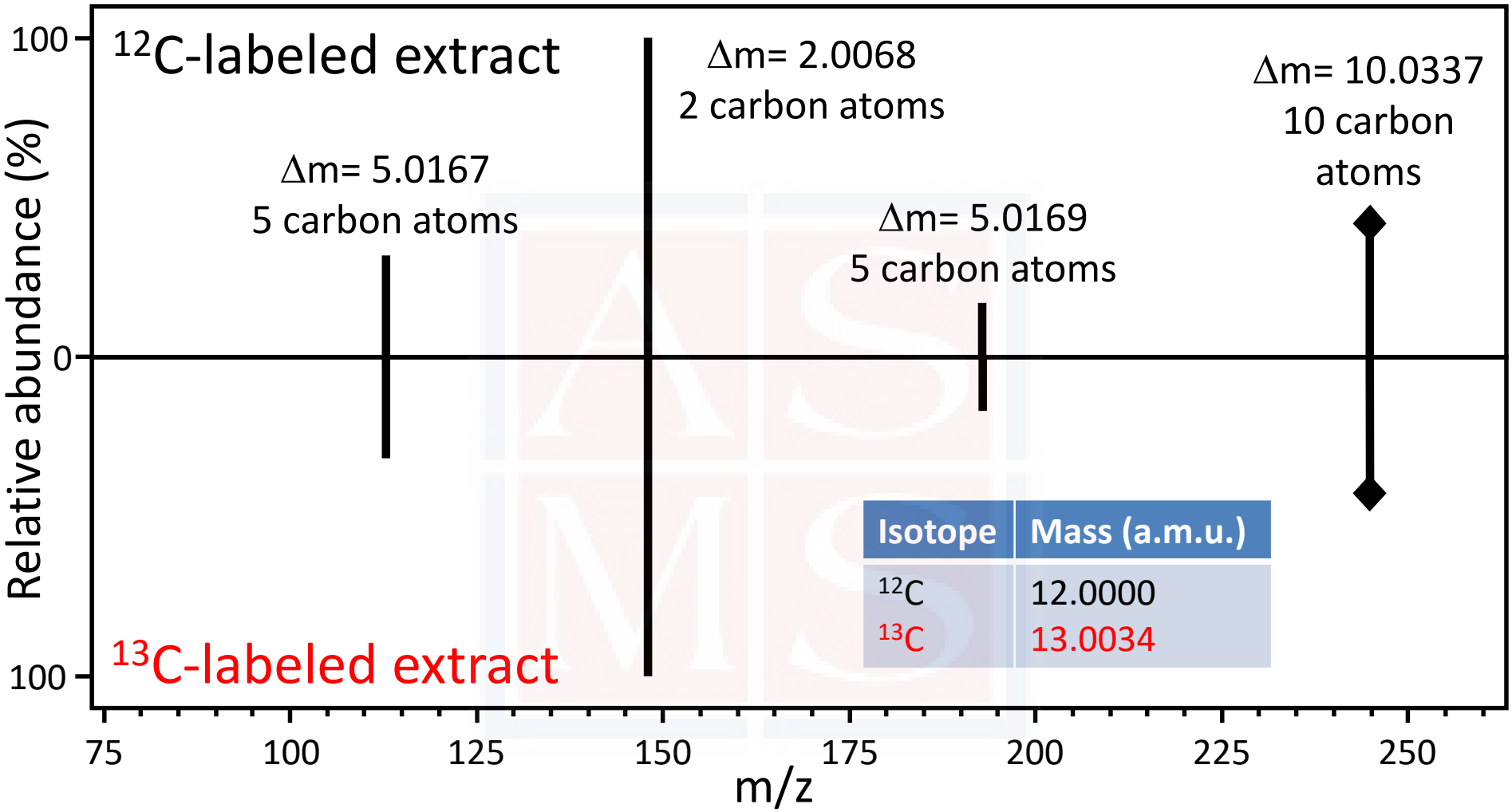
Feature (*m/z*)	Isotopologues (*m/z*)
90.0546	90.0546
	93.0645
104.1066	104.1066
	109.1231

5. List of isotopologues

6. Analysis of MS/MS spectra

X^{13}CMS: Huang et al., Anal. Chem., 2014.

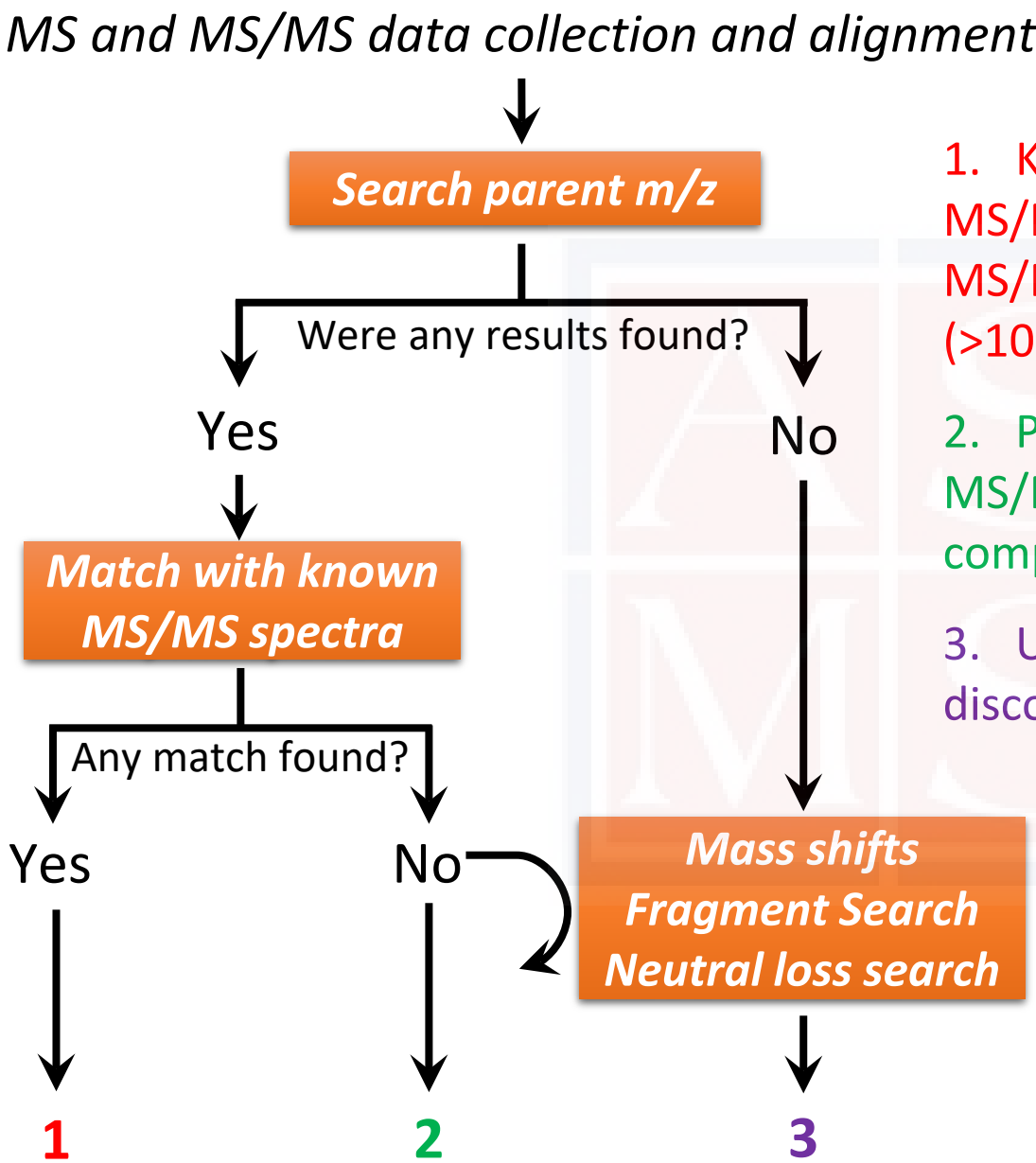
Analysis of MS/MS spectra: overview



Isotope	Mass (a.m.u.)
^{12}C	12.0000
^{13}C	13.0034

Use this information to gain insight about structural properties of molecules

Workflow for the identification of uniformly-labeled metabolites



1. Known metabolite with known MS/MS spectra→ Generation of MS/MS data of the ^{13}C isotopomers (>100 incorporated into isoMETLIN).

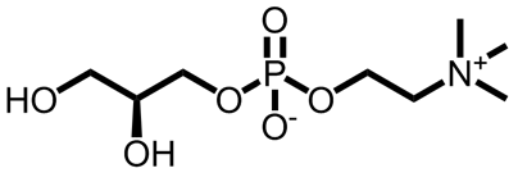
2. Putative metabolite with unknown MS/MS spectra→ Identification of compounds.

3. Unknown metabolite→ Tentative discovery of compounds.

TABLE OF CONTENTS

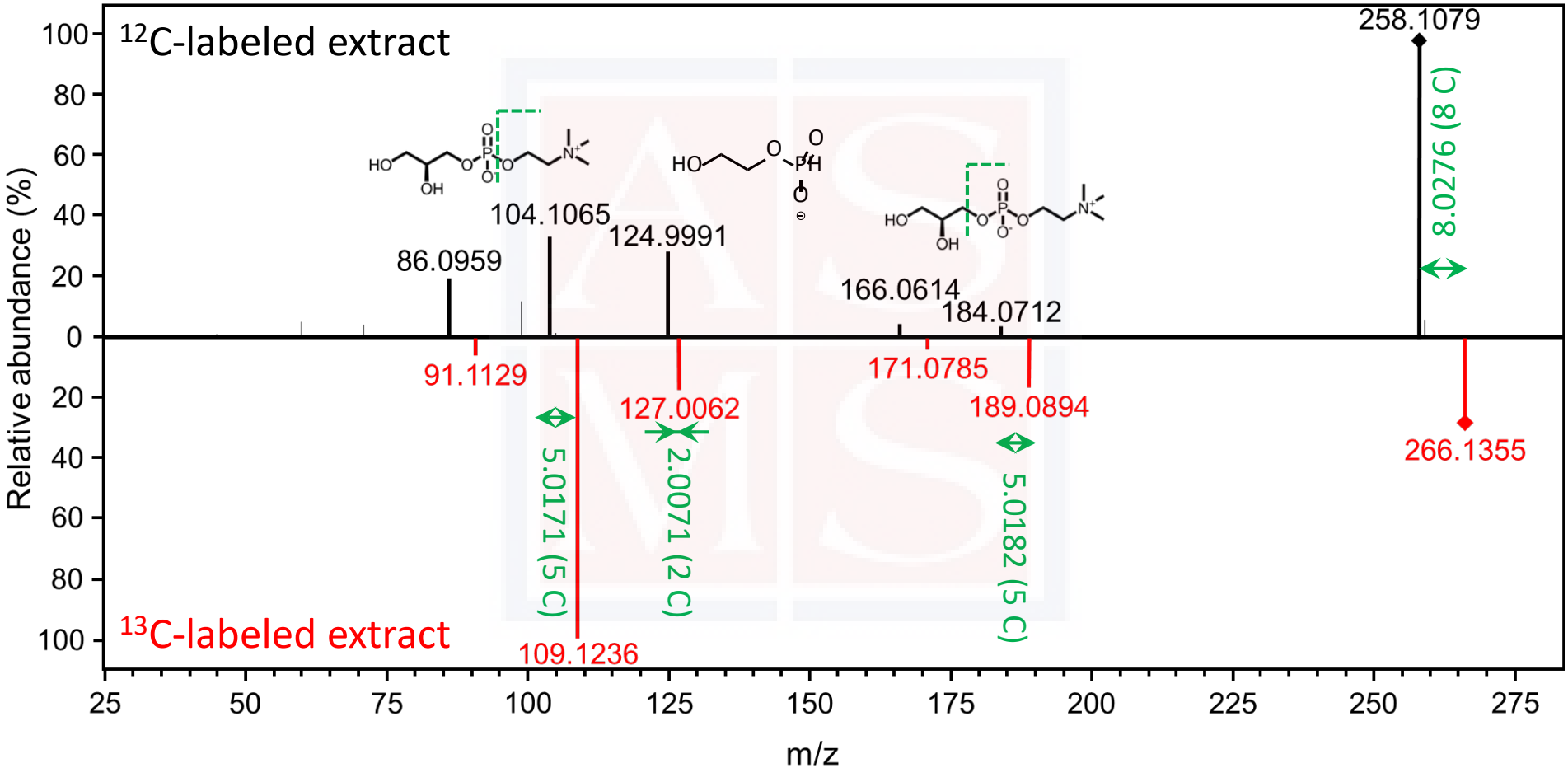
- ENDOGENOUS ISOTOPIC DISTRIBUTION OF A FEATURE.
- GENERATION OF ISOTOPIC DATA USING UNIFORMLY-LABELED MICROORGANISMS.
- IDENTIFICATION OF UNKNOWNS USING ISOTOPES.

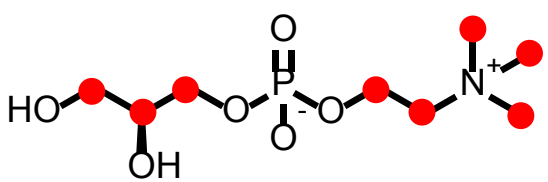
Example I. Known metabolite, known MS/MS



Chemical Formula: $C_8H_{20}NO_6P$

sn-Glycero-3-phosphocholine





Chemical Formula: $^{13}C_8H_{20}NO_6P$

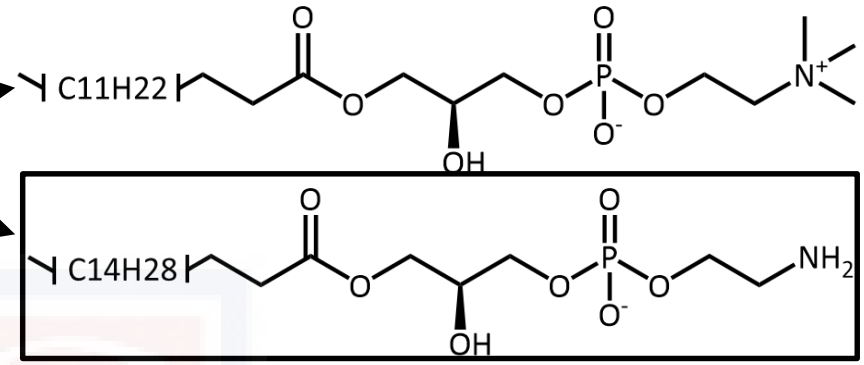
^{13}C-*sn*-Glycero-3-phosphocholine

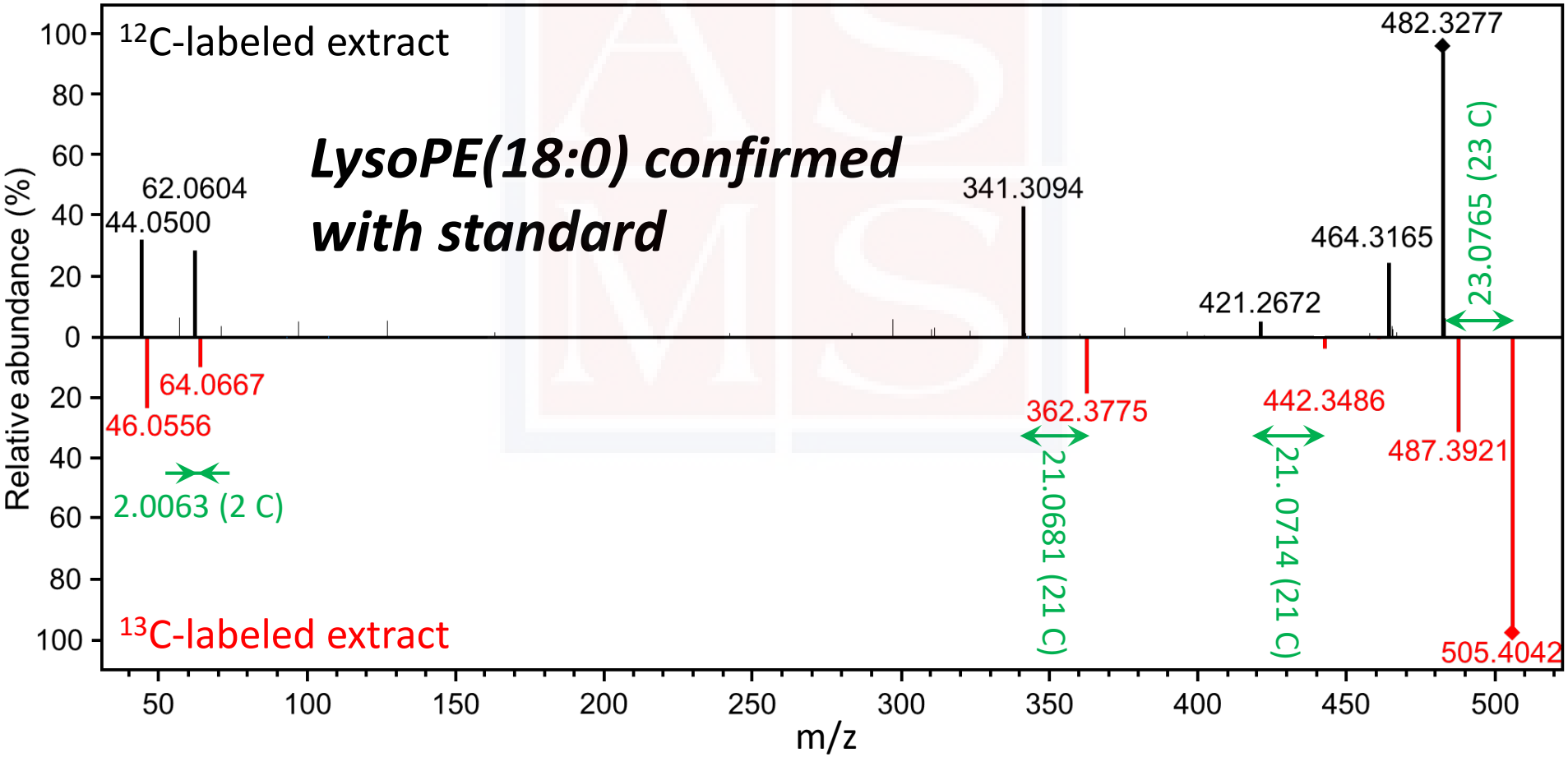
Example II. Putative metabolite, unknown MS/MS

Search of *m/z*:
482.3277
Error: 10 ppm

List of putative metabolites:

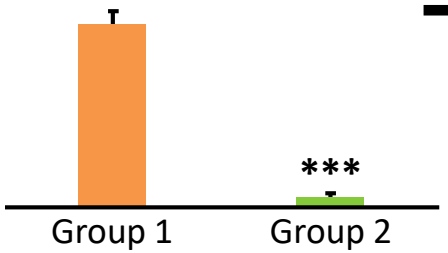
- LysoPC(15:0) (C23H48NO7P)
- LysoPE(18:0) (C23H48NO7P)
- PUBCHEM_54608258 (C24H54BrN2S)
- PUBCHEM_71749542 (C30H43NO4)

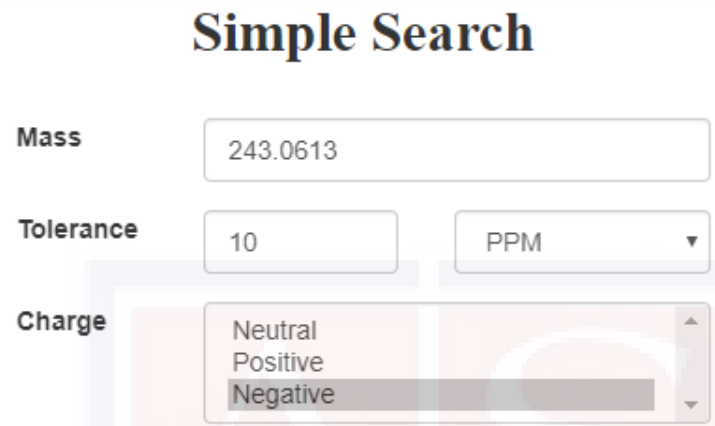




Example III. Putative metabolite, unknown MS/MS

Unknown feature
m/z=243.0613
Found in a different project

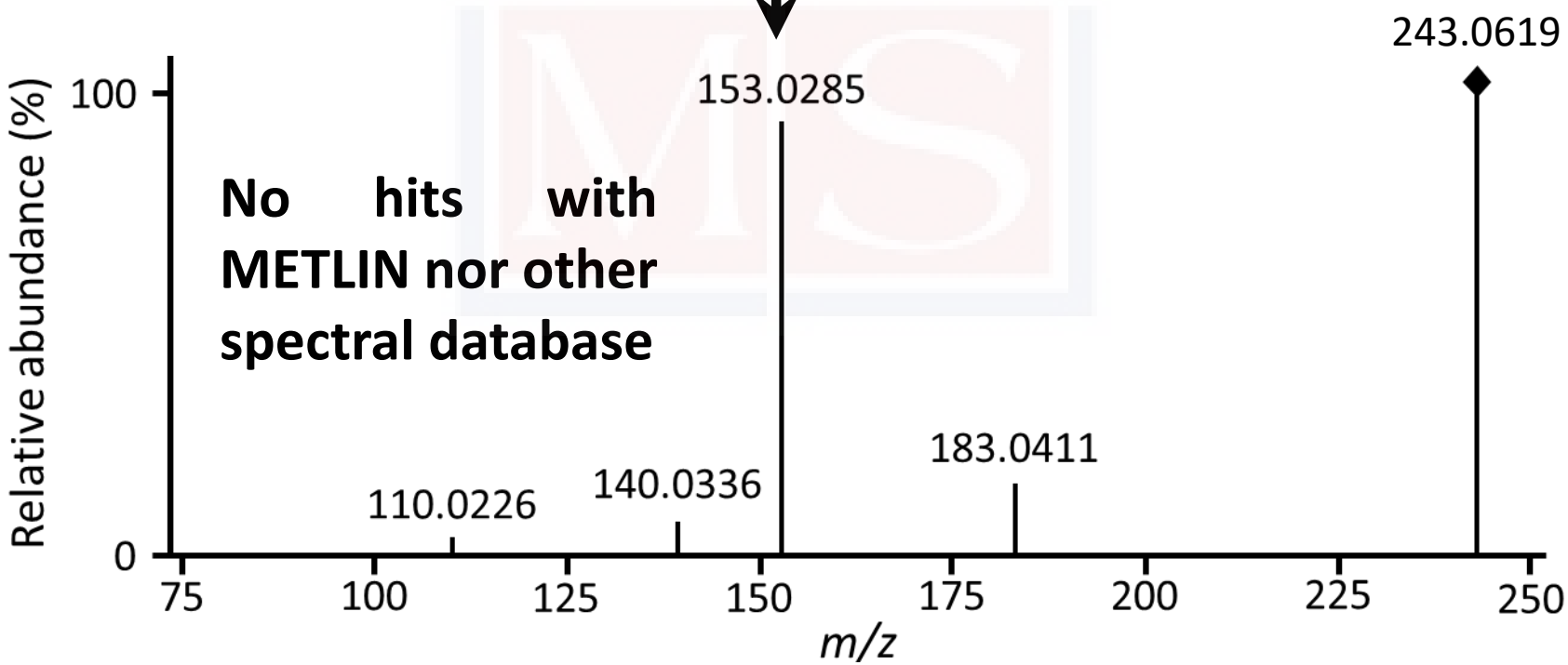




List of putative metabolites:

- Uridine (C9H12N2O6)
- Pseudouridine (C9H12N2O6)
- C10H8N6O2
- C11H17O2PS
- C13H12N2OS
- Others

MS/MS

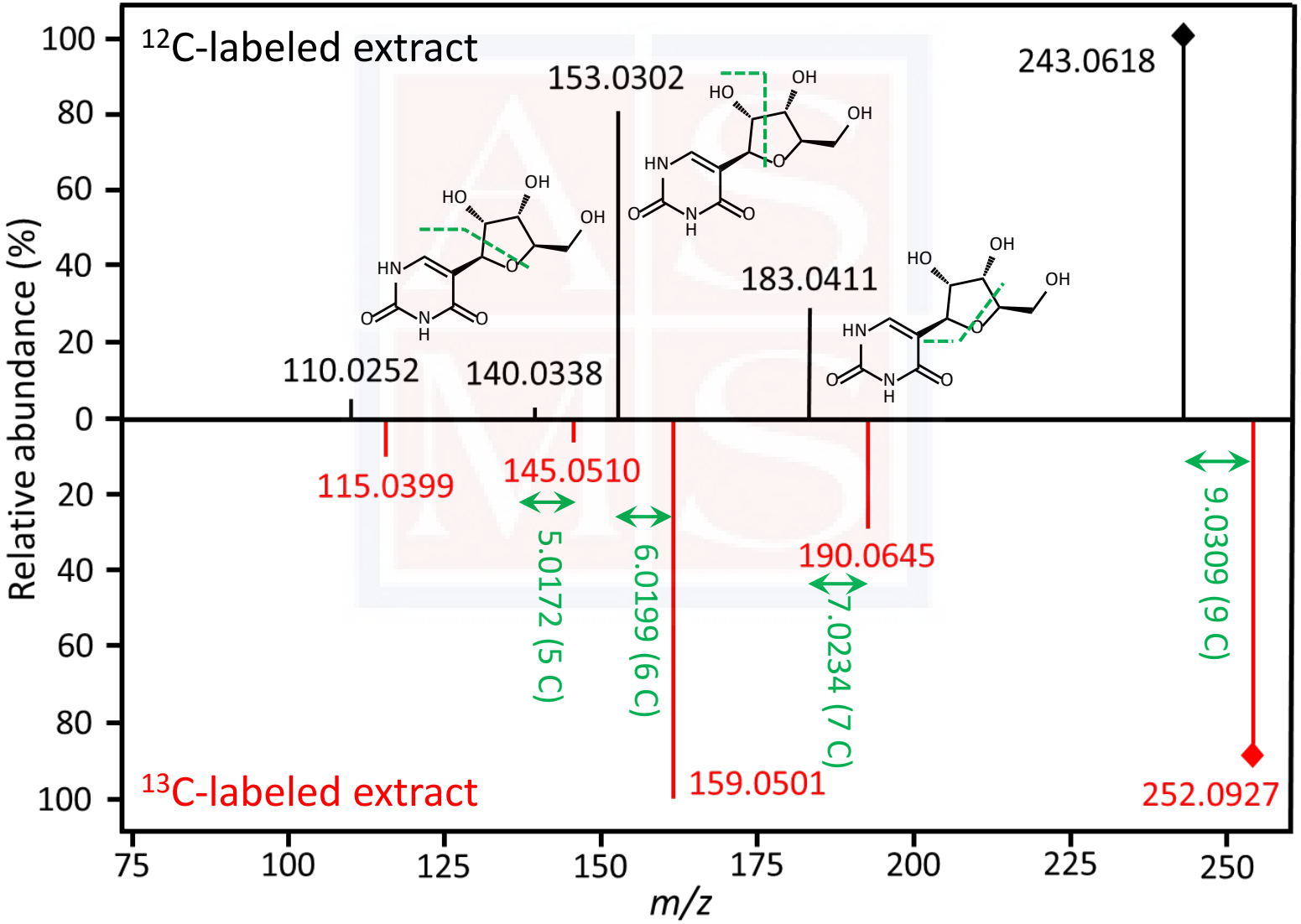


No hits with METLIN nor other spectral database

Example III. Putative metabolite, unknown MS/MS

- Uridine (C9H12N2O6)
- Pseudouridine (C9H12N2O6)
- C10H8N6O2
- C11H17O2PS
- C13H12N2OS
- Others

Pseudouridine confirmed with standard

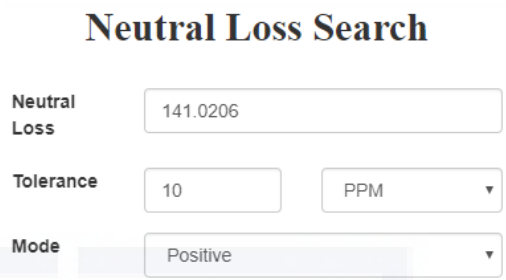


Example IV. Unknown metabolite. Tentative discovery

Search of *m/z*:
608.3972
Error: 10 ppm
No hits in METLIN and PubChem

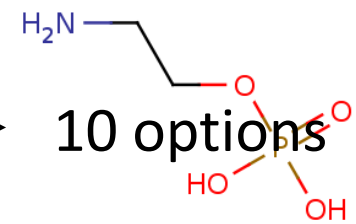
→

List of putative formulas:
42 molecular formulas
With 30 carbon atoms:
- C30H53N7O4S
- C30H58NO9P



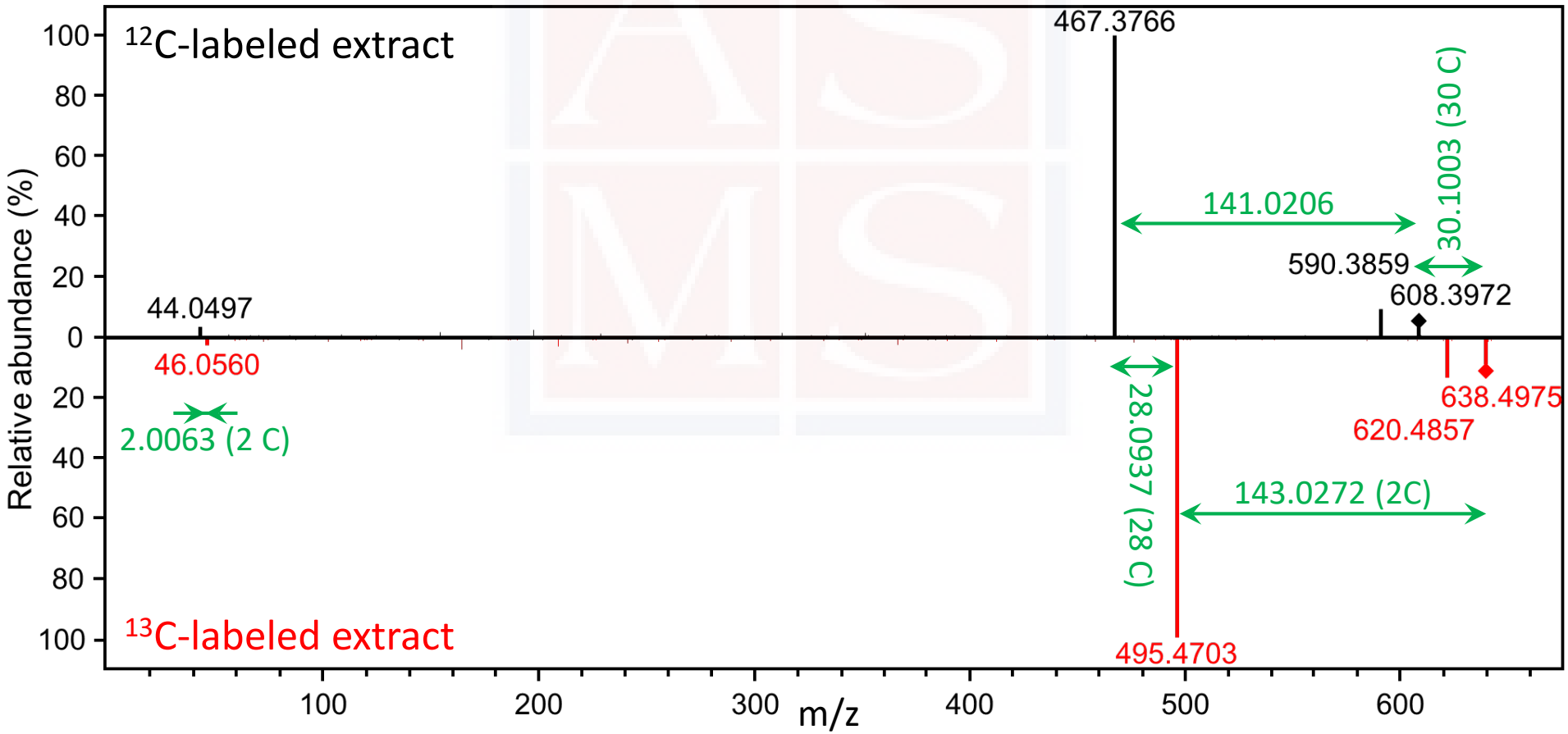
- 2 carbon atoms
- Intensity> 30%

→



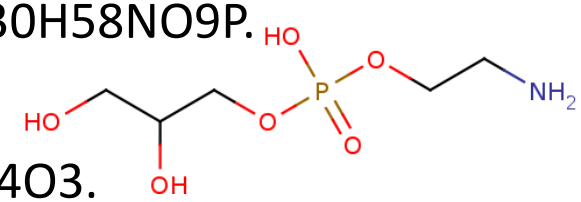
10 options

Phosphatidylethanolamine

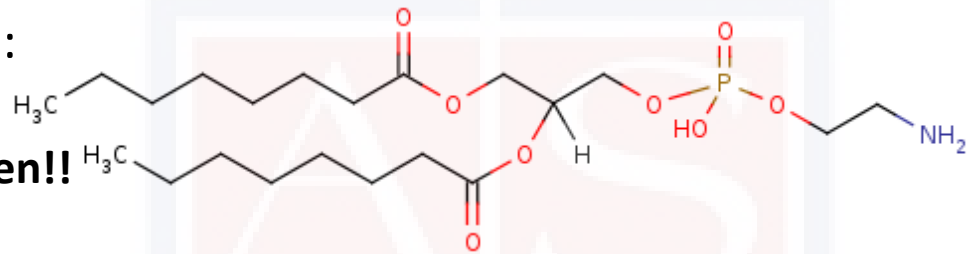


Example IV. Unknown metabolite. Tentative discovery

- ✓ Molecule with 30 carbon atoms. Molecular formula: C30H58NO9P.
- ✓ Phosphatidylethanolamine group (C5H14NO6P):



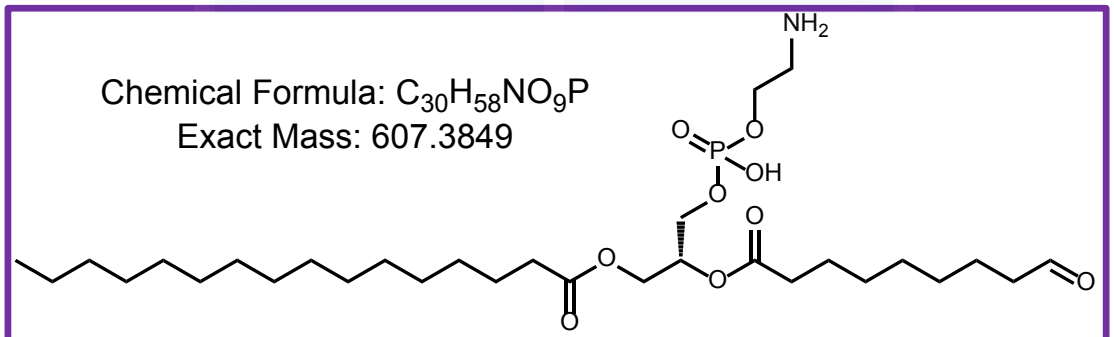
- ✓ Rest of molecule: C30H58NO9P – C5H14NO6P = C25H44O3.
- ✓ PE phospholipid:



One more oxygen!!

- ✓ PE phospholipid with an <u>oxidized fatty acid</u>. Total number of carbon atoms= 25.
- ✓ Biological context:
  - ❑ Oxidized PE are products of ozonolysis in bronchoalveolar lavage (*Almstrand, et al., Anal. Biochem., 2015*).
  - ❑ Palmitoyl-9-oxo-nonanoyl-PC is a product of lung surfactant phospholipid oxidation in smokers (*Kimura, et al., Lung, 2012*).

