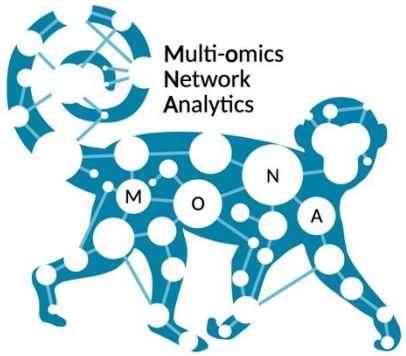
Informatics Platform and MoNA

An introduction to Metabolomics

Date: 03 June 2026



Henry Webel, Senior Data Scientist

Felicia Cara Schulz, PhD student

Maria Barranco Altirriba, PostDoc

Alberto Pallejà, Team Lead Data Science Platform

Introduction to metabolomics

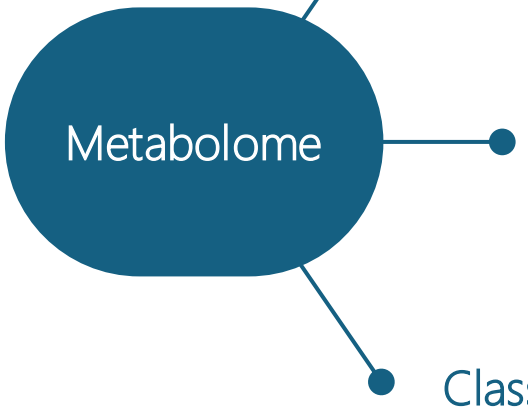
Complete set of small molecules (usually <1500 Da) present in a biological system at a given time

Introduction to metabolomics

Complete set of small molecules (usually <1500 Da) present in a biological system at a given time

Sources: exogenous, endogenous, microbiome-derived

Introduction to metabolomics



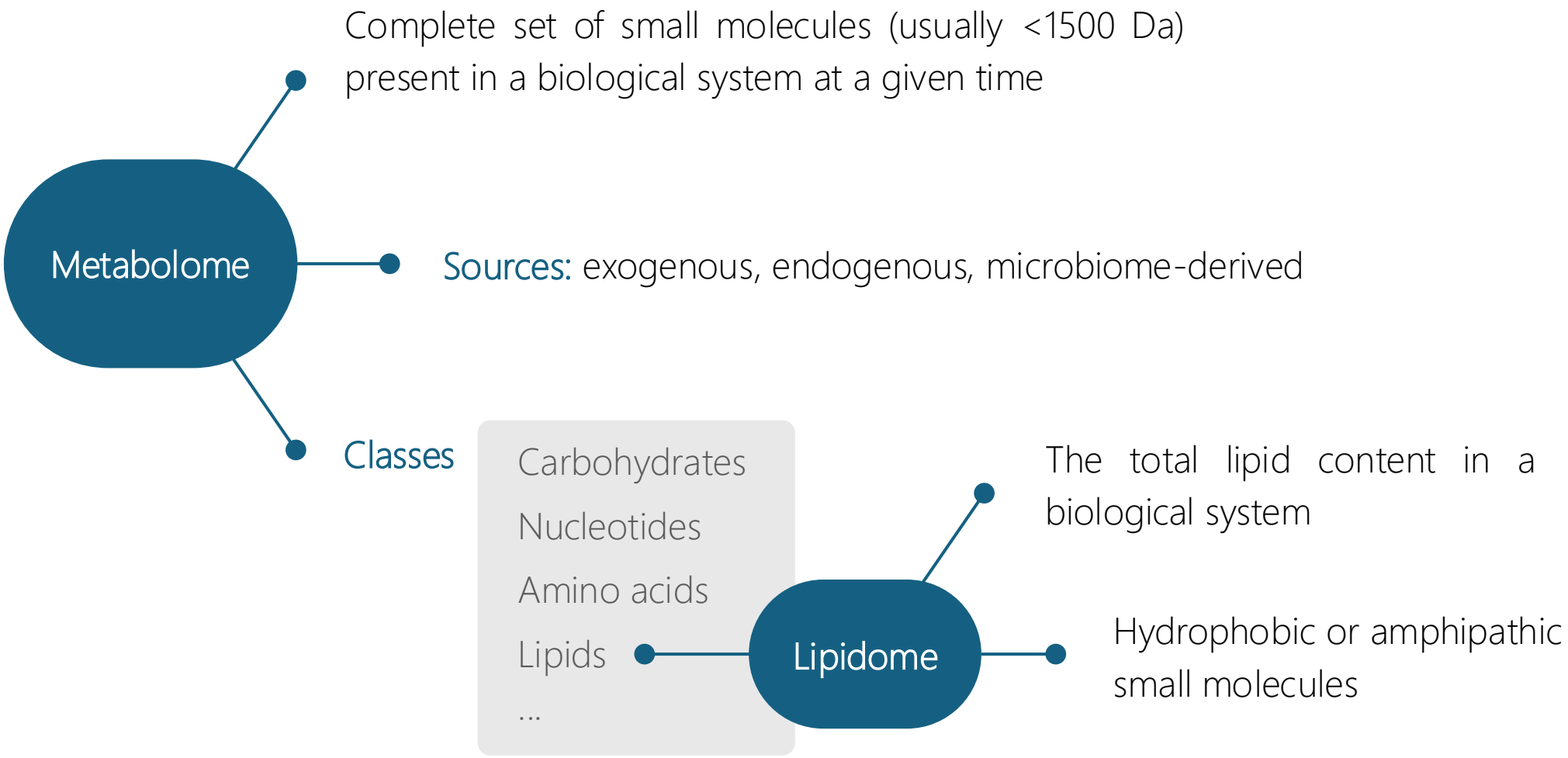
Complete set of small molecules (usually <1500 Da) present in a biological system at a given time

Sources: exogenous, endogenous, microbiome-derived

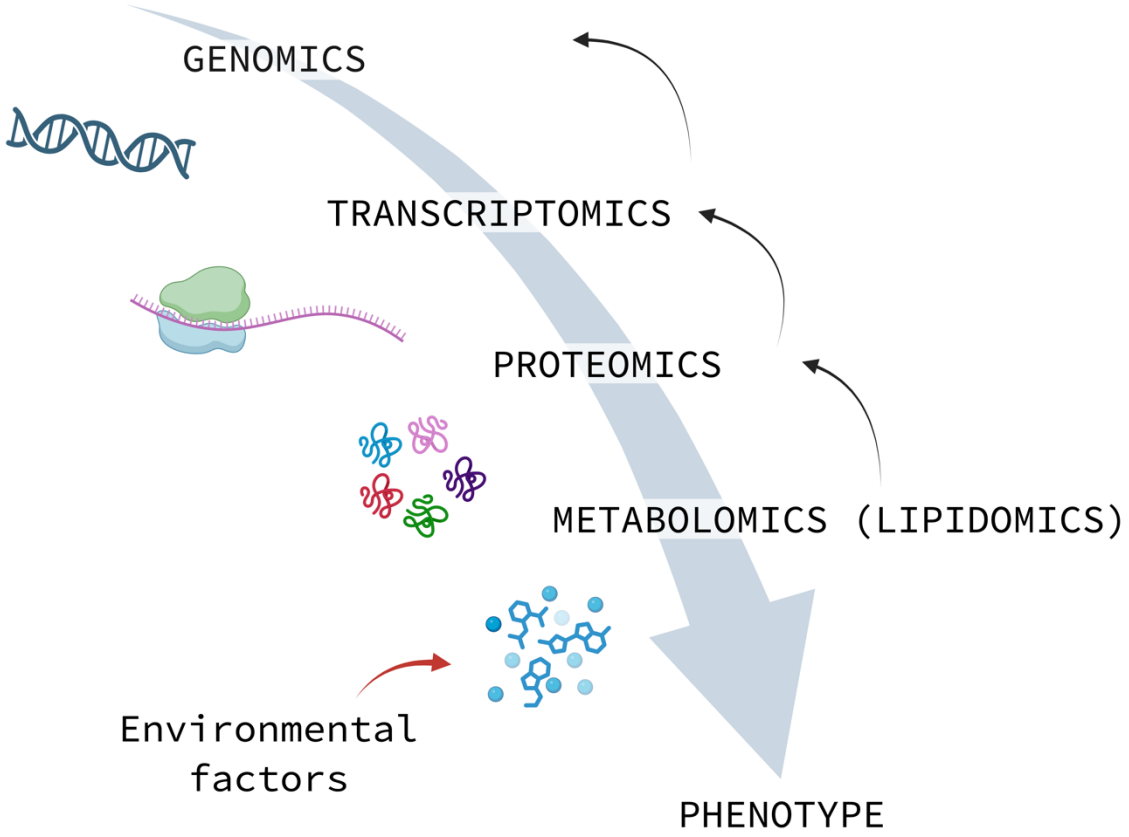
Classes

- Carbohydrates
- Nucleotides
- Amino acids
- Lipids
- …

Introduction to metabolomics



Introduction to metabolomics



Introduction to metabolomics

Applications

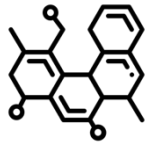
1. Discovery of new disease biomarkers
2. The understanding of molecular mechanisms in pathophysiological processes
3. Nutrition research
4. Drug discovery
5. Fermentation
6. Environmental Monitoring and Ecotoxicology

Introduction to metabolomics

Applications examples

Proposed biosynthetic pathway for tetrangulol in Streptomyces sp. KL110A

Identification of biosynthetic gene clusters (BGCs) in Streptomyces.



Metabolites produced by the strain

Comparison with previously characterized pathways

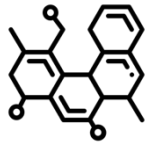
Trejo-Alarcon, L.M., Cano-Prieto, C., Calheiros de Carvalho, A. et al. Integrative metabolo-genomics suggests a biosynthetic pathway for tetrangulol in Streptomyces sp. KL110A. World J Microbiol Biotechnol (2025).

Introduction to metabolomics

Applications examples

Proposed biosynthetic pathway for tetrangulol in Streptomyces sp. KL110A

Identification of biosynthetic gene clusters (BGCs) in Streptomyces.



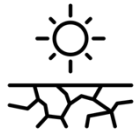
Metabolites produced by the strain

Comparison with previously characterized pathways

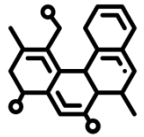
How do different wheat-related species respond metabolically to drought?

7 *Triticae* species (drought tolerant types and cultivated types) were studied.

Normal conditions



Drought stress conditions



Sugars, aa and organic acids were consistently increased (stress protection/osmotic adjustment).

Trejo-Alarcon, L.M., Cano-Prieto, C., Calheiros de Carvalho, A. et al. Integrative metabolo-genomics suggests a biosynthetic pathway for tetrangulol in Streptomyces sp. KL110A. World J Microbiol Biotechnol (2025).

Ullah N, Yüce M, Neslihan Öztürk Gökçe Z, Budak H. Comparative metabolite profiling of drought stress in roots and leaves of seven Triticeae species. BMC Genomics (2017).

Introduction to metabolomics

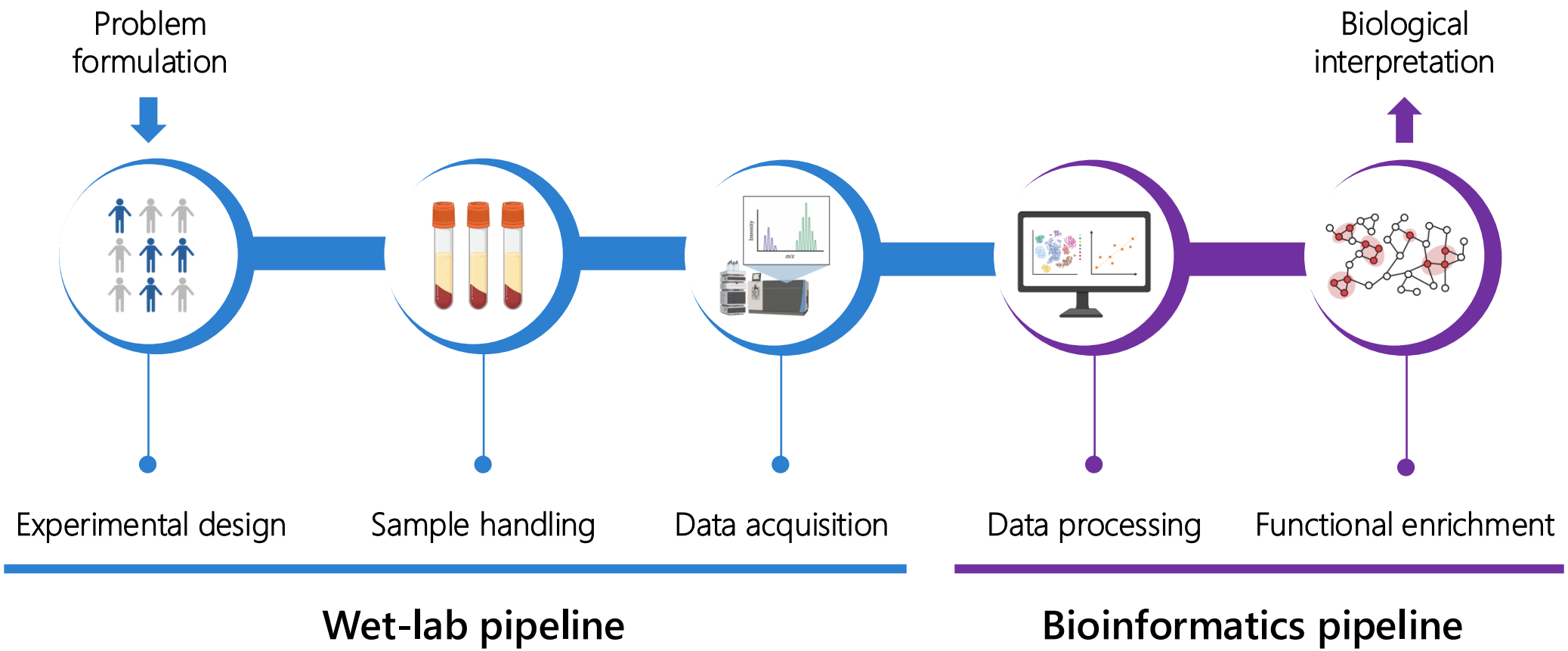
Study types

1. **Untargeted** – the goal is to characterise the broadest range of metabolites present in a sample. It allows to discover unanticipated alterations; however, it results in complex datasets.

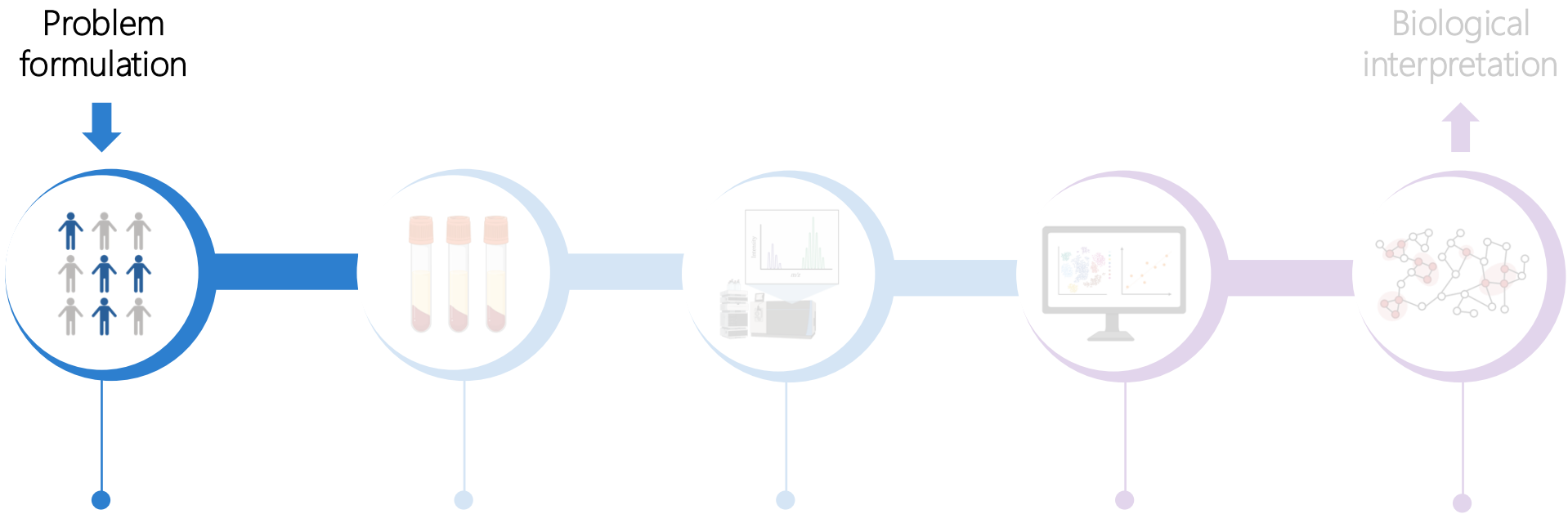
2. **Semi-targeted** – the goal is to identify and quantify a large set of small molecules. It allows the characterization of a pathway or some specific classes of small molecules (e.g., aminoacids).

3. **Targeted** – the goal is to determine absolute concentrations of a predefined set of metabolites (selected based on prior knowledge). It is usually used to validate untargeted or semi-targeted metabolomics studies.

Metabolomics pipeline



Metabolomics pipeline



Experimental design

Confounding factors:

- Analytical
- Physiological

Sample handling

- Collection, storage and preparation
- Tissues, cells, biofluids
- Isolate molecules of interest.

Data acquisition

- LC-MS
- GC-MS
- NMR

Data processing

- Pre-processing
- Data filtering and normalization
- Statistical analysis
- Peak annotation

Functional enrichment

- Over-representation analysis
- Functional class scoring
- Pathway topology

Metabolomics pipeline

Experimental design

Deal with confounding factors that could correlate or mask the biology under-study:

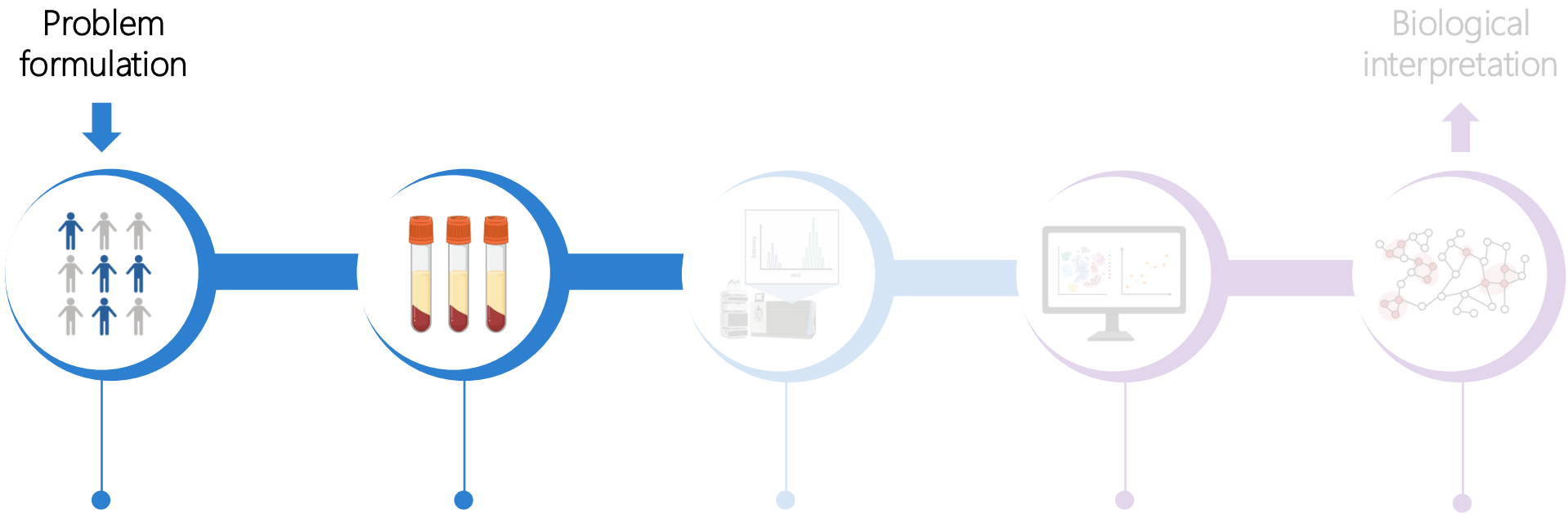
- Analytical: batch effects, injection order, etc.
- Biological: sex, age, body mass index, etc.

Biological samples randomization based on their characteristics

Quality control (QC) samples are typically prepared by taking aliquots from a pool of all the study samples.

QC samples are injected at the start and periodically throughout the acquisition to stabilize the signal and correct for technical variability.

Metabolomics pipeline



Experimental design

Confounding factors:

- Analytical
- Physiological

Sample handling

- Collection, storage and preparation
- Tissues, cells, biofluids
- Isolate molecules of interest.

Data acquisition

- LC-MS
- GC-MS
- NMR

Data processing

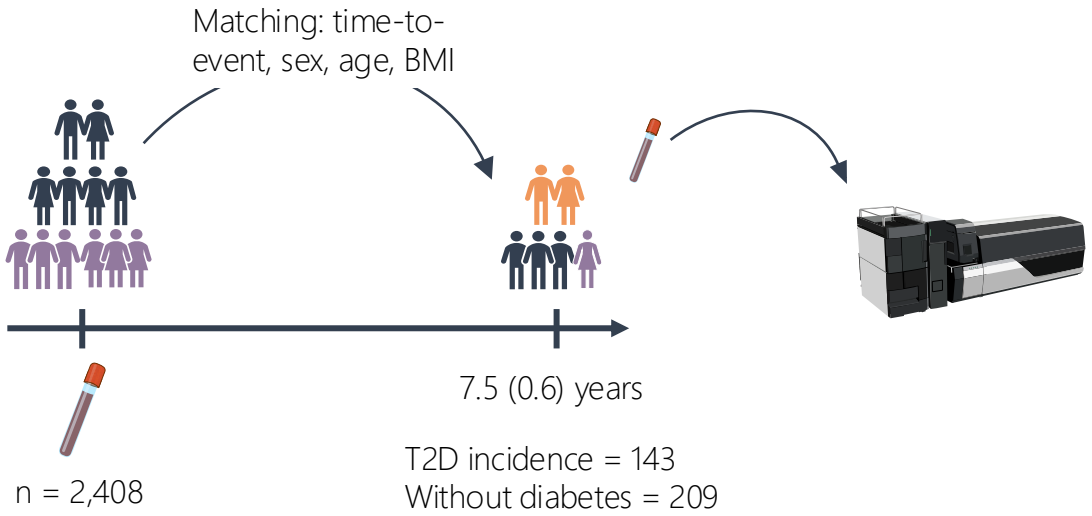
- Pre-processing
- Data filtering and normalization
- Statistical analysis
- Peak annotation

Functional enrichment

- Over-representation analysis
- Functional class scoring
- Pathway topology

Metabolomics pipeline

Study example – Diabetes incidence

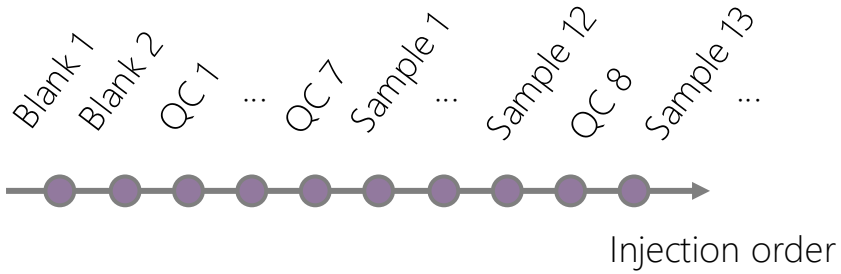


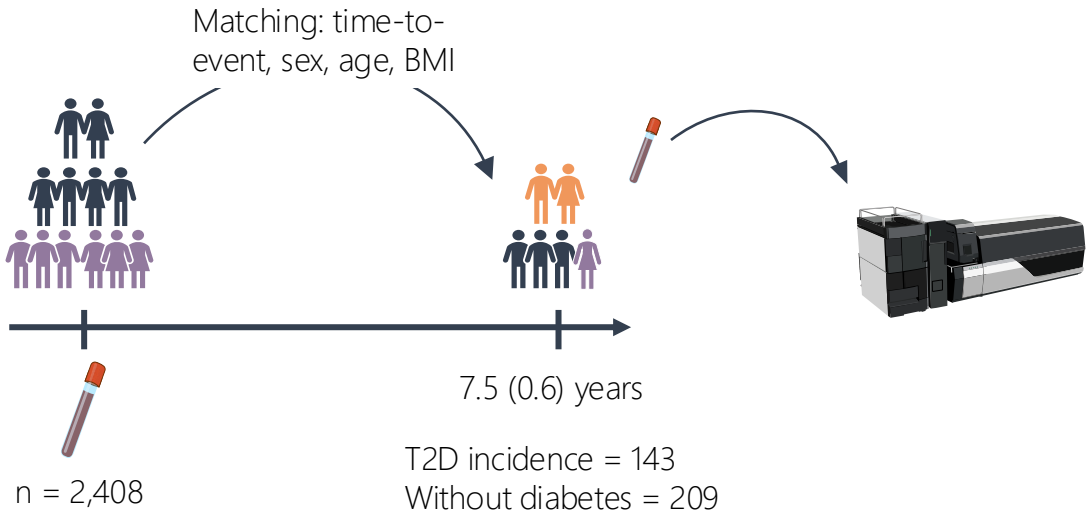
Barranco-Altirriba et al. "Guanine and pregnenolone sulfate are associated with incident type 2 diabetes in two independent populations". Frontiers in Endocrinology (2025).

Metabolomics pipeline

Study example – Diabetes incidence

Total of:
- 45 QC samples
- 4 blanks
- 352 samples



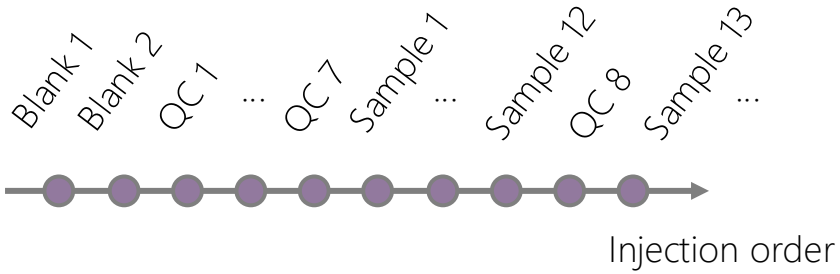


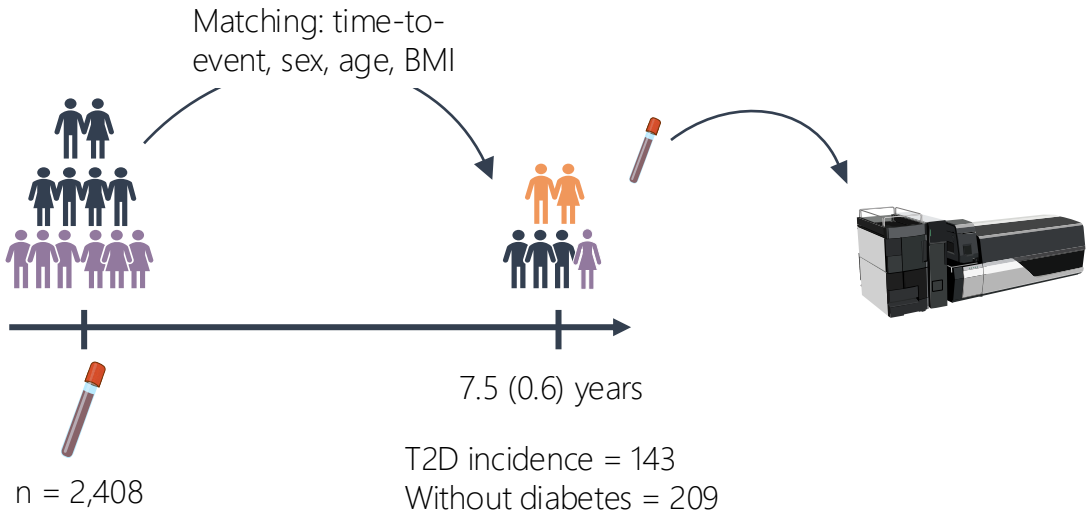
Barranco-Altirriba et al. "Guanine and pregnenolone sulfate are associated with incident type 2 diabetes in two independent populations". Frontiers in Endocrinology (2025).

Metabolomics pipeline

Study example – Diabetes incidence

Total of:
- 45 QC samples
- 4 blanks
- 352 samples





	Without T2D (N=286)	With T2D (N=143)	P-value
Sex (women)	157 (54.9%)	79 (55.2%)	
Age (years)	57.5 (12.8)	56.6 (12.0)	0.442
BMI (kg/m²)	30.5 (4.56)	30.8 (4.91)	0.483
HT (yes)	160 (55.9%)	85 (59.4%)	0.558
Insulin Resistance	2.38 (1.49)	3.05 (1.83)	**<0.001**
Glucose (mg/dL)	95.0 (10.6)	103 (13.2)	**<0.001**
Family history (yes)	134 (46.9%)	90 (63.4%)	**0.002**

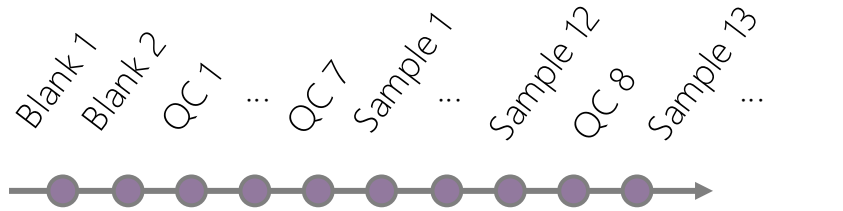
Barranco-Altirriba et al. "Guanine and pregnenolone sulfate are associated with incident type 2 diabetes in two independent populations". Frontiers in Endocrinology (2025).

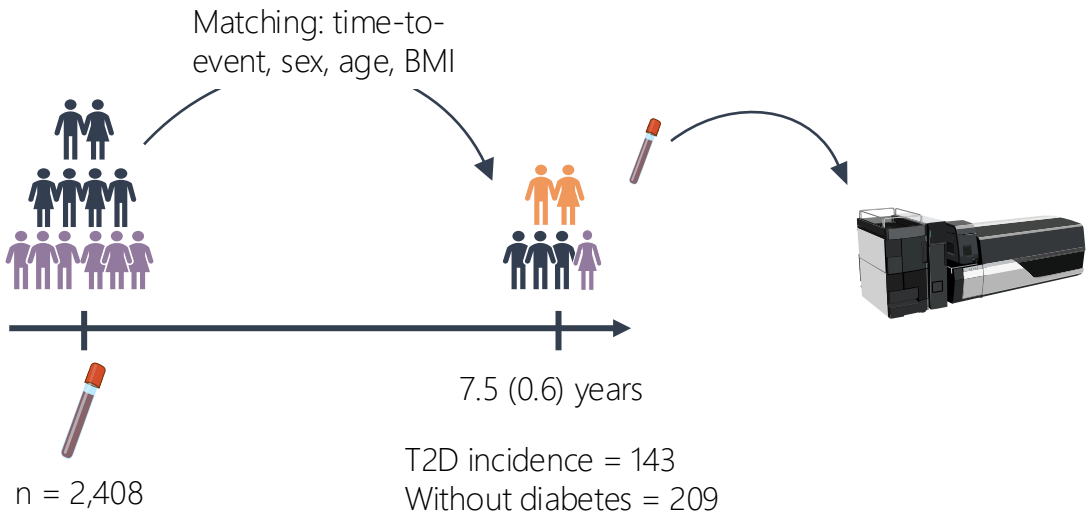
Metabolomics pipeline

Study example – Diabetes incidence

Total of:

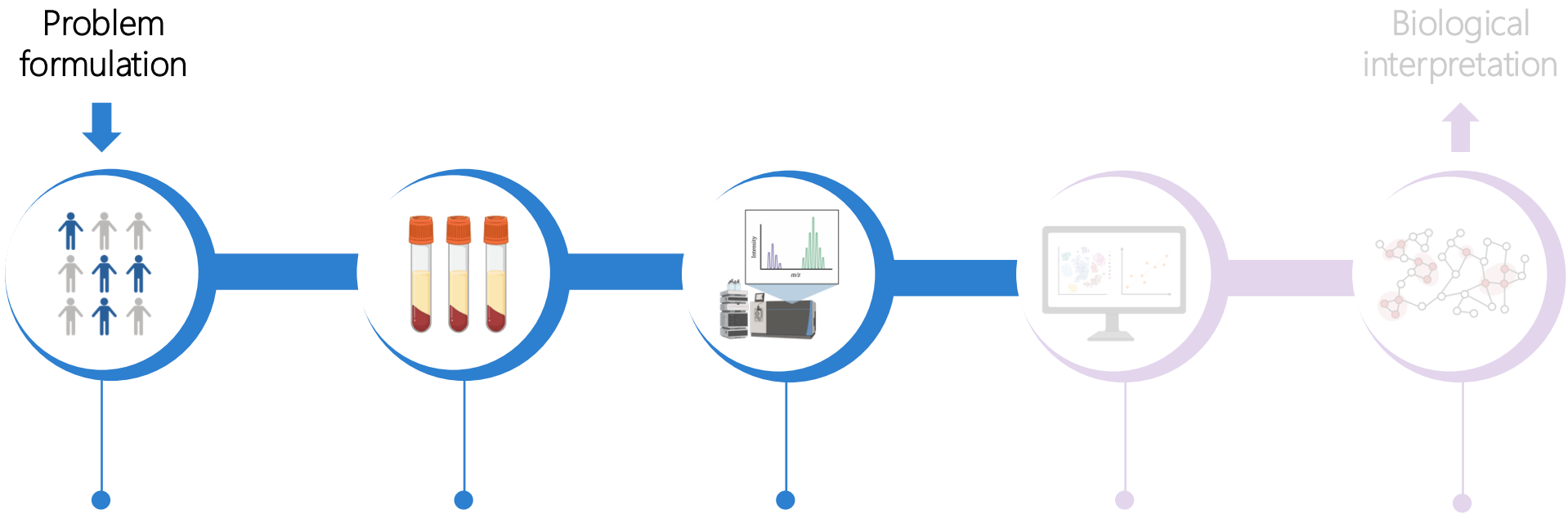
- 45 QC samples
- 4 blanks
- 352 samples





	Without T2D (N=286)	With T2D (N=143)	P-value
Sex (women)	157 (54.9%)	79 (55.2%)	
Age (years)	57.5 (12.8)	56.6 (12.0)	0.442
BMI (kg/m²)	30.5 (4.56)	30.8 (4.91)	0.483
HT (yes)	160 (55.9%)	85 (59.4%)	0.558
Insulin Resistance	2.38 (1.49)	3.05 (1.83)	**<0.001**
Glucose (mg/dL)	95.0 (10.6)	103 (13.2)	**<0.001**
Family history (yes)	134 (46.9%)	90 (63.4%)	**0.002**

Metabolomics pipeline



Experimental design

Confounding factors:

- Analytical
- Physiological

Sample handling

- Collection, storage and preparation
- Tissues, cells, biofluids
- Isolate molecules of interest.

