FORUM LABO LYON

FROM RESEARCH TO INDUSTRY

cea

METABOHUB

4 Wm Workflow4metabolomics

ifb

PhenoMeNal

www.cea.fr

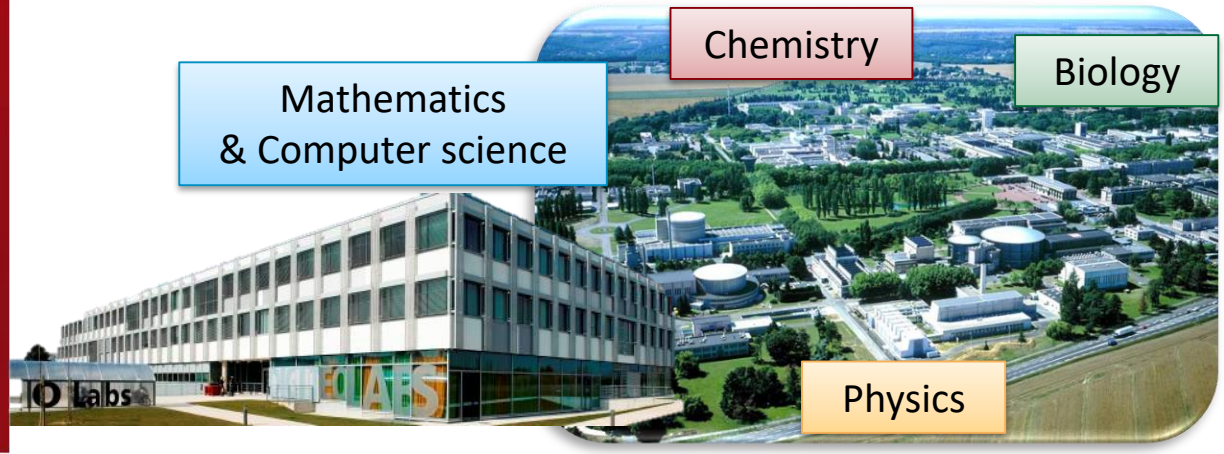
Processing and analysis of metabolomics data

Etienne Thévenot *et al.*

CEA, LIST (Saclay, France)
Laboratory for Data Analysis and Systems' Intelligence
MetaboHUB

etienne.thevenot@cea.fr

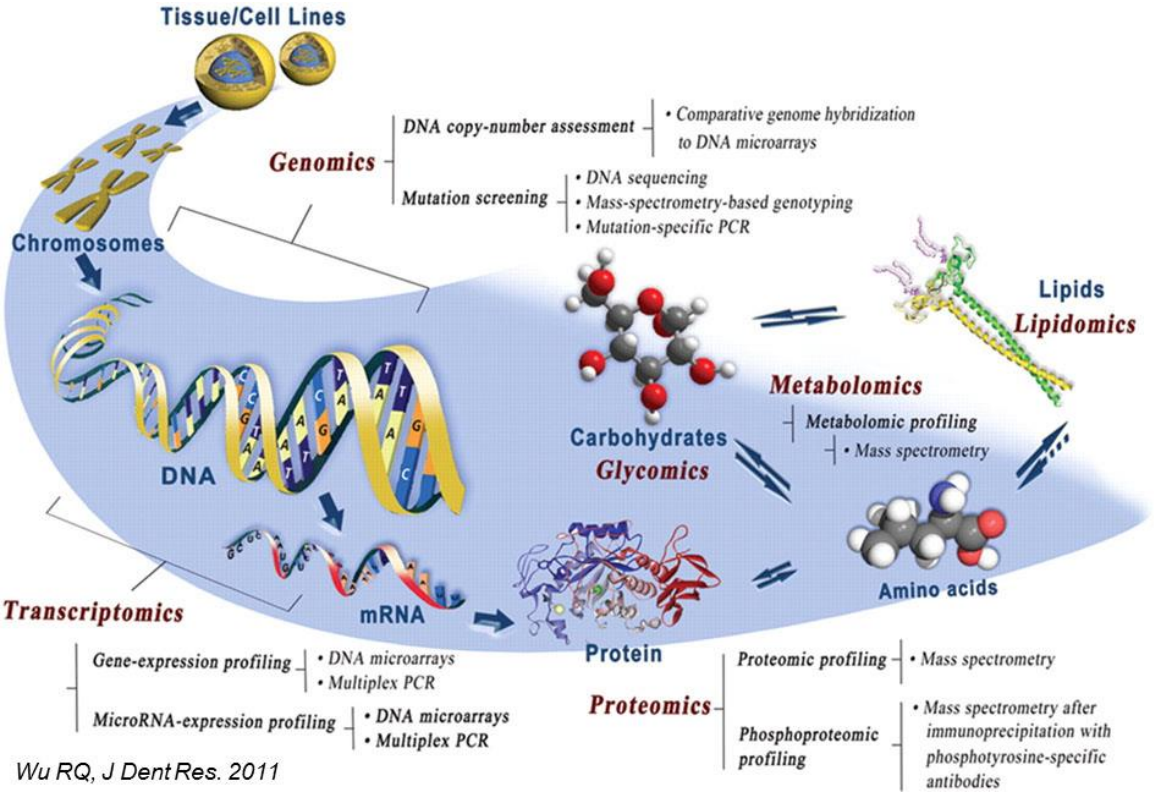
http://etiennethevenot.pagesperso-orange.fr/



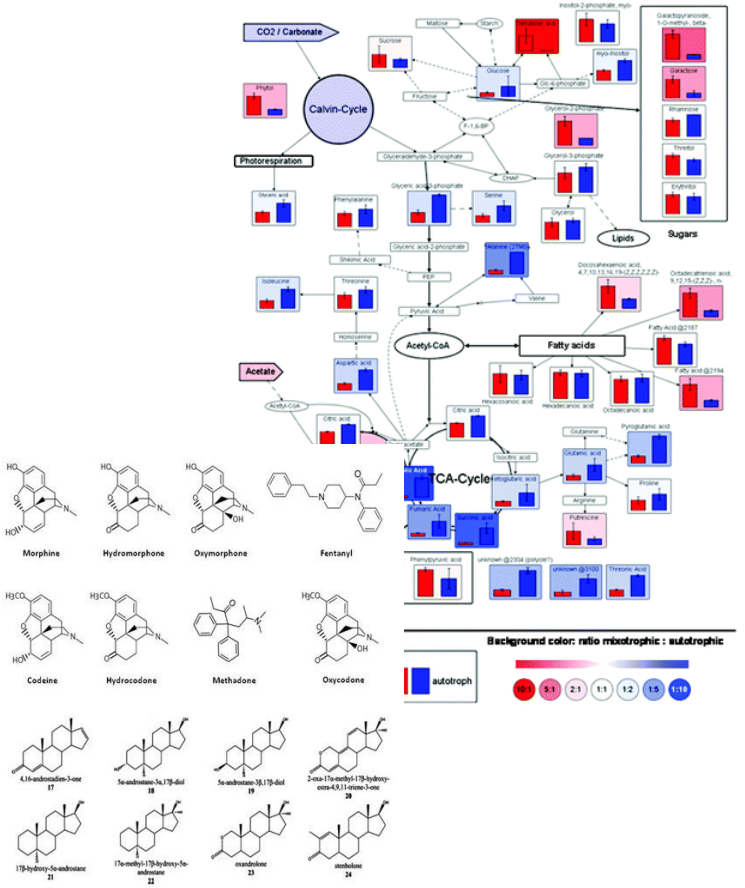
- **Computational metabolomics**
- Preprocessing
- Statistical analysis
- Annotation
- Workflow management
- Bridging proteomics & metabolomics

Metabolomics

- omics science
- dedicated to small molecules (< 1kDa)
- involved in metabolic chemical reactions

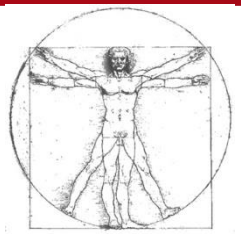


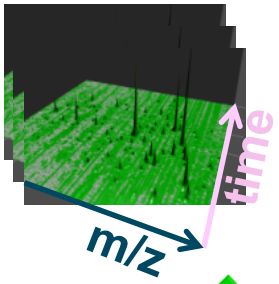
Wu RQ, J Dent Res. 2011

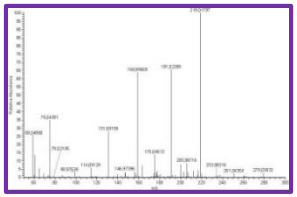


Computational steps

Biological question







n raw data

Preprocessing

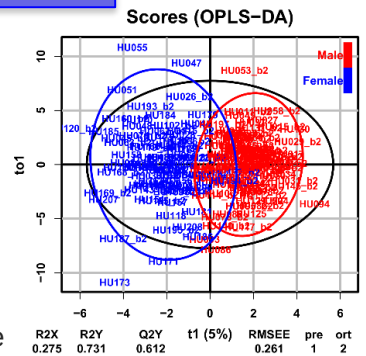
n samples

p variables

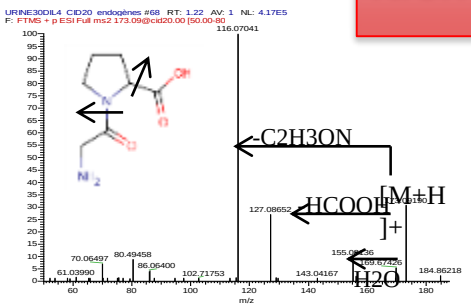
mz	rt	Db_015	...	Db_068
75.0322	41.28	22162	...	48575
75.0441	174.83	1371	...	820
75.0634	56.23	49111	...	91769
...	...	...	...	...
999.6653	844.61	571	...	636
999.6759	844.61	711	...	665
999.6865	844.61	698	...	612