Chemical Formula: $C_{30}H_{58}NO_9P$
Exact Mass: 607.3849



1-hexadecanoyl-2-(9-oxo-nonanoyl)-*sn*-glycero-3-phosphoethanolamine

CONCLUSIONS

- ✓ **The endogenous isotopic distribution of a feature may help with its annotation and identification:**

  1. **The M+1, M+2 and M+3 relative intensity aids to narrow down the possible molecular formula of a feature.**
  2. **The mass defect is an indicator of the presence of different atoms in the molecule, helping with its annotation and the analysis of its MS/MS spectra.**
  3. **The generation of multiple charged ions is a valuable tool for the analysis of metabolites with high molecular weight, but also a double-edged sword.**

- ✓ **A precious endogenous Vs isotope-labeled MS/MS data can be generated from extracts of isotope-labeled microorganisms. This data can be used to:**

  1. **Generate MS/MS data of isotopes of known molecules for quantitative metabolomics.**
  2. **Help with the identification of metabolites that lack of MS/MS data in databases.**
  3. **Discover new metabolites.**

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

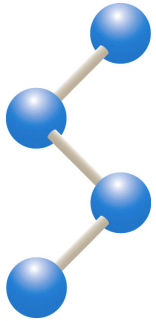
---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics



- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Autonomous Strategies

- Moving metabolomics into the future
  - Automation
    - Robots
    - Netflix
  - Smart systems
    - Simple literature reviews
    - Watson work
    - Wrapping your own cognitive learning suit

Autonomous Strategies

- How many people have a robot in their labs?

- How many have automated workflows?

Autonomous Strategies

transcriptic

DISCOVERY BIOLOGY ON DEMAND

The Emerald Cloud Lab

At your command

Brian Frezza former Ghadiri lab

The Automation of Science

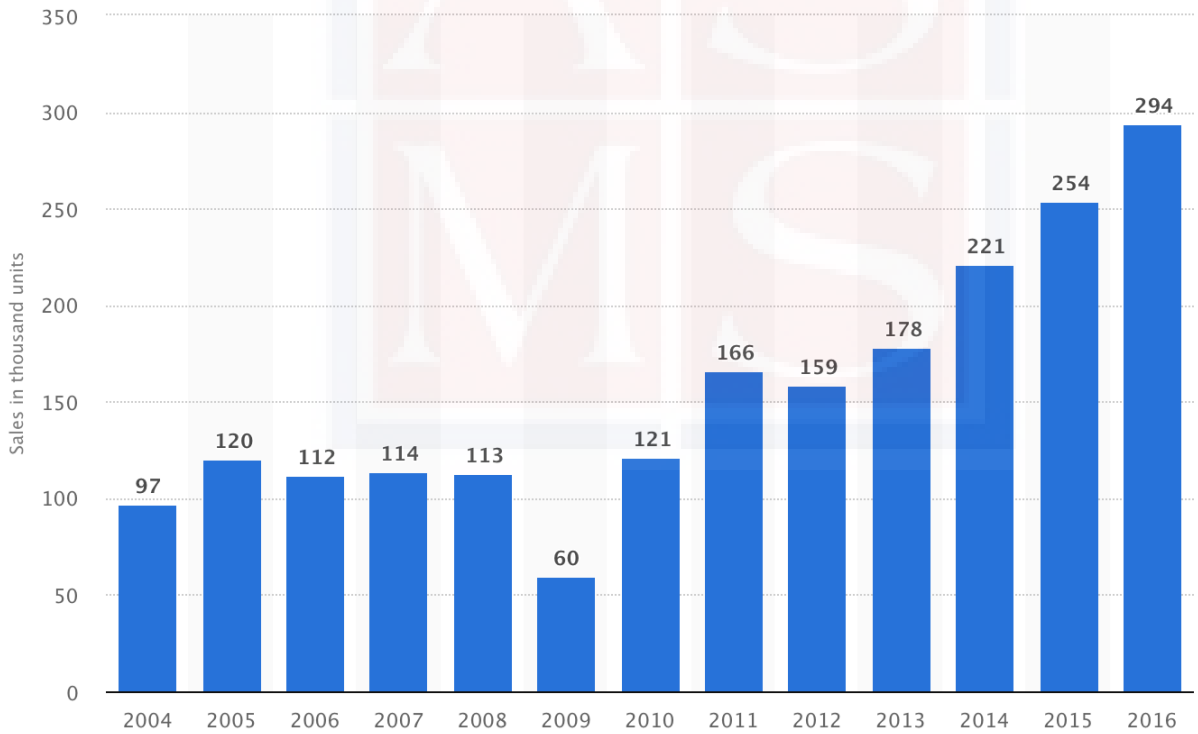
Ross D. King,[1*] Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research, we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested treelike structure, 10 levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalization describes how a machine contributed to scientific knowledge.

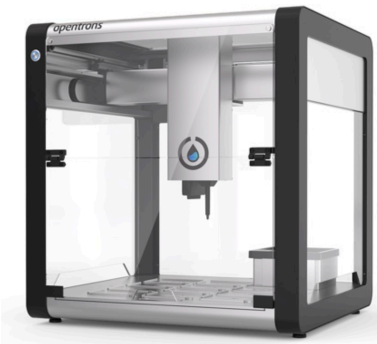
www.sciencemag.org SCIENCE VOL 324 3 APRIL 2009

Autonomous Strategies – robots

- Robot prices have dropped !
  - No excuse to not have automation
- Software has been greatly improved
  - Drag and drop systems



Industrial robot sales – statista.com



opentrons

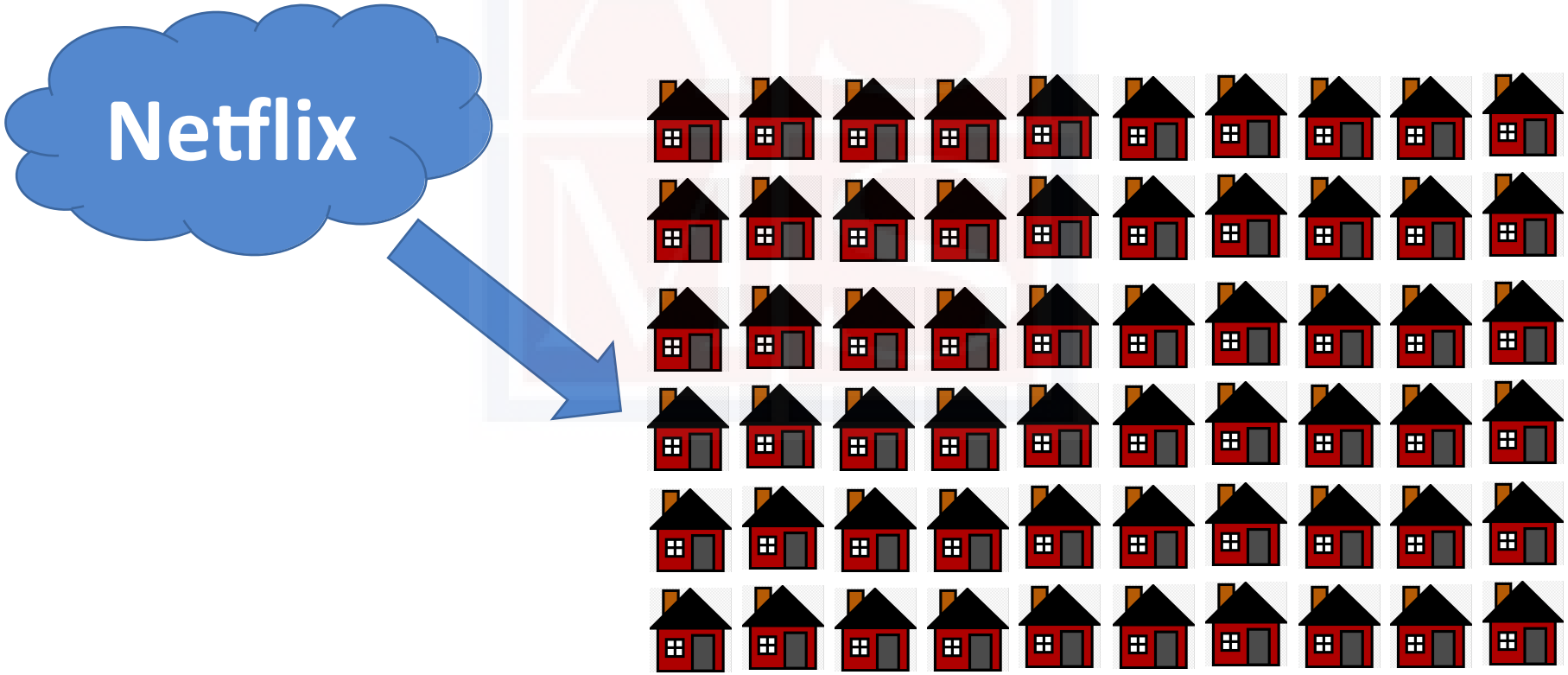
Autonomous Strategies

- Easy automation via IFTTT (if that then do this)
  - Get notifications
  - Start a second

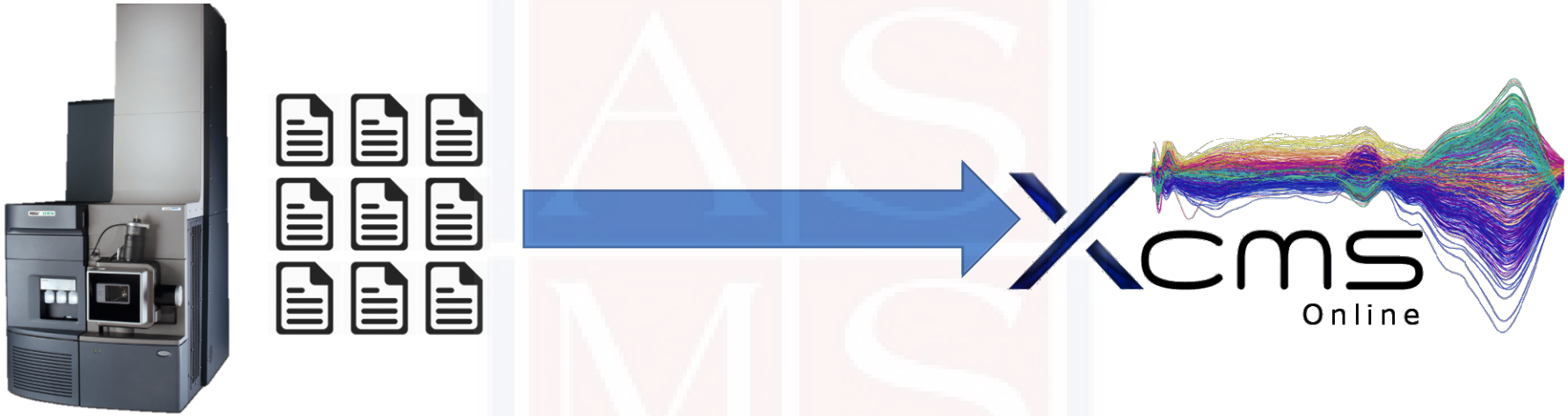
microbot

Autonomous Strategies - netflix

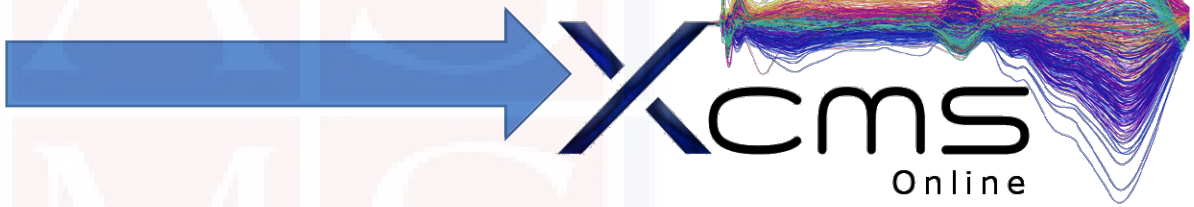
- Netflix has a problem.
  - How to get hundreds of videos to hundreds of people simultaneously around the world ?



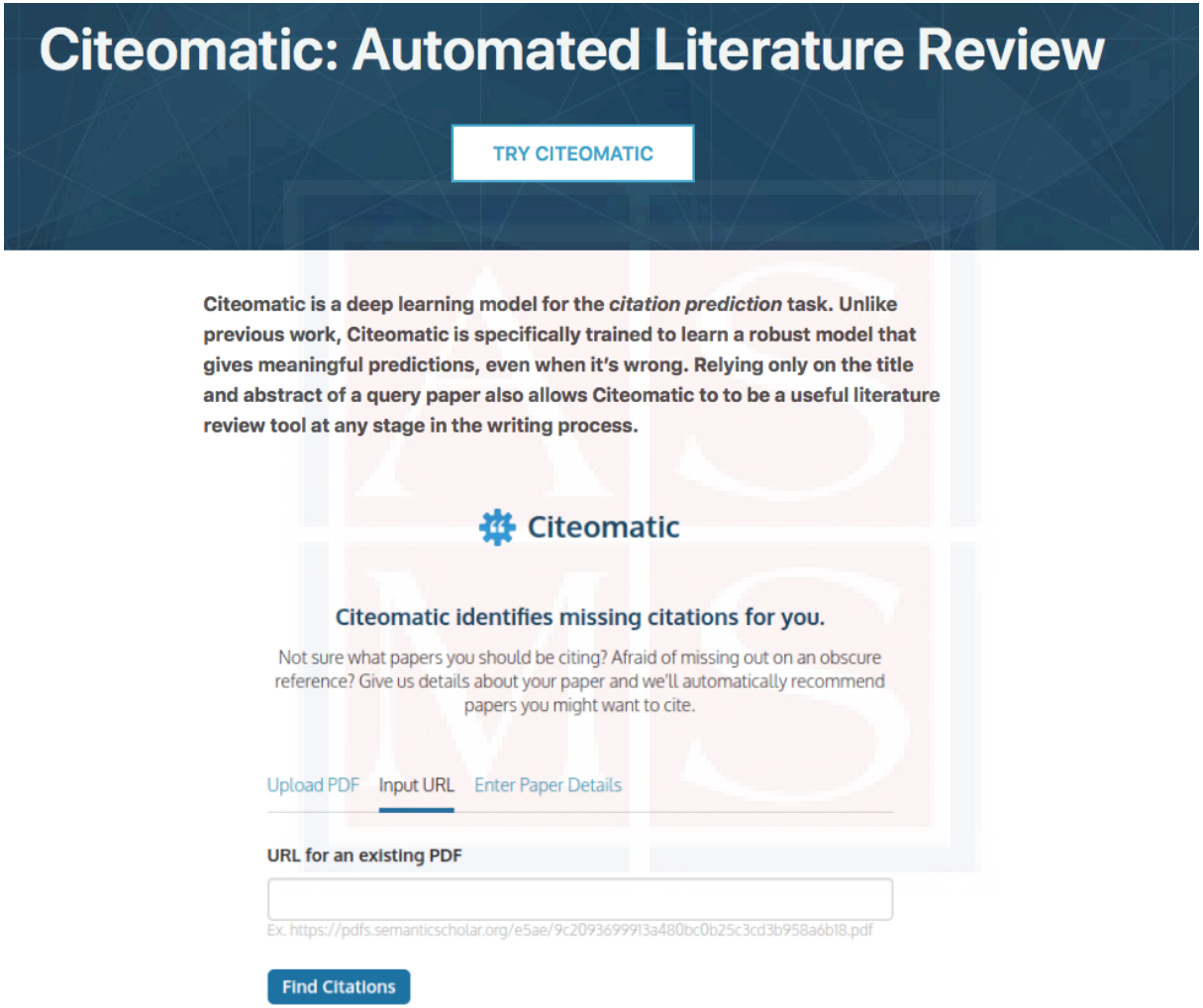
Autonomous Strategies – Netflix



Autonomous Strategies – Netflix



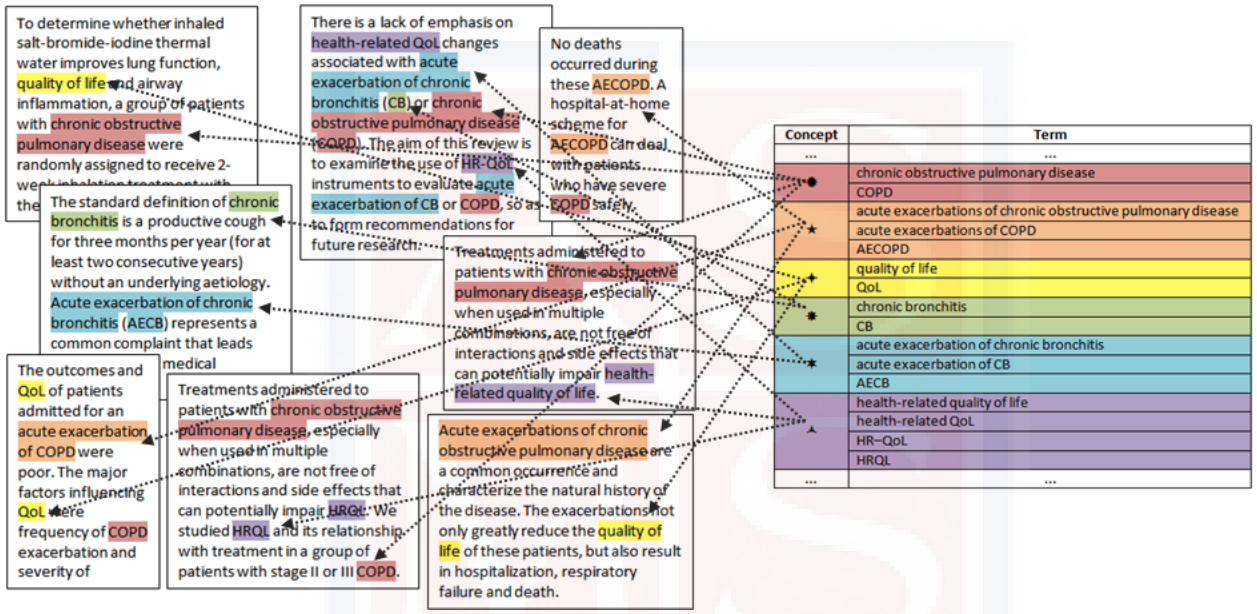
Autonomous Strategies – Simple Literature reviews



- Not quite as good as it sounds.
- Looks historically

Autonomous Strategies – simple literature reviews

FlexiTerm



- Downloadable open source
- Neutralises source documents

Autonomous Strategies - IBM

analytical chemistry

Article

Cite This: *Anal. Chem.* 2017, 89, 11505-11513

pubs.acs.org/ac

Exposome-Scale Investigations Guided by Global Metabolomics, Pathway Analysis, and Cognitive Computing

Benedikt Warth,[*,†,§] Scott Spangler,[‖] Mingliang Fang,[†,⊥] Caroline H. Johnson,[#] Erica M. Forsberg,[†] Ana Granados,[†] Richard L. Martin,[‖] Xavier Domingo-Almenara,[†] Tao Huan,[†] Duane Rinehart,[†] J. Rafael Montenegro-Burke,[†] Brian Hilmers,[†] Aries Aisporna,[†] Linh T. Hoang,[†] Winnie Uritboonthai,[†] H. Paul Benton,[†] Susan D. Richardson,[▽] Antony J. Williams,[○] and Gary Siuzdak[*,†,‡]

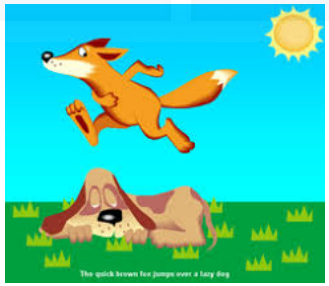
IBM Watson

Autonomous Strategies - IBM

The red fox jumped over the lazy brown dog

The **red** ? jumped over the lazy **brown** 🐶

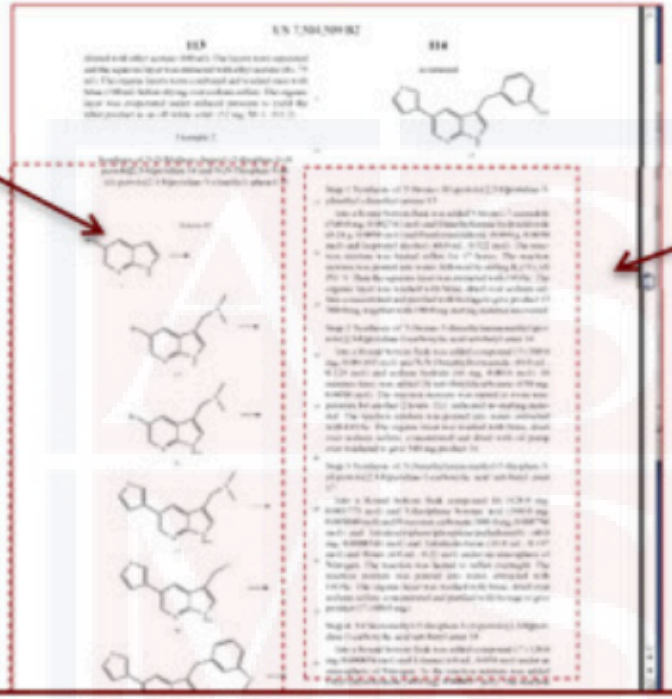
The **red** ?   *the* 😴 **brown** 🐶



Autonomous Strategies - IBM

As bitmap images

Picture of chemicals found in the document Images



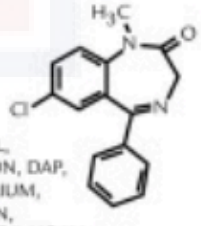
As text

Chemical names found in the text of documents

Partents also have (Manually Created) Chemical Complex Work Units (CWU's)

Nomenclature issuses: Valium has > 149 " names "

Valium (Trade Name) = Diazepam (Generic Name) = CAS # 439-14-5 (Chemical ID #) =



ALBORAL, ALISEUM, ALUPRAM, AMIPROL, ANSIOLIN, ANSIOLISINA, APALIRIN, APOZEPA, ASSIVAL, ATENSINE, ATILEN, BIALZEPAM, CALMOCITENE, CALMPOSE, CERCINE, CEREGULART, CONDITION, DAP, DIACEPAN, DIAPAM, DIAZEMULS, DIAZEPAN, DIAZETARD, DIENPAX, DIPAM, DIPEZONA, DOMALIUM, DUKSEN, DUXEN, E-PAM, ERIDAN, EVACALM, FAUSTAN, FREUDAL, FRUSTAN, GIHITAN, HORIZON, KITRIUM, LA-III, LEMBROL, LEVIUM, LIBERETAS, METHYL, DIAZEPINONE, MOROSAN, NEUROLYTRIL NOAN NSC-77518 PACITRAN PARANTEN PAXATE PAXEL PLIDAN QUETINIL QUIATRIL QUIEVITA RELAMINAL RELANIUM RELAX RENBORIN RO 5-2807 S.A.R.L SAROMET SEDAPAM SEDIPAM SEDUKSEN SEDUXEN, SERENACK SERENAMIN SERENZIN SETONIL SIBAZON SONACON STESOLIN, TENSOPAM TRANIMUL TRANQDYN TRANQUASE TRANQUIRIT, TRANQUO-TABUNEN, YMBRIUM UNISEDIL USEMPAXAP VALEO VALITRAN VALRELEASE VATRAN VELIUM, VIVAL VIVOL WY-3467

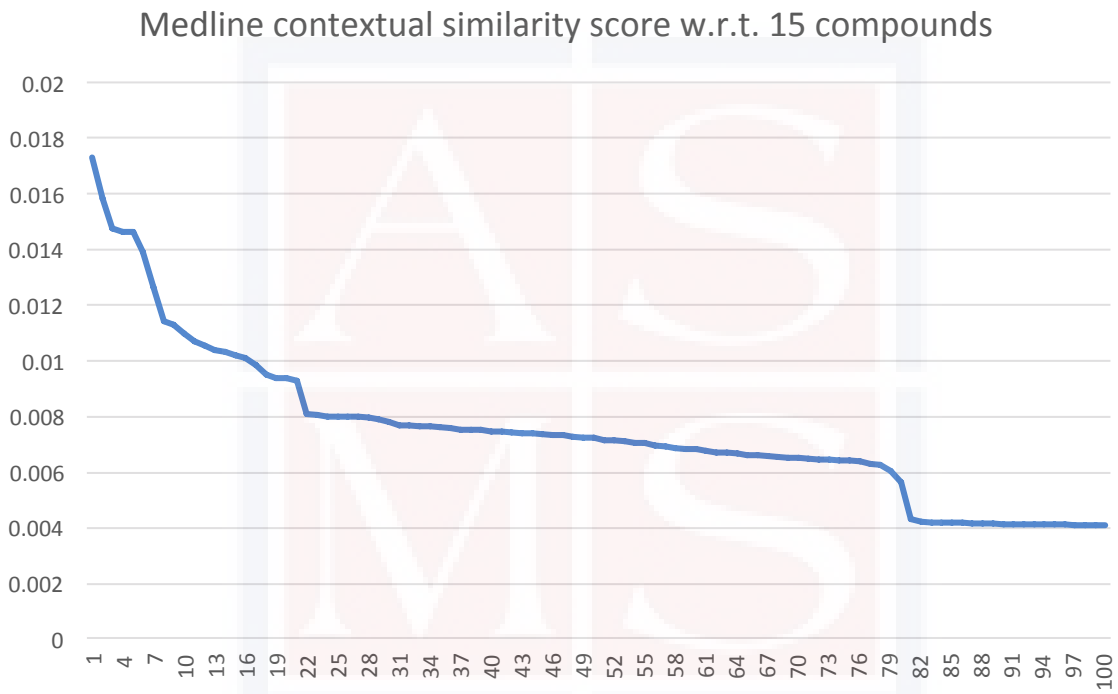
Chemical nomenclature can be daunting

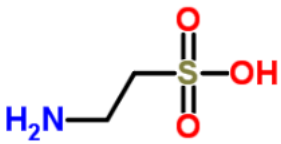
Autonomous Strategies - IBM

well as documents containing any other synonymous chemical representation. A reader capable of absorbing 10 papers per day would need nearly 100 years to go through this potentially relevant literature, which is an unrealistic feat. Instead, Watson[47] mines text so as to create a model for each metabolite that represents all the terms present in the abstracts of the papers that specifically mention the named metabolite.

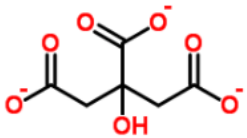
We believe that this example demonstrates the predictive potential of Watson in finding new potential EDCs similar to the training set. More importantly, cognitive computing is not limited to a specific mode of action and may be extended to other toxicant classes such as carcinogens or genotoxic compounds. Therefore, the tested machine-learning strategy provides a valuable resource for future identification of suspects and literature-based priority ranking. This holds the potential to screen tens to hundreds of thousands of chemicals in some hours/days which would not be possible manually. Especially when merging this kind of prioritizing with screening of untargeted LC–MS data as outlined above, this technology opens up for new and unexpected discoveries regarding both exposure and effect. This approach is of special value to the

Autonomous Strategies - IBM

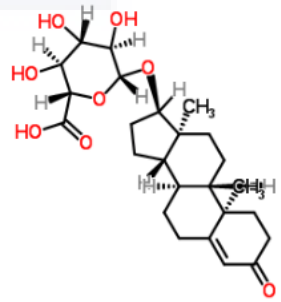




Taurine



Citrate



Epitestosterone Glucronide

Autonomous Strategies – roll your own

The ContentMine

The ContentMine is extracting 100 million facts from the academic literature

UK http://contentmine.org team@contentmine.org

DeepDive

Quick Start | Documentation | Showcase | Papers | Data | Chat | Forum

Fork me on GitHub

- Launching or Installing DeepDive
  - Quick launching
  - Quick installation
- Running your first DeepDive app
- Next steps

DeepDive Quick Start

DeepDive helps you extract structured knowledge from less-structured data with statistical inference without having to write any sophisticated machine learning code. Here we show how you can quickly install and run your first DeepDive application.

Launching or Installing DeepDive

Quick launching

First, you can quickly launch DeepDive with minimal installation using Docker by running the following command:.

```
$ bash <(curl -fsSL git.io/getdeepdive)
```

Then selecting the `deepdive_docker_sandbox` option:

```
### DeepDive installer for Mac
1) deepdive                  5) jupyter_notebook
2) deepdive_docker_sandbox   6) postgres
3) deepdive_example_notebook 7) run_deepdive_tests
4) deepdive_from_release     8) spouse_example
# Install what (enter to repeat options, a to see all, q to quit, or a number)? 2
```

Now, point your web browser to a terminal with shell access to an environment where DeepDive is installed. You will find our examples included there as well.

```
$ cd deepdive-examples/spouse
```

You can also see a notebook version of the spouse example tutorial if you point your browser to the tutorial notebook.

Home | About | Contact | Disclaimer | Privacy | Notes

NIH NATIONAL CANCER INSTITUTE

NCI/CADD Group

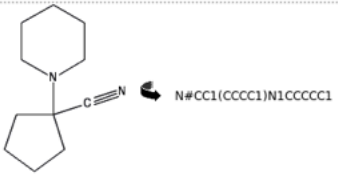
OSRA: Optical Structure Recognition Application

News | Dependencies | Compilation | Usage | License | Download | Web Interface | Validation | Author

Description

OSRA is a utility designed to convert graphical representations of chemical structures, as they appear in journal articles, patent documents, textbooks, trade magazines etc., into SMILES (Simplified Molecular Input Line Entry Specification - see http://en.wikipedia.org/wiki/SMILES) or SD files - a computer recognizable molecular structure format. OSRA can read a document in any of the over 90 graphical formats parseable by ImageMagick - including GIF, JPEG, PNG, TIFF, PDF, PS etc., and generate the SMILES or SDF representation of the molecular structure images encountered within that document.

Note that any software designed for optical recognition is unlikely to be perfect, and the output produced might, and probably will, contain errors, so curation by a human knowledgeable in chemical structures is highly recommended.



Autonomous Strategies - Deepdive

- DeepDive is a general natural language processing software
- Nice tutorials online
- Python based
- Has backend data to download

PMC-OA (PubMed Central Open Access Subset)

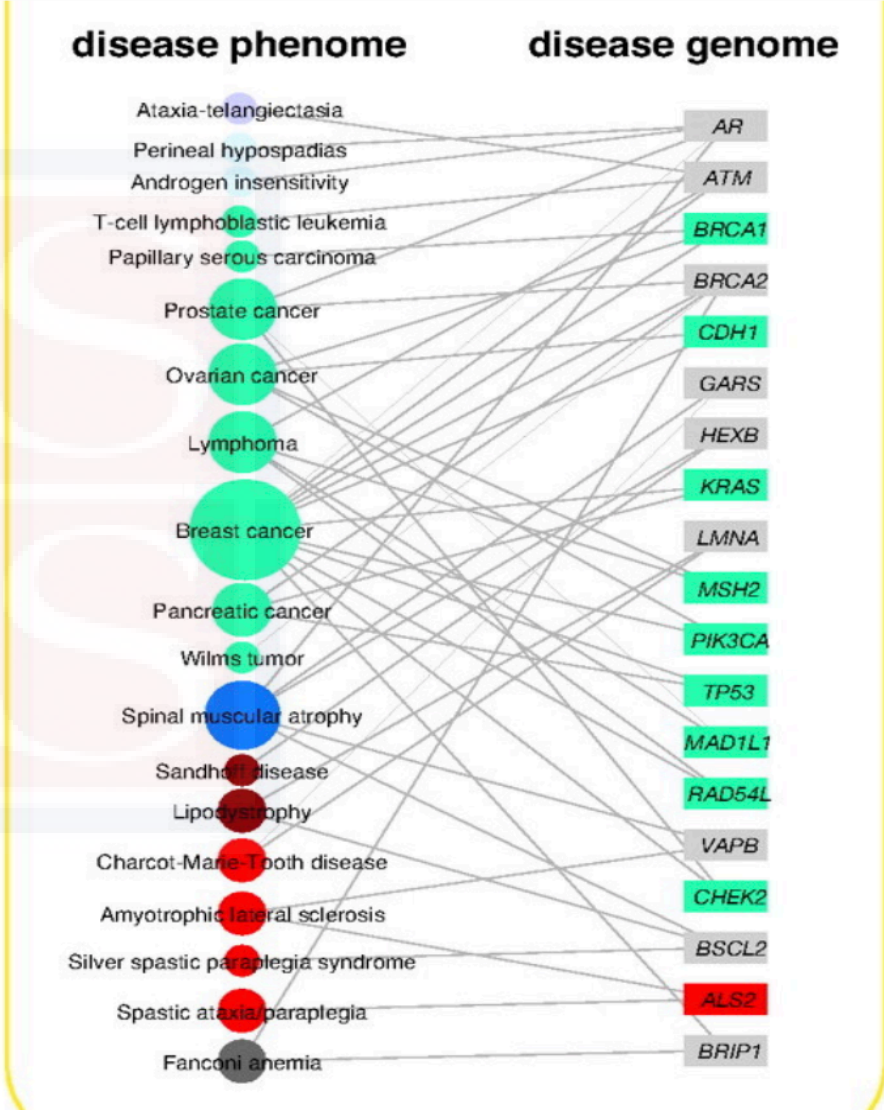
Quick Statistics & Downloads

Pipeline	HTML > STRIP (html2text) > NLP (Stanford CoreNLP 1.3.4)		
Size	70 GB	**Document Type**	Journal Articles
# Documents	359,324	**# Machine Hours**	100 K
# Words	2.7 Billion	**# Sentences**	110 Million
Downloads	Download Full Corpus	Download Small Teaser	

PATENT (Google Patents)

Quick Statistics & Downloads

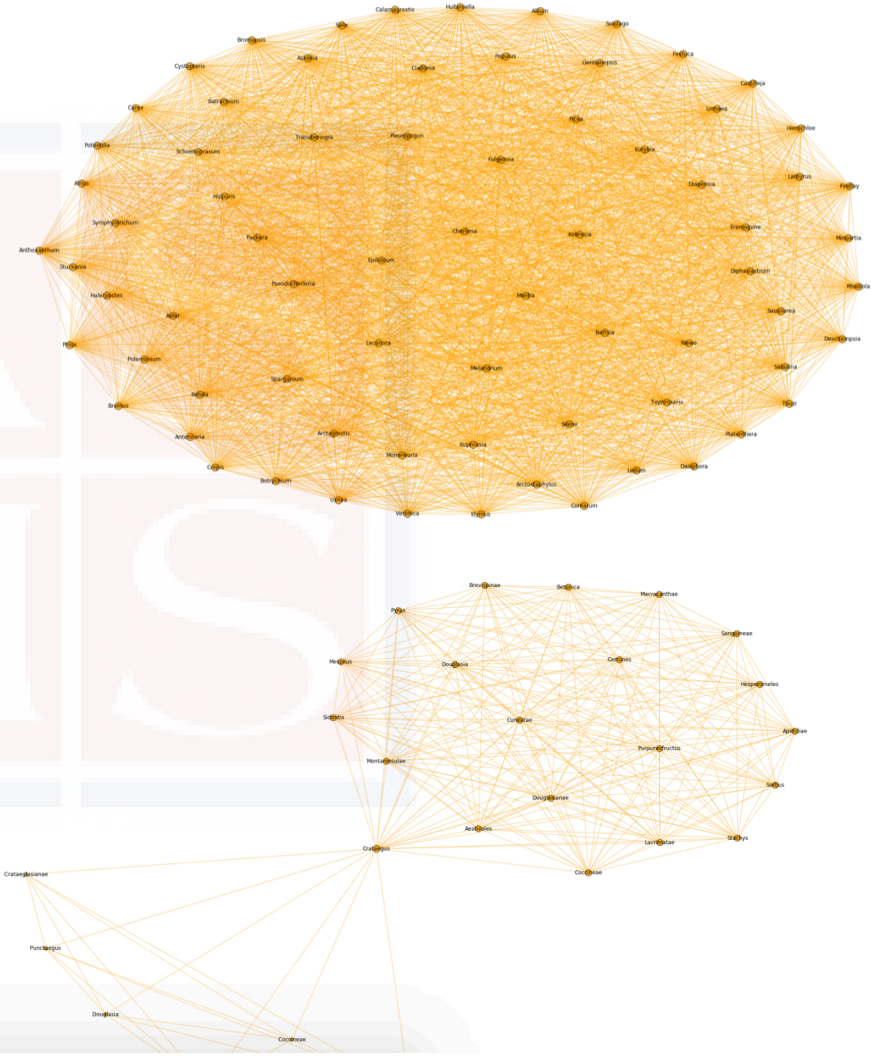
Pipeline	OCR'ed Text > NLP (Stanford CoreNLP 3.5.1)		
Size	428 GB	**Document Type**	Government Document
# Documents	2,437,000	**# Machine Hours**	100 K
# Sentences	248 Million	**# Words**	7.7 Billion
Downloads	Download Full Corpus	Download Small Teaser	



Autonomous Strategies - ContentMine

- Pulls directly fror literature
  - Licensing ?

