



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

MS-based metabolite profiling in plant

luigi.lucini@unicatt.it





What is a Metabolite?

- Any organic molecule detectable in a system with a MW < 1500 Da
- Includes peptides, oligonucleotides, sugars, nucleosides, organic acids, ketones, aldehydes, amines, amino acids, lipids, steroids, alkaloids, foods, chemical additives, toxins, pollutants,
- Includes human & microbial products
- Concentration > detectable (1 pM)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Metabolite ID

- Two scenarios identification of “known unknowns” and “unknown unknowns”
- For “known unknowns” use spectral or metabolite libraries to ID and quantify via spectral deconvolution
- For “unknown unknowns” (truly novel) use computer-aided structure elucidation methods (CASE)

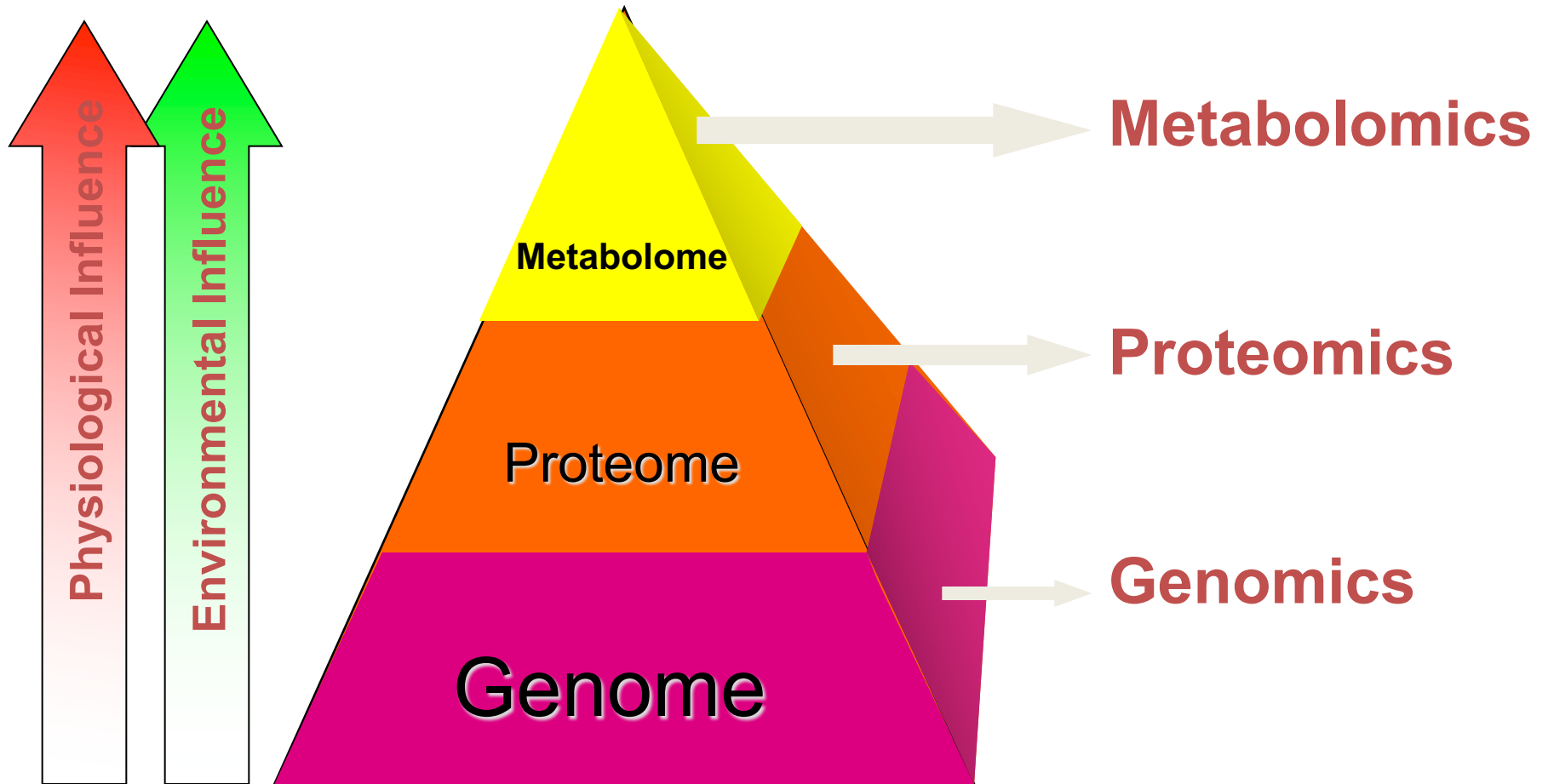
“...there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know.”





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

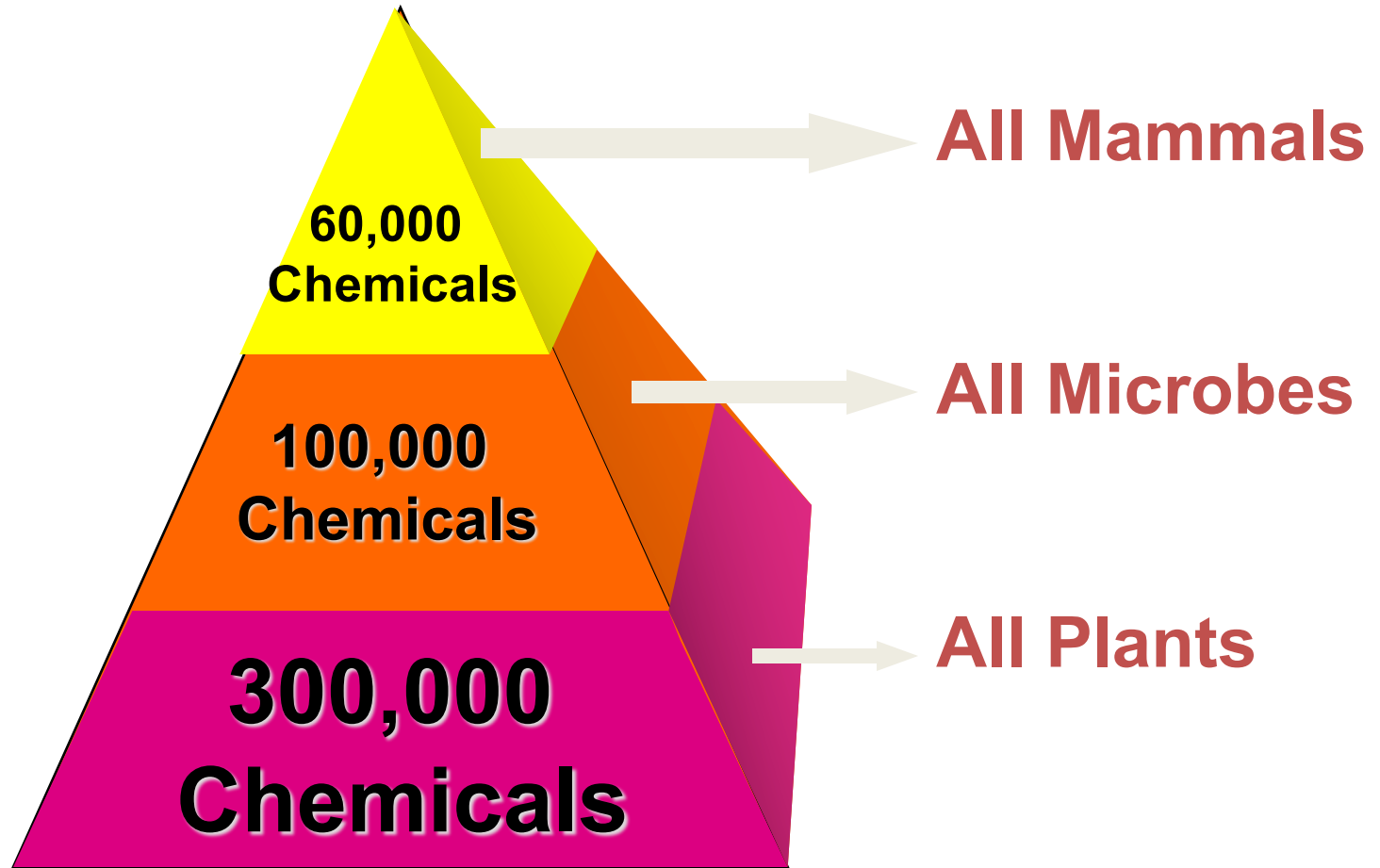
The Pyramid of Life





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Different Metabolomes

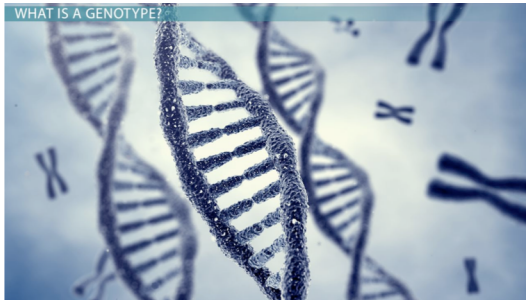


The Pyramid of Life



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

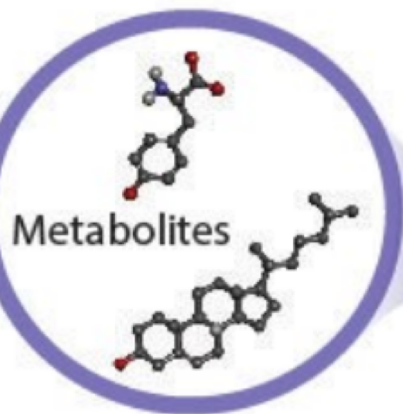
Chemical fingerprints



X



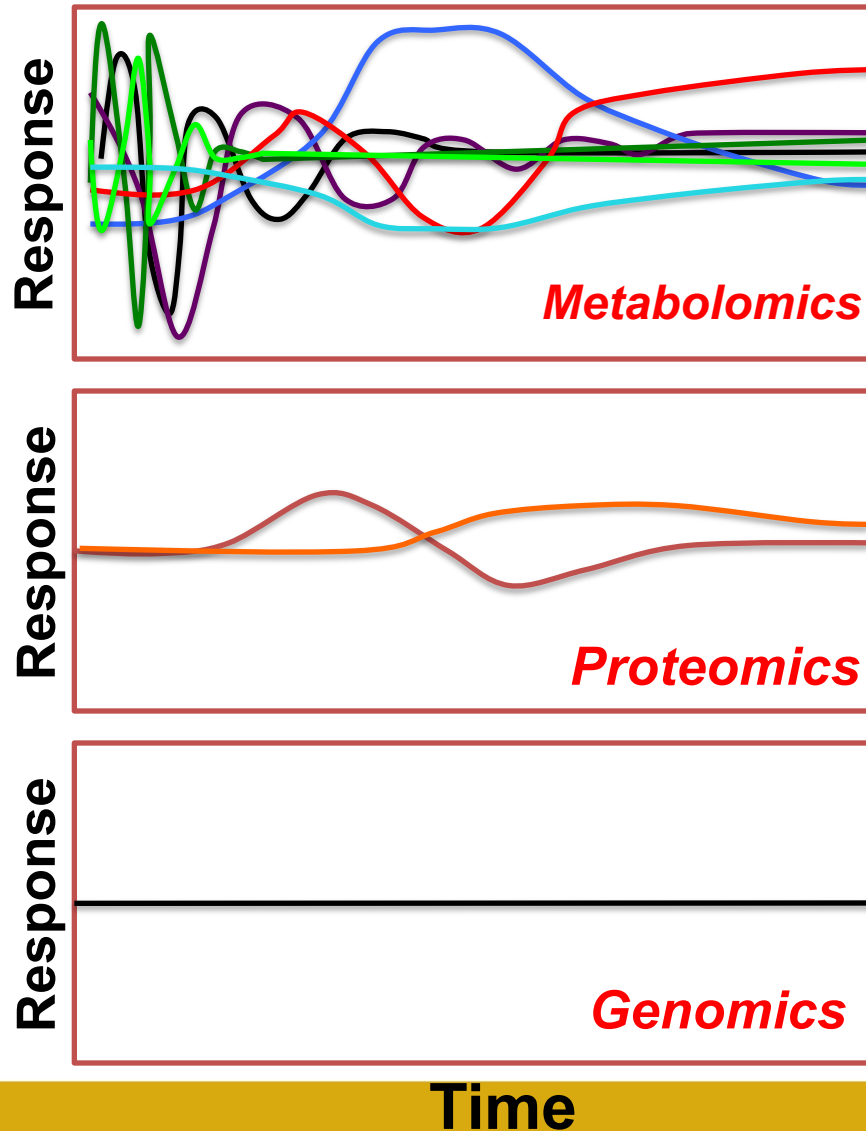
=





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Metabolomics is More Time Sensitive Than Other “Omics”





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Metabolomics Workflow



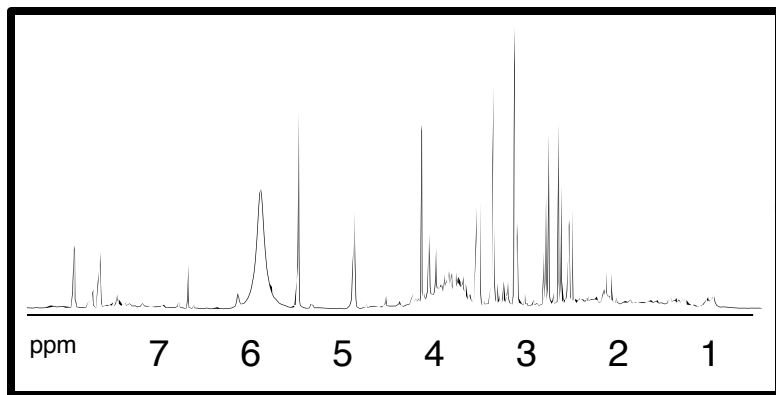
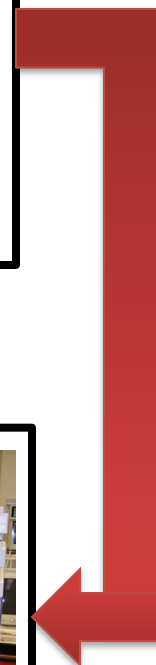
Biological or Tissue Samples