1 peak table

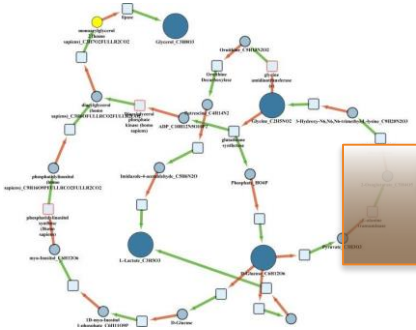
Statistics



Identification



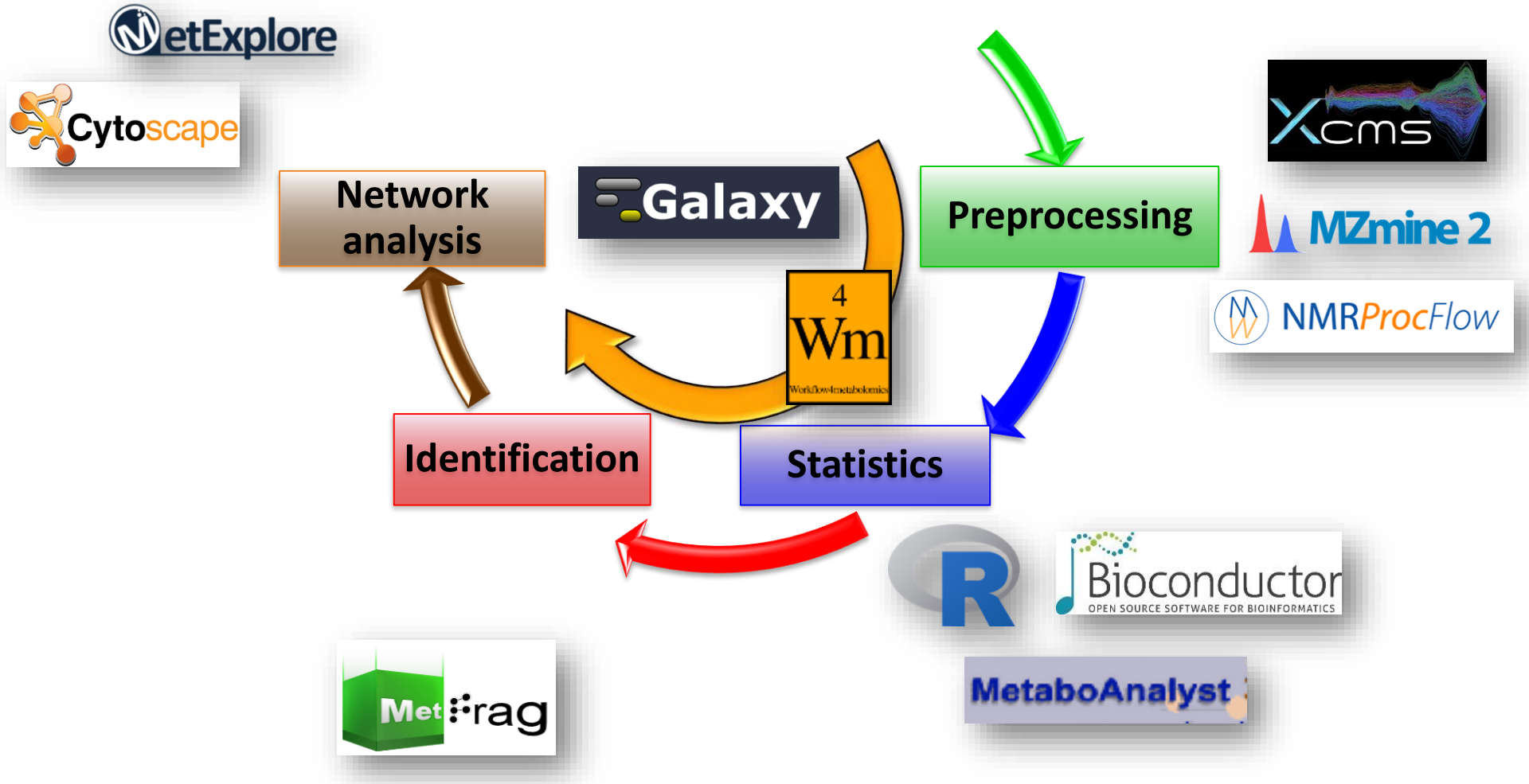
Chemical structures



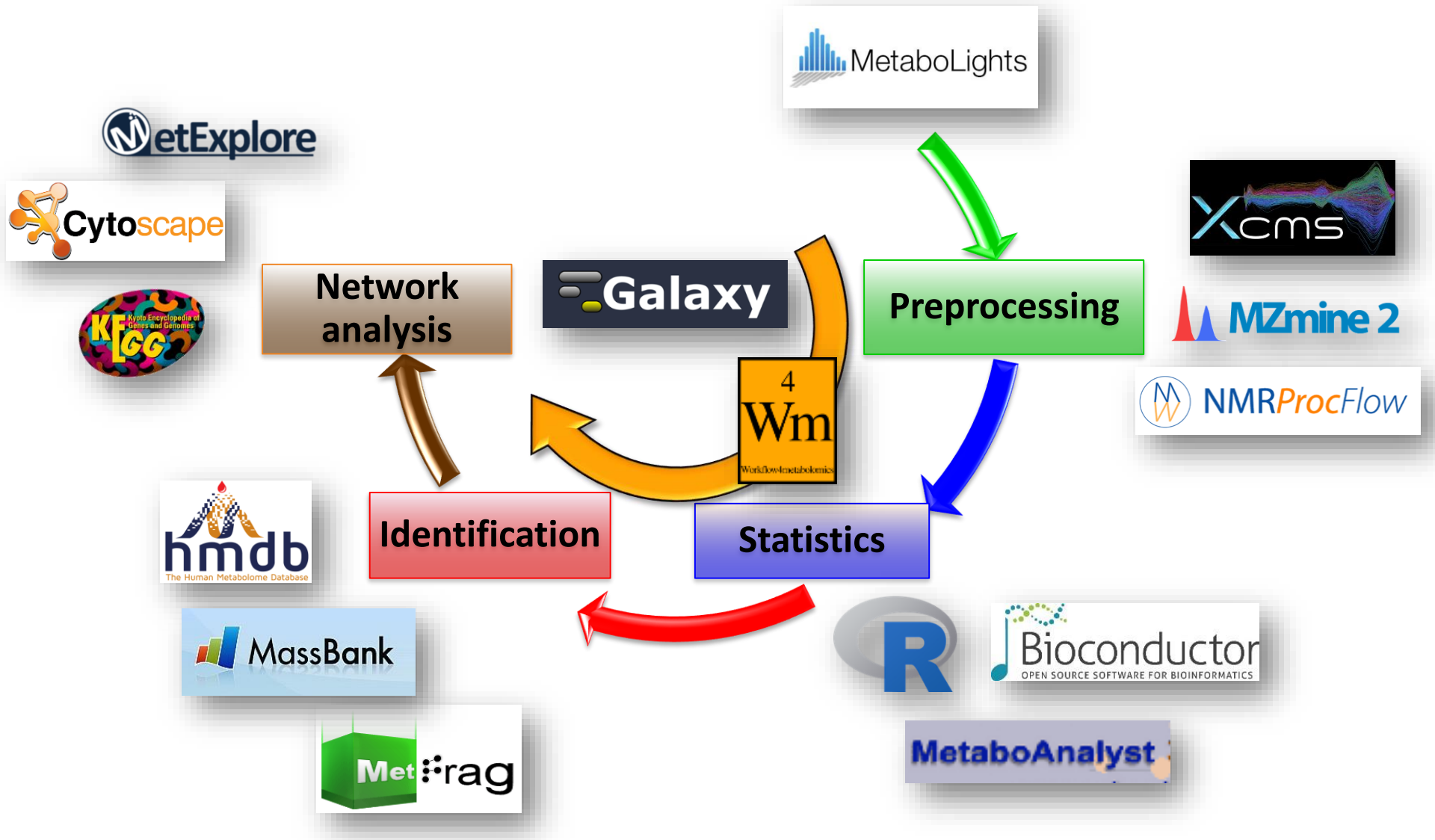
Biological pathways

Network analysis

Many software tools



Many software tools and databases



Computational metabolomics

➢ **Preprocessing**

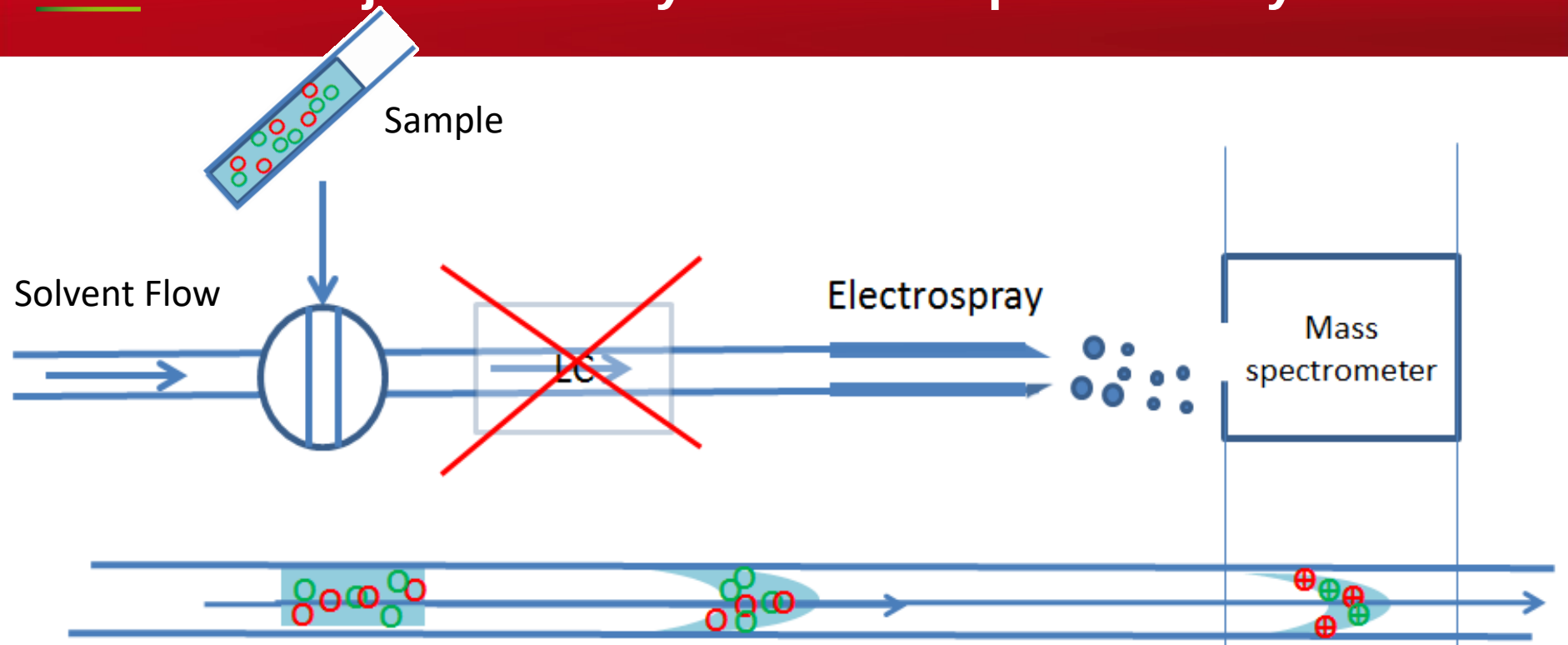
Statistical analysis

Annotation

Workflow management

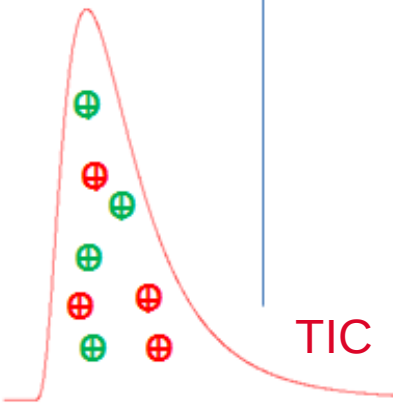
Bridging proteomics & metabolomics

Flow injection analysis - mass spectrometry



	Pro	Cons
LC-MS	• Sensitive • Wide dynamic range	• Slow (10-60 min) • Expensive and difficult chromatography
FIA-MS	• Fast (1-3 min)	• Matrix effect • Isomers not separated

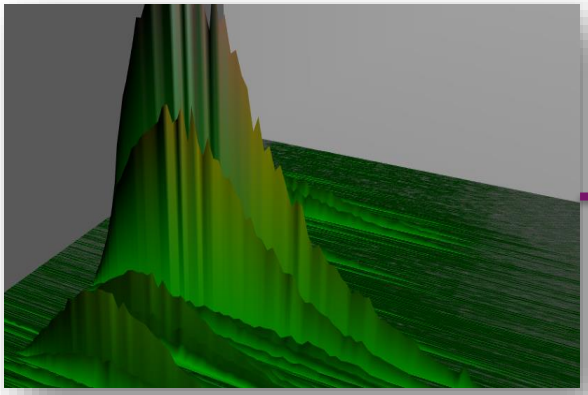
FIA is adapted to high-throughput screening



The *proFIA* workflow

Alexis Delabrière

Raw files



Detection of m/z bands

Noise estimation

Peak model estimation