Data acquisition

- LC-MS
- GC-MS
- NMR

Data processing

- Pre-processing
- Data filtering and normalization
- Statistical analysis
- Peak annotation

Functional enrichment

- Over-representation analysis
- Functional class scoring
- Pathway topology

Metabolomics data acquisition

- Nuclear Magnetic Resonance (NMR)
- Gas chromatography-Mass spectrometry (GC-MS)
- Liquid chromatography-Mass spectrometry (LC-MS)

Metabolomics data acquisition

Nuclear Magnetic Resonance (NMR)

STRENGTHS

1. Non-destructive
2. Easily quantifiable
3. No prior separation required
4. Identification of novel compounds
5. High reproducibility

LIMITATIONS

1. Low-sensitivity
2. High detection limit
3. Large sample volumes

Metabolomics data acquisition

Gas chromatography-Mass spectrometry (GC-MS)

STRENGTHS

1. More mature and cost-effective
2. Excellent separation reproducibility
3. Availability of spectral libraries facilitates metabolite identification

LIMITATIONS

1. Sample derivatization required
2. Smaller coverage
3. Longer experimental times

Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS)

STRENGTHS

1. Broader range of metabolites
2. Simpler sample preparation
3. High sensitivity

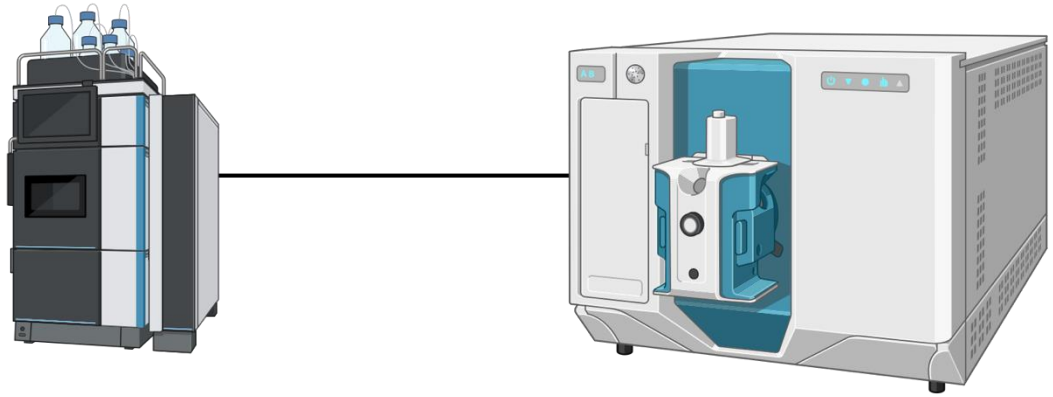
LIMITATIONS

1. Higher cost
2. Lower reproducibility
3. Reduced compatibility with volatile compounds

Preferred choice in metabolomics: **more than 70%** of the metabolomics studies were using LC-MS in 2022 .

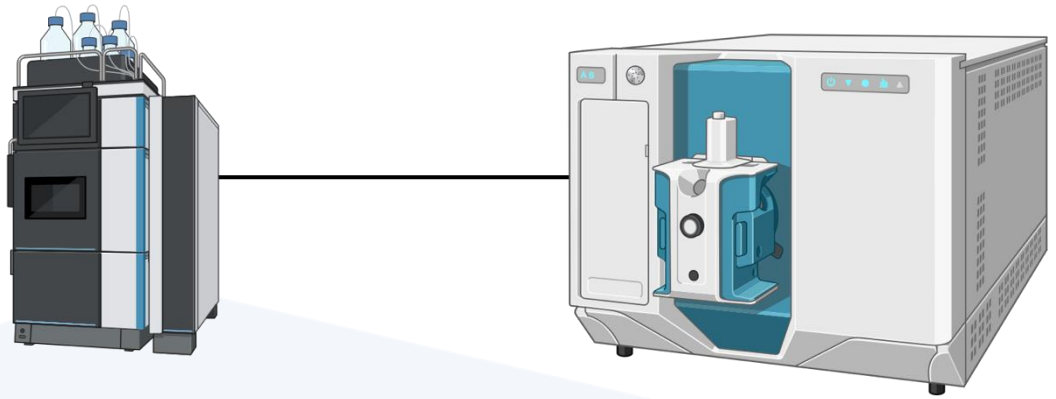
Metabolomics data acquisition

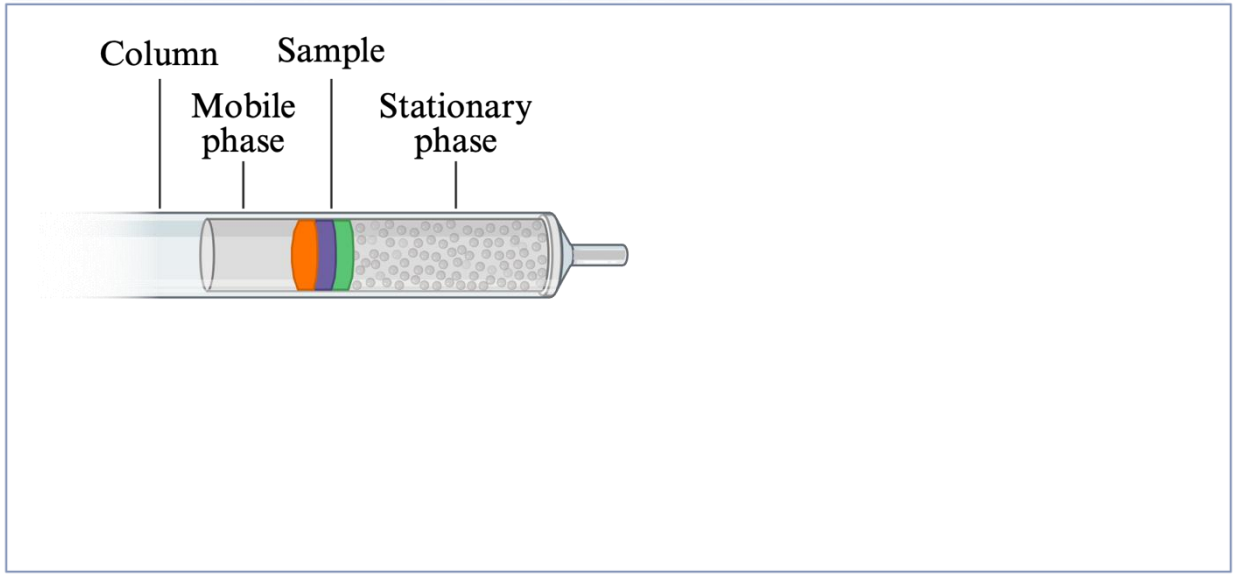
Liquid chromatography-Mass spectrometry (LC-MS/MS)



Metabolomics data acquisition

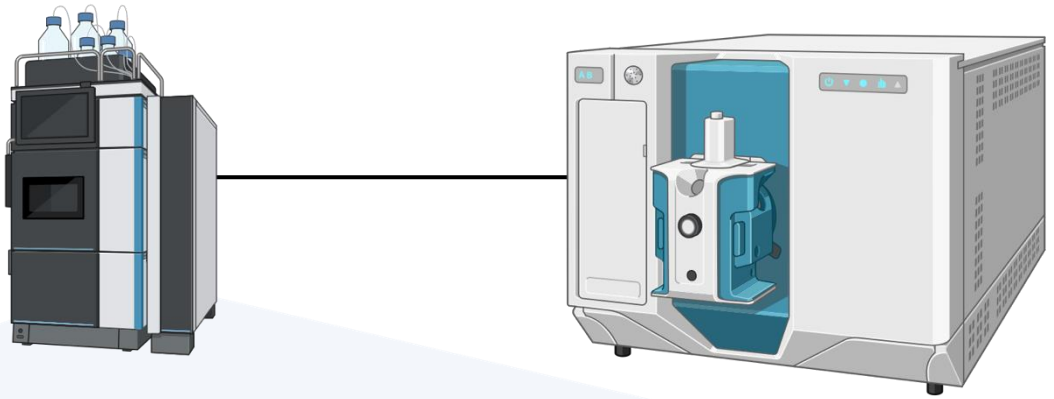
Liquid chromatography-Mass spectrometry (LC-MS/MS)

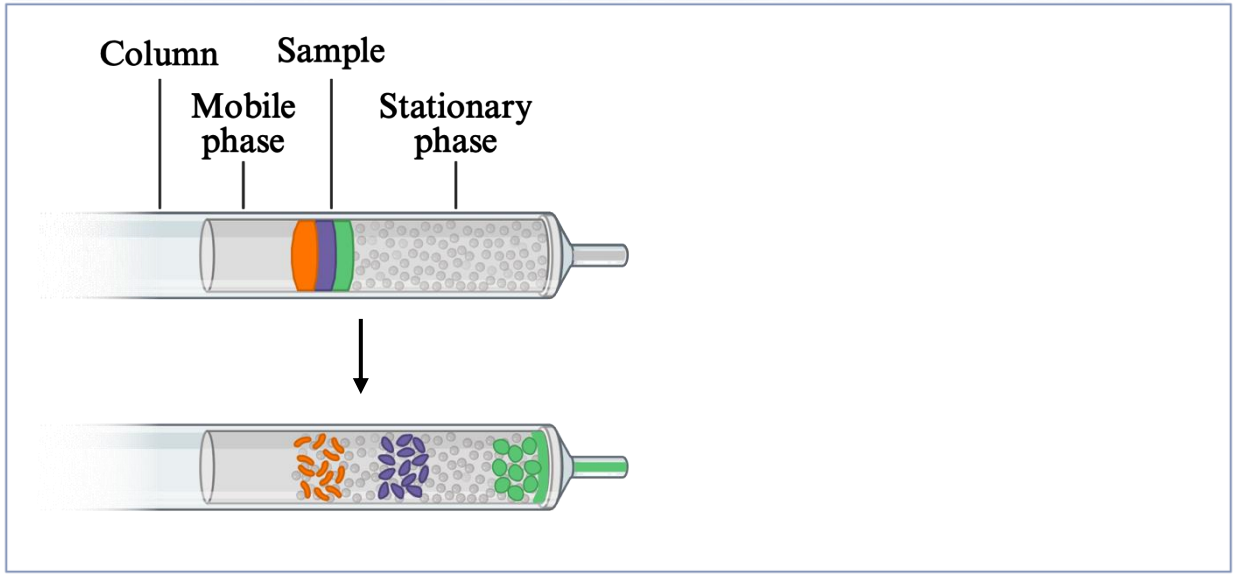




Metabolomics data acquisition

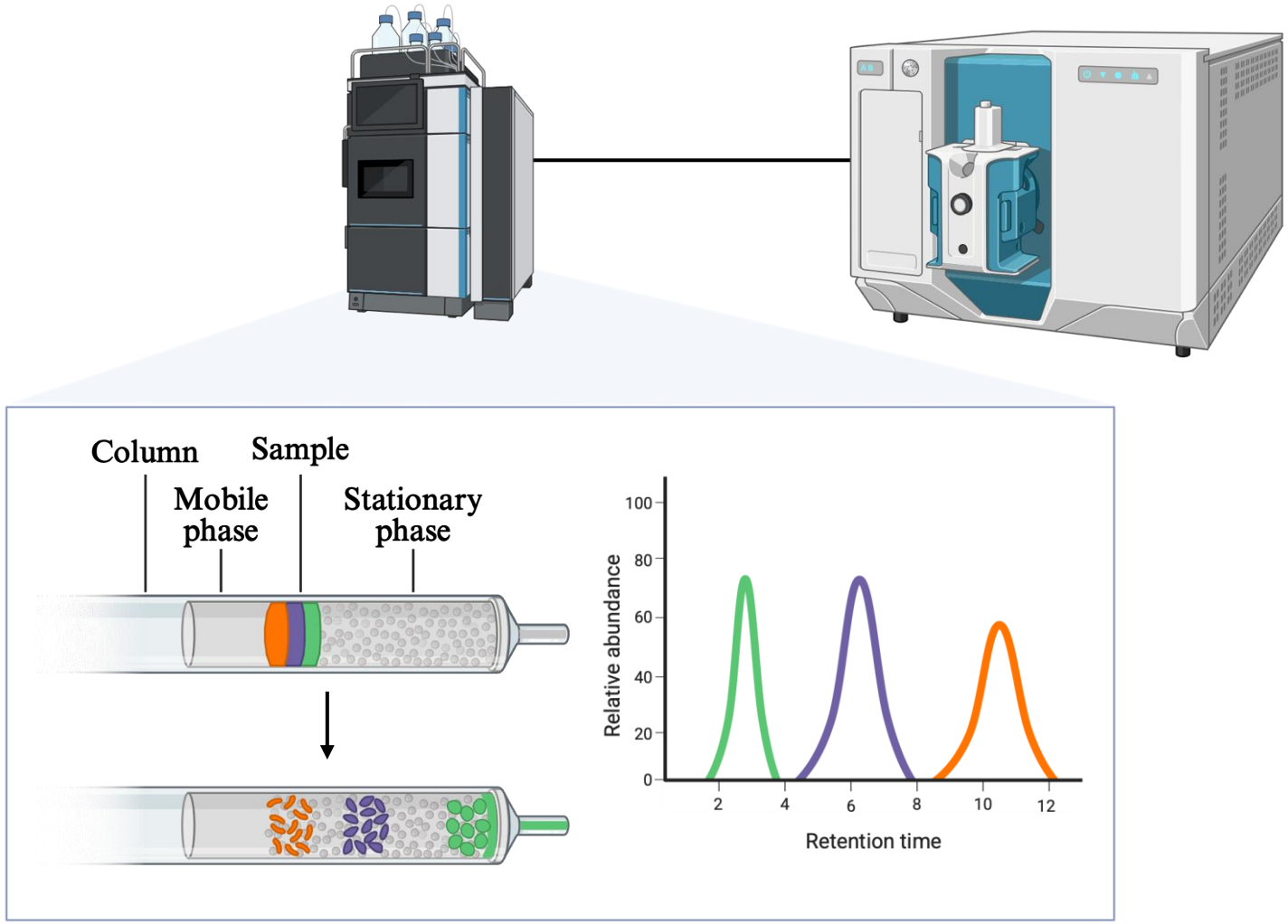
Liquid chromatography-Mass spectrometry (LC-MS/MS)





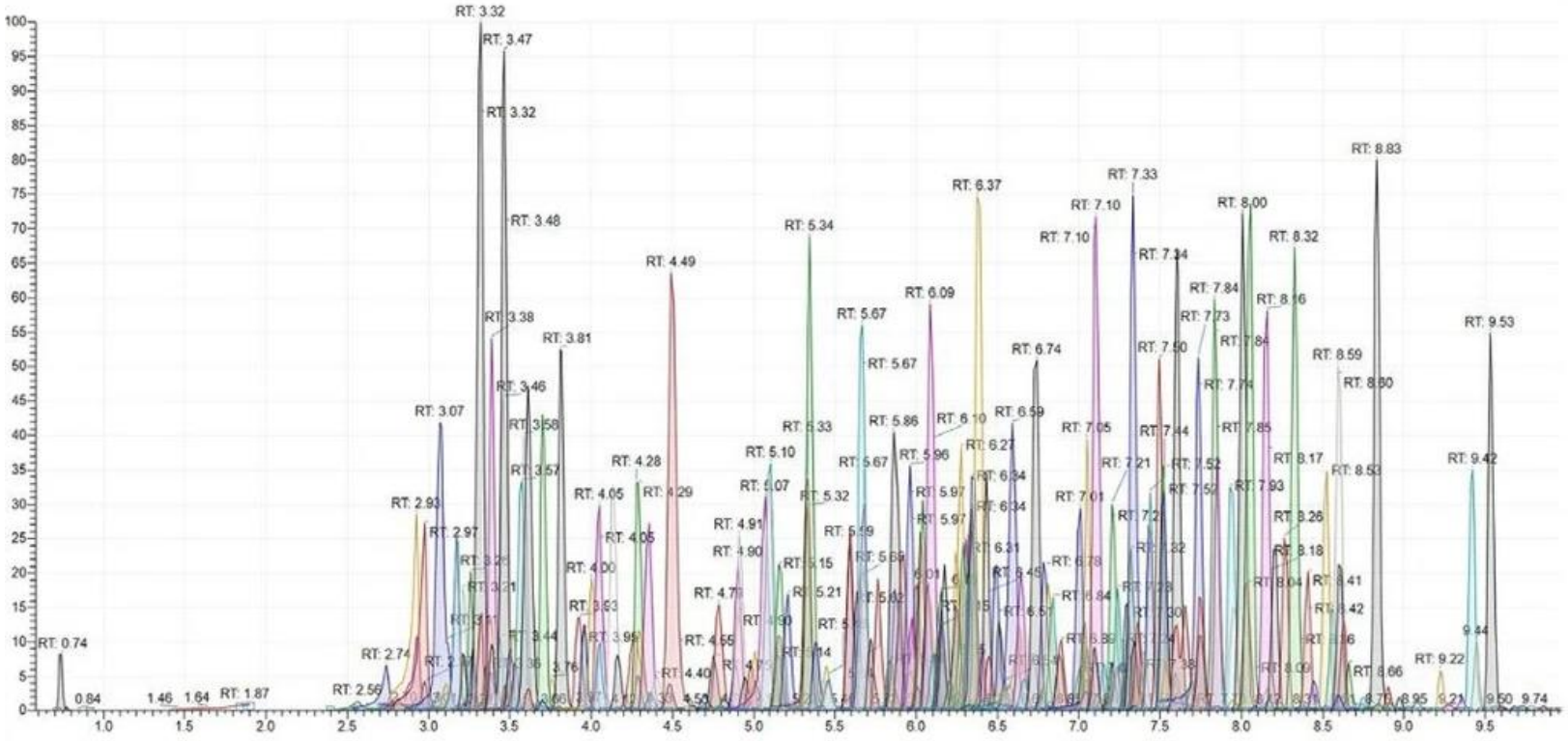
Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)



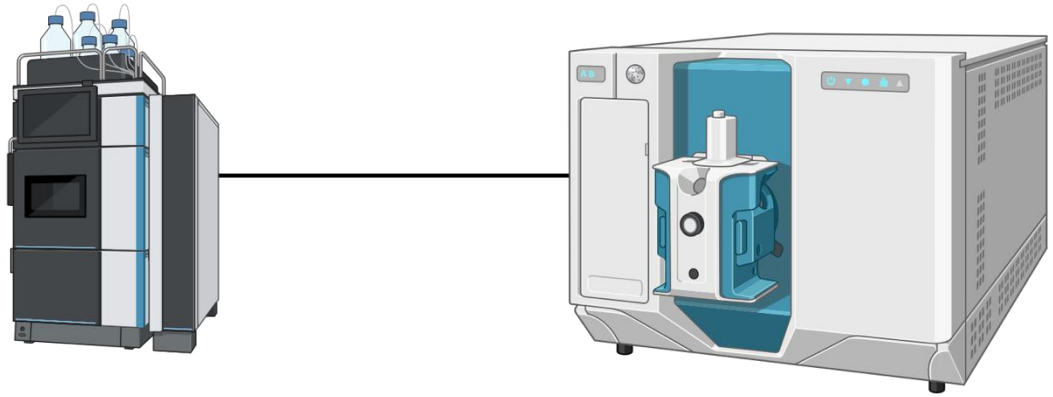
Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)



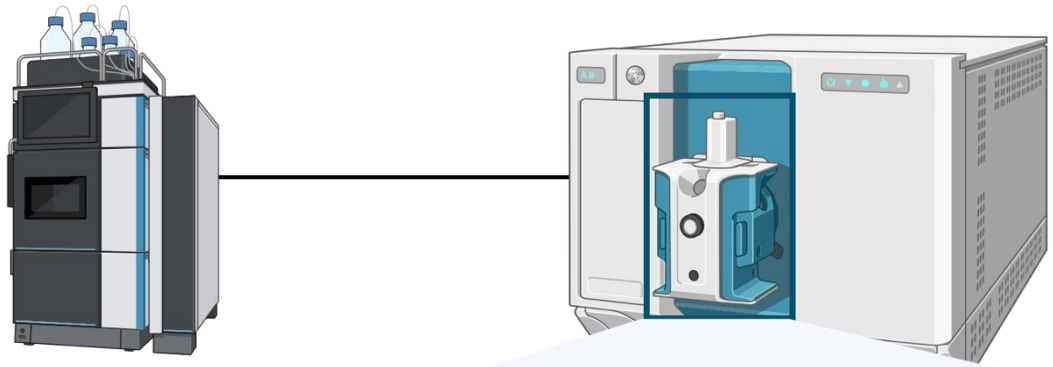
Metabolomics data acquisition

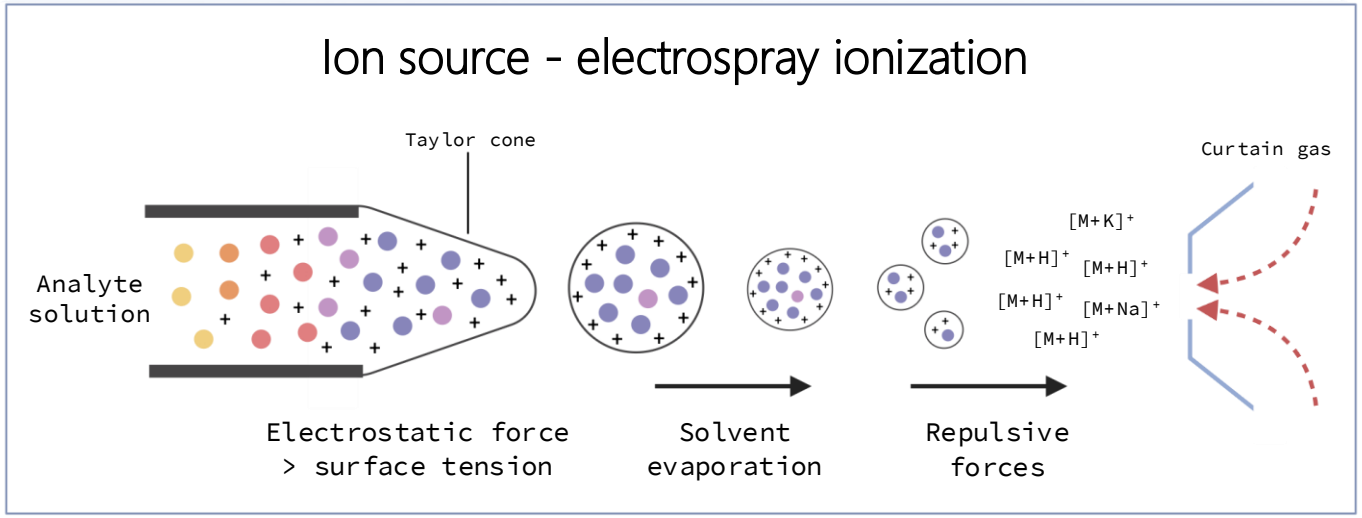
Liquid chromatography-Mass spectrometry (LC-MS/MS)



Metabolomics data acquisition

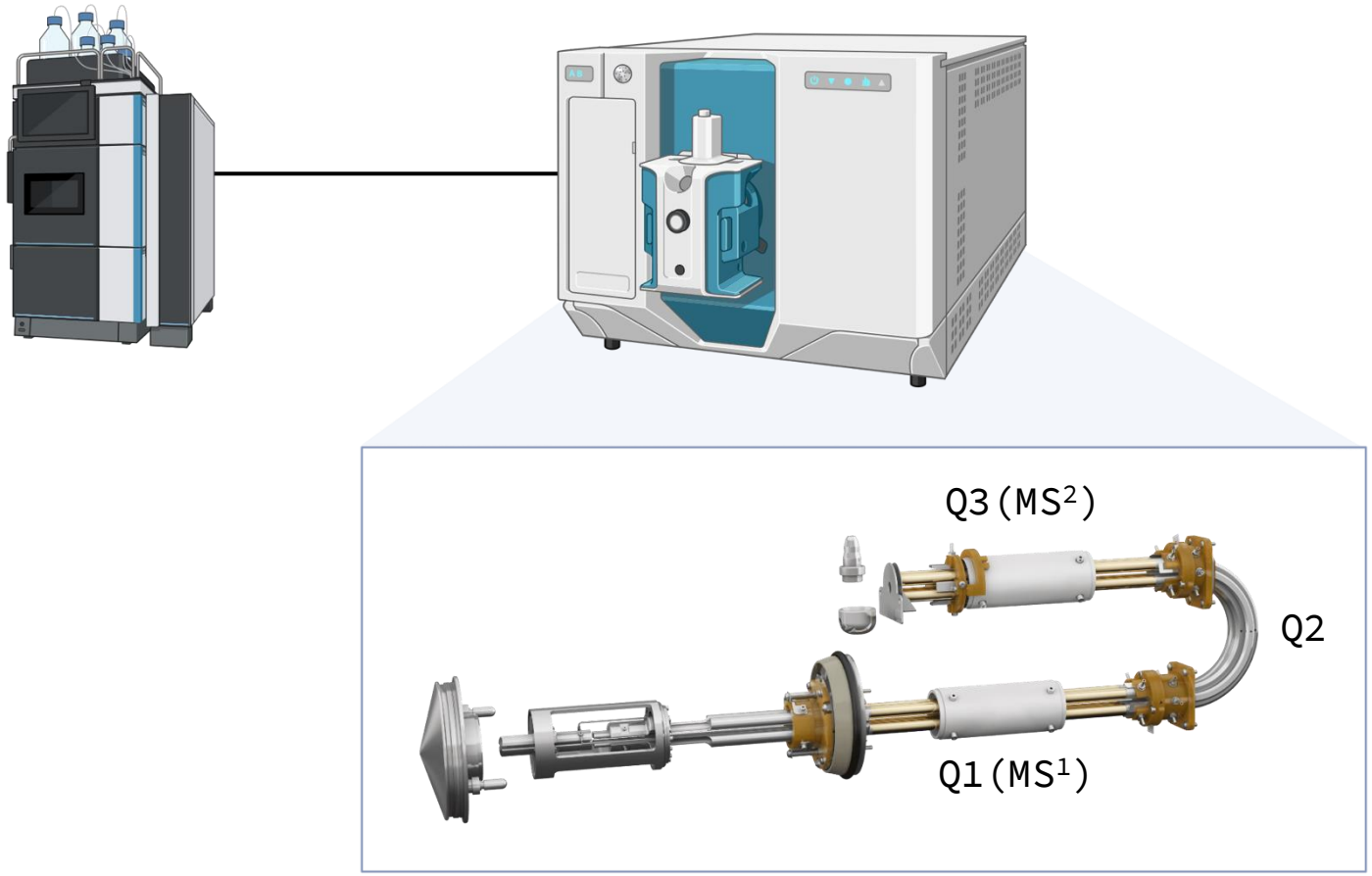
Liquid chromatography-Mass spectrometry (LC-MS/MS)





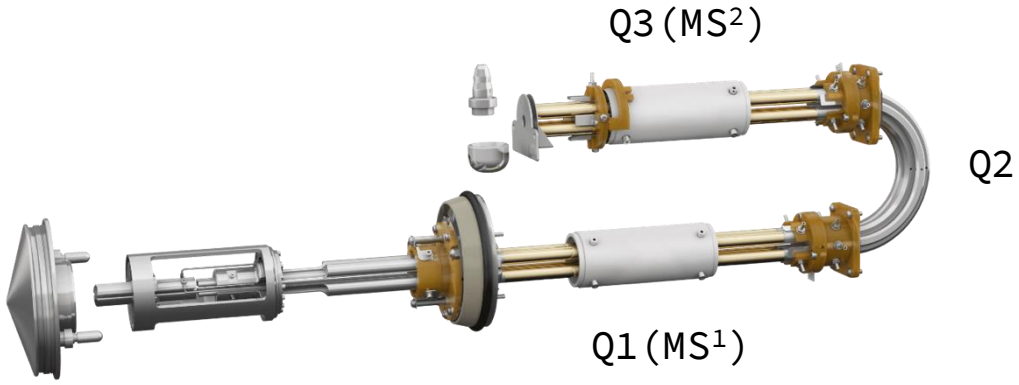
Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)



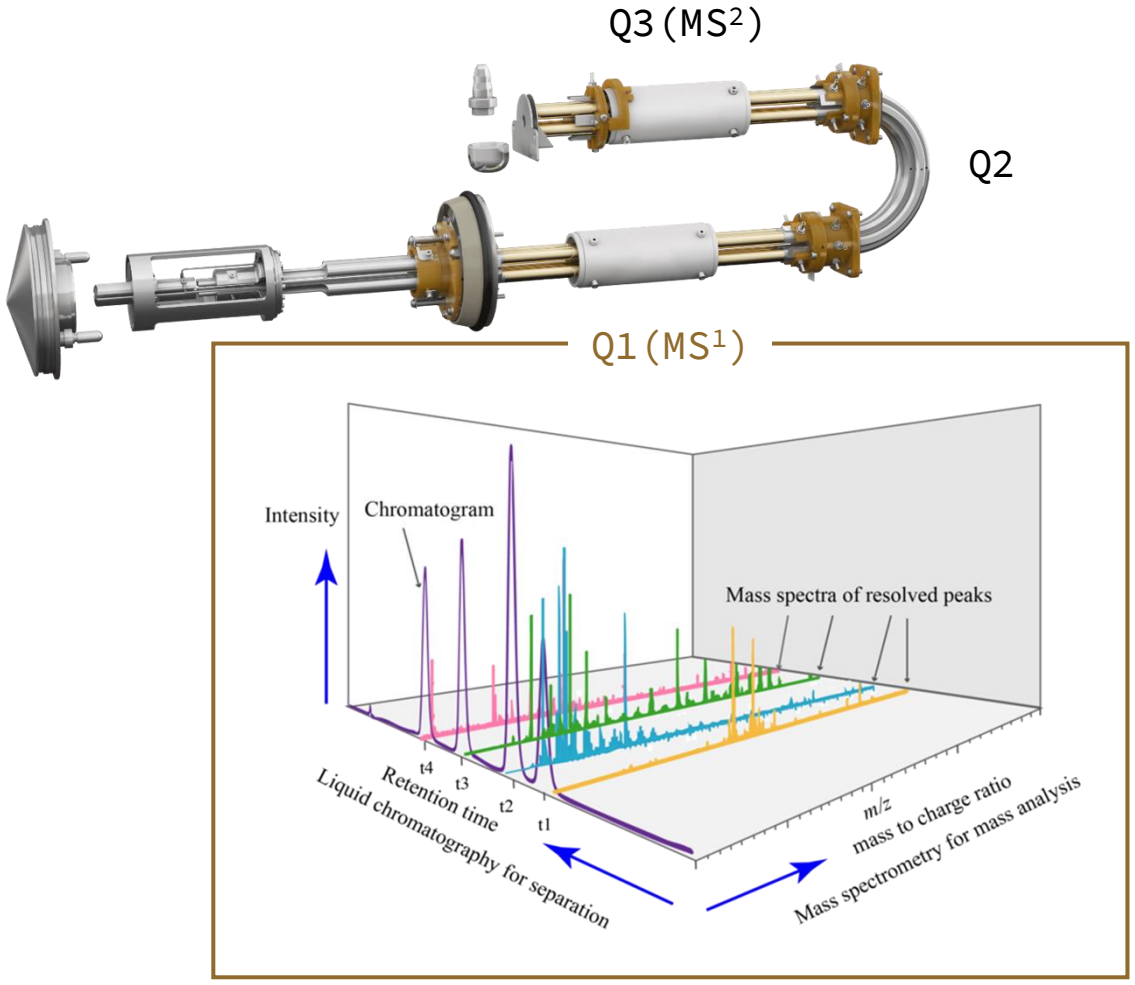
Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)



Metabolomics data acquisition

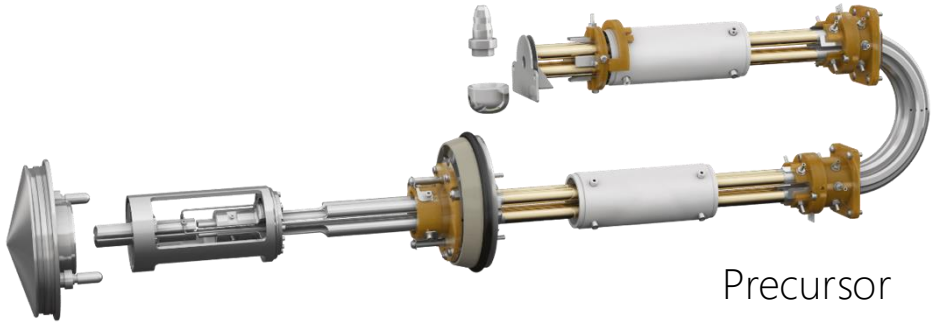
Liquid chromatography-Mass spectrometry (LC-MS/MS)



Metabolomics data acquisition

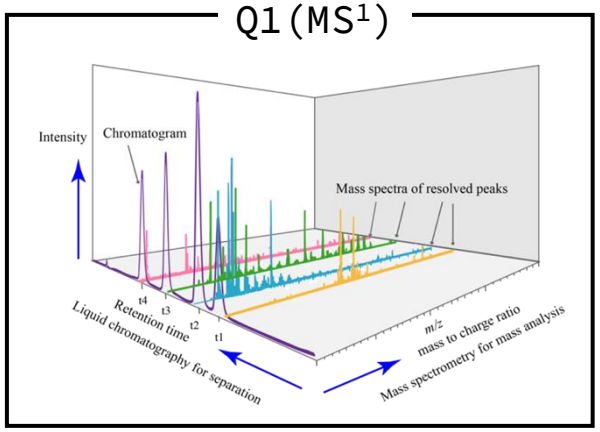
Liquid chromatography-Mass spectrometry (LC-MS/MS)

Q3(MS²)



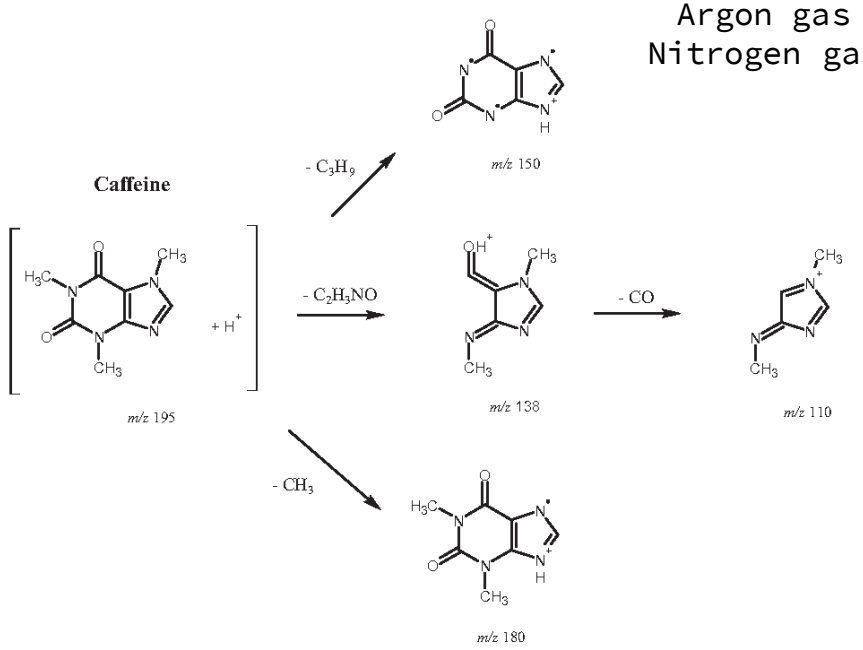
Precursor

Q1(MS¹)



Q2(collision chamber)

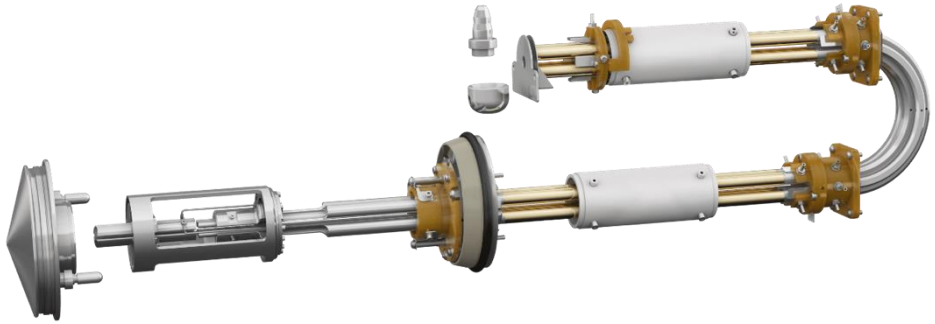
Argon gas
Nitrogen gas



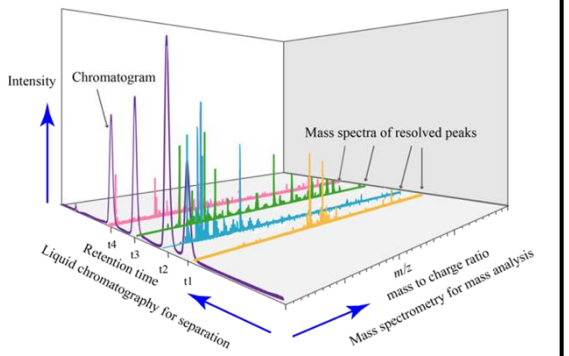
Pellegrini, Manuela et al. Rapid communications in mass spectrometry (2007).