- Python based
- Many tutorials



Thank you and Questions?

- Automation is coming and getting better for metabolomics
  - Lets do more

- Cognitive Natural language processing is getting better and is a quick way of understanding and dealing with large data.

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

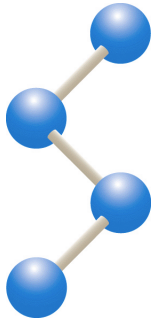
---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics



- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

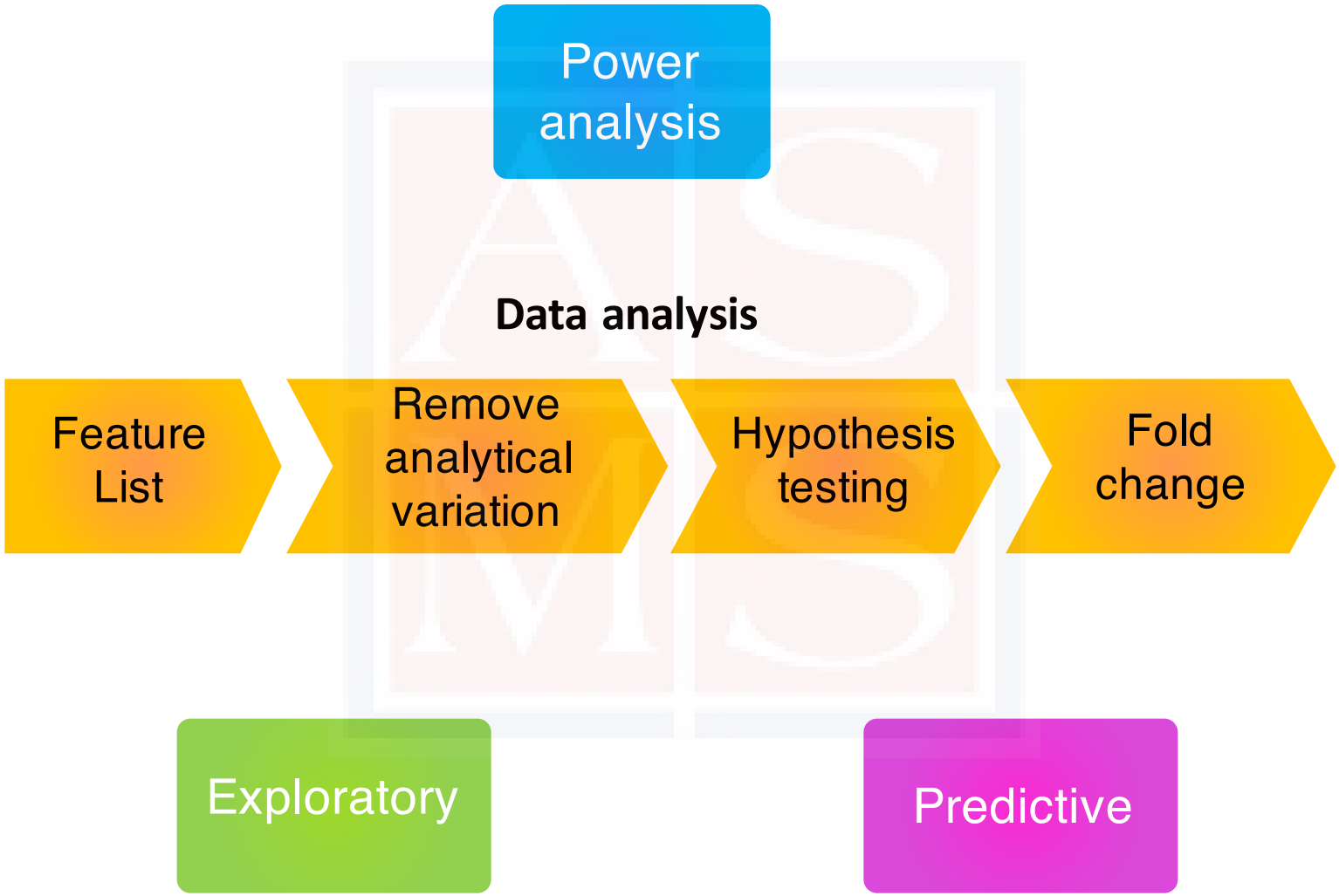
---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Contents

- **Do's and don'ts of statistics**
  - Power analysis
  - Analytical variation
  - Multiple testing correction
  - Parametric vs non-parametric
- **Multivariate Methods and Machine Learning**
  - PCA
  - Machine learning in a nutshell

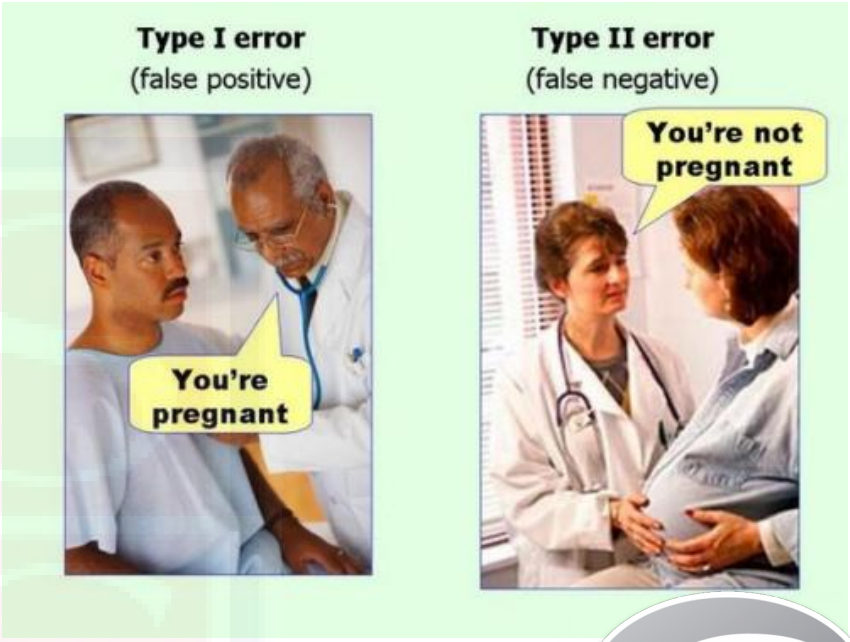
The statistics workflow



Power analysis: how many samples?

Power depends upon:

- Effect size -> $(\mu_0 - \mu_1)/SD$
- Type of experiment/hypothesis
- Sample size
- Error types




```
> power.t.test(delta=1, sd=0.5, sig.level=0.05,
power=0.8)

Two-sample t test power calculation
      n = 5.090008
```

1-β

α

Power analysis: how many samples?

Question

Sridhar Radhakrishnan
26.85 · Research Diets, Inc.

How can I perform power calculations for untargeted metabolomics experiment, which is a hypothesis generating experiment?

There is not such power analysis technique to calculate the power in advanced in untargeted metabolomics[1]

Some guidelines…

[1] Vinaixa, M. et al. Metabolites, 2 (2012) 775-795

Power analysis: how many samples?

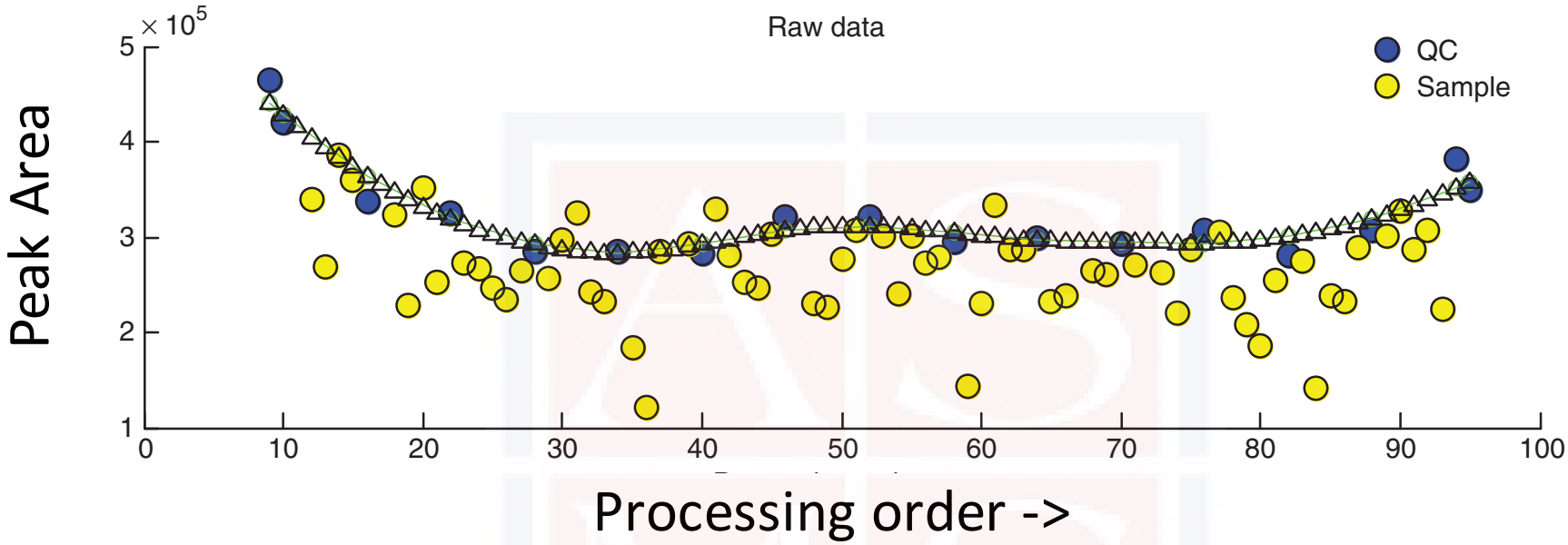
1) Our untargeted experiment is a pilot study (hypothesis generating) and we are going to validate it (QqQ)[1]

2) 20 samples rule of thumb[2-5]

3) Consider a ‘custom’ effect size[6]. (Delta=1 and SD=0.5), (D=1 and SD=1), (D=1 and SD=2) …

[1] Vinaixa, M. *et al*. Metabolites, 2 (2012) 775-795
[2] B. J. Blaise *et al*. Anal. Chem. 88 (2016) 5179-5188
[3] Lenth, R. V. Am. Stat. 55 (2001), 187–193.
[4] Hajian-Tilaki, K. Casp. J. Int. Med. 2 (2011), 289–298.
[5] Wong, M. Y.; Day, N. E.; Wareham, N. J. Statist. Med. 18 (1999), 2831–2845.
[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New Jersey: Lawrence Erlbaum.

Analytical variation



0) Always randomize samples!

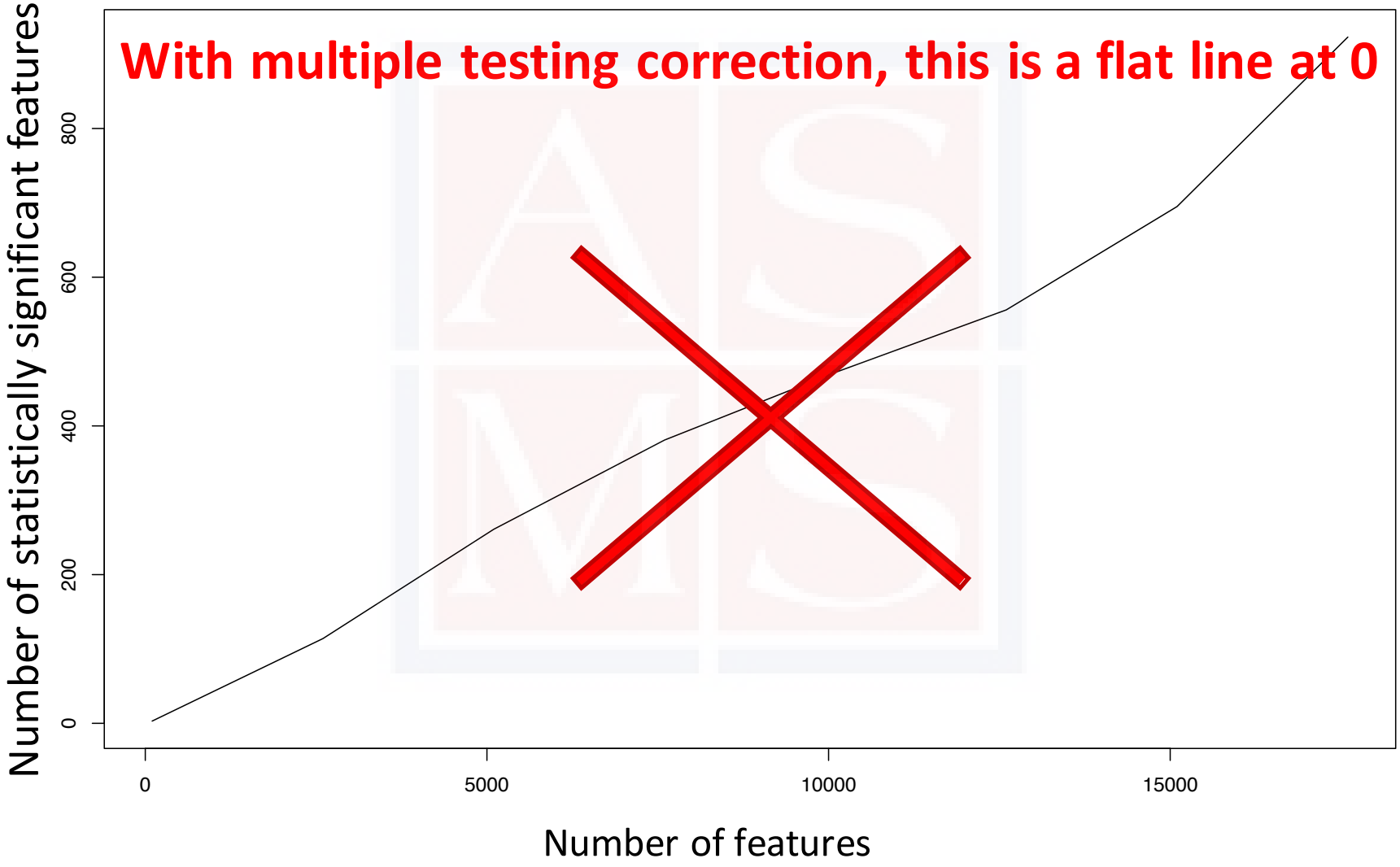
1) Remove features detected in less than 50% of QC sample[1]

2) Determine the **CV** for each feature **within QC class**.

3) Remove features with **CV > 20% within QC class**[2].

[1] W. D. Dunn et al., Nat. Protoc., 30 (2011) 1060-1083 (Figure taken from)
[2] Vinaixa, M. *et al.* Metabolites, 2 (2012) 775-795

Hypothesis testing



Hypothesis testing

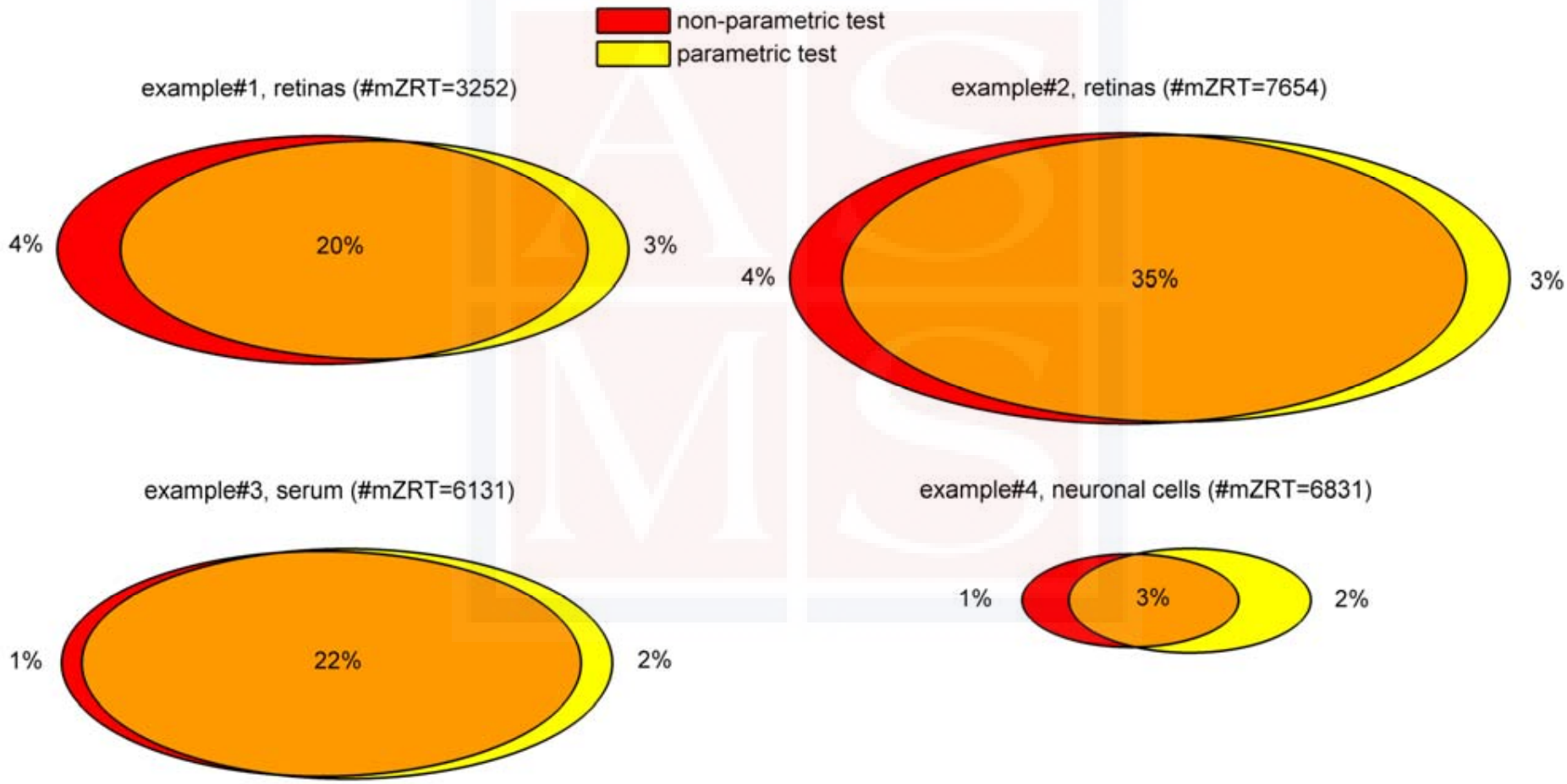
Different multiple testing methods:

- Holm, Hochberg, Hommel…
- Bonferroni
- False Discovery Rate (FDR) (q-values)

```
> p.adjust(c(0.002,0.89,0.03,0.0002,0.76,0.05,0.89),
method='bonferroni')

[1] 0.0140 1.0000 0.2100 0.0014 1.0000 0.3500 1.0000
```

Hypothesis testing

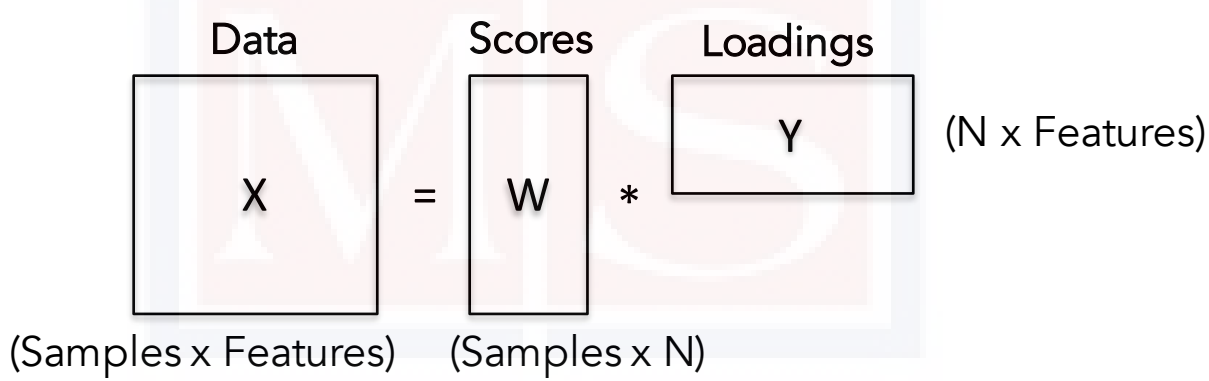
Parametric or non-parametric?



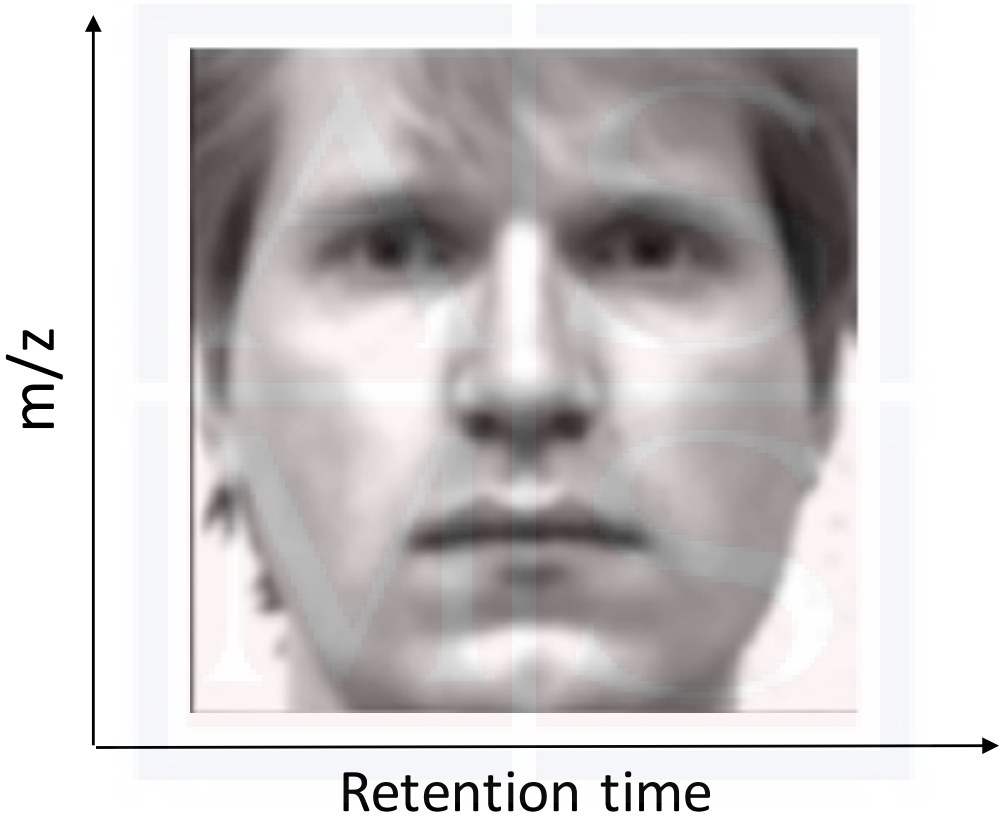
Picture from Vinaixa, M. et al., Metabolites, 2 (2012) 775-795

Multivariate analysis: PCA

- One of the most used methods in metabolic profiling
- Dimensionality reduction
- Groups data into sets (principal components) of correlated variables
- Principal components are uncorrelated



Multivariate analysis: PCA



Multivariate analysis: PCA

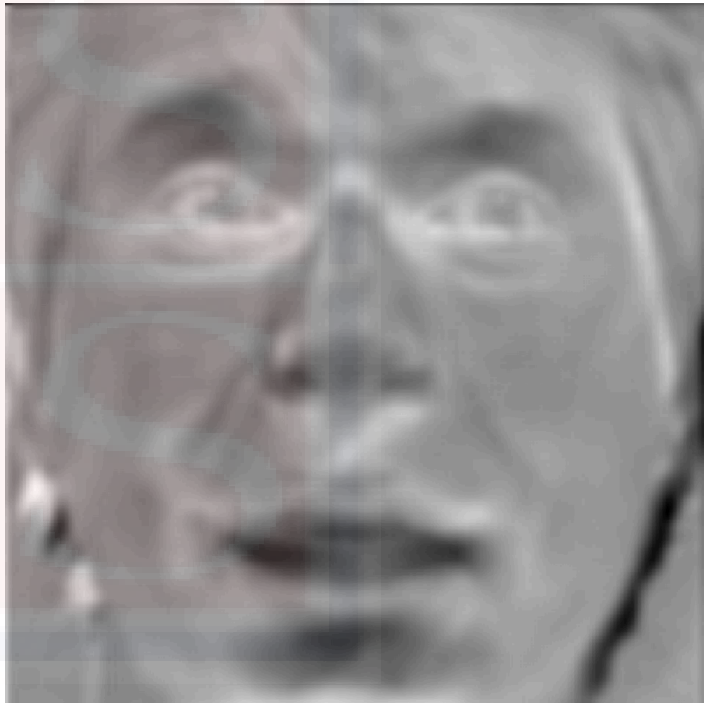
An example...

Multivariate analysis: PCA

Loadings

PC1

PC2



Multivariate analysis: PCA

Loadings

PC3

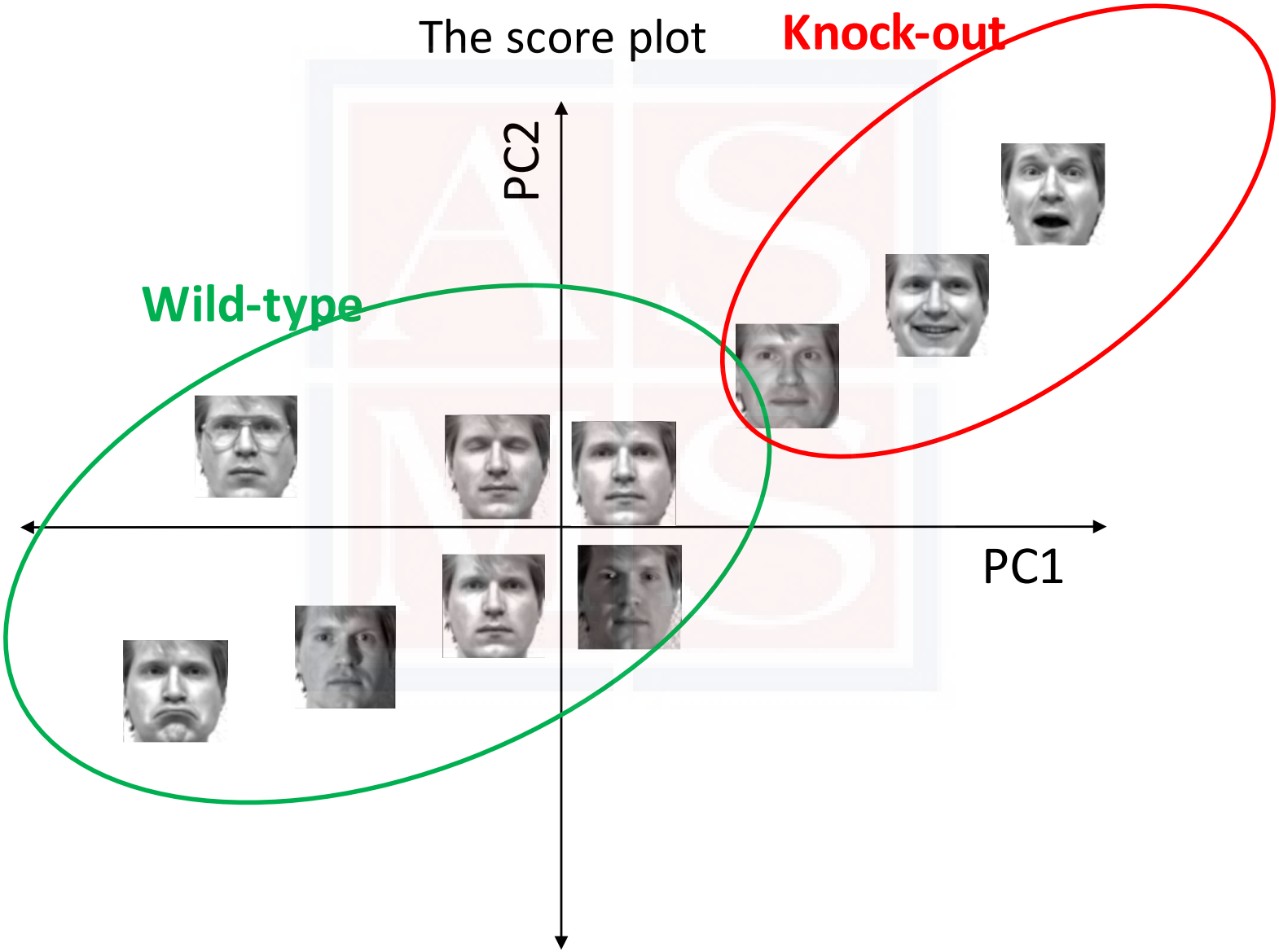
PC4

Multivariate analysis: PCA

Wild-type
(Sad)

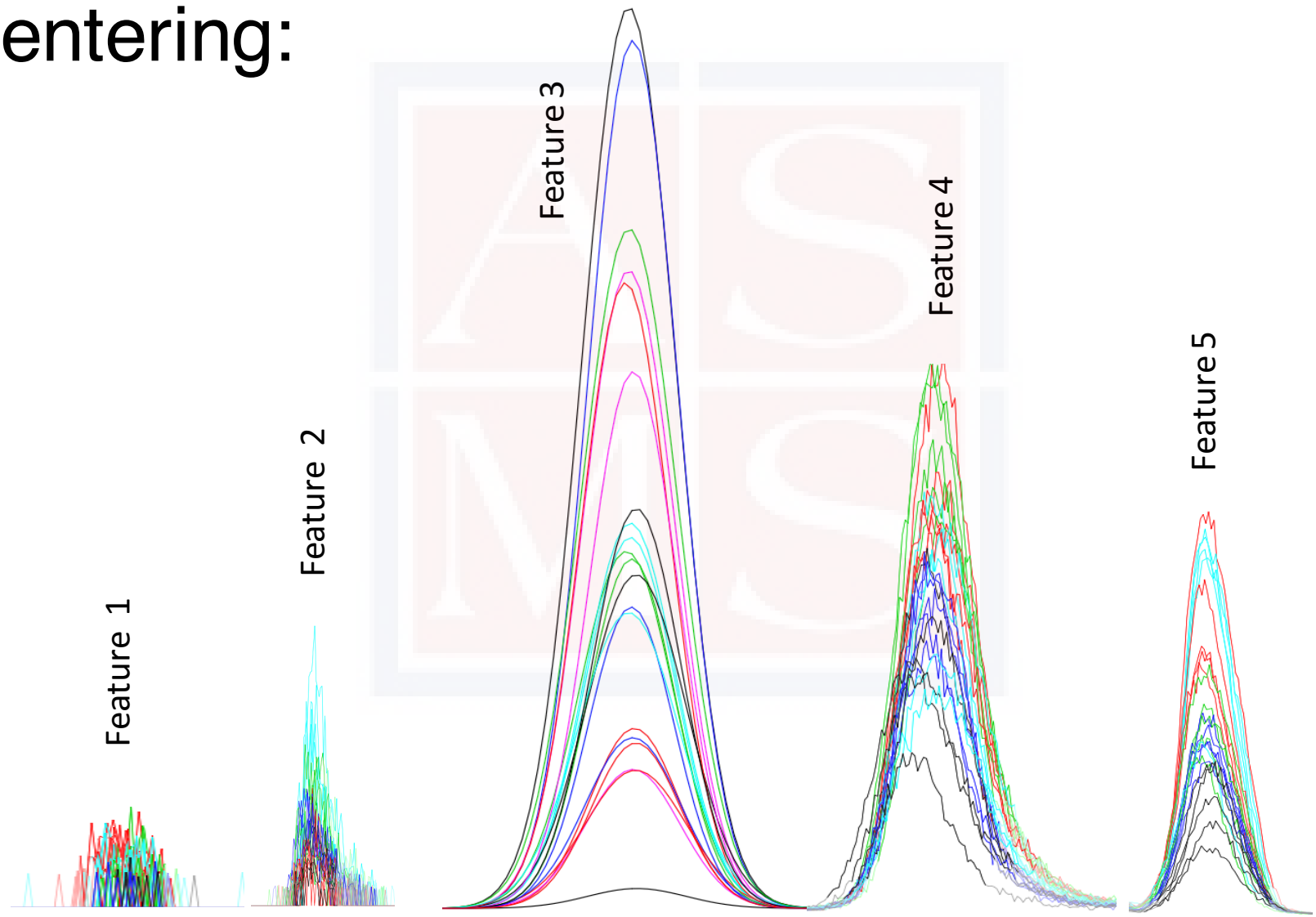
Knock-out
(Happy)