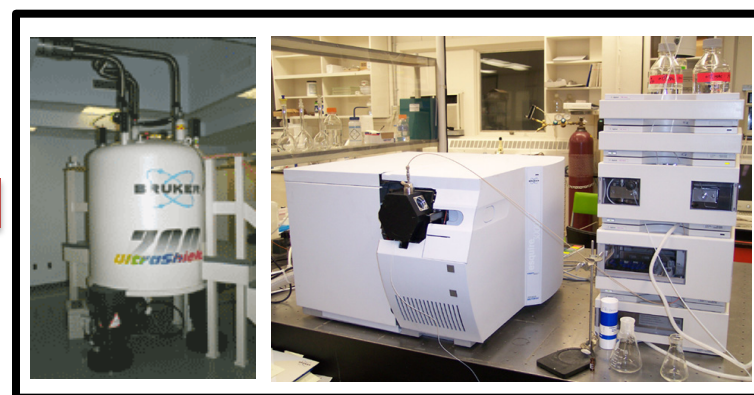
Extraction



Extracts



Data Analysis

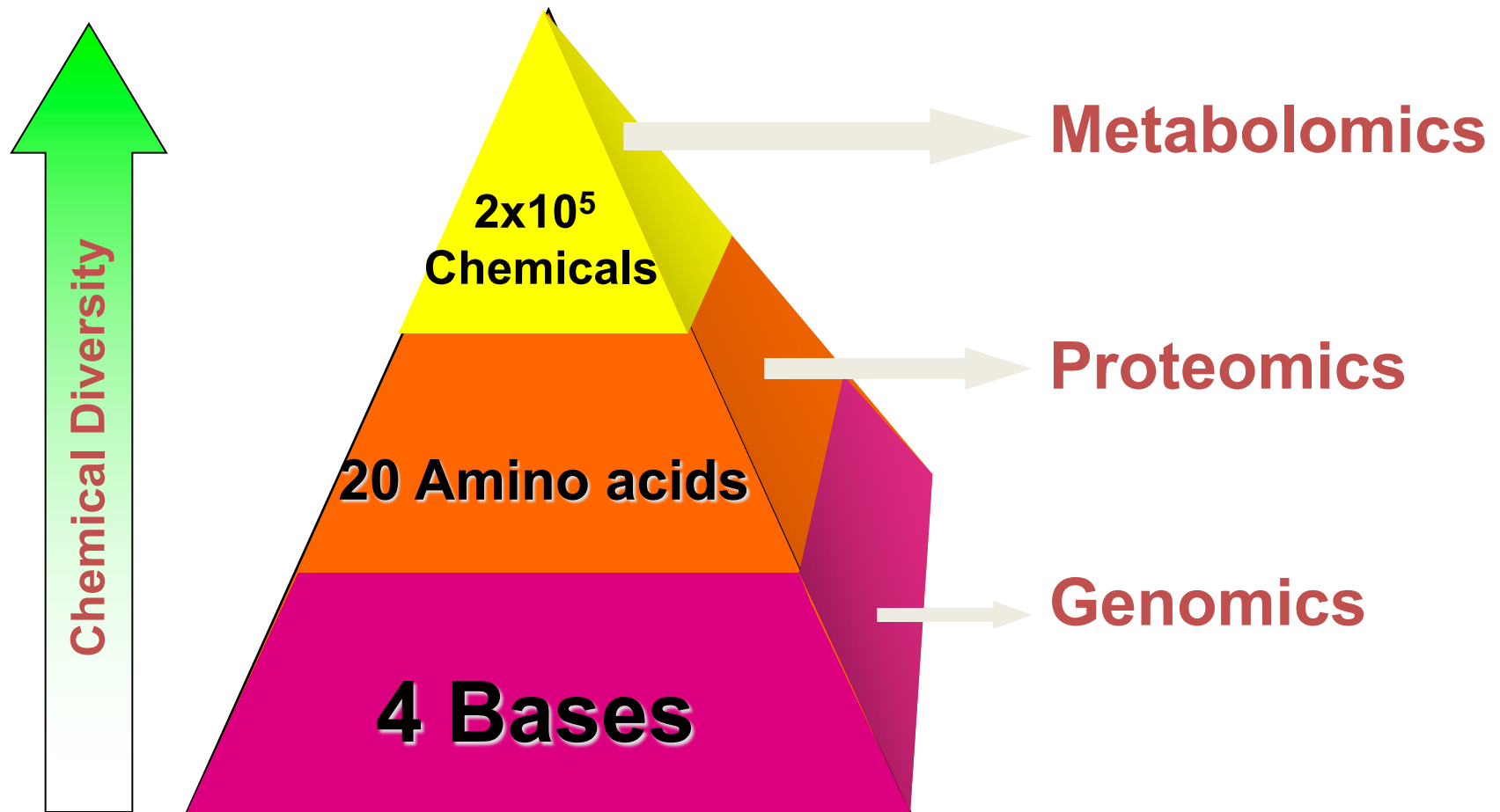


Chemical Analysis



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Chemical diversity



The Pyramid of Life



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

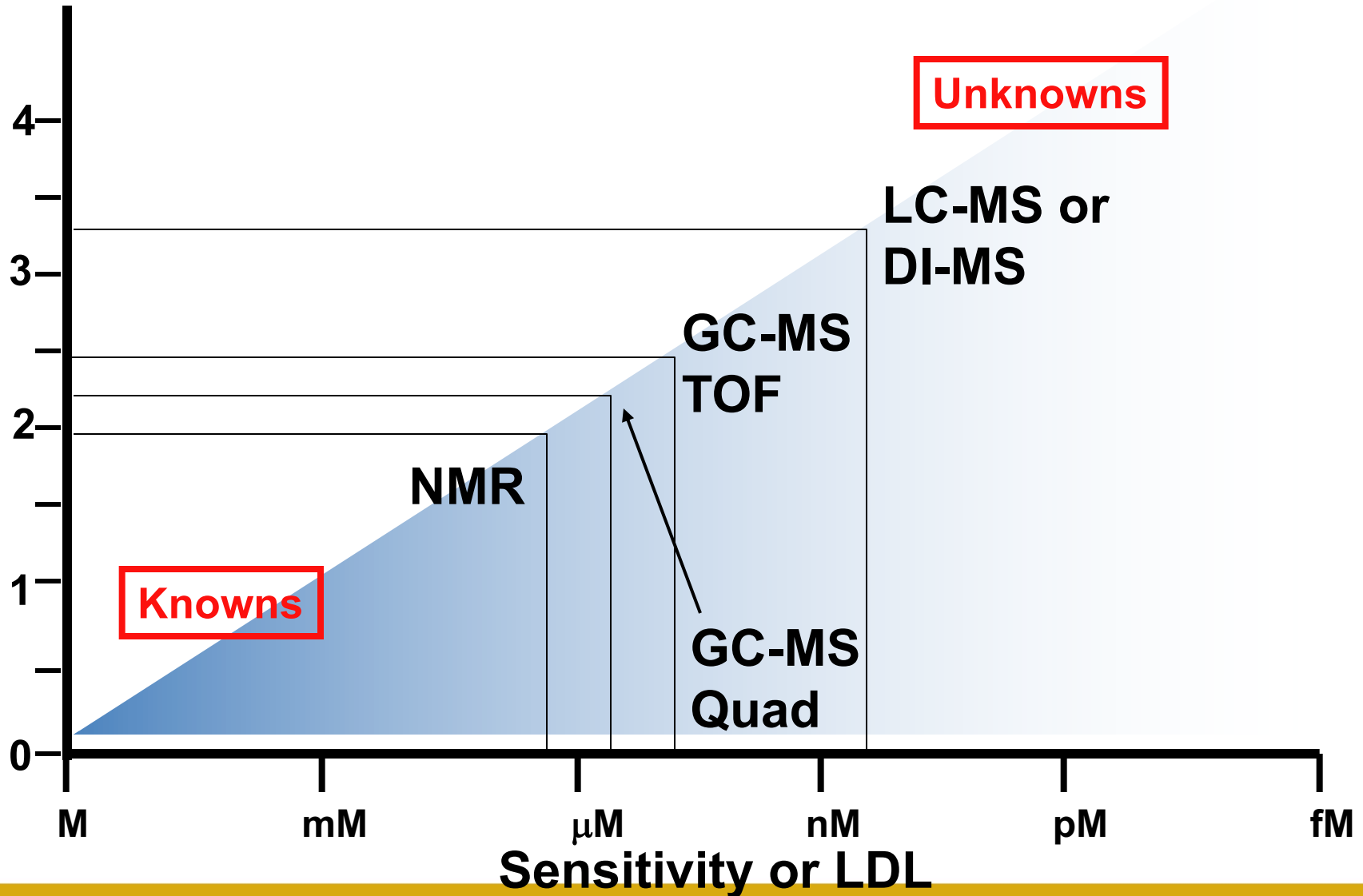
Metabolomics Technologies



- UPLC, HPLC
- CE/microfluidics
- LC-MS
- FT-MS
- QqQ-MS
- NMR spectroscopy
- X-ray crystallography
- GC-MS
- FTIR



Metabolites or Features detected (Log₁₀)



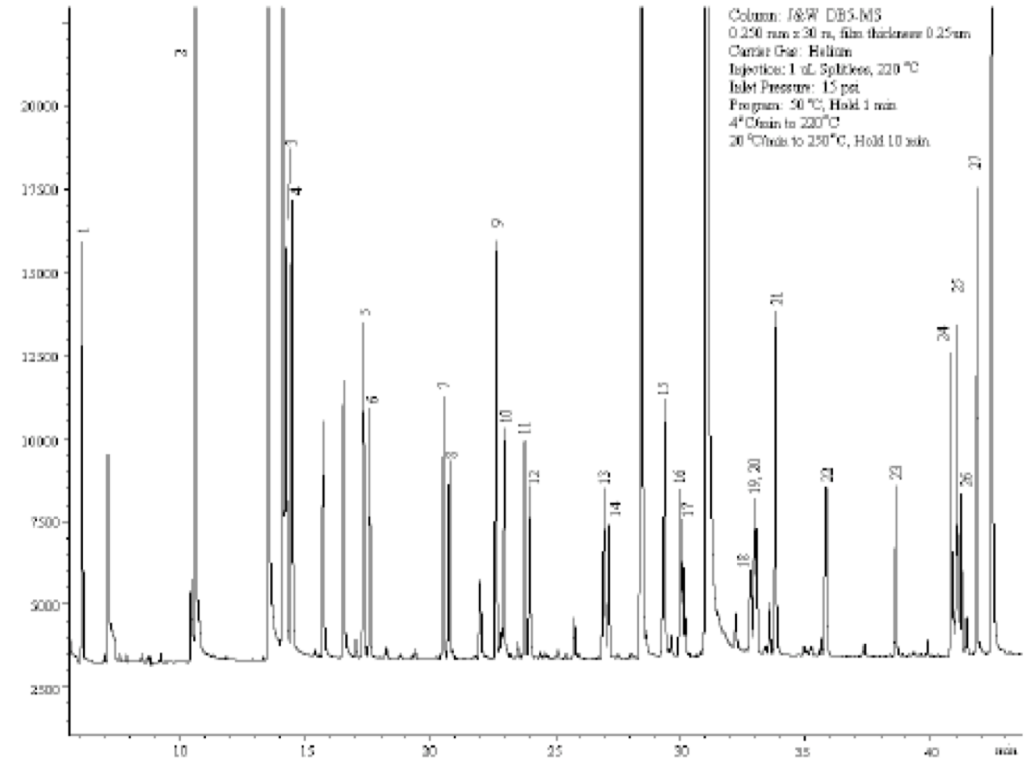
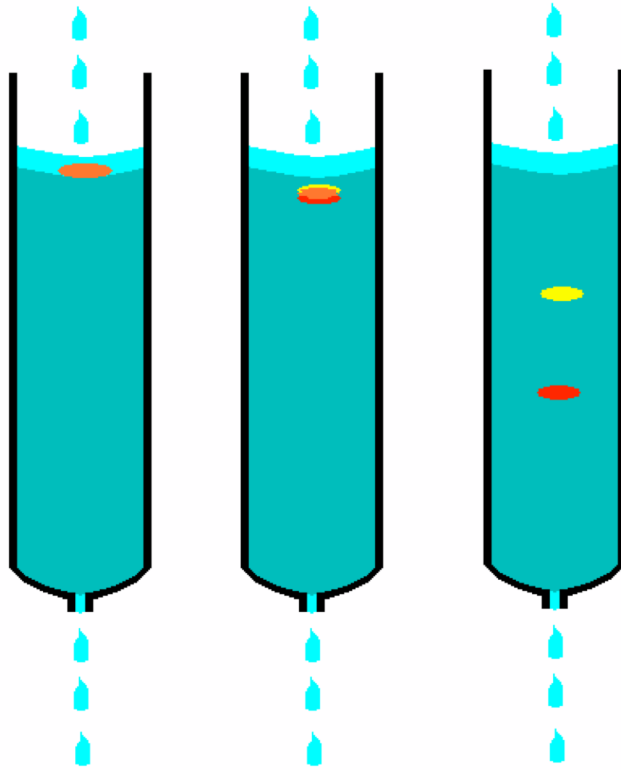
High resolution MS metabolomics

(UHPLC-QTOF)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

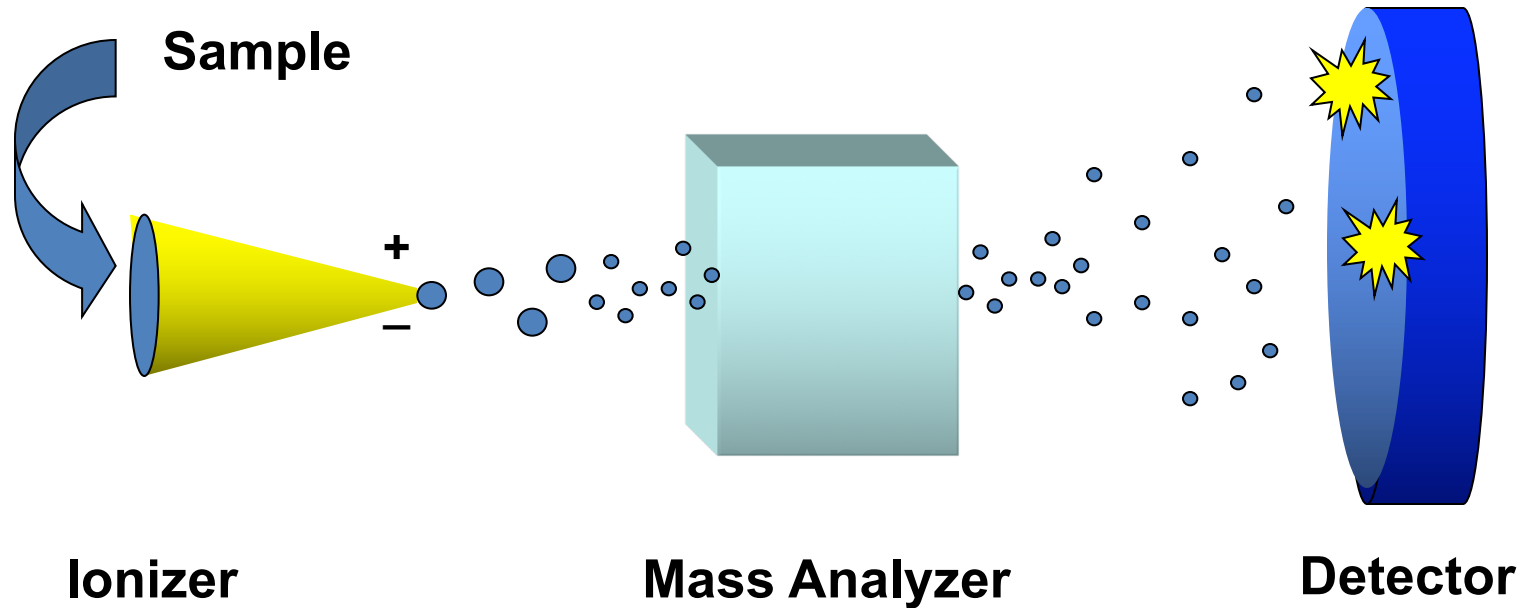
Chromatography





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Mass Spec Principles





Mass Analyzers

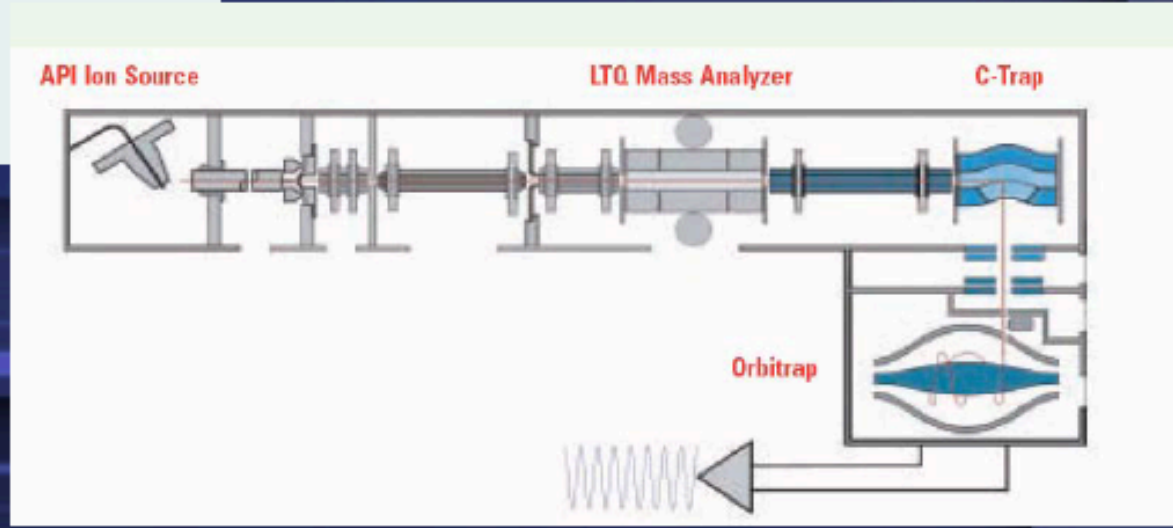
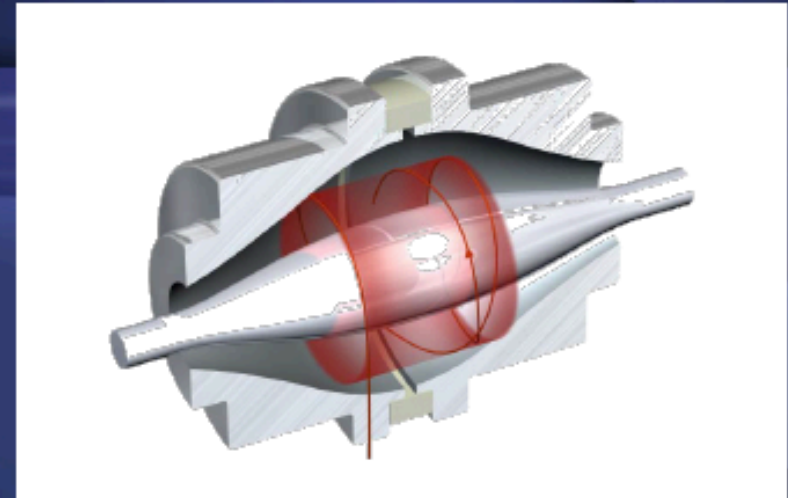
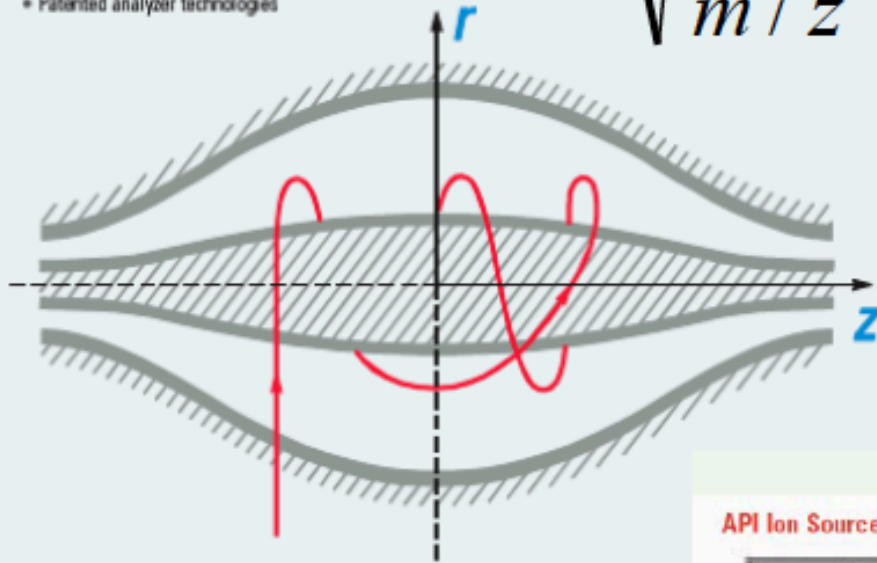
- Magnetic Sector Analyzer (**MSA**)
 - High resolution, exact mass, original MA
- Quadrupole Analyzer (**Q** or **Q***)
 - Low (1 amu) resolution, fast, cheap
- Time-of-Flight Analyzer (**TOF**)
 - No upper m/z limit, high throughput
- Ion Cyclotron Resonance (**FT-ICR**)
 - Highest resolution, exact mass, costly

Orbitrap Mass Analyzer

Orbitrap: A Breakthrough Electrostatic Ion Trap

- Highest ion trapping efficiencies
- Large ion capacity
- Stable and robust operation
- Patented analyzer technologies

$$\omega = \sqrt{\frac{k}{m/z}}$$





orbitrap

The orbital mass analyser is based on the orbital trapping of ions.

Injected ions cycle around the central electrode and at the same time oscillate along the horizontal axis

Orbitrap – Principle of Operation

Kingdon Trap 1923

Diagram illustrating the principle of operation of an Orbitrap mass filter. The central electrode is surrounded by two outer electrodes. The ions are trapped in the center, oscillating along the horizontal axis (z-axis) and rotating around the central electrode. The diagram shows the ion's path (green and red rings) and the resulting signal waveform (red arrow pointing to a blue waveform).