Metabolomics data acquisition

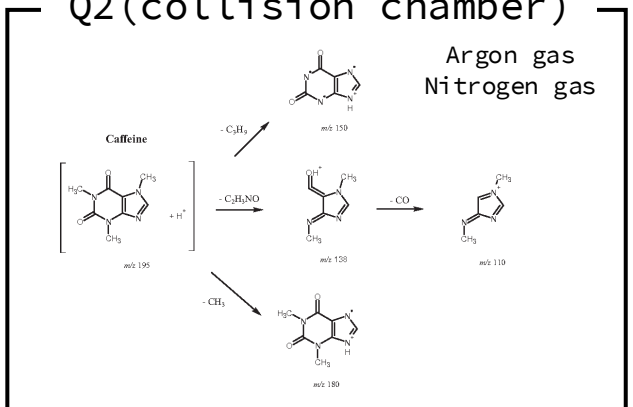
Liquid chromatography-Mass spectrometry (LC-MS/MS)



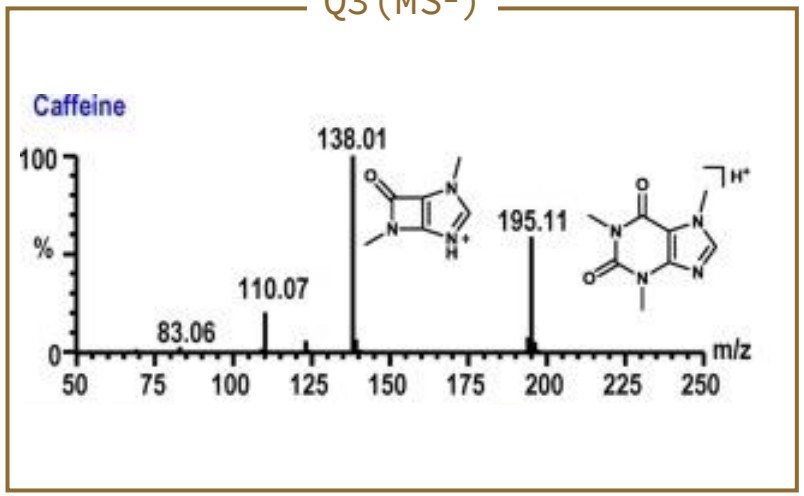
Q1(MS¹)



Q2(collision chamber)



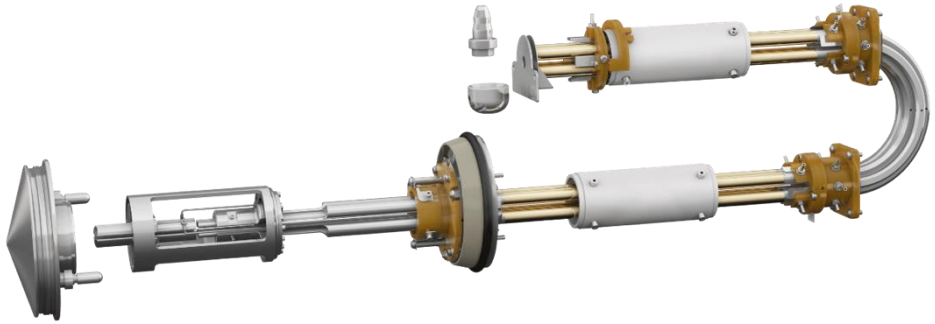
Q3(MS²)

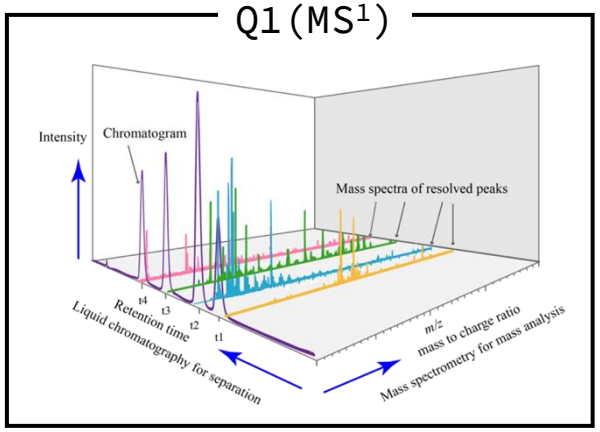


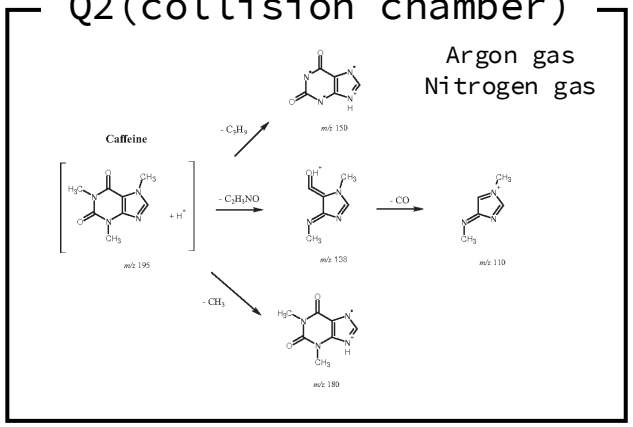
Pellegrini, Manuela et al. Rapid communications in mass spectrometry (2007).

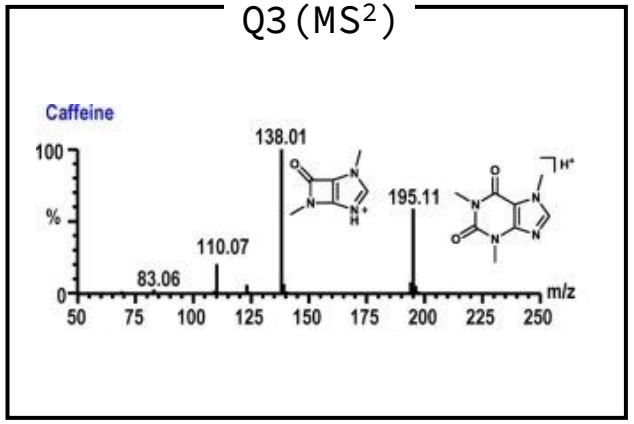
Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)









Pellegrini, Manuela et al. Rapid communications in mass spectrometry (2007).

Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)

Data Dependent Acquisition (DDA)
Ions are selected based on intensity.
The MS makes real-time decisions during the run: MS^1 scan → Selection of Top K most intense ions

Pros
High-quality MS/MS spectra

Cons
Biased toward abundant metabolites

Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)

Data Dependent Acquisition (DDA)
Ions are selected based on intensity.
The MS makes real-time decisions during the run: MS^1 scan → Selection of Top K most intense ions

Pros
High-quality MS/MS spectra

Cons
Biased toward abundant metabolites

Data Independent Acquisition (DIA)
It fragments all ions within predefined m/z windows.

Pros
Much more comprehensive coverage

Cons
MS/MS spectra are complex and overlapping

Metabolomics data acquisition

Liquid chromatography-Mass spectrometry (LC-MS/MS)

Data Dependent Acquisition (DDA)
Ions are selected based on intensity.
The MS makes real-time decisions during the run: MS^1 scan → Selection of Top K most intense ions

Pros
High-quality MS/MS spectra

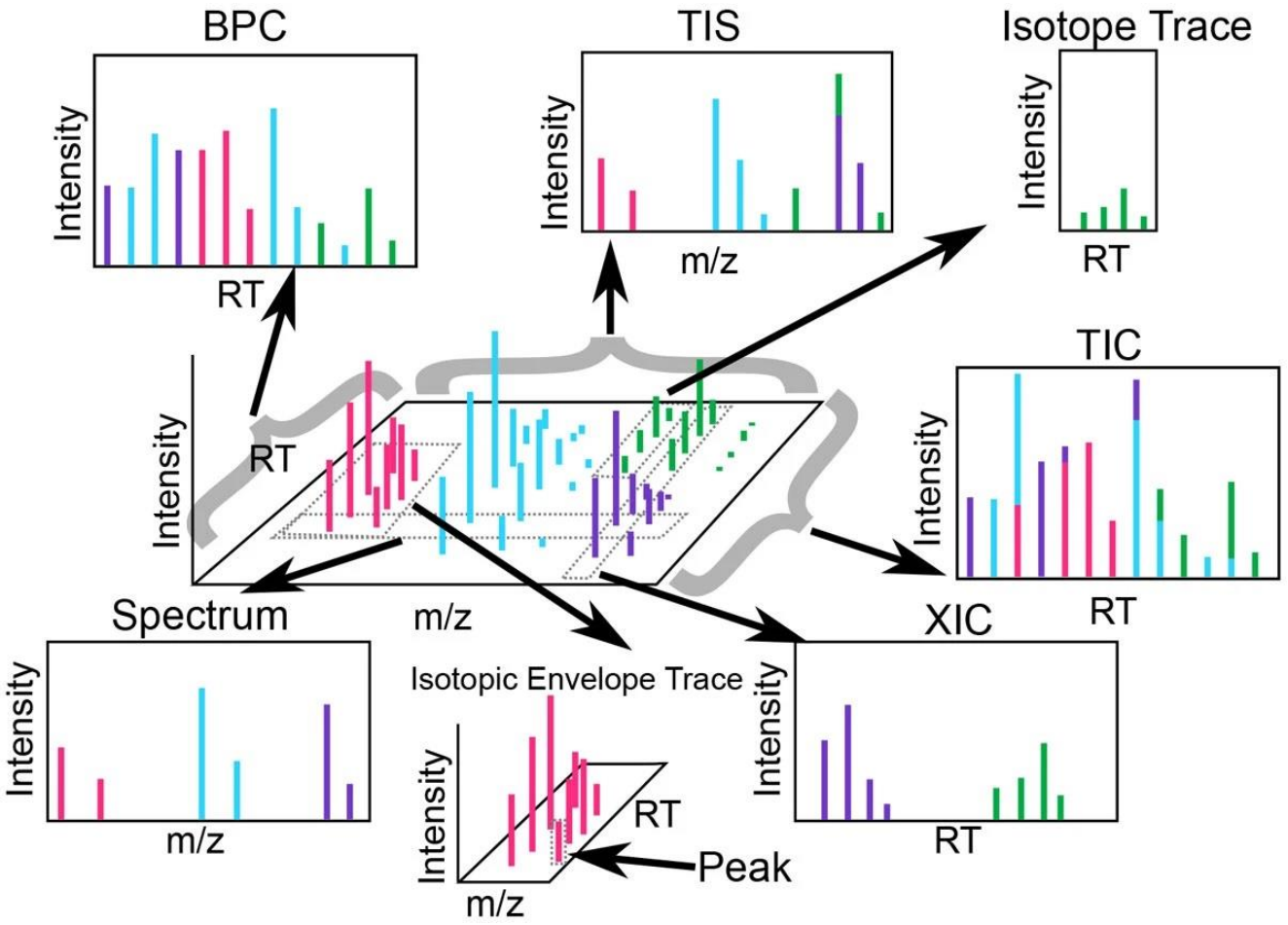
Cons
Biased toward abundant metabolites

Data Independent Acquisition (DIA)
It fragments all ions within predefined m/z windows.

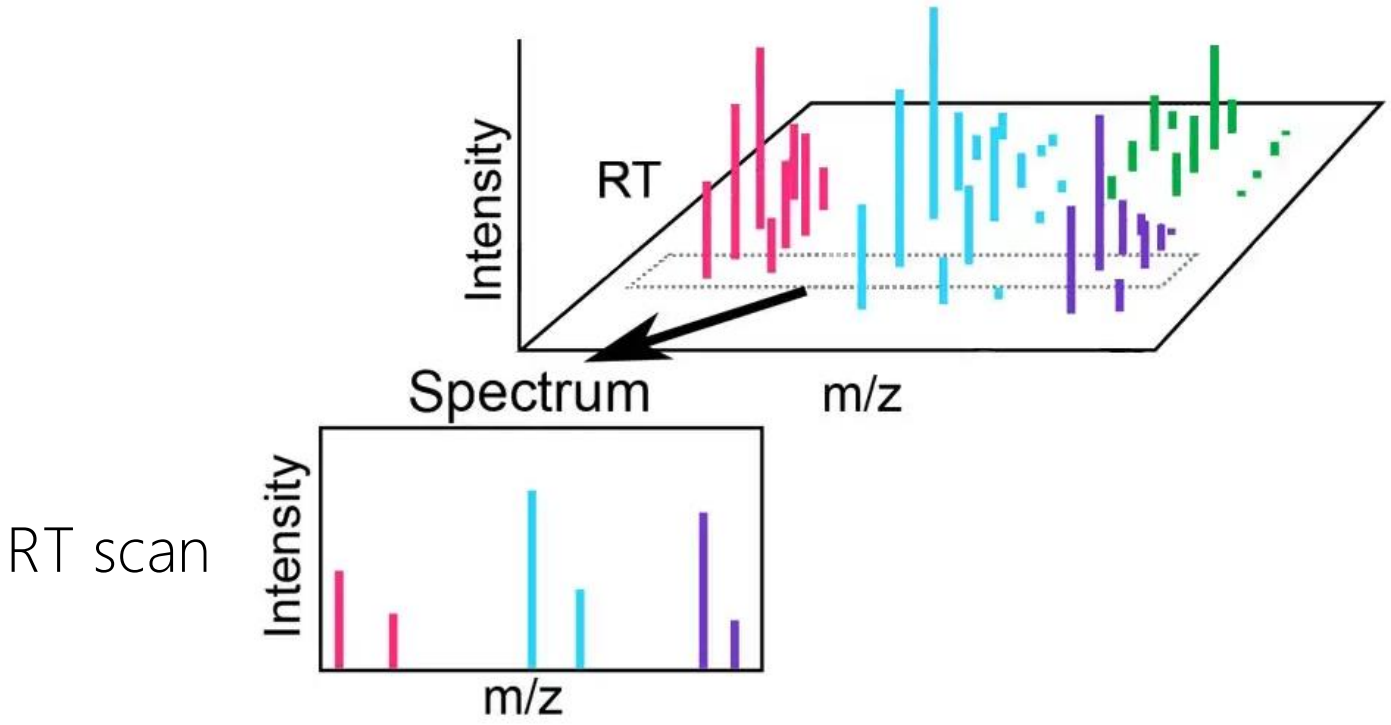
Pros
Much more comprehensive coverage

Cons
MS/MS spectra are complex and overlapping

LC-MS data

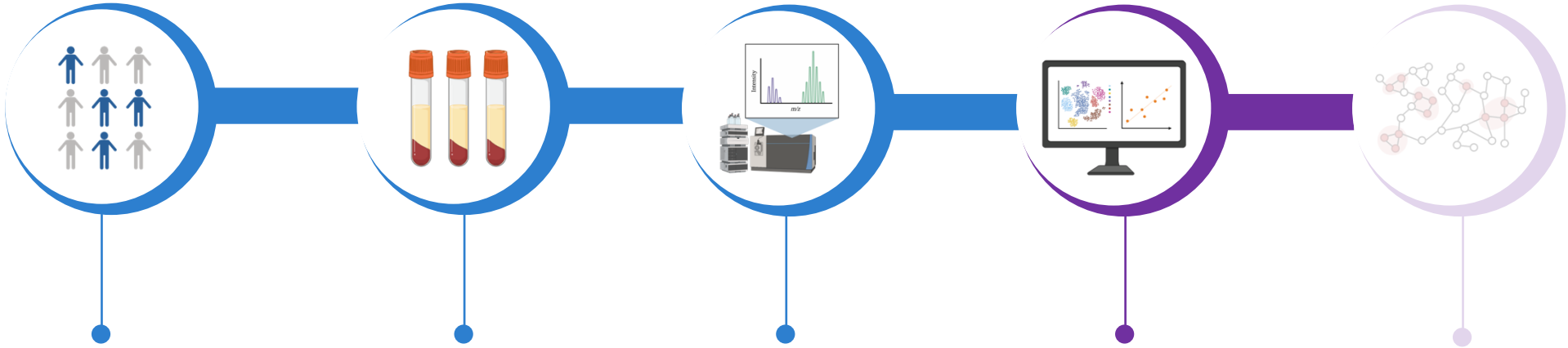

BPC
Intensity
RT
TIS
Intensity
m/z
Isotope Trace
Intensity
RT
TIC
Intensity
RT
Intensity
RT
m/z
Spectrum
Intensity
m/z
Isotopic Envelope Trace
Intensity
RT
Peak
m/z
XIC
Intensity
RT

LC-MS data



Metabolomics pipeline

Problem formulation



Biological interpretation

Experimental design

Confounding factors:

- Analytical
- Physiological

Sample handling

- Collection, storage and preparation
- Tissues, cells, biofluids
- Isolate molecules of interest.

Data acquisition

- LC-MS
- GC-MS
- NMR

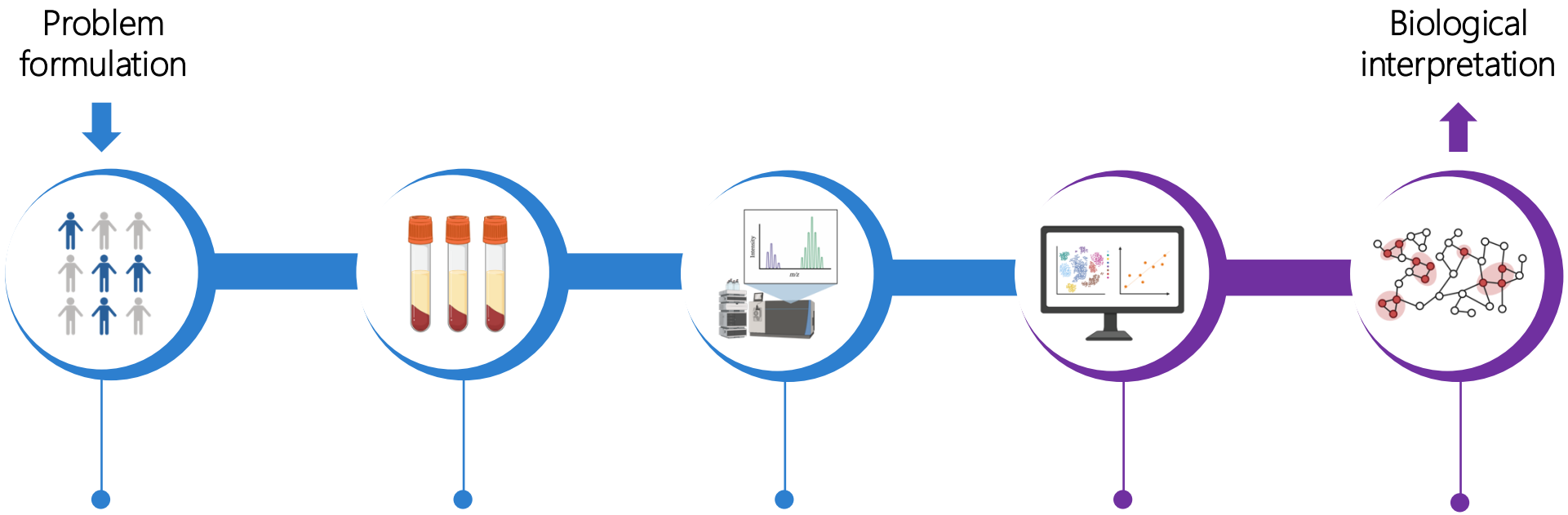
Data processing

- Pre-processing
- Data filtering and normalization
- Statistical analysis
- Peak annotation

Functional enrichment

- Over-representation analysis
- Functional class scoring
- Pathway topology

Metabolomics pipeline



Experimental design

Confounding factors:

- Analytical
- Physiological

Sample handling

- Collection, storage and preparation
- Tissues, cells, biofluids
- Isolate molecules of interest.

Data acquisition

- LC-MS
- GC-MS
- NMR

Data processing

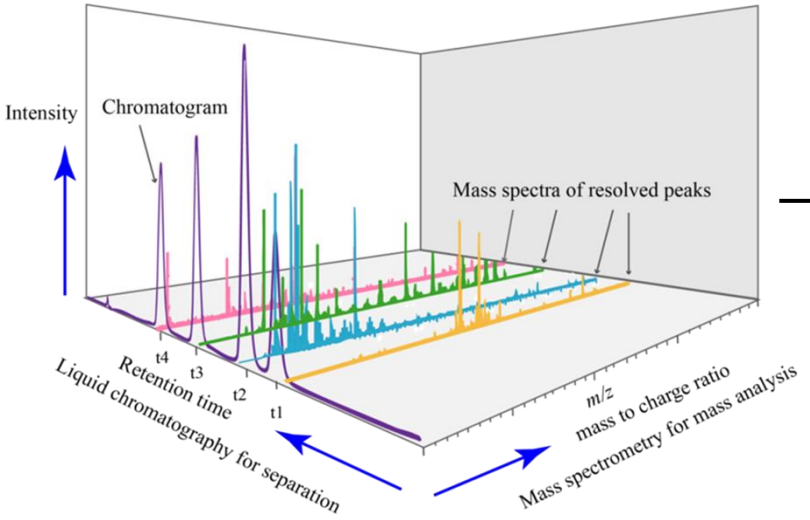
- Pre-processing
- Data filtering and normalization
- Statistical analysis
- Peak annotation

Functional enrichment

- Over-representation analysis
- Functional class scoring
- Pathway topology

Pre-processing

The goal is to convert **raw LC-MS data** into a **table of n peaks** defined by m/z, rt and an intensity value for each sample k.



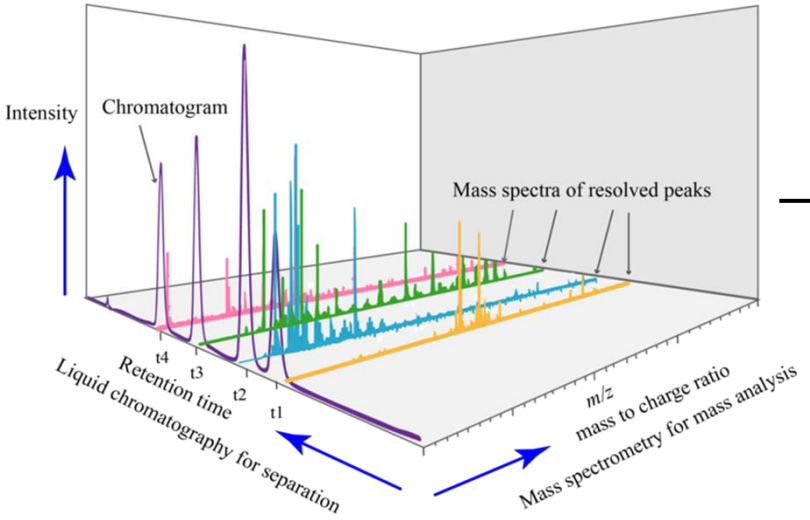
— Pre-processing →

m/z	RT	I_1	…	I_k
m/z_1	RT_1	I_{11}	…	I_{1k}
m/z_2	RT_2	I_{21}	…	I_{2k}
…	…	…	…	…
m/z_n	RT_n	I_{n1}	…	I_{nk}

1. Peak detection
2. Peak alignment and matching

Pre-processing

The goal is to convert **raw LC-MS data** into a **table of n peaks** defined by m/z, rt and an intensity value for each sample k.



— Pre-processing →

m/z	RT	I_1	…	I_k
m/z_1	RT_1	I_{11}	…	I_{1k}
m/z_2	RT_2	I_{21}	…	I_{2k}
…	…	…	…	…
m/z_n	RT_n	I_{n1}	…	I_{nk}

1. Peak detection
2. Peak alignment and matching

EXAMPLE

Positive acquisition mode: 2545 peaks

Negative acquisition mode: 2187 peaks

Pre-processing

Centroiding

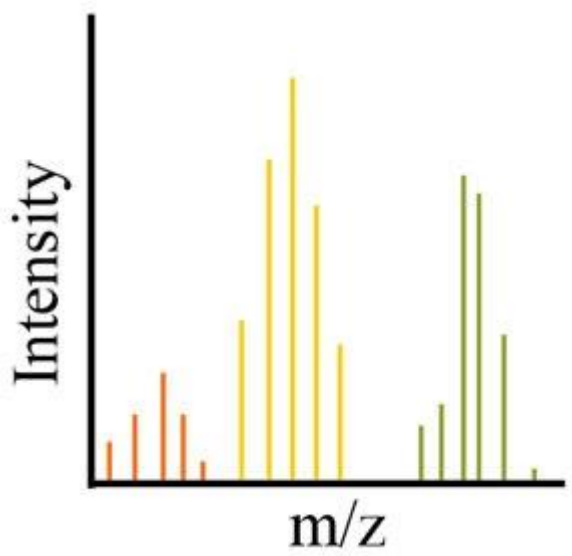
- The MS has finite resolving power → ions are detected as profile peaks around their true m/z rather than as a single exact m/z
- Centroiding reduces the amount of data without much loss of information.
- The next processing steps require centroided data.
- Many manufacturers already apply centroiding of the profile data (during acquisition or immediately after).

Pre-processing

Centroiding with OpenMS PeakPickerHiRes

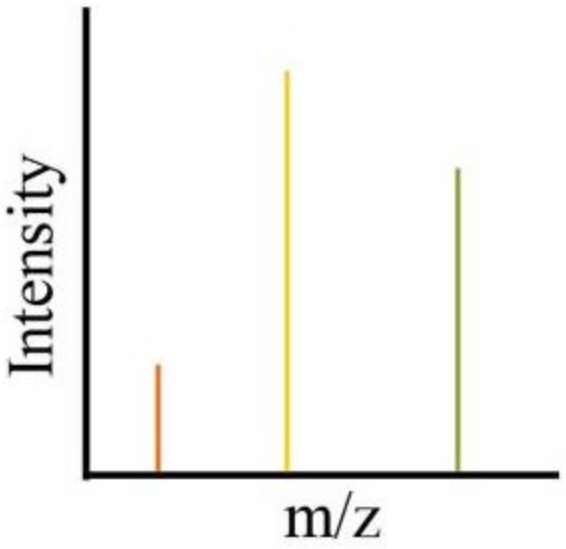
Profile mode

MS1 spectrum (scan at RT = 5.2 min)



Centroid mode

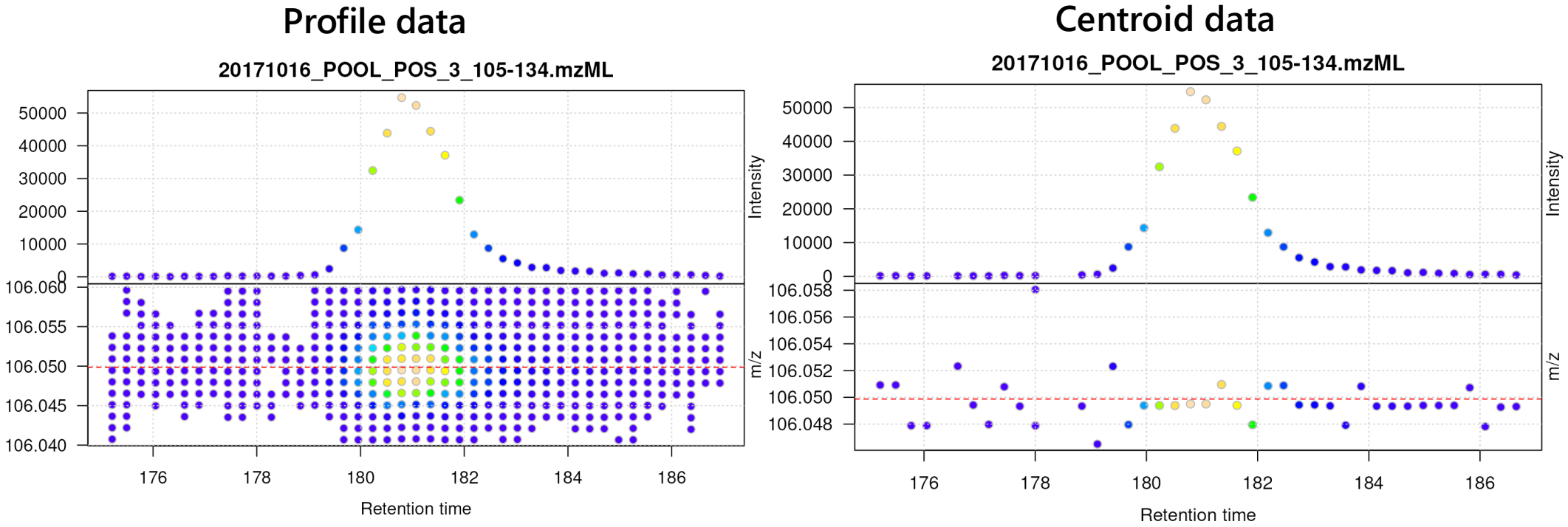
MS1 spectrum (scan at RT = 5.2 min)



An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics

Pre-processing

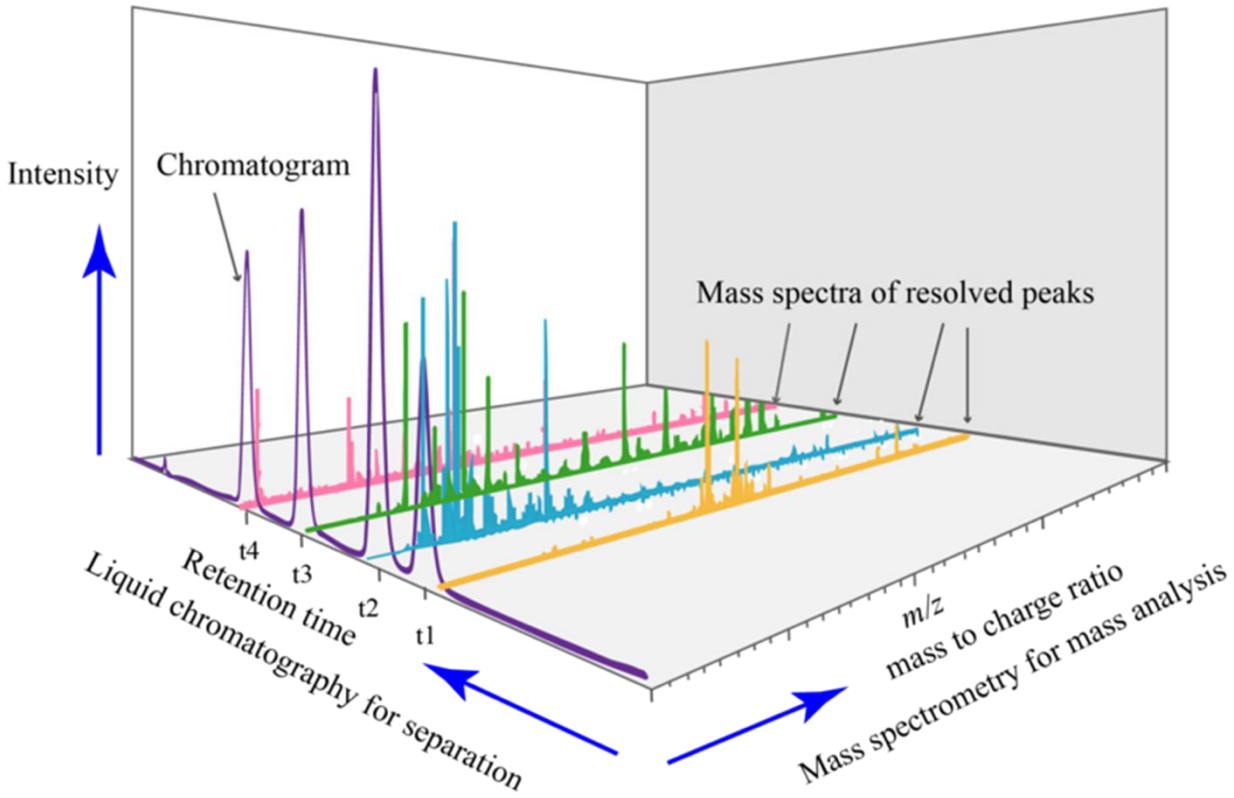
Peak detection



The **goal** is to identify the **chromatographic peaks** (peaks across the retention time axis) and return a **peak feature** defined by m/z, rt and intensity per sample.

Pre-processing

Peak detection with OpenMS FeatureFinderMetabo



1. Mass trace detection

   `MassTraceDetection()`

2. Elution peak detection

   `ElutionPeakDetection()`

3. Feature assembly

   `FeatureFindingMetabo()`

Kenar, E et al. (2014). Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. Molecular & cellular proteomics.