Multivariate analysis: PCA



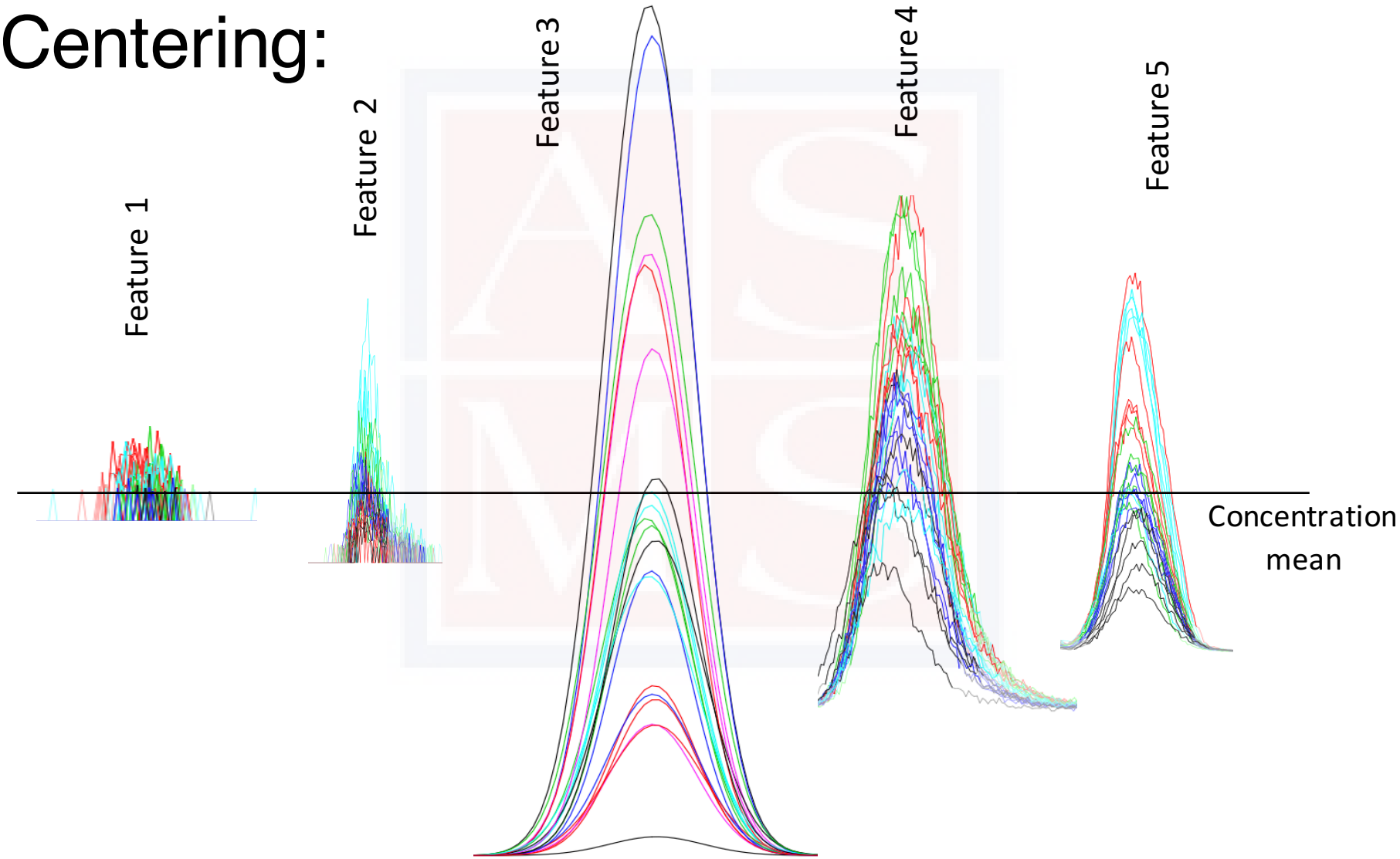
Multivariate analysis: PCA

Centering:

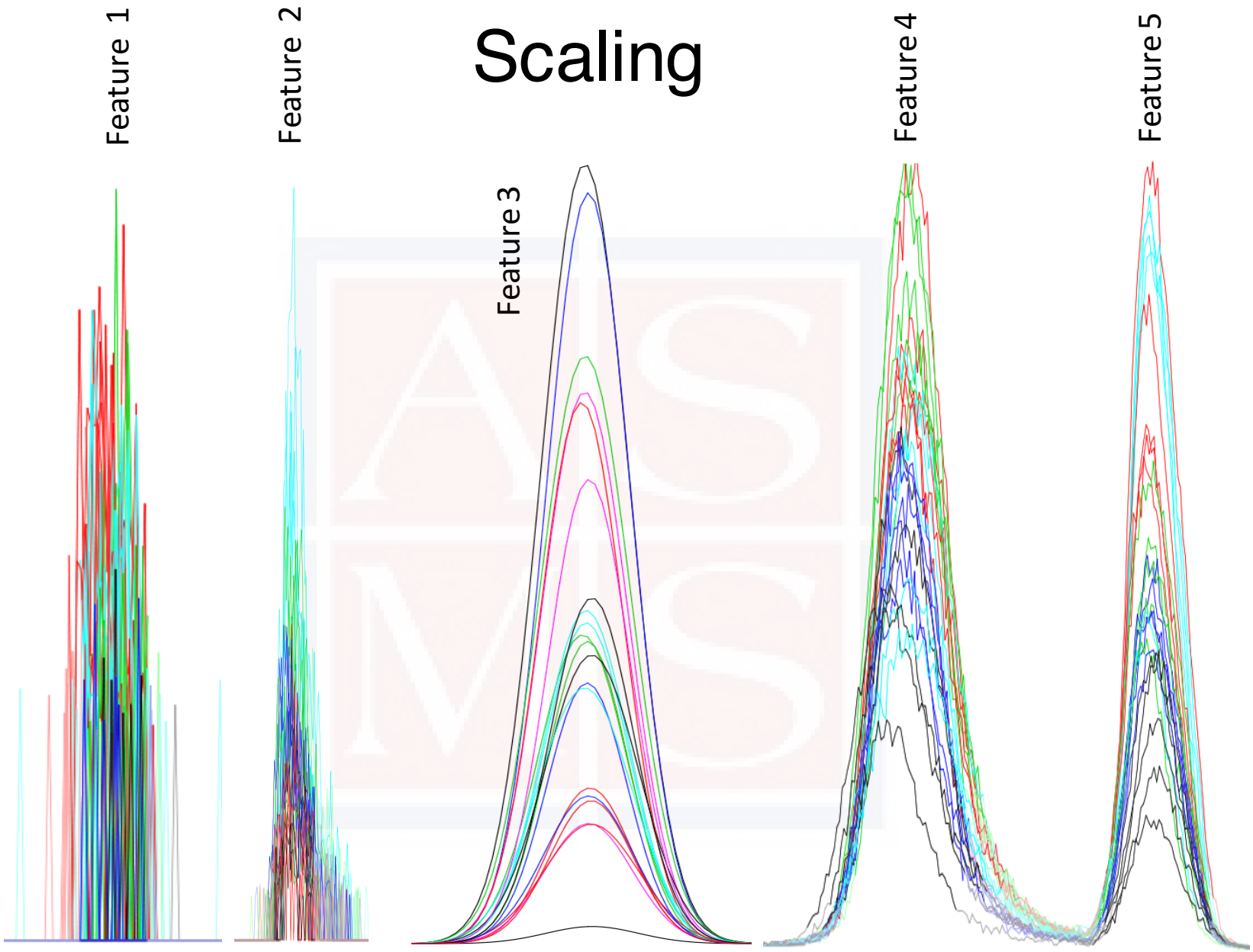


Multivariate analysis: PCA

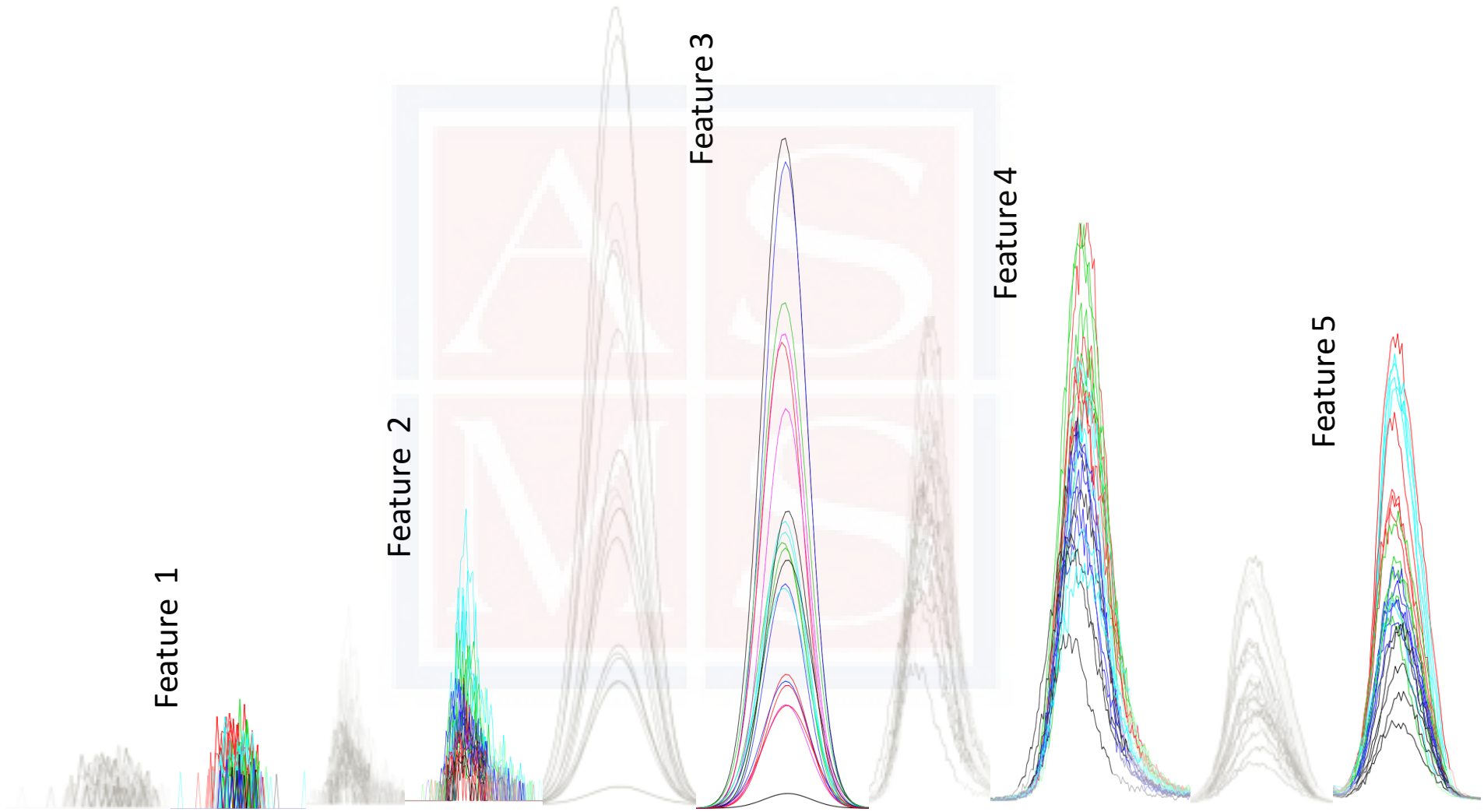
Centering:



Scaling



Scaling: pareto-scale



[1] van den Berg, R. A. *et al*. BMC Genomics. 7 (2006) 142

Multivariate analysis: PCA

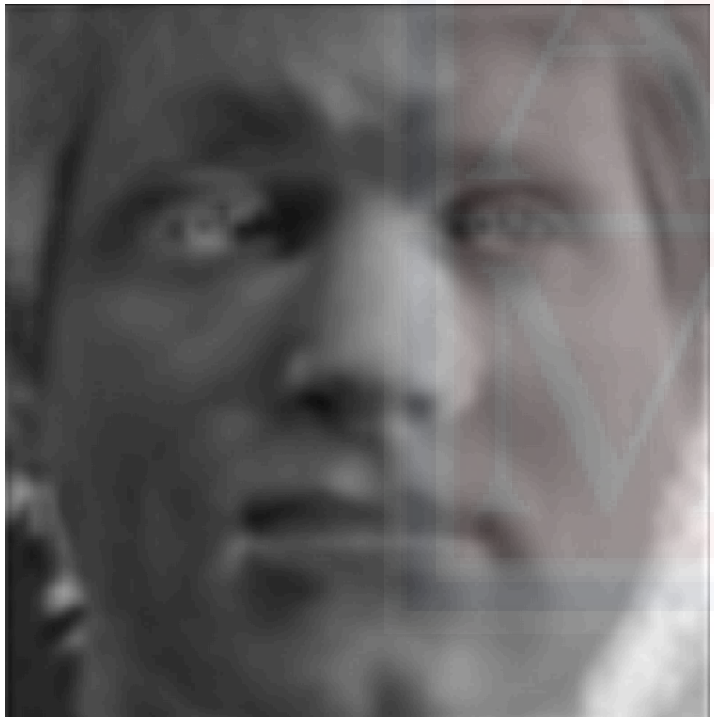
PC1

PC2

Multivariate analysis: PCA

Misinterpretation of PCA

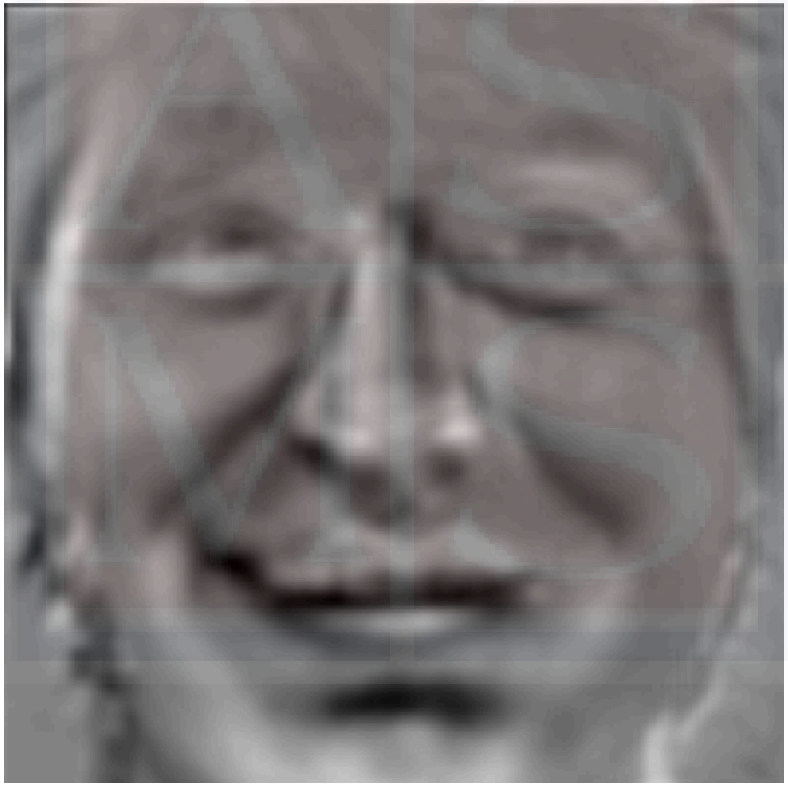
PC1



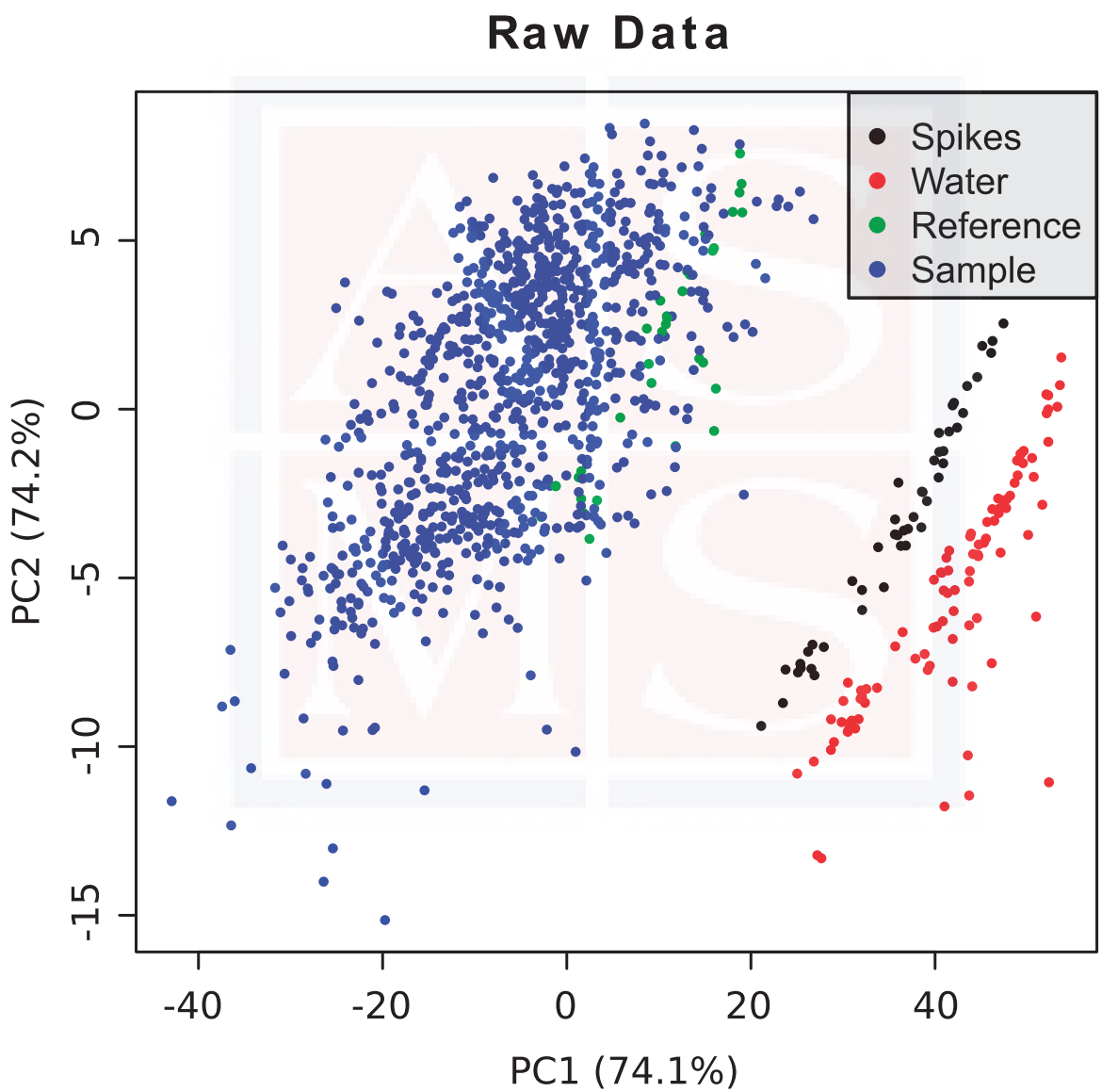
PC2

Multivariate analysis: PCA

PC5!!!



Multivariate analysis: PCA



Picture from Fernández-Albert et al., Bioinformatics, 30 (2014) 2899-2905

Alternatives to PCA

PCA is powerful, but it is exploratory, not predictive

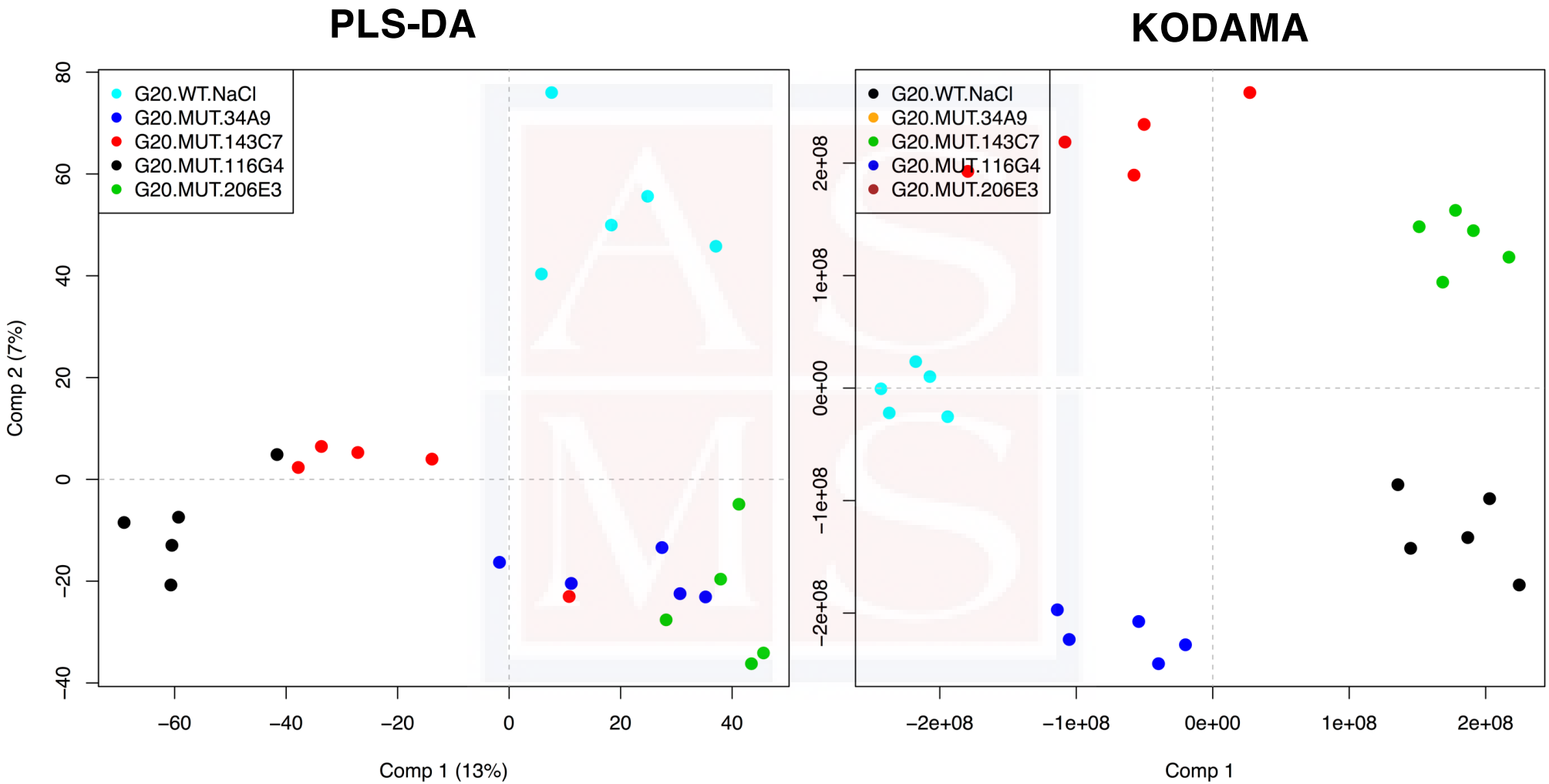
Exploratory multivariate methods:

- Data reduction
- Pulls out and prioritizes what features play the most important role in our phenotype
- Detects important or analytical drifts
- Allow revealing signatures rather than just statistically significant disregulated metabolites (p-values)

Alternatives to PCA:

- Linear Discriminant Analysis (LDA)
- Partial Least Squares (PLS) and PLS Discriminant Analysis (PLS-DA)
- **Machine learning**
  - **Knowledge discovery by accuracy maximization (KODAMA)**
  - **KNN**
  - **Random Forest**

KODAMA



Cacciatore S, *et al*. Knowledge discovery by accuracy maximization. Proc Natl Acad Sci. 111 (2014) 5117-22.
Cacciatore S, *et al*. KODAMA: an R package for knowledge discovery and data mining. Bioinformatics 33 (2016) 621-623.

KODAMA

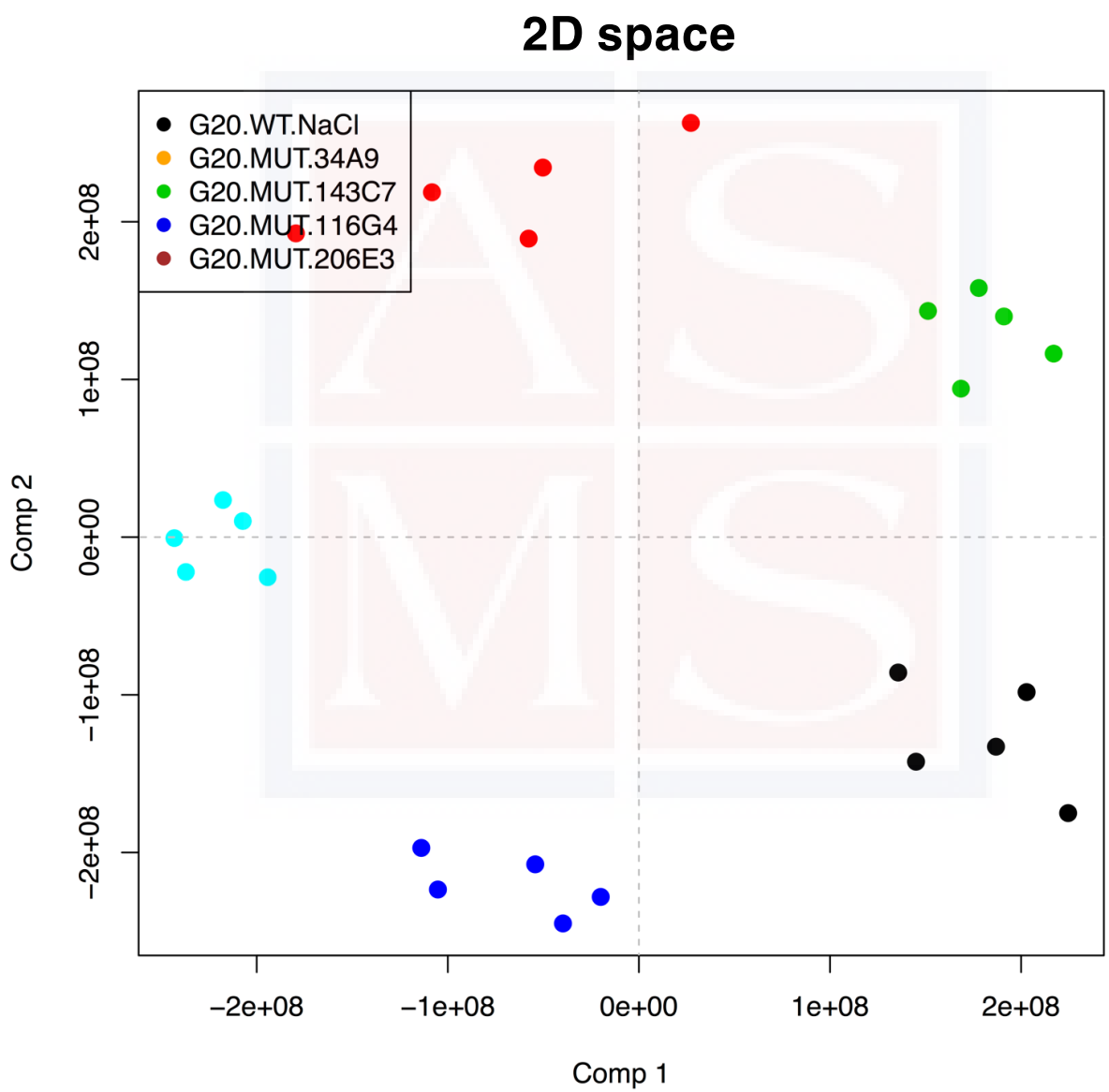
```
> library(openxlsx)
> metdRaw <- read.xlsx('XCMS.diffreport.MultiClass.xlsx')
> metd <- t(metdRaw[,40:64])
```

```
> rownames(metd)
[1] "G20.WT.NaCl.r001" "G20.WT.NaCl.r002" "G20.WT.NaCl.r003"
"G20.WT.NaCl.r004" "G20.WT.NaCl.r005" "G20.MUT.34A9.r001"
"G20.MUT.34A9.r002" …

> metClass <- sapply(rownames(metd), function(x)
paste(strsplit(x, '\\.')[[1]][-4], collapse='.'),
USE.NAMES=FALSE)
> metCol <- as.numeric(as.factor(metClass))
```

KODAMA

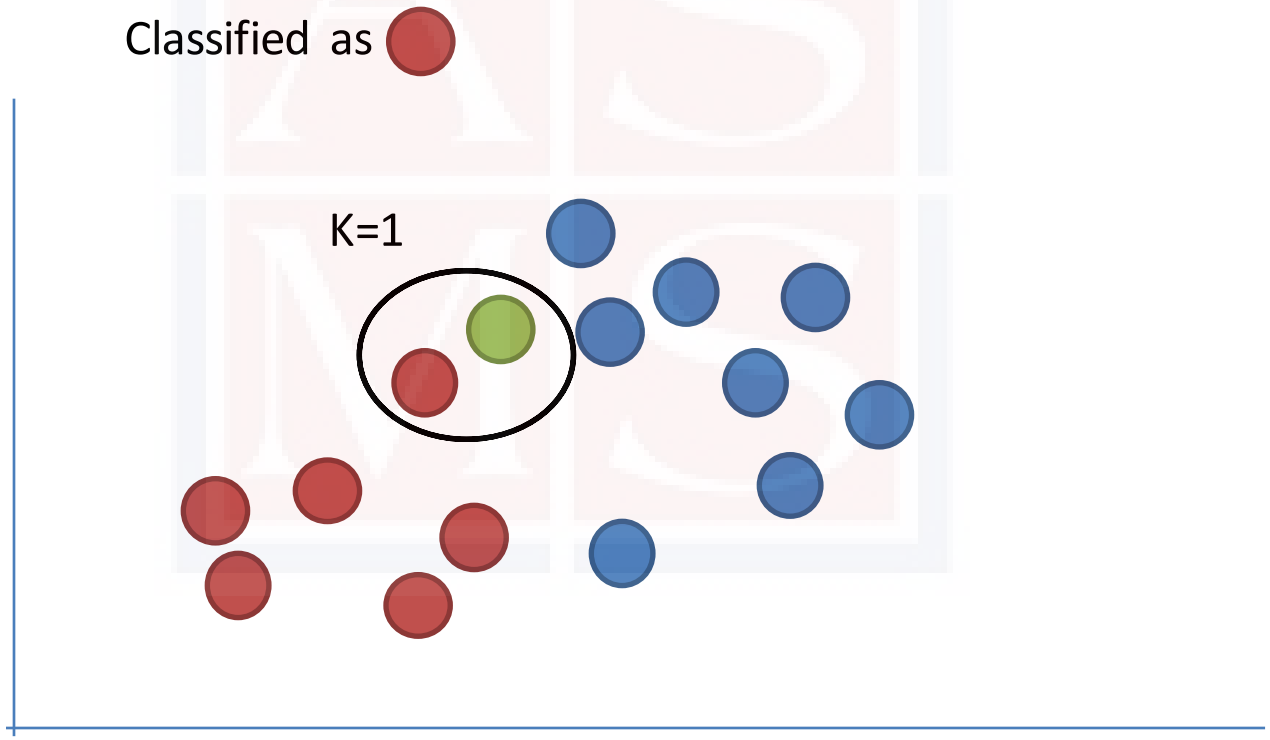
```
> library(KODAMA)
> kod.out <- KODAMA(metd, constrain=metClass)
> plot(kod.out$pp, col=metCol, pch=19, xlab='Comp 1',
ylab='Comp 2’)

> library(mixOmics)
> pls.out <- plsda(metd, Y=metClass, ncomp=4)
```

Classification



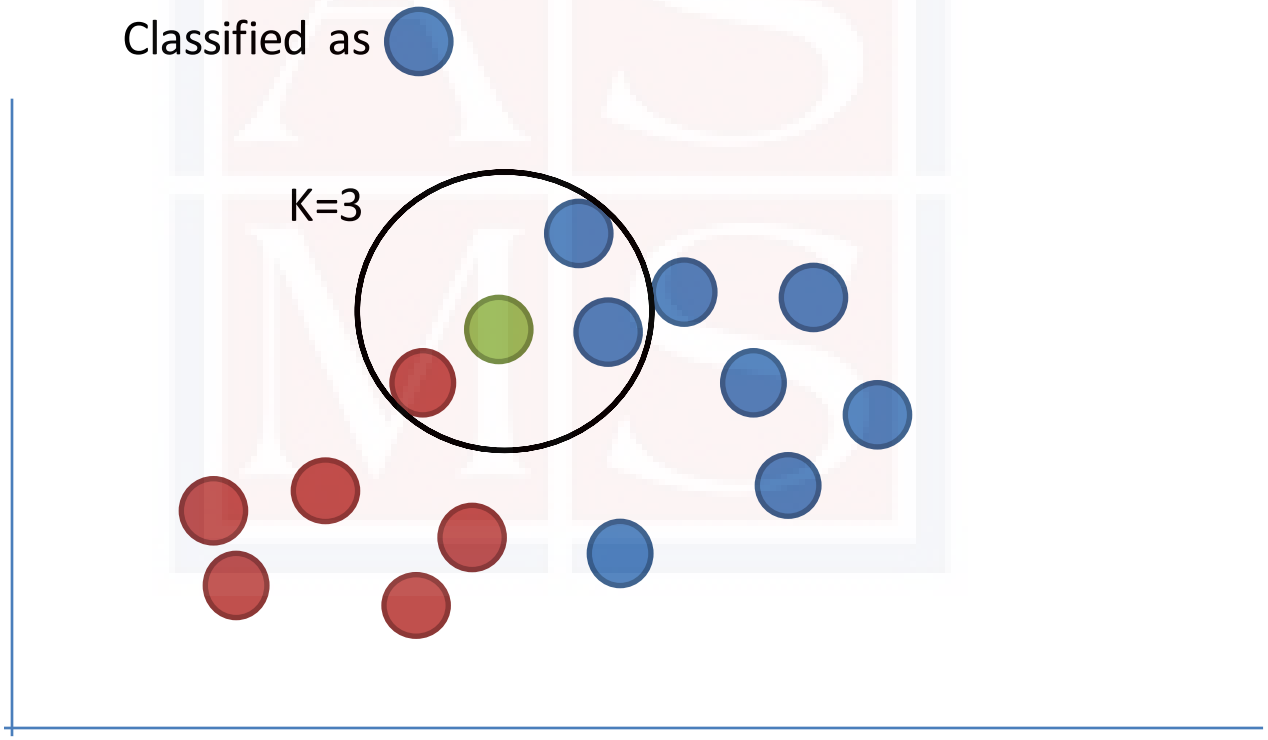
KNN: *k*-nearest neighbors algorithm

KNN is a classification method that takes into account the closest neighbors to classify a new data observation



KNN: *k*-nearest neighbors algorithm

KNN is a classification method that takes into account the closest neighbors to classify a new data observation



KNN: *k*-nearest neighbors algorithm

KNN is a classification method that takes into account the closest neighbors to classify a new data observation

Advantages:

- KNN's decision boundary is highly flexible

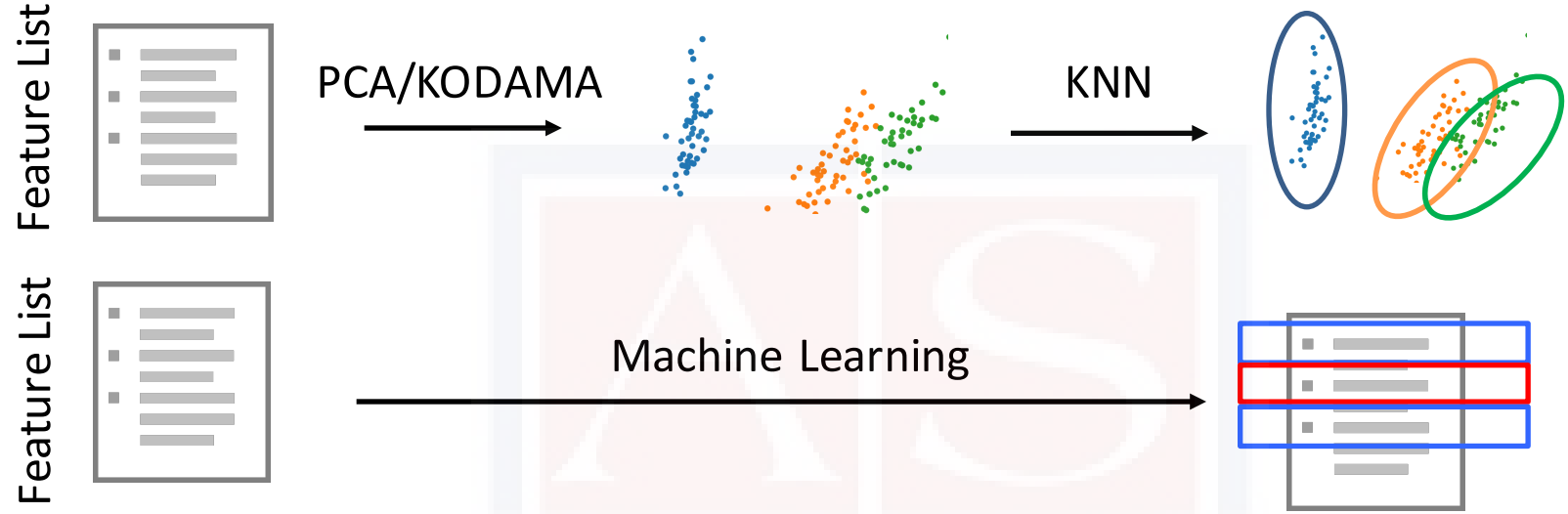
Drawbacks:

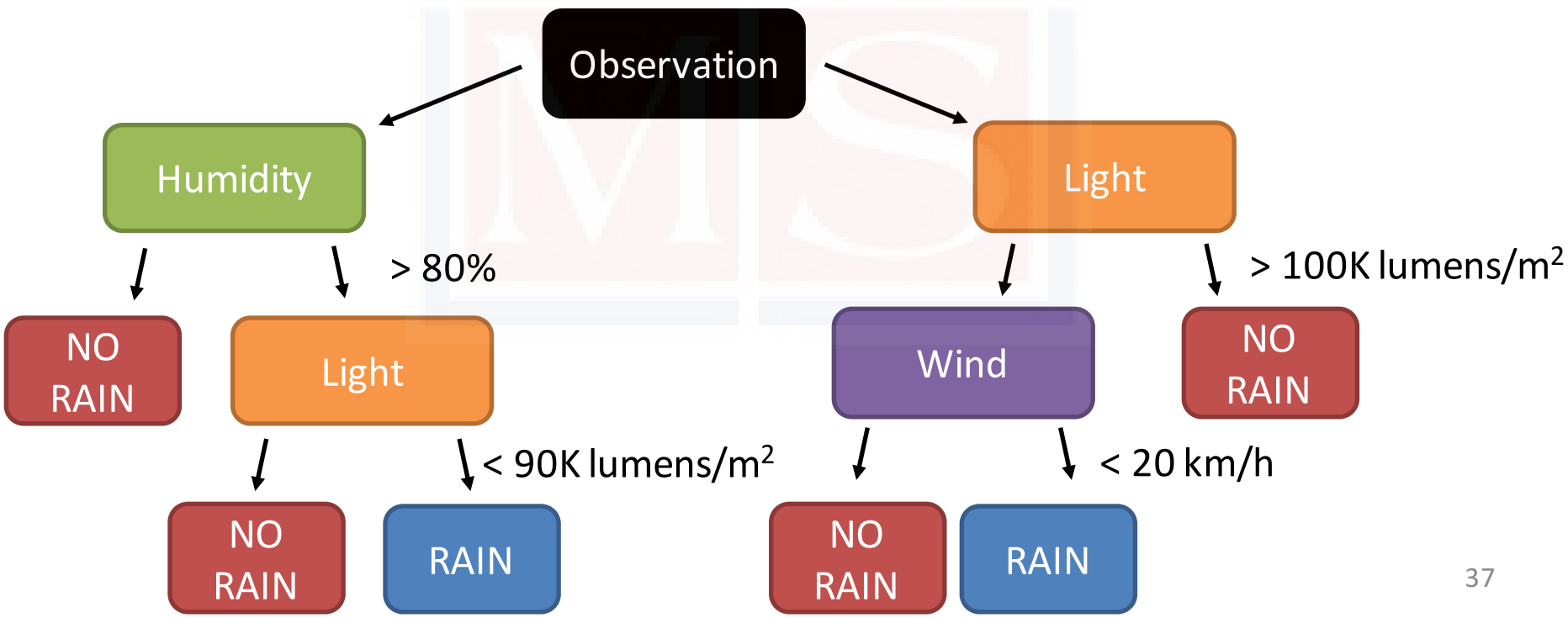
- Slow
- KNN gives the same importance to all the variables (best performance over already reduced data)
- Need to estimate k (overfitting)

KNN: *k*-nearest neighbors algorithm

```
> library(class)
> modKNN <- knn(train=kod.out$pp[-c(1,10),],
test=kod.out$pp[c(1,10),], cl=metClass[-c(1,10)], k=1)

> table(metClass[c(1,10),], modKNN)

                 G20.MUT.34A9 G20.WT.NaCl
G20.MUT.34A9            1           0
G20.WT.NaCl             0           1
```

Random Forest





Random Forest

An ensemble approach that uses decision rules to predict a specific class

Advantages:

- Runs efficiently on large data bases
- Out of bag (OOB) estimates can be used for model validation
- Decorrelates trees (good for metabolomics)

Drawbacks:

- The more the number of trees, the more slow
- Bad predictions outside the 'learning' limits when used for regression

Random Forest

```
> library(randomForest)
> metdf <- cbind(as.data.frame(metd), metClass)
> colnames(metdf)[-ncol(metdf)] <- paste('V',
colnames(metdf)[-ncol(metdf)], sep='')
```

```
> gam1 <- randomForest(metClass~., data=metdf, ntree=50)
> gam1

          G20.MUT.116G4 G20.MUT.143C7 G20.MUT.206E3 G20.MUT.34A9 G20.WT.NaCl class.error
G20.MUT.116G4     5             0             0             0           0         0.0
G20.MUT.143C7     0             3             0             1           1         0.4
G20.MUT.206E3     0             1             3             1           0         0.4
G20.MUT.34A9      0             0             2             3           0         0.4
G20.WT.NaCl       0             0             0             0           5         0.0
```

ROC curve

Receiver operating characteristic (ROC) curves are used to see how well your classifier can separate positive and negative examples (specially when comparing two classifiers) and to identify the best threshold for separating them.

1) Score > 95 F

TP: 50 FP: 50
TN:0 FN:0

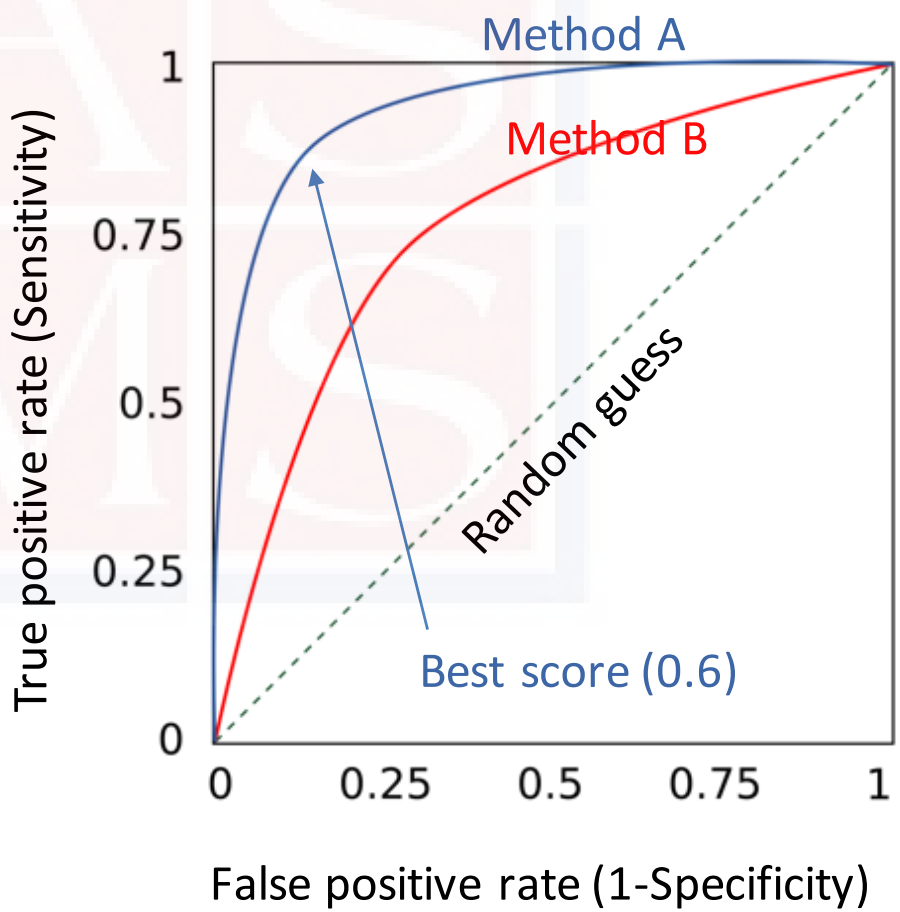
2) Score > 100 F

TP: 20 FP: 6
TN:25 FN:4

3) Score > 105F

… …

FPR = FP/(FP+TN)
TPR=TP/(FN+TP)



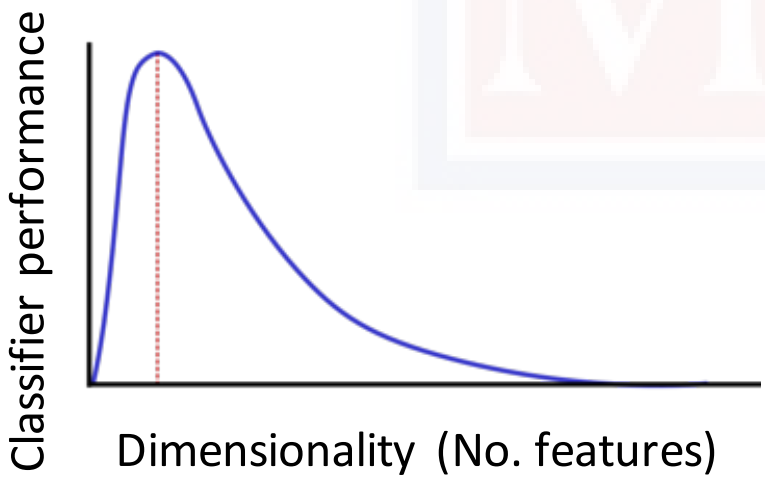
Overfitting

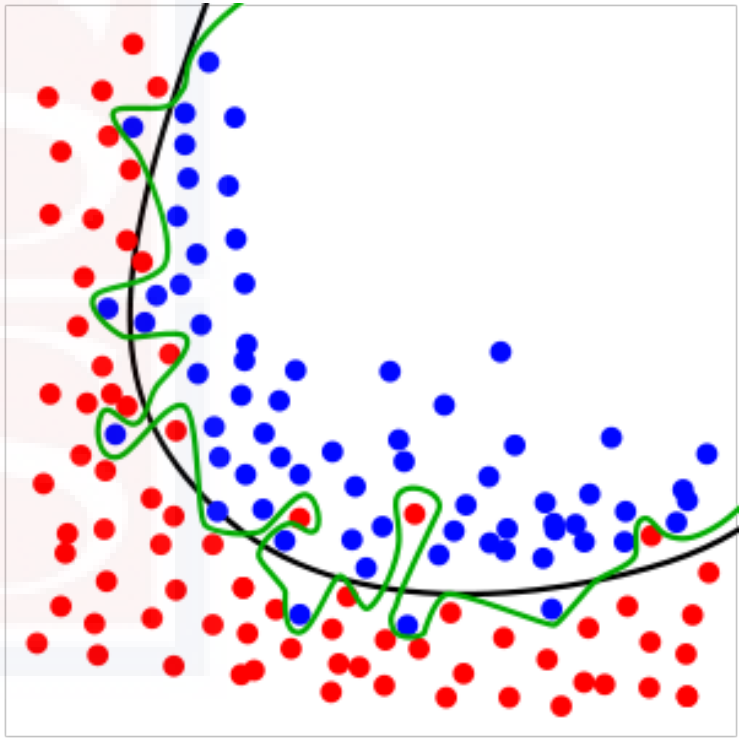
1) Use an exclusive test set for validation

Training set	Test set

Both training set and test set should have the same performance

2) Cross-validation
3) Use enough data examples
4) Remove variables/features (curse of dimensionality!)





Garbage in – garbage out principle!

Thank you for your attention!

Questions?

Advanced Metabolomics

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

---- 02:15 pm Break ----

Advanced Metabolomics

- *Primary Experimental and Informatic Challenges*
- *Key Algorithms in Creating Reproducible Data*
- *Computational Metabolite Data Annotation*
- *Pathway Analysis & Multi-Omic Integration*
- *Identifying Metabolites from Scratch*
- *Statistics in Design & Interpretation*
- *Activity Metabolomics*

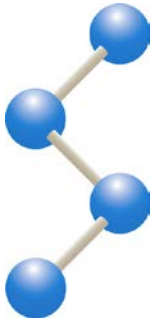
June 3rd

---- 09:00 am Begin ----

---- 10:15 am Break ----

---- 12:00 pm Lunch ---

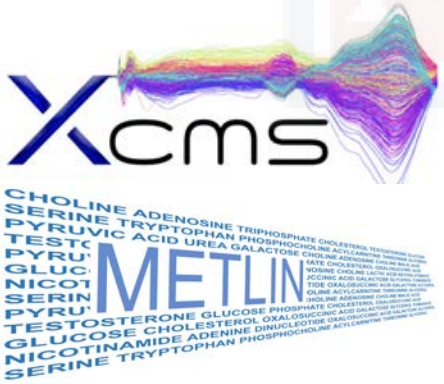
---- 02:15 pm Break ----

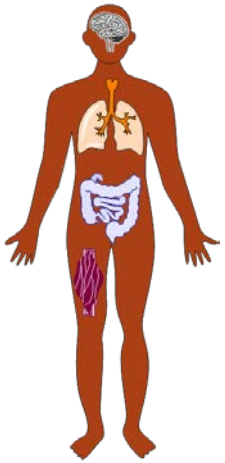


Advanced Metabolomics

May 16th-17th

INTEGRATING METABOLOMICS INTO BIOLOGY





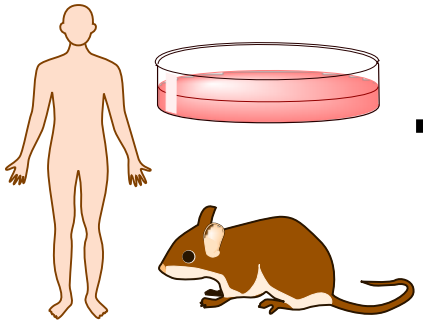
CARLOS GUIJAS

- **BIOLOGICAL MODEL IS PIVOTAL TO HANDLE OUR METABOLOMICS DATA AT ALL ANALYTICAL LEVELS.**

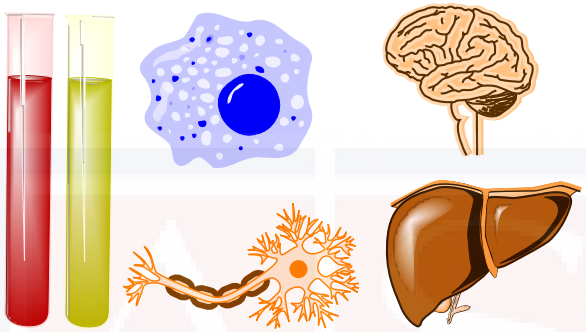
- **USING METABOLOMICS OUTPUT IN FUNCTIONAL ASSAYS: BEYOND THE DISCOVERY OF BIOMARKERS.**

The untargeted metabolomics workflow: an overview

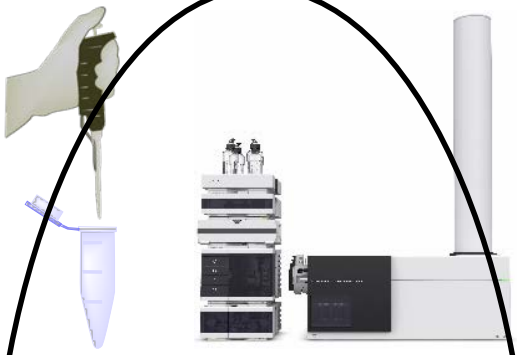
Biological model



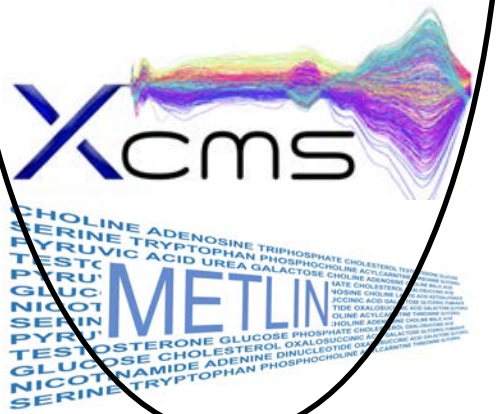
Biosource



Analytical process



Bioinformatics



List of dysregulated metabolites and pathways

- 7-ketocholesterol
- Tryptophan
- Citric acid
- Nicotine degradation II
- Citrulline biosynthesis
- Lactose degradation III

- Sample prep in metabolomics involves several decision-making steps dependent on the biological problem:

1. Selection of the biosource to be analyzed.

   a) Which biosource is the most suitable to find differences in the metabolome and provide as most as possible information?

   b) If we have a cell-based system, should we analyze supernatants in addition to cell extracts?

   c) If we have a whole-organism system, is it worthwhile processing biofluids?

2. Selection of the type of extraction. Tightly related with previous selection.

3. Normalization of results: which parameters should we measure to use the same amount of starting material across all samples?

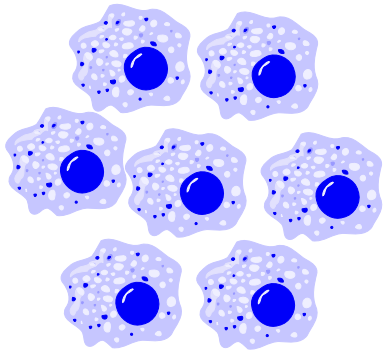
Example I. Do we analyze supernatants in cell-based models?

Model of study	Analysis of supernatants	Reason
Endocrine cells: pancreatic cells, enterocytes, hepatocytes, macrophages, adipocytes, ovary/testis cells		
Skeletal muscle cell		
Osteoclast		
Stem cell differentiation		

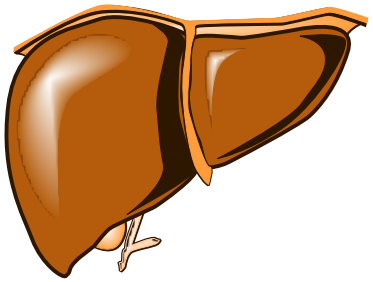
Example II. Do we analyze biofluids in whole-organism models?

Model of study	Analysis of plasma/urine	Reason
Endocrine organs/tissues: pancreas, liver, adipose tissue, ovaries/testis		
Localized diseases: some skin diseases, alopecia		
Gut microbiome modifications		
Kidney disease		
Endotoxic shock by intraperitoneal injection of LPS		

Sample prep. Normalization parameters



- Protein content
- Number of cells



- Tissue mass

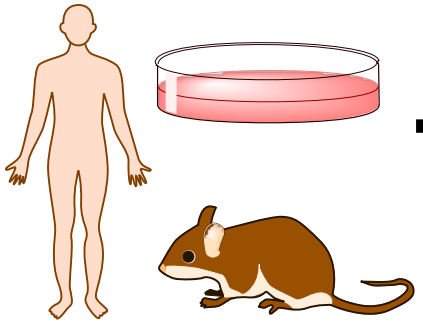
- Volume of biofluid

Example III. Choice of normalization magnitude

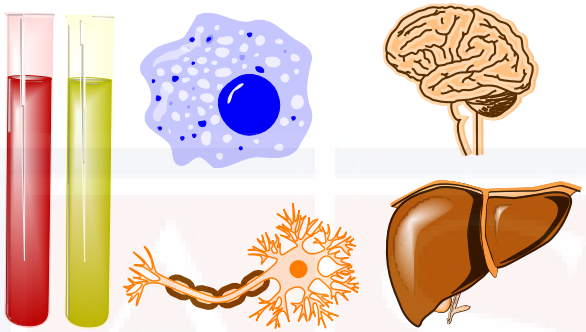
Source	Norma-lization	Pitfall	Alternative
Cells	Cell number		
Cells	Protein content		
Tissue	Mass		
Fecal matter	Mass		
Biofluids	Volume		
Urine	Volume		

The untargeted metabolomics workflow: an overview

Biological model

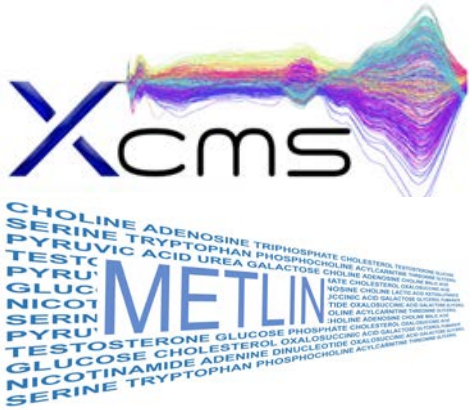


Biosource



Analytical process

Bioinformatics



List of dysregulated metabolites and pathways

- 7-ketocholesterol
- Tryptophan
- Citric acid
- Nicotine degradation II
- Citrulline biosynthesis
- Lactose degradation III

XCMS processing. Selection of species

General | Feature Detection | Retention Time Correction | Alignment | Statistics | Annotation | Identification | Visualization | Miscellaneous

Option	Value	Note:
ppm	2	tolerance for database search
adducts	[M+H]+ [M+NH4]+ [M+Na]+ [M+H-H2O]+ [M+H-2H2O]+ [M+K]+ [M+ACN+H]+ [M+ACN+Na]+ [M+2Na-H]+ [M+2H]2+	adducts to be considered for database search
sample biosource	SELECT BIOSOURCE set default SELECTED: default-HUMAN	Select your species/cell line, etc. that correspond to your samples. Default human.
pathway ppm deviation	5	metabolite pathway lookup

▸ View Advanced Options

Select	ID	Biosource	Strain
SELECT	YEAST	S.cerevisiae	
SELECT	DALA207559	D.alaskensis	G20
SELECT	HUMAN	H.sapiens	
SELECT	FLY	Drosophila melanogaster	
SELECT	ECOLI	E.coli K-12 substr. MG1655	

Showing 1 to 5 of 7,627 entries

Previous 1 2 3 4 5 … 1526 Next

XCMS processing. Selection of species

Pathway	Overlapping putative metabolites[1]	All metabolites[2*]	p-values
1D-*myo*-inositol hexakisphosphate biosynthesis II (mammalian)	2	2	1.7e-4
D-*myo*-inositol (3,4,5,6)-tetrakisphosphate biosynthesis	2	2	1.7e-4
\'\"inosine-5\'\'-phosphate biosynthesis II\"\'	2	2	1.7e-4
purine and pyrimidine metabolism	4	13	5.6e-4

***Mus musculus* (Correct biosource)**

Pathway	Overlapping putative metabolites[1]	All metabolites[2*]	p-values
1D-*myo*-inositol hexakisphosphate biosynthesis V (from Ins(1,3,4)P3)	2	2	4.2e-4
1D-*myo*-inositol hexakisphosphate biosynthesis II (mammalian)	2	2	4.2e-4
D-*myo*-inositol (3,4,5,6)-tetrakisphosphate biosynthesis	2	2	4.2e-4
D-*myo*-inositol (1,4,5,6)-tetrakisphosphate biosynthesis	2	2	4.2e-4

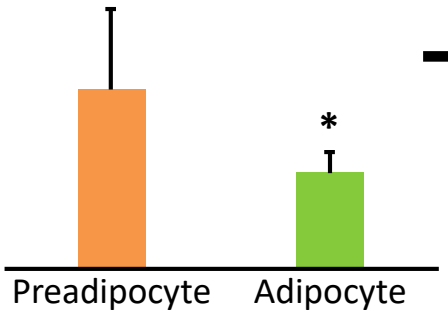
Homo sapiens

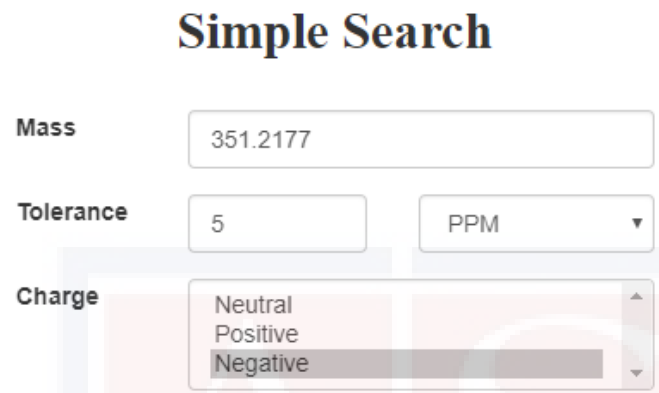
Pathway	Overlapping putative metabolites[1]	All metabolites[2*]	p-values
adenosylcobalamin salvage from cobinamide I	2	2	8.5e-6
guanine and guanosine salvage	2	3	4.9e-5
\'\"inosine-5\'\'-phosphate biosynthesis I\"\'	2	3	4.9e-5
guanosine nucleotides degradation III	2	4	2.1e-4

Escherichia coli

METLIN identification

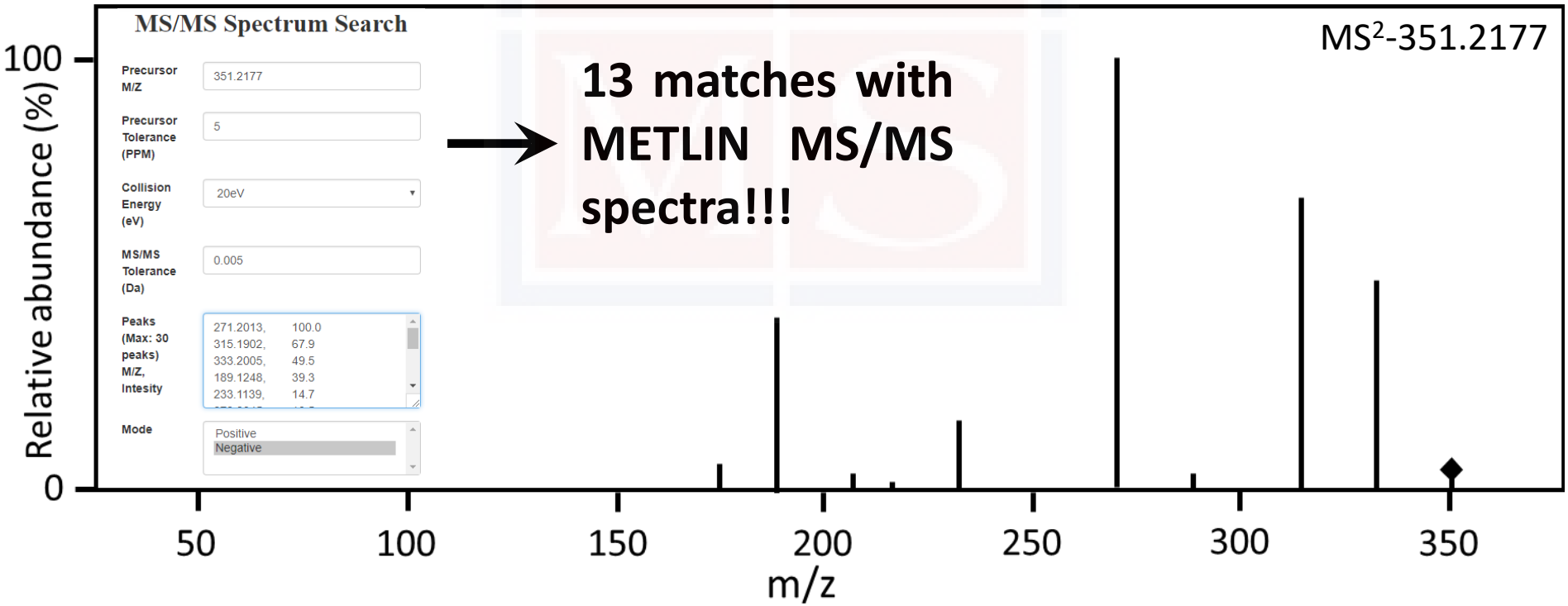
Unknown feature
m/z=351.2177





136 putative metabolites:

- PGE_2
- PGD_2
- PGI_2
- $PGF_{2\alpha}$ derivatives
- PGH_2
- Lipoxin A_4



13 matches with METLIN MS/MS spectra!!!

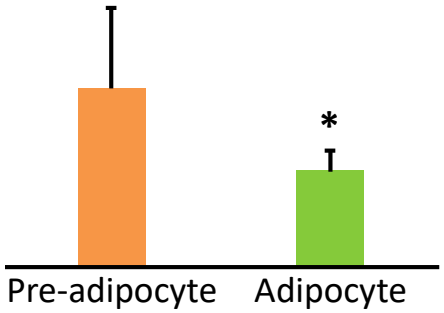
Example IV. Use biology to identify metabolites

- Since MS/MS spectra match with METLIN has multiple hits, the simplest way to identify this molecule is to compare its retention time with the retention time of authentic standards.

- Purchasing 13 standards is expensive→ Use of biological information to narrow down the candidates.

- Relevant biological information:

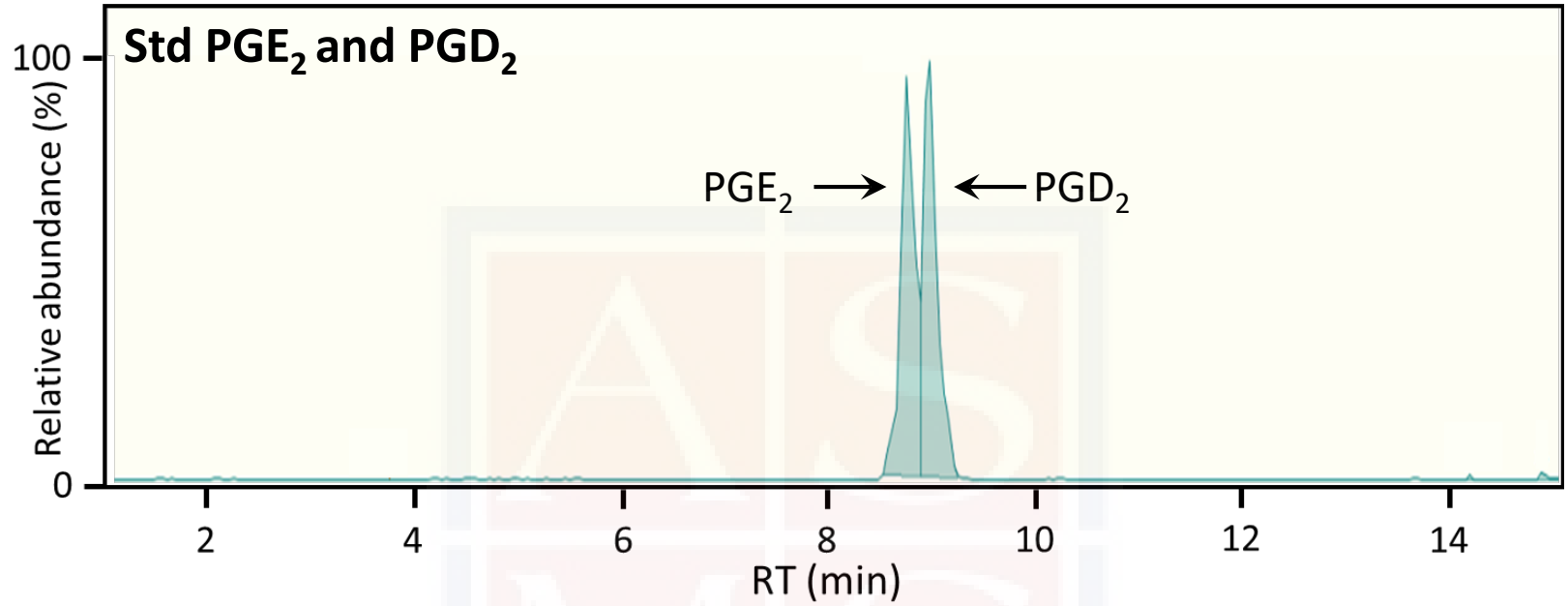
1. Pre-adipocytes in differentiation to adipocytes.
2. Molecule found in supernatants only, not in cell extracts.

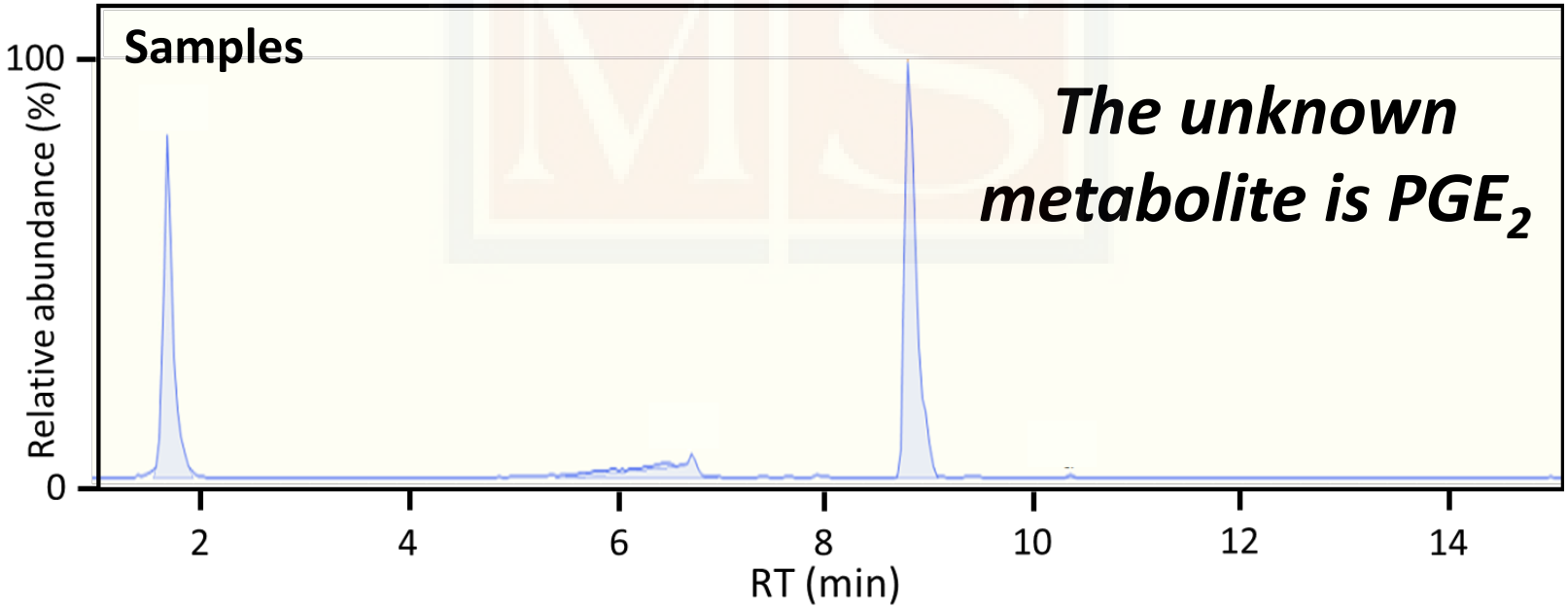


3. Bibliography. Search for the involvement of those 13 metabolites in adipocyte differentiation:

   - PGE_2 suppresses 3T3-L1 pre-adipocyte differentiation (*Tsuboi, et al., Biochem. Biophys. Res. Comm., 2004).* PGE_2 blocks pre-adipocyte differentiation into white adipocytes (*Garcia-Alonso, et al., J. Biol. Chem., 2013*).
   - PGE_2 is the major AA derivative produced in multiple cell types (*Dennis & Norris, Nat. Rev. Immunol, 2015*).
   - PGD_2 (PGE_2 isomer) shows a very similar MS/MS spectra and elutes very close to PGE_2 in reversed-phase chromatography (*Dumlao, et al., BBA, 2011*).