Kingdon Trap 1923

Characteristic frequencies:

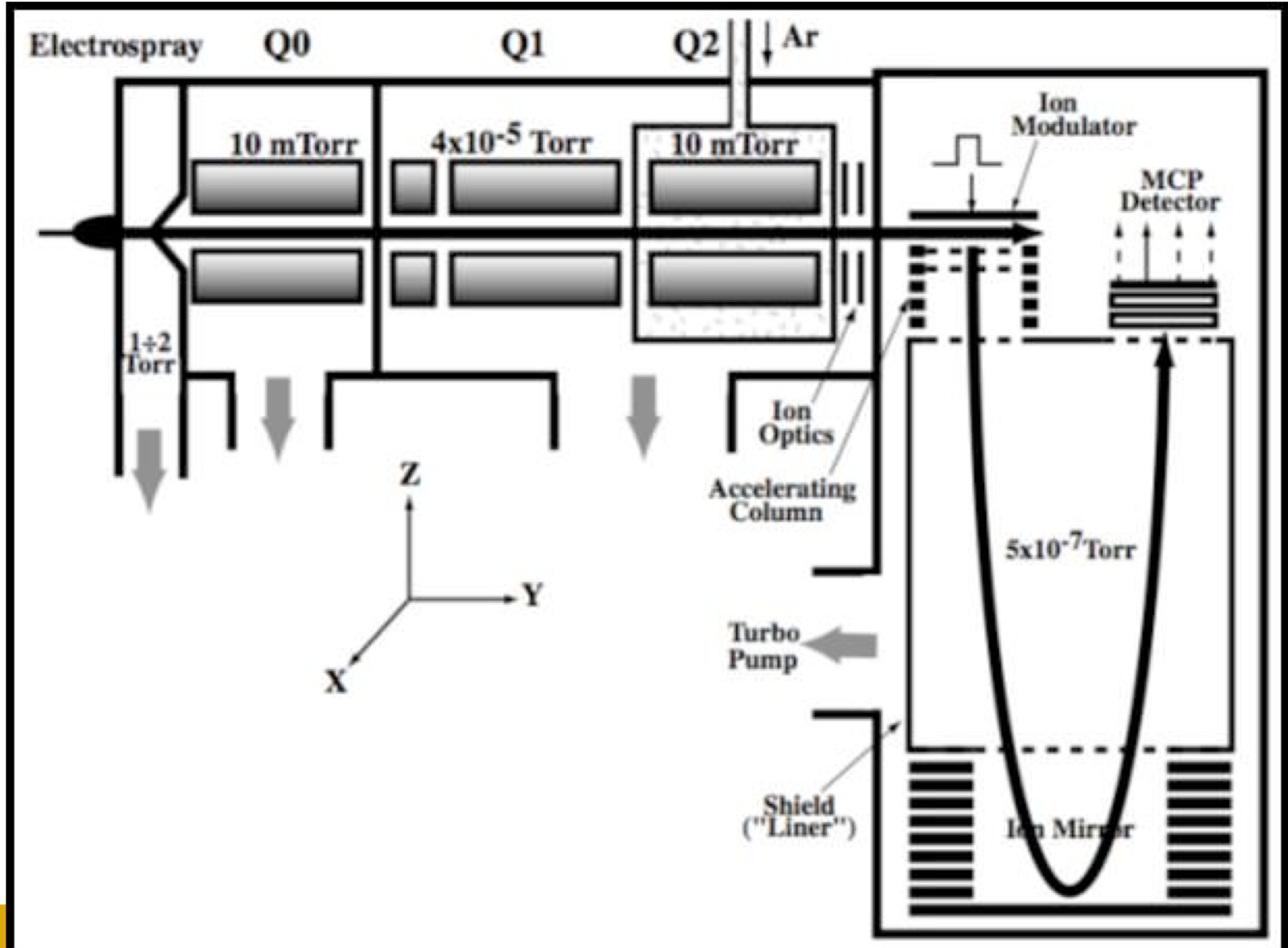
- Frequency of rotation ω_ϕ
- Frequency of radial oscillations ω_r
- Frequency of axial oscillations ω_z

Equation for axial oscillation frequency:

$$\omega_z = \sqrt{\frac{k}{m/q}}$$

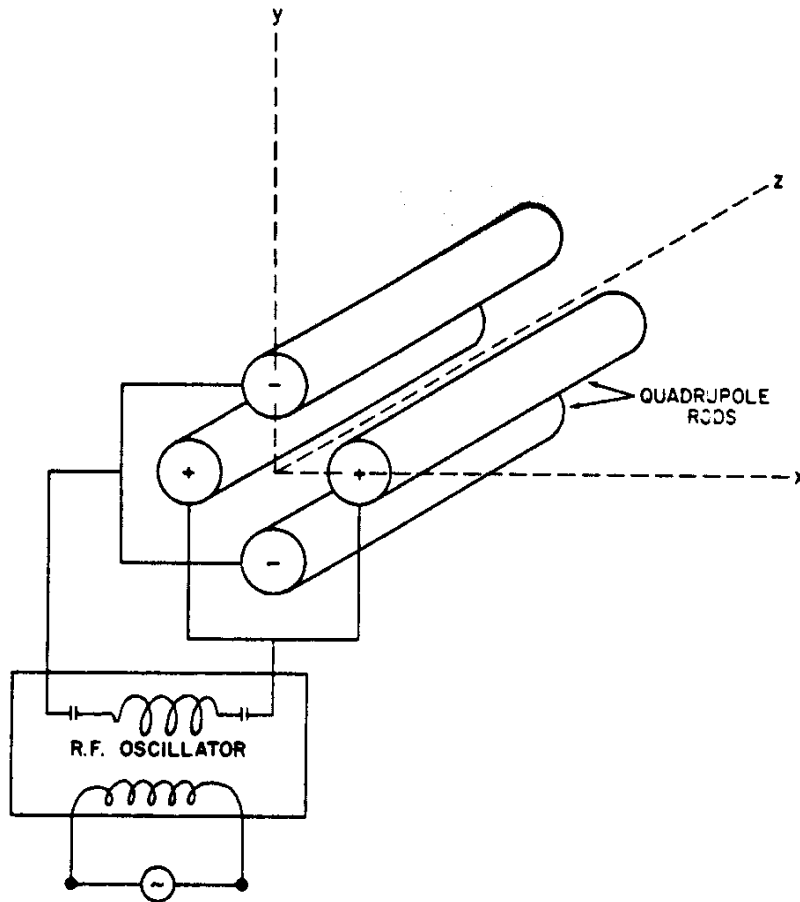


QTOF





Quadrupole Mass Analyzer



Uses a combination of RF and DC voltages to operate as a mass filter.

- Has four parallel metal rods.
- Lets one mass pass through at a time.
- Can scan through all masses or sit at one fixed mass.



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Quadrupoles have variable ion transmission modes



mass scanning mode

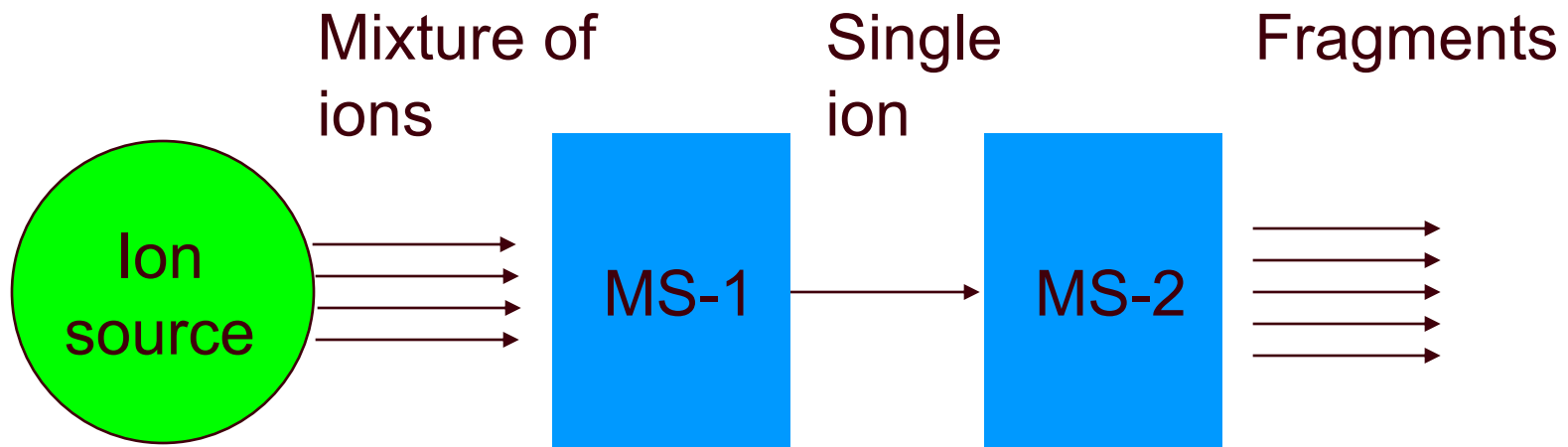


single mass transmission mode



What is MSMS?

MS/MS means using two mass analyzers (combined in one instrument) to select an analyte (ion) from a mixture, then generate fragments from it to give structural information.

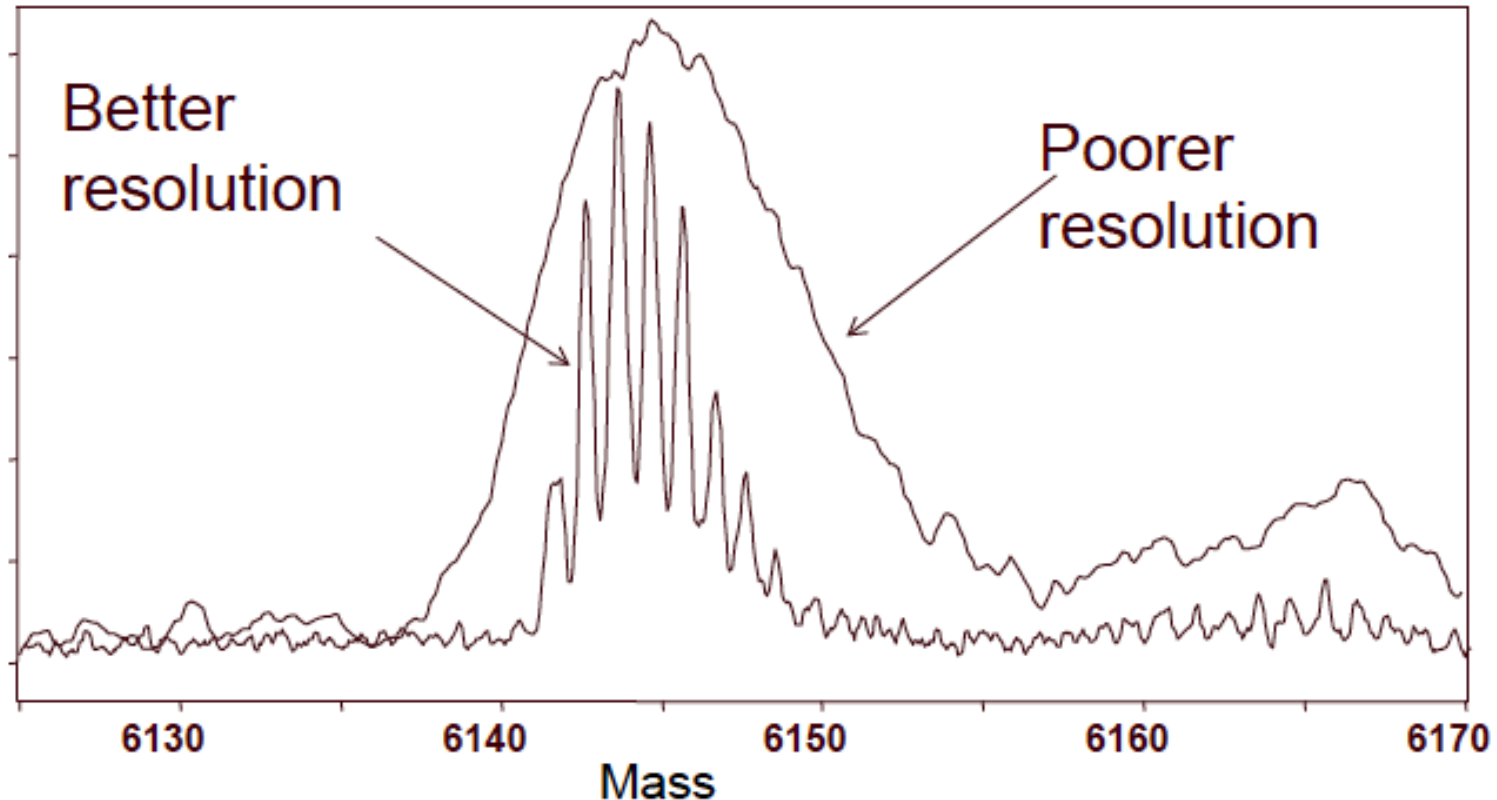




Why HRMS?

What if the resolution is not so good?

At lower resolution, the mass measured is the average mass.





High Resolution MS

Massa isotopica	n. isomeri possibili	Risoluzione
350	3207	400
350,6	298	3500
350,68	27	7000
350,686	4	11500
350,6861	1	44000

$C_{15}H_{94}NP_2$

**Profilo
metabolico**



- **Tracciabilità**
- **Adulterazione**
- **Trattamenti tecnologici**
- **Biomarkers**
- ...



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Why HRMS?

Isotopes

+Most elements have more than one stable isotope.

For example, most carbon atoms have a mass of 12 Da, but in nature, 1.1% of C atoms have an extra neutron, making their mass 13 Da.

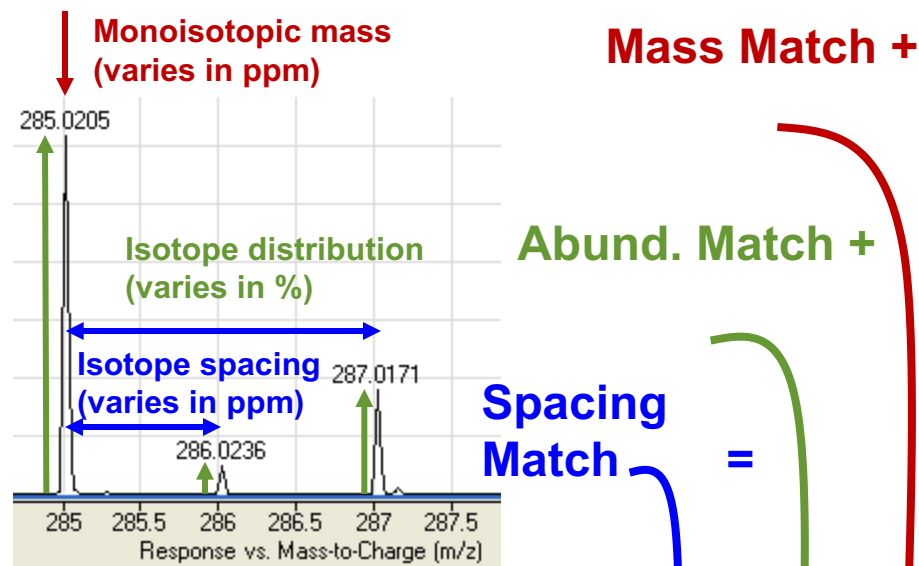
+Why do we care?

Mass spectrometers can “see” isotope peaks if their resolution is high enough.

If an MS instrument has resolution high enough to resolve these isotopes, better mass accuracy is achieved.



Scoring based on:



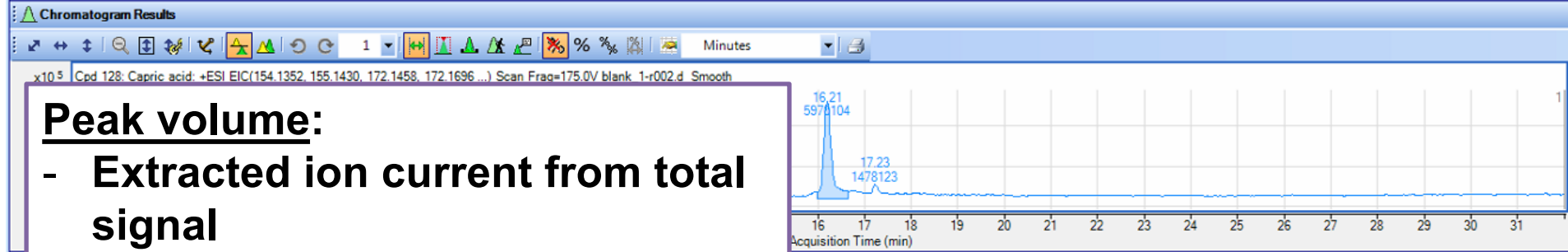
Overall Score

MS Formula Results: Cpd 2: C₁₀H₉CIN₄O₂S

Best	Formula (M)	Calc m/z	Score	Cross Score	Mass	Calc Mass	Diff (ppm)	Abs Diff (p)	Spacing Matc	Abund Matc	Mass Match	m/z	DBE
<input checked="" type="checkbox"/>	C ₁₀ H ₉ CIN ₄ O ₂ S	285.0208	99.55		284.0137	284.0135	-0.71	0.71	99.19	99.26	99.69	285.021	8
<input type="checkbox"/>	C ₇ H ₁₂ N ₂ O ₆ S ₂	285.021	77.28		284.0137	284.0137	0.01	0.01	99.54	1.93	100	285.021	3
<input type="checkbox"/>	C ₇ H ₁₃ CIN ₄ O ₂ S ₂	285.0241	75.57		284.0137	284.0168	11.12	11.12	99.87	83.87	46.22	285.021	3

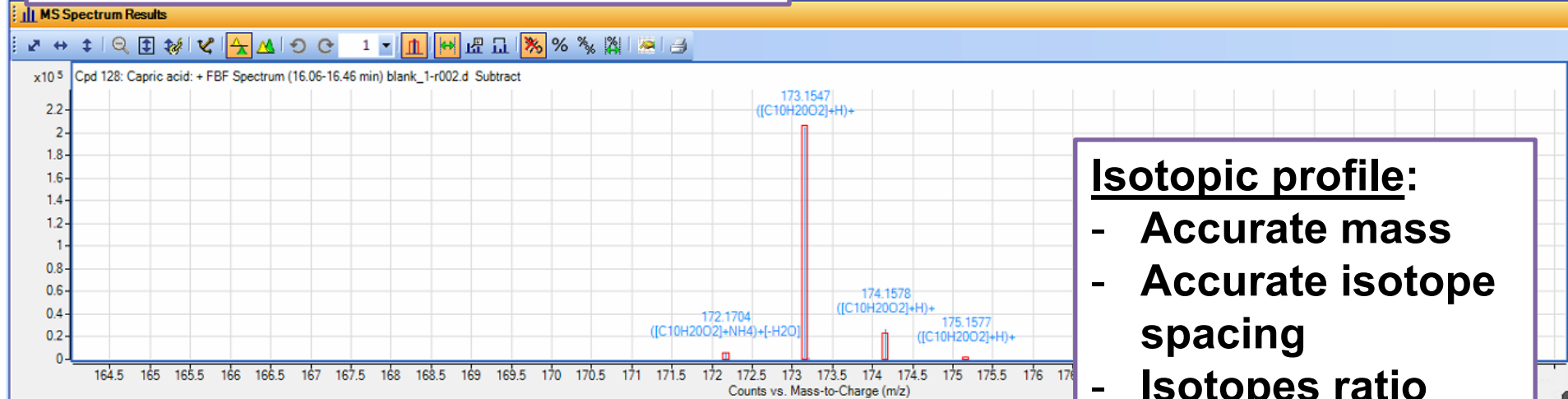


45	L-valine	Cpd 45: L-valine	C5 H11...	11...	11...	117...	3.4	99...	118...	7...	35...	36...	5...	11...	7	aminoacids
1.	Indol	Cpd 106: Indol	C8 H7 N	11...	11...	117...	4.62	76...	118...	3...	16...	16...	0...	53...	4	antimicrobials
1.	Capric acid	Cpd 128: Capric acid	C10 H2...	17...	17...	172...	4.21	93...	249...	1...	48...	20...	0...	59...	7	fatty acids
1.	Myristic acid	Cpd 130: Myristic acid	C14 H2...	22...	22...	228...	-0.06	85...	344...	2...	97...	40...	0...	92...	5	fatty acids
1.	Stearic acid	Cpd 132: Stearic acid	C18 H3...	28...	28...	284...	-0.07	99...	413...	2...	16...	97...	0.1	10...	8	fatty acids



Peak volume:

- Extracted ion current from total signal



Isotopic profile:

- Accurate mass
- Accurate isotope spacing
- Isotopes ratio

Text/Synonym Search

Quick Search

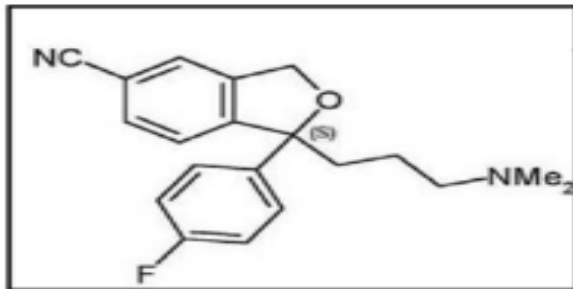
Keywords :

Location (optional) :

*Go

Location Hint: Enter a city, a state, city comma state, or zip code.