Pre-processing

Peak detection with OpenMS FeatureFinderMetabo

1. Mass trace detection

$$P = \{p_k\} \text{ where } p_k = \{rt_k, m/z_k, I_k\}$$

Scan	RT	m/z	Intensity
1	240	181.0710	100
2	241	181.0711	200
3	242	181.0709	500
4	243	181.0710	300
1	240	455.2802	80
2	241	455.2800	120
3	242	455.2803	300

Pre-processing

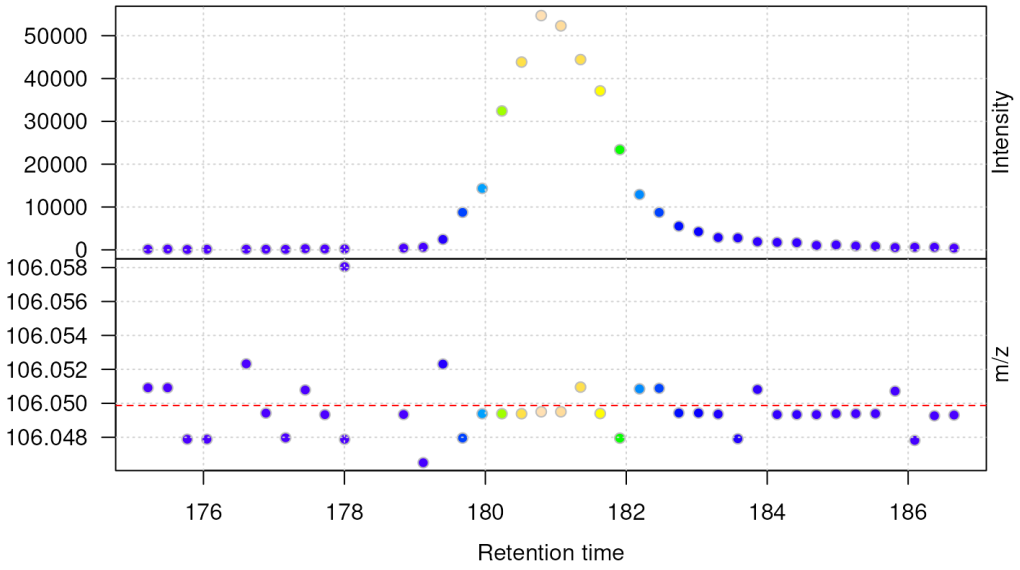
Peak detection with OpenMS FeatureFinderMetabo

1. Mass trace detection

$P = \{p_k\}$ where $p_k = \{rt_k, m/z_k, I_k\}$

Scan	RT	m/z	Intensity
1	240	181.0710	100
2	241	181.0711	200
3	242	181.0709	500
4	243	181.0710	300
1	240	455.2802	80
2	241	455.2800	120
3	242	455.2803	300

Mass trace T: List of n peaks $p_k \in P$ with similar m/z and that occur in adjacent scans of an LC-MS run (they elute at similar RT).



Pre-processing

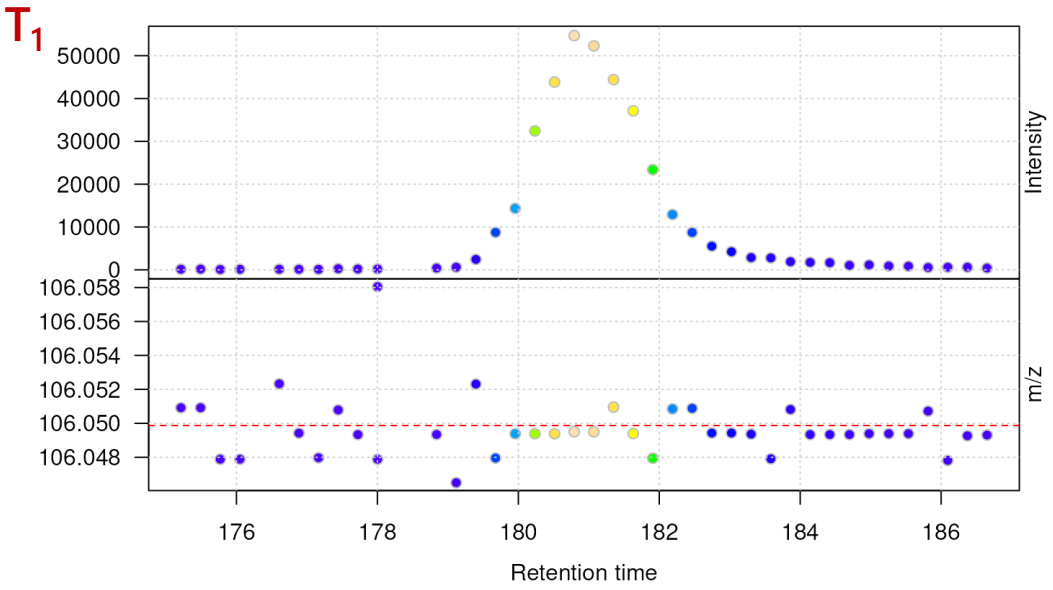
Peak detection with OpenMS FeatureFinderMetabo

1. Mass trace detection

$P = \{p_k\}$ where $p_k = \{rt_k, m/z_k, I_k\}$

Scan	RT	m/z	Intensity
1	240	181.0710	100
2	241	181.0711	200
3	242	181.0709	500
4	243	181.0710	300
1	240	455.2802	80
2	241	455.2800	120
3	242	455.2803	300

Mass trace T: List of n peaks $p_k \in P$ with similar m/z and that occur in adjacent scans of an LC-MS run (they elute at similar RT).



Pre-processing

Peak detection with OpenMS FeatureFinderMetabo

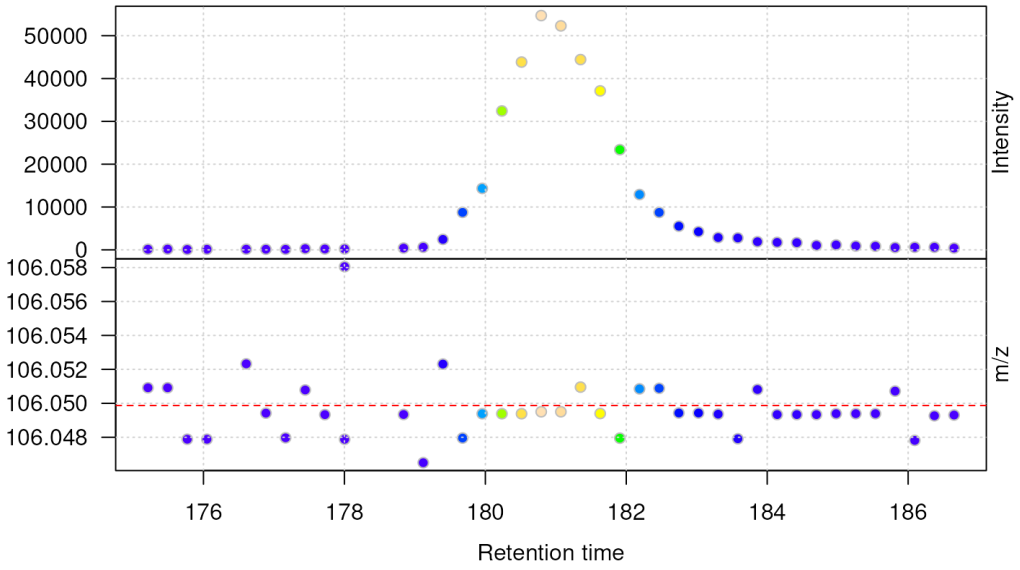
1. Mass trace detection

$P = \{p_k\}$ where $p_k = \{rt_k, m/z_k, I_k\}$

Scan	RT	m/z	Intensity
1	240	181.0710	100
2	241	181.0711	200
3	242	181.0709	500
4	243	181.0710	300
1	240	455.2802	80
2	241	455.2800	120
3	242	455.2803	300

T_2

Mass trace T: List of n peaks $p_k \in P$ with similar m/z and that occur in adjacent scans of an LC-MS run (they elute at similar RT).



Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

2. Elution peak detection

Does this mass trace contain multiple chromatographic peaks?

Pre-processing – Peak detection

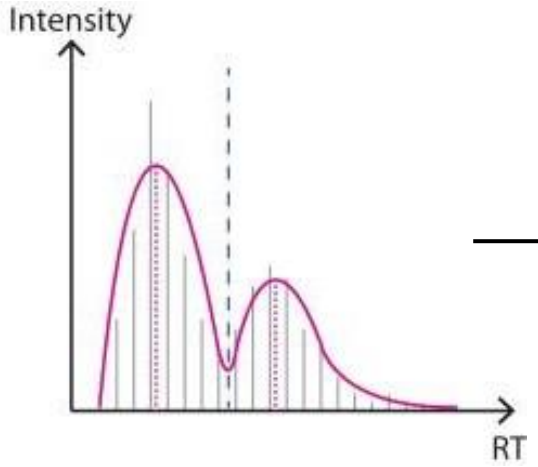
OpenMS FeatureFinderMetabo

2. Elution peak detection

Does this mass trace contain multiple chromatographic peaks?

Two compounds can have:

- Almost identical m/z
- Overlap in RT



1. Smooth the signal with LOWESS
2. Find:
   - Local maxima (peaks)
   - Minima between maxima (separation points)
3. Each split becomes a final mass trace

Intensity: Area under the chromatographic peak
Retention time: Position of the chromatographic peak maximum

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature
(find isotopic patterns)?

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

Scan	RT	m/z	Intensity
1	240	181.0710	100
2	241	181.0711	200
3	242	181.0709	500
4	243	181.0710	300
1	240	455.2802	80
2	241	455.2800	120
3	242	455.2803	300

Trace	RT	m/z
T_1	241.85	181.0710
T_2	243	455.2800

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

Trace	RT	m/z
T_1	241.85	181.0710
T_3	242.75	182.0740
T_4	242.02	183.0770
T_5	242.52	184.0800

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

^{13}C vs ^{12}C

Trace	RT	m/z	
T_1	241.85	181.0710	
			+1.003
T_3	242.75	182.0740	
			+1.003
T_4	242.02	183.0770	
			+1.003
T_5	242.52	184.0800	

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

^{13}C vs ^{12}C

Trace	RT	m/z	
T_1	241.85	181.0710	
			+1.003
T_3	242.75	182.0740	
			+1.003
T_4	242.02	183.0770	
			+1.003
T_5	242.52	184.0800	

Mass traces co-elute
Correct m/z distances
Correct isotopic abundance ratios

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

Trace	RT	m/z
T_1	241.85	181.0710
T_3	242.75	182.0740
T_4	242.02	183.0770
T_5	242.52	184.0800

^{13}C vs ^{12}C

+1.003

+1.003

+1.003

→ Feature 1

Mass traces co-elute
Correct m/z distances
Correct isotopic abundance ratios

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

3. Feature assembly

Which traces belong together as one metabolite feature (find isotopic patterns)?

Trace	RT	m/z
T_1	241.85	181.0710
T_3	242.75	182.0740
T_4	242.02	183.0770
T_5	242.52	184.0800

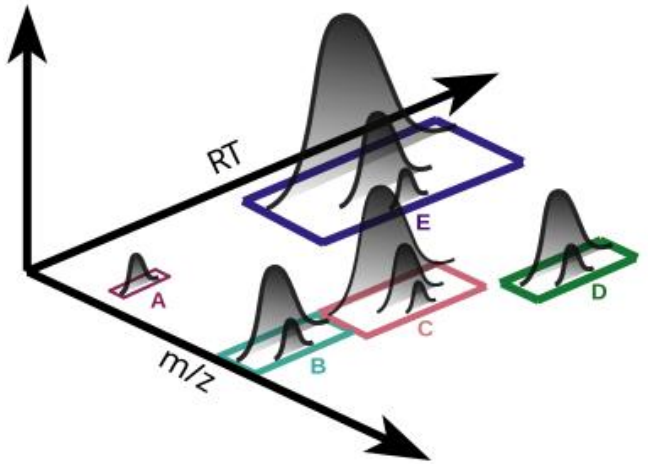
^{13}C vs ^{12}C

+1.003

+1.003

+1.003

→ Feature 1



Mass traces co-elute
Correct m/z distances
Correct isotopic abundance ratios

Pre-processing – Peak detection

OpenMS FeatureFinderMetabo

What do we have now?

Sample 1

m/z	RT	Intensity
m/z_1	RT_1	I_1
m/z_2	RT_2	I_2
…	…	…
m/z_n	RT_n	I_n

Sample 2

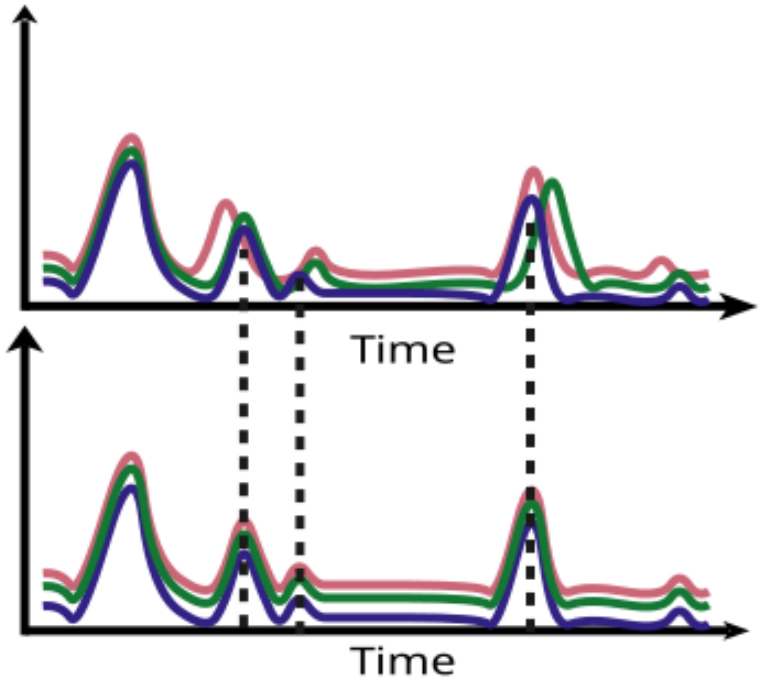
m/z	RT	Intensity
m/z_1	RT_1	I_1
m/z_2	RT_2	I_2
…	…	…
m/z_t	RT_t	I_t

Sample k

m/z	RT	Intensity
m/z_1	RT_1	I_1
m/z_2	RT_2	I_2
…	…	…
m/z_m	RT_m	I_m

Pre-processing – Peak alignment and matching

Goal: Match features across different samples together



1. Peak alignment across the retention time axis

   `MapAlignmentAlgorithmPoseClustering()`

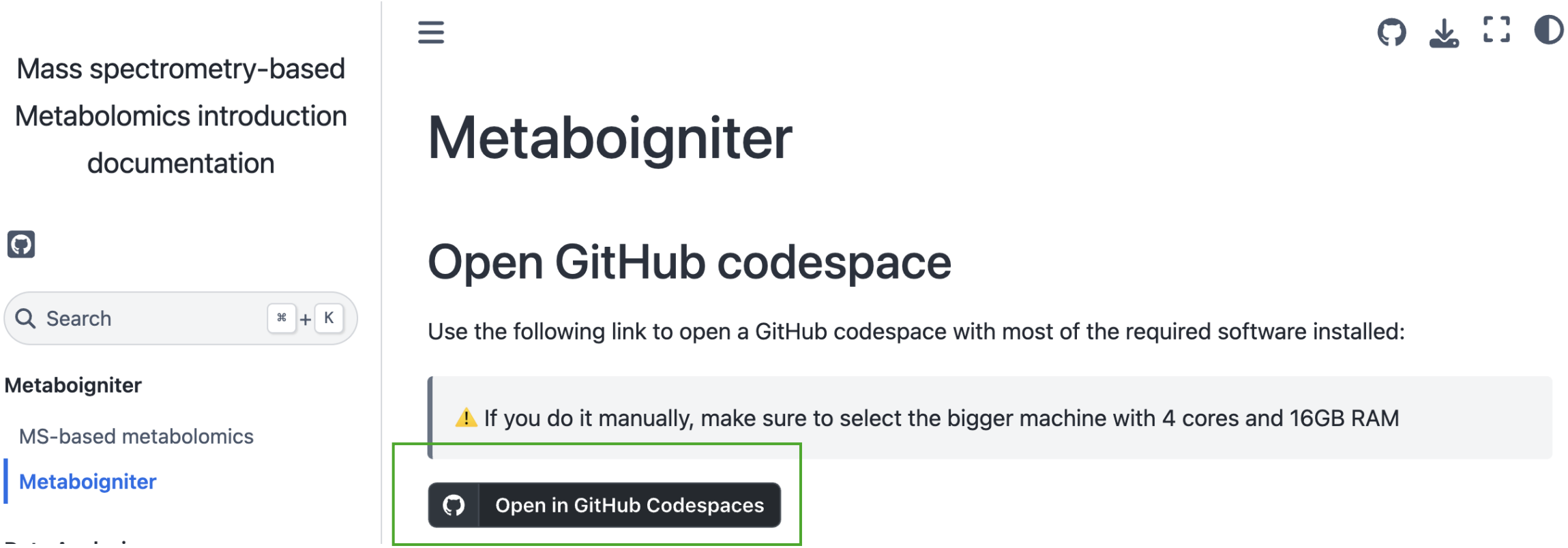
2. Grouping corresponding features in multiple runs

   `FeatureGroupingAlgorithmKD()`

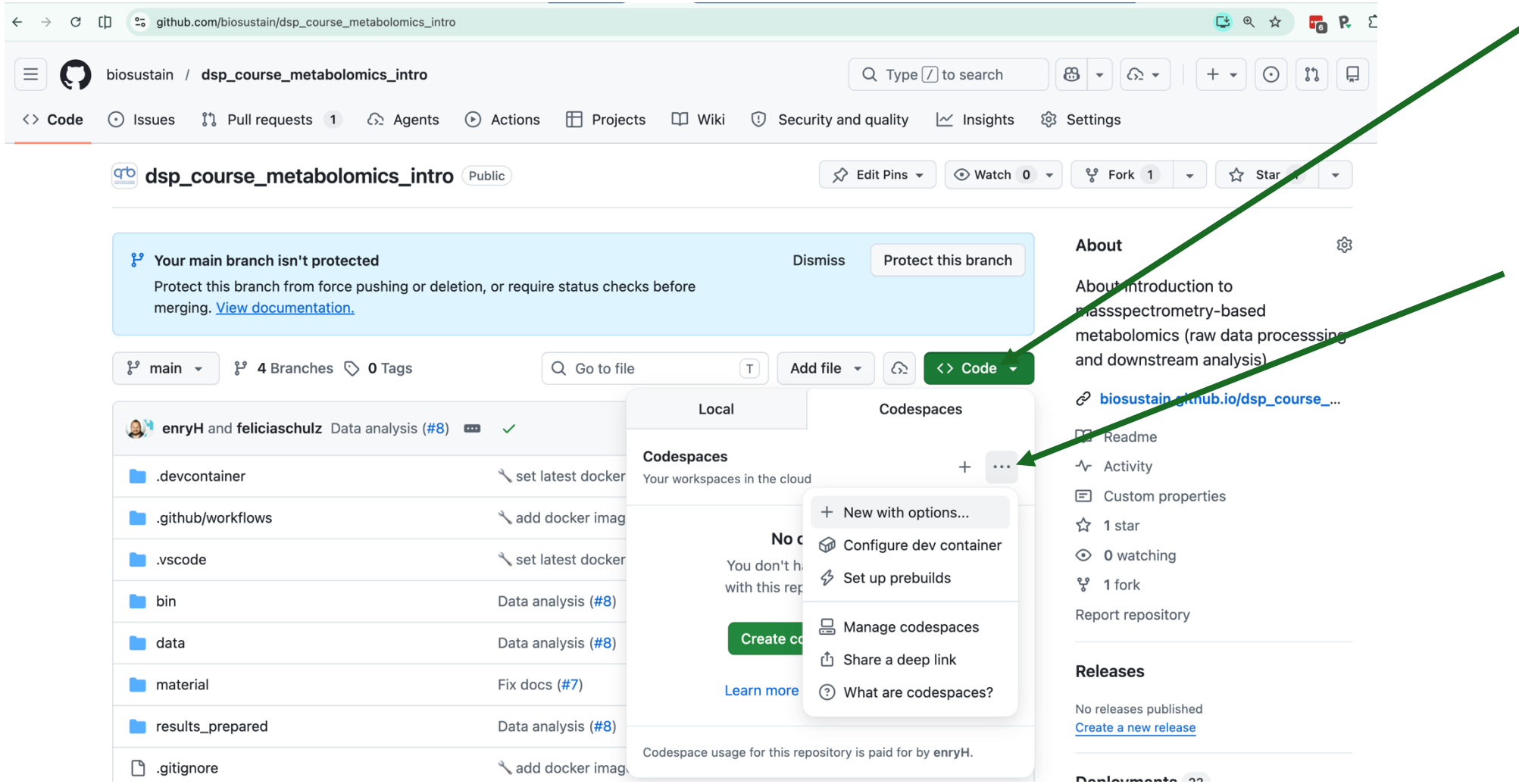
Break!

Pre-processing – Hands on

Click on: Open in GitHub Codespaces



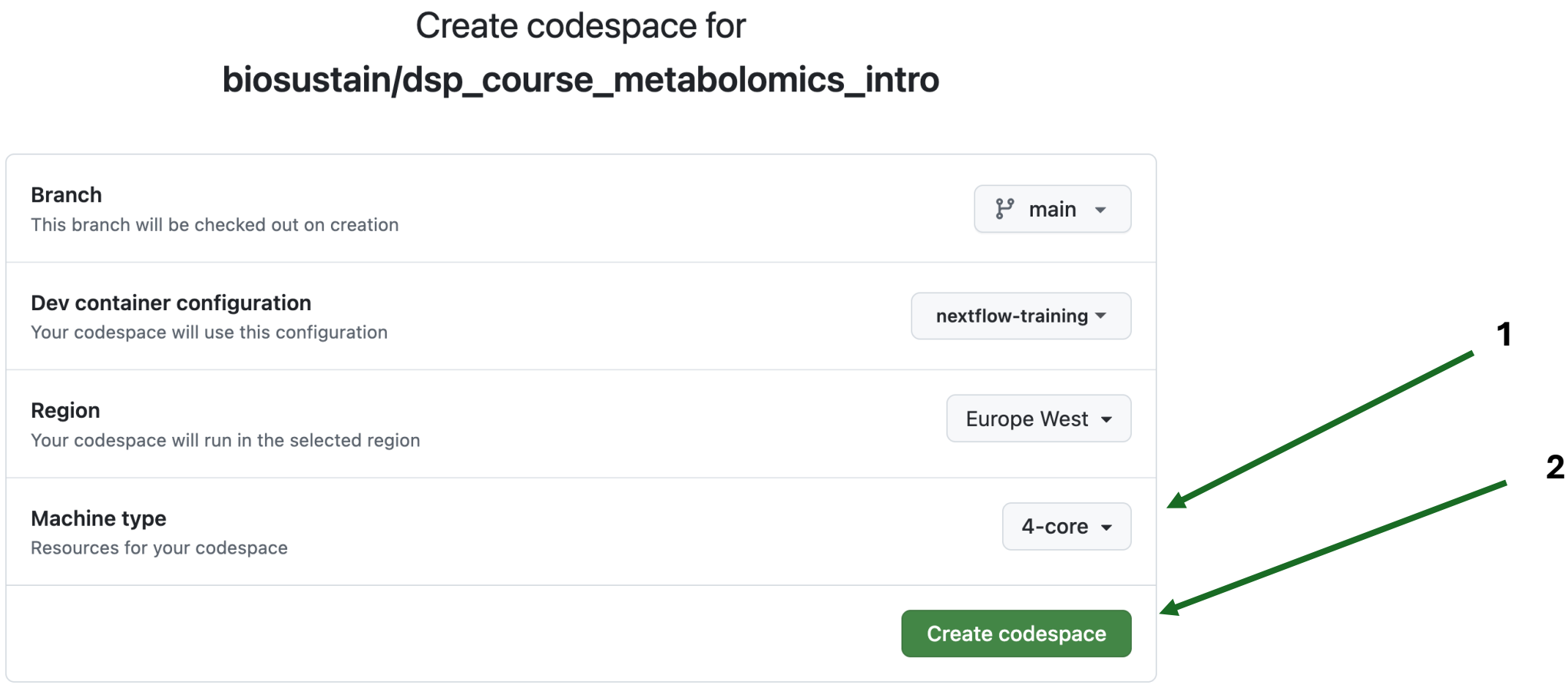
Alternative setup of codespace

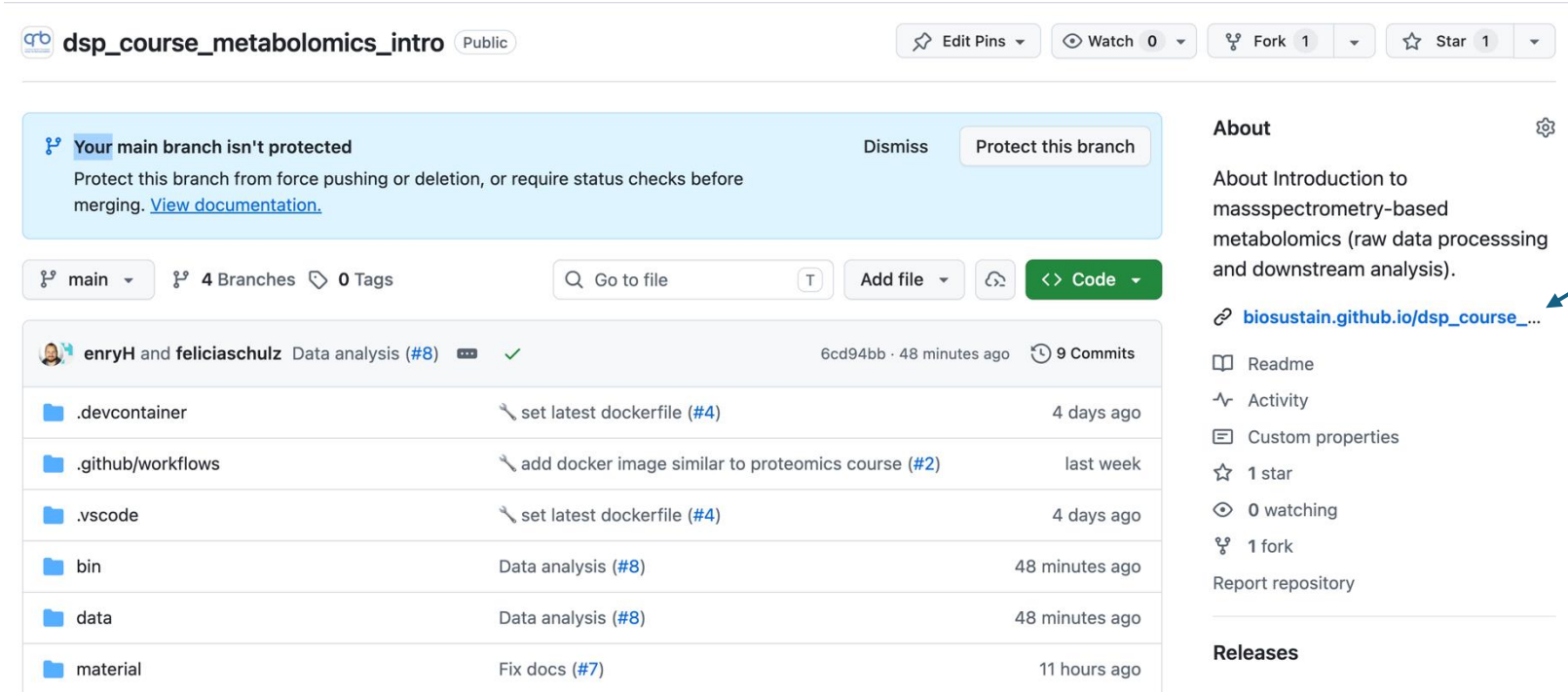


https://github.com/biosustain/dsp_course_proteomics_intro

Codespace with 4 cores (and 16GB of memory)

- **Not Safari**
- **Ignore Errors (todo tree)**

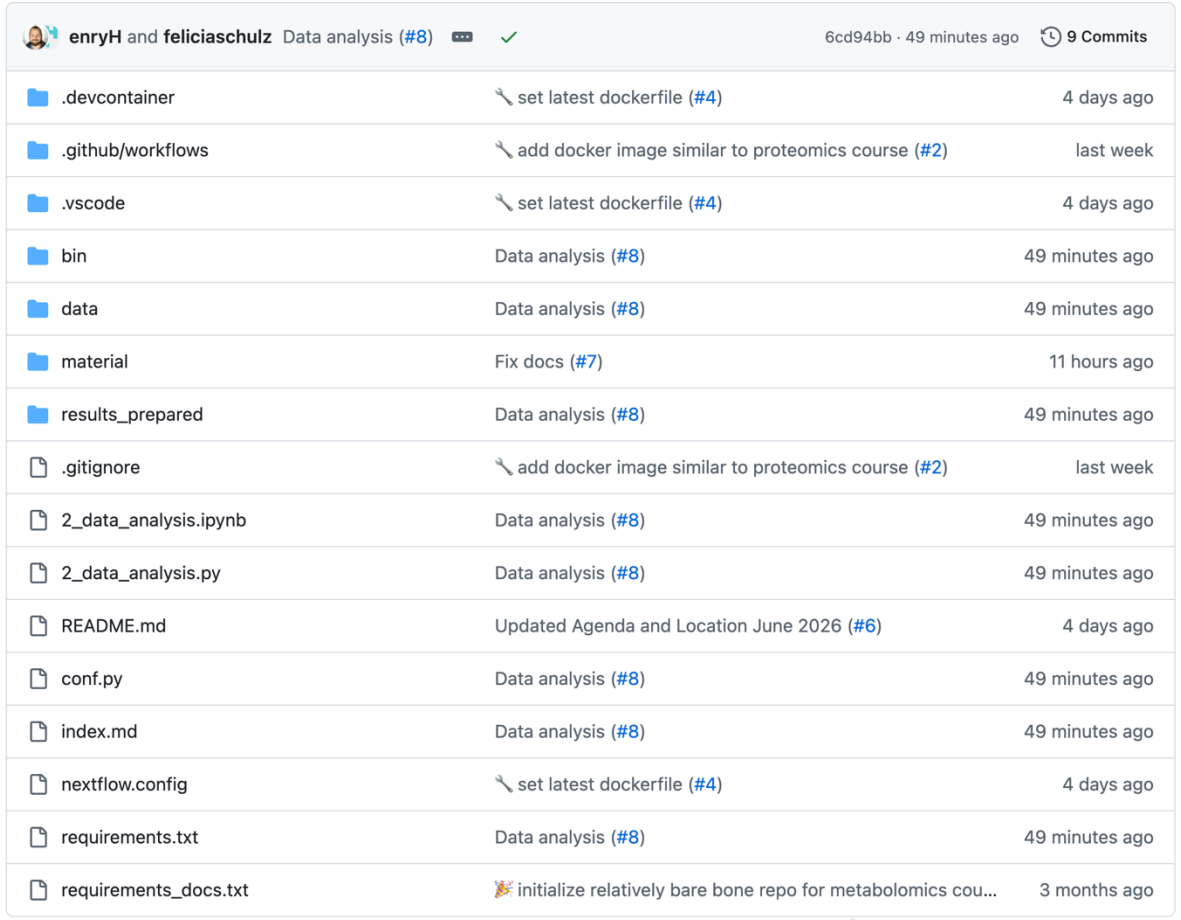




Go to website

Repo Overview

Feel free to ask questions at any point



Nextflow and nf-core pipeline template

nextflow

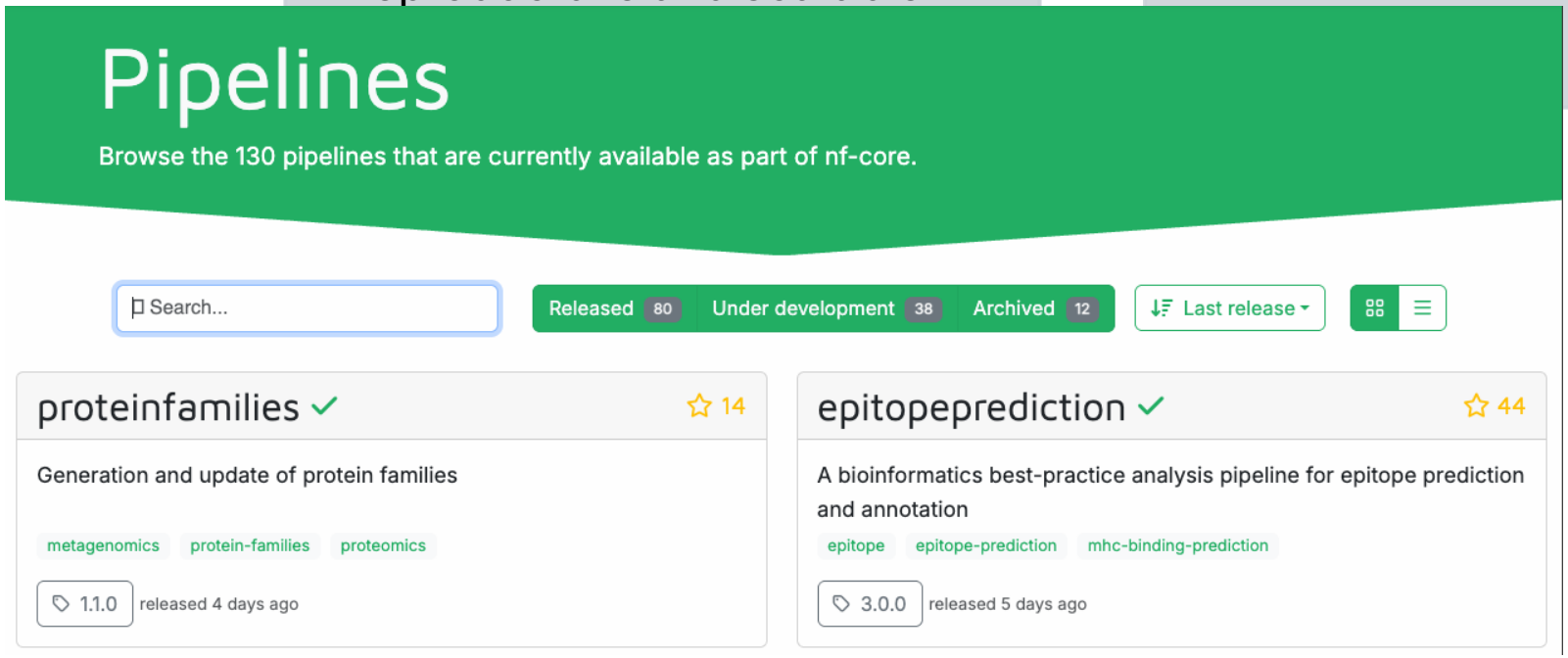
Nextflow is a workflow executor (analysis steps combined in an execution graph)

- Reproducibile and scalable

Nf-core is a collection of workflows maintained by nextflow and an open-source community

nf-core

A global community effort to collect a curated set of open-source analysis pipelines built using Nextflow.

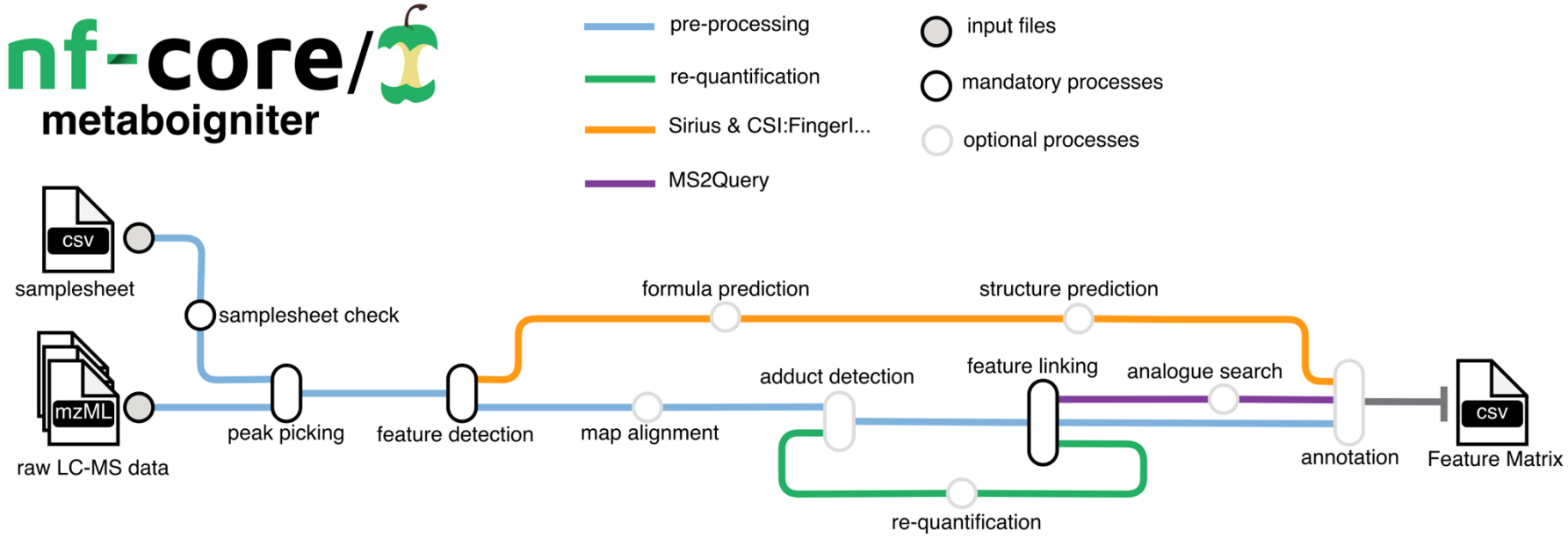


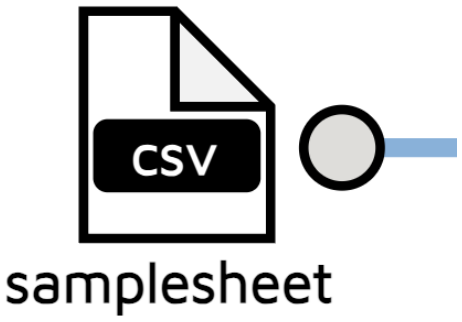
Works with

- Linux machines (on servers)
- on MacOS

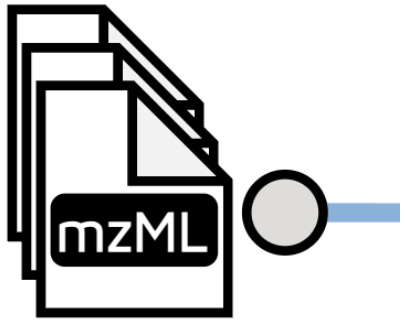
INPUTS

Metaboigniter – Metro Map



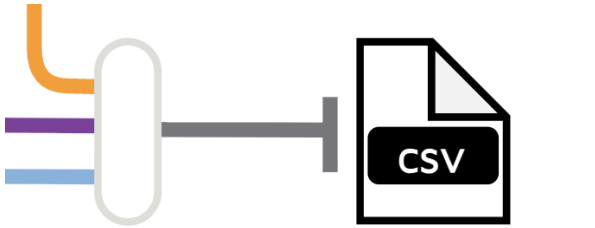


samplesheet



raw LC-MS data

(...)



annotation Feature Matrix

Samplesheet

sample	level	type	msfile
MS_A_POS	MS1	normal	data/MTBLS8735/MS_A_POS.mzML
MS_B_POS	MS1	normal	data/MTBLS8735/MS_B_POS.mzML
MS_C_POS	MS1	normal	data/MTBLS8735/MS_C_POS.mzML
MS_D_POS	MS1	normal	data/MTBLS8735/MS_D_POS.mzML
MS_E_POS	MS1	normal	data/MTBLS8735/MS_E_POS.mzML
MS_F_POS	MS1	normal	data/MTBLS8735/MS_F_POS.mzML
MS_QC_POOL_1_POS	MS1	normal	data/MTBLS8735/MS_QC_POOL_1_POS.mzML
MS_QC_POOL_2_POS	MS1	normal	data/MTBLS8735/MS_QC_POOL_2_POS.mzML
MS_QC_POOL_3_POS	MS1	normal	data/MTBLS8735/MS_QC_POOL_3_POS.mzML
MS_QC_POOL_4_POS	MS1	normal	data/MTBLS8735/MS_QC_POOL_4_POS.mzML
MSMS_2_A_CE20_POS	MS2	normal	data/MTBLS8735/MSMS_2_A_CE20_POS.mzML
MSMS_2_A_CE30_POS	MS2	normal	data/MTBLS8735/MSMS_2_A_CE30_POS.mzML
MSMS_2_A_CES_POS	MS2	normal	data/MTBLS8735/MSMS_2_A_CES_POS.mzML
MSMS_2_E_CE20_POS	MS2	normal	data/MTBLS8735/MSMS_2_E_CE20_POS.mzML
MSMS_2_E_CE30_POS	MS2	normal	data/MTBLS8735/MSMS_2_E_CE30_POS.mzML
MSMS_2_E_CES_POS	MS2	normal	data/MTBLS8735/MSMS_2_E_CES_POS.mzML

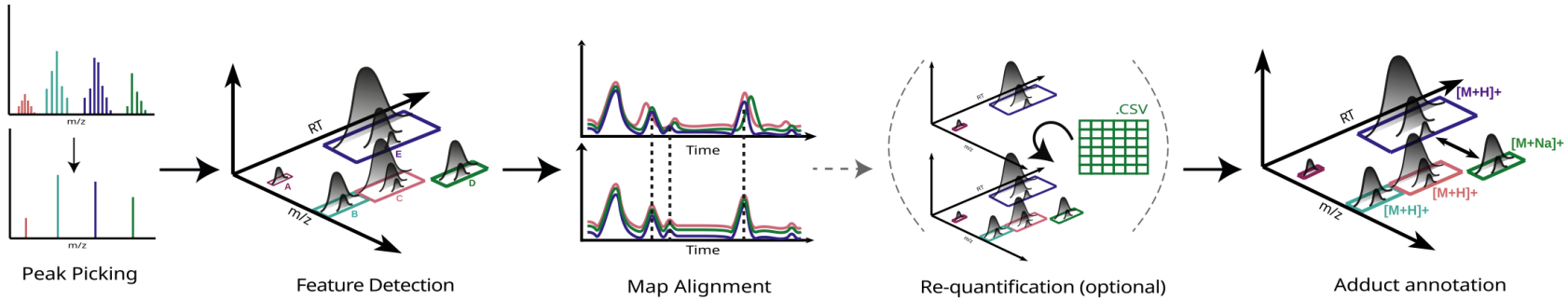
- Open standard, text based:
  - mzML files (indexed)
- Needs to be created from .d (bruker), .raw (Thermo), etc. vendor specific file formats
  - Conversion not (yet) integrated into workflow

One MS1 spectrum (DDA)

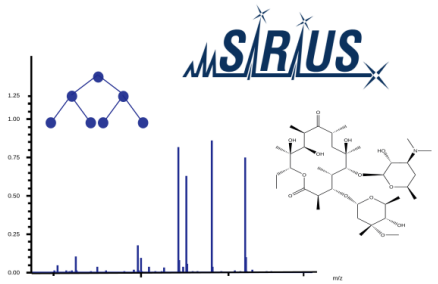

```xml
<spectrum id="controllerType=0 controllerNumber=1 scan=6" index="5" defaultArrayLength="736">
  <cvParam cvRef="MS" accession="MS:1000511" value="1" name="ms level" />
  <cvParam cvRef="MS" accession="MS:1000579" value="" name="MS1 spectrum" />
  <cvParam cvRef="MS" accession="MS:1000130" value="" name="positive scan" />
  <cvParam cvRef="MS" accession="MS:1000285" value="15248891" name="total ion current" />
  <cvParam cvRef="MS" accession="MS:1000127" value="" name="centroid spectrum" />
  <cvParam cvRef="MS" accession="MS:1000504" value="401.923461914063" name="base peak m/z" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
  <cvParam cvRef="MS" accession="MS:1000505" value="2704116.25" name="base peak intensity" unitAccession="MS:1000131" unitName="number of detector counts" unitCvRef="MS" />
  <cvParam cvRef="MS" accession="MS:1000528" value="375.884124755859" name="lowest observed m/z" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
  <cvParam cvRef="MS" accession="MS:1000527" value="1665.40612792969" name="highest observed m/z" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
  <scanList count="1">
    <cvParam cvRef="MS" accession="MS:1000795" value="" name="no combination" />
    <scan instrumentConfigurationRef="IC1">
      <cvParam cvRef="MS" accession="MS:1000016" value="6.0323342" name="scan start time" unitAccession="UO:0000031" unitName="minute" unitCvRef="UO" />
      <cvParam cvRef="MS" accession="MS:1000512" value="FTMS + p NSI Full ms [375.0000-1800.0000]" name="filter string" />
      <cvParam cvRef="MS" accession="MS:1000927" value="50" name="ion injection time" unitAccession="UO:0000028" unitName="millisecond" unitCvRef="UO" />
      <scanWindowList count="1">
        <scanWindow>
          <cvParam cvRef="MS" accession="MS:1000501" value="375" name="scan window lower limit" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
          <cvParam cvRef="MS" accession="MS:1000500" value="1800" name="scan window upper limit" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
        </scanWindow>
      </scanWindowList>
    </scan>
  </scanList>
  <binaryDataArrayList count="2">
    <binaryDataArray encodedLength="3152">
      <cvParam cvRef="MS" accession="MS:1000514" value="" name="m/z array" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />
      <cvParam cvRef="MS" accession="MS:1000523" value="" name="64-bit float" />
      <cvParam cvRef="MS" accession="MS:1000574" value="" name="zlib compression" />
      <binary>eJwt03lYVXUawPHjVumjGA1lG5dTGtMi3MYUG32496hJOiNX3EvlcjLNsEDNckGWIyoGxaIwiU3AsUWcZ2TJEnEm5Wimg48C466j15M12ZRgU1paCTPP+/399Xne9fceFk3TzIeylxna//0Vjck5ovM26h1oDLd
    </binaryDataArray>
    <binaryDataArray encodedLength="4356">
      <cvParam cvRef="MS" accession="MS:1000515" value="" name="intensity array" unitAccession="MS:1000131" unitName="number of counts" unitCvRef="MS" />
      <cvParam cvRef="MS" accession="MS:1000523" value="" name="64-bit float" />
      <cvParam cvRef="MS" accession="MS:1000574" value="" name="zlib compression" />
      <binary>eJwtWHlcT+kXviLFSI2MTJi5ZWIaTTGDsl+EkT3KVtypUVl+dhr5TG4qo7Jk8FPWm32pjC1kuxUlEu0JuSVDpozK0pdkPj3P/PV+3vu+73nPec45zznvFQRBuRdyQxIEQWzvf6ppFBz63sIYceZq06jafJHbNGr
    </binaryDataArray>
  </binaryDataArrayList>
</spectrum>
```

Metaboigniter - Overview

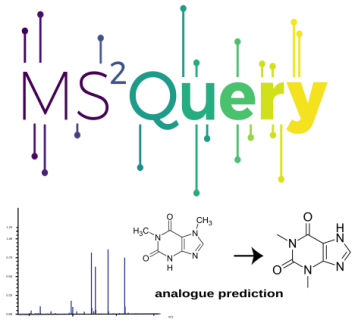
Pre-Processing



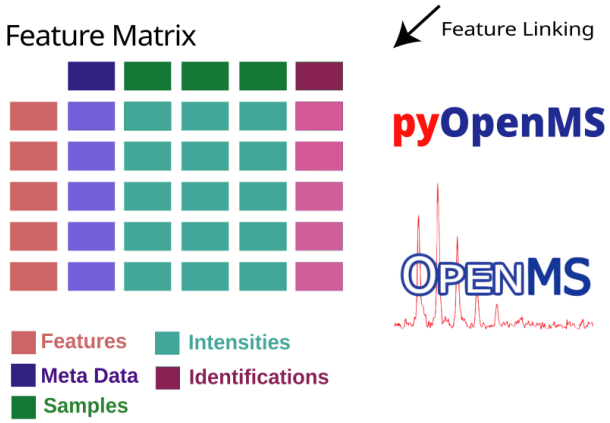
Identification (optional)



Predict metabolite formulas and structures with SIRIUS and CSI:FingerID.



Detect analogue molecules from your MS2 spectra with the machine learning tool MS2Query.



OpenMS and pyOpenMS

OpenMS 3.5.0 documentation

Search

ABOUT

Installation

Community

Learning

GETTING STARTED

Introduction

WebApps

Workflow Editor

TOPPView

TOPP Tools

pyOpenMS

MANUAL

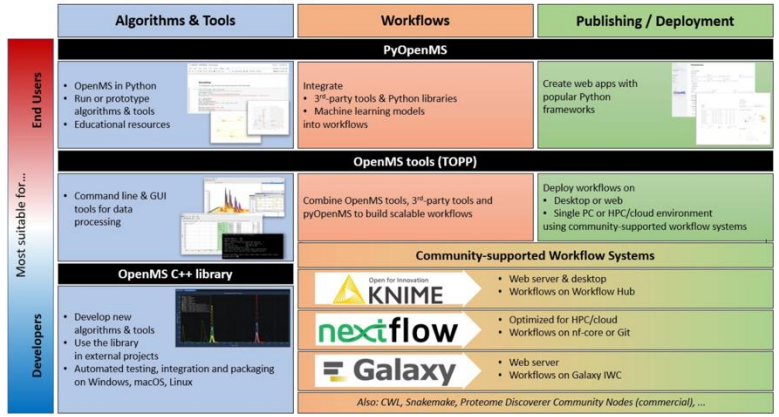
Contribute

Developers

Additional

1. What is OpenMS

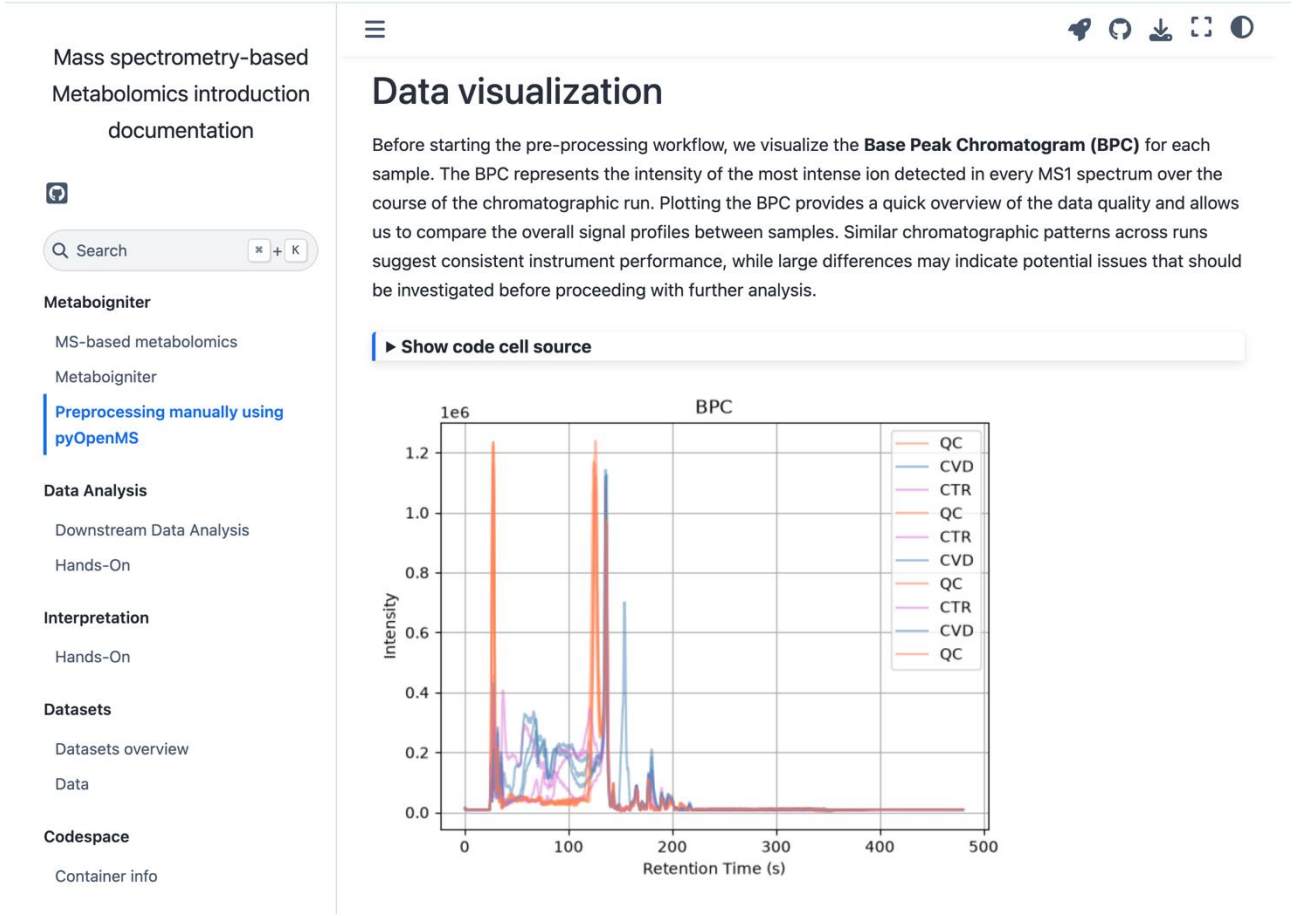
OpenMS is a free, open-source framework based on a C++ library with Python bindings. It is commonly used for liquid chromatography-mass spectrometry (LC-MS) data management and analyses. OpenMS provides an infrastructure for the rapid development of mass spectrometry related software as well as a rich toolset built on top of it. OpenMS is available under the three clause BSD licence and runs under Windows, macOS, and Linux operating systems.



OpenMS developers can create new C++ algorithms and tools, while users can execute tools or implement new algorithms or scripts in Python. Workflows integrate pyOpenMS scripts and OpenMS tools with third-party tools and external Python libraries to create scalable data-processing pipelines. For deployment, users can use pyOpenMS with web frameworks or deploy workflows on desktop, high-performance computing (HPC) or cloud infrastructure using one of the community-supported workflow systems.

OpenMS supports the Proteomics Standard Initiative (PSI) formats for MS data. The main contributors of OpenMS are currently the Eberhard-Karls-Universität in Tübingen, the Freie Universität Berlin, and the University of Toronto.

pyOpenMS functions highlighted



Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Input

Peak	m/z	rt	I_1	...	I_k
P_1	m/z_1	rt_1	I_{11}	...	I_{1k}
...	...	...	...	...	...
P_n	m/z_n	rt_n	I_{n1}	...	I_{nk}

Missingness due to errors or real absence of the molecule

Data analysis - Filtering and normalization

Filtering using the 80% rule

Peak	m/z	rt	missingness (%)	I_1	...	I_k
P_1	m/z_1	rt_1	5%	I_{11}	...	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	...	I_{2k}
P_3	m/z_3	rt_3	22%	I_{31}	...	I_{3k}
...	...	...	...	...	...	...
P_n	m/z_n	rt_n	42%	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Filtering using the 80% rules

Peak	m/z	rt	missingness (%)	I_1	...	I_k
P_1	m/z_1	rt_1	5%	I_{11}	...	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	...	I_{2k}
P_3	m/z_3	rt_3	22%	I_{31}	...	I_{3k}
...	...	...	...	...	...	...
P_n	m/z_n	rt_n	42%	I_{n1}	...	I_{nk}

Data analysis - Filtering and normalization

Filtering using the 80% rules

Peak	m/z	rt	missingness (%)	I_1	...	I_k
P_1	m/z_1	rt_1	5%	I_{11}	...	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	...	I_{2k}
P_3	m/z_3	rt_3	22%	I_{31}	...	I_{3k}
...	...	...	...	...	...	...
P_n	m/z_n	rt_n	42%	I_{n1}	...	I_{nk}

Positive acquisition mode: 2297 peaks

Negative acquisition mode: 2063 peaks

Data analysis - Filtering and normalization

Filtering using the modified 80% rule

Peak	m/z	rt	D_1	D_2	D_k	D:missingness (%)	C_1	C_2	C_k	C:missingness (%)	…
P_1	m/z_1	rt_1	D_{11}	D_{12}	D_{1k}	5%	C_{11}	C_{12}	C_{1k}	7%	…
P_2	m/z_2	rt_2	D_{21}	D_{22}	D_{2k}	2%	C_{21}	C_{22}	C_{2k}	83%	…
P_3	m/z_3	rt_3	D_{31}	D_{32}	D_{3k}	22%	C_{31}	C_{32}	C_{3k}	18%	…
…	…	…	…	…	…	…	…	…	…	…	…
P_n	m/z_n	rt_n	D_{41}	D_{42}	D_{4k}	42%	C_{41}	C_{42}	C_{4k}	21%	…

Data analysis - Filtering and normalization

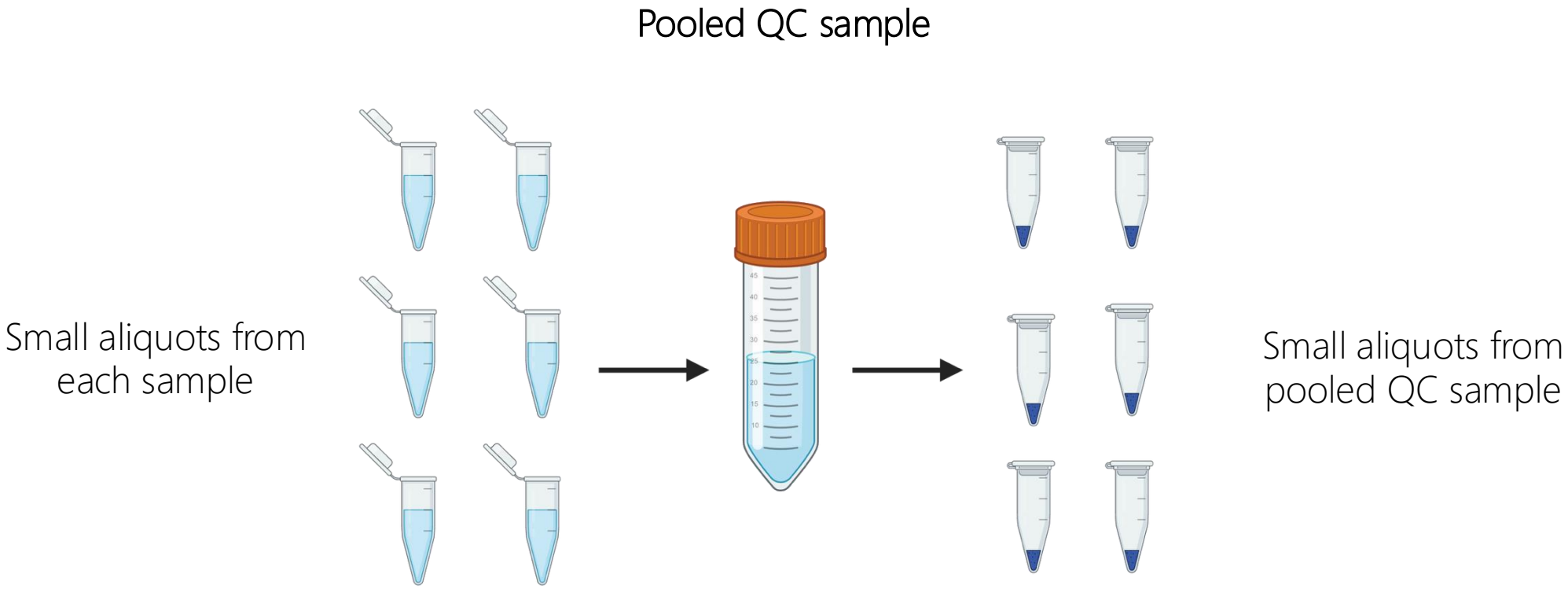
Filtering using the modified 80% rule

Peak	m/z	rt	D_1	D_2	D_k	D:missingness (%)	C_1	C_2	C_k	C:missingness (%)	…
P_1	m/z_1	rt_1	D_{11}	D_{12}	D_{1k}	5%	C_{11}	C_{12}	C_{1k}	7%	…
P_2	m/z_2	rt_2	D_{21}	D_{22}	D_{2k}	2%	C_{21}	C_{22}	C_{2k}	83%	…
P_3	m/z_3	rt_3	D_{31}	D_{32}	D_{3k}	22%	C_{31}	C_{32}	C_{3k}	18%	…
…	…	…	…	…	…	…	…	…	…	…	…
~~P_n~~	~~m/z_n~~	~~rt_n~~	~~D_{41}~~	~~D_{42}~~	~~D_{4k}~~	~~42%~~	~~C_{41}~~	~~C_{42}~~	~~C_{4k}~~	~~21%~~	…

"Perfect biomarker" gets retained

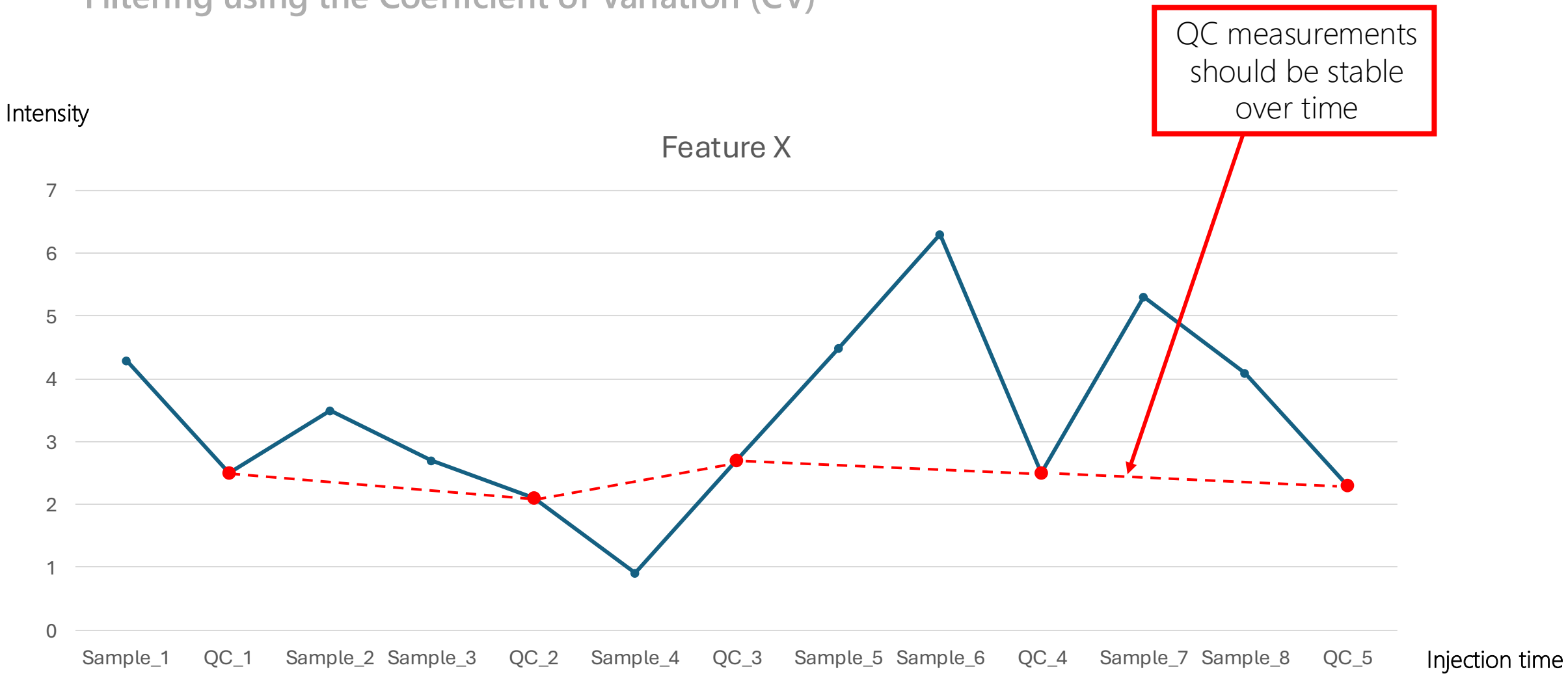
Data analysis - Filtering and normalization

Filtering using the Coefficient of Variation (CV)



Data analysis - Filtering and normalization

Filtering using the Coefficient of Variation (CV)



Data analysis - Filtering and normalization

Filtering using the Coefficient of Variation (CV)

$$CV\ (\%) = \left(\frac{Standard\ deviation}{Mean}\right) \times 100$$

- If the QC CV > S CV: remove feature
- If the QC CV > threshold %: remove feature

Peak	m/z	rt	S_1	S_2	S_k	S:CV	QC_1	QC_2	QC_k	QC:CV	…
P_1	m/z_1	rt_1	S_{11}	S_{12}	S_{1k}	8%	QC_{11}	QC_{12}	QC_{1k}	7%	…
P_2	m/z_2	rt_2	S_{21}	S_{22}	S_{2k}	10%	QC_{21}	QC_{22}	QC_{2k}	5%	…
~~P_3~~	~~m/z_3~~	~~rt_3~~	~~S_{31}~~	~~S_{32}~~	~~S_{3k}~~	~~5%~~	~~QC_{31}~~	~~QC_{32}~~	~~QC_{3k}~~	~~20%~~	…
…	…	…	…	…	…	…	…	…	…	…	…
~~P_n~~	~~m/z_n~~	~~rt_n~~	~~S_{41}~~	~~S_{42}~~	~~S_{4k}~~	~~42%~~	~~QC_{41}~~	~~QC_{42}~~	~~QC_{4k}~~	~~35%~~	…

Data analysis - Filtering and normalization

Filtering using the Blanks control

Blank extracts

- Prepared with same materials, chemicals, consumables as study samples
- But: no biological sample

Why? Check for…

- … compounds from other sources than tested biological material
- … carryover in instrument

Data analysis - Filtering and normalization

Filtering using the Blanks control

If the ratio of total blanks intensities / total samples intensities for a feature exceeds a certain threshold, remove

Blank extracts

- Prepared with same materials, chemicals, consumables as study samples
- But: no biological sample

Peak	m/z	rt	S_1	S_2	S_k	S:Total	B	B_2	B_k	B:Total	B_Total/ S_Total
P_1	m/z_1	rt_1	S_{11}	S_{12}	S_{1k}	…	B_{11}	B_{12}	B_{1k}	…	0.2
~~P_2~~	~~m/z_2~~	~~rt_2~~	~~S_{21}~~	~~S_{22}~~	~~S_{2k}~~	~~…~~	~~B_{21}~~	~~B_{22}~~	~~B_{2k}~~	~~…~~	~~1~~
P_3	m/z_3	rt_3	S_{31}	S_{32}	S_{3k}	…	B_{31}	B_{32}	B_{3k}	…	0.001
…	…	…	…	…	…	…	…	…	…	…	…
~~P_n~~	~~m/z_n~~	~~rt_n~~	~~S_{41}~~	~~S_{42}~~	~~S_{4k}~~	~~…~~	~~B_{41}~~	~~B_{42}~~	~~B_{4k}~~	~~…~~	~~2~~

Data analysis - Imputation

Imputation with zeros

After filtering: still some missing values ⟶ need imputation

Peak	m/z	rt	missingness (%)	I_1	I_2	I_3	...	I_k
P_1	m/z_1	rt_1	5%	I_{11}	N/A	I_{13}	...	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	I_{22}	N/A	...	I_{2k}
P_3	m/z_3	rt_3	3%	N/A	I_{32}	I_{33}	...	I_{3k}
...	...	...	...	...	...	...	...	...
P_n	m/z_n	rt_n	2%	I_{n1}	I_{n2}	I_{n3}	...	I_{nk}

Data analysis - Imputation

Imputation with zeros

After filtering: still some missing values ⟶ need imputation

Common imputation strategy: **Imputation with zeros**

Peak	m/z	rt	missingness (%)	I_1	I_2	I_3	...	I_k
P_1	m/z_1	rt_1	5%	I_{11}	0	I_{13}	...	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	I_{22}	0	...	I_{2k}
P_3	m/z_3	rt_3	3%	0	I_{32}	I_{33}	...	I_{3k}
...	...	...	...	...	...	...	...	...
P_n	m/z_n	rt_n	2%	I_{n1}	I_{n2}	I_{n3}	...	I_{nk}

Data analysis - Imputation

Imputation with zeros

Alternative strategy: Half-minimum imputation

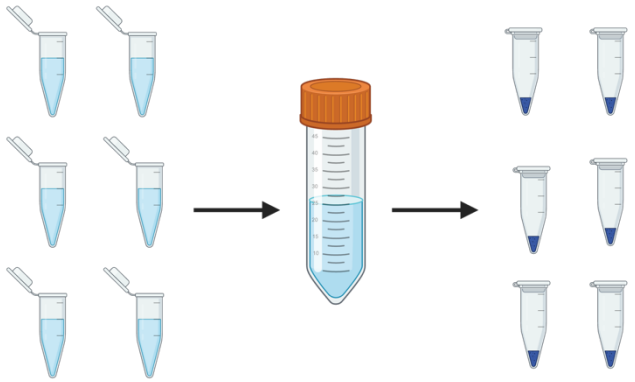
Assumption: Feature data is **Missing Not at Random** (MNAR) = metabolite could be just below the limit of detection ⟶ limit of detection is …?