Example IV. Use biology to identify metabolites





The unknown metabolite is PGE_2

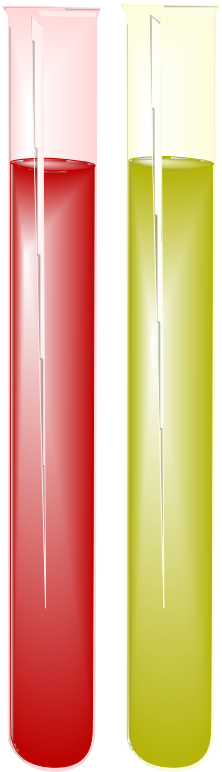
- **BIOLOGICAL MODEL IS PIVOTAL TO HANDLE OUR METABOLOMICS DATA AT ALL ANALYTICAL LEVELS.**

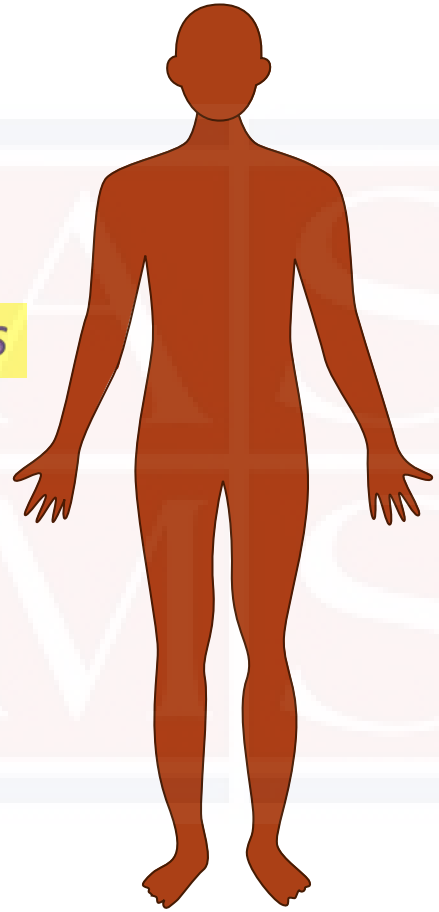
- **USING METABOLOMICS OUTPUT IN FUNCTIONAL ASSAYS: BEYOND THE DISCOVERY OF BIOMARKERS.**

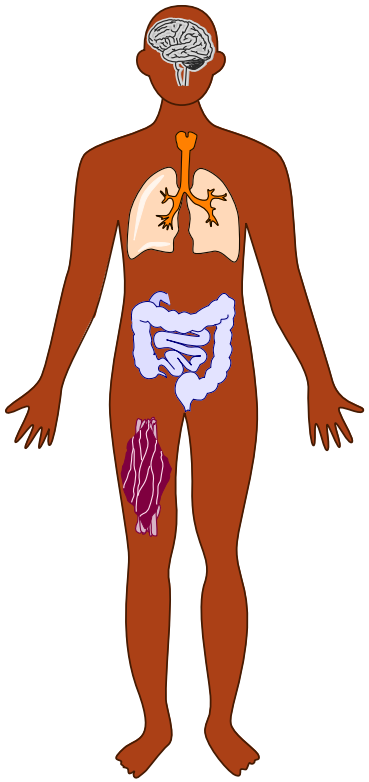
GENOMICS

TRANSCRIPTOMICS

PROTEOMICS







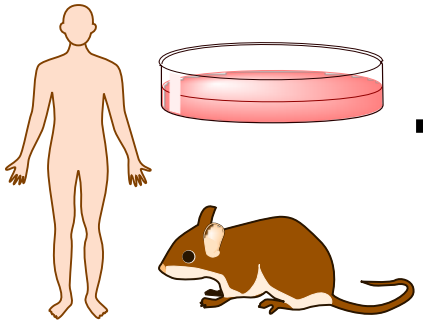
- Biomarkers
- Mechanism of disease

METABOLOMICS

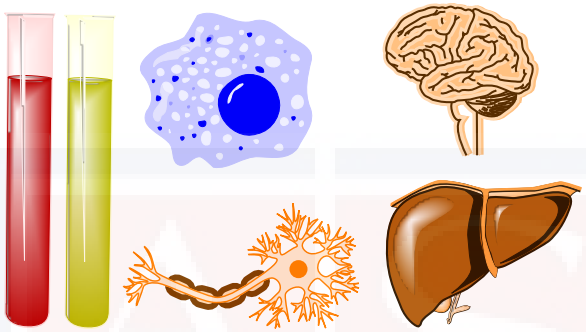
Discovery and prioritization of metabolites to be used as phenotype modulators.

The untargeted metabolomics workflow

Biological model

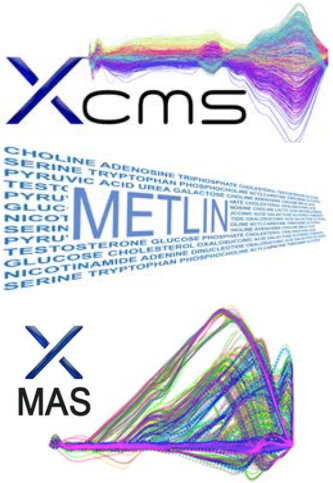


Biosource



Analytical process

Bioinformatics

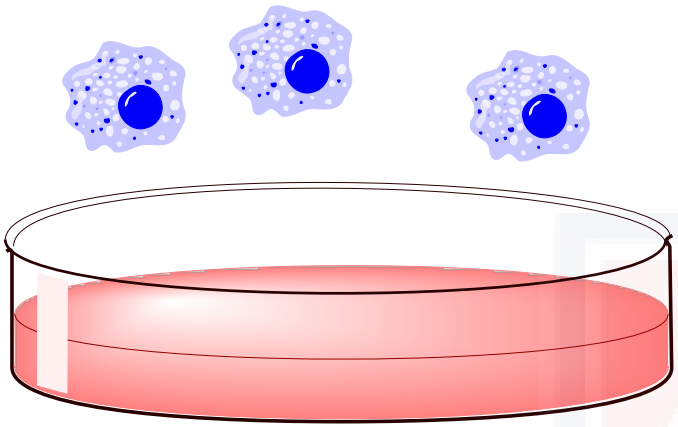


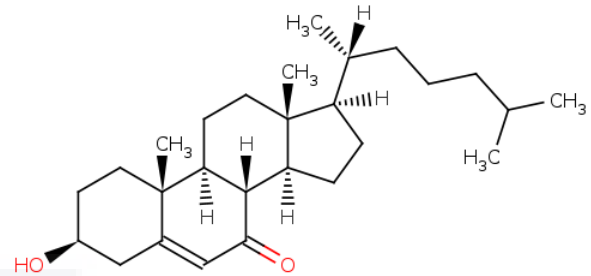
List of dysregulated metabolites and pathways

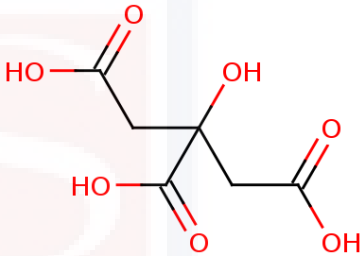
- 7-ketocholesterol
- Tryptophan
- Citric acid
- Nicotine degradation II
- Citrulline biosynthesis
- Lactose degradation III
- Taurine

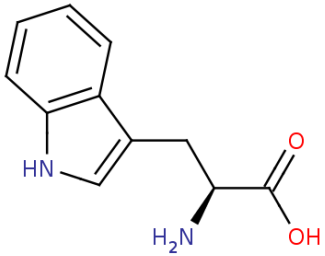
Cell-based functional assays

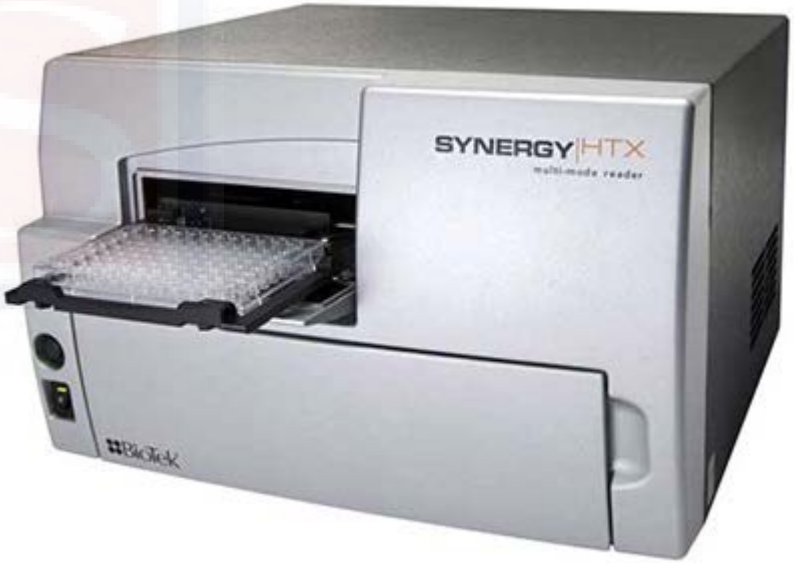
Metabolomics Activity Screening: cell-based functional assays









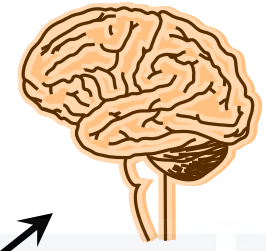


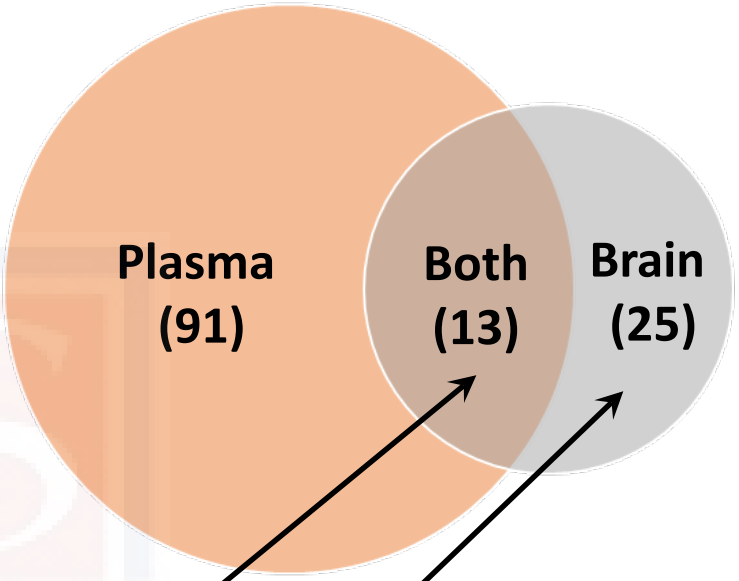
Metabolomics Activity Screening: cell-based functional assays

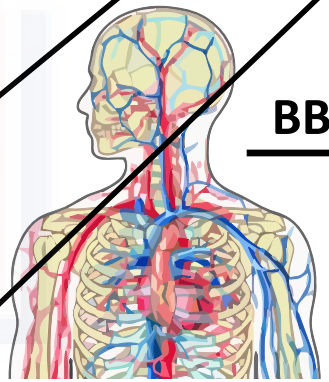
Functional assay	Description	Well-suited for
Cytotoxicity, proliferation and viability	Measurement of number of viable cells (trypan blue), metabolic activity (MTT) or DNA synthesis (BrdU).	Preliminary assays to assess toxicity of metabolites
Apoptosis	Measurement of membrane asymmetry (annexin V) or mitochondrial degradation (cytochrome C oxidase).	Tumor cells, but virtually all cell types
Glucose uptake	Measurement of uptake of fluorescent glucose analogs (2-NBDG).	Adipocytes, muscle
Modulation of the inflammatory response	Measurement of NF-κB activity (NF-κB reporter luciferase assay). Simultaneous detection of multiple cytokines.	Immune cells, endothelial cells, adipocytes, fibroblasts
GPCR signaling	Measurement of intracellular calcium (FLIPR).	Virtually all cell types
Autophagy	Measurement of autophagic vacuoles (monodansylcadaverine).	Cancer cells, degenerative diseases
ROS	Measurement of ROS through multiple probes (oxidized/reduced glutathione, catalase, superoxide anion).	Immune cells, tumor cells, neurons

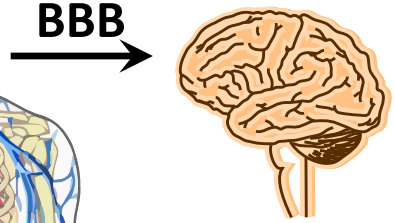
Example I. Ruling out metabolites

Model of disease: Calorie restriction in mice.







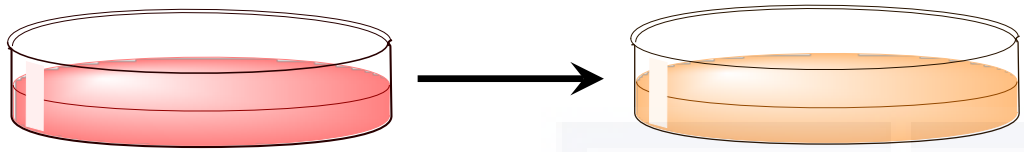


Objective: To **return back** to mice the **dysregulated metabolites** to study their **modulating effect** in caloric restriction**.** We cannot buy the 103 metabolites:

1. As food supplements.
2. Through intracerebral injection.

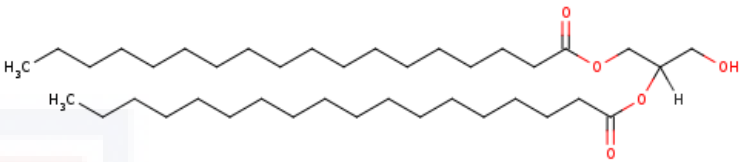
Example II. Ruling out metabolites

Model of disease: Pre-adipocyte differentiation.



Pre-adipocytes Adipocytes

Diacylglycerol molecules



Objective: To **add** dysregulated **metabolites** to the **pre-adipocytes** to **study** their role in **cell differentiation**.

1. Biology: DAG are **intracellular** signaling molecules, known to be activators of several PKC isoforms, which play roles in cell differentiation (Newton, JLR., 2009).

2. Biophysics: Due to its hydrophobicity (logP>14) and the absence of transporters, when it is added exogenously, DAG is accumulated between the two leaflets of plasma membrane.

3. It is not worthwhile to screen the effect of these metabolites, unless:

   a) Use a short-chain analogue, able to pass through the plasma membrane.

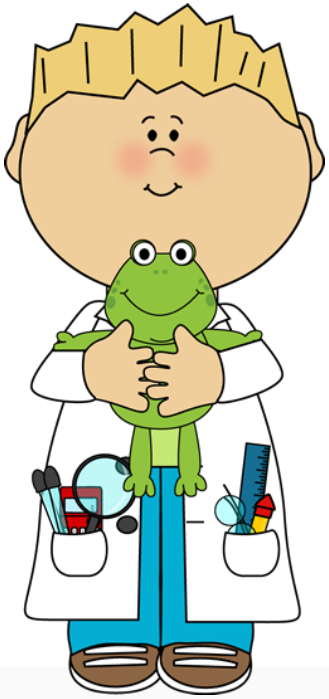
   b) Derivatize the compound to let it enter the cell with groups that are hydrolyzed once within the cell.

   c) Use carriers (mixed micelles).

CONCLUSIONS

- ✓ **The nature of the biological system we are analyzing is essential for the design of our metabolomics workflow and the use of our output data.**

1. **Sample prep.**
2. **Data processing.**
3. **Metabolite identification.**
4. **Functional assays.**



He put the frog on the ground and told it to jump. The frog jumped.

So the scientist cut off one of the frog's legs. The scientist told the frog to jump, the frog jumped.

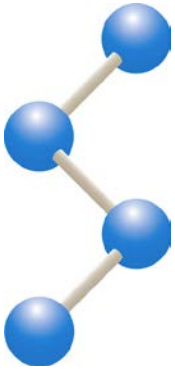
The scientist cut off another leg. He told the frog to jump. Frog jumped again.

The scientist cut off one more leg. He told the frog to jump. Frog jumped again.

So the scientist cut off his last leg.

He told the frog to jump, but the frog didn't. He tried again, but nothing.

So the scientist wrote in his notebook, "Frog with no feet, goes deaf."



Advanced Metabolomics

May 17th

- ***Primary Experimental and Informatic Challenges***
- ***Key Algorithms in Creating Reproducible Data***
- ***Computational Metabolite Data Annotation***
- ***Pathway Analysis & Multi-Omic Integration***
- ***Identifying Metabolites from Scratch***
- ***Statistics in Design & Interpretation***
- ***Activity Metabolomics***

May 17th

---- 09:00 am Begin ----

---- 10:30 am Break ----

---- 12:00 pm Lunch ---

---- 02:30 pm Break ----

---- 04:00 pm Finish ----

Advanced Metabolomics

June 3rd 2018

Relevant Metabolomics Papers

Systems Biology guided by Metabolomics	**Nature Methods 2017**
Metabolomics: Beyond Biomarkers and Towards Mechanisms	**Nature Reviews 2016**
XCMS: Processing MS Data using Nonlinear Alignment and Metabolite ID	**Analytical Chemistry 2006**
Mzmine 2: Modular framework for processing, visualizing, and analyzing MS data	**BMC Bioinformatics 2010**
Bioinformatics: The Next Frontier of Metabolomics	**Analytical Chemistry 2015**
Predicting Network Activity from High Throughput Metabolomics	**PLOS Computational Biology 2013**
Interactive XCMS Online: Simplifying Advanced Data Processing and Statistical	**Analytical Chemistry 2014**
Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling	**Analytical Chemistry 2015**
Thermal Degradation of Small Molecules: A Global Metabolomics Investigation	**Analytical Chemistry 2015**
Arteriovenous Blood Metabolomics: A Readout of Intra-Tissue Metabostasis	**Scientific Reports 2015**
Metabolism Links Bacterial Biofilms and Colon Carcinogenesis	**Cell Metabolism 2015**
CFM-ID: a web server for annotation, prediction and metabolite ID from tandem mass spectra	**Nucleic Acid Research 2014**
Determining Conserved Metabolic Biomarkers from a Million Database Queries	**Bioinformatics 2015**
Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling	**Analytical Chemistry 2015**
Metabolomic data streaming for biology-dependent data acquisition	**Nature Biotechnology 2014**
Comprehensive bioimaging with fluorinated nanoparticles	**Nature Comm. 2015**
Liquid chromatography quadrupole time-of-flight mass spectrometry	**Nature Protocols 2013**
Multivariate Analysis in Metabolomics	**Current Metabolomics 2013**
Intra- and Interlaboratory Reproducibility of UPLC TOF MS for Urinary Metabolic Profiling	**Analytical Chemistry 2012**
A Guideline to Univariate Statistical Analysis for LC/MS	**Metabolites 2012**
An accelerated workflow for untargeted metabolomics using METLIN database	**Nature Biotechnology 2012**
Within-Day Reproducibility of an HPLC-MS-Based Method	**Journal Proteome Research 2007**
HMDB: the Human Metabolome Database	**Nucleic Acid Research 2007**
XCMS: Processing MS Data using Nonlinear Alignment and Metabolite ID	**Analytical Chemistry 2006**
METLIN: A Mass Spectral Database	**Therapeutic Drug Monitoring 2005**

nature biotechnology

Metabolomics activity screening for identifying metabolites that modulate phenotype

Carlos Guijas[1,4], J Rafael Montenegro-Burke[1,4], Benedikt Warth[1,2,4], Mary E Spilker[1] & Gary Siuzdak[1,3]

Metabolomics, in which small-molecule metabolites (the metabolome) are identified and quantified, is broadly acknowledged to be the omics discipline that is closest to the phenotype[1–3]. Although appreciated for its role in biomarker discovery programs, metabolomics can also be used to identify metabolites that could alter a cell's or an organism's phenotype. Metabolomics activity screening (MAS) as described here integrates metabolomics data with metabolic pathways and systems biology information, including proteomics and transcriptomics data, to produce a set of endogenous metabolites that can be tested for functionality in altering phenotypes. A growing literature reports the use of metabolites to modulate diverse processes, such as stem cell differentiation, oligodendrocyte maturation, insulin signaling, T-cell survival and macrophage immune responses. This opens up the possibility of identifying and applying metabolites to affect phenotypes. Unlike genes or proteins, metabolites are often readily available, which means that MAS is broadly amenable to high-throughput screening of virtually any biological system.

Historically, metabolites have been either supplemented or eliminated from growth media and diets to modulate cellular activity and affect phenotype. For example, in the phenylalanine hydroxylase deficiency disease phenylketonuria, deficient metabolism of phenylalanine results in severe and adverse symptoms that can only be ameliorated by strict adherence to a low-phenylalanine diet from birth[4]. A prominent example of a frequently supplemented metabolite is niacin (vitamin B_3), which has an important role in energy transfer and maintenance of metabolic activity[5]. Metabolites can also function as metabolic coenzymes (e.g., coenzyme Q10 (CoQ10) and thiamine) and modulation of coenzymes can alter phenotypes by altering regulation of enzyme reactions. For example, statins, a class of cholesterol-lowering drugs, have the side effect of inhibiting the endogenous synthesis of CoQ10 (ref. 6). CoQ10 (ubiquinone) is a commonly prescribed supplement for patients receiving statins to regain mitochondrial energy homeostasis.

Metabolomics is used to identify the set of metabolites that are associated with physiological conditions or aberrant processes. To date, the main focus of the field has been on using this information to identify biomarkers and active or dysregulated pathways. In this Perspective, we discuss how to screen metabolomics data for metabolites that can be used to either induce or suppress biological functions. Unlike proteins, or genes, endogenous metabolites are readily amenable to biological testing and clinical applications.

Metabolomics activity screening

Untargeted (global) metabolomics uses liquid chromatography high-resolution mass spectrometry (LC-MS) to carry out comprehensive comparative analysis of metabolites. LC-MS is well-suited to metabolomic analyses, because it has high sensitivity, specificity, and reproducibility. It enables a broad statistical assessment of the metabolites extracted from a sample, and can be used to reveal unanticipated metabolic perturbations. There are numerous commercial and freely available data-processing packages, such as XCMS Online[7], Mzmine[8], and MetaboAnalyst[9], that can be applied to analyze LC-MS data. These suites of algorithms can identify chromatographic peaks, align them, and then statistically assess the comparative data, based on calculated probability, fold change, and intensity. Metabolites that are differentially regulated can be identified using databases (e.g., METLIN (https://metlin.scripps.edu), the human metabolome database (HMDB; http://www.hmdb.ca), and LIPID MAPS; http://www.lipidmaps.org/)[10–12], whose features and limitations have been reviewed[13]. The main advantage of untargeted LC-MS metabolomics is that it is an unbiased way to identify metabolites associated with a particular condition, whether it is stem-cell differentiation[14,15], immune-cell activation[16–19], remyelination in multiple sclerosis[20], chronic pain[21], or type 2 diabetes[22,23], to name but a few of the hundreds of examples that have been reported.

Endogenous metabolites identified in metabolomics data sets can be screened to identify metabolites that modulate phenotype. Unlike genes and proteins, metabolites are readily available, typically inexpensive, and have relatively simple structural features making them very amenable to screening.

Various MAS workflows can be designed to identify metabolites from metabolomics experiments for activity testing (**Fig. 1**). The most straightforward approach selects metabolites based on statistical significance and fold change, which is also the standard method for screening metabolites in global metabolomics experiments. For example, in a comparative analysis using a cell model, any metabolites

[1]The Scripps Research Institute, Scripps Center for Metabolomics and Mass Spectrometry, La Jolla, California, USA. [2]University of Vienna, Department of Food Chemistry and Toxicology, Vienna, Austria. [3]The Scripps Research Institute, Department of Integrative and Computational Biology, La Jolla, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to G.S. (siuzdak@scripps.edu).

Received 28 June 2017; accepted 14 February 2018; published online 5 April 2018; doi:10.1038/nbt.4101

that have statistical significance represented by a *P*-value lower than 0.001, and fold changes greater than two, would qualify for further testing, although these values are user-defined and can vary. A secondary level of candidate selection would be to test metabolites from pathways identified as being active, a feature that has been recently automated in XCMS Online[24]. This 'biologically driven' selection method would include metabolites identified as dysregulated and metabolites involved in pathways of interest. Metabolites can be plotted onto pathway maps and ranked on the basis of the number of pathways involving each metabolite, leveraging pathway specificity. A third level of candidate selection can be mediated by manipulating the activity of enzymes in pathways using inhibitors or molecular biology approaches.

An important part of metabolite selection, beyond evaluating statistical significance, fold change, and pathways, is metabolite identification. For this purpose, multiple databases have been created that allow metabolites to be putatively identified using accurate mass and tandem mass spectrometry data[10]. Metabolite identification is validated by comparison with an authentic standard with tandem MS data generation as well as chromatography retention time (when available). Further validation in which experimental samples are analyzed using a targeted approach with triple quadrupole MS to compare against the original quantification (performed in the untargeted experiments) can also be used. These multiple levels of authentication help minimize misidentifications, which commonly occurred in the past when only precursor *m*/*z* values were used.

It is worth noting that while databases for initial identification information are not complete, their growth has been tremendous in the last decade. Currently, users examine multiple databases when performing searches because the databases are not completely overlapping[25].

Phenotype-modulating metabolites identified using MAS

Metabolomics has been applied to provide insights into immunomodulation[16–19], cardiovascular disease[26–28], and diabetes[22,23], with specific examples from our group, including stem cell differentiation (G.S. and colleagues)[14], the role of microbiome metabolism (G.S. and colleagues)[29], molecular origins of chronic pain (G.S. and colleagues)[21], and, most recently, remyelination for neuron repair (J.R.M.-B., G.S, and colleagues)[20]. Comparatively, though, little effort has been dedicated to examining the activity of these biomarkers. In the following paragraphs, we briefly outline five examples of biologically active metabolites as unraveled by MAS.

Modulating stem cell differentiation. One of our (G.S. and colleagues)[14,30] earliest efforts in stem cell analysis was designed to identify metabolites associated with cell differentiation. In these experiments, the metabolome of pluripotent stem cells, differentiated neurons and cardiomyocytes were quantitatively compared. Globally, the differentially regulated metabolites indicated that oxidation was a primary driver for cell differentiation. For example, arachidonic acid, a polyunsaturated fatty acid and the metabolic precursor to >100 functionally diverse metabolites, is highly upregulated in stem cells. Arachidonic acid in stem cells is important for maintaining 'chemical plasticity' and in mediating differentiation by regulation of redox status and activation of oxidative pathways. A crucial downstream molecule in these experiments, protectin D1 (derived from docosahexaenoic acid, also a polyunsaturated fatty acid) was used to promote differentiation and neurogenesis at concentrations as low as 50 nM (**Fig. 2a**).

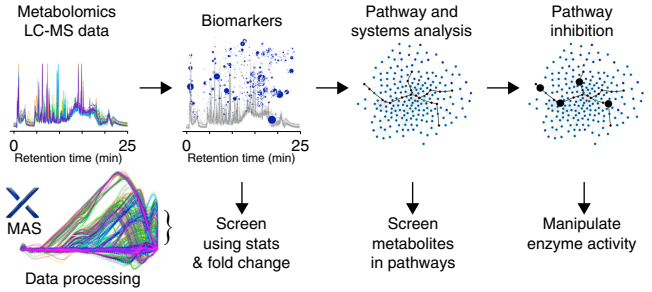


Figure 1 MAS for the identification of endogenous metabolites that modulate phenotype. Metabolomics data analysis and identification of candidates for screening are carried out by XCMS Online or other data-processing approaches. Initial candidates are generated using statistical and fold-change cut-offs and can then be further investigated using high-throughput screening to identify biologically active metabolites. Pathway analysis can provide additional metabolite candidates, while a third level of screening would identify candidates following perturbations with known pathway inhibitors.

Overall, the results from these experiments suggested that the activation of oxidation is a metabolic requirement of stem-cell differentiation. Specifically, endogenous metabolites that promote pluripotency induce stem cells to a more reduced state whereas those that promote differentiation induce a more oxidized state. Moreover, it is well known that hypoxia maintains the pluripotent and undifferentiated phenotype of stem or precursor cells both, *in vitro* and *in vivo*[31]. Interestingly, these results also showed that endogenous metabolites are not merely substrates and products of metabolic reactions, but rather are involved in modulating stem cell differentiation and can be used to enhance their regenerative potential.

Modulating type 2 diabetes. Branched fatty acid esters of hydroxy fatty acids (FAHFAs) were discovered as dysregulated metabolites in mice protected against diabetes and further used to modulate the type 2 diabetes phenotype[22]. A class of uncharacterized endogenous metabolites were found to be highly upregulated in the adipose tissue and plasma of mice overexpressing the glucose transporter Glut4 compared to their wild-type littermates, in an untargeted metabolomics study. Even though the *m*/*z* of these compounds did not correspond to any known metabolite in METLIN and LIPID MAPS, its structure was characterized as FAHFA using fragmentation spectra in negative-ion mode. Glut4-overexpressing transgenic mice have an elevated lipogenesis and glucose tolerance, despite being obese, with elevated levels of circulating fatty acids. Hence, it was hypothesized that FAHFAs could affect glucose and insulin homeostasis. Once chemically characterized and synthesized, palmitic acid 9-hydroxystearic acid (9-PAHSA), one of the most abundant FAHFAs, was tested in an *in vivo* model of type 2 diabetes. Diabetic mice fed a high-fat diet, that were orally administered 9-PAHSA, showed an overall higher glucose tolerance and insulin sensitivity compared with controls (**Fig. 2b**). Moreover, in adipocytes, the improvement in glucose metabolism resulted from 9-PAHSA-triggered binding and activation of the GPR120 receptor, a well-known anti-inflammatory and insulin-sensitizing mediator in response to omega-3 fatty acids.

Because type 2 diabetes is accompanied by a low-grade inflammation in adipose tissue that may contribute to the insulin-resistant state, 9-PAHSA was further tested as a possible immunomodulator

of the adipose-tissue-associated inflammatory response. Mice orally supplemented with 9-PAHSA showed an effective reduction in the *in vivo* inflammatory response of adipose tissue macrophages to a high-fat diet. In summary, 9-PAHSA was discovered and tested as a possible phenotype modulator. When exogenously administered, 9-PAHSA increased insulin sensitivity and glucose tolerance in a mouse model of type 2 diabetes[22].

Modulating T-cell survival and anti-tumor activity. Metabolic modulation through L-arginine prompted a central memory-like T-cell phenotype with enhanced survival capacity and anti-tumor activity both *in vitro* (human) and *in vivo* (mouse model)[19]. In that study, untargeted flow injection metabolomics analyses[32] were performed to determine the dynamic changes in arginine metabolism during a time-course experiment. Results were validated by monitoring cell uptake of isotopically labeled L-arginine to determine its fate/flux as well as enzyme inhibitors and clones.

These observations were then further explored to demonstrate that higher L-arginine levels induced structural alterations in three transcriptional regulators (BAZ1B, PSIP1, and TSN) and modulated T-cell metabolic 'fitness' and survival (**Fig. 2c**).

Modulating innate immune response. Correct regulation of the innate immune response is a key factor in the maintenance of whole-body homeostasis. Dysregulation of the immune response may underpin several illnesses related to an excessive or chronic activation or immunosuppression. Relevant to this, the uncommon phosphatidylinositol species 1,2-diarachidonyl-glycero-3-phosphoinositol (PI(20:4/20:4)) was found to be upregulated in mouse-resident peritoneal macrophages stimulated with the yeast cell wall preparation zymosan, a classic stimulus of the innate immune response[16]. This lipid species, previously characterized using the LIPID MAPS database, is rapidly formed and degraded upon stimulation, suggesting a role in regulating cell signaling events, such as generation of reactive oxygen species and secretion of lysozyme, two pivotal molecules produced by macrophages for pathogen killing. When added exogenously, macrophages incorporate this molecule into their phospholipid pool and show a higher superoxide anion production and lysozyme secretion than control cells and macrophages enriched with a scrambled phosphatidylinositol species (**Fig. 2d**), suggesting this molecule plays a key role in the coordination of the macrophage response to zymosan.

Modulating oligodendrocyte maturation. We (J.R.M.-B., G.S., and collaborators)[20] have also used MAS to analyze oligodendrocyte precursor cell (OPC) differentiation in multiple sclerosis, an autoimmune disease characterized by demyelination of axons and neuronal dysfunction. Disease remission in multiple sclerosis is dependent on remyelination, which involves the differentiation of OPCs and leads to the formation of mature oligodendrocytes[33]. Premyelinating oligodendrocytes are present in chronic lesions of patients and inhibition of OPC differentiation is associated with multiple sclerosis disease progression. Therefore, a promising complementary treatment of multiple sclerosis involves the identification of pharmacological agents that stimulate remyelination by enhancing OPC differentiation. Multiple drug candidates have been identified using high-throughput screening[34], which induce OPC differentiation *in vitro* and enhance remyelination *in vivo*.

We used MS-based metabolomics to investigate how endogenous metabolites play a role in the process of OPC differentiation[20]. Among other related metabolites, taurine, an amino sulfonic acid, was found to be significantly elevated (~20-fold) over the course of *in vitro* oligodendrocyte differentiation (**Fig. 2e**). When added exogenously at physiologically relevant concentrations, taurine not only enhanced drug-induced OPC differentiation but also facilitated the *in vitro* myelination of co-cultured axons. Unlike in L-arginine T-cell modulation, and overturning the common assumption that upregulated metabolites are end-point biomarkers, the addition of upregulated taurine had a positive effect, further stimulating remyelination during OPC differentiation. Mechanistically, taurine-induced activities that enhance OPC differentiation and myelination appear to be driven by taurine's ability to increase serine levels, which is an initial building block required for the synthesis of the glycosphingolipid components of myelin.

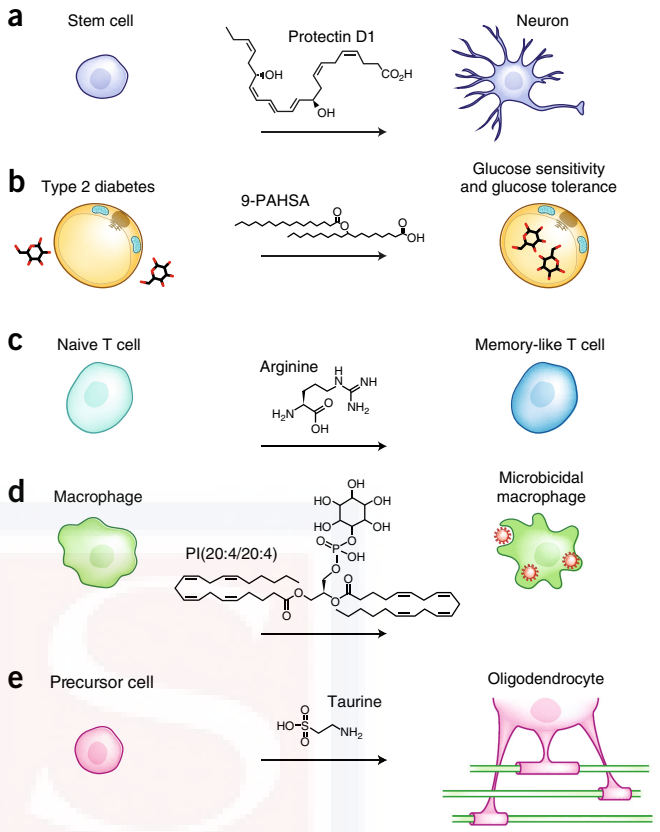


Figure 2 MAS demonstrated in stem-cell differentiation, a mouse model of type 2 diabetes, T-cell function and activity, macrophage response to a fungal stimulus, and a remyelination model for multiple sclerosis. (**a**) Experiments with embryonic stem cells identified the metabolites involved in their differentiation. Among them, protectin D1, a lipid, was found to enhance differentiation to neurons by a factor of 15 (ref. 14). (**b**) 9-PAHSA was discovered in adipose tissue and plasma of glucose-tolerant mice. This metabolite was identified as a key molecule that maintains correct glucose homeostasis in a model of type 2 diabetes induced by a high-fat diet[22]. (**c**) L-arginine levels decreased in activated naive T-cells. When L-arginine levels were raised externally, this amino acid actively increased survival and anti-tumor activity of T cells by modulating the activity of several transcriptional factors[19]. (**d**) Minor phospholipid species PI(20:4/20:4) is actively synthesized by activated macrophages. When exogenously added, this lipid amplified microbicidal capacity of macrophages in response to the fungal stimulus zymosan[16]. (**e**) Taurine, that was observed to be highly upregulated during OPC differentiation, enhanced the effect of a novel drug treatment (miconazole) to induce OPC differentiation into mature oligodendrocytes, a promising cell target for multiple sclerosis treatment[20].