Signal filtering

Peak alignment

Missing values imputation

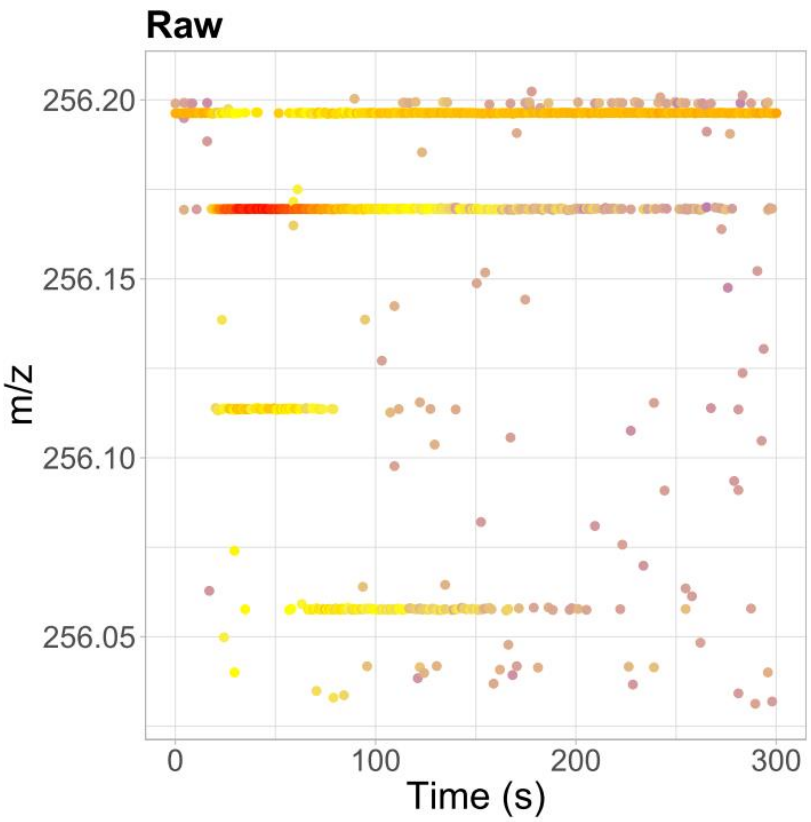
Original work
Major modification of existing algorithm

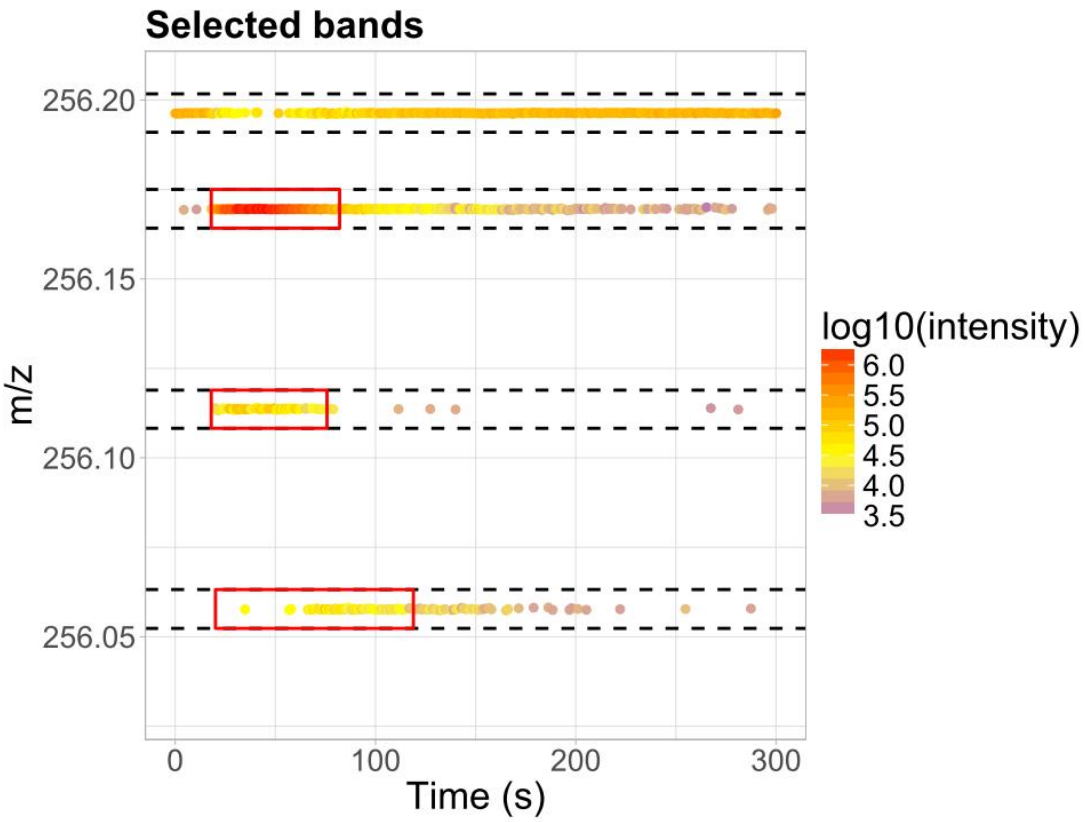
Variable by sample table of peak intensities

	1	2	3	4	5	6
1	mzMed	mzMin	mzMax	meanSolvent	corMean	UlH344_12_A
2	163.02777	163.02767	163.02782	845186.744	0.42770334	2105187
3	163.03893	163.03889	163.039	0	0.71292516	10484
4	163.11577	163.11562	163.11586	0	0.54386442	183310
5	164.02923	164.02919	164.0294	0	0.6646757	119575

Delabriere et al. (2017). *proFIA*: A data preprocessing workflow for Flow Injection Analysis coupled to High-Resolution Mass Spectrometry. *Bioinformatics*.

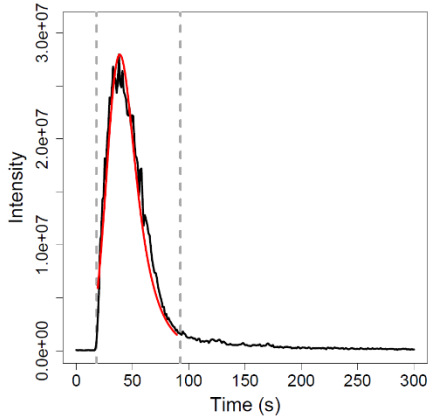
Detection of m/z bands



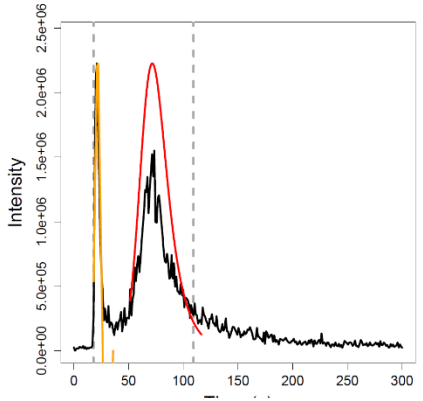


Signal filtering

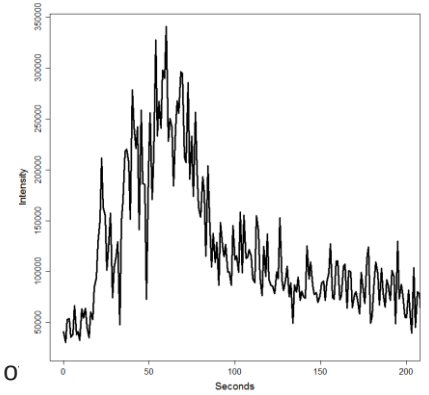
- This peak model is used to perform matching filtration on the signal



- The match can be extended if a second maximal is found on the filter. If not, a triangular filter is used for coarser grain



- A statistical test has been developed to discard signals too close to the baseline



Computational metabolomics

Preprocessing

➢ **Statistical analysis**

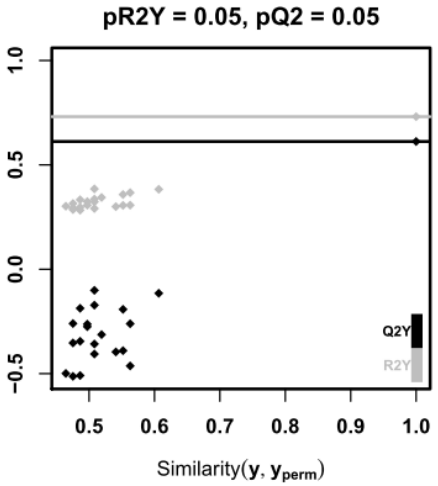
Annotation

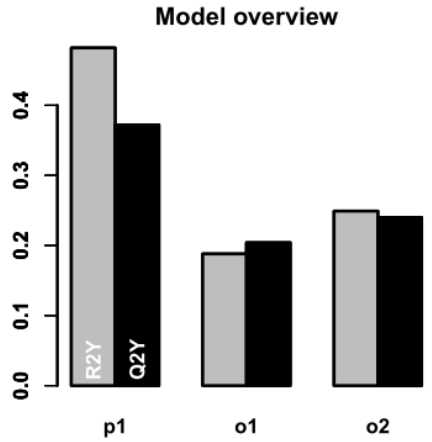
Workflow management

Bridging proteomics & metabolomics

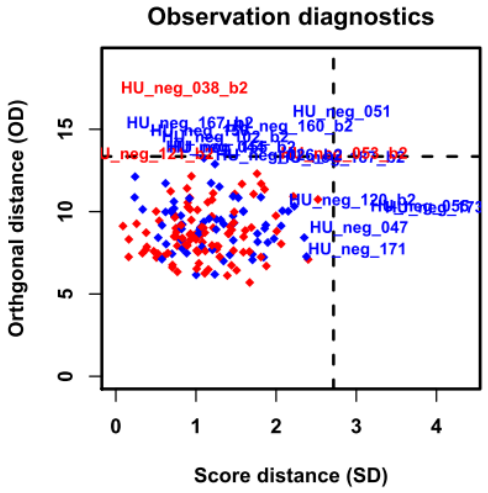
ropls package: R implementation of the (O)PLS(-DA) modeling algorithms

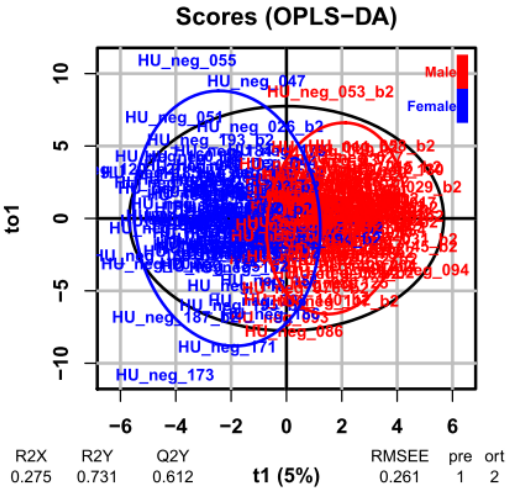
- Full diagnostics
  - outliers
  - permutation testing





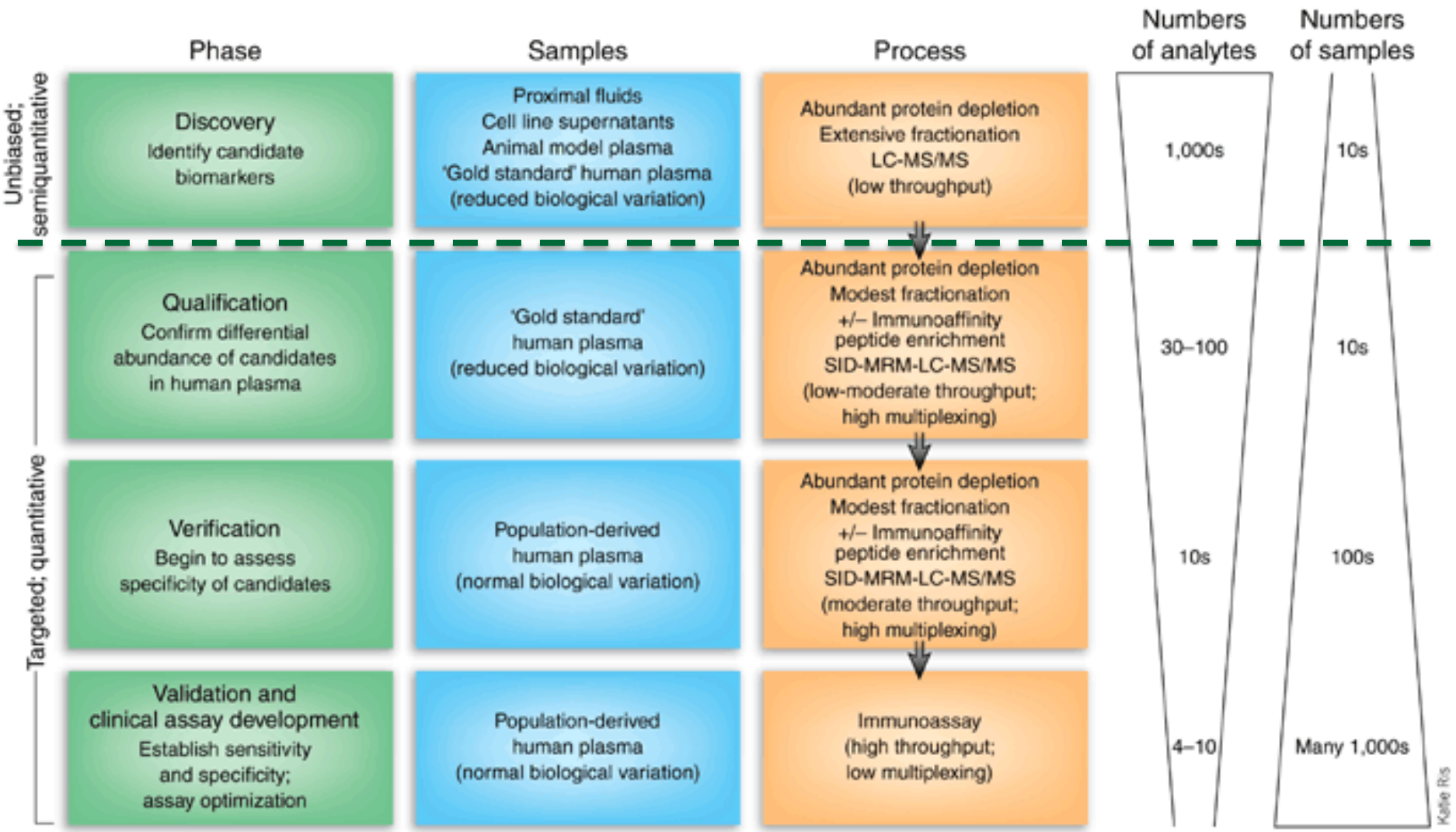
- Full numerical and graphical results
  - R2X, R2Y, Q2Y
  - VIPs





Thevenot et al. (2017). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses.

Feature selection: from biomarker discovery to clinical diagnostics



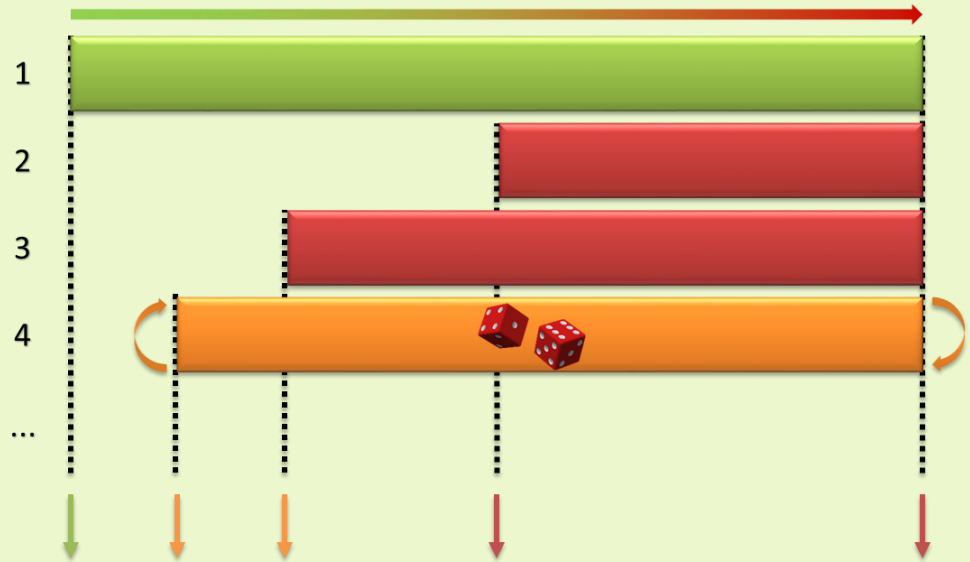
[Rifai et al. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility]

The *biosigner* wrapper algorithm

Philippe Rinaudo

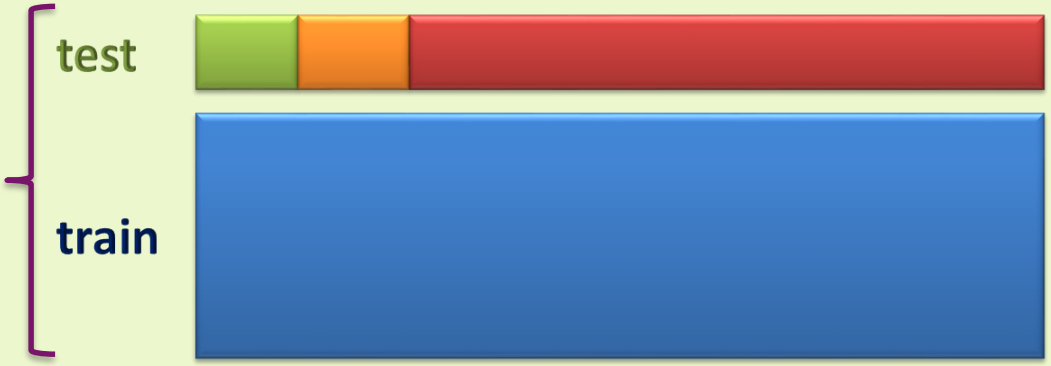
Repeat until selected subset is stable:

2) features ranked by their importance for the classifier on the train subsets



3) half-interval search for the largest irrelevant feature subset

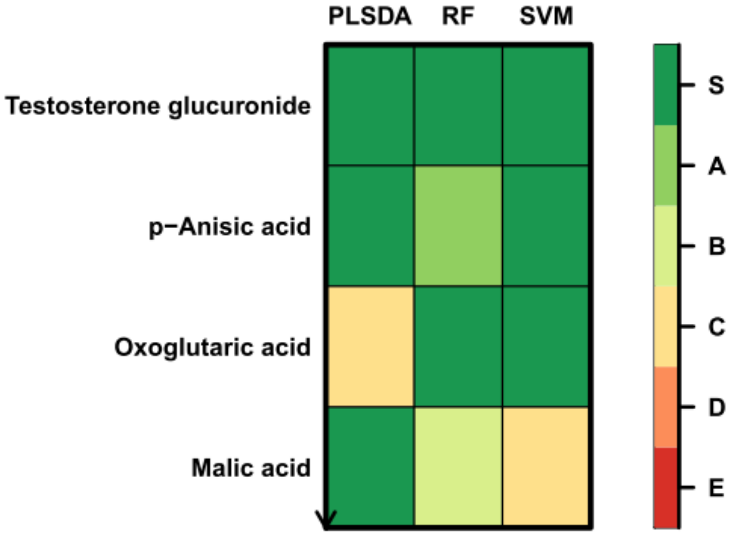
1) resampling (bootstrap)



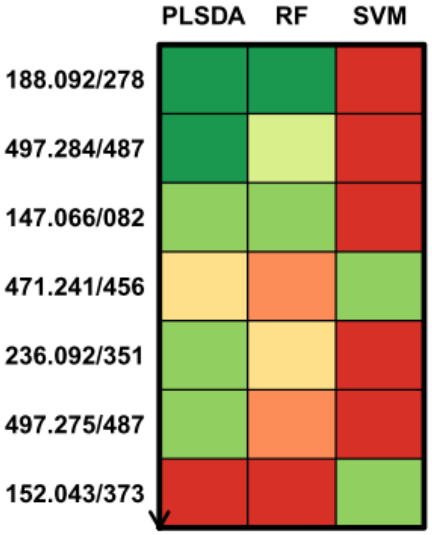
4) discard irrelevant features

biosigner package: model performances

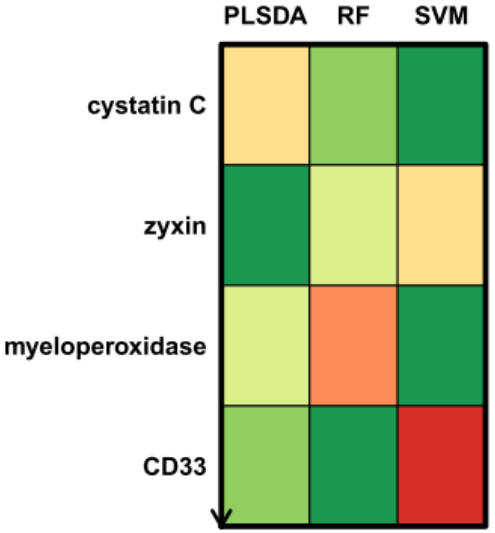
sacurine *(ropls)*



diaplasma *(biosigner)*



leukemia *(golubEsets)*



		sacurine	diaplasma	leukemia
factor		gender	diabetic type	ALL/AML
samples		183	69	72
features		109	5,501	7,129
signatures		[2-3]	[0-2]	[1-2]
performances (full -> restricted)	PLS-DA	87% -> **89%**	83% -> **91%**	95% -> **87%**
	Random Forest	86% -> **86%**	81% -> **81%**	92% -> **92%**
	SVM	88% -> **89%**	83% -> **na**	93% -> **95%**

Computational metabolomics

Preprocessing

Statistical analysis

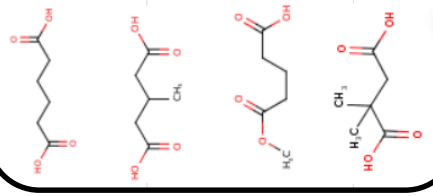
➢ **Annotation**

Workflow management

Bridging proteomics & metabolomics

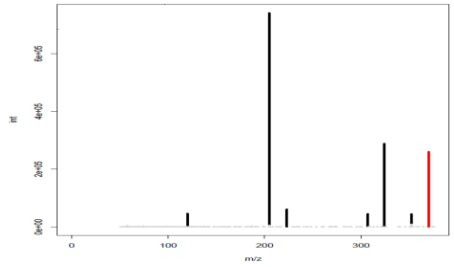
Spectrum identification

Molecular and/or spectral **database**



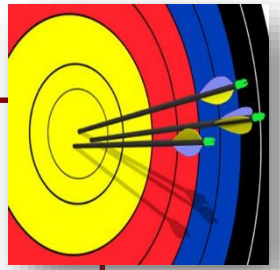
Input:

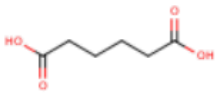
- A mass spectrum
- A precursor m/z



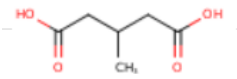
Identification sofware*

Output :
Scored molecules

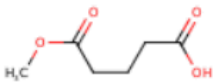




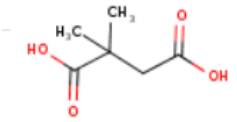
0,98



0,85

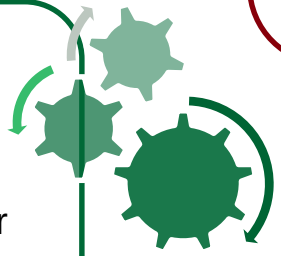


0,63



0,82

Model learned on a database.
CFM-ID: Transition probabilities
CSI-FingerID: Linear transformation between kernel spaces

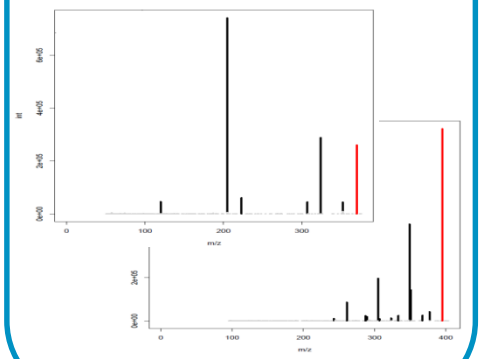


- CSI-FingerID (Shen, 2014)
- CFM-ID (Allen, 2014)
- MetFrag (Wolf, 2010)
- Etc…

Mining spectral libraries

Input:

- A set of spectra

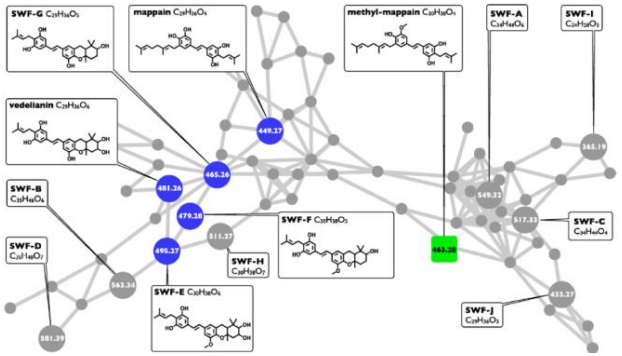


Spectral Mining software

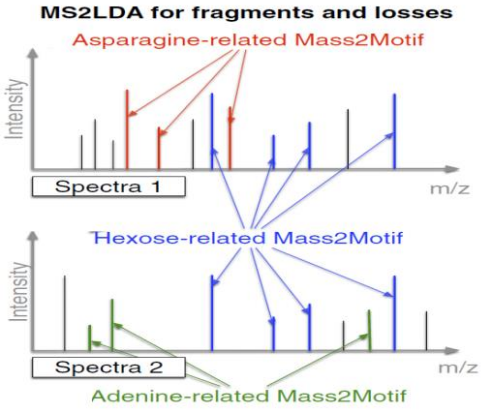
Output:

Information about structural similarities

- **Networks** (GNPS, Wang, 2014)



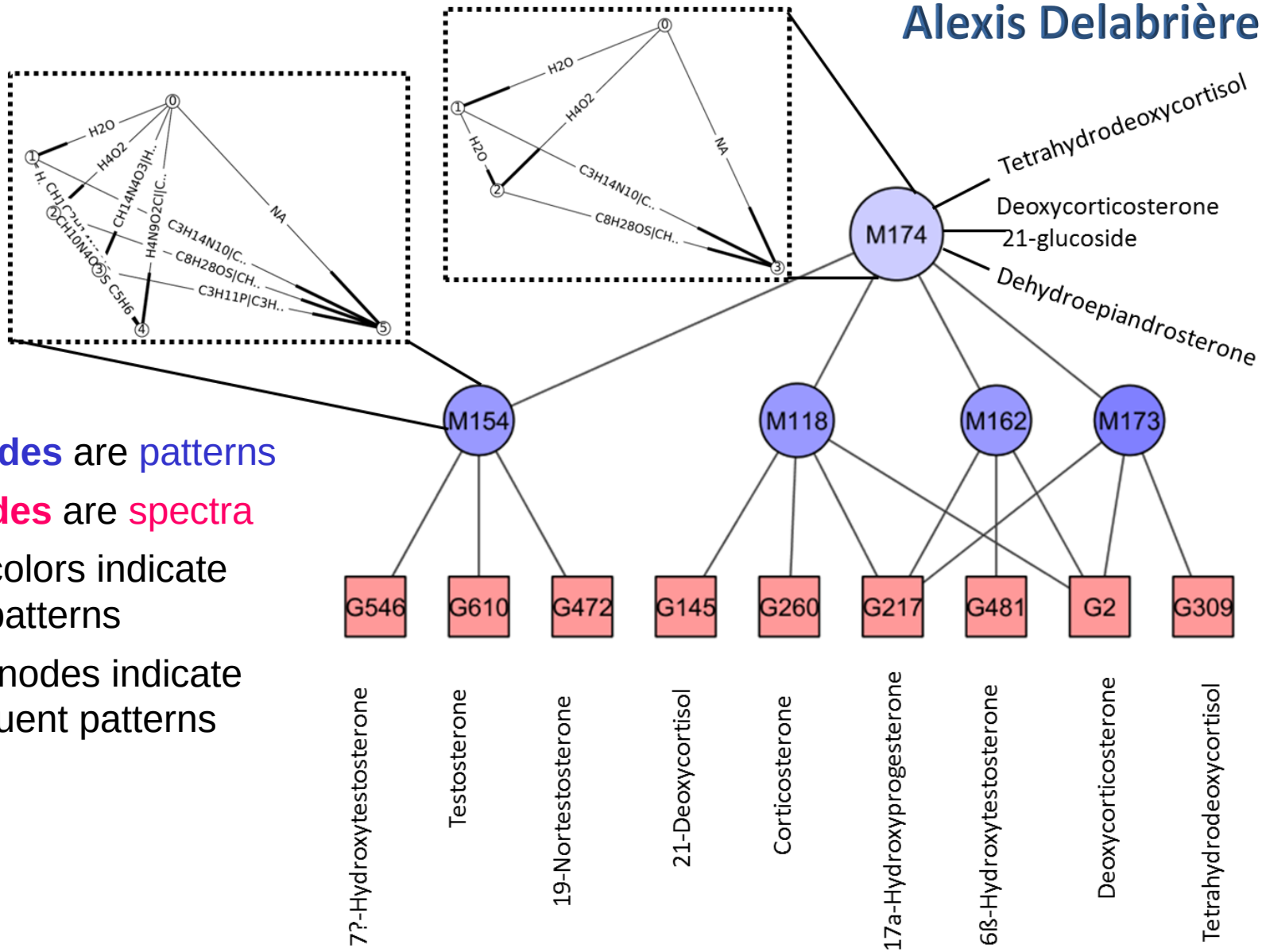
- **Motifs** discovery (MS2LDA, Van der Hooft, 2016)



Frequent subgraph mining

Alexis Delabrière

- **Blue nodes** are patterns
- **Red nodes** are spectra
- Lighter colors indicate smaller patterns
- Smaller nodes indicate less frequent patterns



Computational metabolomics

Preprocessing

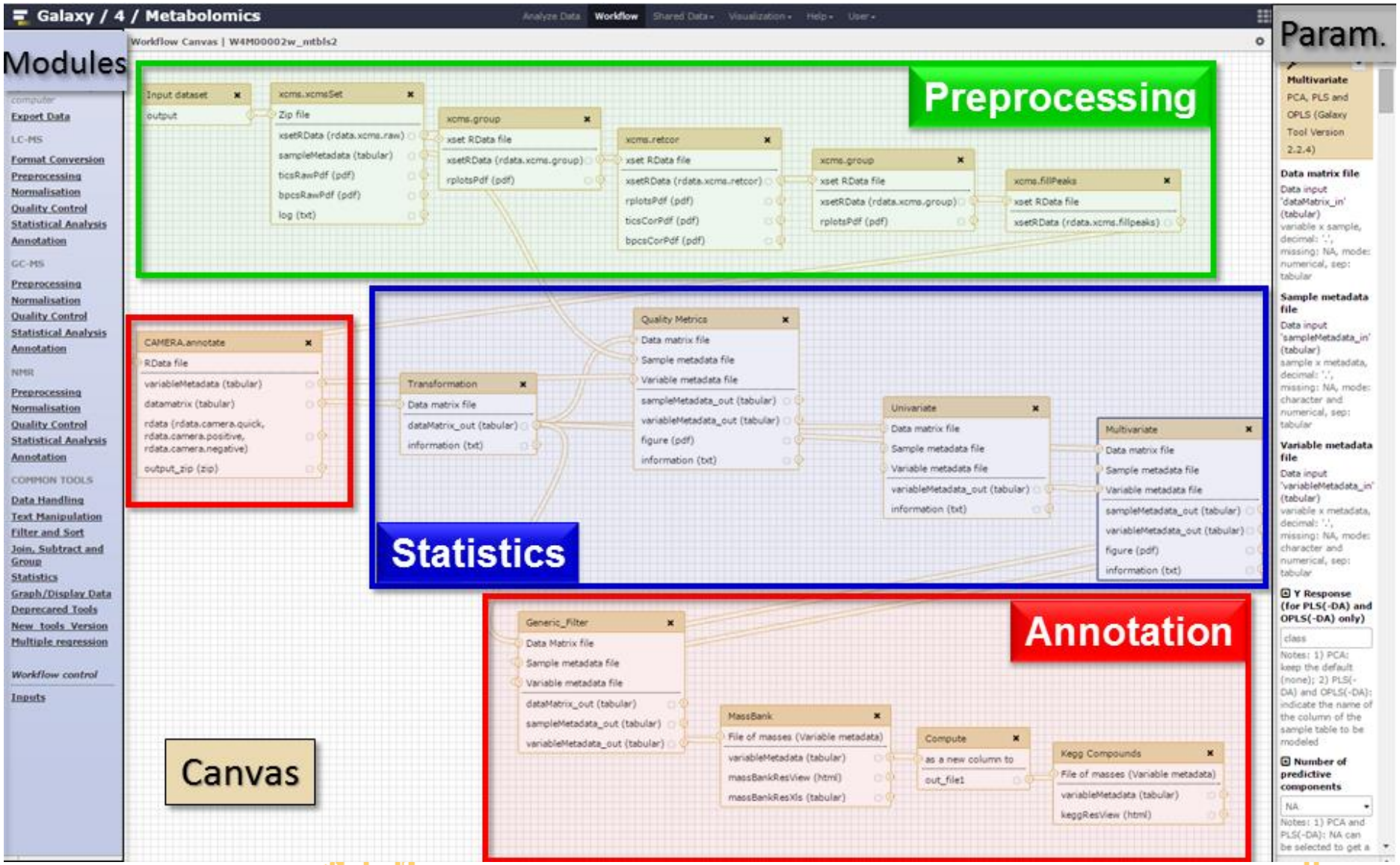
Statistical analysis

Annotation

➢ **Workflow management**

Bridging proteomics & metabolomics

Managing workflows with Galaxy



Workflow4metabolomics

Main menu

- Home
- Events
- History
- Introduction
  - The Galaxy environment
  - The LC-MS workflow
  - The GC-MS workflow
  - The NMR workflow
  - References
- HowTo
- Download
  - Datasets
- Referenced WorkFlows and Histories
- How to contribute?
- Developer resources
  - Source-code
  - Virtual environments

Workflow4Metabolomics 3.0

Welcome to the collaborative portal dedicated to metabolomics data processing, analysis and annotation for Metabolomics community.

" We are happy to announce the next **Workflow4Experimenters (W4E) international course 2018**: *Using Galaxy and the Workflow4metabolomics infrastructure to analyse metabolomics data.*
Please save the date: ***8-12 October 2018*** at Pasteur Institute, Paris - France

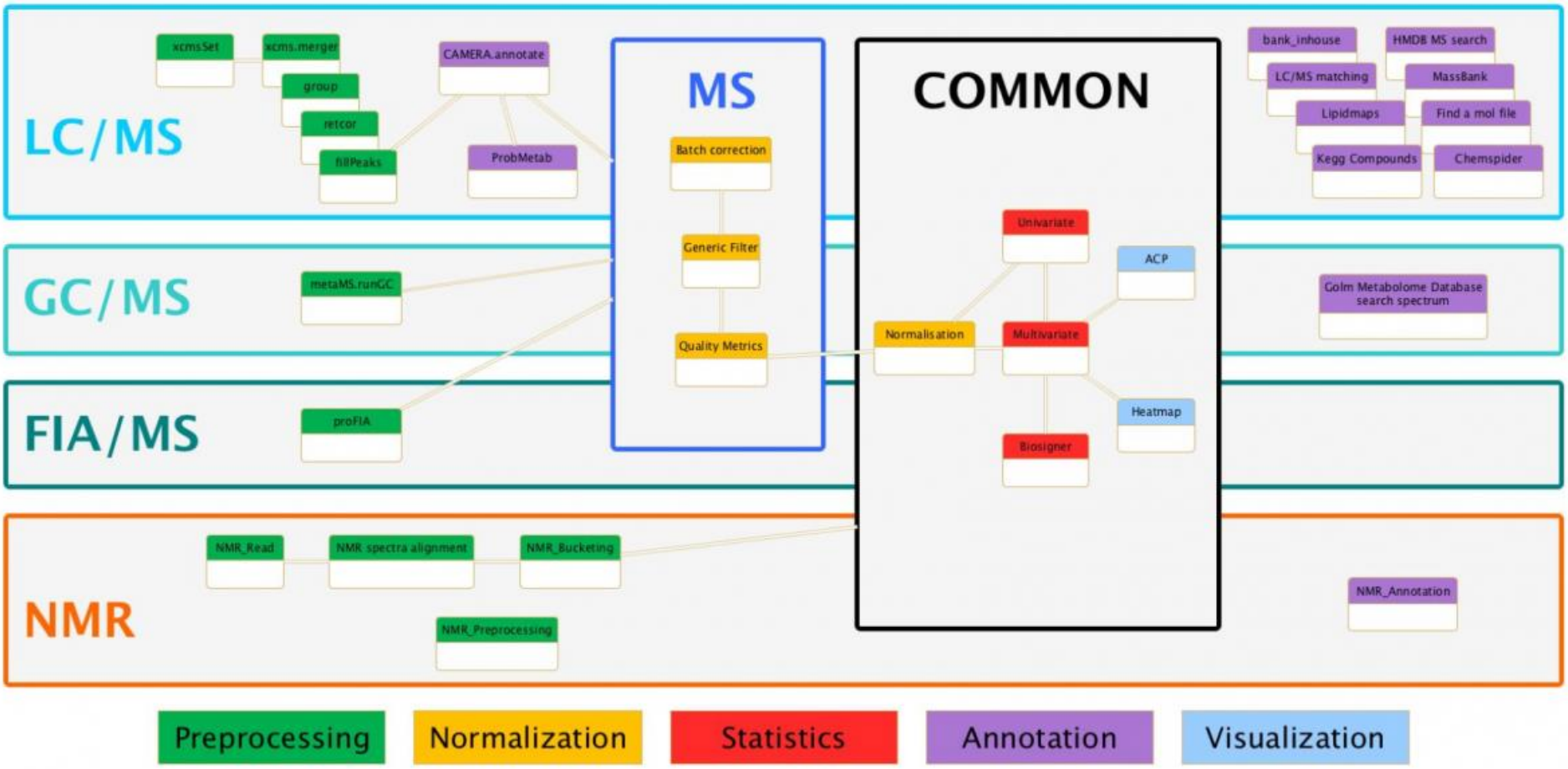
More news in April ! "

Follow us on Twitter @workflow4metabo

STEP 1 STEP 2 STEP 3 STEP 4

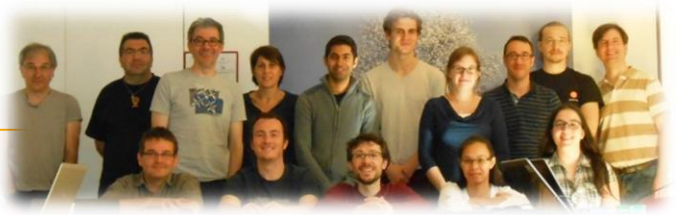
[Giacomoni et al. (2015). Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics.]

W4M tools



Guitton et al. (2017). Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics.

W4M offer



The W4M Core Team

- Private account
- Computation and storage resources
- Help desk

- Sharing and referencing of histories and workflows (DOI)

- Annual courses (tutoring on your own data)

Save the date: 8-12 October 2018, Pasteur Institute (Paris)

- Installation of local instances NEW

contact@workflow4metabolomics.org

A team work from talented people

- The CEA team
  - Pierrick Roger-Mele, Natacha Lenuzza, Alexis Delabrière, Philippe Rinaudo *et al.*

- The Workflow4Metabolomics Core Team
  - Christophe Caron, Franck Giacomoni, Gildas Le Corguillé, Yann Guitton, Marie Tremblay-Franco, Jean-François Martin, Mélanie Pétéra, Nils Paulhe, Christophe Dupérier, Cécile Canlet, *et al.*

- The PhenoMeNal consortium
  - Christoph Steinbeck, Steffen Neumann, Namrata Kale, Pablo Moreno, Kenneth Haug, Reza Salek, Tim Ebbels, Philippe Rocca-Serra, Michael van Vliet, Marta Cascantes, Pedro de Atauri, Christoph Ruttkies, Kristian Peters, Daniel Schober, Luca Pirredu, Gianluigi Zanetti, Merlijn van Rijswijk, Ola Spjuth, Ralf Weber, Mark Viant, *et al.*

Computational metabolomics

Preprocessing

Statistical analysis

Annotation

Workflow management

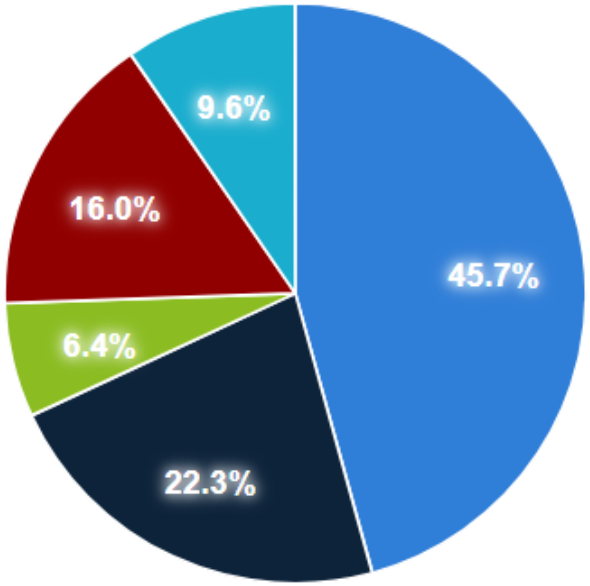
➢**Bridging proteomics & metabolomics**

Audience

- ~ 100-120 attendees
- 40% from industry

Domaine

Chart options »



biologie	43
chimie	21
mathématiques	6
informatique	15
autre	9

Overview of some recent publications

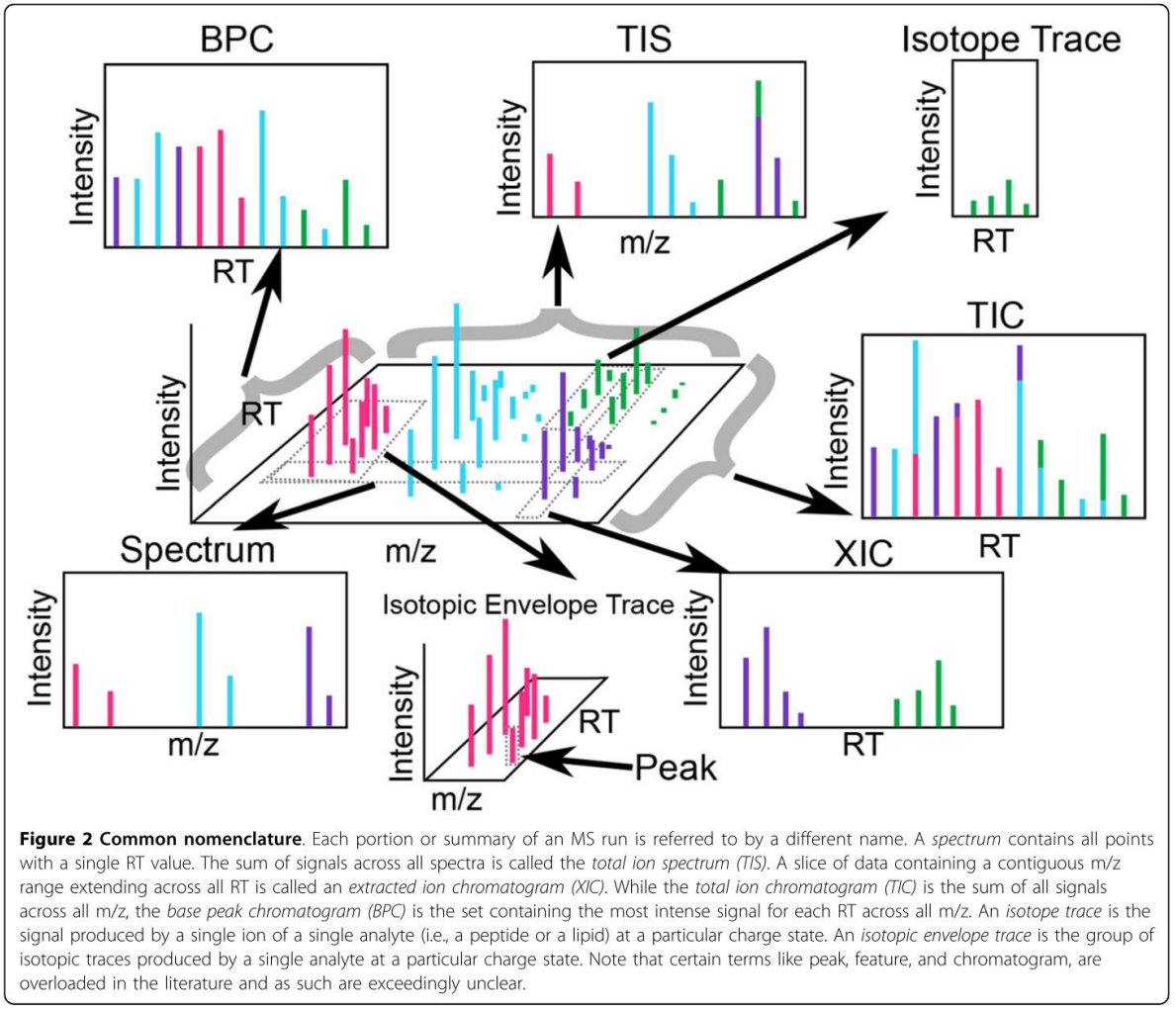
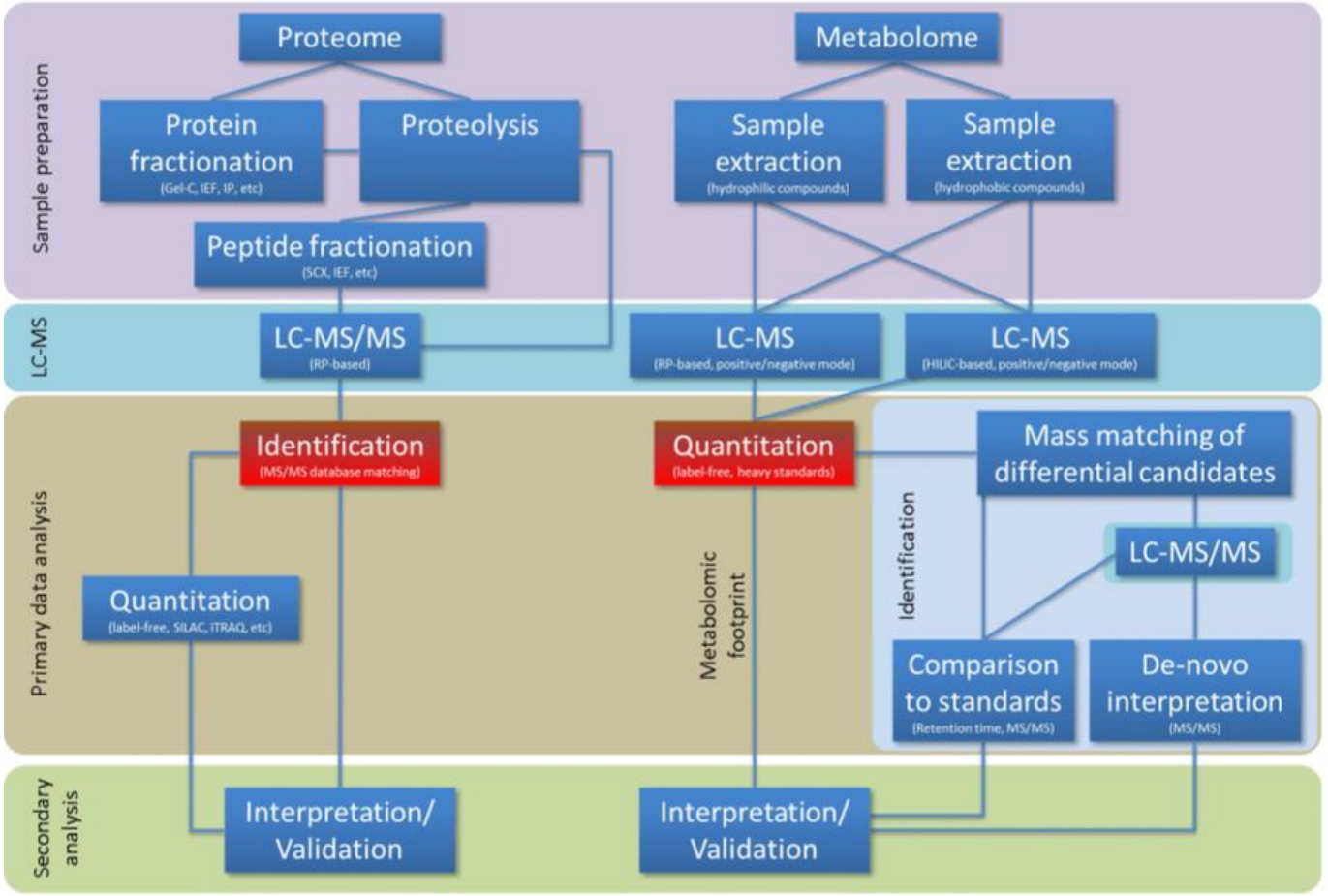


Figure 2 Common nomenclature. Each portion or summary of an MS run is referred to by a different name. A *spectrum* contains all points with a single RT value. The sum of signals across all spectra is called the *total ion spectrum (TIS)*. A slice of data containing a contiguous m/z range extending across all RT is called an *extracted ion chromatogram (XIC)*. While the *total ion chromatogram (TIC)* is the sum of all signals across all m/z, the *base peak chromatogram (BPC)* is the set containing the most intense signal for each RT across all m/z. An *isotope trace* is the signal produced by a single ion of a single analyte (i.e., a peptide or a lipid) at a particular charge state. An *isotopic envelope trace* is the group of isotopic traces produced by a single analyte at a particular charge state. Note that certain terms like peak, feature, and chromatogram, are overloaded in the literature and as such are exceedingly unclear.

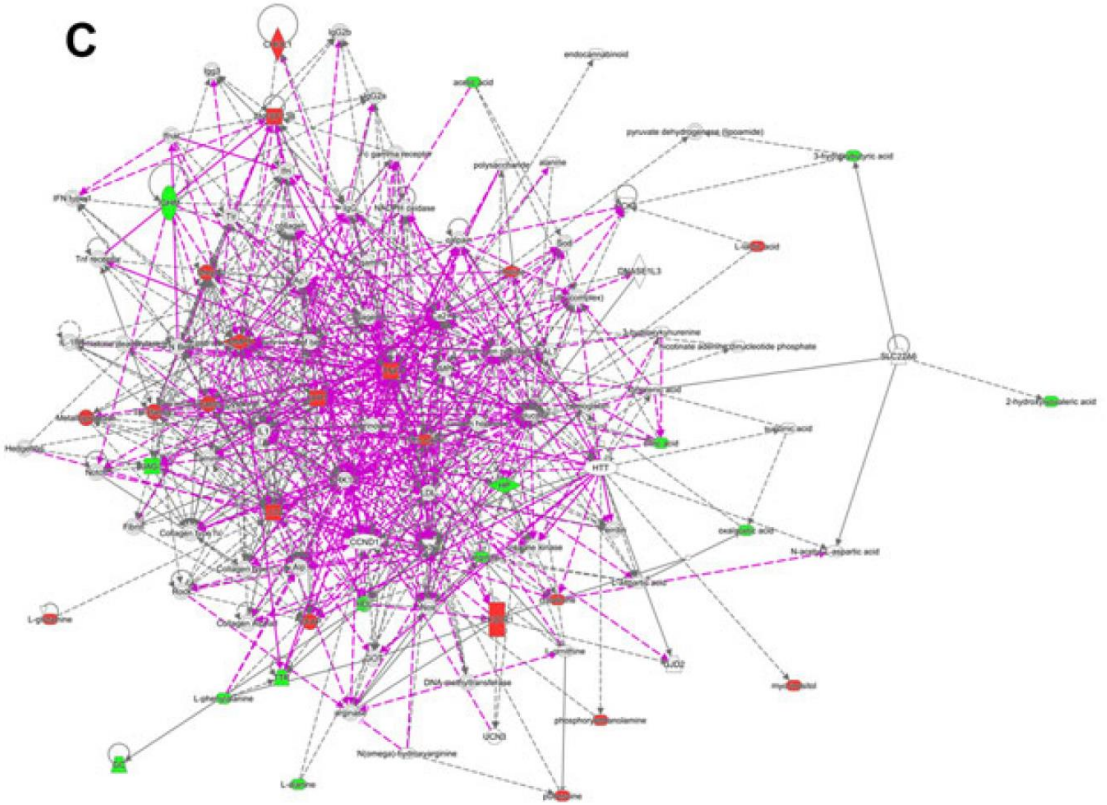
Smith *et al.* (2014). Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics*, 15. DOI:10.1186/1471-2105-15-S7-S9

Overview of some recent publications



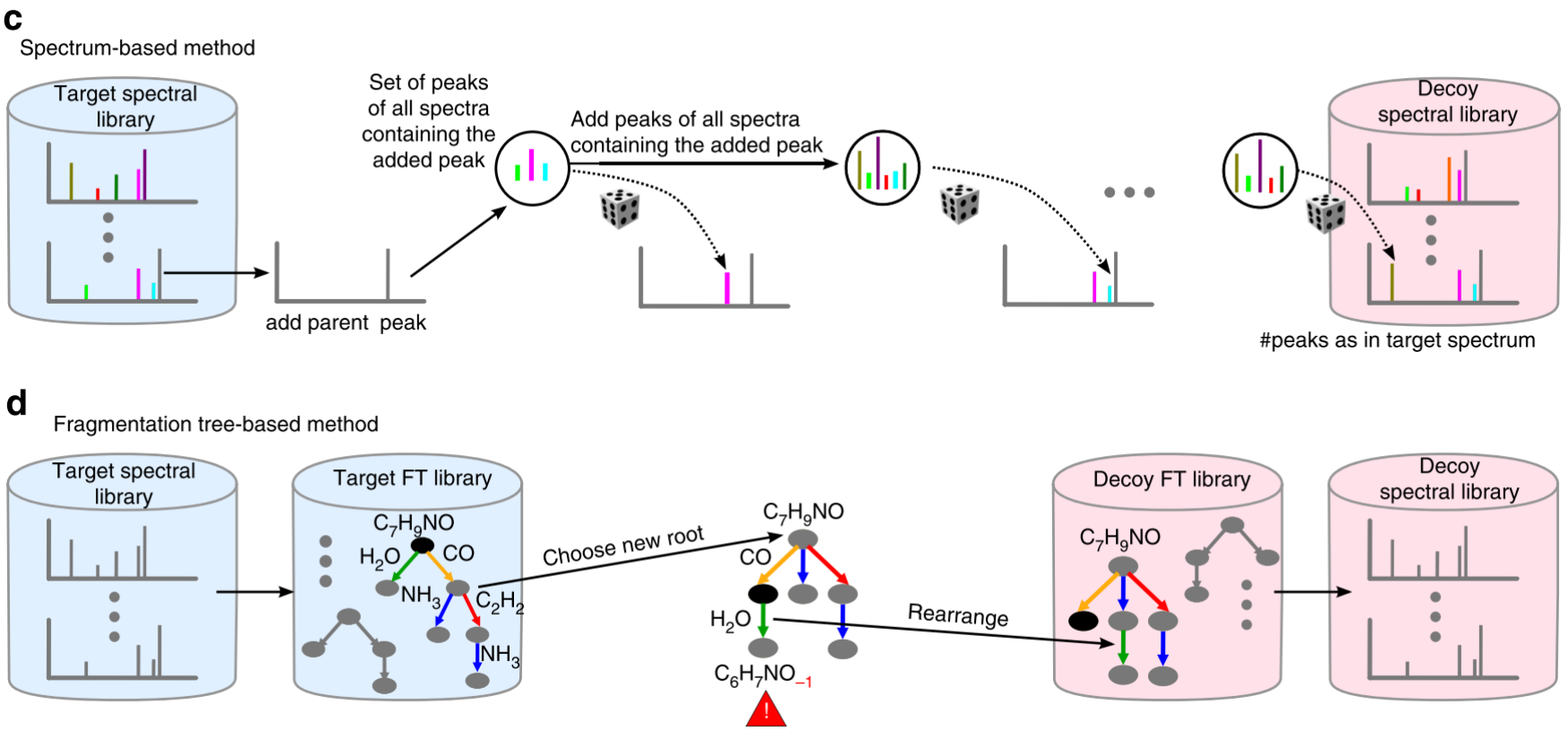
Fischer *et al.* (2013). Two birds with one stone: Doing metabolomics with your proteomics kit. *Proteomics*, 13:3371-3386, DOI:10.1002/pmic.201300192.

Overview of some recent publications



Del Boccio *et al.* (2016). Integration of metabolomics and proteomics in multiple sclerosis: From biomarkers discovery to personalized medicine. *Proteomics Clinical Applications*, 10(4), 470–484. https://doi.org/10.1002/prca.201500083

Overview of some recent publications



Scheubert *et al.* (2017). Significance estimation for large scale metabolomics annotations by spectral matching. *Nature Communications*, 8(1), 1494. https://doi.org/10.1038/s41467-017-01318-5

Bridges between proteomics & metabolomics

- Preprocessing
- Statistics
- Identification
- Data integration
- Metabolic networks
- Formats (MSnBase)
- Databases
- Biology
- Analytical chemistry
- …