Structure Search



Mol Weight Search

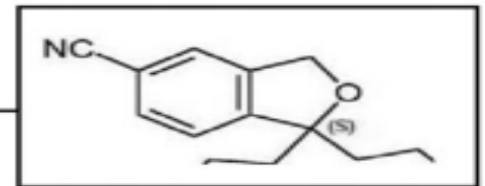


Smiles & Formula Search

Smiles & Formula Search

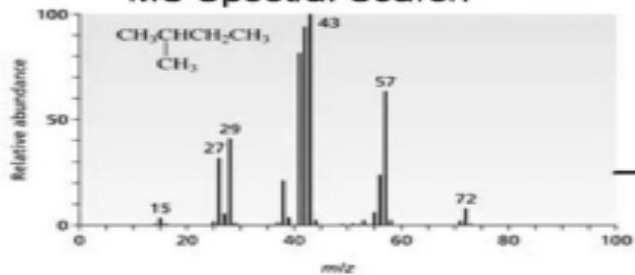
CN1C=C(N=C1)

Substructure Search

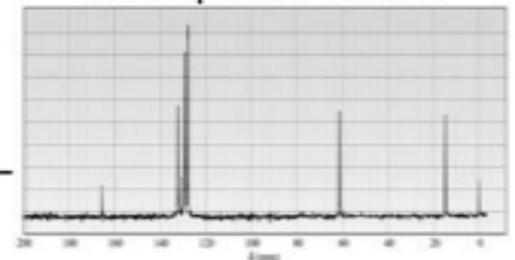


Human
Metabolome
Database
(HMDB)

MS Spectral Search



NMR Spectral Search



METQPRDSKNACDEG
LIGGEDCANKSDRPQ
VTREYQPAS FHGFDL
PIYTREWQQNMLKAF

Sequence Search



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

MS Spectral DBs

NIST/AMDIS

The NIST 08 Mass Spectral Library (NIST/EPA/NIH)--New 2008 Version

Scientific Instrument Services, Inc.

Search

Home Products Catalog Services Machining References Customers Dealers Company

Mass Spec | Gas Chrom | Liquid Chrom | Thermal Description | Vacuum Systems | General Lab | Cleaning | Software

Software

MS/GC Software

Highlighted Products

EMULSION 3D - The industry standard for electron and ion simulation.

NIST 08 - The fully evaluated collection of electron impact (EI) mass spectra.

The NIST 08 mass spectral database, the successor to the NIST 05, is a fully evaluated collection of electron impact (EI) mass spectra. It also contains MS/MS spectra and GC data. It is the product of a two decade, comprehensive evaluation and expansion of the world's most widely used mass spectral reference library by a team of experienced mass spectrometrists in which each spectrum was examined for correctness.

NIST Components

NIST is not just a mass spectral library. It contains these components:

- **UPDATED Electron Impact (EI) mass spectral library** - 220,460 spectra of 192,108 unique compounds, with identifications and usually chemical structures. You may search names of compounds online.
- **UPDATED MS/MS library** - 14,852 spectra of 5,308 precursor ions (3,898 cations and 1,410 anions).
- **UPDATED Gas Chromatography (GC) data library** - 224,038 Kovats retention index values for 21,847 compounds in the EI library, now on both polar and non-polar columns. Includes retention indices with GC column conditions and literature citations.
- **UPDATED NIST MS Search software** - software for searching (identifying) compounds from their mass spectra and for browsing mass spectral libraries. Also includes MS interpretation programs for analyzing mass spectra on the basis of chemical structure, molecular formula, isotopic patterns, and more.
- **UPDATED AMDIS software** - software for deconvoluting gas/liquid chromatograms
- **UPDATED Documentation** - Approximately 50 page printed and electronic manual on

Figure: EI spectra, structure, and corresponding data for a sample compound in the NIST database.

Metlin

Scripps Center For Metabolomics
METLIN: Metabolite and Tandem MS Database

MS HOME Overview Search XCMSOnline Software/Services Metabolomics Science Publications

CHOLINE ADENOSINE SERINE TRYPTOPHAN PYRUVIC ACID TESTOSTERONE PYRUVIC ACID PYRUVIC ACID GLUCOSE NICOTINAMIDE SERINE TRYPTOPHAN PYRUVIC ACID TESTOSTERONE GLUCOSE INFLAMMATION NICOTINAMIDE SERINE TRYPTOPHAN

Statistics

- # Metabolites: 240,478
- # High Resolution MS/MS Spectra: 60,266
- # Metabolites w/ High Resolution MS/MS: 11,793

example | details...

Functionality

- Single & Batch Precursor Ion (m/z) searching
- Single & Multiple Fragment Ion (m/z) searching
- Neutral Loss searching
- De Novo Fragment Characterization

GolmDB

GolmDB - the Golm Metabolome Database

GOLM METABOLOME DATABASE

Mass Spectrum of Alanine (2TMS)

reference spectrum, search or click to enlarge

Many current approaches in metabolomics are characterized by an ongoing transition from qualitative towards quantitative methods, similar to the previous development in genomics, transcriptomics and proteomics. The application of state-of-the-art high-throughput technologies results in a significant growth in size and complexity of the generated data leading to an increased demand for computational methods of visualization, annotation and data mining.

The GMD incorporates quantitative data of metabolite pool size changes. As such, the GMD constitutes a storage system for data on the molecular level, providing access for computational methods based on the analytical results.

The GMD facilitates orthogonal metabolite profiling; this is to say, to profile a single metabolite across multiple experimental conditions or to profile all metabolites within a single experimental setup. Using this unique data set as a reference, the GMD classifies unknown metabolic signatures with respect to species and experimental factors.

Your use of this site is governed by our Terms of Use linked below.

This project is supported and funded by the Max Planck Society, the German Research Foundation, DFG and the European Union.

MassBank

MassBank - High Resolution Mass Spectral Database

MassBank | High Resolution Mass Spectral Database

Most Visited Getting Started Latest Headlines

MassBank | High Resolution Mass

MassBank High Resolution Mass Spectral Database

Database Service

- Statistics
- Documents
- Download
- Manuals
- About MassBank
- Contact
- Consortium Members
- Site Map

News

- May 31, 2011 Manual page was updated. [new](#)
- May 12, 2011 Manual page was updated. [new](#)
- Mar 28, 2011 Data (64 spectra) from BMK, CAMS & PLUMC, China was added. [new](#)
- Apr 05, 2011 Data (141 spectra) from PFOS research group was added. [new](#)
- Mar 24, 2011 Document page was updated. [new](#)

Database Service

- Spectrum Search
- Quick Search
- Peak Search
- Substructure Search
- Advanced Search
- Spectral Browser
- Batch Service
- Browse Page
- Record Index

MassBank is financially supported from National BioScience Database Center, Japan Science and Technology Agency (2011-2013). The Mass Spectrometry Society of Japan officially supports MassBank. Please cite the article (DOI) when using MassBank.



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Comprehensive MetDBs

KEGG

KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions, and relations

KEGG PATHWAY BRITE MODULE KO GENOME GENES LIGAND DISEASE DRUG DBGET

Select profile: Enter keywords: Go Help

[New maps | Update history]

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

1. **Metabolism**
Global/overview, Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Other amino, Glycan, Cofactor/vitamin, Terpenoid/PC, Other secondary metabolite, Xenobiotics, Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

7. **Drug Development**

Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretation of higher-level systemic functions.

- Search Pathway - basic pathway mapping tool
- SearchColor Pathway - advanced pathway mapping tool

EcoCyc

EcoCyc: Encyclopedia of *E. coli* Genes and Metabolic Pathways

EcoCyc *E. coli* Database

EcoCyc is a scientific database for the bacterium *Escherichia coli* K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic pathways.

New to EcoCyc? Take the [guided tour](#) of the EcoCyc.org Web site, watch our [free online instructional videos](#), or read our 2013 article: ["EcoCyc: using model organism databases with systems biology"](#)

[EcoCyc User Guide >>](#)

Gene Expression Data Analysis
Multiple tools are available in this website for analysis of gene expression data.

[Learn More](#)

MetaboLights

MetaboLights - Metabolomics experiments and derived information

MetaboLights

MTBLS1 from University of Cambridge
NMR based metabolomics of Human Type 2 Diabetes urine samples

MetaboLights

Download

Pre-packaged ISACcreator download. To make it easy for new users, please download and just unzip our pre-packaged ISACcreator with plugin and configurations.

Submit a new study
Use this option if your study has not been submitted before.

Update an existing study
Use this option if you like to update a previously submitted study.

HumanCyc

HumanCyc: Encyclopedia of Human Genes and Metabolism

HumanCyc: Encyclopedia of Human Genes and Metabolism

HumanCyc provides an encyclopedic reference on human metabolic pathways. It provides a zoomable human metabolic map diagram, and it has been used to generate a steady-state quantitative model of human metabolism.

Getting Started

New to HumanCyc? Typical usage includes:

- Search for a gene or pathway using the Quick Search box at the upper right, or use the Search menu
- Analyze human metabolomics and gene-expression data

For more information on HumanCyc, see our article ["Computational prediction of human metabolic pathways from the complete human genome"](#) in *Genome Biology*.

[User Guide >>](#)

RouteSearch

RouteSearch: Search for Paths through the Metabolic Network

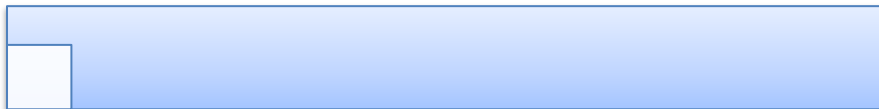
Search for lowest-cost paths through the metabolic network of the selected organism. Or, design lowest-cost pathways to novel compounds by adding reactions from MetaCyc.

[Learn More](#)



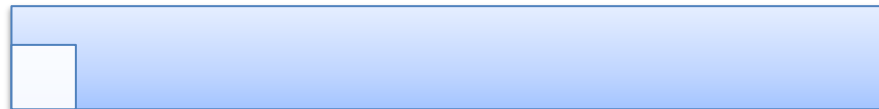
Naive vs forced approach

Forced: Find-by-Formula



- Needs a formula or a source of formula (database)
- Deeper in compounds fishing
- Higher number of false positive

Naive: Find-by-Molecular-Feature



- Needs an isotopic model (e.g. glycans, peptides, common organic molecules)
- Often requires a following identification step
- Lower number of false positives, higher number of false negative



From meaningless to meaningful data

Raw data

- Huge amount of information
- Redundancy? False positive?

dataset

- Data gathering
- Alignment and filtering

Interpret.

- Statistics
- Chemometrics & bioinformatics

Answer(s) to a biological problem?



1- Review compounds at a glance:

- Check presence/absence across treatments
- Preliminary investigate alignment and check for peak shape

2- Align and filter

3- Export for recursive analysis

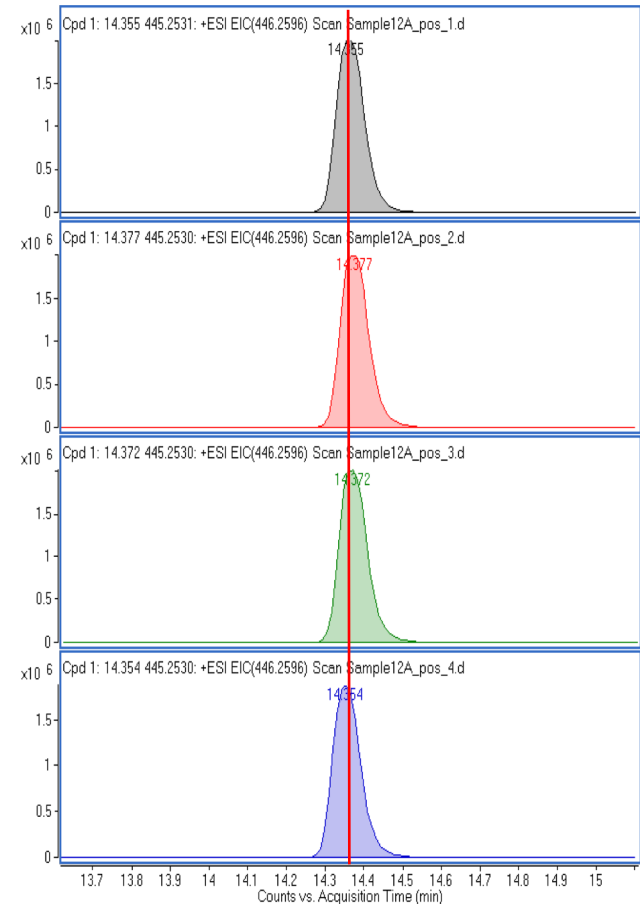
RT (med)	Missed	%RSD
1.064	9	
22.841	0	
31.757	0	
22.842	0	
23.085	0	
0.939	0	
23.083	0	
0.892	10	
18.406	0	
19.535	0	
5.219	10	
16.351	0	
31.759	0	
8.787	10	
27.502	0	
1.595	10	
27.473	0	
24.637	0	
1.089	8	
28.514	0	
27.443	0	

Tgt	Score (Tgt)	Score (M)
no EIC peaks		
no EIC peaks		
	99.93	
	88.01	
	88.05	
	88.03	
	88.06	
	88.04	
	99.94	
	88.06	
	88.07	





- 1. Alignment (optionally using internal standards) - Retention time tolerance**
- 2. Filter by mass tolerance**
- 3. Remove irreproducible compounds – Filter by Frequency**
- 4. Remove weak compounds – Filter by Abundance**
- 5. Remove highly variable compounds – Filter by Sample Variability**



SPECTRAL LIBRARIES BASED METABOLOMICS

(GC-MS)



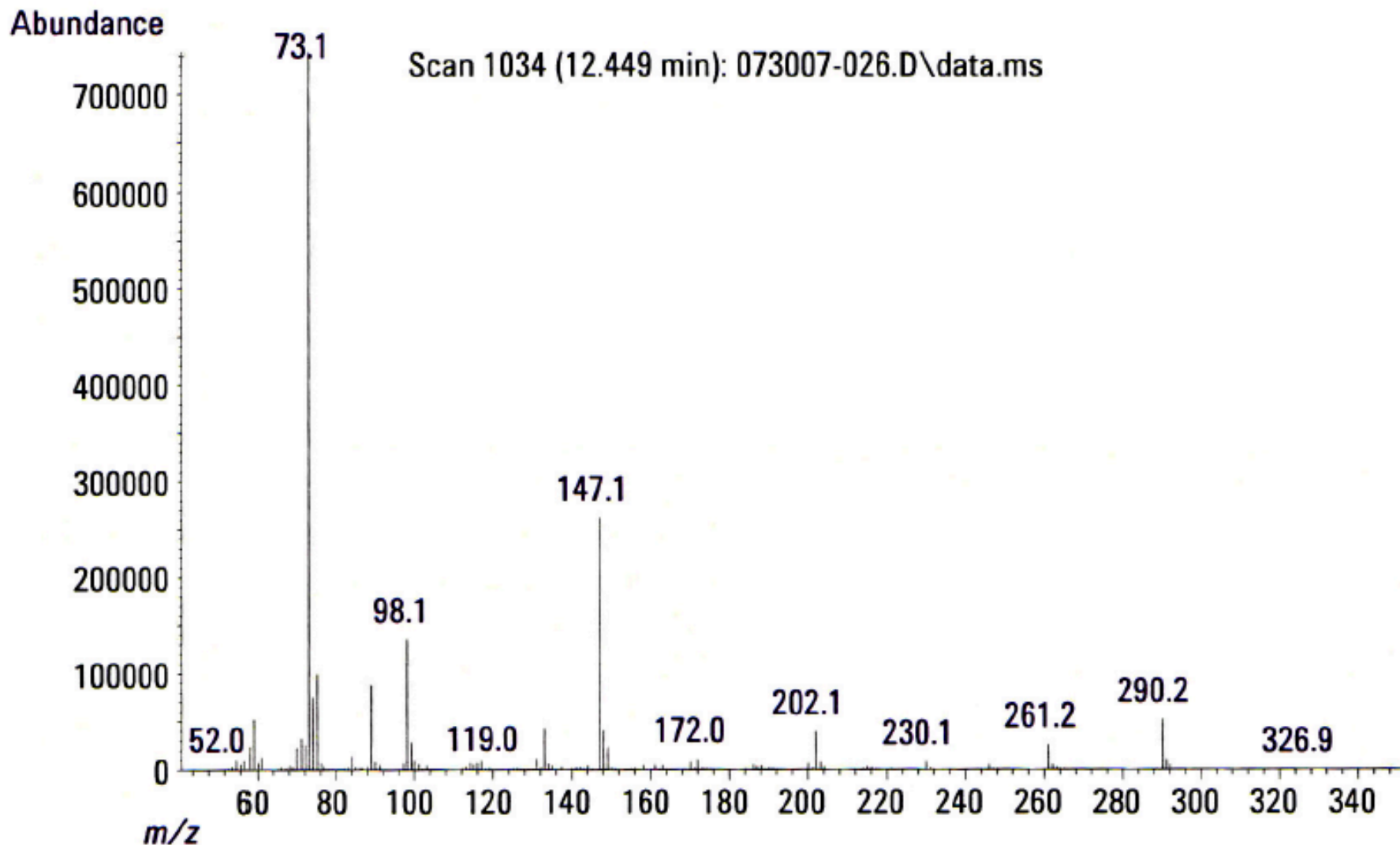
ID by GC-MS

- GC-MS is often best for identification of amino acids, organic acids, sugars, fatty acids and molecules with $MW < 500$
- GC has higher resolution and better reproducibility than LC
- EI-MS is more standardized than soft ionization methods, so EI spectra are more comparable.



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

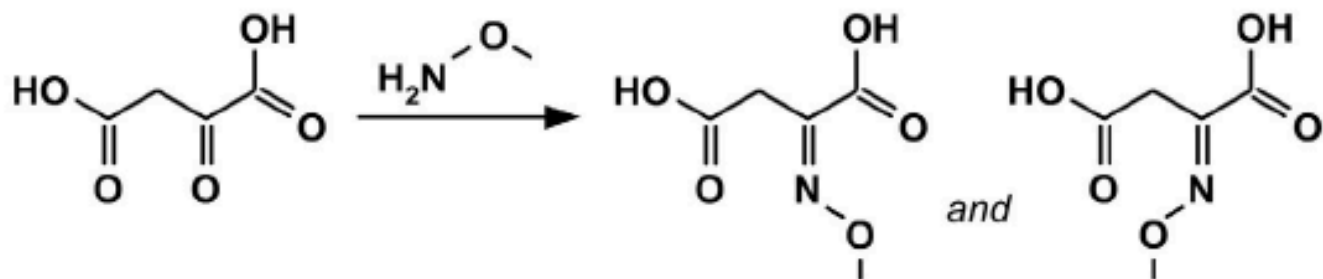
Electron impact MS spectra



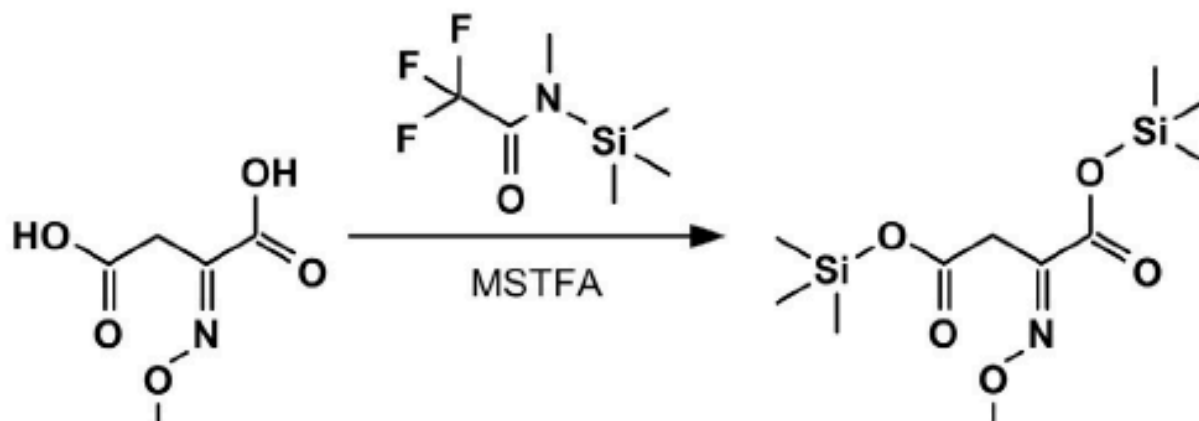


Need for derivatization

Step I. Methoxyamination to protect carbonyl groups



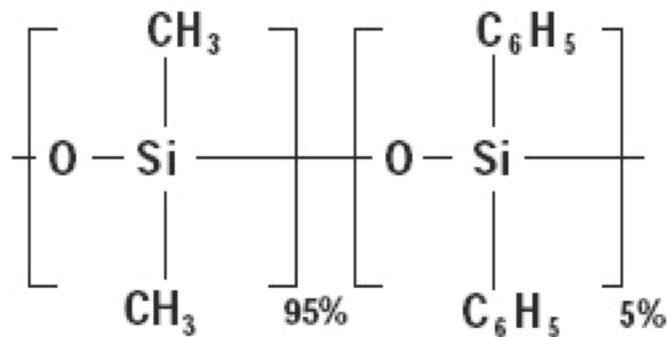
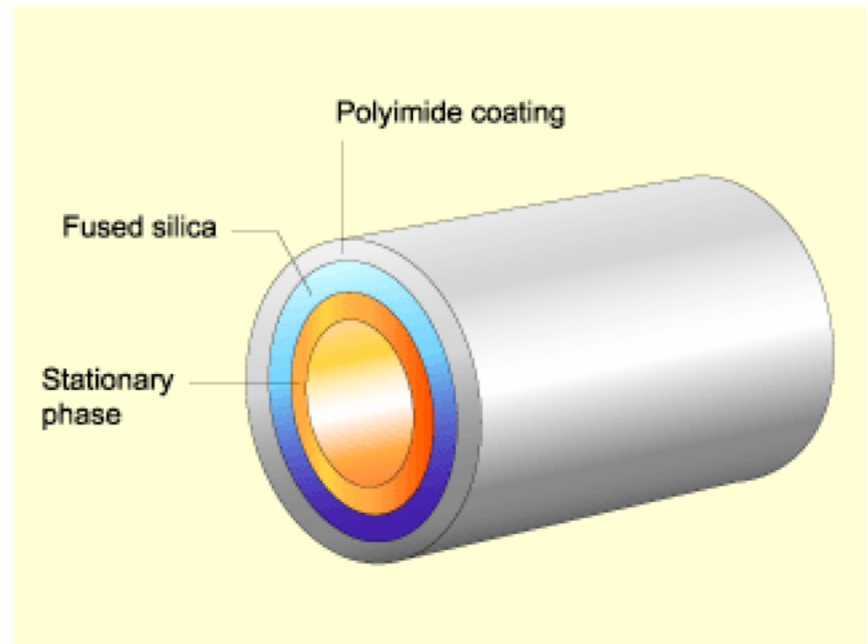
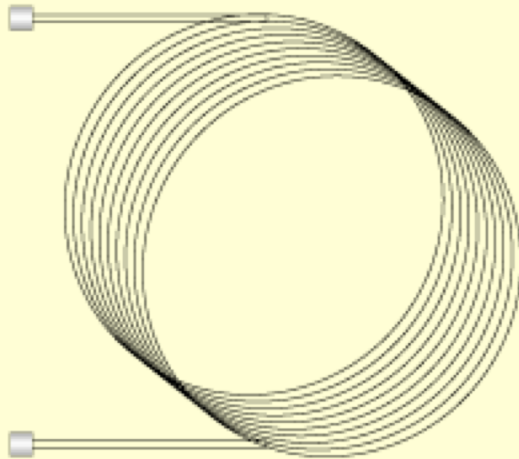
Step II. Trimethylsilylation to decrease the boiling point





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

GC-Columns



Polysiloxane



Retention Time/Index

- Retention time (RT) is the time taken by an analyte to pass through a column
- RT is affected by compound, column (dimensions and stationary phase), flow rate, pressure, carrier, temp.
- Comparing RT from a standard sample to an unknown allows compound ID
- Retention index (RI) is the retention time normalized to the retention times of adjacently eluting n-alkanes



RT locking

Agilent MassHunter Qualitative Analysis B.06.00 - Default.m

File Edit View Find Identify Spectra Chromatograms Method Wizards Actions Configuration Tools Help

Data Navigator

Sort by Data File

- ✓ Cpd 11: 10.448
- ✓ Cpd 12: [C10] Methyl Caprate [10.647]
- ✓ Cpd 13: 10.733
- ✓ Cpd 14: 11.325
- ✓ Cpd 15: 12.017
- ✓ Cpd 16: [C12] Methyl Laurate [13.250]
- ✓ Cpd 17: [C14] Methyl Myristate [15.597]
- ✓ Cpd 18: [RTL] Myristic Acid d27 [16.727]
- ✓ Cpd 19: 16.723
- ✓ Cpd 20: [C16] Methyl Palmitate [17.723]
- ✓ Cpd 21: [C18] Methyl Stearate [19.663]
- ✓ Cpd 22: [C20] Methyl Eicosanoate [21.441]
- ✓ Cpd 23: [C22] Methyl Docosanoate [23.082]
- ✓ Cpd 24: [8343] dioctyl phthalate [23.163]
- ✓ Cpd 25: [C24] Methyl Linocerate [24.603]
- ✓ Cpd 26: [C26] Methyl Hexacosanoate [26.023]
- ✓ Cpd 27: [C28] Methyl Octacosanoate [27.349]
- ✓ Cpd 28: [C30] Methyl Triacosanoate [28.723]
- ✓ Cpd 29: 29.357
- ✓ Cpd 30: 29.725

Compound List

Automatically Show Columns

Label	Show/Hide	Cpd	Name	RT /	CAS	RT (DB)	Score	Formula	St
Cpd 16: [C12] Methyl L...	✓	16	[C12] Methyl Laurate [13.250]	13.243	111-32-3	1200	88.85	C13H26O2	
Best	▽	ID Source	Name	Formula	Score	RT Diff	Score (Lib)	RT (DB)	Library
○	LibSearch	[C12] Methyl Laurate [13.250]	C13H26O2	88.85	-1186.757	88.85	1200	Fiehr	
○	LibSearch	[16213400] Cholesterol D6 [27.5...	C27H40O6...	55.97	-2809.297	55.97	2822.54	Fiehr	
○	LibSearch	[C9] Methyl Pelargonate [9.248]	C10H20O2	54.74	-886.757	54.74	900	Fiehr	
○	LibSearch	[C10] Methyl Caprate [10.647]	C11H22O2	54.33	-986.757	54.33	1000	Fiehr	
○	LibSearch	[C8] Methyl Caprylate [7.812]	C9H18O2	51.74	-786.757	51.74	800	Fiehr	

Label	Show/Hide	Cpd	Name	RT /	CAS	RT (DB)	Score	Formula	St
Cpd 17: [C14] Methyl L...	✓	17	[C14] Methyl Myristate [15.597]	15.596	124-10-7	1400	86.79	C15H30O2	
Cpd 18: [RTL] Myristi...	✓	18	[RTL] Myristic Acid d27 [16.727]	16.709	60658-4-	1503	57.65	C15H27O2	
Cpd 19: 16.723	✓	19		16.723					
Cpd 20: [C16] Methyl...	✓	20	[C16] Methyl Palmitate [17.723]	17.728	112-39-0	1600	85.09	C17H34O2	

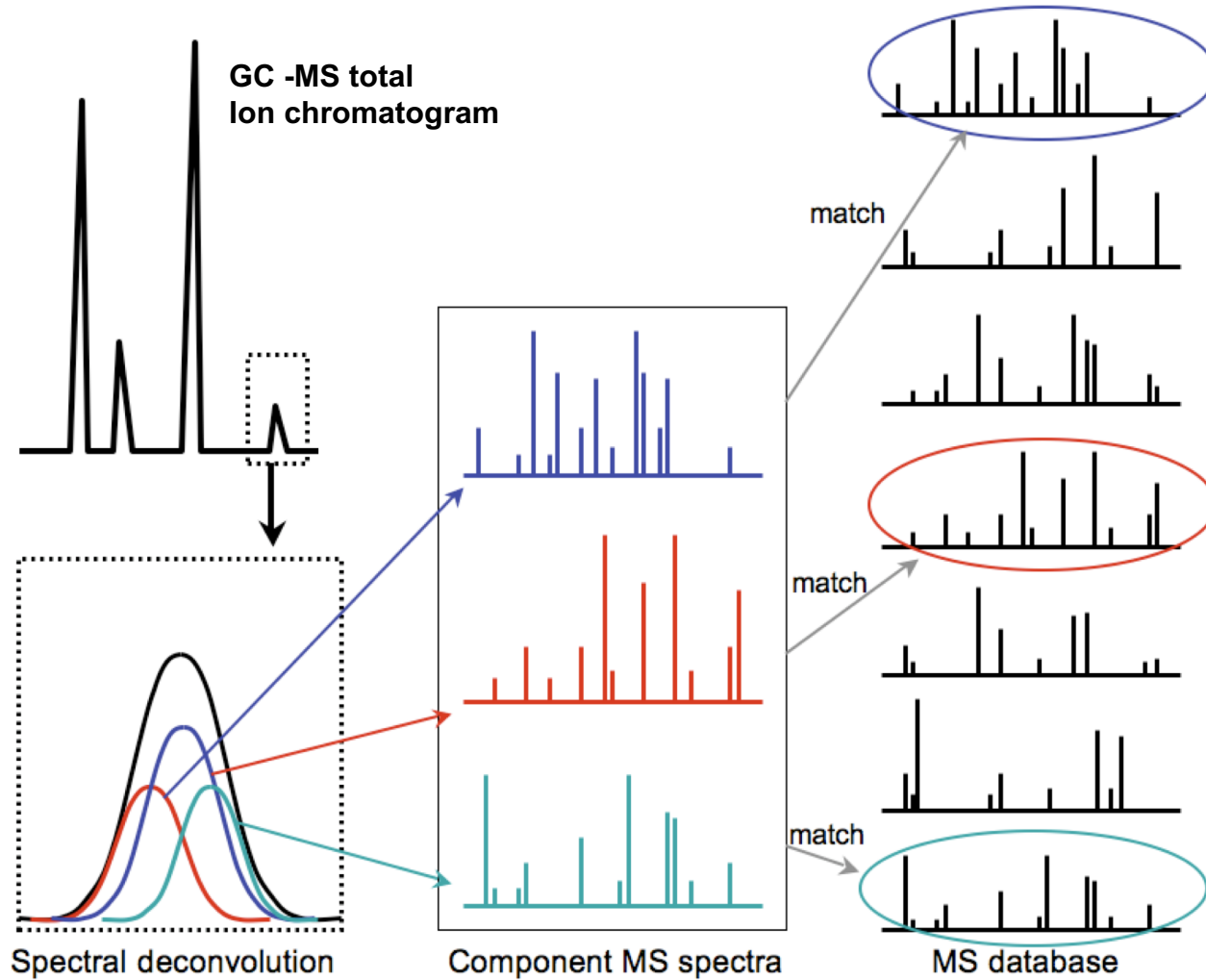
Chromatogram Results

Cpd 16: [C12] Methyl Laurate [13.250] + ECC Scan RI-CALIBRATION.D



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

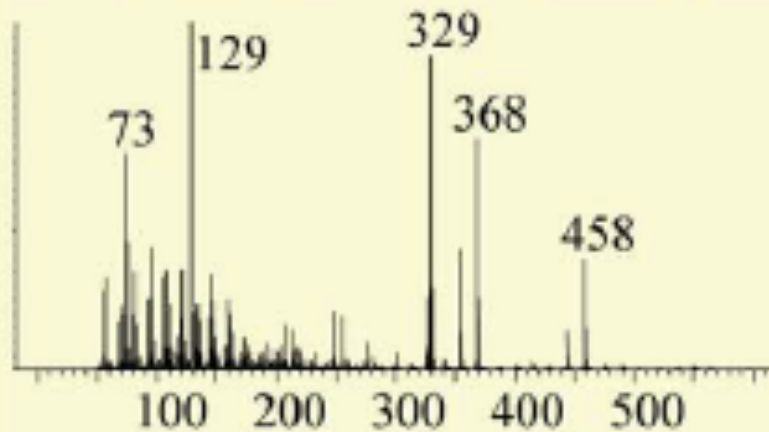
Metabolite ID by GC-MS



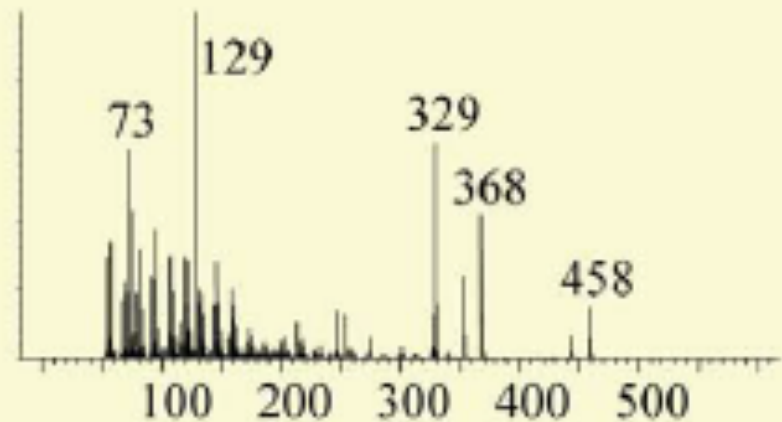


Spectral comparison

Scan 3448 (27.549 min): 042508-08.D



#1012: [304] cholesterol [27.555]



Scan 3448 (27.549 min): 042508-08.D\data.ms blood plasma488755

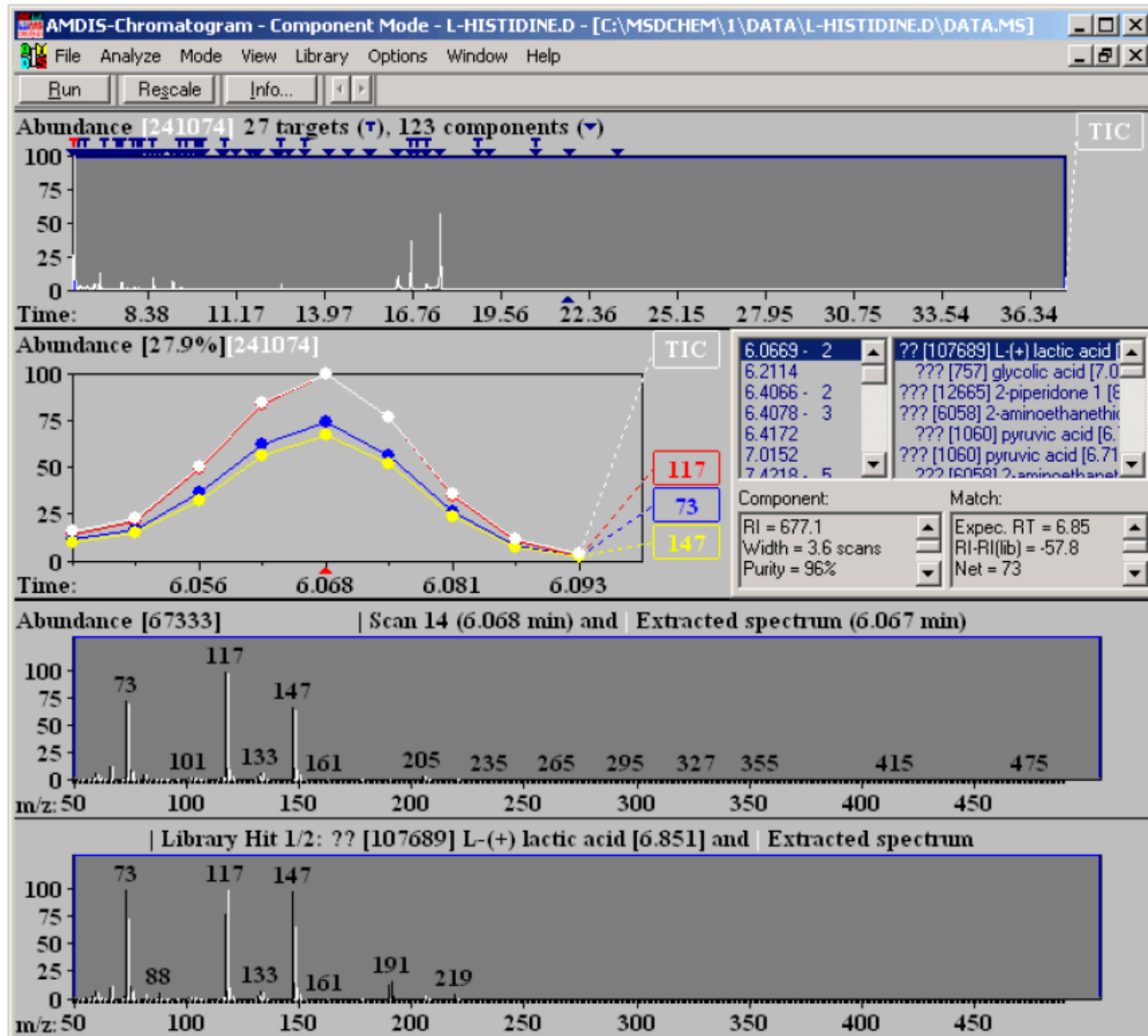
Name

MW Formula Qual CAS#

1	[304] cholesterol	[27.555]	9999	C ₂₇ H ₄₆ O	91	000057-88-5
2	[637775] 3,5-dimethoxy-4-hydroxycinnamic acid	[20.65]	9999	C ₁₁ H ₁₂ O ₅	7	000530-59-6



Library search



DATA REDUCTION and INTERPRETATION



Data analysis

- Data collected represented in a matrix

Variables going across in different columns

Objects going down in different rows	X-var 1	X-var 2	X-var 3	Y-var 1	Y-var 2
Sample 1					
Sample 2...					

A propositional approach to describing and using metabolomics data (the x-data) for analyzing complex systems. These may have other specific properties (the y-data) which one may also wish to 'explain' in terms of the x-data.

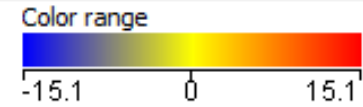
- Chemometric Approach
 - Principle Component Analysis (PCA)
 - Partial Least-Squares (PLS) Method
 - Orthogonal PLS (OPLS)
- Targeted Profiling



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Unsupervised cluster analysis

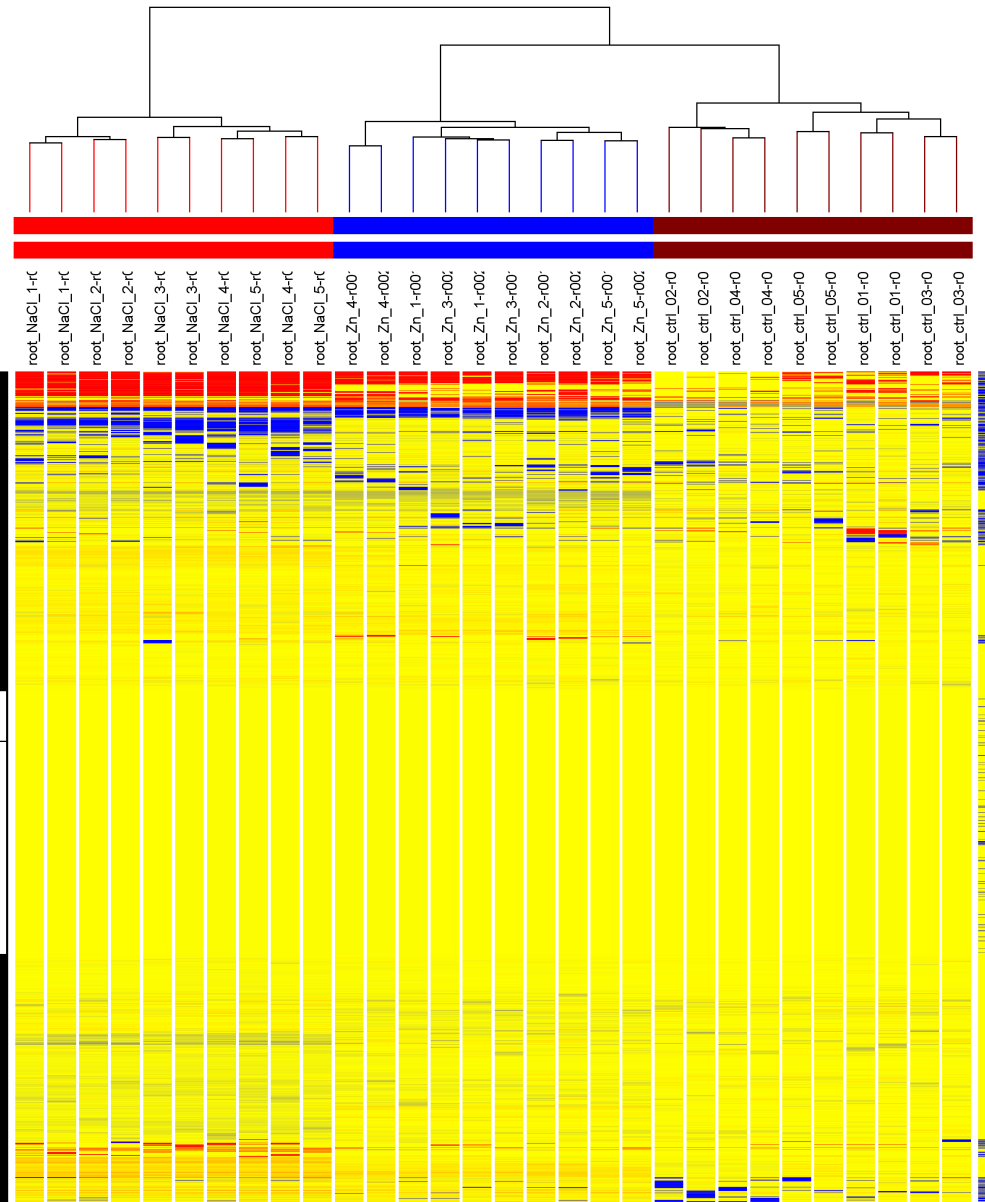
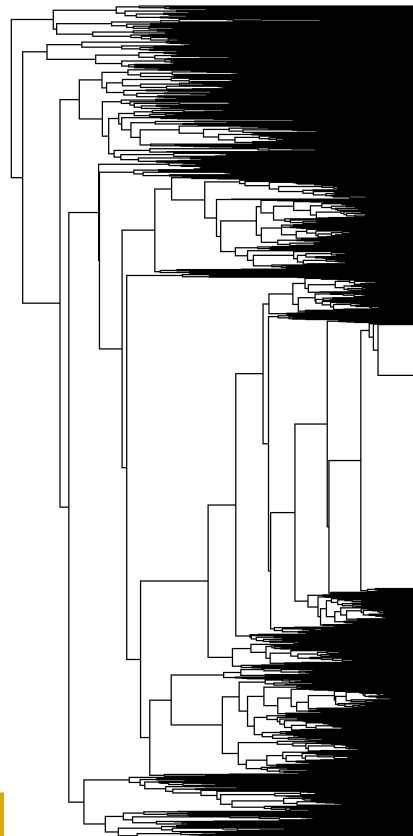
Legend - Hierarchical Combined Tree



Condition

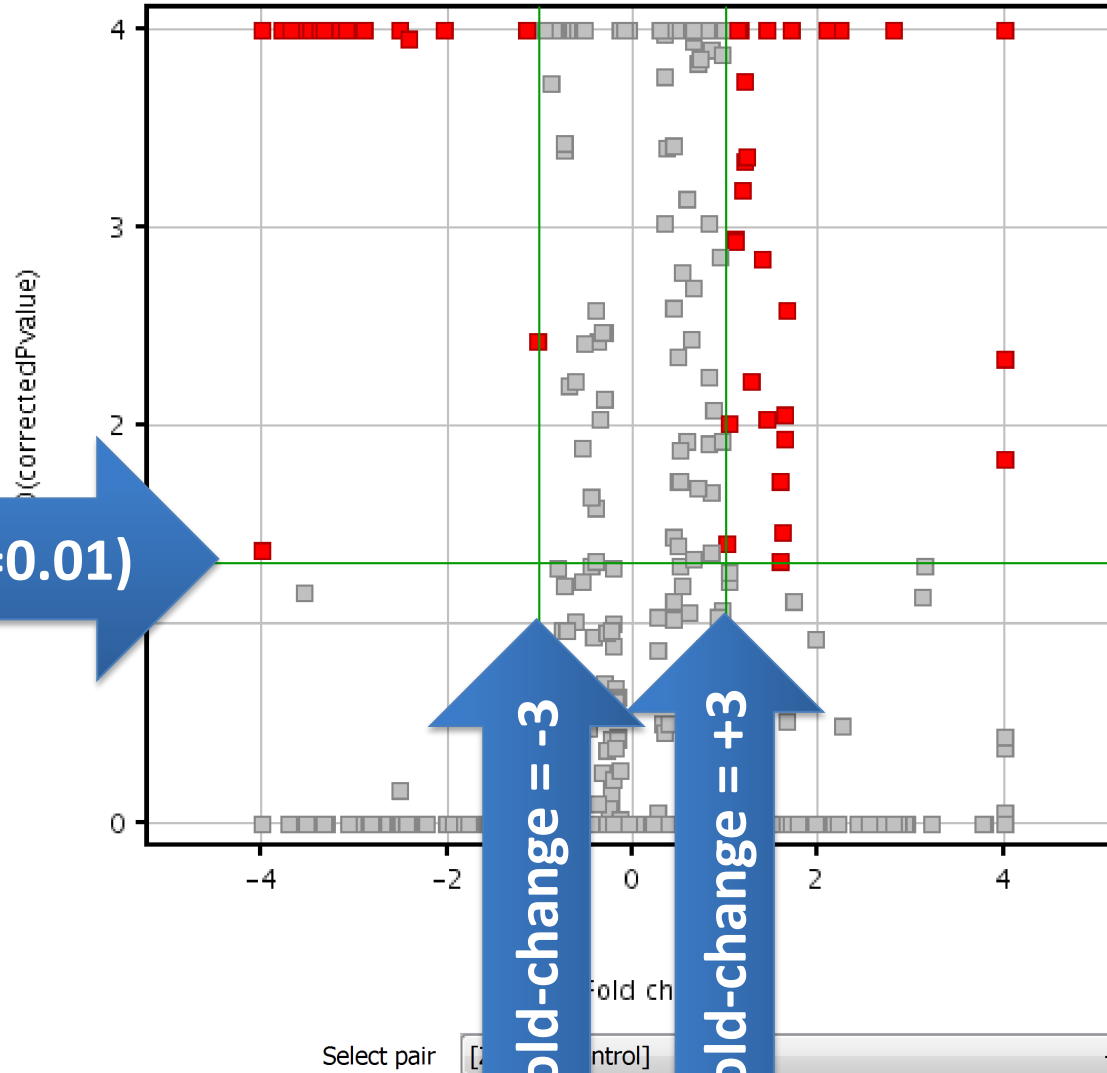
- [NaCl]
- [Zn]
- [control]

Condition
treatment



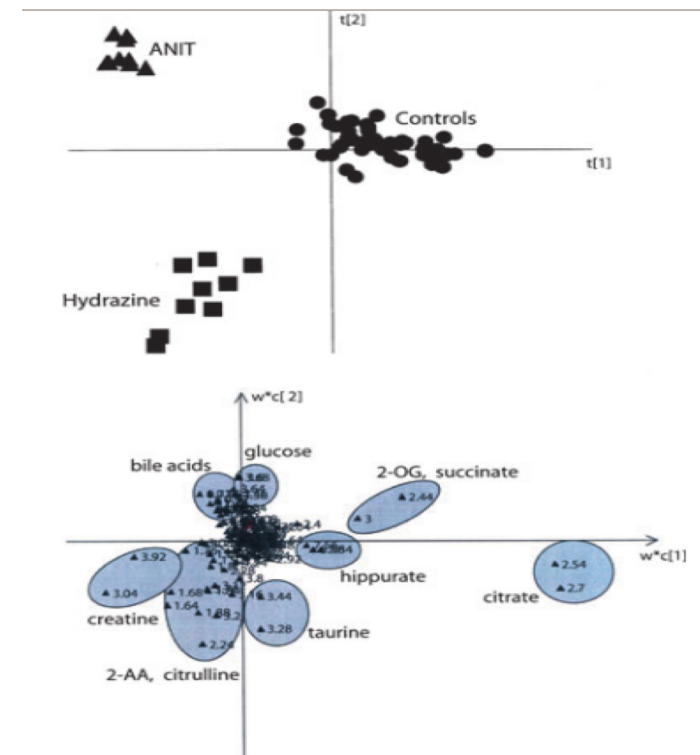


Volcano analysis



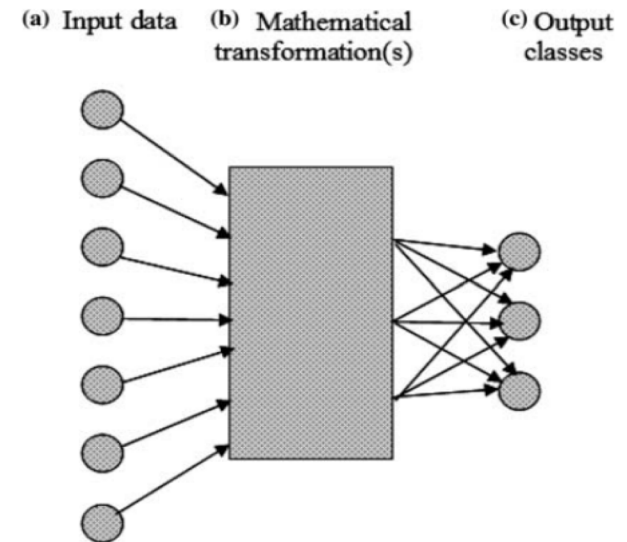


- Unsupervised
- Multivariate analysis based on projection methods
- Extract and display the systematic variation in the data
- Each Principle Component (PC) is a linear combination of the original data parameters
- PCs Orthogonal to each other
- Conversion of original data leads to two matrices, known as scores and loadings
- The scores(T) represent a low-dimensional plane that closely approximates X. Linear combinations of the original variables. Each point represents a single sample spectrum.
- A loading plot/scatter plot(P) shows the influence (weight) of the individual X-variables in the model. Each point represents a different spectral intensity.
- The part of X that is not explained by the model forms the residuals(E)
- $X = TP^T = t_1p_1^T + t_2p_2^T + \dots + E$





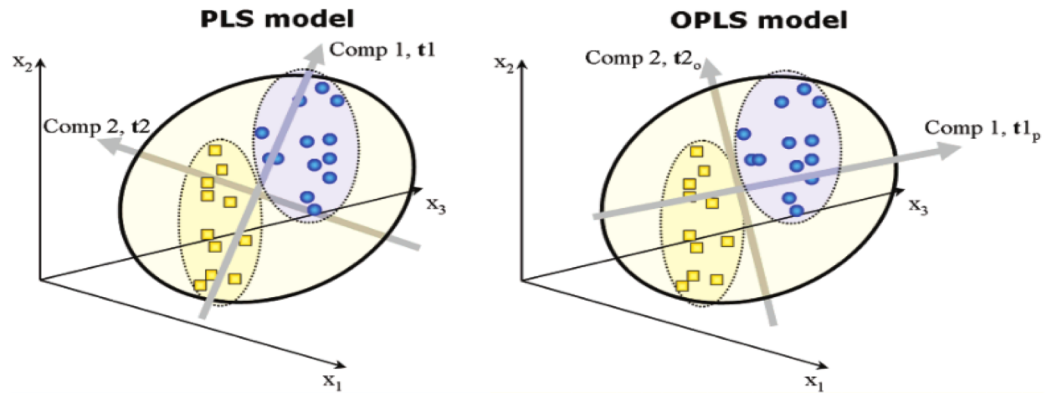
- Supervised learning method.
- Principles that of PCA. But in PLS, a second piece of information is used, namely, the labeled set of class identities.
- Two data tables considered namely X (input data from samples) and Y (containing qualitative values, such as class belonging, treatment of samples)
- The PLS algorithm maximizes the covariance between the X variables and the Y variables



The class assignment problem. The inputs can be considered, and are referred to, as the “explanatory variables” or “x-data” whereas the functional or the other classes of interest, which are still variables associated with the samples, are referred to as “dependent variables” or “y-data” and are to be obtained as the outputs.



OPLS



A geometrical illustration of the difference between the PLS-DA and OPLS-DA models. In the left panel, the PLS components cannot separate the between-class variation from the within-class variation, and the resulting PLS component loadings mixes both types of variations. In the right panel, the OPLS components are able to separate these two different variations. Component 1 (t_{1p}) is the predictive component and displays the between-class ([blue circles], [yellow squares]) variation of the samples. The corresponding loading profile can be used for identifying variables important for the class separation. Component 2 (t_{2o}) is the Y-orthogonal component and models the within group (within-class) variation.

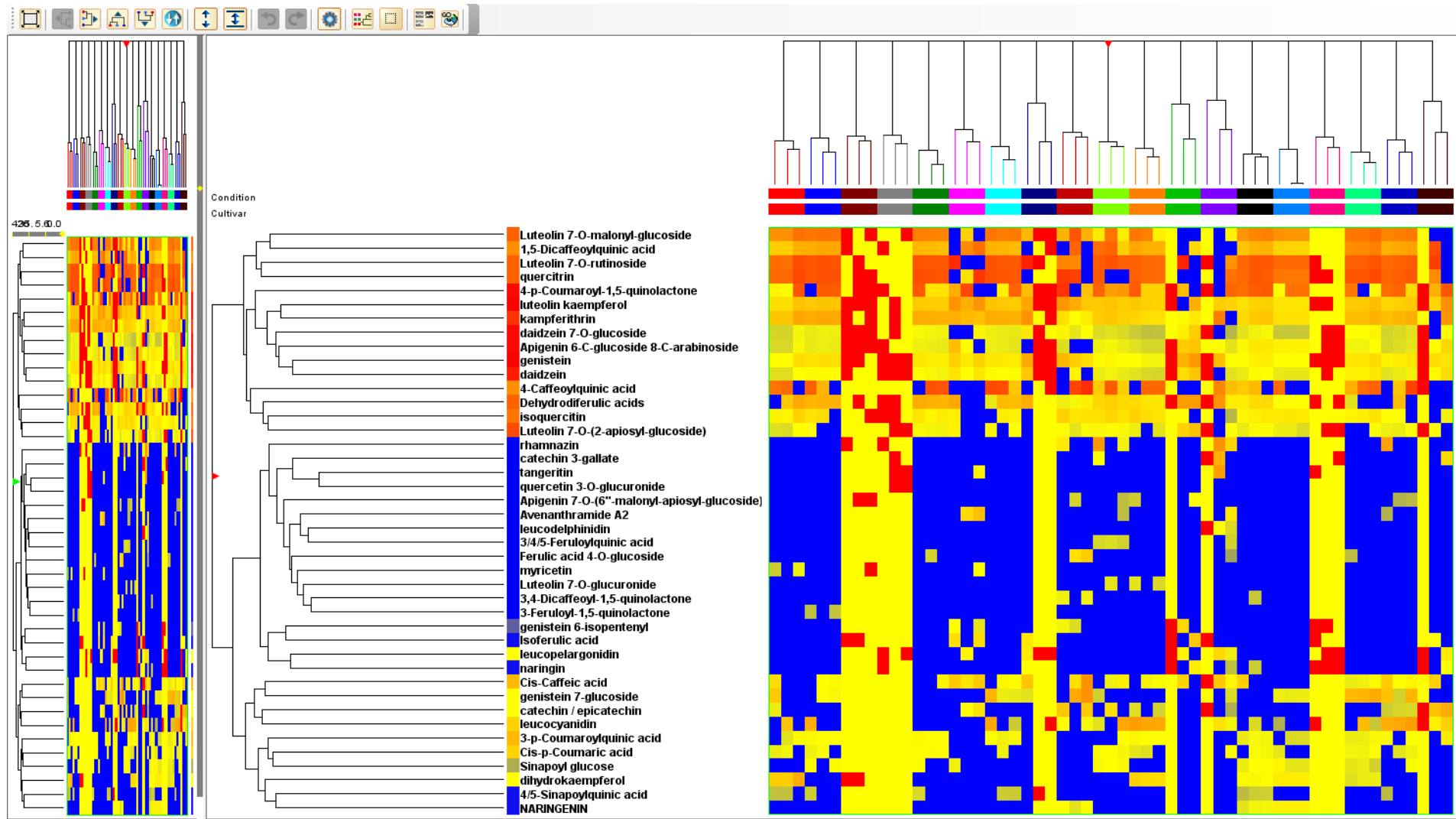
- **OPLS method is a recent modification of the PLS method to help overcome pitfalls**
- **Main idea to separate systematic variation in X into two parts, one linearly related to Y and one unrelated (orthogonal).**
- **Comprises two modeled variations, the Y-predictive ($T_p P_p^T$) and the Y-orthogonal ($T_o P_o^T$) components.**
- **Only Y-predictive variation used for modeling of Y.**
- **OPLS-DA compared to PLS-DA**

APPLICATIONS



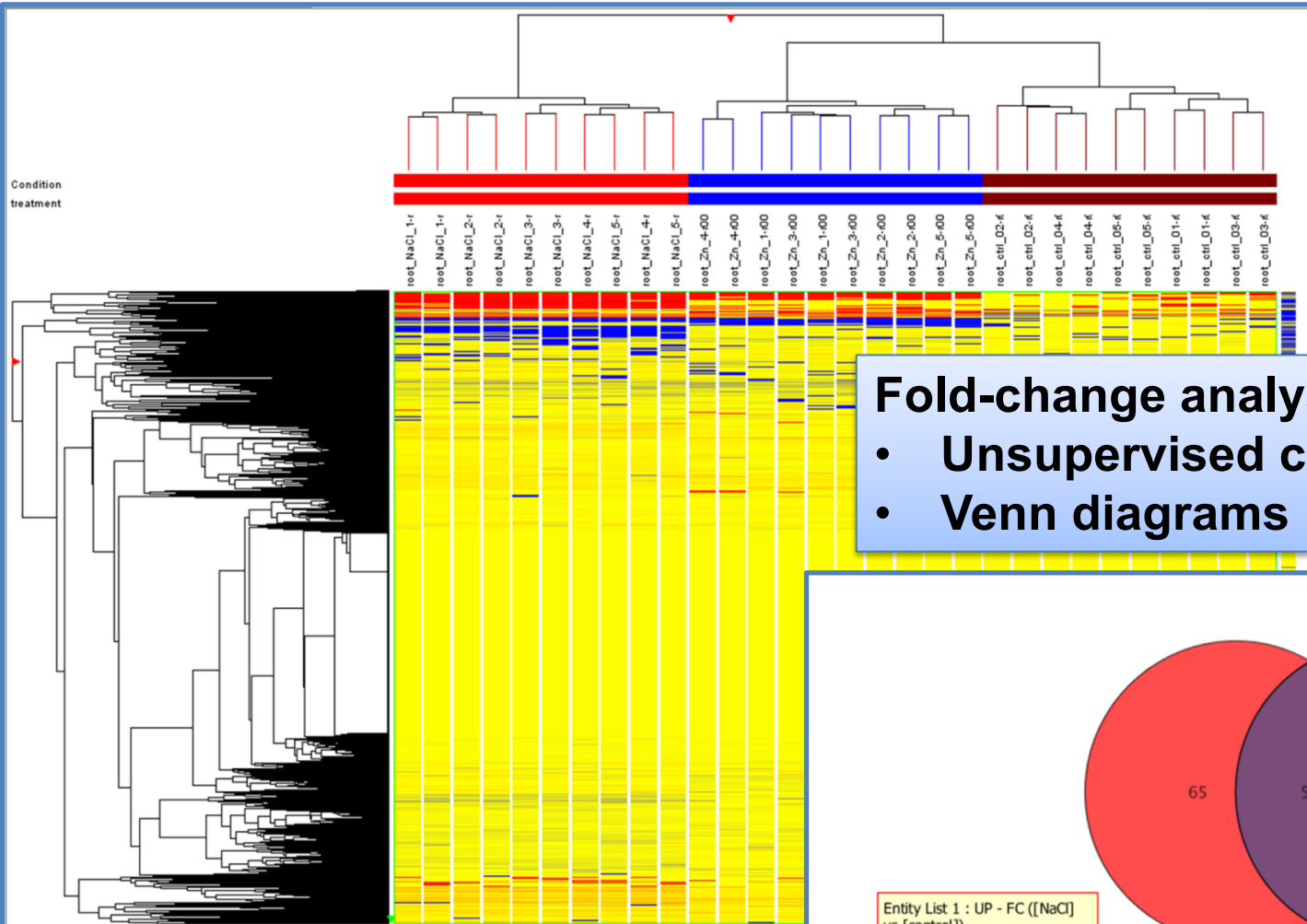
UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Cluster analysis varietà carciofo (phenolic profiling)



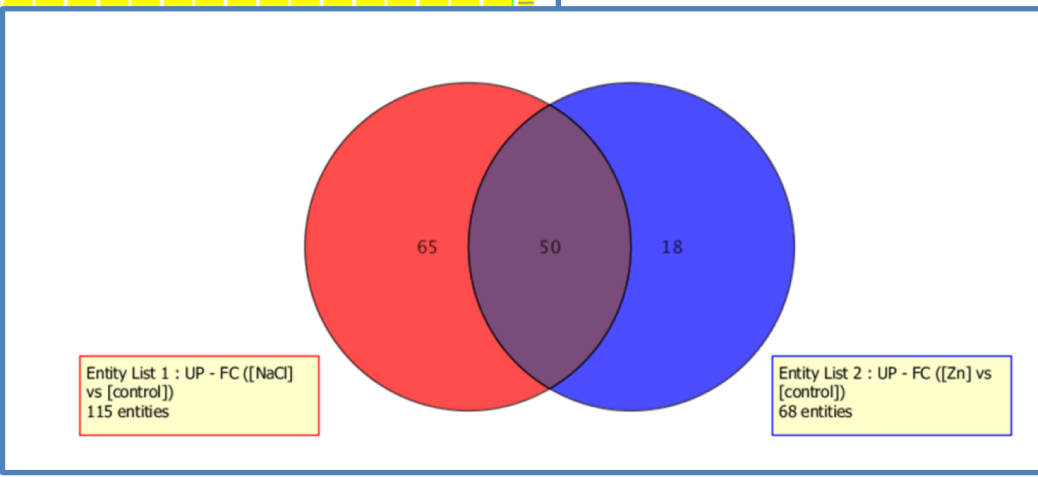


Effect of abiotic stress on plant metabolism (Zn vs salinity vs control)



Fold-change analysis let to perform:

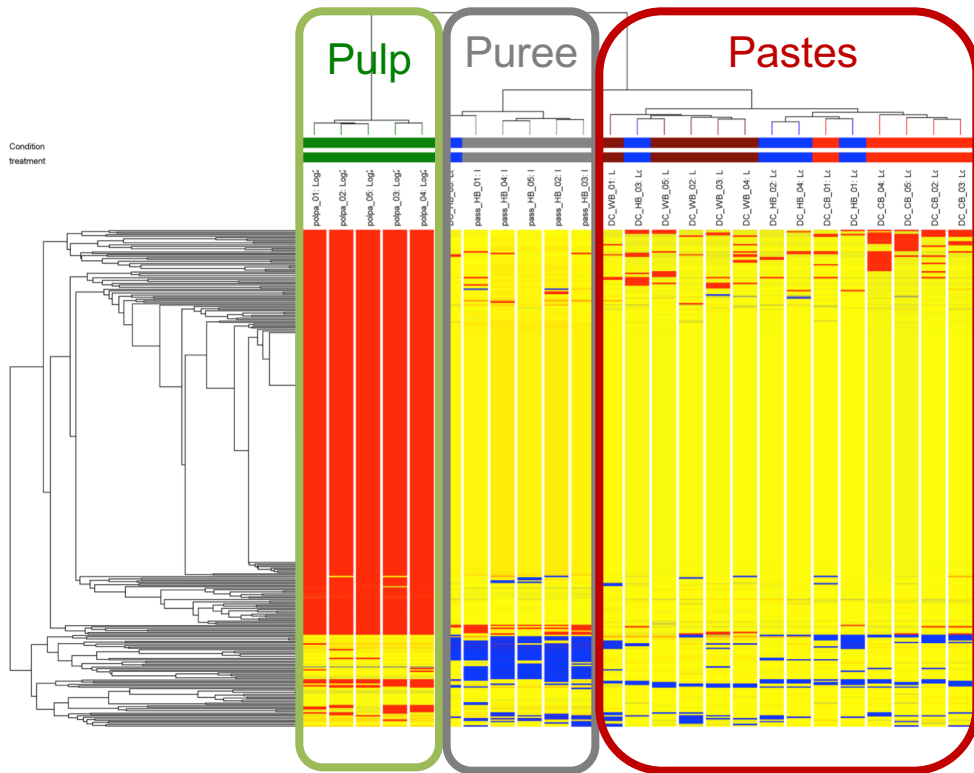
- Unsupervised cluster analysis
- Venn diagrams



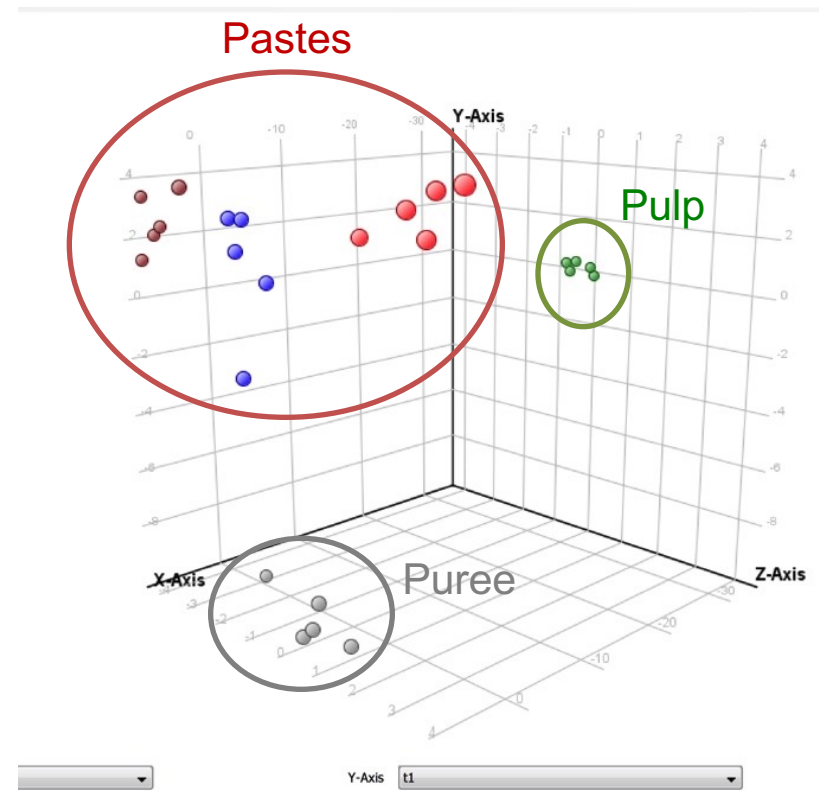


Discrimination of different processing technologies

Unsupervised cluster analysis

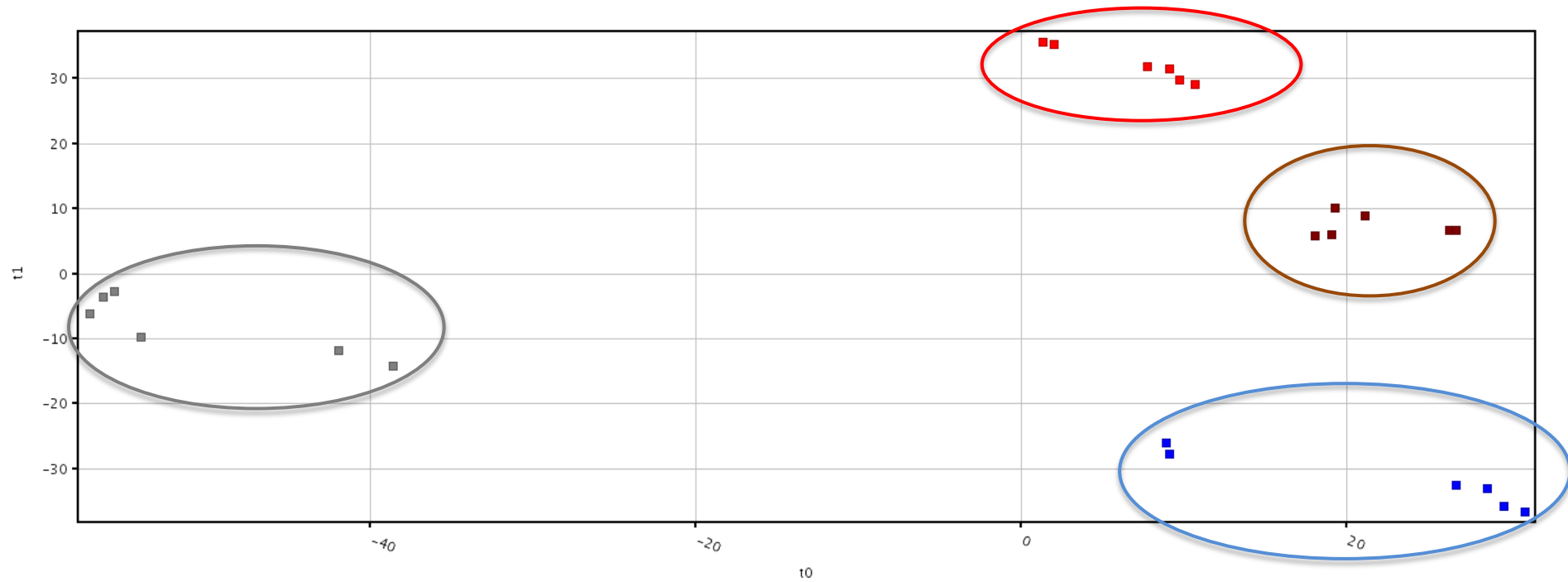


Partial Least Square Discriminant Analysis (PLS-DA)





Rintracciabilità dei processi fermentativi (fave cacao)



Legend - Scores

Color by treatment (Non-averaged)

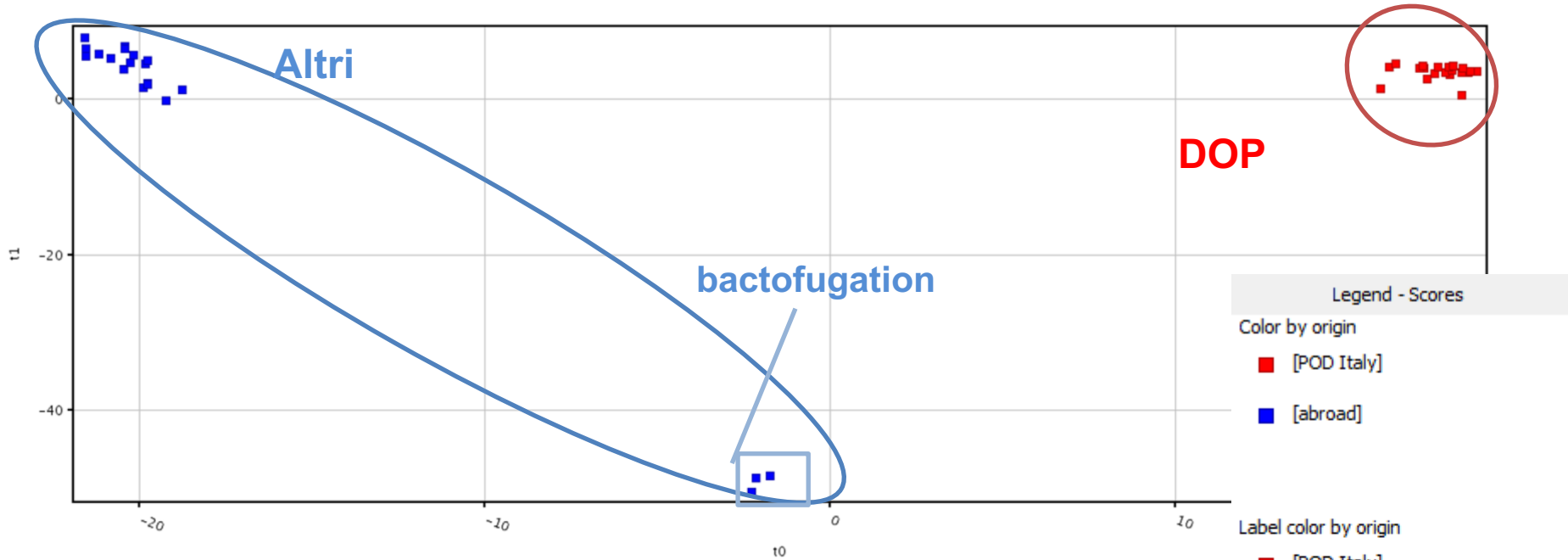
- [T1]
- [T2]
- [T3]
- [control]

X-Axis

Y-Axis



Tutela produzioni – Grana DOP

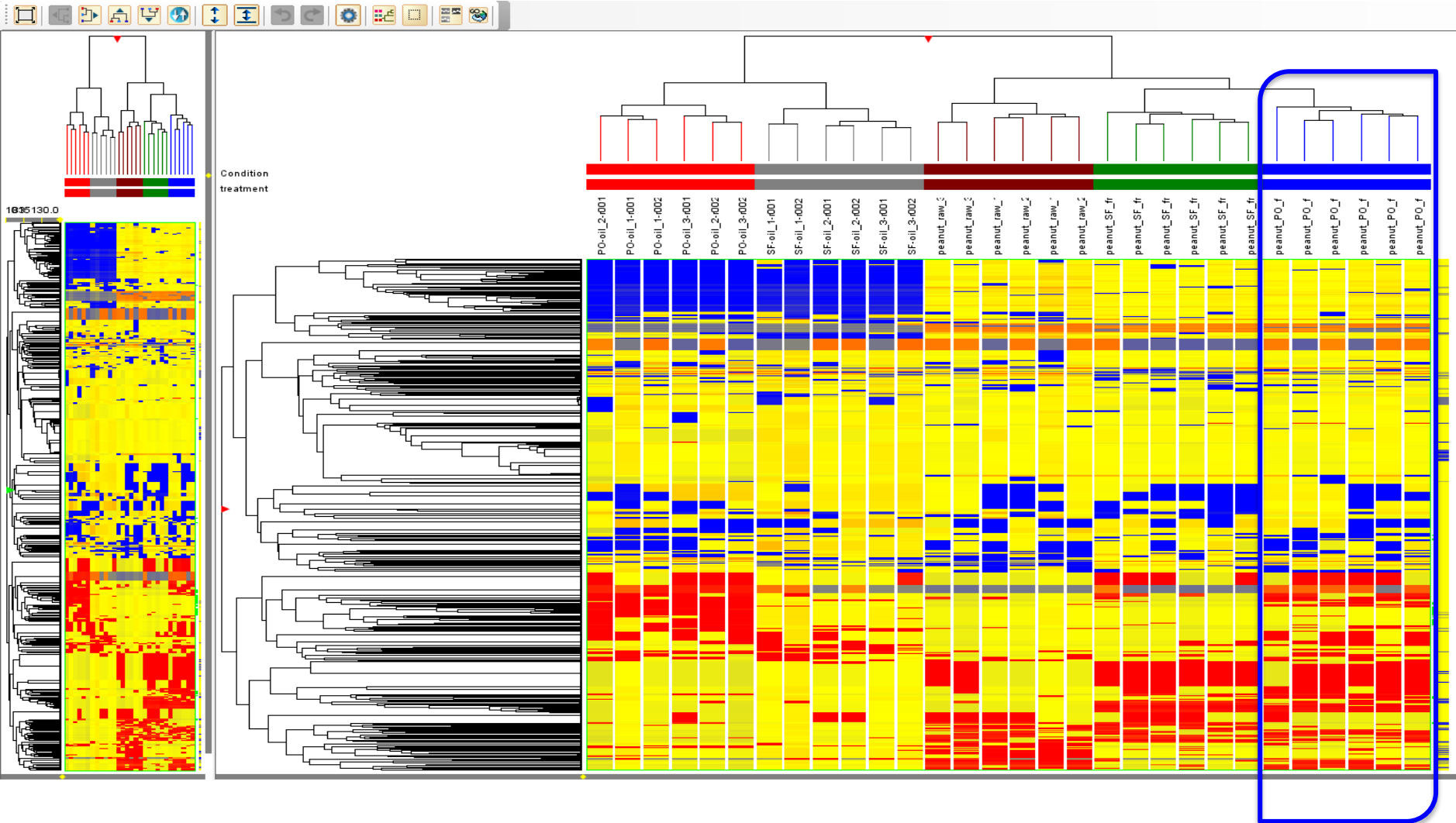


X-Axis: t0
Y-Axis: t1

[abroad] (Predicted)	
20	0
0	20



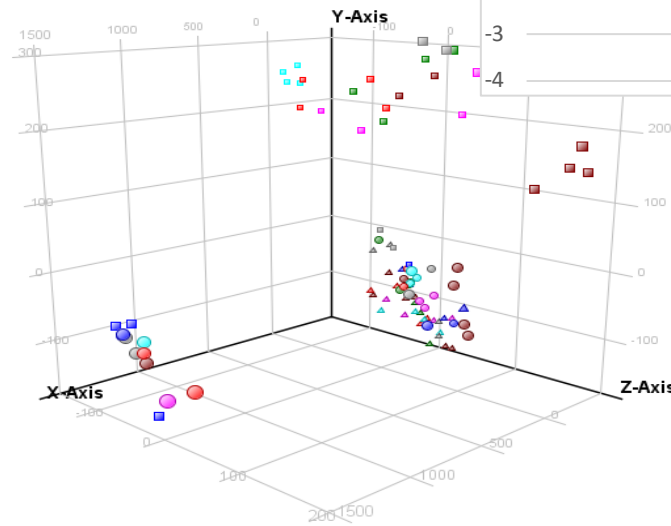
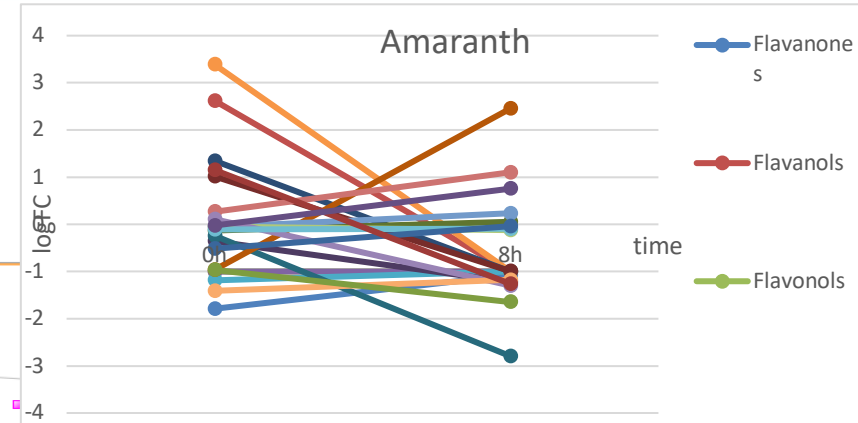
Utilizzo olio di palma





UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Fermentazione intestinale farine gluten free

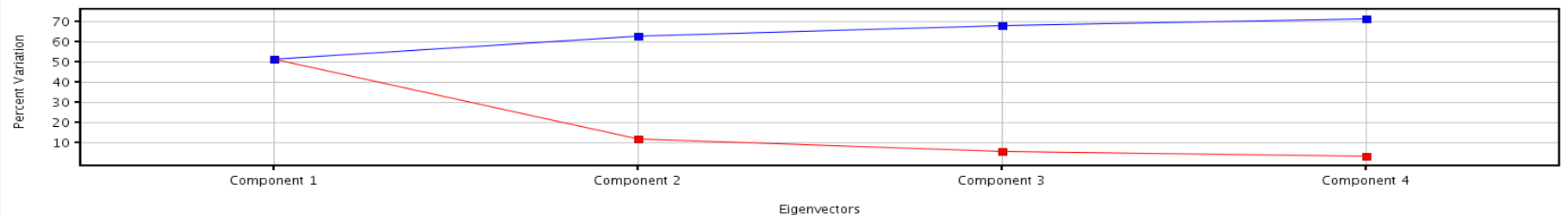


X-Axis Component 1 (51.34%)

Y-Axis Component 2 (11.79%)

Z-Axis Component 3 (5.35%)

PCA Loadings 3D PCA Loadings 3D PCA Scores C-C Plot PCA Scores





- “Systematic bond-breaking” approach [Hill and Mortishire-Smith, 2005] to correlate tandem MS data with chemical structures
- Scores are generated from each product ion, the mass accuracy of the fragments, and the overall percentage of ions intensity being plausibly explained with sub-structures

Agilent MassHunter Molecular Structure Correlator 8.05.00 - white-M+H.ces10

Compound formula

ID	Formula	Isomers	Total Gps	d[M]ppm	d[M]ppm	Product
1	C20H34O8	71	56	1.7	1.7	
2	C19H30O8	5	4	1.7	1.7	
3	C19H30O7	9	3	3.3	3.3	
4	C20H34O4	649	313	5.0	5.0	
5	C19H30O5	66	60	4.0	3.4	
6	C19H30O6	9	0	3.4	3.4	
7	C19H30O5	2	2	3.8	3.8	

Fragment formulas for C20H34O8

m/z	Intensity	nom. inters	formula	d[M]ppm
111.068	52.16	12.49	C8H10O3	7.8
129.0179	87.16	20.87	C8H10O4	2.6
129.1000	23.13	5.54		
139.0200	186.88	44.75	C8H10O4	4.2
157.0101	349.55	83.70	C8H10O5	0.3
185.0804	417.62	100.00	C9H12O4	2.4
206.0305	17.41	4.17		
213.0762	214.41	51.34	C10H13O5	-2.1
217.0338	183.86	44.03	C8H10O7	2.2
259.1527	312.02	74.71	C13H20O5	5.0
269.1372	91.25	21.86	C14H21O5	4.3
273.0903	112.89	27.81	C12H17O7	-0.1
329.1599	148.82	35.64	C18H29O7	-1.3
343.2102	16.44	3.94	C18H29O6	3.8
361.2223	119.38	28.89	C18H29O7	-0.6

12 mixed compound formulas

Mass	Intensity	Weight%	No. of cand.	Best score
111.068	52.16	0.0	9	63.1
129.0179	87.16	0.1	14	96.3
129.1000	23.13	0.0	0	0.0
139.0200	186.88	0.2	5	86.4
157.0101	349.55	0.7	10	94.2
185.0804	417.62	2.3	22	97.1
206.0305	17.41	0.2	0	0.0
213.0762	214.41	2.8	24	93.6
217.0338	183.86	2.7	4	97.8
259.1527	312.02	10.1	19	93.2
269.1372	91.25	4.8	12	90.9
273.0903	112.85	6.5	9	85.5
329.1599	148.82	26.2	6	98.1
343.2102	16.44	3.7	6	94.5

Structure Search

Parameters: 12/774

ChemSpider (Web)

Sort by: Score

Structure: C20H34O8

Standard IChKey: KCZLLKLYDRIWELD-UHFFFAFYSAN

Score: 97.11

ChemSpider: 2283362

Structure: C20H34O8

Standard IChKey: GZCLKYGREVAFR-UHFFFAFYSAN

Score: 97.11

ChemSpider: 629

Structure: C20H34O8

Standard IChKey: FESZQJRWVWYNE-UHFFFAFYSAN

Score: 96.78

ChemSpider: 828526

Structure: C20H34O8

Standard IChKey: PHRBDLZBFJDOT-MVVOOKESAN

Score: 84.63

Penalty=5.5 dM=7 Appm Score=83.1

Penalty=5.5 dM=7 Appm Score=83.1

Penalty=5.5 dM=7 Appm Score=83.1

Penalty=5.5 dM=7 Appm Score=83.1



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Structure elucidation & metabolite ID

Structural elucidation to discriminate isobaric compounds

Compound formula: **C₁₁H₁₆N₂O₂**

Structure Search
Parameters Compatibles/Total: 2/2 Add Structure
IPU_&_metab Go Sort by Score

Fragments of structure #1 -- elucidated: 53 Display F

	Mass	Intensity	Weight(%)	No. of c
1	94.0644	3442.63		1.8
2	151.0863	2030.22		18.7
3	134.0958	610.53		2.7
4	89.0590	594.81		0.2
5	195.1221	513.65		21.9
6	117.0892	468.73		0.9
7	152.1066	361.81		3.5
8	99.0433	354.10		0.3

1-OH-MDIPU

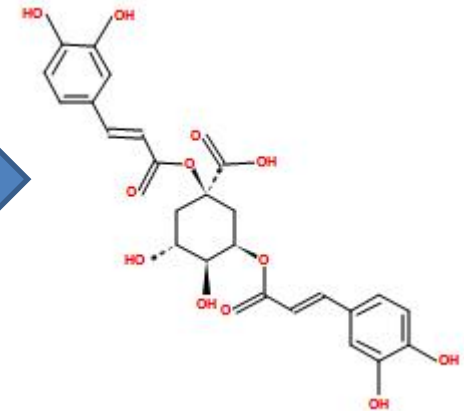
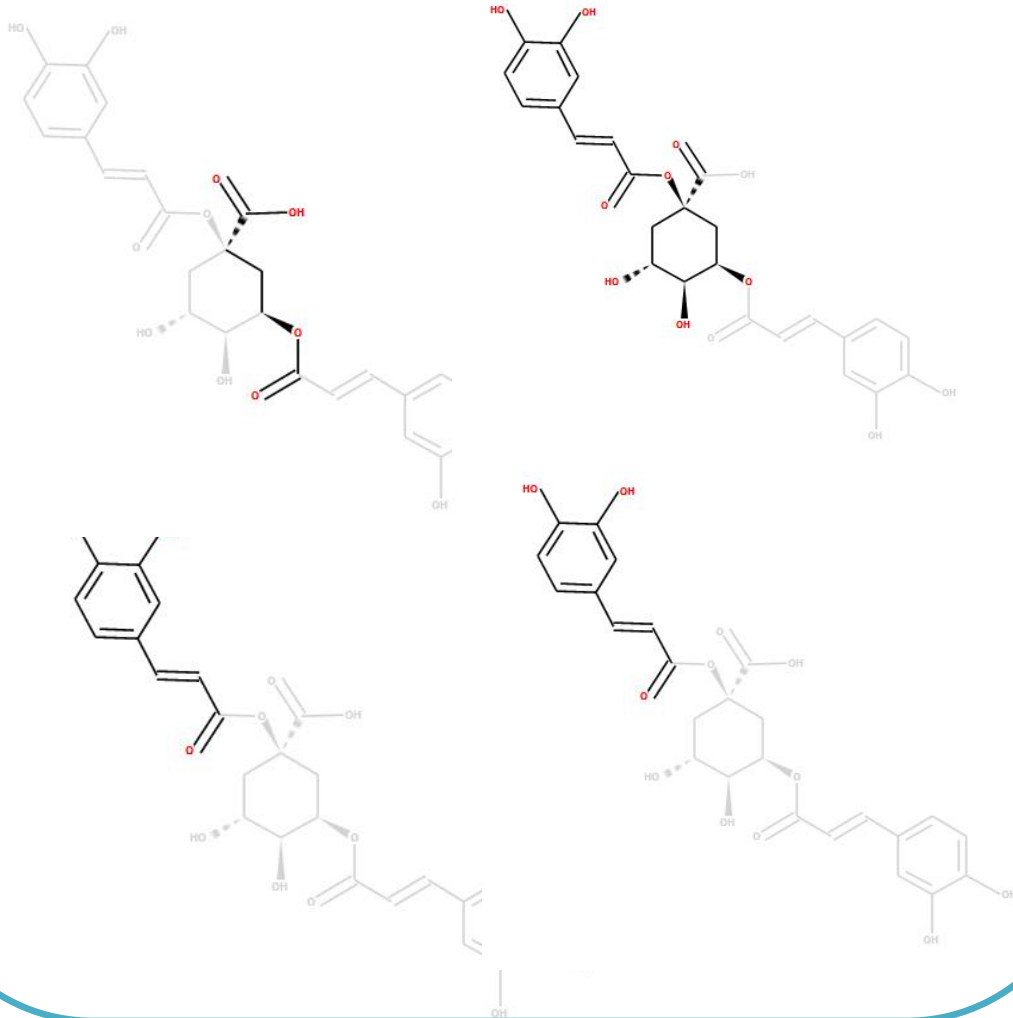
2-OH-MDIPU

Penalty=2.0 dM=4.7ppm Score=97.2 1 Of 1 Penalty=3.5 c
C9H13N-H C9H13N-H

Penalty=4.5 dM=4.7ppm Score=91.5 1 Of 1
C9H13N-H



Structural confirmation



**Confirm structure of
selected differential
metabolites**



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

From totally unknown to structures

Agilent MassHunter Molecular Structure Correlator B.05.00 -- white--M+H; ce=10

File Settings Help

Compound formula

M = 402.2247; 12 formula candidates from MFG

ID	Formula	Isomers	Taut. Gaps	dM(ppm)	IdM(ppm)	Product
1	C20H34O8	71	58	-1.7	1.7	5
2	C17H26N1O02	5	4	-1.7	1.7	6
3	C17H34N6O52	0	0	-2.9	2.9	7
4	C21H30N4O4	649	513	5.0	5.0	8
5	C16H30N6O6	66	60	-5.0	5.0	10
6	C13H34N6O6S	0	0	3.4	3.4	9
7	C21H38O3S2	2	2	3.8	3.8	2

Fragment formulas for C20H34O8

#	m/z	intensity	nom. intens.	formula	dM(ppm)
1	111.0068	52.16	12.49	C5H8O3	7.8
2	129.0179	87.16	20.87		
3	129.1000	23.13	5.54		
4	139.0020	186.88	44.75		
5	157.0131	349.55	83.70		
6	185.0804	417.62	100.00		
7	206.8935	17.41	4.17		
8	213.0762	214.41	51.34		
9	217.0338	183.86	44.03		
10	259.1527	312.02	74.71		
11	269.1372	91.25	21.85		
12	273.0969	112.80	27.01		
13	329.1599	148.82	35.64		
14	343.2102	16.44	3.94		
15	361.2223	119.38	28.59		

Structure Search

Parameters Compatibles/Total: 10/774

ChemSpider (Web) Go

Sort by Score Sort

C20H34O8; m= 402.2254

Standard InChIKey: KCCLEAUZFNBELO-UHFFFAOYSA-N

Score: 97.11

More Info

MSC Save Delete

ChemSpider: 2285382

12 mixed compound formulas

Fragments of structure #1 -- elucidated: 54.2% ions, 99.8% Weight

#	Mass	Intensity	Weight(%)	No. of candid.	Best score
1	111.0068	52.16	0.0	9	83.1
2	129.0179	87.16	0.1	14	96.3
3	129.1000	23.13	0.0	0	0.0
4	139.0020	186.88	0.2	5	86.4
5	157.0131	349.55	0.7	10	94.2
6	185.0804	417.62	2.3	22	97.1
7	206.8935	17.41	0.2	0	0.0
8	213.0762	214.41	2.8	24	93.6
9	217.0338	183.86	2.7	4	97.8
10	259.1527	312.02	13.1	19	93.2
11	269.1372	91.25	4.8	12	90.9
12	273.0969	112.80	6.5	9	98.5
13	329.1599	148.82	26.2	6	98.1
14	343.2102	16.44	3.7	6	94.5

Penalty=6.5 dM=7.8ppm Score=83.1

1 Of 4

Penalty=6.5 dM=7.8ppm Score=83.1

1 Of 2

C5H8O3-5H

C5H8O3-5H

Penalty=6.5 dM=7.8ppm Score=83.1

1 Of 1

Penalty=6.5 dM=7.8ppm Score=83.1

1 Of 2

C5H8O3-5H

C5H8O3-5H

Identification of compounds in inks for food packaging material has been possible using MSC interfaced to ChemSpider