Outlook

In the past, several metabolites have been discovered as effective phenotype modulators, using approaches other than MAS, including cellular fractioning, ligand-binding assays, and enzymatic assays. Examples include sphingosine-1-phosphate (immunomodulation)[35], docosahexaenoic acid (cognitive function)[36], carnitine (fertility)[37], anandamide (neurological disorders)[38], and melatonin (sleep)[39], to name a few. However, the use of metabolomics data to characterize dysregulated metabolites of interest is gaining more attention because this approach is able to detect a wide range of small molecules at low concentrations, increasing throughput. Thus, metabolomics has been successful in identifying active metabolites for phenotype modulation (**Table 1**). It is clear that metabolomics can enable identification of molecules with interesting and potentially beneficial functions.

Notwithstanding MAS's clear utility, challenges exist that could impede the broad its implementation. It is unclear how to accurately identify either the best candidate molecules for further testing or which molecule among the numerous other dysregulated metabolites is likely to be the most effective phenotype modulator. Statistical analyses, metabolite classification schemes based on prior metabolite activity knowledge, and pathway redundancy have all been used to prioritize the best candidates and reduce the need for large-scale biological validation experiments. Follow-up experiments including the use of pathway metabolites, pathway inhibitors and stable isotope labeling as well as flux analysis are valid strategies to further reduce the list of candidate metabolites. Another challenge for untargeted metabolomics studies is the identification of 'unknown' molecules. This is attributed to the chemical diversity and heterogeneity of the metabolome, and substantial effort has been dedicated to the development of advanced computational tools for tandem MS prediction and metabolite characterization. Although thousands of metabolites are commercially available, a limiting aspect of MAS can be the lack of commercially available reference materials for activity validation, particularly for lesser known or characterized metabolites. In those cases, the potential solutions available at present involve synthetic or isolation strategies, and for the most part, the former is the more common approach because large amounts of sample are rarely available to undertake isolation attempts[40].

Discovery metabolomics has largely been used to identify biomarkers and characterize mechanisms of biological action. Going forward, the use of MAS to identify biologically active endogenous metabolites that can be used to intentionally alter phenotype might prove to be a far more effective application of metabolomics. Metabolites identified using MAS can be used to induce a phenotypic response alone, or in conjunction with a drug. MAS might conceivably permit dose or side effect reduction while maintaining or even improving therapeutic outcomes.

Applications of MAS could be expanded to disease modulation, biofilm initiation or suppression, drug–exposome interactions, plant biology and immunotherapy. Perhaps what is most intriguing is that rather than identifying metabolites to understand pathways, we can apply metabolites to modulate physiology, thereby turning the tables on conventional thinking.

Table 1 Metabolomics activity screening examples

Metabolite	System	Original observation	Induced phenotype	Reference
Protectin D1	Stem cell differentiation	Polyunsaturated fatty acid precursors decrease during differentiation	Promotes differentiation into neurons	14
cis-9,10-octadecenoamide (oleamide)	Sleep induction	Accumulated in cerebrospinal fluid of sleep-deprived felines	Induces sleep	41
Trimethylamine *N*-oxide	Cardiovascular disease	Augmented in plasma of subjects with cardiovascular risk	Increases scavenger receptors expression, foam cell formation and atherosclerotic lesions	27,28
N,N-dimethylsphingosine	Chronic pain	Increased in a rat model of chronic neuropathic pain	Elicits neuropathic pain behavior and cytokine release	21
PI(20:4/20:4)	Pathogen killing	Upregulated in zymosan-stimulated macrophages	Increases superoxide anion production and lysozyme release	16
3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid	Gestational and type 2 diabetes	Elevated in plasma in human and mice models of gestational diabetes and type 2 diabetes	Impairs glucose tolerance and β-cell function	23
9-PAHSA	Type 2 diabetes	Increased in plasma and adipose tissue of diabetes-protected mice. Decreased in diabetic humans	Improves glucose metabolism and insulin sensitivity	22
S-adenosyl methionine	Stem cell differentiation	Downregulated in naive embryonic stem cells	Induces differentiation to primed stem cells	15
cis-7-hexadecenoic acid (16:1n-9)	Cardiovascular disease	Elevated in atherosclerosis-initiating foamy monocytes and macrophages	Decreases inflammatory response to bacterial lipopolysaccharide in monocytes and macrophages	17
Dioxolane A3	Acute inflammation	Increased in thrombin-activated platelets	Promotes neutrophil recruitment and activation	18
Proline, isoleucine, and phenylalanine	Synthetic mutualism	Secreted by *Zymomonas mobilis*	Results in rescue and growth of *Escherichia coli* auxotrophs in co-culture with *Z. mobilis*	42
L-arginine	Adaptive immune response	Decreased in activated naive T cells	Induces differentiation into memory-like cell, increases survival and anti-tumor activity	19
Taurine	Multiple sclerosis	Upregulated during oligodendrocyte precursor cell differentiation	Enhances oligodendrocyte differentiation and myelination	20

ACKNOWLEDGMENTS
We gratefully acknowledge financial support from the National Institutes of Health (Grants R01 GM114368-03, P30 MH062261-10, P01 DA026146-02), and support was also received from Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley Laboratory for the US Department of Energy, Office of Science, Office of Biological and Environmental Research under Contract DE-AC02-05CH11231.

COMPETING INTERESTS
The authors declare no competing interests.

1. Johnson, C.H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
2. Patti, G.J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
3. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
4. Woolf, L.I., Griffiths, R. & Moncrieff, A. Treatment of phenylketonuria with a diet low in phenylalanine. *BMJ* **1**, 57–64 (1955).
5. Kamanna, V.S. & Kashyap, M.L. Mechanism of action of niacin. *Am. J. Cardiol.* **101** 8A, 20B–26B (2008).
6. Banach, M. *et al.* Statin therapy and plasma coenzyme Q10 concentrations--A systematic review and meta-analysis of placebo-controlled trials. *Pharmacol. Res.* **99**, 329–336 (2015).
7. Tautenhahn, R., Patti, G.J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84**, 5035–5039 (2012).
8. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič , M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
9. Xia, J., Sinelnikov, I.V., Han, B. & Wishart, D.S. MetaboAnalyst 3.0-making metabolomics more meaningful. *Nucleic Acids Res.* **43**, W251–W257 (2015).
10. Tautenhahn, R. *et al.* An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* **30**, 826–828 (2012).
11. Wishart, D.S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
12. Fahy, E. *et al.* A comprehensive classification system for lipids. *J. Lipid Res.* **46**, 839–861 (2005).
13. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *Trends Analyt. Chem.* **78**, 23–35 (2016).
14. Yanes, O. *et al.* Metabolic oxidation regulates embryonic stem cell differentiation. *Nat. Chem. Biol.* **6**, 411–417 (2010).
15. Sperber, H. *et al.* The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. *Nat. Cell Biol.* **17**, 1523–1535 (2015).
16. Gil-de-Gómez, L. *et al.* A phosphatidylinositol species acutely generated by activated macrophages regulates innate immune responses. *J. Immunol.* **190**, 5169–5177 (2013).
17. Guijas, C., Meana, C., Astudillo, A.M., Balboa, M.A. & Balsinde, J. Foamy monocytes are enriched in cis-7-hexadecenoic fatty acid (16:1n-9), a possible biomarker for early detection of cardiovascular disease. *Cell Chem. Biol.* **23**, 689–699 (2016).
18. Hinz, C. *et al.* Human platelets utilize cycloxygenase-1 to generate dioxolane A3, a neutrophil-activating eicosanoid. *J. Biol. Chem.* **291**, 13448–13464 (2016).
19. Geiger, R. *et al.* L-arginine modulates t cell metabolism and enhances survival and anti-tumor activity. *Cell* **167**, 829–842.e13 (2016).
20. Beyer, B.A. *et al.* Metabolomics-based discovery of a metabolite that enhances oligodendrocyte maturation. *Nat. Chem. Biol.* **14**, 22–28 (2018).
21. Patti, G.J. *et al.* Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. *Nat. Chem. Biol.* **8**, 232–234 (2012).
22. Yore, M.M. *et al.* Discovery of a class of endogenous mammalian lipids with anti-diabetic and anti-inflammatory effects. *Cell* **159**, 318–332 (2014).
23. Prentice, K.J. *et al.* The furan fatty acid metabolite CMPF is elevated in diabetes and induces β cell dysfunction. *Cell Metab.* **19**, 653–666 (2014).
24. Huan, T. *et al.* Systems biology guided by XCMS Online metabolomics. *Nat. Methods* **14**, 461–462 (2017).
25. Wohlgemuth, G. *et al.* SPLASH, a hashed identifier for mass spectra. *Nat. Biotechnol.* **34**, 1099–1101 (2016).
26. Koeth, R.A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576–585 (2013).
27. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011).
28. Wang, Z. *et al.* Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* **163**, 1585–1595 (2015).
29. Wikoff, W.R. *et al.* Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. USA* **106**, 3698–3703 (2009).
30. Panopoulos, A.D. *et al.* The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Res.* **22**, 168–177 (2012).
31. Ezashi, T., Das, P. & Roberts, R.M. Low O_2 tensions and the prevention of differentiation of hES cells. *Proc. Natl. Acad. Sci. USA* **102**, 4783–4788 (2005).
32. Fuhrer, T., Heer, D., Begemann, B. & Zamboni, N. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Anal. Chem.* **83**, 7074–7080 (2011).
33. Franklin, R.J.M. & Ffrench-Constant, C. Remyelination in the CNS: from biology to therapy. *Nat. Rev. Neurosci.* **9**, 839–855 (2008).
34. Deshmukh, V.A. *et al.* A regenerative approach to the treatment of multiple sclerosis. *Nature* **502**, 327–332 (2013).
35. Fernández-Pisonero, I. *et al.* Synergy between sphingosine 1-phosphate and lipopolysaccharide signaling promotes an inflammatory, angiogenic and osteogenic response in human aortic valve interstitial cells. *PLoS One* **9**, e109081 (2014).
36. Cardoso, C., Afonso, C. & Bandarra, N.M. Dietary DHA and health: cognitive function ageing. *Nutr. Res. Rev.* **29**, 281–294 (2016).
37. Ng, C.M., Blackman, M.R., Wang, C. & Swerdloff, R.S. The role of carnitine in the male reproductive system. *Ann. NY Acad. Sci.* **1033**, 177–188 (2004).
38. Wise, L.E., Shelton, C.C., Cravatt, B.F., Martin, B.R. & Lichtman, A.H. Assessment of anandamide's pharmacological effects in mice deficient of both fatty acid amide hydrolase and cannabinoid CB1 receptors. *Eur. J. Pharmacol.* **557**, 44–48 (2007).
39. Hardeland, R. *et al.* Melatonin—a pleiotropic, orchestrating regulator molecule. *Prog. Neurobiol.* **93**, 350–384 (2011).
40. Cohen, L.J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549**, 48–53 (2017).
41. Cravatt, B.F. *et al.* Chemical characterization of a family of brain lipids that induce sleep. *Science* **268**, 1506–1509 (1995).
42. Kosina, S.M. *et al.* Exometabolomics assisted design and validation of synthetic obligate mutualism. *ACS Synth. Biol.* **5**, 569–576 (2016).

INNOVATION

Metabolomics: beyond biomarkers and towards mechanisms

Caroline H. Johnson, Julijana Ivanisevic and Gary Siuzdak

Abstract | Metabolomics, which is the profiling of metabolites in biofluids, cells and tissues, is routinely applied as a tool for biomarker discovery. Owing to innovative developments in informatics and analytical technologies, and the integration of orthogonal biological approaches, it is now possible to expand metabolomic analyses to understand the systems-level effects of metabolites. Moreover, because of the inherent sensitivity of metabolomics, subtle alterations in biological pathways can be detected to provide insight into the mechanisms that underlie various physiological conditions and aberrant processes, including diseases.

Metabolites are the substrates and products of metabolism that drive essential cellular functions, such as energy production and storage, signal transduction and apoptosis. In addition to being produced directly by the host organism, metabolites can derive from microorganisms, as well as from xenobiotic, dietary and other exogenous sources[1].

The biochemical actions of metabolites are far-reaching. To start, metabolites can regulate epigenetic mechanisms and maintain the pluripotency of embryonic stem cells (ES cells)[2–6]. It has also been well established that metabolites such as ATP, acetyl-CoA, NAD^+, and *S*-adenosyl methionine (SAM) can function as co-substrates, regulating post-translational modifications that affect protein activity[7,8]. In addition, fatty acids and hormones can interact with plasma proteins to enable their transport in the bloodstream[9,10]. Furthermore, metabolite–protein interactions can aid in facilitating cellular responses by initiating signalling cascades, thus evidencing the role of metabolites in signal transduction[11,12].

Indirectly, metabolites affect the environment in which they are produced. Under normal conditions, homeostatic controls exist to counteract any adverse biological consequences of such effects. For example, acidic metabolites decrease the pH of the microenvironment[13,14], and high concentrations of these acidic metabolites are found, for instance, in the colon, owing to bacterial fermentation of dietary carbohydrates that leads to the production of short-chain fatty acids. These are, however, efficiently neutralized by mucosal production of bicarbonate. Notably, such homeostatic controls can be compromised with age and during disease, leading to functional decline and a failure to return to steady state. In addition, the adaptation of aberrant glycolytic cancer cells to the large amounts of lactate and protons that they produce occurs through modification of the activity of transporters, exchangers, pumps and carbonic anhydrases, which all help to maintain the intracellular pH and enable cells to survive the acidic microenvironment[15]. Thus, as metabolites can have a wide range of functions in the cell and organism, there is growing motivation to better ascertain their specific functions, as well as to understand their physiological roles. This can be done by implementing various metabolomic approaches to identify metabolites and metabolic pathways that are associated with particular phenotypes, and then integrating this knowledge with functional and mechanistic biological studies.

The main methodologies that are used for metabolite recovery and identification are untargeted (global) and targeted mass spectrometry-based metabolomics, which are discussed in more detail in BOX 1.

Untargeted metabolomics aims to measure the broadest range of metabolites present in an extracted sample without *a priori* knowledge of the metabolome. The types of metabolites that are recovered are influenced by the extraction and analytical method of choice, but they result in a complex data set that requires computational tools to identify and correlate metabolites between samples and to examine their interconnectivity in metabolic pathways in relation to the phenotype or aberrant process (see BOX 2 and Supplementary information S1 (box)). By contrast, targeted metabolomics provides higher sensitivity and selectivity than untargeted metabolomics, but metabolites are analysed on the basis of *a priori* information, whereby methods are developed and optimized for the analysis of specific metabolites and metabolic pathways of interest. Targeted analysis also constitutes an important part of a metabolomics workflow to validate and expand upon results from untargeted analysis[16].

The types of samples that can be analysed using metabolomics are wide-ranging and include tissues, cells and biofluids. Tissue analysis, in particular, is perhaps the most powerful approach for studying localized and specific responses to stimuli and pathogenesis, yielding explicit biochemical information about the mechanisms of disease. Traditionally, tissue analysis involves extraction of the complete tissue material into a liquid form, from which the metabolite changes are averaged across the different cell types and regions of the analysed organ. In addition to this total tissue analysis, subregional, cellular and even subcellular metabolite profiles can provide further insight into structure-to-function relationships; this is particularly valuable in the case of heterogeneous tissues such as brain and cancers[17]. Simultaneous sampling of arterial blood (entering the organ) and venous blood (draining the organ), followed by paired analysis, can also have value in the investigation of tissue metabolic activity[16]. This paired arteriovenous approach provides information about the metabolite uptake and release patterns across the

Box 1 | **Mass spectrometry in metabolomics**

Mass spectrometry

Mass spectrometry is an excellent analytical platform for metabolomic analysis, as it provides high sensitivity, reproducibility and versatility. It measures the masses of molecules and their fragments to determine their identity. This information is gained by measuring the mass-to-charge ratio (*m*/*z*) of ions that are formed by inducing the loss or gain of a charge from a neutral species. The sample, comprising a complex mixture of metabolites, can be introduced into the mass spectrometer either directly or preceded by a separation approach (using liquid chromatography or gas chromatography). Direct injection has been successfully implemented for high-throughput metabolomics. However, as thousands of ions can be present in metabolomic experiments, chromatographic separation before entering the mass spectrometer minimizes signal suppression and allows for greater sensitivity, and — by providing a retention time identifier — it can further aid metabolite identification. In addition to *m*/*z* and retention time information, the identification of an ion is facilitated by fragmentation pattern information that is acquired by tandem mass spectrometry[83].

Untargeted metabolomics

Untargeted or global metabolomic analysis allows for an assessment of the metabolites extracted from a sample and can reveal novel and unanticipated perturbations. Untargeted analyses are most effective when implemented in a high-resolution mass spectrometer, to facilitate structural characterization of the metabolites. Its primary advantage is that it offers an unbiased means to examine the relationship between interconnected metabolites from multiple pathways. However, it is not yet possible to obtain all metabolite classes simultaneously, as many factors affect metabolite recovery, depending on the functional group of the metabolite. In addition, there are a large number of unknown metabolites that remain unannotated in metabolite databases[35]. Thus, depending on the pH, solvent, column chemistry and ionization technique used, untargeted metabolomics can provide a detailed assessment of the metabolites in a sample, revealing a wide range of metabolite classes.

Targeted metabolomics

Targeted metabolomic analyses measure the concentrations of a predefined set of metabolites. A standard curve for a concentration range of the metabolite of interest is prepared, so that accurate quantification can be gained. This type of analysis can be used to obtain exact concentrations of metabolites identified by untargeted metabolomics, providing analytical validation.

Imaging metabolomics

It is also possible to reveal the localization of selected metabolites within a tissue sample using imaging mass spectrometry techniques, such as matrix-assisted laser desorption ionization (MALDI)[84], nanostructure-imaging mass spectrometry (NIMS)[70,85], desorption electrospray ionization mass spectrometry (DESI)[86] and secondary ion mass spectrometry (SIMS)[87], among others. NIMS and DESI are especially suited to the analysis of small molecules.

tissue of interest and therefore gives insight into tissue metabostasis. The power of this paired analysis allows for the measurement of metabolite arteriovenous differences or ratios and offers a compelling compromise with sampling effort, compared to the traditional approach of venous blood analysis.

During the past few years, substantial progress has been made in metabolomic analysis by improving instrument performance, experimental design and sample preparation, ultimately facilitating broader analytical capabilities. Moreover, the surge in new chemoinformatic (computational approaches for handling chemical information) and bioinformatic (computational approaches for handling biological information) tools has provided extensive support for data acquisition, analysis and integration. This has greatly enhanced our ability to identify metabolites in various samples and allowed us to correlate these metabolites with particular phenotypes, thus establishing useful biomarkers that are indicative of particular physiological states or aberrations. The ultimate challenge now is to move beyond simply identifying metabolites and using them as biomarkers, and to start establishing the direct physiological roles of metabolites and their involvement in metabolic networks, as well as determining how changes in their levels are implicated in different phenotypic outcomes. This Innovation article focuses on how this most relevant hurdle for metabolomics can be overcome. We describe how advances in technologies that are used in metabolite identification and analysis, experimental design and pathway mapping are helping us to gain more meaningful data, revealing important nodes for further investigation. We also discuss how this information, when combined with traditional biological methods, can enable us to ascertain molecular mechanisms and begin to infer biological causality.

Current challenges in metabolomics

During the past few years, metabolomics has evolved considerably to overcome challenges that initially confounded analysis[18]. A major challenge still exists for the identification of metabolites and validation of metabolites in human populations. However, the most important challenge is to develop workflows for assigning biological meaning to metabolites and to move towards finding mechanisms of disease.

Metabolite identification and validation. The initial focus of metabolomics has been on biomarker discovery, with the aim of identifying metabolites that are correlated with various diseases and environmental exposures. This has, for example, led to the identification of plasma trimethylamine *N*-oxide (TMAO) and urinary taurine as markers of cardiovascular disease (CVD)[19] and ionizing radiation exposure[20–22], respectively. In order to correlate metabolites with a phenotype, the two biggest hurdles faced are metabolite identification and biomarker validation. In any given untargeted metabolomics experiment, only a subset of all metabolite features present can be positively identified. This has been facilitated by novel *in silico* tools[23–25] (see below, as well as BOX 2 and Supplementary information S1 (box)), the expansion and development of metabolite databases[26] (see BOX 2 and Supplementary information S1 (box)) and the synthesis of previously unattainable standard compounds that can confirm the identification of the metabolite (these standards are either novel compounds or were previously not available in an isotope-labelled form)[27].

Biomarker validation can be challenging, owing to difficulties in measuring subtle differences in metabolite concentrations between control and aberrant conditions, and because of the lack of follow-up with targeted metabolomic experiments (BOX 1). These follow-up experiments should be carried out in an additional cohort of biological samples for validation of the metabolite changes with the phenotype. Moreover, one of the largest challenges to biomarker validation is overcoming inter-individual metabolite variation, which arises owing to differences in genetic factors and environmental exposures. All of these influences result in significantly different metabolic responses in population studies[1], making it extremely difficult to pinpoint metabolites that are correlated with a particular condition and, ultimately, to provide clinical biomarkers. This is the case

especially when examining a multifaceted disease such as cancer. There are a number of methods that can be applied before and after analysis to overcome some of the biological variation associated with human studies. Establishing appropriate experimental design and statistical power for the study, and using patient questionnaires with subsequent population stratification, as well as regression modelling, can allow for the extraction of important metabolites[28]. These types of approaches can remove confounding samples from the analysis and help to streamline the data to identify metabolites that are correlated with the biological stimulus and not another influence. In addition, using appropriate metabolite normalization strategies, such as analysing metabolite ratios or normalizing to creatinine in urine studies, may help. Developing databases to collect data on the normal fluctuations in metabolite concentration ranges that occur in response to factors such as diet[29], age, gender, circadian rhythm and exercise, which are frequent causes of sample-to-sample variability, would also be useful. Indeed, some databases that contain information on specific metabolite concentration ranges in human biofluids and in dietary components — the Human Metabolome Database (HMDB)[30] and FooDB, respectively — have already been developed.

Functional analysis of metabolites. Perhaps the largest challenge that metabolomic researchers face in any study is relating the identified metabolites to their biological roles, which is a necessary step for moving beyond biomarkers and towards mechanisms. Biomarkers obtained from human population studies can provide a starting point for finding links between diseases and metabolic pathways[31], and further mechanistic work can be carried out using *in vitro* and animal-based studies, as previously shown[32]. Furthermore, patient-derived primary cell lines and xenografts can provide more reliable models for finding relatable data, as such samples make it possible to control for genetic and environmental influences.

However, to evaluate the biological roles of one or several metabolites (a metabolic signature), one first has to determine their functions in metabolic pathways and their interconnectivity, and, more broadly, determine which metabolic pathways are perturbed by the aberrant condition[33]. Only such a multi-level analysis can provide a comprehensive understanding of the systemic biological changes that are associated with particular metabolites and potentially direct further mechanistic studies. Determining the interactions of metabolites in metabolic pathways is particularly challenging. Metabolic pathway maps currently include ~2,000 metabolites; however, similar to metabolite databases, they are somewhat incomplete, as some metabolites have not yet been characterized[34,35]. Novel molecules are regularly being discovered, adding to the pool of known metabolites[22,36]. Multi-layered approaches that integrate metabolomic and other 'omics' data (see below) acquired from the same samples provide an opportunity to investigate the system-wide changes in a disease and to delve further into metabolic pathway interactions and the mechanisms of disease development and progression[37,38]. In addition, novel experimental approaches, such as stable isotope-assisted analysis (see below), can trace metabolite utilization in pathways in a temporal manner.

Recent technical advancements

Developments in innovative informatics strategies have been a major driver in overcoming some of the challenges presented with metabolomic analysis[33]. Advances in data processing, statistical analysis and metabolite characterization have enabled the identification of more metabolites that are associated with a particular phenotype than was ever previously achievable. Moving towards mechanistic investigations, novel metabolic pathway analysis tools that assess the interconnectivity of these metabolites can provide important insights, particularly

Box 2 | **Computational tools in metabolomics**

Metabolomic analyses, and untargeted metabolomics in particular, result in the generation of complex data sets; therefore, computational tools are crucial to process and interpret these results. The problems associated with big data processing, statistical analyses, metabolite identification and biological interpretation are not trivial, but there are now some novel tools available that accelerate and automate the computational workflows, providing user-friendly tools for both novice and expert bioinformaticians (for further details, refer to Supplementary information S1 (box)).

Data processing and statistical analysis

After data upload, mass spectral peaks are picked, realigned and annotated. The data is deconvoluted using computational tools to remove instrumental and chemical noise, thus providing only the biologically relevant information.

The types of statistical analyses that can be implemented for metabolomics data are vast, and choosing the correct test can be challenging. Online tools such as XCMS Online[42], DeviumWeb MetaboAnalyst[43] and many others give researchers the ability to carry out a wealth of tests. Some of the most recent advances are tools that provide false discovery rate measurements to ensure that the data have statistical power. Other concepts that are especially useful for finding biologically relevant metabolites are multi-group and meta-analyses, which can reveal shared metabolic changes across multiple experiments[88].

Metabolite identification and databases

Initial putative metabolite identifications can be made on the basis of the accurate mass-to-charge ratio (*m/z*) of the mass spectral ion. This is aided by the use of comprehensive metabolite databases such as METLIN[89], HMDB[90], MassBank[91] and GMD[26,92]. Tandem mass spectrometry experiments can then be carried out on the isolated ion, followed by matching with an authentic standard, in order to obtain characteristic fragments and retention time information to distinguish the ion from structural isomers. *In silico* prediction tools provide further insight into metabolite identification when a particular *m/z* or tandem mass spectrometry fragmentation pattern does not provide a match[24,93]. A recent innovation in ion mobility mass spectrometry, the rotationally averaged cross-collisional section (CCS), provides another level of metabolite identification, and databases containing CCS information are currently in the early stages of development[94]. Despite all of these innovations, some metabolite features cannot be assigned to a molecular structure. It is therefore important that they are published (databases for these have already been set up on METLIN) to aid in their future identification and correlation to phenotypes.

Biological interpretation

Network modelling and pathway-mapping tools can help us to understand the parts that metabolites play in relation to each other and in biological aberrations. Thereafter, metabolites can be placed into context with upstream genes and proteins to lead mechanistic investigations[47]. As well as the established and comprehensive metabolic network resources Kegg[95], Recon1 (REF. 34) and Biocyc[96], there are several recently developed programs that use novel methods to find pathway connectivity, as well as aiding in metabolite identification. These include mummichog[46] and metabolite set enrichment analysis (MSEA)[97]. In addition, stable isotope metabolomics[56,57] and omics-scale big data integration can reveal interconnectivity between metabolites and their relationships with genes and proteins (see also main text).

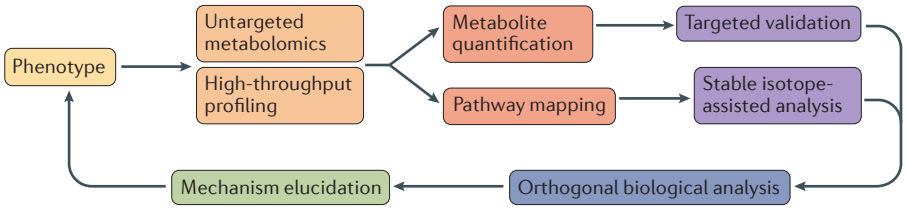


Figure 1 | **From metabolites to pathways and mechanisms.** The workflow outlines a holistic approach that begins with high-throughput untargeted metabolite profiling. Analysis of biofluids, cells or tissues reveals quantitative metabolite changes (as a result of a stimulus) that can be validated further. Metabolites can be mapped and analysed within metabolic pathways to relate the metabolites to each other, and within interconnected biological pathways, providing potential targets for further mechanistic studies. The combination of metabolomic, orthogonal biological analysis and isotope-assisted deciphering of pathways allows the mechanism of the aberrant phenotype to be ascertained.

when paired with advanced metabolomic techniques such as stable isotope tracing and integration with other orthogonal data sets, ultimately providing systems-level analyses (FIG. 1).

Informatics. The development of computational and chemoinformatic tools for metabolomics can effectively support experimental data upload, processing, statistical analyses and metabolite identification, and, when used in conjunction with bioinformatic tools, can place metabolites into biological context (see BOX 2 and Supplementary information S1 (box)). Metabolomic data sets obtained by mass spectrometry (BOX 1) contain information on thousands of ions that are generated in the mass spectrometer from each sample, in which the ions represent the precursor intact metabolite or its fragments, adducts or isotopes. Computational tools are thus essential for reducing the redundancy in these complex data sets and facilitating identification of the most relevant metabolites.

For researchers in the field of metabolomics, computational resources are growing at a rapid rate, and many of these have been discussed in detail elsewhere[33,39]. However, metabolomic analysis remains a time-consuming process, and metabolite identification is still a limiting factor. Therefore, computational workflows that significantly speed up the process of data upload and data mining, with novel methods for automated or *in silico* metabolite identification and biological interpretation, are needed. Such automated computational workflows — allowing data streaming from the instrument to the software, automated qualitative and quantitative metabolite characterization, calculation of fold change and statistical significance, and, importantly, metabolite pathway analysis — have recently been developed (for more detail, see Supplementary information S1 (box)).

As metabolomics is highly interdisciplinary, and not all laboratories have personnel that are specialized in all areas of the experimental workflow, it is often the case that some of these computational tools are out of reach for those not specialized in informatic approaches or new to the metabolomics field. Fortunately, this is beginning to change, with several resources provided through the US National Institutes of Health (NIH) Common Fund Metabolomics Program. This programme funds six regional comprehensive metabolomic resource cores, a data repository and a coordination centre, to enable hands-on and online training in a range of areas, including data processing and interpretation. Another initiative, the Coordination of Standards in Metabolomics (COSMOS), is also helping to promote the standardization of metabolomics, by providing both experimental and data sharing, thus aiding new researchers in the field[40] (see Supplementary information S1 (box)). There are several tools, including the workflows mentioned, that are user-friendly but have advanced parameters for expert users, thus providing a resource for all levels of expertise[41,42]. Some of these are available as part of the mass spectrometry vendor software, whereas other tools are provided as open-access software that can be utilized from data upload through to the metabolite pathway analysis[42,43]. These tools have already been successfully used to correlate single or multiple validated metabolites to a biological aberration. For example, MZmine 2 was used to show the interaction between dietary lipids and gut microbiota for regulating cholesterol metabolism[44], and metabolomic analysis using both XCMS Online and MetaboAnalyst revealed metabolic dysregulation in ischaemic retinopathy[45].

As discussed above, to move from using metabolites as predictive biomarkers to leading mechanistic investigations, the metabolites need to be put into their biological context by identifying their roles in metabolic pathways, their interconnectivity with other metabolites, and their relationships to upstream genes and proteins. Informatics approaches can greatly facilitate these analyses and can help to reveal broad potential metabolite activity across multiple metabolites and pathways[46], and can also provide big data integration across different -omics technologies (see below)[47] such as the systems biology approach recently developed on XCMS Online. As an example, a recent study took advantage of various bioinformatics tools to analyse genetic influences on metabolites in human blood. For this, a network of genetic–metabolic interactions was generated, first using Gaussian graphical models to connect biochemically related metabolites and then connecting metabolites with genetic loci from a genome-wide association study[38]. Novel concepts such as these have maximized the ability to extract important biological information from metabolites.

Stable isotope-assisted metabolomics. One of the most promising ways to ascertain the roles of metabolites in metabolic pathways is to track their utilization with stable isotope tracers. These experiments make use of commercially available metabolites labelled with stable isotopes such as carbon (^{13}C), nitrogen (^{15}N) or deuterium (^{2}H). The design of stable isotope-assisted experiments is based on *a priori* information for a particular metabolite or metabolic pathway of interest; these studies can thus be led by information obtained from untargeted metabolomic analysis (BOX 1).

The results from targeted and/or untargeted metabolomic analysis do not provide information on intracellular metabolic rates and relative pathway activities, and, for example, increased levels of one metabolite can be caused by increased activity of metabolite-producing enzymes or decreased activity of metabolite-consuming enzymes[49]. Following up with stable isotope-labelling experiments provides additional information on how a particular compound (nutrient or substrate) is metabolized with respect to a particular phenotype and can help to identify the pathways that contribute the most to substrate utilization. Thus, stable isotope-assisted tracing of a labelled substrate can reveal its metabolic fate.

There are several ways to carry out a stable isotope-assisted experiment. In metabolic steady state experiments, the measured metabolite pools (or levels) are equilibrated, and fluxes (or conversion rates) are roughly constant[35]. In addition, the labelling enrichment becomes stable over time (from a labelled nutrient into a given metabolite) to reach the isotopic steady state. The interpretation of isotope-enriched data in such conditions can provide information on relative pathway activity, such as the relationship between metabolites, and it also allows quantification of nutrient contributions to the production of different metabolites[49]. By contrast, in kinetic (or dynamic) flux experiments, the system has yet to reach steady state, and flux refers to the *in vivo* velocities of the individual metabolic reactions[35]. Thus, kinetic flux analysis provides dynamic labelling patterns, which allow quantification of metabolite flux when combined with intracellular metabolite concentrations[48,49]. As a notable example, kinetic flux revealed mechanisms for NADPH metabolism, including the contribution of the 10-formyl-tetrahydrofolate pathway to NADPH production[50]. Steady state flux analyses have also contributed to revealing important substrate utilization, with a recent clinical example uncovering selective activation of pyruvate carboxylase over glutaminase 1 in early-stage non-small-cell lung cancer[51].

Stable isotope-assisted metabolomics can be used to calculate flux within a specific set of related pathways — or, on a larger scale, it can encompass multiple metabolites, labelled precursors and pathways. However, such analyses are computationally highly complex for dynamic experiments, leading to a decrease in accuracy[35]. In order to overcome this, algorithms have recently been developed that combine both stable isotope analysis and untargeted metabolomics[52–55]. This technology, called global isotope metabolomics, provides comprehensive differential labelling between two biological conditions, offering further understanding of metabolism at a systems level. Even though untargeted stable isotope metabolomics is a relatively new tool, its value has already been demonstrated in several studies[56,57]. It also provides yet another example of the power of informatics in metabolomic analyses.