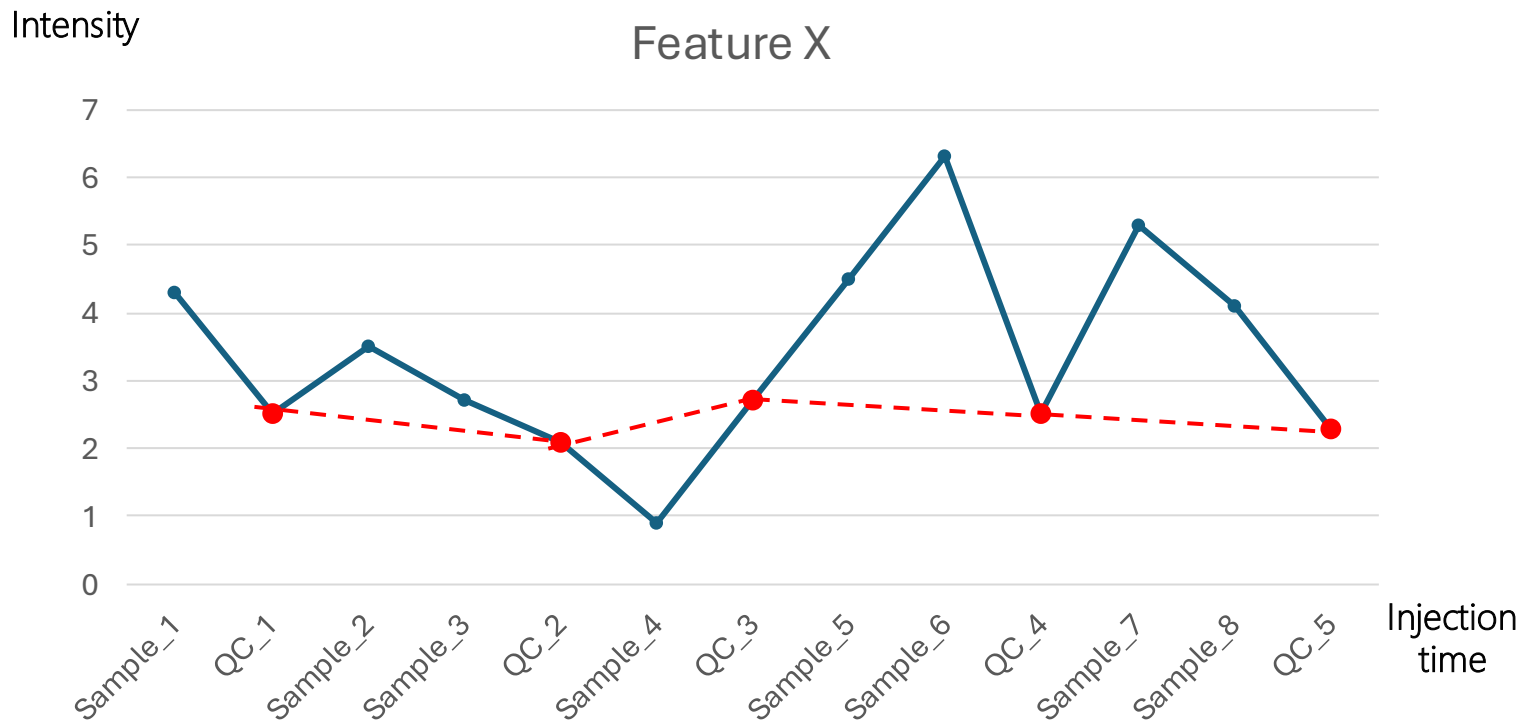
For each feature, find the minimum intensity value recorded, and divide it in half

Peak	m/z	rt	missingness (%)	I_1	I_2	I_3	…	I_k
P_1	m/z_1	rt_1	5%	I_{11}	$I_{22}/2$	I_{13}	…	I_{1k}
P_2	m/z_2	rt_2	8%	I_{21}	I_{22}	$I_{73}/2$	…	I_{2k}
P_3	m/z_3	rt_3	3%	$I_{51}/2$	I_{32}	I_{33}	…	I_{3k}
…	…	…	…	…	…	…	…	…
P_n	m/z_n	rt_n	2%	I_{n1}	I_{n2}	I_{n3}	…	I_{nk}

Data analysis - Filtering and normalization

Drift correction

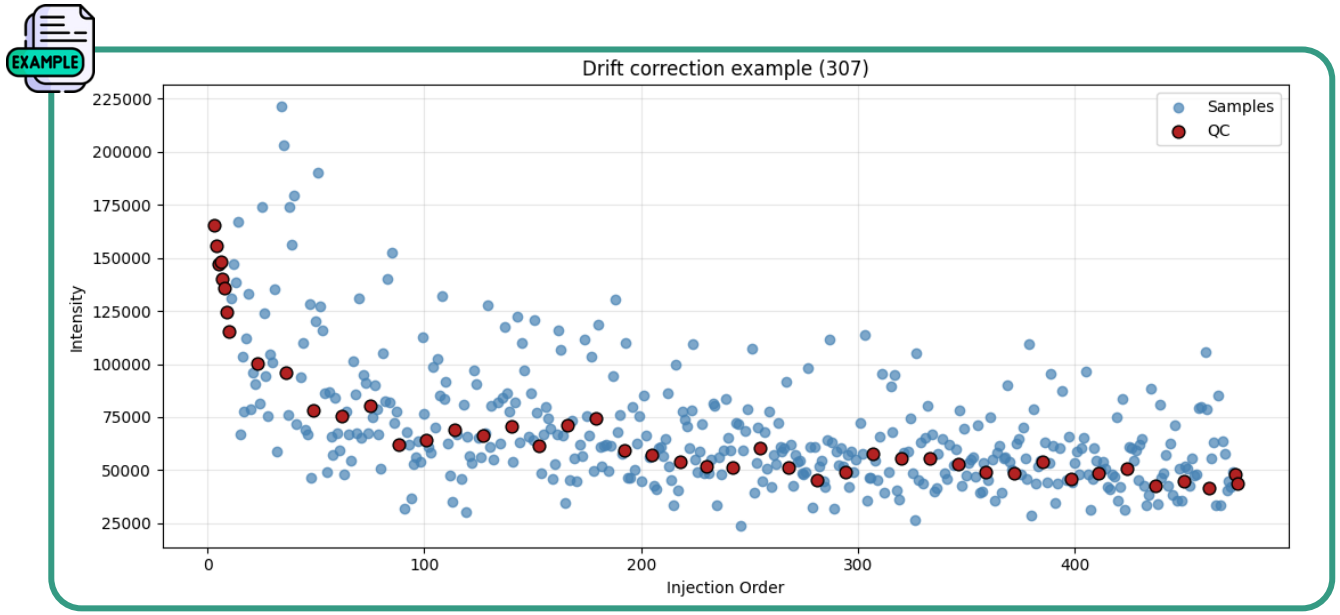




Data analysis - Filtering and normalization

Drift correction: LOESS

- We want to use QC samples to estimate artificial variation that comes from instrumental drift over time
- We want to subtract that variation from our data points

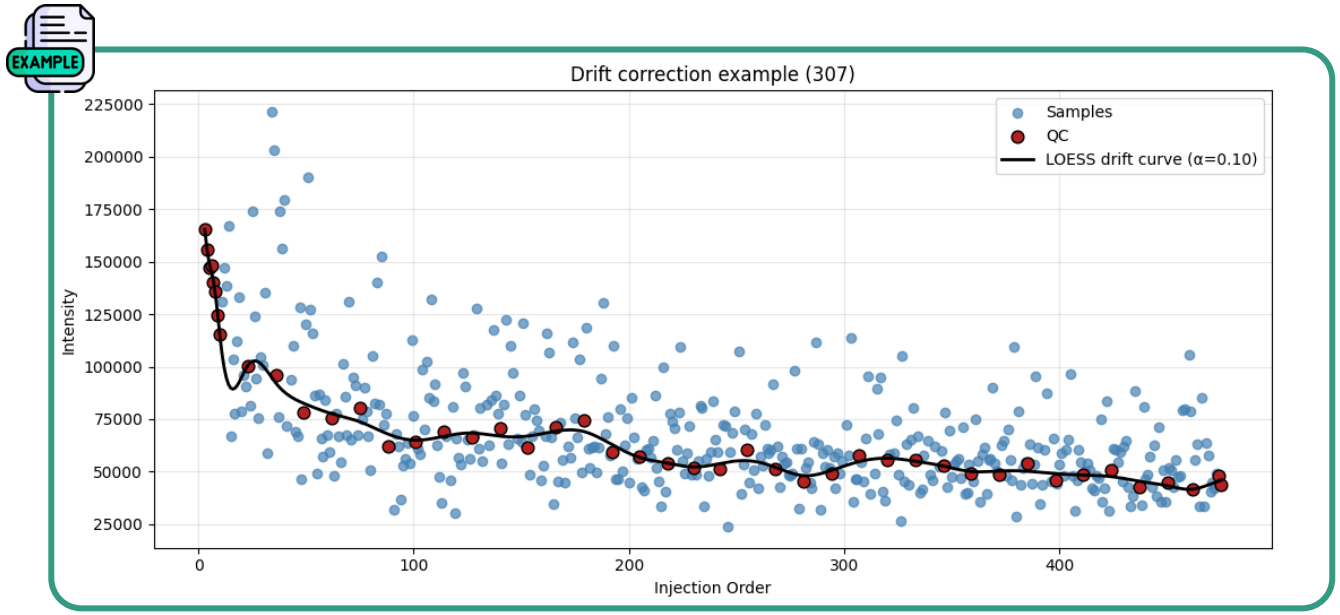


Data analysis - Filtering and normalization

Drift correction: LOESS

LOESS = Locally Estimated Scatterplot Smoothing

- Interpolate a value for QC samples at the run order value of every biological sample

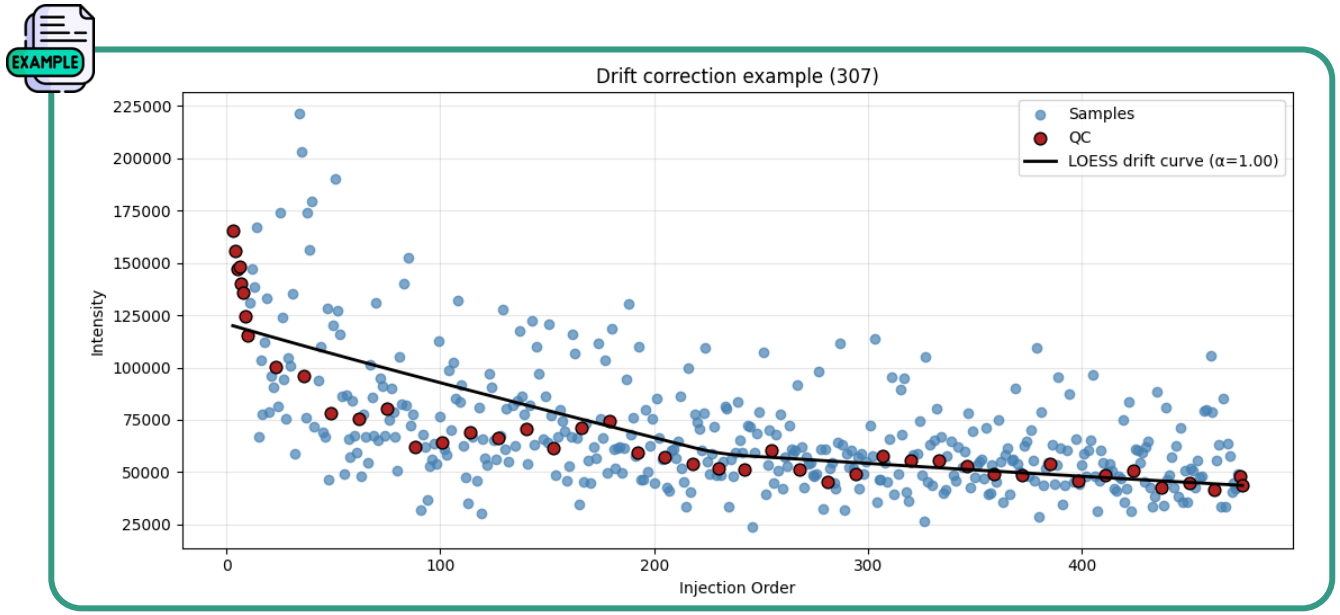


Data analysis - Filtering and normalization

Drift correction: LOESS

LOESS = Locally Estimated Scatterplot Smoothing

- Interpolate a value for QC samples at the run order value of every biological sample
- To reduce influence of extreme values, use LOESS Smoothing before interpolation

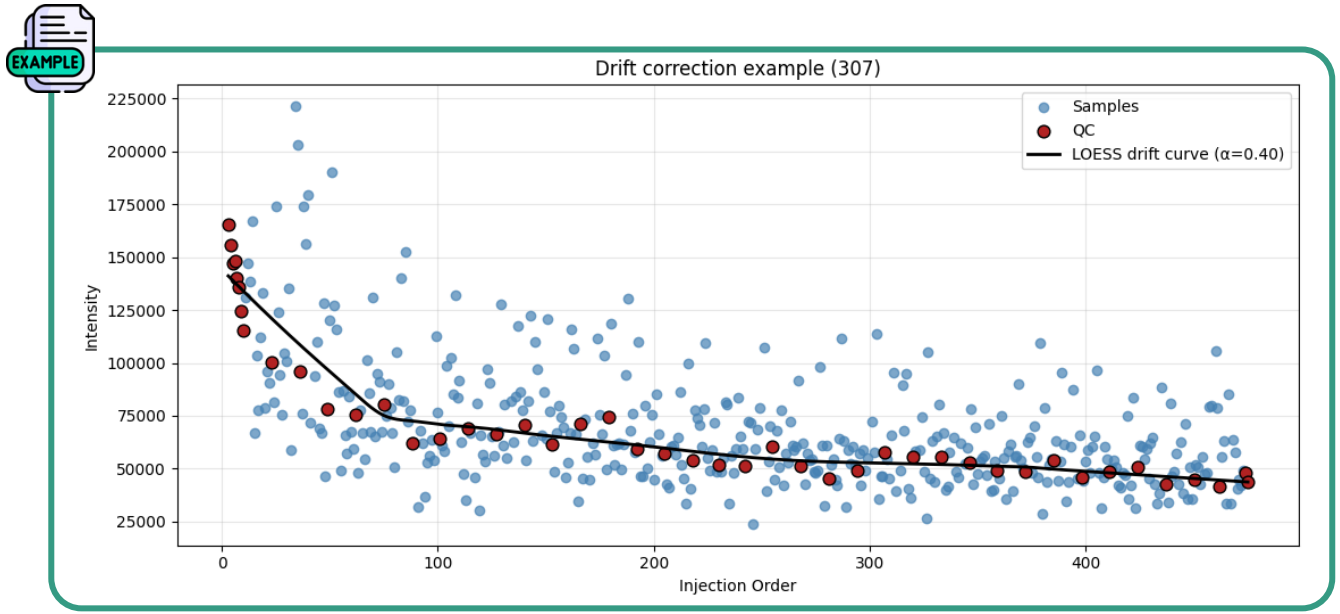


Data analysis - Filtering and normalization

Drift correction: LOESS

LOESS = Locally Estimated Scatterplot Smoothing

- Interpolate a value for QC samples at the run order value of every biological sample
- To reduce influence of extreme values, use LOESS Smoothing before interpolation
- Adjust level of smoothing with parameter alpha

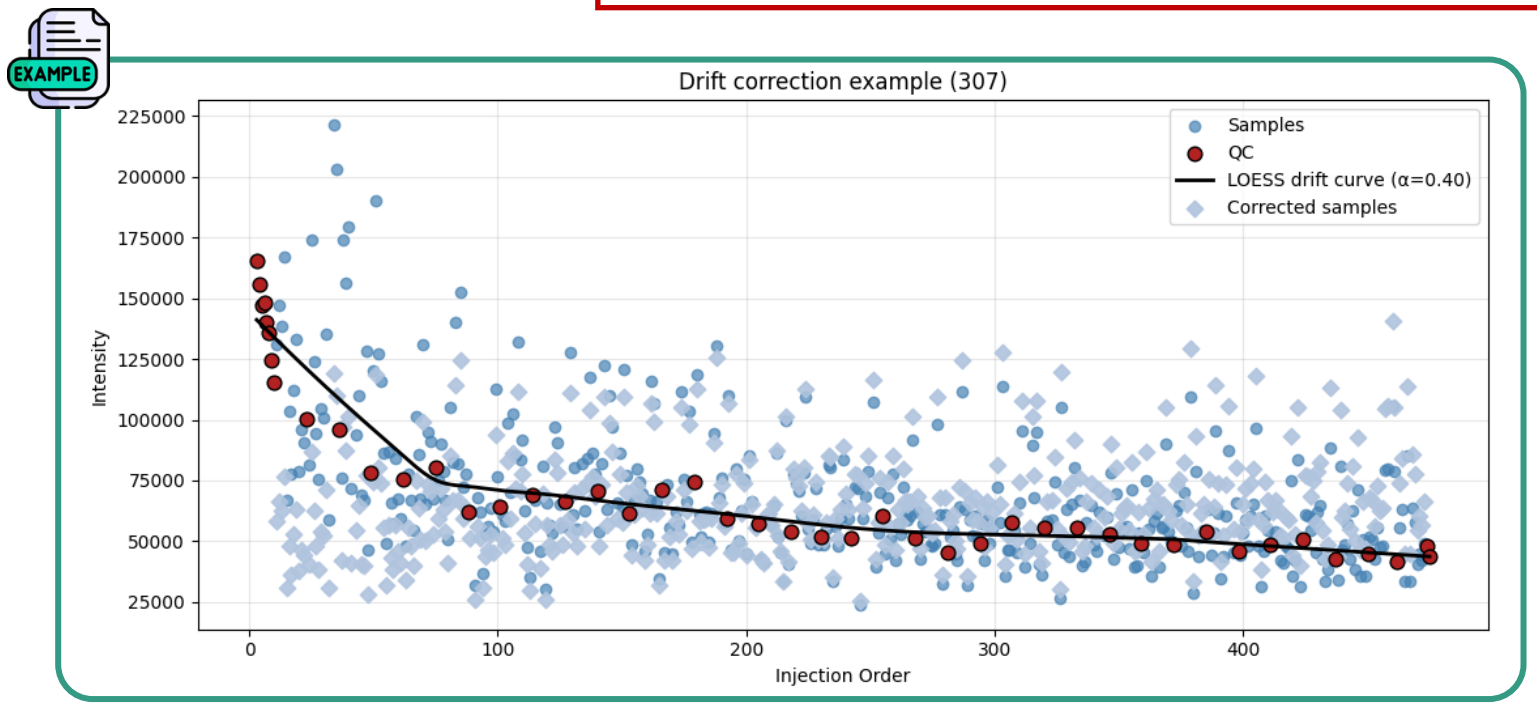


Data analysis - Filtering and normalization

Drift correction: LOESS

Now that we have an artificial QC value for each injection time point, can correct our biological sample values

$$\text{new values} = \frac{\text{original values}}{\text{drift curve}} \times \text{median(QCs)}$$



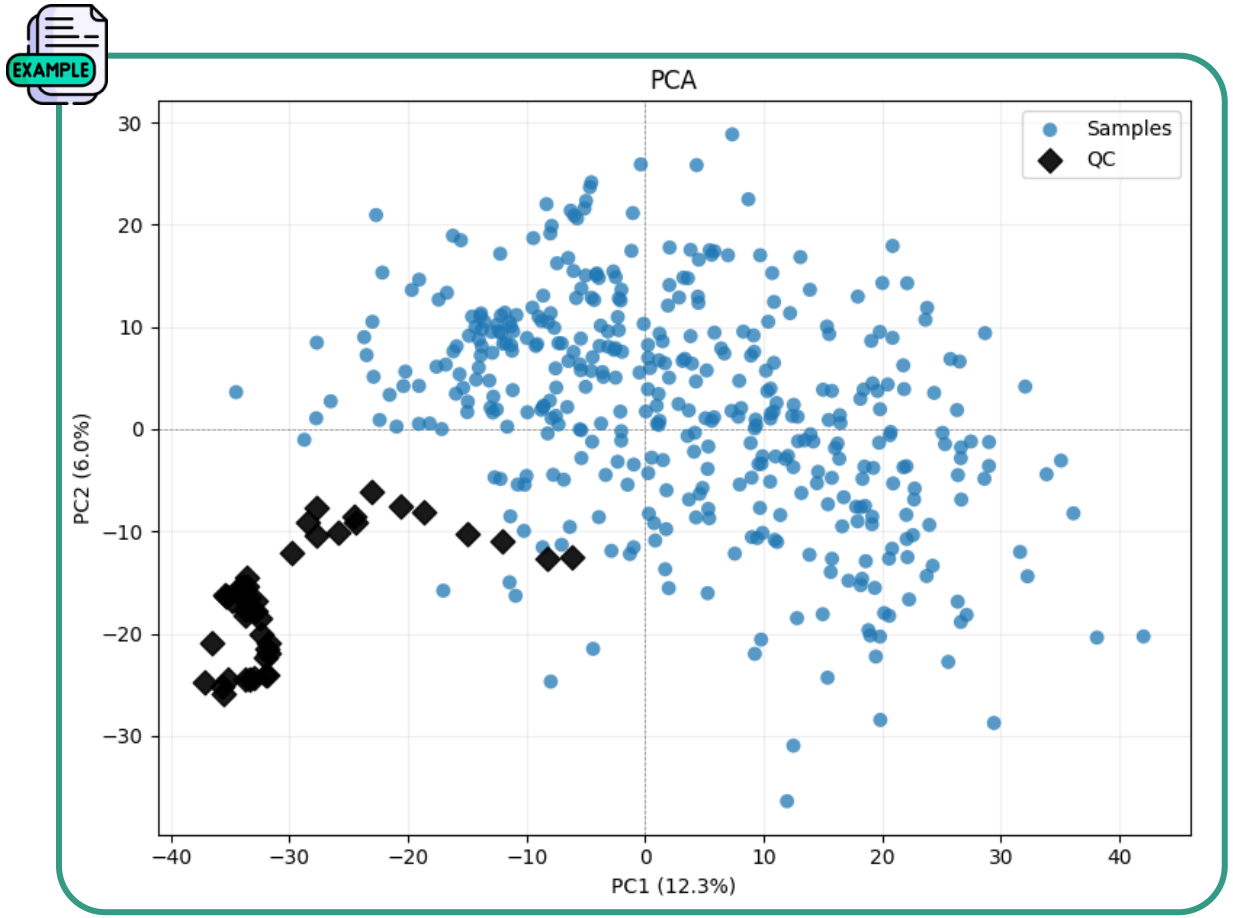
- Assumes drift is multiplicative: scales signal up or down
- Divide by drift curve to remove that factor
- Have relative intensity: how strong signal was relative to instrument at that moment
- Scale back to this specific feature's intensities with QC median

Data analysis - Filtering and normalization

Drift correction: CPCA

Common Principal Components Analysis (CPCA)

- Finds variation that is common across all groups
- Assumption: technical variation (drift) is shared across all groups, whereas biological variation is unique
- Want to use CPCA to identify and then remove this variation

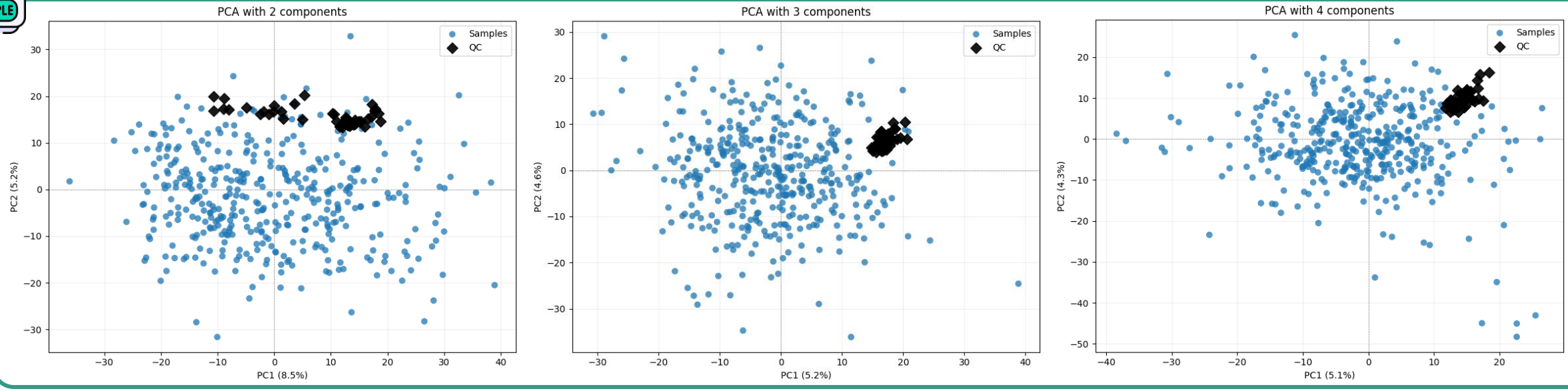


Data analysis - Filtering and normalization

Drift correction: CPCA

- Can try removing different numbers of components
- Choose the one where the cluster of QCs after removal of principal component is the tightest
- Can calculate the distance of points to cluster centroid to assess objectively

EXAMPLE

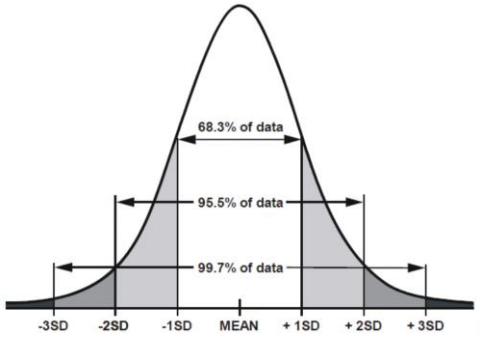


Data analysis - Filtering and normalization

Normalization

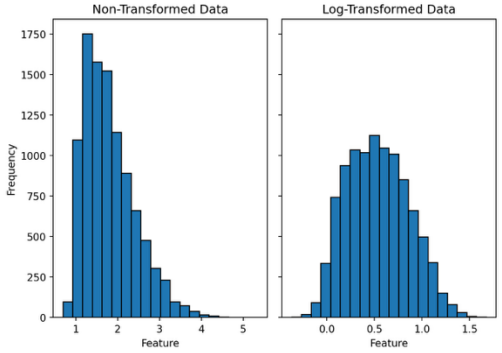
Z-score normalization (scaling)

$$z = \frac{x - \mu}{\sigma}$$



- Metabolite intensities can have vastly different magnitudes
- Gives every feature equal weight

Log transformation



- Metabolomics data is right-skewed (lower bounded at 0, but no upper bound)
- Stabilizes variance
- Makes fold changes more interpretable

Internal standards

$$\text{Normalized intensity} = \frac{\text{Analyte intensity}}{\text{Internal standard intensity}}$$

- Known amount
- Usually not used in untargeted metabolomics, there are too many metabolites

Data analysis - Statistical Analysis

Univariate analyses – one metabolite at a time

Comparing two groups

- T-test on log-transformed data (parametric)
- Mann-Whitney U test, if normality can't be assumed (non-parametric)

Comparing more than two groups

- ANOVA (ANalysis Of VAriance) (parametric)
- Kruskal-Wallis, if normality can't be assumed (non-parametric)

Continuous outcomes

- Linear regression
- Correlation

Adjusting for covariates

- ANCOVA (ANalysis of COVAriance)
- Multiple linear regression

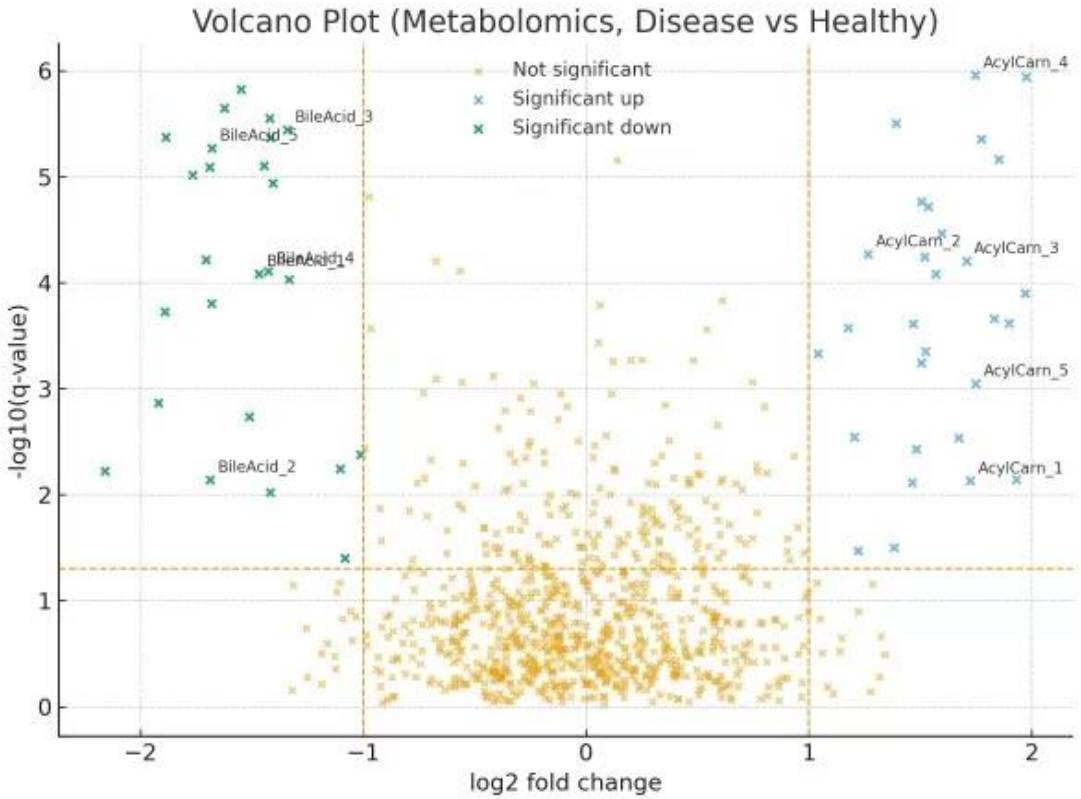
Data analysis - Statistical Analysis

Univariate analyses – one metabolite at a time

Fold changes = difference in the average metabolite level between sample groups

Important: multiple testing correction

- Huge numbers of features, so p-values need to be corrected
- Common: Benjamini-Hochberg (BH) False Discovery Rate
- Rather consider adjusted p-values (or q-values)



Data analysis - Statistical Analysis

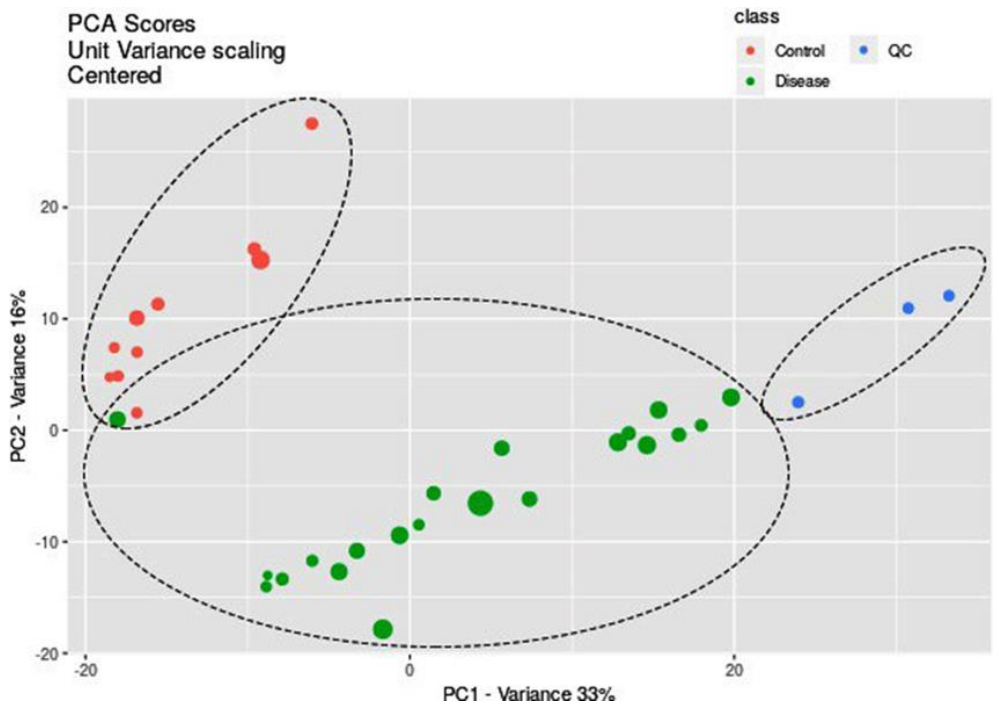
Multivariate analysis – all metabolites together

Principal Component Analysis (PCA)

- Reduces high dimensionality of metabolomics data into principal components (PCs)
- PCs capture largest sources of variance
- Can use this to see if controls cluster well

Partial Least-Squares Discriminant Analysis (PLS-DA)

- Discovers which metabolites drive the most separation between groups
- Latent variables are rotated specifically to discriminate between classes



Data analysis – Example Pipeline

1. Filtering
   - Modified 80%-rule
   - Blanks filtering
   - CV-filtering
2. Imputation
   - Half-minimum imputation
3. Drift correction
   - LOESS smoothing-based correction & CPCA
   - Assess which one works better
4. Normalization
   - Log transforming data
5. Statistical analysis
   - ANOVA
   - Volcano plot

& using PCAs throughout to check data

Break!

Data analysis – Hands on

Metabolomics Interpretation

Peak annotation vs peak identification

Identification levels according to the Chemical Analysis Working Group of the Metabolomics Standards Initiative:

Level 1 – Identified compounds require comparison of at least two independent properties (e.g., retention time and MS/MS spectra) with those of an authentic standard analysed under identical conditions.

Metabolomics Interpretation

Peak annotation vs peak identification

Identification levels according to the Chemical Analysis Working Group of the Metabolomics Standards Initiative:

Level 1 – Identified compounds require comparison of at least two independent properties (e.g., retention time and MS/MS spectra) with those of an authentic standard analysed under identical conditions.

Level 2 – Putatively annotated compounds based on similarity to MS/MS data in commercial or public databases (reference standards are not available).

Metabolomics Interpretation

Peak annotation vs peak identification

Identification levels according to the Chemical Analysis Working Group of the **Metabolomics Standards Initiative**:

Level 1 – Identified compounds require comparison of at least two independent properties (e.g., retention time and MS/MS spectra) with those of an authentic standard analysed under identical conditions.

Level 2 – Putatively annotated compounds based on similarity to MS/MS data in commercial or public databases (reference standards are not available).

Level 3 – Putatively characterized compound class.

Level 4 – Unknown feature.

Metabolomics Interpretation

Peak annotation vs peak identification

Identification levels according to the Chemical Analysis Working Group of the **Metabolomics Standards Initiative**:

Level 1 – Identified compounds require comparison of at least two independent properties (e.g., retention time and MS/MS spectra) with those of an authentic standard analysed under identical conditions.

Level 2 – Putatively annotated compounds based on similarity to MS/MS data in commercial or public databases (reference standards are not available).

Level 3 – Putatively characterized compound class.

Level 4 – Unknown feature.

LC-MS[1]

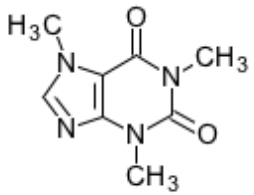
Metabolomics Interpretation

Peak annotation using LC-MS[1]

Peak annotation is a major bottleneck and direct matching of masses is not enough:

1

A single metabolite produces multiple peaks due to adducts and fragments.



Caffeine
Exact mass = 194.0804

LC-MS +

$[M+H]^+$

$[M+Na]^+$

$[M+K]^+$

$[M+NH_4]^+$

Metabolomics Interpretation

Peak annotation using LC-MS[1]

Peak annotation is a major bottleneck and direct matching of masses is not enough:

2 DB matching results in hundreds of potential candidates.

m/z = 195.0877 → hmdb (The Human Metabolome Database)

Positive mode
10 different adducts

635 Results:
Caffeine $[M+H]^+$ - mass = 194.0804
Enprofylline $[M+H]^+$ - mass = 194.0804
Luminol $[M+NH_4]$ – mass = 177.0538
…

Metabolomics Interpretation

Peak annotation using LC-MS[1]

Peak annotation is a major bottleneck and direct matching of masses is not enough:

1. A single metabolite produces multiple peaks due to adducts and fragments.
2. DB matching results in hundreds of potential candidates.
3. The metabolome is not completely annotated.

⇩

Computational algorithms that use m/z, rt and intensity to return putative annotations.

Metabolomics Interpretation

Peak annotation using LC-MS[1]

Peak annotation is a major bottleneck and direct matching of masses is not enough:

1. A single metabolite produces multiple peaks due to adducts and fragments.
2. DB matching results in hundreds of potential candidates.
3. The metabolome is not completely annotated.

Computational algorithms  Low confidence level putative annotations.

Metabolomics Interpretation

Peak annotation vs peak identification

Identification levels according to the Chemical Analysis Working Group of the **Metabolomics Standards Initiative**:

Level 1 – Identified compounds require comparison of at least two independent properties (e.g., retention time and MS/MS spectra) with those of an authentic standard analysed under identical conditions.

Level 2 – Putatively annotated compounds based on similarity to MS/MS data in commercial or public databases (reference standards are not available).

$LC\text{-}MS^2$

Level 3 – Putatively characterized compound class.

Level 4 – Unknown feature.

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

1. Spectral library matching

2. In-silico fragmentation

3. Fragmentation tree methods

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

1. Spectral library matching
    - Match our query spectrum with known MS^2 spectra

    We only have annotated spectra for a fraction of the known metabolome

    We don't know the whole metabolome

2. In-silico fragmentation

3. Fragmentation tree methods

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

1. Spectral library matching
   - Match our query spectrum with known MS^2 spectra

     We only have annotated spectra for a fraction of the known metabolome

     We don't know the whole metabolome

2. In-silico fragmentation
   - Predicting MS^2 spectra from chemical structures to increase spectral libraries

     Still very unreliable, especially for large complex molecules and lipids

3. Fragmentation tree methods

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

1. Spectral library matching
   - Match our query spectrum with known MS^2 spectra

   We only have annotated spectra for a fraction of the known metabolome

   We don't know the whole metabolome

2. In-silico fragmentation
   - Predicting MS^2 spectra from chemical structures to increase spectral libraries

   Still very unreliable, especially for large complex molecules and lipids

3. Fragmentation tree methods
   - To predict molecular formulas, fingerprints or even novel structures.

   Excellent results for smaller metabolites (<400 Da), but unreliable for larger molecules

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

1. Spectral library matching
   - Match our query spectrum with known MS^2 spectra

We only have annotated spectra for a fraction of the known metabolome

We don't know the whole metabolome

2. In-silico fragmentation
   - Predicting MS^2 spectra from chemical structures to increase spectral libraries

Still very unreliable, especially for large complex molecules and lipids

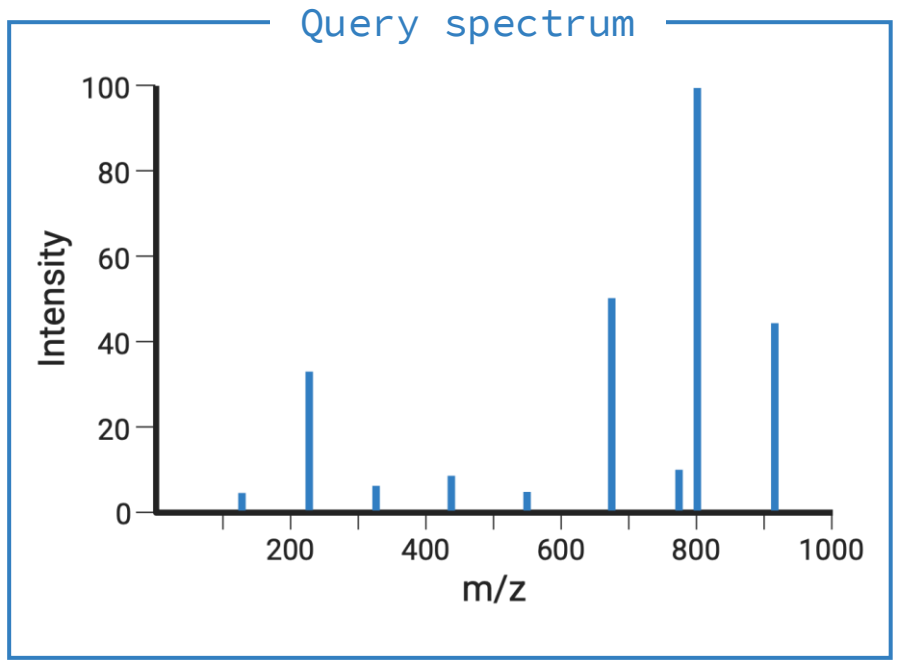
3. Fragmentation tree methods
   - To predict molecular formulas, fingerprints or even novel structures.

Excellent results for smaller metabolites (<400 Da), but unreliable for larger molecules

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

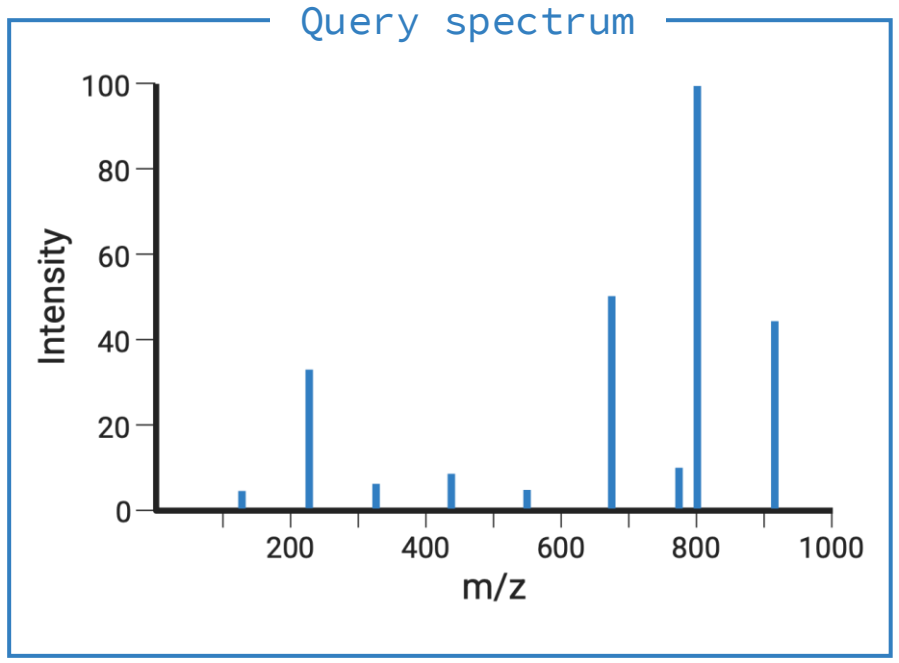
Spectral library matching

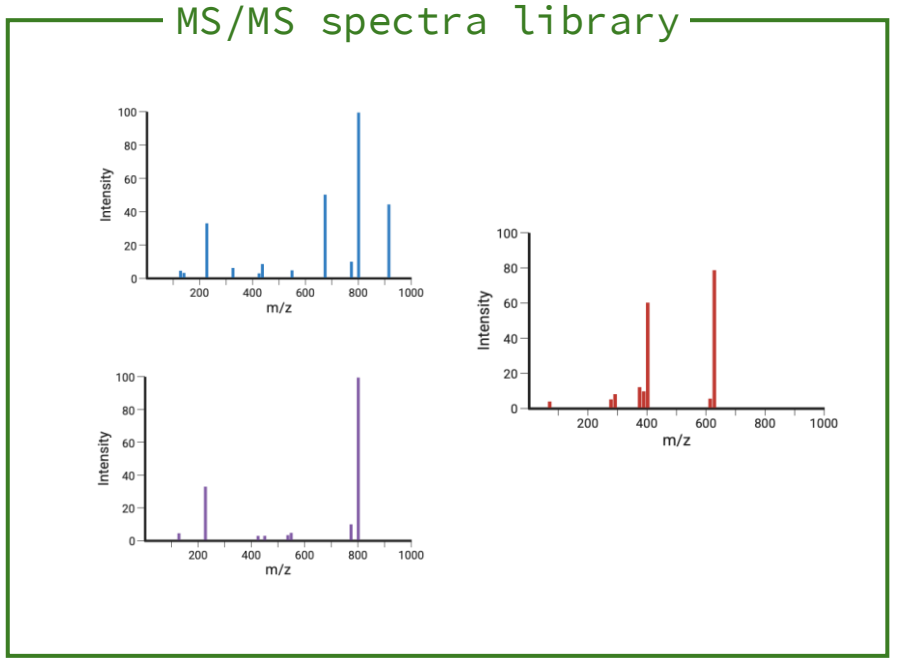


Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Spectral library matching

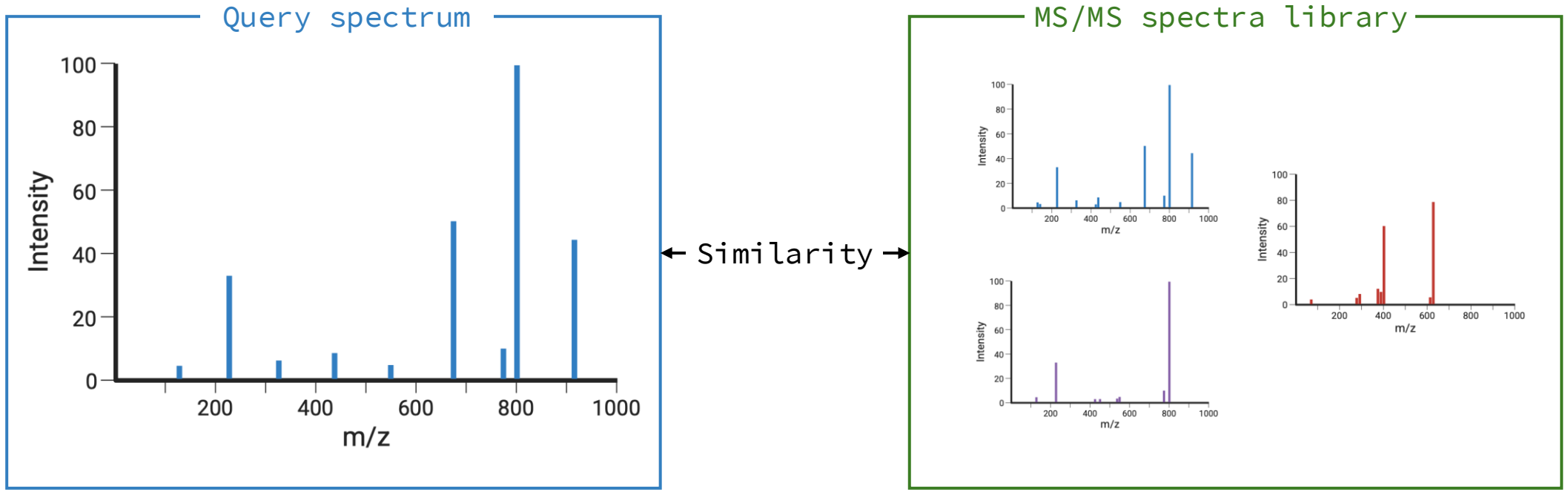




Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

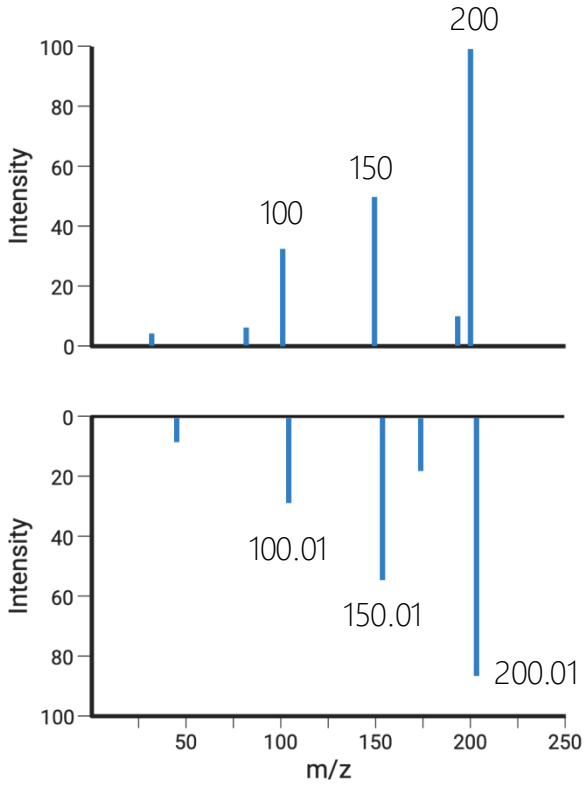
Spectral library matching



Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Cosine similarity

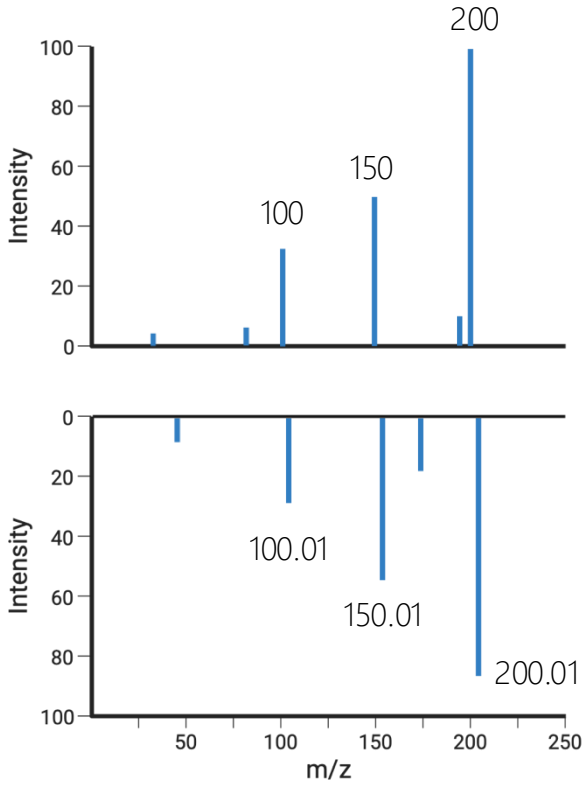


Query spectrum

Reference spectrum

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

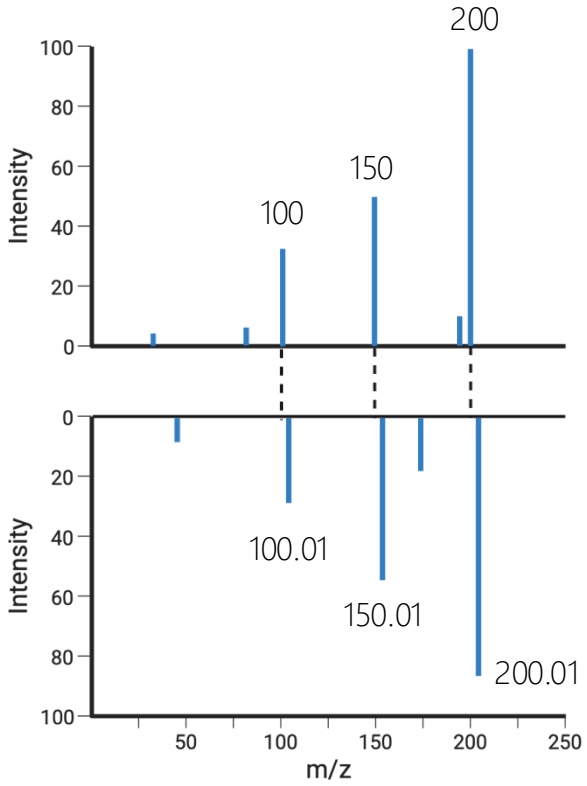


Cosine similarity

1. Find peaks that match: $|mz(p) - mz(p')| < t \ (\text{e.g.}, 0.05 \text{ Da})$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra



Cosine similarity

1. Find peaks that match: $|mz(p) - mz(p')| < t$ $(\mathrm{e.g.}, 0.05\ \mathrm{Da})$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Cosine similarity

1. Find peaks that match: $|mz(p) - mz(p')| < t \ (\mathrm{e.g.}, 0.05 \ \mathrm{Da})$
2. Intensities of matching peaks are converted to vectors

I = [38, 46, 101]

I' = [30, 59, 85]

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Cosine similarity

I = [38, 46, 101]

I′ = [30, 59, 85]

1. Find peaks that match: $|mz(p) - mz(p')| < t \ (\mathrm{e.g.}, 0.05 \ \mathrm{Da})$
2. Intensities of matching peaks are converted to vectors
3. Cosine similarity:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} = 0.98$$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

What if … Two molecules differ by a methyl group:

Molecule A precursor = 300 Da
Molecule B precursor = 314 Da

⚠️ Difference = 14 Da

Many fragments may shift +14 Da

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

What if … Two molecules differ by a methyl group:

Molecule A precursor = 300 Da

Molecule B precursor = 314 Da

⚠️ Difference = 14 Da

Many fragments may shift +14 Da

— Cosine similarity approach → No matching peaks

$\mathbf{Similarity = 0}$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

What if … Two molecules differ by a methyl group:

Molecule A precursor = 300 Da

Molecule B precursor = 314 Da

⚠️ Difference = 14 Da

Many fragments may shift +14 Da

— Cosine similarity approach → No matching peaks

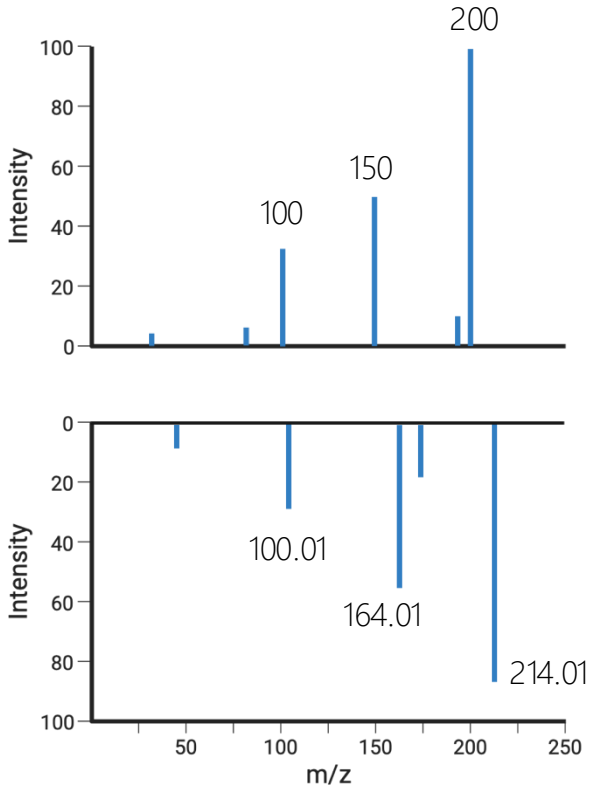
$\mathbf{Similarity = 0}$

But molecules are structurally very similar… Modified cosine similarity

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Modified cosine similarity

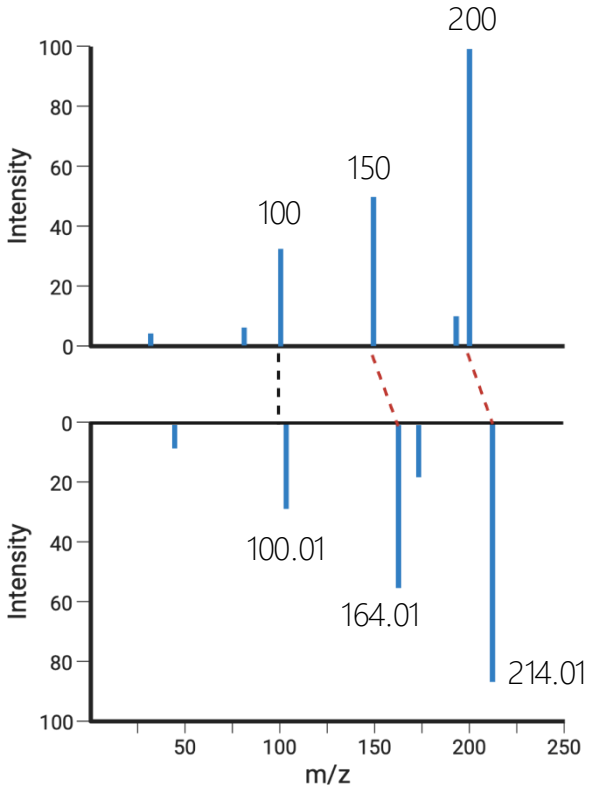


Query spectrum

Reference spectrum

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra



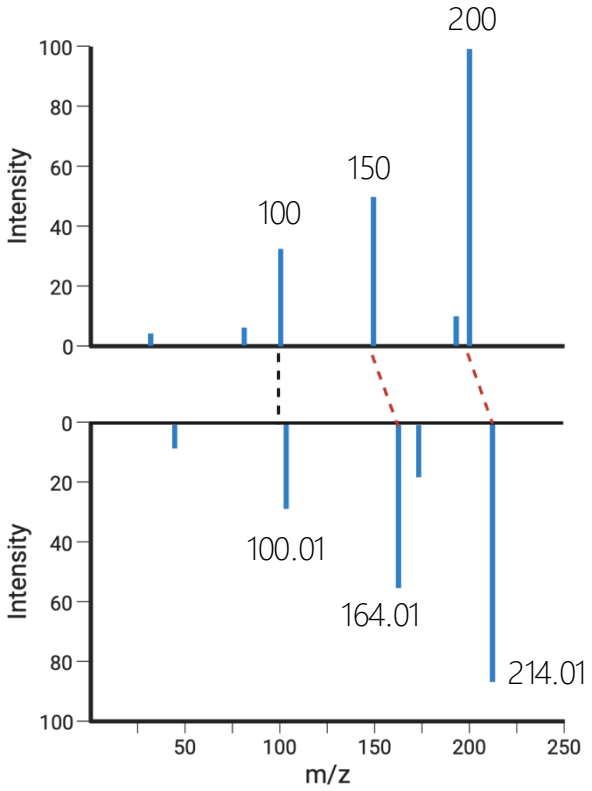
Modified cosine similarity

1. Find peaks that match: $|mz(p) + M - mz(p')| < t$

   Precursor mass difference $M = PM(S') - PM(S) = 14\ \mathrm{Da}$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra



Modified cosine similarity

1. Find peaks that match: $|mz(p) + M - mz(p')| < t$

   Precursor mass difference $M = PM(S') - PM(S) = 14\ \text{Da}$

2. Intensities of matching peaks are converted to vectors

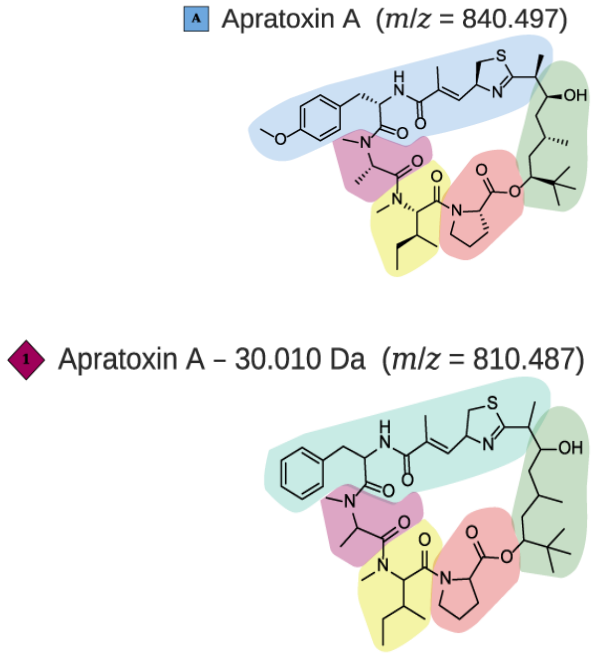
3. Cosine similarity:

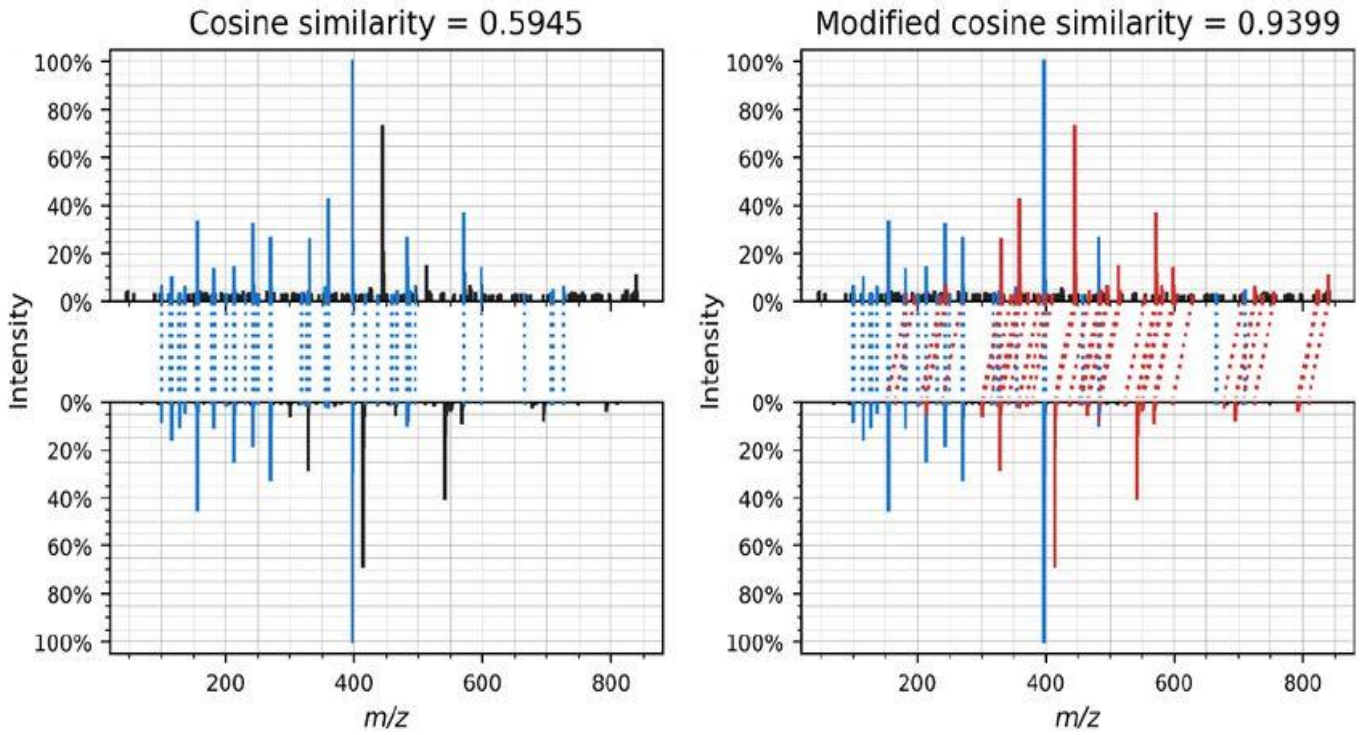
$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Spectral library matching





Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

Spectral library matching

However…

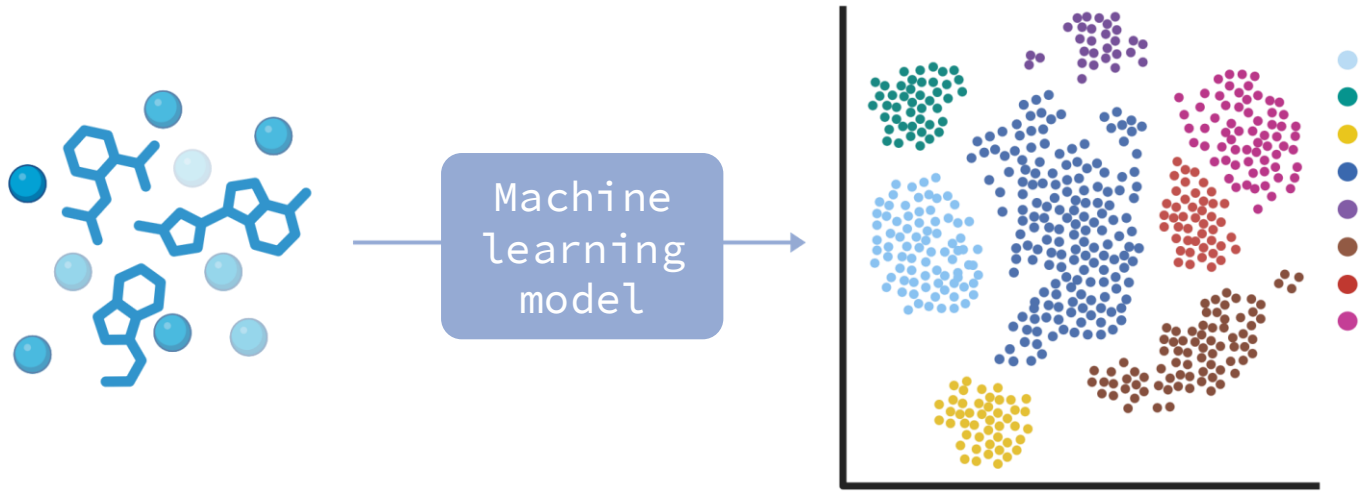
Small chemical modifications ⟶ Low modified cosine scores

Ideally, we want an approach that understands chemical structure!

Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

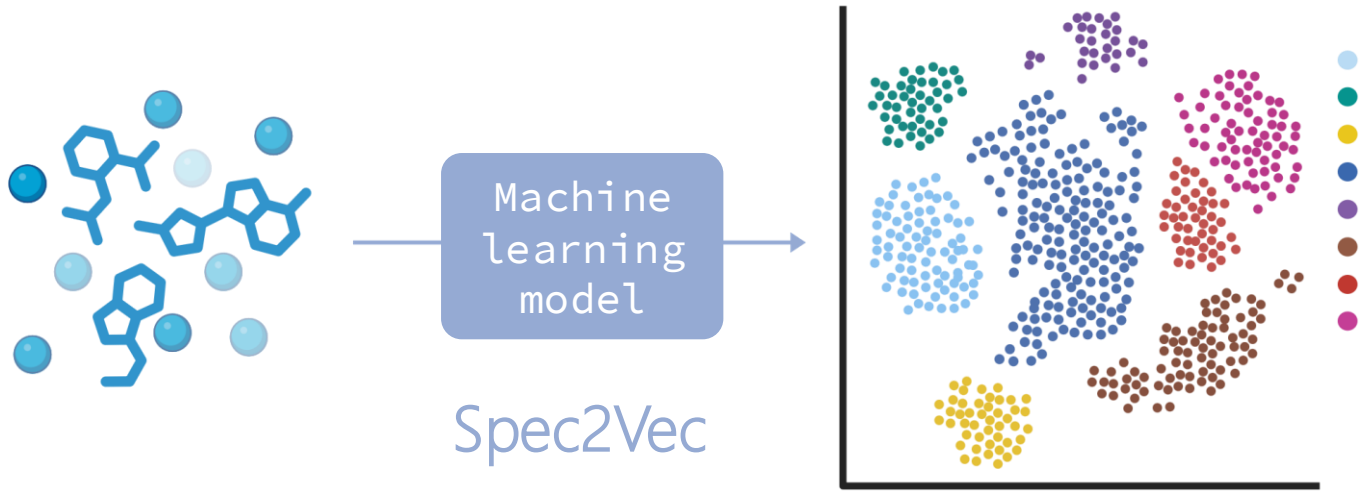
Machine learning-based similarities



Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

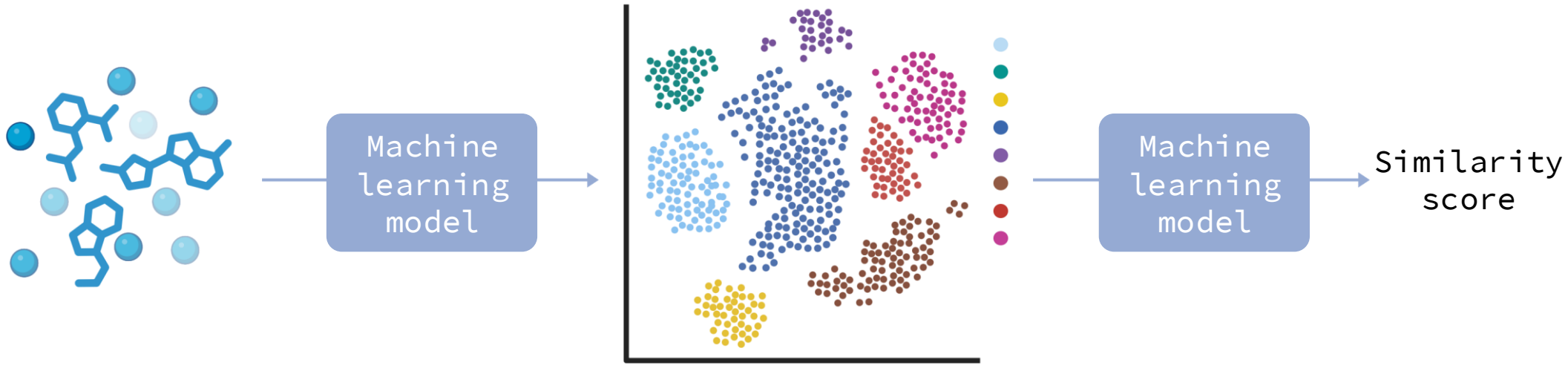
Machine learning-based similarities



Metabolomics Interpretation

Peak annotation using LC-MS^2 spectra

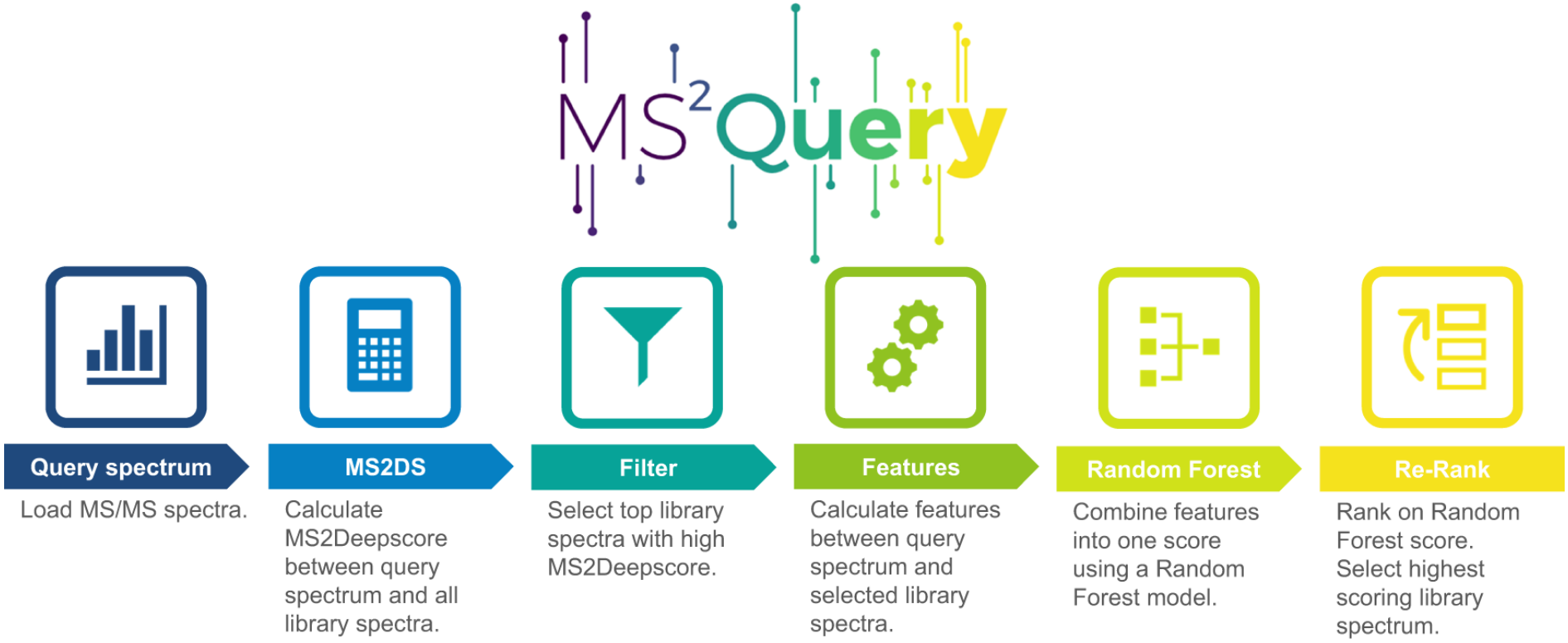
Machine learning-based similarities



MS2Deepscore

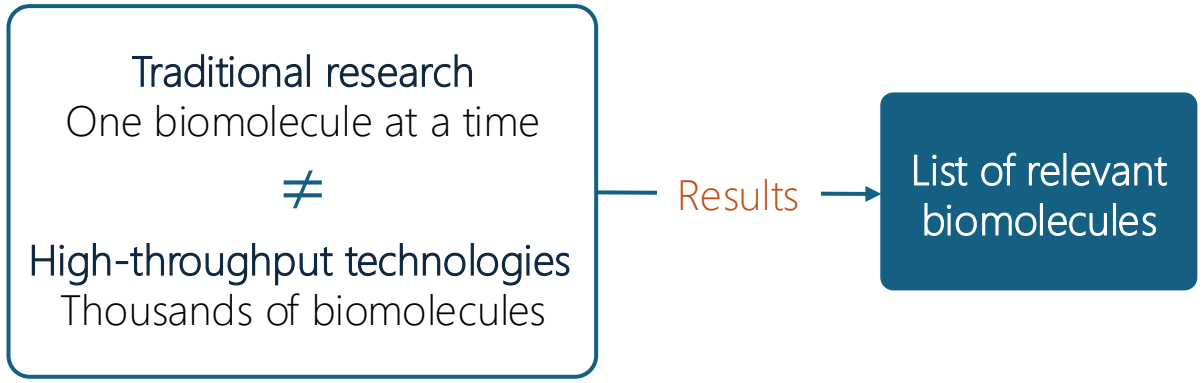
Metabolomics Interpretation

Peak annotation with MS2Query



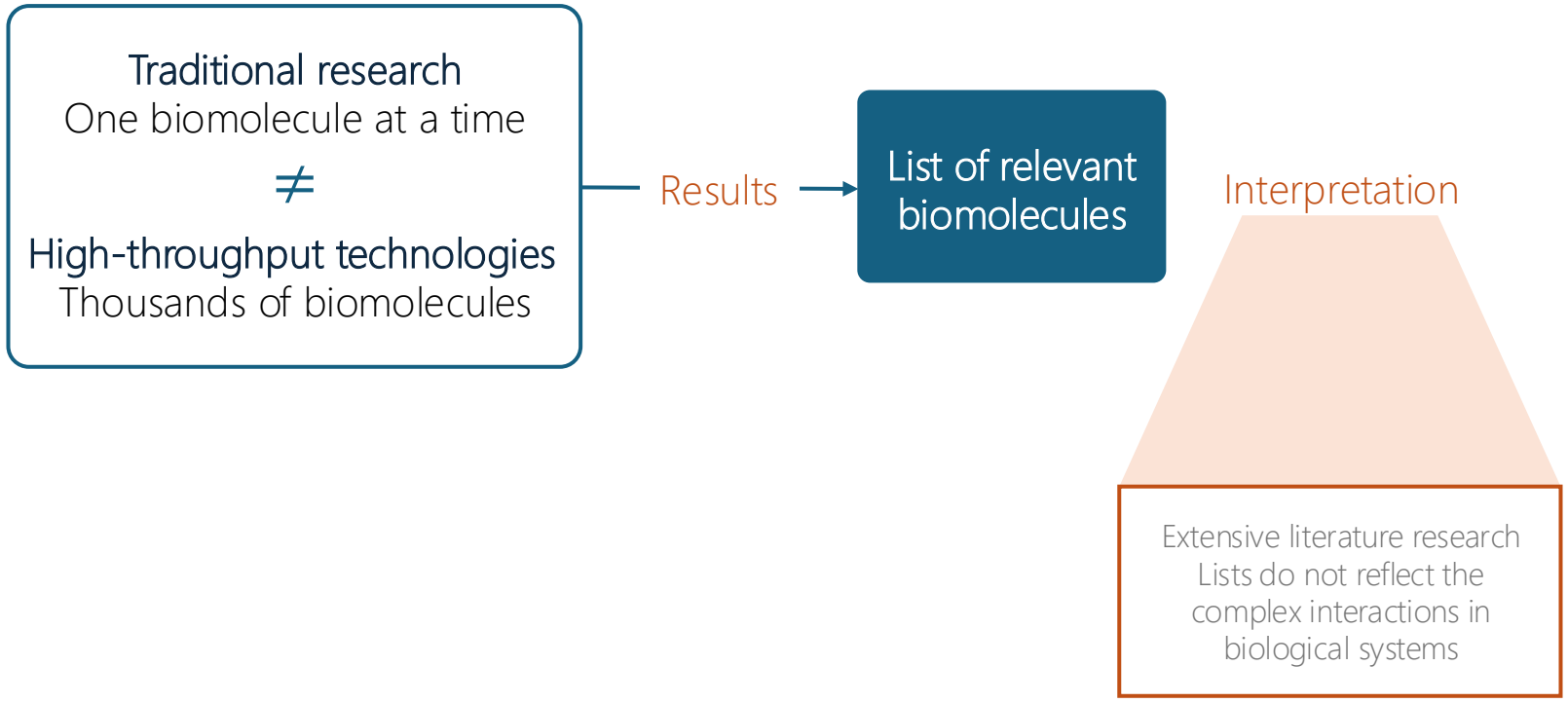
Metabolomics Interpretation

Pathway enrichment



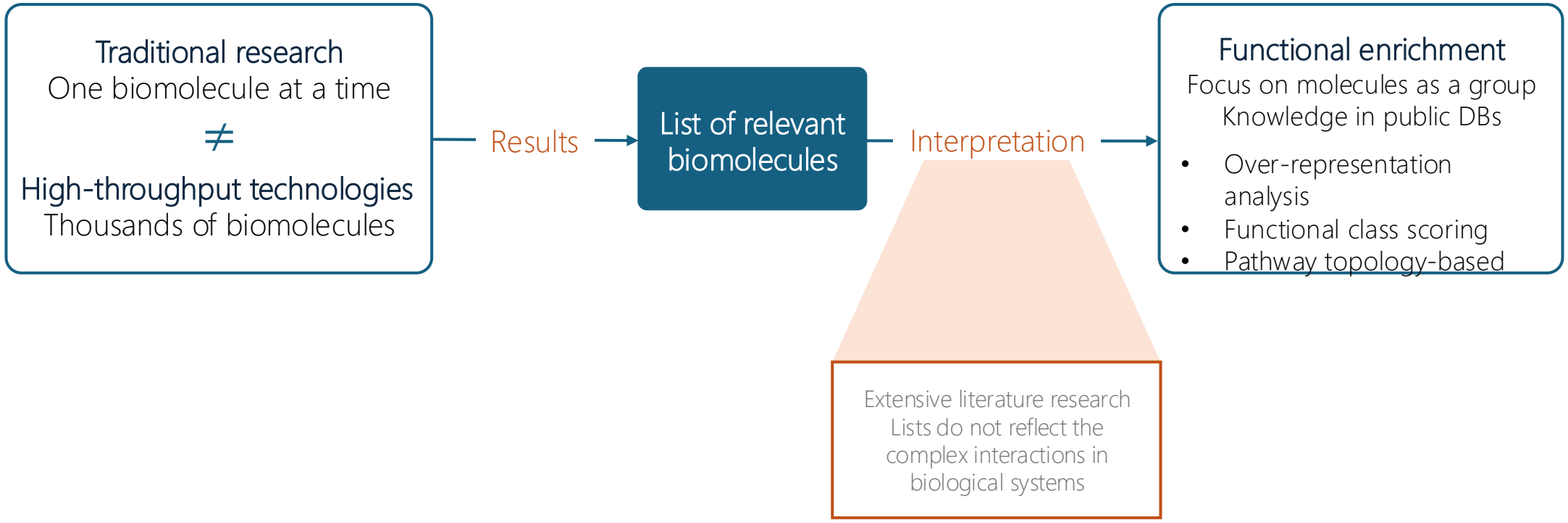
Metabolomics Interpretation

Pathway enrichment



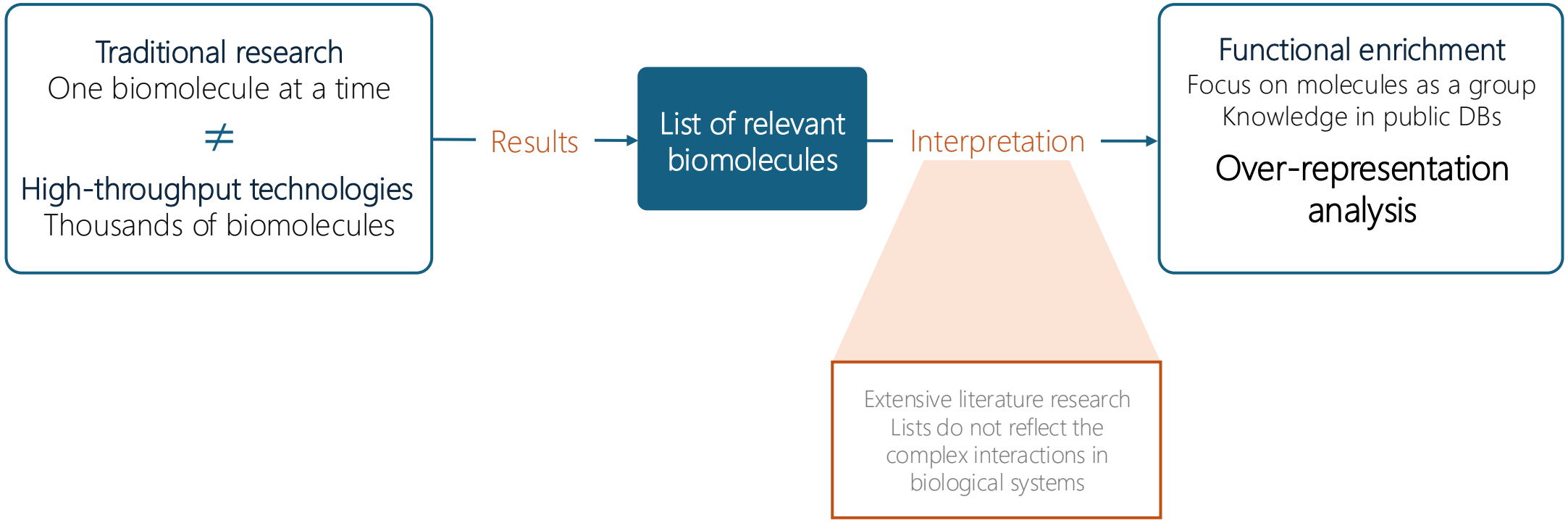
Metabolomics Interpretation

Pathway enrichment



Metabolomics Interpretation

Pathway enrichment



Metabolomics Interpretation

Over-Representation Analysis (ORA)

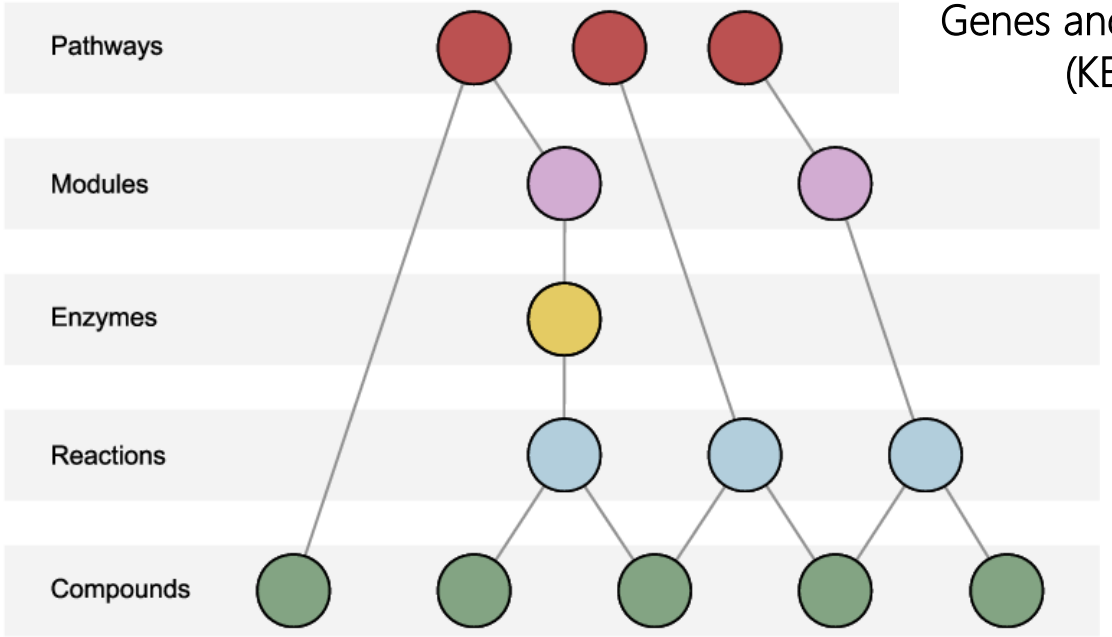
Test if the proportion of differentially expressed molecules within a particular pathway is statistically higher or lower than would be expected by chance.

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Test if the proportion of differentially expressed molecules within a particular pathway is statistically higher or lower than would be expected by chance.

Kyoto Encyclopedia of Genes and Genomes (KEGG)



Picart-Armada, Sergio, et al. "FELLA: an R package to enrich metabolomics data." BMC bioinformatics

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Background: 100 metabolites measured (that belong to 3 pathways)

Pathway	Total # of metabolites
Pathway A	10
Pathway B	20
Pathway C	30
Not in a pathway	40
Total	100

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Background: 100 metabolites measured (that belong to 3 pathways)

10 differentially expressed metabolites

Pathway	Total # of metabolites	Significant metabolites
Pathway A	10	5
Pathway B	20	2
Pathway C	30	1
Not in a pathway	40	2
Total	100	10

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Background: 100 metabolites measured (that belong to 3 pathways)

10 differentially expressed metabolites

How many metabolites would we expect by chance?

Pathway	Total # of metabolites	Significant metabolites	Random
Pathway A	10	5	$10 \times (10/100) = 1$
Pathway B	20	2	$10 \times (20/100) = 2$
Pathway C	30	1	$10 \times (30/100) = 3$
Not in a pathway	40	2	
Total	100	10	

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Background: 100 metabolites measured (that belong to 3 pathways)

10 differentially expressed metabolites

How many metabolites would we expect by chance?

Pathway	Total # of metabolites	Significant metabolites	Random
Pathway A	10	5	$10 \times (10/100) = 1$
Pathway B	20	2	$10 \times (20/100) = 2$
Pathway C	30	1	$10 \times (30/100) = 3$
Not in a pathway	40	2	
Total	100	10	

Metabolomics Interpretation

Over-Representation Analysis (ORA)

Contingency table for Pathway A

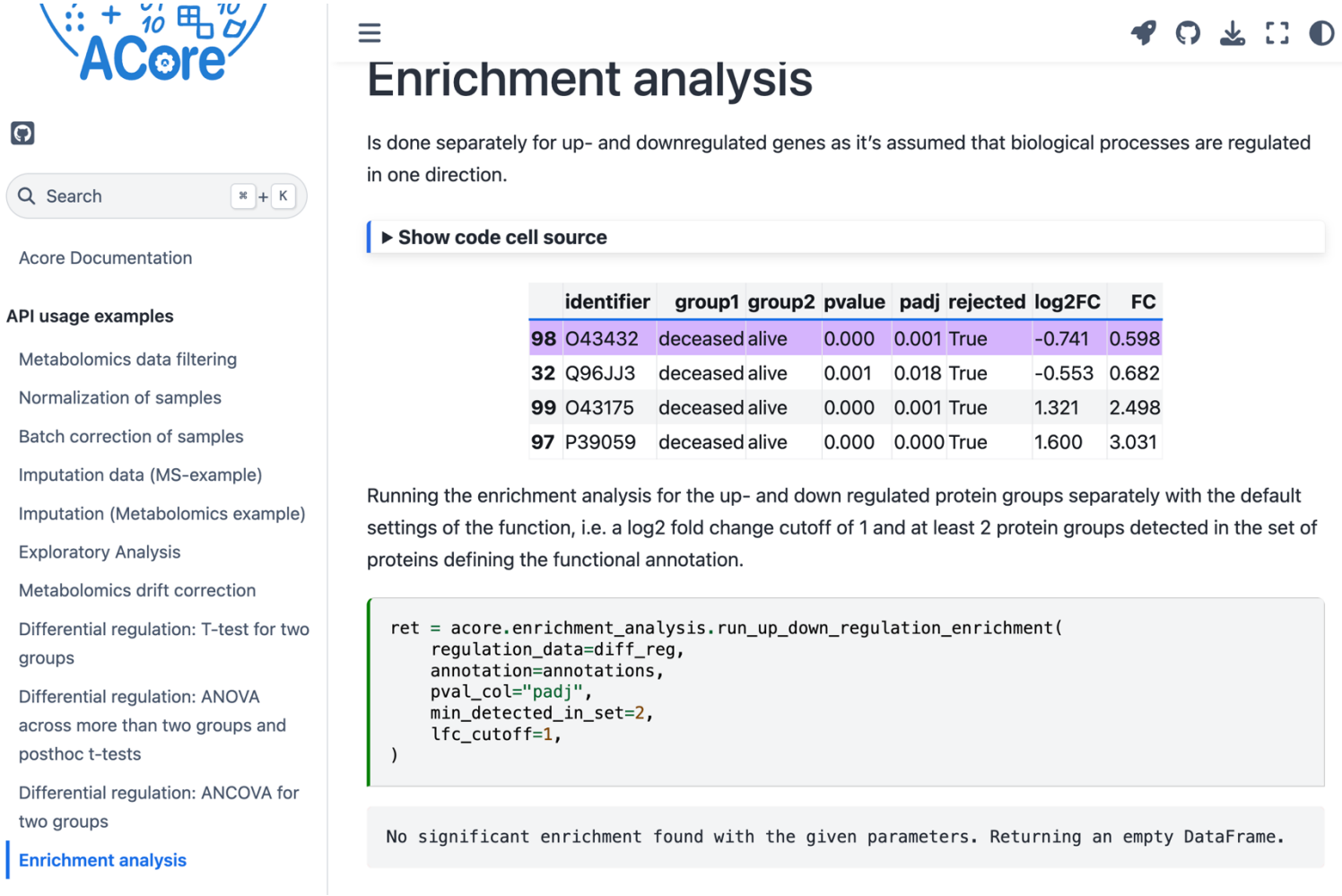
	In A	Not in A	Total
Significant	5	5	10
Not significant	5	85	90
Total	10	90	100

Hypergeometric distribution: $P(X \geq 5)$

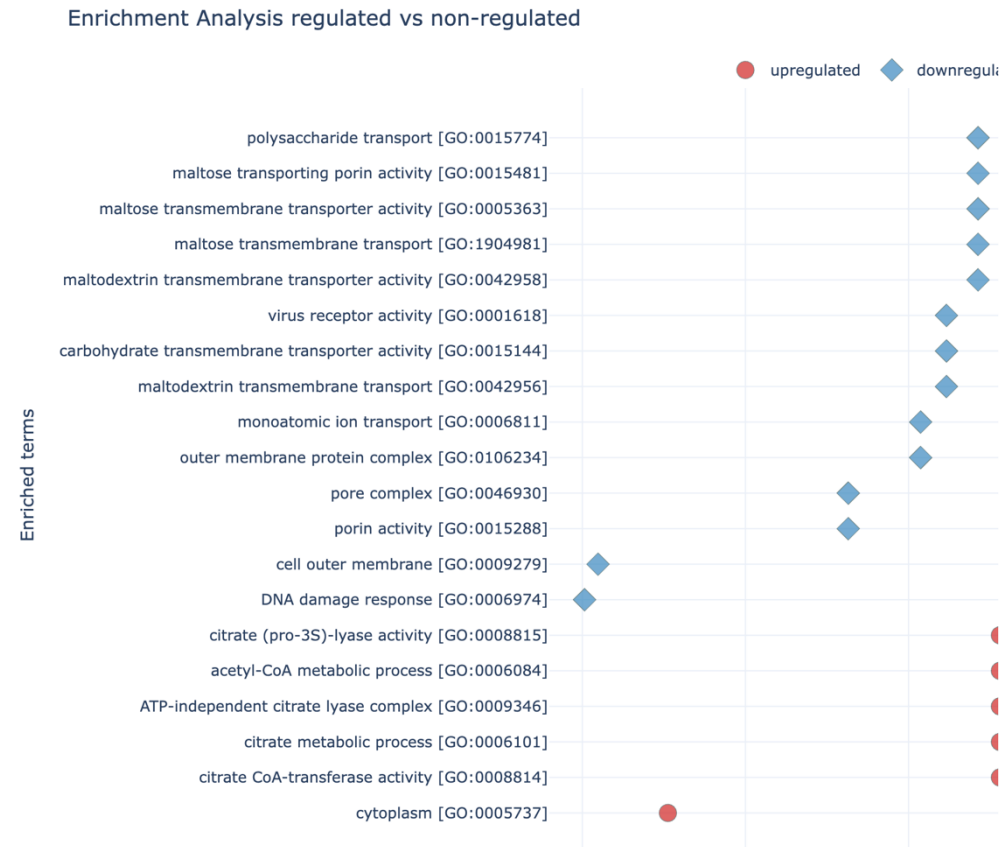
Break!

Interpretation – Hands on

Acore enrichment functionality



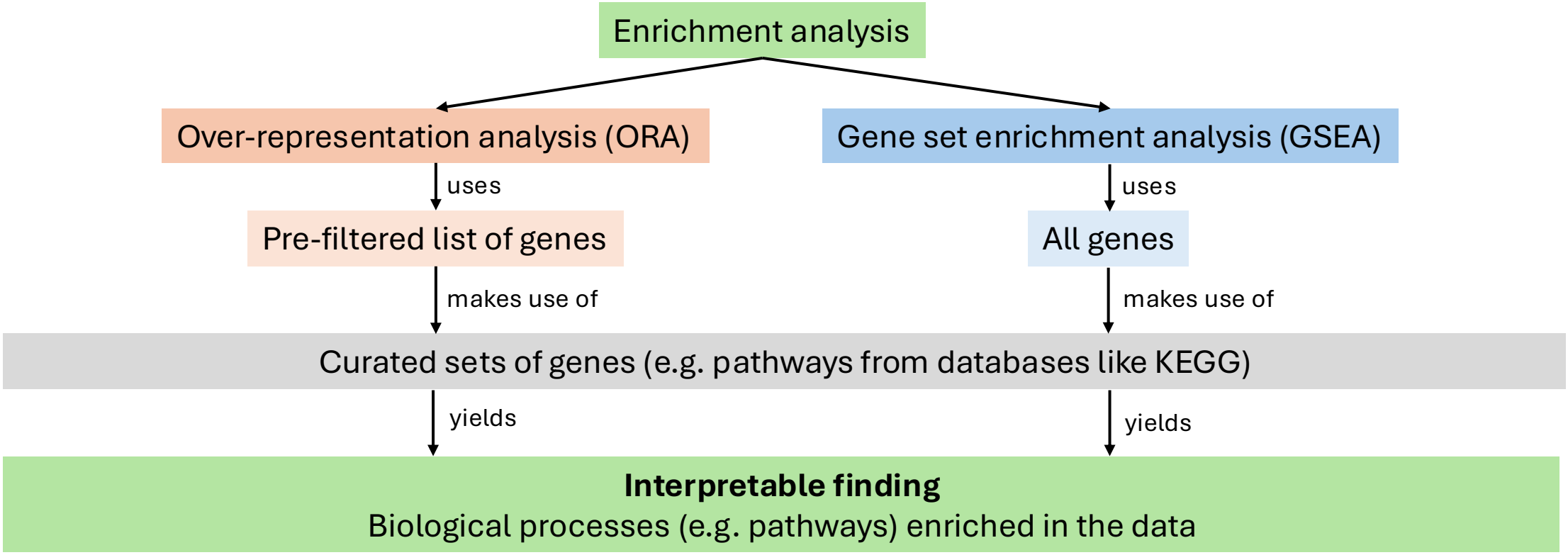
▶ Show code cell source



Some slides from transcriptomics course

Functional analysis for biological interpretation

- **Enrichment analysis**
  - Knowledge-based approach for biological interpretration of omics data, e.g. RNAseq results
  - **Goal:** find biological terms (curated gene sets for biological processes) that are enriched in the data for biological interpretation (no single-gene interpretation)
  - Biological term/pathway lists are obtained from KEGG, GO etc or can be manually curated (often .gmt files)
  - Different types of enrichment analysis



Functional analysis for biological interpretation

- **Gene-set enrichment analysis (GSEA)** – all genes ranked by fold-change and p-value

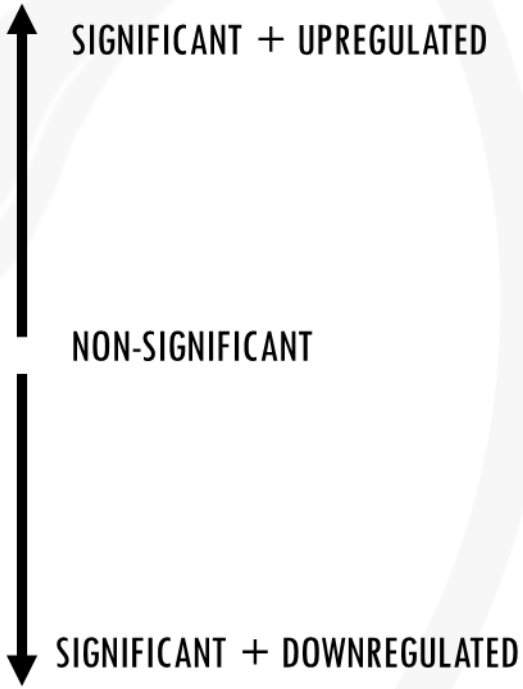
List of **all** genes

	Gene symbol	p-val	Differentially expressed	FC
1	*HAO1*	0.0001	Yes	-5.6
2	*RDH16*	0.00045	Yes	-7.6
3	*ALDOB*	0.00009	Yes	-6.7
4	*AKR1C4*	0.0051	Yes	
5	*ABCB11*	0.067	Yes	
6	*BAAT*	0.0034	Yes	23.5
7	*SLC2A2*	0.00990	Yes	2
8	*FCER1G*	0.00673	Yes	6.7
		...		
6532	*OSM*	0.00009	Yes	9
6533	*VEGFA*	0.067	Yes	8.5
6534	*ZAP70*	0.0034	Yes	7.8

$$RANKING = SIGN(FC) * -\log_{10}(p\text{-value})$$

Ranked list of genes

	Gene symbol	Ranking
1	*JAK9*	12.9
2	*RDH16*	11.4
3	*ALDOB*	11.3
5	*AKR1C4*	10.7
6	*ABCB11*	8.6
7	*BAAT*	3.6
8	*SLC2A2*	0.0099
9	*FCER1G*	0.00673
	...	
6532	*MyD88*	-9
6533	*TRIFF*	-9.37
6534	*IRAK1*	-15.6

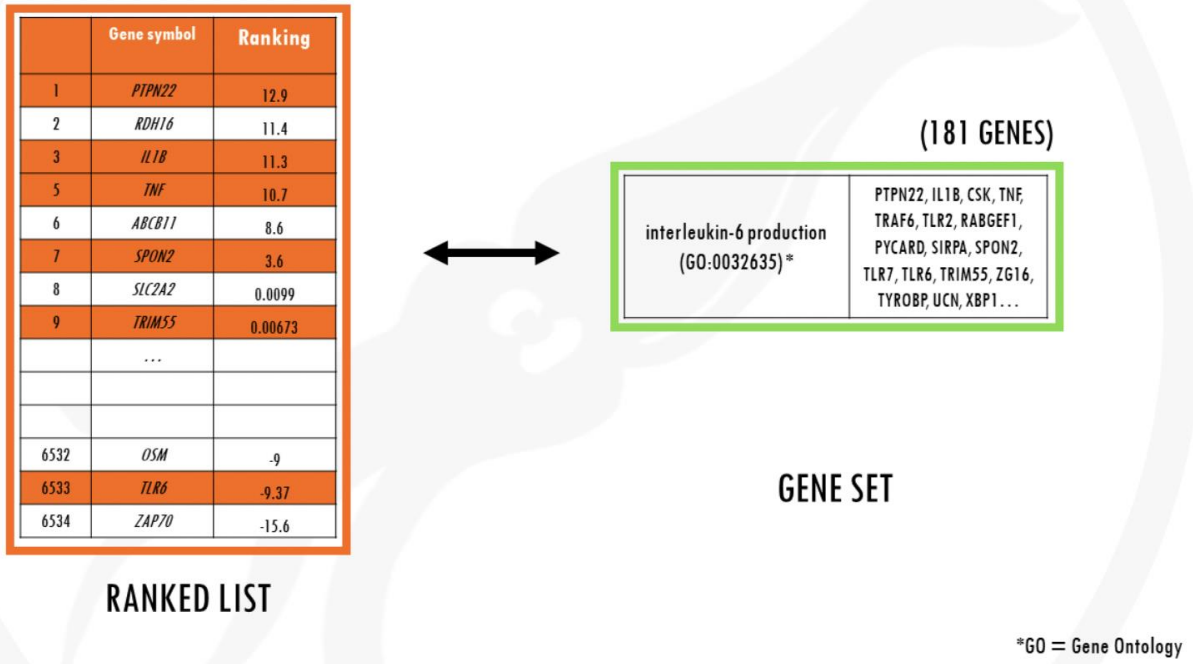


Modified from available illustrations.

https://www.pnas.org/doi/10.1073/pnas.0506580102

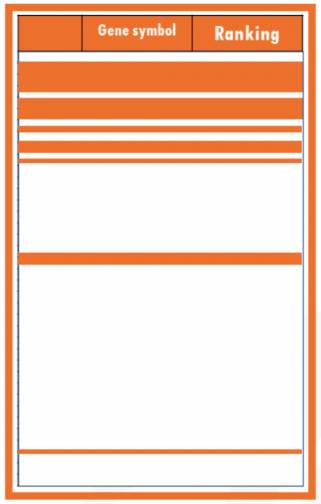
Functional analysis for biological interpretation

- **Gene-set enrichment analysis (GSEA)** – Which biological processes are enriched at the top and at the bottom of the ranked gene list?

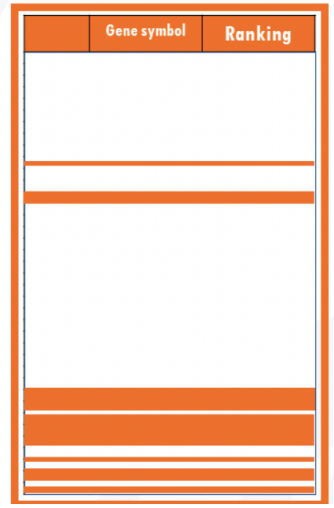


Result will show genes and biological terms that are

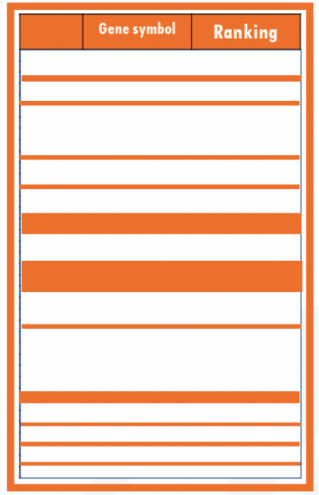
Up-regulated



Down-regulated



Not enriched



Statistical metrics:

- Einrichment score (ES) of a biological term represents the running sum when walking down the list
- Normalised enrichment score (NES): ES is normalised to to make biological terms with different number of genes comparable
- p-value from permutations (adjusted)

Modified from available illustrations.

https://www.pnas.org/doi/10.1073/pnas.0506580102

Functional analysis for biological interpretation

- **Enrichment analysis – details and summary**

	Over-representation analysis (ORA)	Gene set enrichment analysis (GSEA)
Input gene list	Filtered list, e.g. p_{adj}<0.05 and log2FC>|1| → Genes not passing the thresholds will be missed	Entire gene list ranked by fold-change (with sign) and/or p_{adj} → No pre-filtering of genes is an advantage
Separation of up- and down-regulated genes	Can be done	No
Core statistical result and tests used	• Typically an adjusted p value • Fisher's exact test (one-sided), hypergeometric distribution	• Normalized enrichment score (NES) and adjusted p value • Modified Kolmogorov-Smirnov and permutation test
Software packages (selection)	mulea, g:profiler, Enrichr, clusterProfiler, GSEApy (via Enrichr)	fgsea, clusterProfiler, GSEApy