Orthogonal approaches for mechanistic studies. Owing to the fact that transcript and protein levels have only a modest correlation with each other, and that metabolites can be further modified by enzymatic processes and can originate from and be modified by various internal and external stimuli, it is necessary to introduce metabolomic analysis approaches that provide big data integration across different -omics (genomics, epigenomics, proteomics and transcriptomics) in order to comprehensively determine the consequences of all metabolites on biology (FIG. 2). Such integrative approaches can help to determine the relationships between gene and protein expression and metabolite concentrations, and the balance between production and consumption of metabolites[58]. As an example, by combining metabolomics with metagenomics and metatranscriptomics data, it was possible to elucidate the origins and roles of bacteria-derived metabolites[59,60]. A recent study also revealed that gut bacteria transplanted from thin or obese people recapitulated the respective phenotypes in gnotobiotic mice, with changes to microbial genes and concomitant downstream metabolites[60].

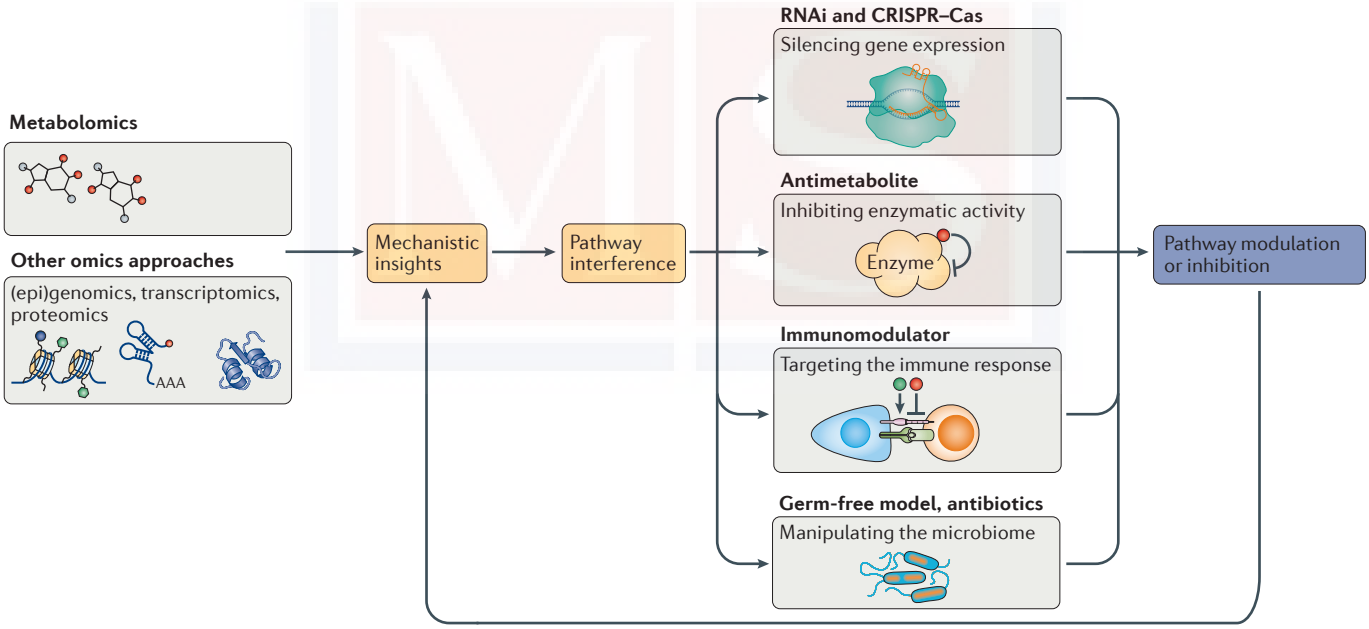


Figure 2 | **Controlling and influencing metabolism: perspectives from metabolomics.** Using various orthogonal techniques, targets identified with metabolomics can be further verified and investigated in more detail. For instance, other 'omics' approaches, including (epi)genomics, transcriptomics and proteomics, can reveal further mechanistic insights into phenotypical changes associated with the metabolite. Various orthogonal techniques also allow targeting of metabolic pathways and can be used to influence metabolite levels and to interfere with metabolic pathways. These approaches can be directed at the gene level and aimed at silencing gene expression, with techniques like CRISPR–Cas-mediated knock outs or RNA interference (RNAi). Alternatively, metabolic pathways can be influenced at at the protein level with the use of antimetabolites. Manipulating sources of exposure to different stimuli can also influence the metabolome, providing further mechanistic insights. For instance, using antibiotics or germ-free models with species-specific inoculation reveals the direct effect of the microbiome on metabolite production. Similarly, immunomodulators can be used to change the efficacy of the host immune system to respond to both the resident microbiota and pathogens, and their metabolic products. This collectively opens up possibilities for better understanding and, eventually, controlling metabolism.

In addition, it was possible to demonstrate that individuals from rural African and African American populations that exchanged diets underwent large changes in their metagenome and metabolome, and this altered their cancer risk[61].

Leading on from multi-layered omics approaches, there are a number of additional orthogonal techniques that can be used to further investigate the biological relationships between metabolites, proteins and genes (FIG. 2). At the gene level, RNA interference (RNAi) or CRISPR–Cas systems can be used to modulate gene expression, and this can help to determine how genes directly affect enzyme activity and metabolite production. Similarly, at the protein level, structural analogues of essential metabolites — so-called antimetabolites — can be used to inhibit a specific metabolic process and attenuate metabolite production or transportation from the cell[62], thereby allowing investigation of the function and importance of specific metabolites[63]. Other approaches that can be used are those that directly change the host metabolome, for instance, through modulating the exposure of the organism to certain stimuli. For example, manipulating the microbiome using germ-free models, antibiotics or immunomodulators (which can change the host response to the resident microbiota) can reveal how bacteria and their metabolites affect the host and their metabolism and can allow us to link these changes to susceptibility to certain diseases[60]. As an example, it has been shown recently that the microbiome is important for the efficacy of immunotherapeutics used in cancer therapy, and that only in individuals harbouring certain bacterial species can these compounds lead to efficient stimulation of cancer-fighting T cells[64,65]. Of note, T cells are known to have distinct energy requirements depending on their activation status, with naive T cells utilizing oxidative phosphorylation for ATP generation, and effector (activated) T cells consuming glucose by aerobic glycolysis and glutaminolysis to support cell growth, in a similar manner to cancer cells[66,67]. Altogether, targeted manipulation of the local cellular environment to affect cellular energy status, in concert with modulation of the microbiome, opens up interesting possibilities to influence the survival of both effector T cells and cancer cells[68].

Novel biological insights

Advances in metabolomic analysis have allowed us to gain a novel understanding of metabolism for various states, processes and diseases, and a few of the most recent studies exemplifying the novel biological insights that can be gained with the use of metabolomics are discussed below. These studies collectively show how information at the metabolite level, particularly when combined with other techniques, can lead to successful association of metabolites with phenotypical causality, thus bringing us closer to a mechanistic understanding of metabolism.

Role of bacterial biofilms in cancer. A recent study carried out on a patient population investigated in more detail a previously validated biomarker for colon cancer, N^1, N^{12}-diacetylspermine (DAS)[69]. In this study, a multidisciplinary approach was used that combined four different metabolomic tools with traditional biochemical techniques. First, it revealed that only DAS, and not its precursors, was correlated with biofilm presence as well as with colon cancer, and that DAS is probably a metabolic end-product of polyamine metabolism. The metabolomic approaches used included untargeted analysis (BOX 1) to compare normal tissues to the tumour tissues, both of which were either associated with or devoid of biofilms. This was followed by a targeted validation step (BOX 1) to confirm the fold change in metabolites and expand the analysis to other metabolites in related pathways. Nanostructure-imaging mass spectrometry (NIMS)[70] (BOX 1) revealed the *in situ* localization of DAS in the mucosal layer of the colon where the biofilms resided. Global isotope metabolomics was further used to investigate the metabolic fate of a stable isotope of DAS in colon cancer cell lines, confirming that it is indeed an end-product of metabolism and is not involved in any other metabolic pathways.

In order to determine the source of the metabolite (the patient versus the biofilm), patients were treated with antibiotics to remove the biofilms (this was confirmed by fluorescent *in situ* hybridization (FISH) analysis), and their samples were analysed for the presence of DAS. In these tissues, DAS concentrations were similar to those previously measured in biofilm-negative patients, showing that the elevated DAS levels seen in biofilm-positive patients originated from the biofilms. In line with this, immunohistochemical analysis of patient samples did not show any change in protein levels of enzymes involved in DAS production. As DAS is a metabolite of polyamine precursors, and polyamines have been associated with various cellular responses including increased cellular proliferation, the propensity of colon cells to overproliferate in the presence of biofilms was investigated and confirmed by immunohistochemistry. In addition, immunofluorescence revealed the presence of pro-inflammatory cytokines in biofilm-covered tissues. This inflammatory state was observed in normal-looking tissues that were associated with biofilms, suggesting that such tissues might be in a pro-carcinogenic state and that biofilm formation indeed promotes colon tumorigenesis[71]. In sum, this example shows how a combination of several metabolomic approaches with orthogonal biological techniques can be used for the initial metabolite discovery, leading to the elucidation of the potential role of biofilms in colon carcinogenesis (FIG. 3). According to this study, colonic bacteria utilize polyamines to build biofilms (producing DAS), and this biofilm formation induces pro-inflammatory and pro-carcinogenic effects in the host tissues, increasing the risk of tumour formation. Interestingly, some metabolomic studies have associated DAS with other cancers, including cancers of the lung[72], breast[73], blood[74] and bladder[75], as well as identifying it as a dietary metabolite[76]. Thus, further studies assessing the roles of diet and bacteria in cancers are of the utmost importance.

Metabolic regulation of cell pluripotency. At the epigenetic level, metabolites have been shown to regulate pluripotency in human ES cells, with a recent study revealing a metabolic switch during the transition between human naive and primed ES cells[2]. It has been found that this switch is regulated by nicotinamide *N*-methyltransferase (NNMT), which controls SAM levels that are required for histone methylation. Analysis of oxygen consumption rates revealed that primed human ES cells have a lower mitochondrial respiration capacity than naive human ES cells, and transcriptomic analysis confirmed a downregulation of mitochondrial electron transport chain genes in the primed state. The transition from naive to primed human ES cells also involved reduced WNT signalling and increased hypoxia-inducible factor 1α (HIF1α) stabilization (shown by proteomic analysis). Untargeted and targeted metabolomics based on gas chromatography and liquid chromatography mass spectrometry (GC–MS and LC–MS) (BOX 1) revealed concomitant changes in metabolic pathways, including glycolysis, fatty acid β-oxidation and lipid biosynthesis. Transcriptomic and genomic analyses showed that the genes involved in these pathways were also changed. The use

of WNT inhibitors and the generation of HIF1α-knockout cells by CRISPR–Cas gene editing further demonstrated that WNT activity is required for the naive state, and that HIF1α is required for human ES cell transition to the primed state. Furthermore, the loss of NNMT in naive human ES cells was associated with an increase in repressive histone marks (histone 3 Lys27 trimethylation; H3K27me3) in developmental and metabolic genes that regulate the metabolic switch in naive to primed cells. Collectively, this comprehensive analysis showed that both NNMT and the metabolic state regulate ES cell development.

Novel therapy for cardiovascular disease. Another example shows how using metabolomics, together with other techniques, can lead to the establishment of a new therapeutic approach — in this case, for decreasing the risk of CVD. Initially, using untargeted metabolomics and then targeted metabolomics for validation and quantification (BOX 1), an association between an increased risk of CVD and plasma concentrations of choline, betaine and TMAO was established[19,77,78]. This was further replicated in apolipoprotein $E^{-/-}$ mice, a mouse model that is highly susceptible to the formation of atherosclerotic plaques — the primary cause of CVD — that were fed high-choline and high-TMAO diets, showing a significant correlation between plasma TMAO and the formation of atherosclerotic plaques. Functional experiments revealed that trimethylamine (TMA)-containing nutrients such as choline, phosphatidylcholine and carnitine are dietary precursors for TMAO, and that liver flavin monooxygenases (FMOs; primarily FMO3) are responsible for converting TMA to TMAO. Analysis of antibiotic-treated mice, together with the observation that the risk of CVD was transmissible upon microbial transfer, led to the conclusion that the microbiome generates TMA. As inhibition of FMO3 can produce side-effects and thus does not provide a sustainable therapy, the next step was to search for an inhibitor of microbial TMA production and investigate its potential as a therapeutic for CVD. Using a structural analogue to choline, 3,3-dimethyl-1-butanol (DMB), found in extra-virgin olive oil, it was possible to inhibit microbial TMA lyases, which are responsible for TMA formation. *In vivo* experiments showed that TMAO levels were indeed reduced in mice fed with high-choline or high-carnitine diets when these mice were simultaneously treated with DMB. Treatment with DMB also prevented atherosclerotic lesion development in apolipoprotein $E^{-/-}$ mice on a choline-enhanced diet[79]. Altogether, this work led to the proposal of a novel therapy for CVD, which bypasses the issues that arise when using inhibitors targeted to a patient's own proteins — an approach potentially resulting in various side-effects for the patient. Instead, this study showed that harmful metabolites can be inhibited at their earliest production, by 'drugging' the gut microbiome, which in the case of CVD is the source of the metabolite contributing to the disease.

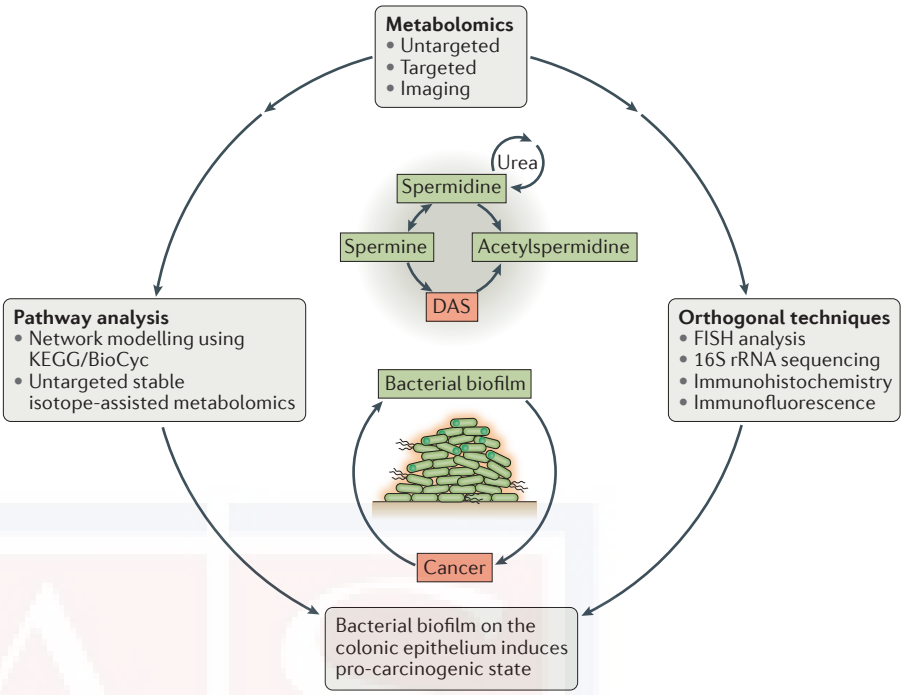


Figure 3 | **Novel biological insights**. Diacetylspermine (DAS) has a role in biofilm-associated colon cancer. Various metabolomic and orthogonal biological techniques contributed to the association of DAS with bacterial biofilms and their role in the pathology of cancer. Fluorescence *in situ* hybridization (FISH) analysis and 16S rRNA sequencing identified the presence of bacterial species and biofilms on colon tissues. Untargeted and targeted metabolomics identified and validated the association of polyamine metabolites with colon cancer tissues. Stratification by biofilm status showed that DAS was upregulated primarily in biofilm-associated tissues, which was confirmed by mass spectrometry imaging. Network modelling using the KEGG and BioCyc databases, and pathway analysis using untargeted stable-isotope assisted metabolomics, showed that DAS is an end-product of polyamine metabolism. For further analysis, orthogonal techniques were used. Immunohistochemistry and immunofluorescence revealed increased cellular proliferation and pro-inflammatory cytokines in biofilm-associated tissues. The combination of these techniques led to the conclusion that bacterial biofilms induce a pro-carcinogenic state in the colon epithelium.

Metabolite-driven regulation of β-cells. An important metabolite, 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid (CMPF), was recently identified in the plasma of humans with gestational diabetes, as well as in those with impaired glucose tolerance and type 2 diabetes[80]. CMPF was identified by untargeted and targeted metabolomic analysis (BOX 1), with further validation by enzyme-linked immunosorbent assay (ELISA).

Mice treated with CMPF at doses comparable to levels found in human individuals with diabetes developed glucose intolerance and impaired insulin secretion after an oral glucose-tolerance test. This was monitored using targeted mass spectrometry and ELISA to measure plasma and tissue CMPF concentrations, and also by glucose-stimulated insulin secretion (GSIS) tests. Mechanistically, CMPF was shown to impair mitochondrial function, decrease glucose-induced ATP synthesis and induce oxidative stress, as assessed by measuring mitochondrial membrane potential and with fluorescence- and bioluminescence-based assays, as well as gene expression analysis. Inhibitors of organic anion transporters (OAT), which are responsible for the clearance of CMPF, blocked the transportation of CMPF into β-cells of the pancreas and prevented β-cell dysfunction. In line with this, treatment of pancreatic islets isolated from OAT3-knockout mouse models with CMPF had no effect on insulin content

or GSIS. Altogether, the metabolite CMPF, identified by metabolomic analysis, provides a mechanistic link between β-cell dysfunction and diabetes and has been shown to function through impairing mitochondrial function and inhibiting insulin biosynthesis.

Mechanism of ischaemia–reperfusion injury. Steady state flux analysis was recently used to help to identify the mechanisms of ischaemia–reperfusion injury, which is a type of tissue damage resulting from oxidative stress and generation of reactive oxygen species (ROS) following the return of circulation to tissue regions previously deprived of oxygen. It was revealed that succinate, which is a metabolite of the tricarboxylic acid (TCA) cycle, is the driver of ROS generation, which can lead to heart attack and stroke following ischaemia–reperfusion injury[81]. The authors also used a combination of untargeted and targeted metabolomics (BOX 1) to reveal an elevation of succinate levels across several organs in a mouse model of ischaemia. Mechanistic studies involving *in silico* modelling, mitochondrial membrane potential measurements, ratiometric assessment and fluorescence assays revealed that in ischaemia, succinate dehydrogenase (SDH) functions in reverse, accumulating succinate from fumarate. Upon reperfusion, succinate is oxidized and drives electrons back through the mitochondrial complex I, thus generating ROS. Together, these findings indicated that SDH could be a target for the prevention of ROS accumulation following reperfusion of ischaemic tissue. Accordingly, antimetabolite inhibitors of SDH prevented succinate accumulation, inhibiting electron flow through complex I and subsequent ROS production, and thereby providing protection from ischaemia–reperfusion injury.

Regulation of cancer cell metabolism. In addition to the previous example, metabolic flux analysis was recently used to investigate the role of mitochondrial enzyme serine hydroxymethyltransferase (SHMT2) in human glioblastoma cells. Specifically, the roles of SHMT2 in central carbon metabolism and in regulating pyruvate kinase M2 (PKM2) activity were investigated and were further linked to glioma cell survival[82]. In these experiments, SHMT2-knockdown cells were treated with uniformly labelled ^{13}C-glucose and showed increased flux from pyruvate to lactate, citrate and alanine, with a concomitant increase in PKM2 activity and oxygen consumption rate. In addition, overexpression of RNAi-resistant SHMT2 cDNA reverted these effects, confirming that SHMT2 negatively affects PKM2. Thus, the stable isotope analysis showed that SHMT2 expression changes the metabolism of cancer cells and limits carbon flux into the TCA cycle via suppression of PKM2. This has been further shown to improve the survival of cells in ischaemic tumour regions. In addition, the study showed that the survival of cancer cells with high SHMT2 expression can be impaired if glycine decarboxylase is inhibited, as this causes accumulation of glycine, which then contributes to the production of toxic metabolites. Altogether, this series of experiments provided novel insights into cancer cell metabolism and demonstrated how metabolic changes can affect cell properties and responses — in this case, cell survival.

Future perspectives

Metabolomics is an exciting and evolving research area, with numerous success stories demonstrating that its power extends from biomarker discovery to understanding the mechanisms that underlie phenotypes. This step towards mechanistic understanding has been made possible by advances in analytical technologies and informatics, and the combination of these tools has generated novel insights into chemical physiology. It has also been made possible as metabolomics has become more widely used in combination with orthogonal technologies, such as genomics, proteomics, structural biology and imaging, as well as with various techniques that allow us to modify gene expression, enzymatic activity, cell signalling or whole metabolic pathways, including the contribution of the naturally occurring microbiota. Thus, the future prospects of metabolomics lie not only in the unique information it provides, but in its integration into systems biology.

Caroline Johnson is at the Department of Environmental Health Sciences, Yale School of Public Health, Yale University, 60 College Street, New Haven, Connecticut 06520, USA.

Julijana Ivanisevic is at the Metabolomics Research Platform, Faculty of Biology and Medicine, University of Lausanne, Rue du Bugnon 19, 1005 Lausanne, Switzerland.

Gary Siuzdak is at the Scripps Center for Metabolomics and Mass Spectrometry, Departments of Chemistry, Molecular and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

Correspondence to G.S.
siuzdak@scripps.edu

doi:10.1038/nrm.2016.25
Published online 16 March 2016

1. Johnson, C. H., Patterson, A. D., Idle, J. R. & Gonzalez, F. J. Xenobiotic metabolomics: major impact on the metabolome. *Annu. Rev. Pharmacol. Toxicol.* **52**, 37–56 (2012).
2. Sperber, H. *et al.* The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. *Nat. Cell Biol.* **17**, 1523–1535 (2015).
3. Yanes, O. *et al.* Metabolic oxidation regulates embryonic stem cell differentiation. *Nat. Chem. Biol.* **6**, 411–417 (2010).
4. Karlic, H. *et al.* Inhibition of the mevalonate pathway affects epigenetic regulation in cancer cells. *Cancer Genet.* **208**, 241–252 (2015).
5. Carey, B. W., Finley, L. W., Cross, J. R., Allis, C. D. & Thompson, C. B. Intracellular α-ketoglutarate maintains the pluripotency of embryonic stem cells. *Nature* **518**, 413–416 (2015).
6. Ulanovskaya, O. A., Zuhl, A. M. & Cravatt, B. F. NNMT promotes epigenetic remodeling in cancer by creating a metabolic methylation sink. *Nat. Chem. Biol.* **9**, 300–306 (2013).
7. Wellen, K. E. *et al.* ATP-citrate lyase links cellular metabolism to histone acetylation. *Science* **324**, 1076–1080 (2009).
8. Nakahata, Y. *et al.* The NAD$^+$-dependent deacetylase SIRT1 modulates CLOCK-mediated chromatin remodeling and circadian control. *Cell* **134**, 329–340 (2008).
9. Gornall, A. G. (ed) *Applied Biochemistry of Clinical Disorders* (Lippincott Williams & Wilkins, 1986).
10. Richieri, G. V. & Kleinfeld, A. M. Unbound free fatty-acid levels in human serum. *J. Lipid Res.* **36**, 229–240 (1995).
11. Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M. & Snyder, M. Extensive *in vivo* metabolite-protein interactions revealed by large-scale systematic analyses. *Cell* **143**, 639–650 (2010).
12. Hubbard, T. D. *et al.* Adaptation of the human aryl hydrocarbon receptor to sense microbiota-derived indoles. *Sci. Rep.* **5**, 12689 (2015).
13. Sharma, M., Astekar, M., Soi, S., Manjunatha, B. S. & Shetty, D. C. pH gradient reversal: an emerging hallmark of cancers. *Recent Pat. Anticancer Drug Discov.* **10**, 244–258 (2015).
14. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661–672 (2014).
15. Brahimi-Horn, M. C., Laferriere, J., Mazure, N. & Pouyssegur, J. in *Tumor Angiogenesis: Basic Mechanisms and Cancer Therapy* (eds Marme, D. & Fusenig, N.) 186 (Springer-Verlag Berlin Heidelberg, 2008).
16. Ivanisevic, J. *et al.* Arteriovenous blood metabolomics: a readout of intra-tissue metabostasis. *Sci. Rep.* **5**, 12757 (2015).
17. Ivanisevic, J. *et al.* Brain region mapping using global metabolomics. *Chem. Biol.* **21**, 1575–1584 (2014).
18. Johnson, C. H. & Gonzalez, F. J. Challenges and opportunities of metabolomics. *J. Cell. Physiol.* **227**, 2975–2981 (2012).
19. Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576–585 (2013).
20. Pannkuk, E. L., Laiakis, E. C., Authier, S., Wong, K. & Fornace, A. J. Jr. Global metabolomic identification of long-term dose-dependent urinary biomarkers in nonhuman primates exposed to ionizing radiation. *Radiat. Res.* **184**, 121–133 (2015).
21. Johnson, C. H. *et al.* Radiation metabolomics. 5. identification of urinary biomarkers of ionizing radiation exposure in nonhuman primates by mass spectrometry-based metabolomics. *Radiat. Res.* **178**, 328–340 (2012).
22. Johnson, C. H. *et al.* Radiation metabolomics. 4. UPLC-ESI-QTOFMS-based metabolomics for urinary biomarker discovery in gamma-irradiated rats. *Radiat. Res.* **175**, 473–484 (2011).
23. Hamdalla, M. A., Ammar, R. A. & Rajasekaran, S. A molecular structure matching approach to efficient identification of endogenous mammalian biochemical structures. *BMC Bioinformatics* **16**, S11 (2015).
24. Wolf, S., Schmidt, S., Muller-Hannemann, M. & Neumann, S. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11**, 148 (2010).
25. Ridder, L. *et al.* Automatic chemical structure annotation of an LC-MSn based metabolic profile from green tea. *Anal. Chem.* **85**, 6033–6040 (2013).

26. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *Trends Analyt. Chem.* http://dx.doi.org/10.1016/j.trac.2015.09.005 (2015).
27. Rocca-Serra, P. *et al.* Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **12**, 14 (2016).
28. Ellis, J. K. *et al.* Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. *BMC Med.* **10**, 61 (2012).
29. Scalbert, A. *et al.* The food metabolome: a window over dietary exposure. *Am. J. Clin. Nutr.* **99**, 1286–1308 (2014).
30. Wishart, D. S. *et al.* HMDB 3.0 — the human metabolome database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
31. Ji, Y. *et al.* Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics. *Clin. Pharmacol. Ther.* **89**, 97–104 (2011).
32. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**, 910–914 (2009).
33. Johnson, C. H., Ivanisevic, J., Benton, H. P. & Siuzdak, G. Bioinformatics: the next frontier of metabolomics. *Anal. Chem.* **87**, 147–156 (2015).
34. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425 (2013).
35. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the metabolome: size, flux, and regulation. *Mol. Cell* **58**, 699–706 (2015).
36. Mathe, E. A. *et al.* Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res.* **74**, 3259–3270 (2014).
37. Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158**, 1415–1430 (2014).
38. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
39. Misra, B. B. & van der Hooft, J. J. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **37**, 86–110 (2016).
40. Salek, R. M. *et al.* COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics* **11**, 1587–1597 (2015).
41. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
42. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84**, 5035–5039 (2012).
43. Xia, J. G., Psychogios, N., Young, N. & Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* **37**, W652–W660 (2009).
44. Caesar, R., Nygren, H., Oresic, M. & Backhed, F. Interaction between dietary lipids and gut microbiota regulates hepatic cholesterol metabolism. *J. Lipid Res.* http://dx.doi.org/10.1194/jlr.M065847 (2016).
45. Paris, L. P. *et al.* Global metabolomics reveals metabolic dysregulation in ischemic retinopathy. *Metabolomics* **12**, 15 (2016).
46. Li, S. *et al.* Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **9**, e1003123 (2013).
47. Cottret, L. *et al.* MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.* **38**, W132–W137 (2010).
48. Metallo, C. M. & Vander Heiden, M. G. Understanding metabolic regulation and its influence on cell physiology. *Mol. Cell* **49**, 388–398 (2013).
49. Buescher, J. M. *et al.* A roadmap for interpreting ^{13}C metabolite labeling patterns from cells. *Curr. Opin. Biotechnol.* **34**, 189–201 (2015).
50. Fan, J. *et al.* Quantitative flux analysis reveals folate-dependent NADPH production. *Nature* **510**, 298–302 (2014).
51. Sellers, K. *et al.* Pyruvate carboxylase is critical for non-small-cell lung cancer proliferation. *J. Clin. Invest.* **125**, 687–698 (2015).
52. Huang, X. *et al.* X13CMS: global tracking of isotopic labels in untargeted metabolomics. *Anal. Chem.* **86**, 1632–1639 (2014).
53. Bueschl, C. *et al.* A novel stable isotope labelling assisted workflow for improved untargeted LC-HRMS based metabolomics research. *Metabolomics* **10**, 754–769 (2014).
54. Creek, D. J. *et al.* Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Anal. Chem.* **84**, 8442–8447 (2012).
55. Capellades, J. *et al.* geoRge: a computational tool to detect the presence of stable isotope labeling in LC/MS-based untargeted metabolomics. *Anal. Chem.* **88**, 621–628 (2016).
56. Chen, Y. J. *et al.* Differential incorporation of glucose into biomass during Warburg metabolism. *Biochemistry* **53**, 4755–4757 (2014).
57. Creek, D. J. *et al.* Probing the metabolic network in bloodstream-form *Trypanosoma brucei* using untargeted metabolomics with stable isotope labelled glucose. *PLoS Pathog.* **11**, e1004689 (2015).
58. Zelezniak, A., Sheridan, S. & Patil, K. R. Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. *PLoS Comput. Biol.* **10**, e1003572 (2014).
59. Hsiao, E. Y. *et al.* Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451–1463 (2013).
60. Ridaura, V. K. *et al.* Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
61. O'Keefe, S. J. *et al.* Fat, fibre and cancer risk in African Americans and rural Africans. *Nat. Commun.* **6**, 6342 (2015).
62. Woolley, D. W. *A Study of Antimetabolites* (John Wiley & Sons, 1952).
63. Johnson, C. H. *et al.* Alterations in spinal cord metabolism during treatment of neuropathic pain. *J. Neuroimmune Pharmacol.* **10**, 396–401 (2015).
64. Sivan, A. *et al.* Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).
65. Vetizou, M. *et al.* Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
66. Frauwirth, K. A. & Thompson, C. B. Regulation of T lymphocyte metabolism. *J. Immunol.* **172**, 4661–4665 (2004).
67. van Stipdonk, M. J. B. *et al.* Dynamic programming of CD8$^+$ T lymphocyte responses. *Nat. Immunol.* **4**, 361–365 (2003).
68. Mockler, M. B., Conroy, M. J. & Lysaght, J. Targeting T cell immunometabolism for cancer immunotherapy; understanding the impact of the tumor microenvironment. *Front. Oncol.* **4**, 107 (2014).
69. Johnson, C. H. *et al.* Metabolism links bacterial biofilms and colon carcinogenesis. *Cell Metab.* **21**, 891–897 (2015).
70. Northen, T. R. *et al.* Clathrate nanostructures for mass spectrometry. *Nature* **449**, 1033–1036 (2007).
71. Dejea, C. M. *et al.* Microbiota organization is a distinct feature of proximal colorectal cancers. *PNAS* **111**,18321–18326 (2014).
72. Wikoff, W. R. *et al.* Diacetylspermine is a novel prediagnostic serum biomarker for non-small-cell lung cancer and has additive performance with pro-surfactant protein B. *J. Clin. Oncol.* **33**, 3880–3886 (2015).
73. Umemori, Y. *et al.* Evaluating the utility of N^1,N^{12}-diacetylspermine and N^1,N^8-diacetylspermidine in urine as tumor markers for breast and colorectal cancers. *Clin. Chim. Acta* **411**, 1894–1899 (2010).
74. Lee, S. H., Suh, J. W., Chung, B. C. & Kim, S. O. Polyamine profiles in the urine of patients with leukemia. *Cancer Lett.* **122**, 1–8 (1998).
75. Stejskal, D. *et al.* Evaluation of urine *N*1,*N*12-diacetylspermine as potential tumor marker for urinary bladder cancer. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech Repub.* **150**, 235–237 (2006).
76. Vargas, A. J., Ashbeck, E. L., Thomson, C. A., Gerner, E. W. & Thompson, P. A. Dietary polyamine intake and polyamines measured in urine. *Nutr. Cancer* **66**, 1144–1153 (2014).
77. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011).
78. Tang, W. H. *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368**, 1575–1584 (2013).
79. Wang, Z. *et al.* Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* **163**, 1585–1595 (2015).
80. Prentice, K. J. *et al.* The furan fatty acid metabolite CMPF is elevated in diabetes and induces β cell dysfunction. *Cell Metab.* **19**, 653–666 (2014).
81. Chouchani, E. T. *et al.* Ischaemic accumulation of succinate controls reperfusion injury through mitochondrial ROS. *Nature* **515**, 431–435 (2014).
82. Kim, D. *et al.* SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance. *Nature* **520**, 363–367 (2015).
83. Siuzdak, G. *The Expanding Role of Mass Spectrometry in Biotechnology* (MCC Press, 2006).
84. Tanaka, K. *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2**, 151–153 (1988).
85. Siuzdak, G. E., Buriak, J. & Wei, J. Desorption/ionization of analytes from porous light-absorbing semiconductor. US Patent 6288390 B1 (2000).
86. Wiseman, J. M., Ifa, D. R., Song, Q. & Cooks, R. G. Tissue imaging at atmospheric pressure using desorption electrospray ionization (DESI) mass spectrometry. *Angew. Chem. Int. Ed. Engl.* **45**, 7188–7192 (2006).
87. Kraft, M. L., Weber, P. K., Longo, M. L., Hutcheon, I. D. & Boxer, S. G. Phase separation of lipid membranes analyzed with high-resolution secondary ion mass spectrometry. *Science* **313**, 1948–1951 (2006).
88. Gowda, H. *et al.* Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal. Chem.* **86**, 6931–6939 (2014).
89. Smith, C. A. *et al.* METLIN — a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747–751 (2005).
90. Wishart, D. S. *et al.* HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
91. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
92. Kopka, J. *et al.* GMD@CSB. DB: the Golm Metabolome Database. *Bioinformatics* **21**, 1635–1638 (2005).
93. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **48**, 291–298 (2013).
94. Paglia, G. *et al.* Ion mobility derived collision cross sections to support metabolomics applications. *Anal. Chem.* **86**, 3985–3993 (2014).
95. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).
96. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
97. Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **38**, W71–W77 (2010).

Acknowledgements
The authors would like to thank Nadine Levin at UCLA for her comments on the manuscript. Funding for this work was supported by US National Institutes of Health (NIH) grants R01 GM114368 and PO1 A1043376-02S1.

Competing interests statement
The authors declare no competing interests.

DATABASES
FooDB: http://foodb.ca/
HMDB: http://www.hmdb.ca/

FURTHER INFORMATION
Common Fund Metabolomics Program: https://commonfund.nih.gov/metabolomics
Coordination of Standards in Metabolomics (COSMOS): http://www.cosmos-fp7.eu/
DeviumWeb: https://github.com/dgrapov/DeviumWeb
MetaboAnalyst: http://www.metaboanalyst.ca/
MZmine 2: http://mzmine.github.io/
XCMS Online: https://xcmsonline.scripps.edu/

SUPPLEMENTARY INFORMATION
See online article: S1 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF