Metabolomics by GC-MS

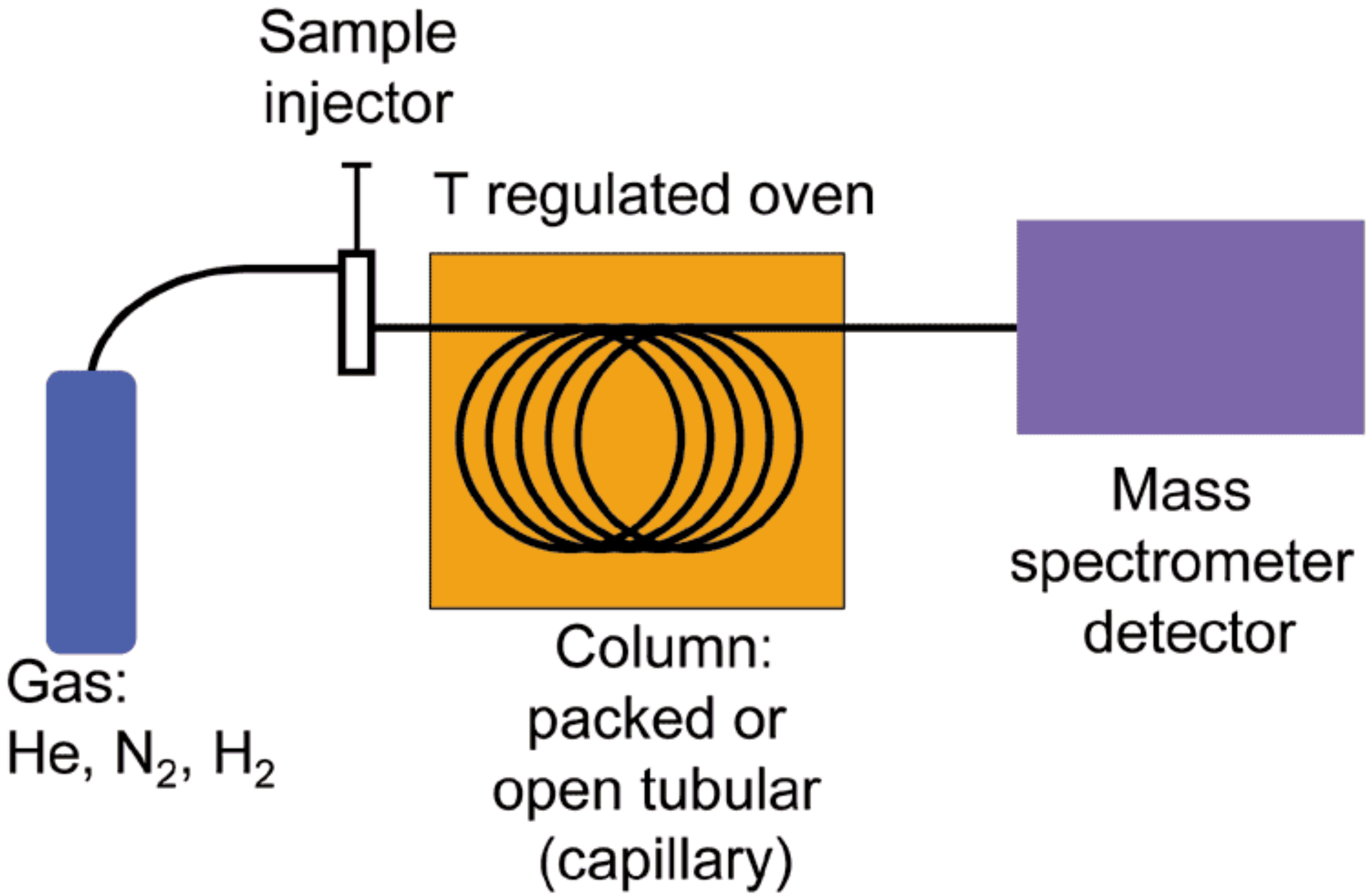
Sara J. Cooper
HudsonAlpha Institute for Biotechnology
Huntsville, AL

January 23, 2015

Outline

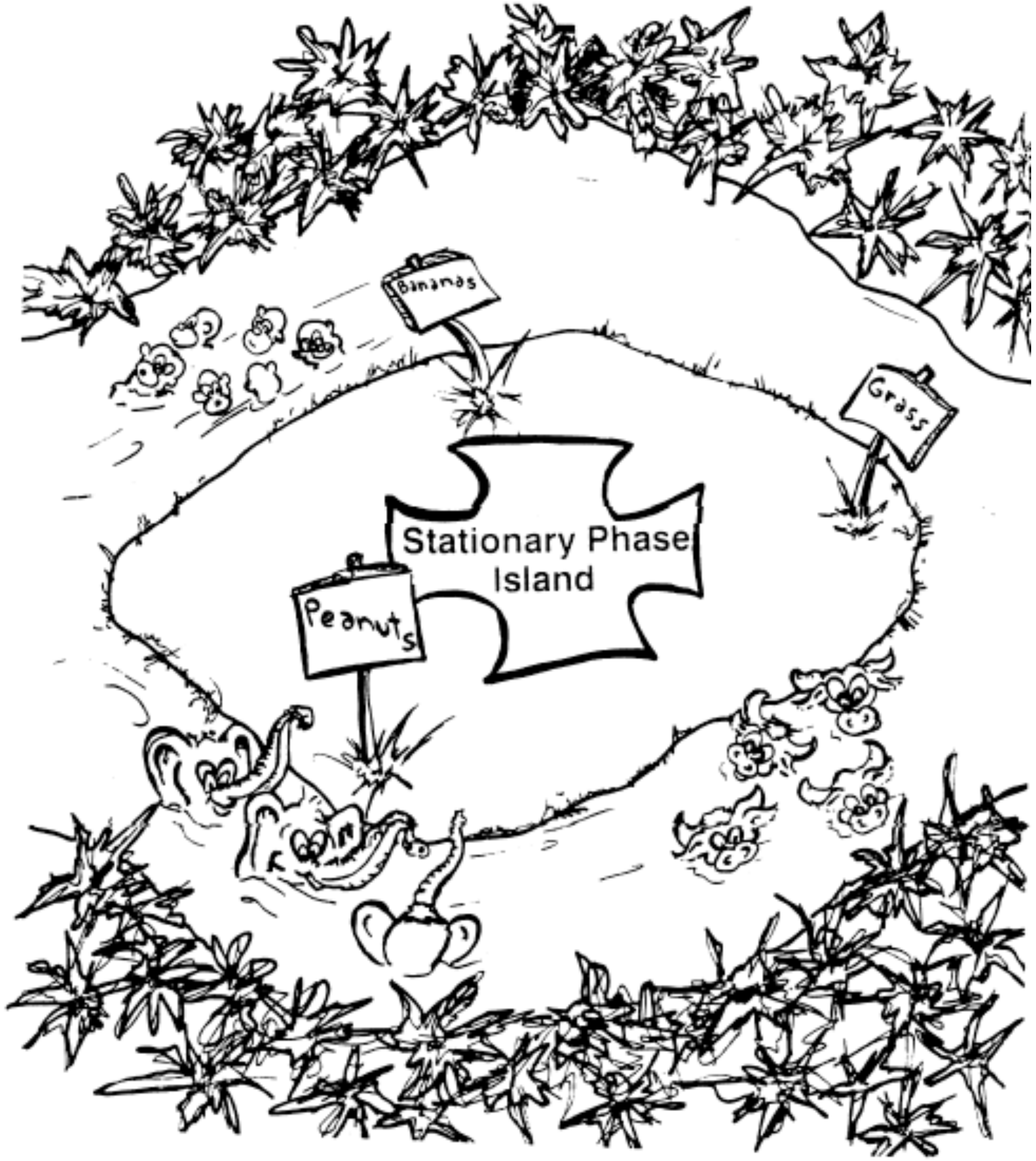
- Basics of GC-MS
  - How it works
  - How it is different from other platforms
- Applications of GC-MS for human health research
  - Designing an experiment
  - Analyzing the data (tools and tricks)
  - Signatures of Disease
  - Integrative analysis

The Nuts and Bolts of GC-MS

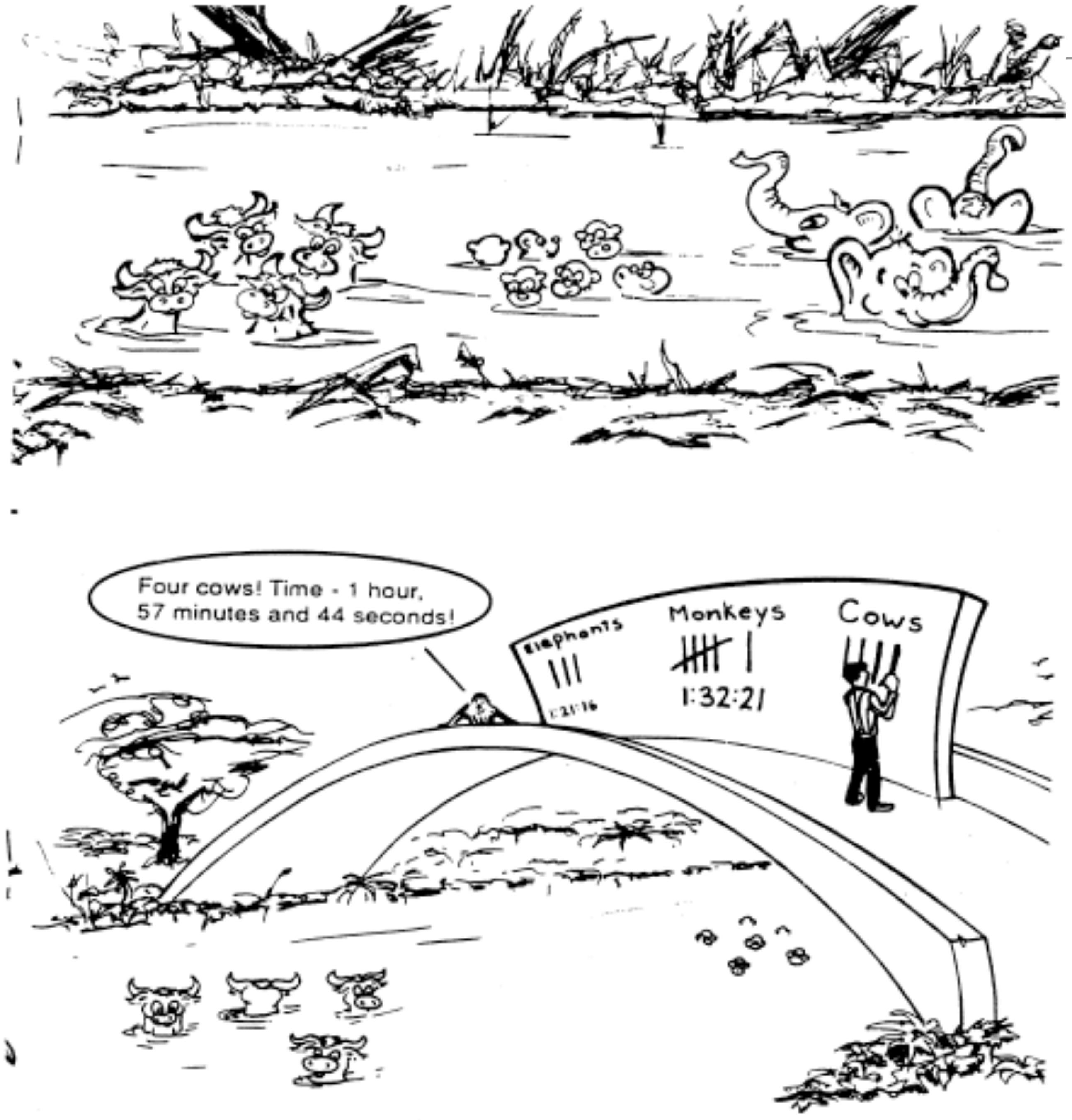

Sample injector
T regulated oven
Gas:
He, N_2, H_2
Column: packed or open tubular (capillary)
Mass spectrometer detector

"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

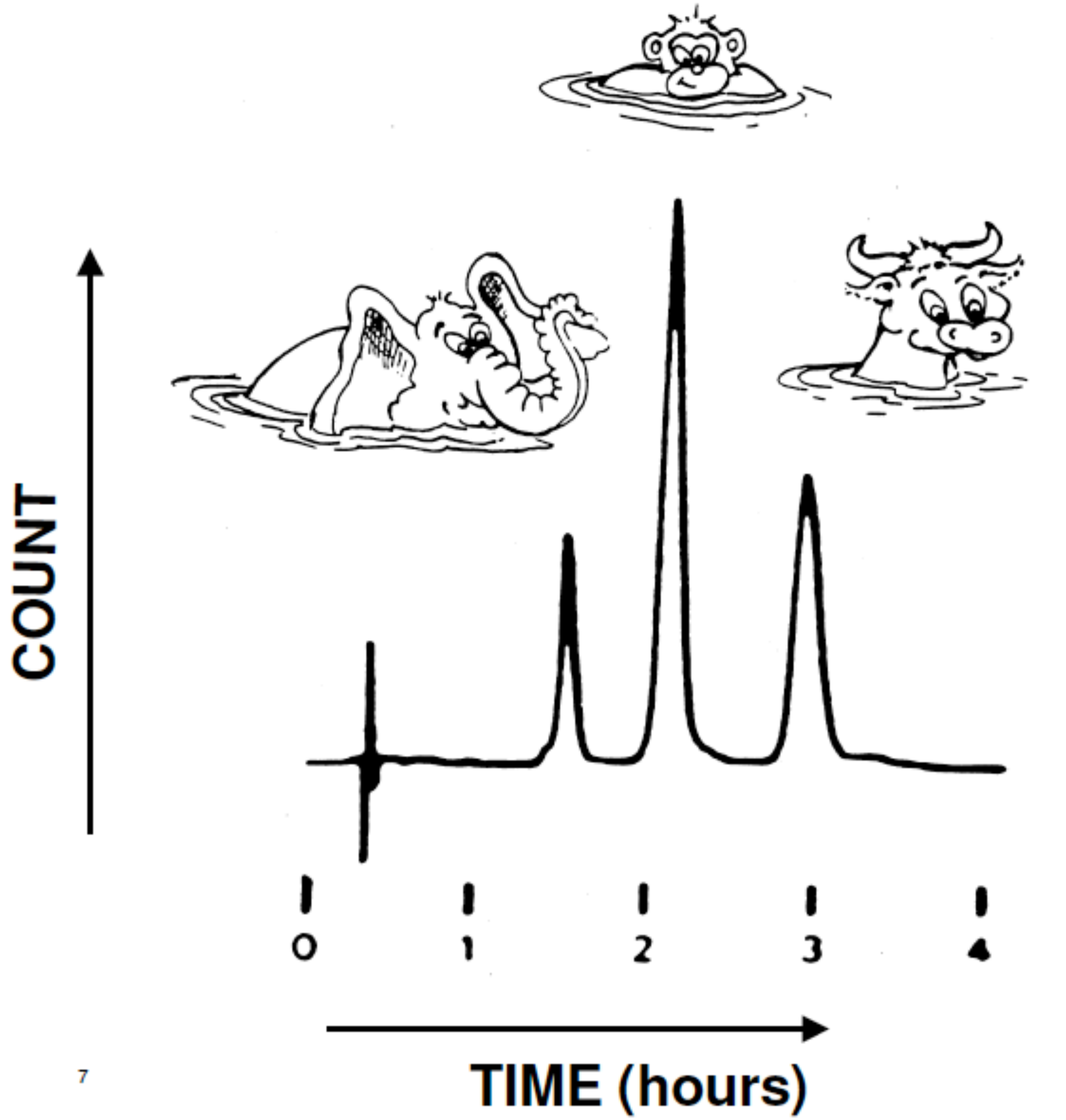
The Principal of GC



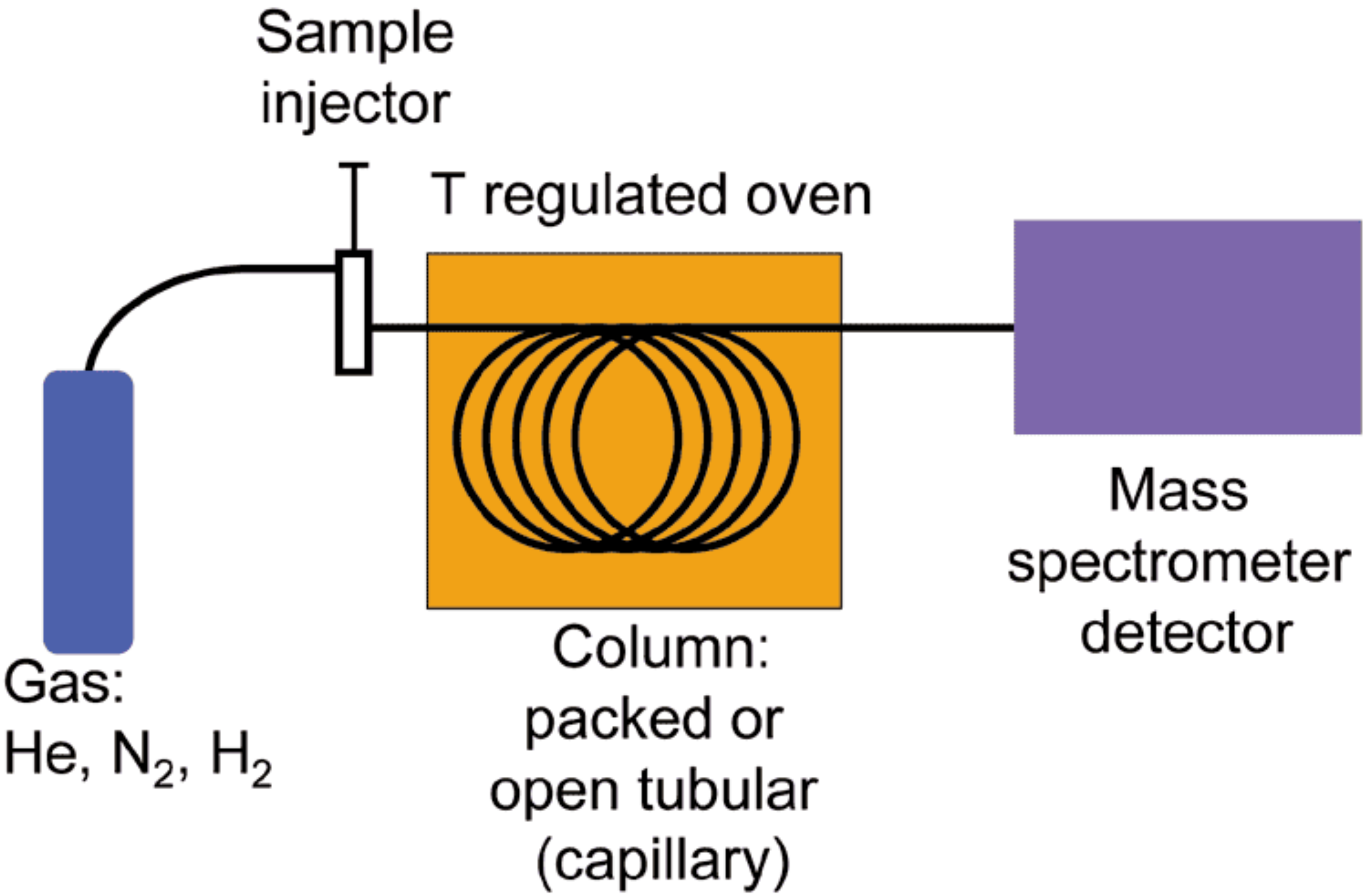
Source-SigmaAldrich ‘thebasicsofgc’



The analysis is now complete.

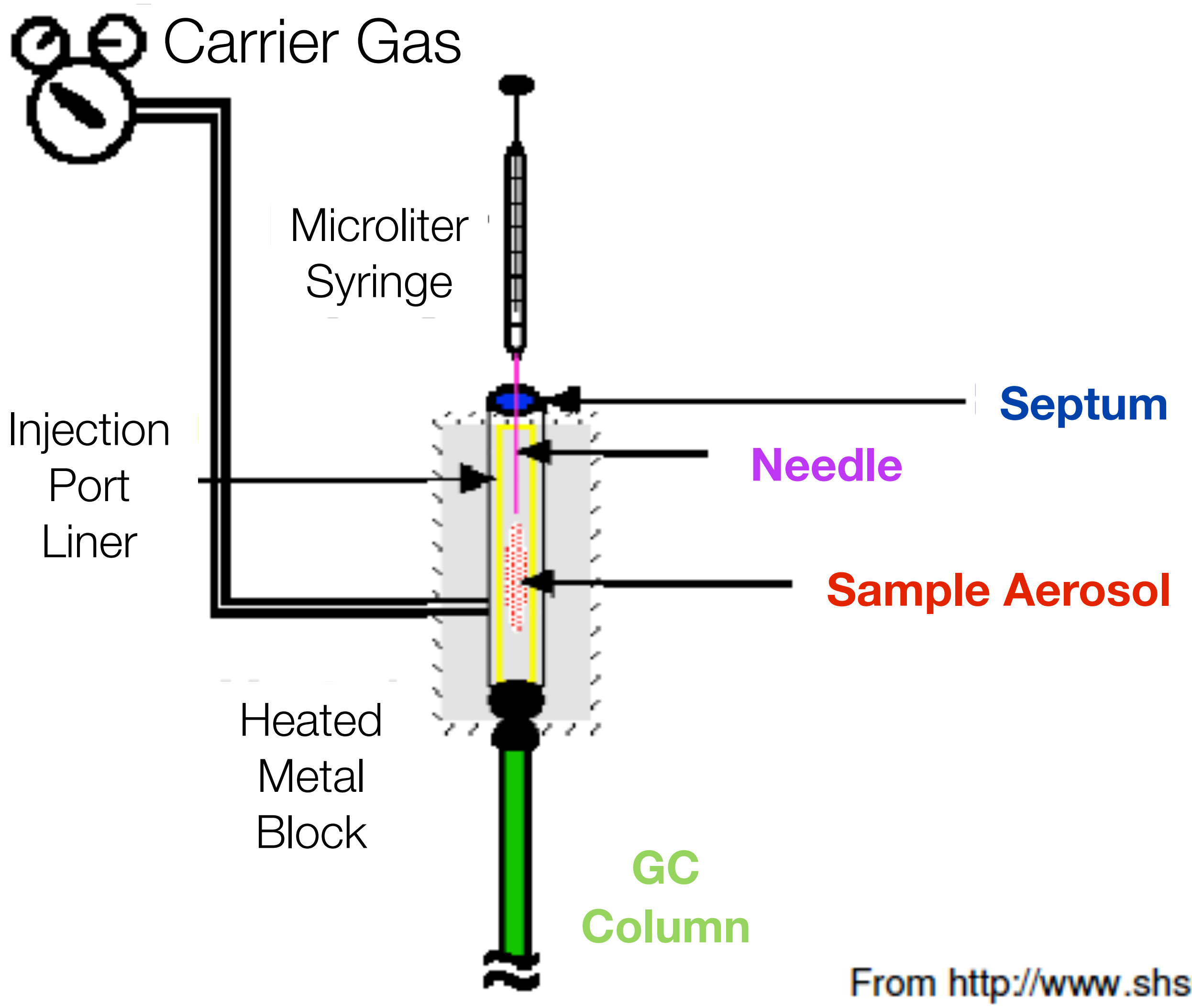


The Nuts and Bolts of GC-MS



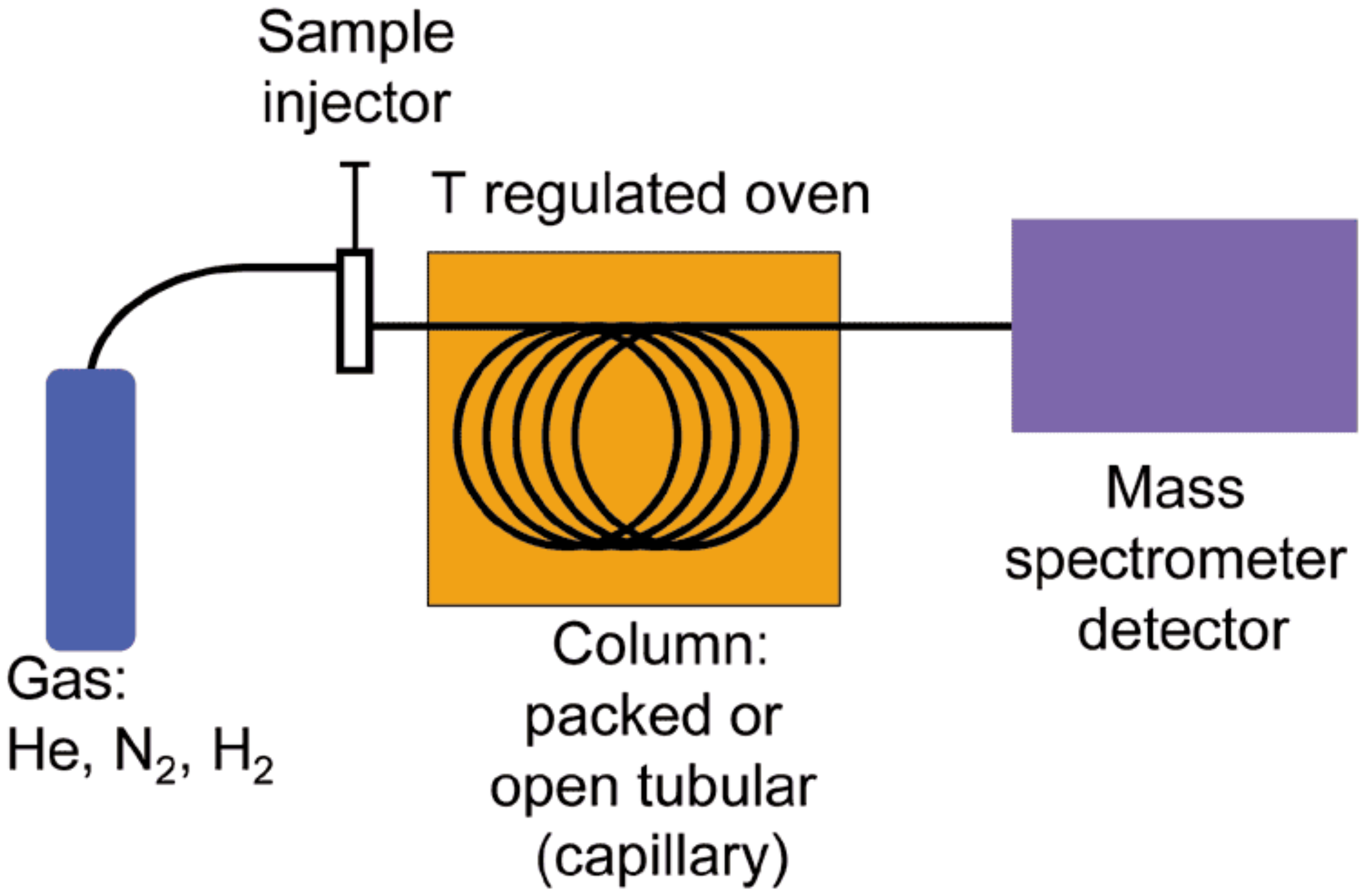
"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Injection



From http://www.shsu.edu/~chemistry/GC/packed.GIF

The Nuts and Bolts of GC-MS



Sample injector
T regulated oven
Mass spectrometer detector
Gas: He, N_2, H_2
Column: packed or open tubular (capillary)

"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Columns: Packed v. Capillary

Packed GC Columns

"Original" GC column
Low efficiency
Coated phase: organic polymers dissolved in solvent and coated onto particles in the tube

Capillary GC Columns

Modern GC column
High efficiency
Usually flexible glass fiber (fused silica) < 1mm ID
Coated phase: organic polymers dissolved in solvent and coated on the inside wall column

Can be 10-30+ meters long
Longer column is better separation, particularly for complex mixtures

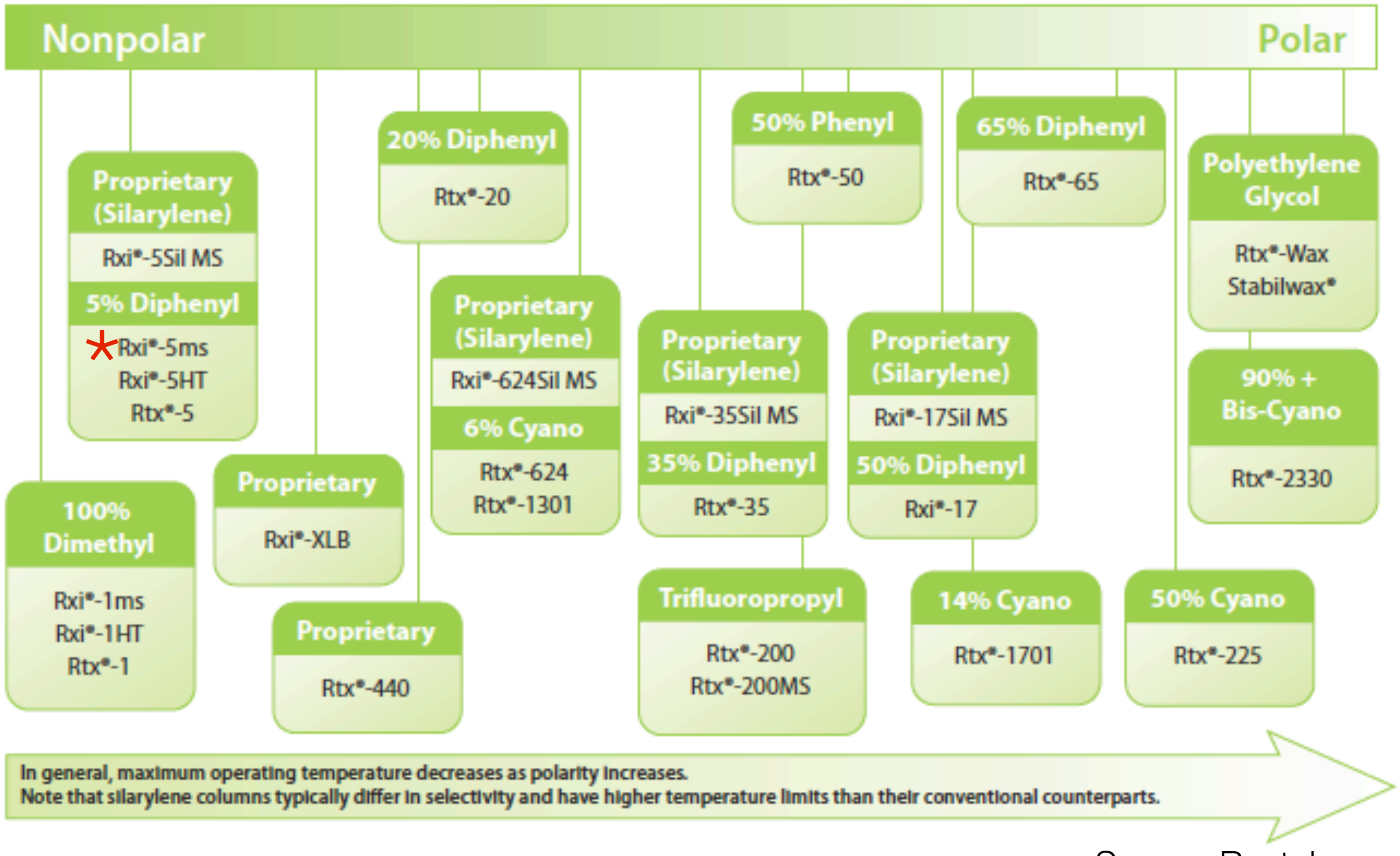
Modified from sigmaaldrich.com

Selecting a column

A nonpolar stationary phase is used for separation of polar analytes
Thickness of the stationary phase affects retention time and column capacity
Inner diameter affects separation and retention times

STATIONARY PHASE

Figure 2: Polarity scale of common stationary phases.



In general, maximum operating temperature decreases as polarity increases.
Note that silarylene columns typically differ in selectivity and have higher temperature limits than their conventional counterparts.

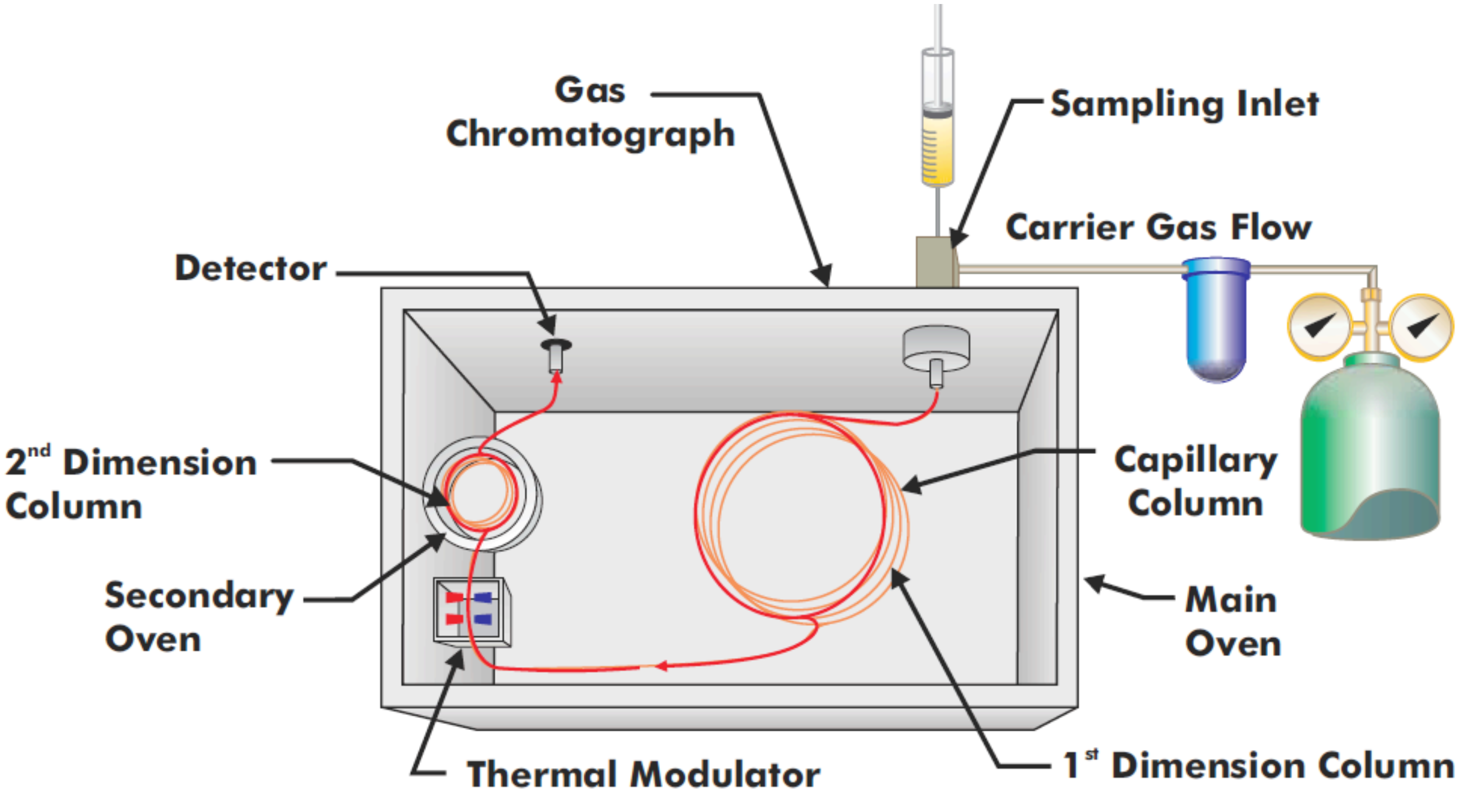
Source-Restek

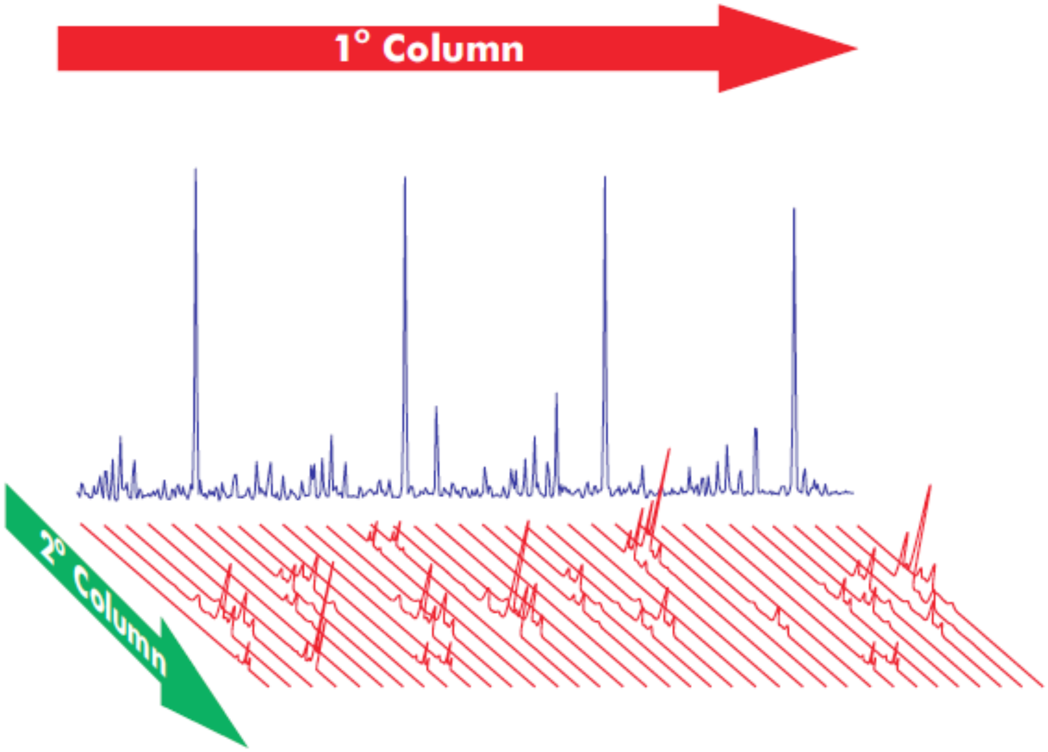
tech tip

Any homologous series of compounds, that is, analytes from the same chemical class (e.g., all alcohols, all ketones, or all aldehydes, etc.) will elute in boiling point order on any stationary phase. However, when different compound classes are mixed together in one sample, intermolecular forces between the analytes and the stationary phase are the dominant separation mechanism, not boiling point.

Two-dimensional chromatography

- GC Columns function in series to improve resolution of chemically similar analytes

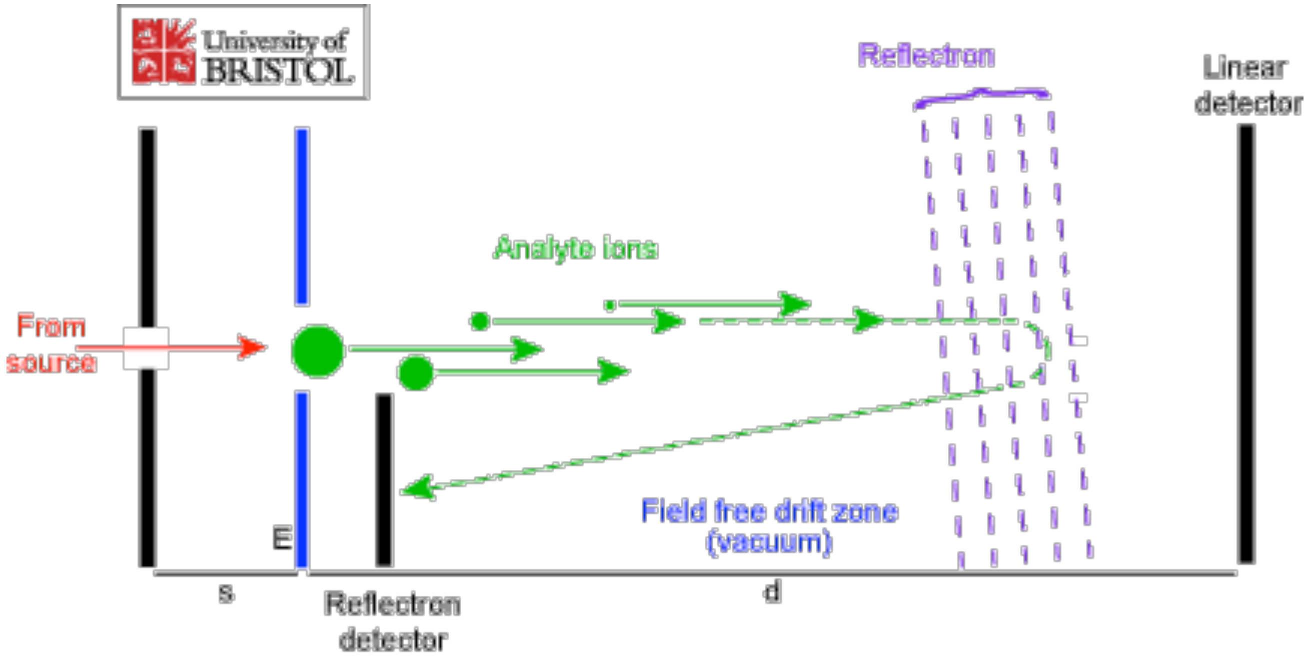




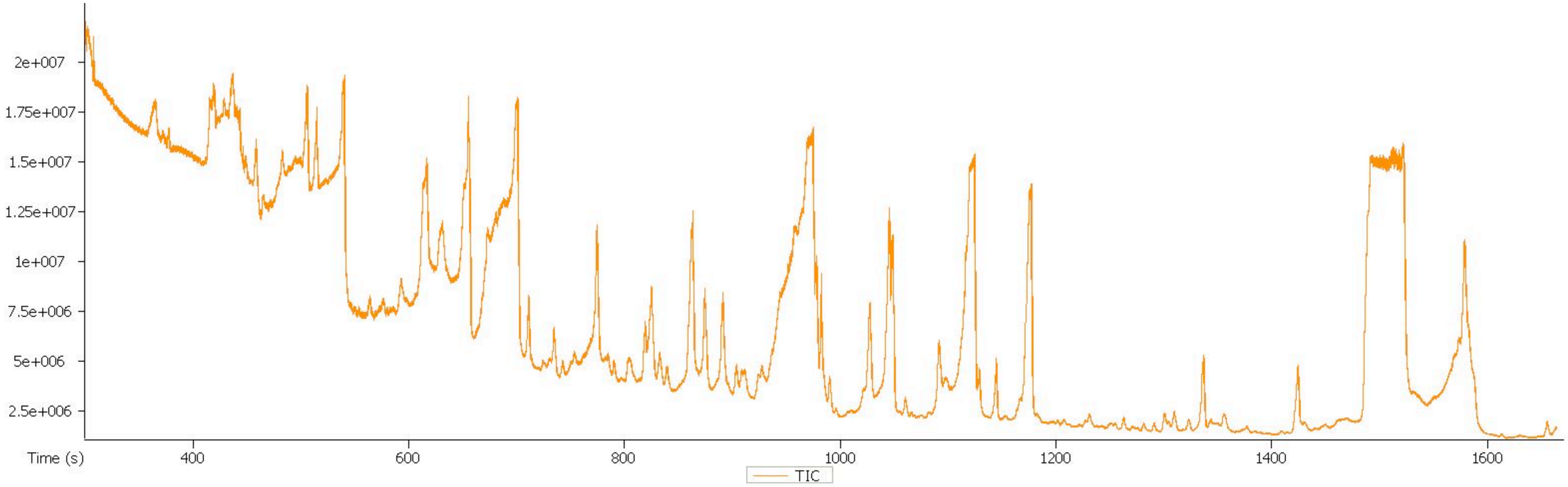
Source: Leco Corp

Mass Spectrometer - Ionization and mass measurement

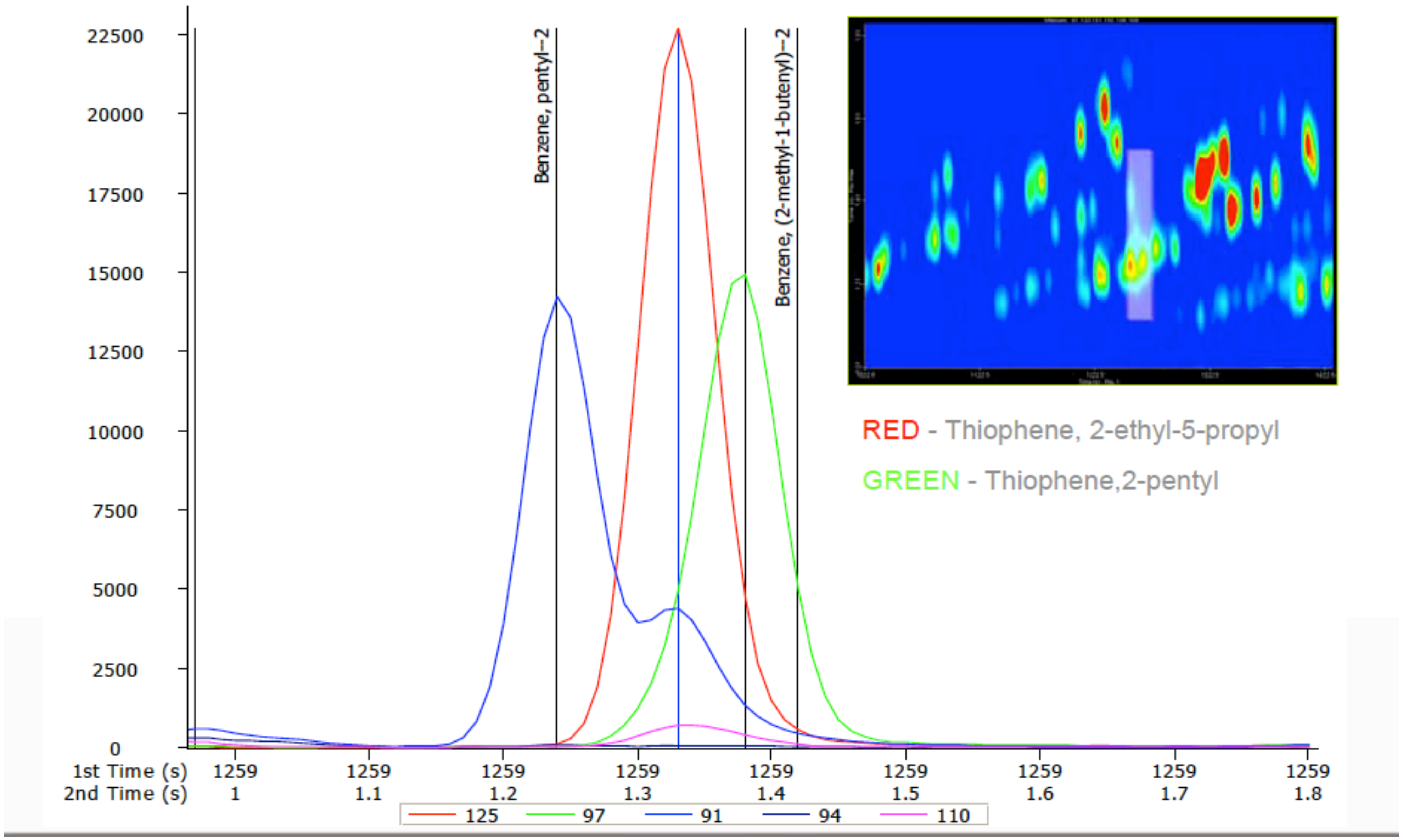
- Ionization
  - Electron Ionization (Standard -70keV)
    - Fragmentation
  - Chemical Ionization (less common)
- Detection
  - Time-of-flight mass spectrometry
    - mass calculated based on time from ionization to reaching detector
  - High-Resolution TOF
    - offers higher mass resolution for metabolite identification



Example data output-Chromatogram



Signal Deconvolution



True Signal Deconvolution®

RED - Thiophene, 2-ethyl-5-propyl

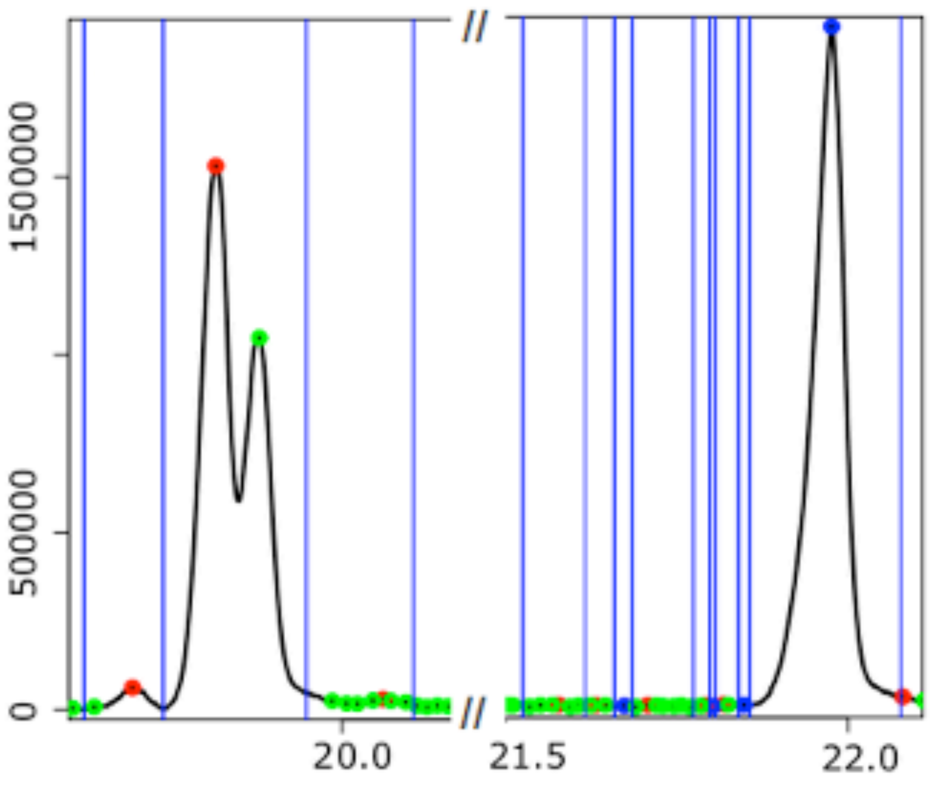
GREEN - Thiophene,2-pentyl

Source: Leco

Principles of Deconvolution

- Generally implemented in AMDIS

- Goal: computationally separate chromatographically overlapping peaks



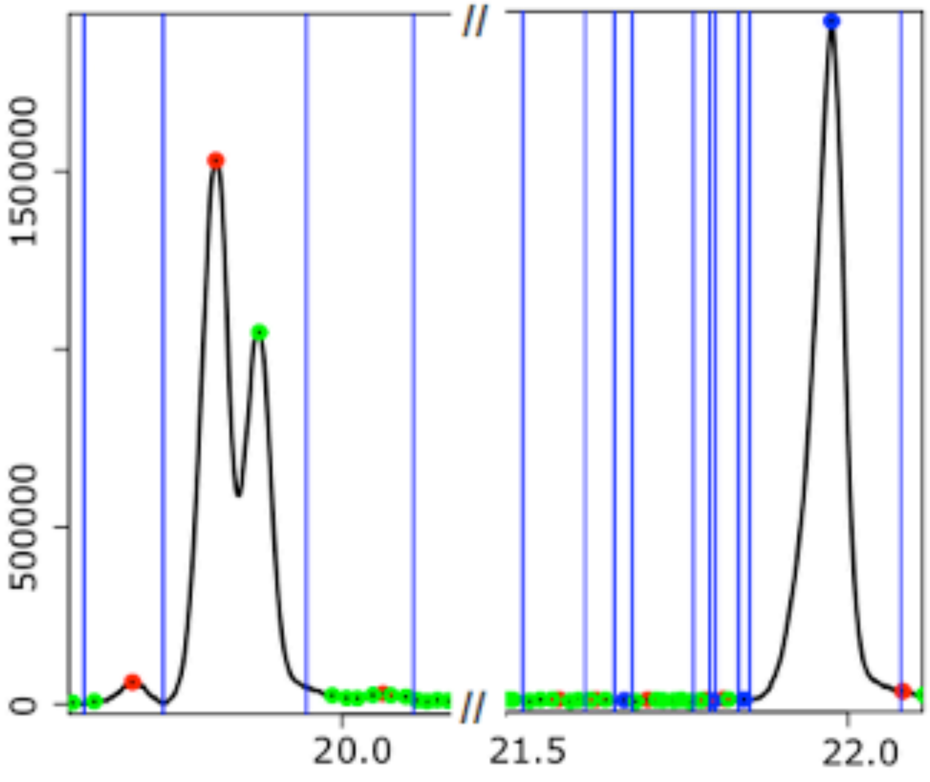
TIC

Source: Du and Zeisel 2013

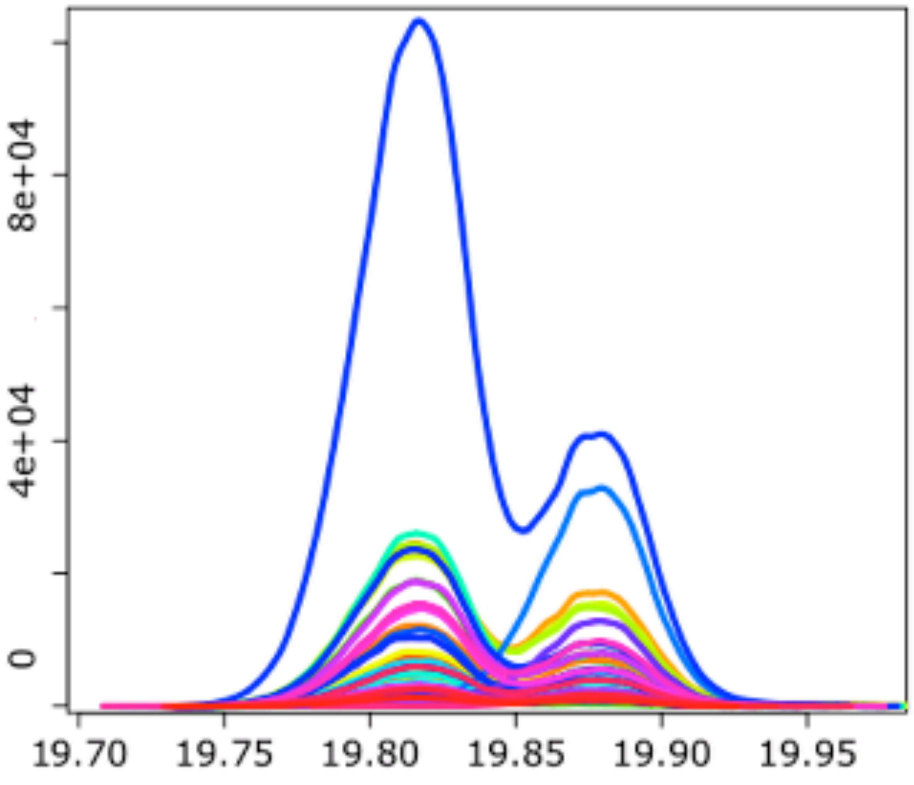
Principles of Deconvolution

- Generally implemented in AMDIS

- Goal: computationally separate chromatographically overlapping peaks



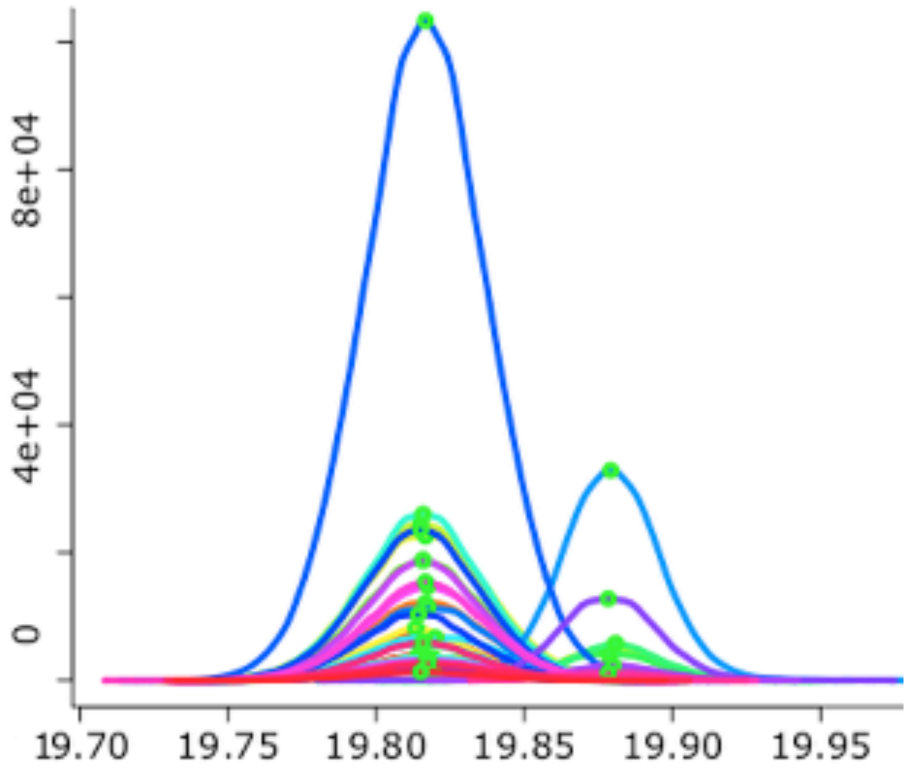
TIC



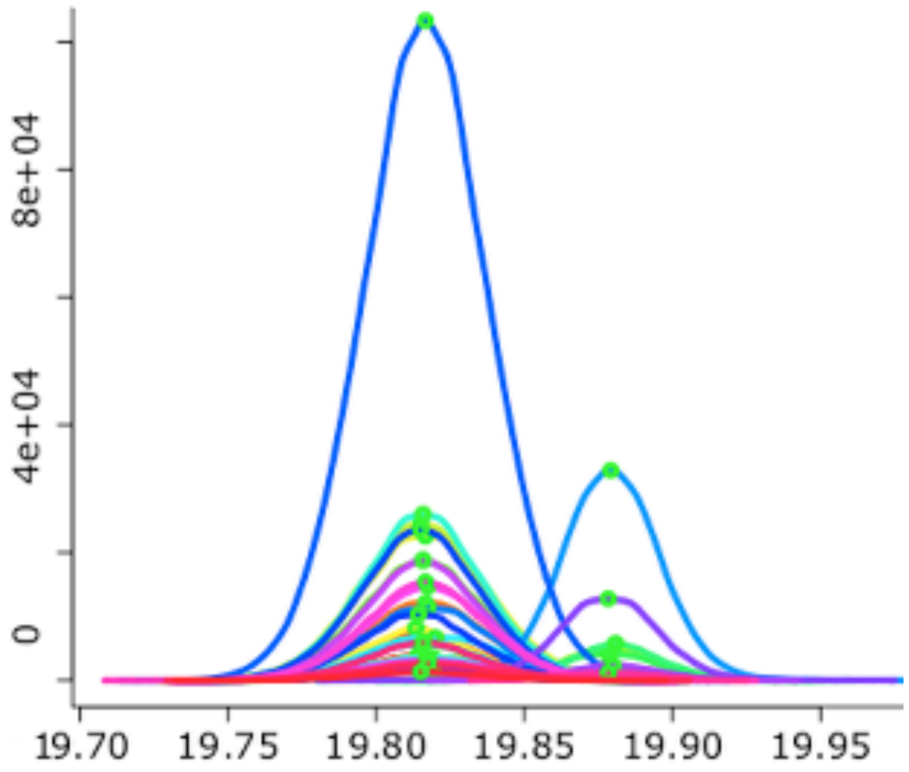
Individual ions

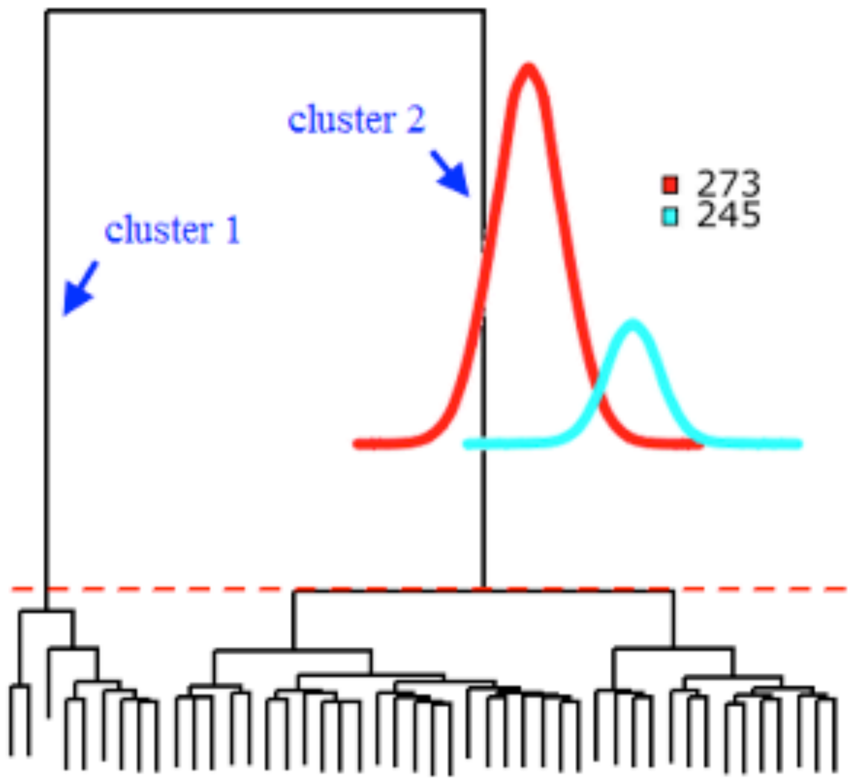
Source: Du and Zeisel 2013

Principles of Deconvolution

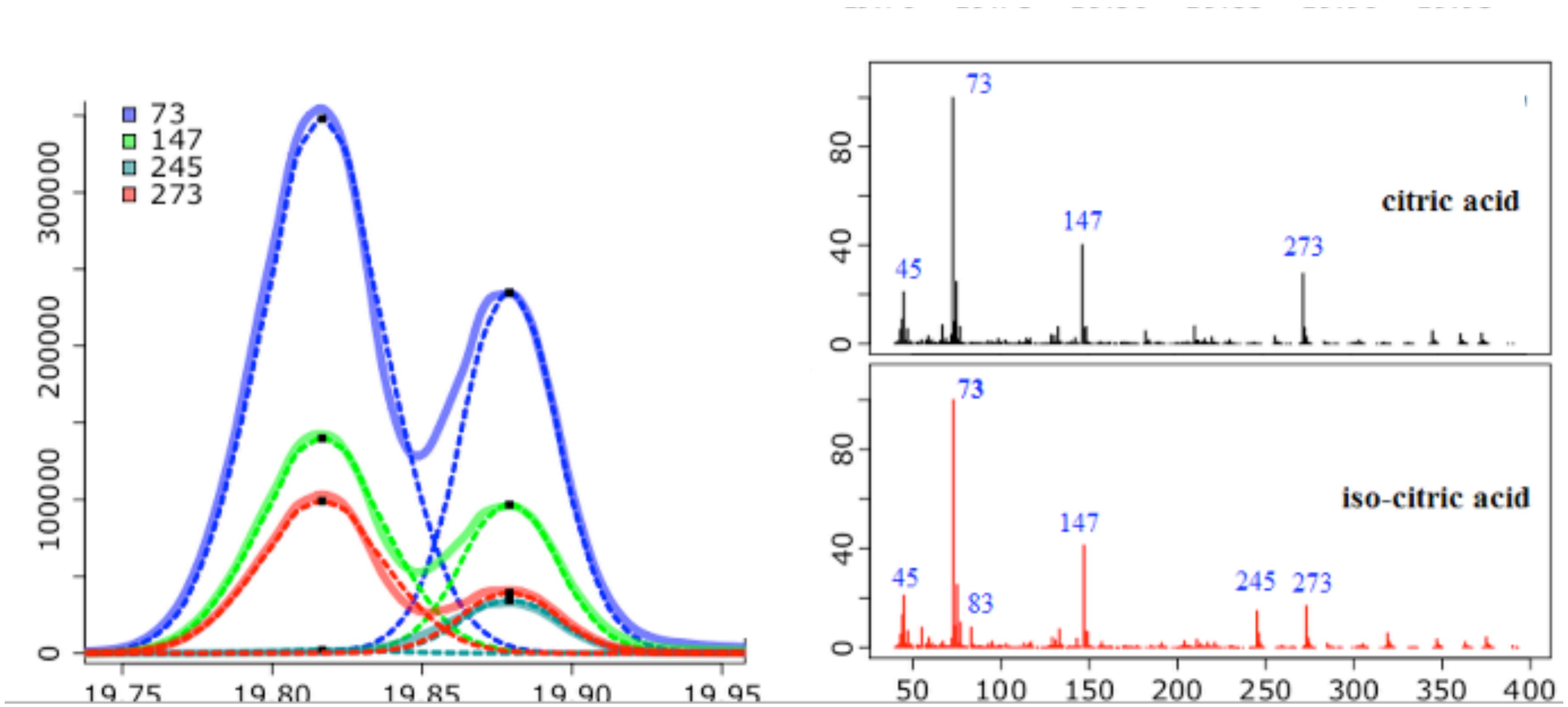


Principles of Deconvolution

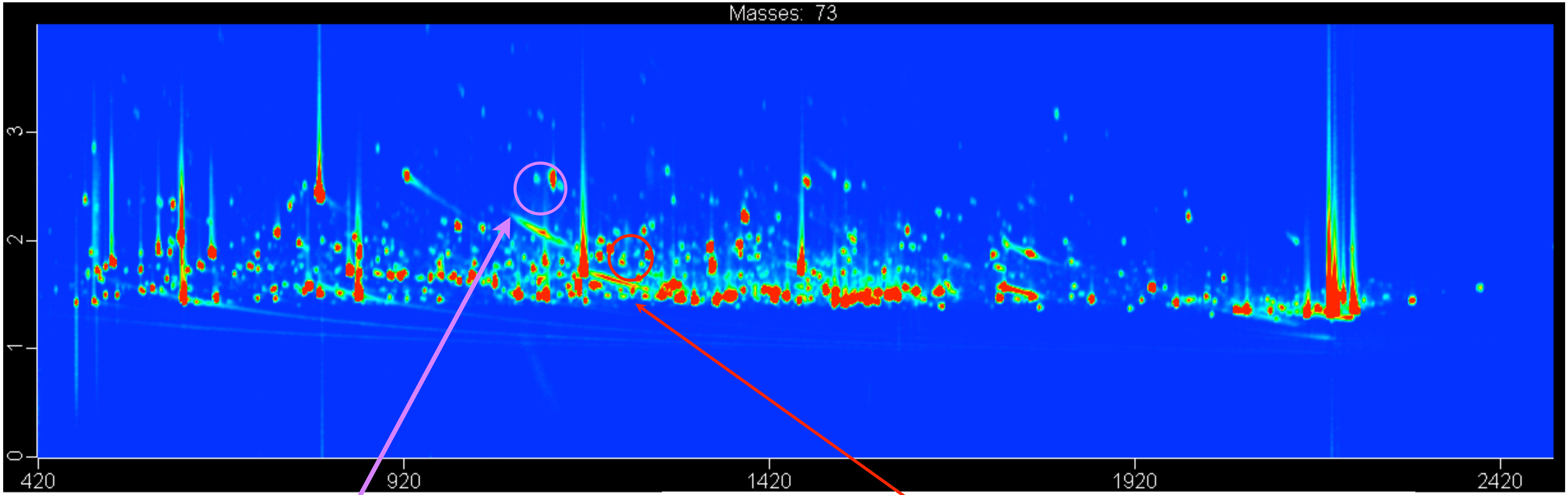




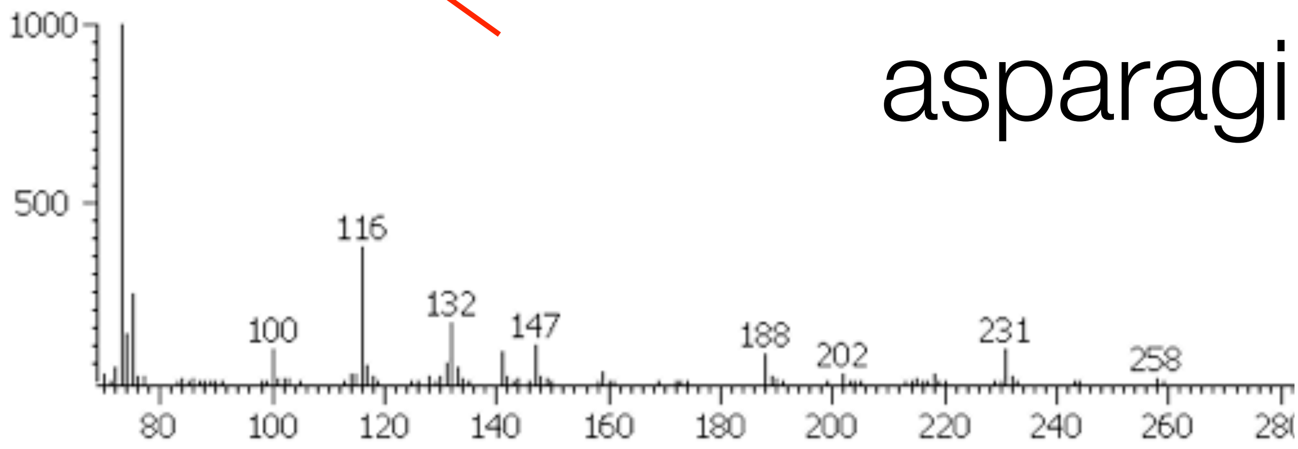
Principles of Deconvolution



Data projected into two dimensions



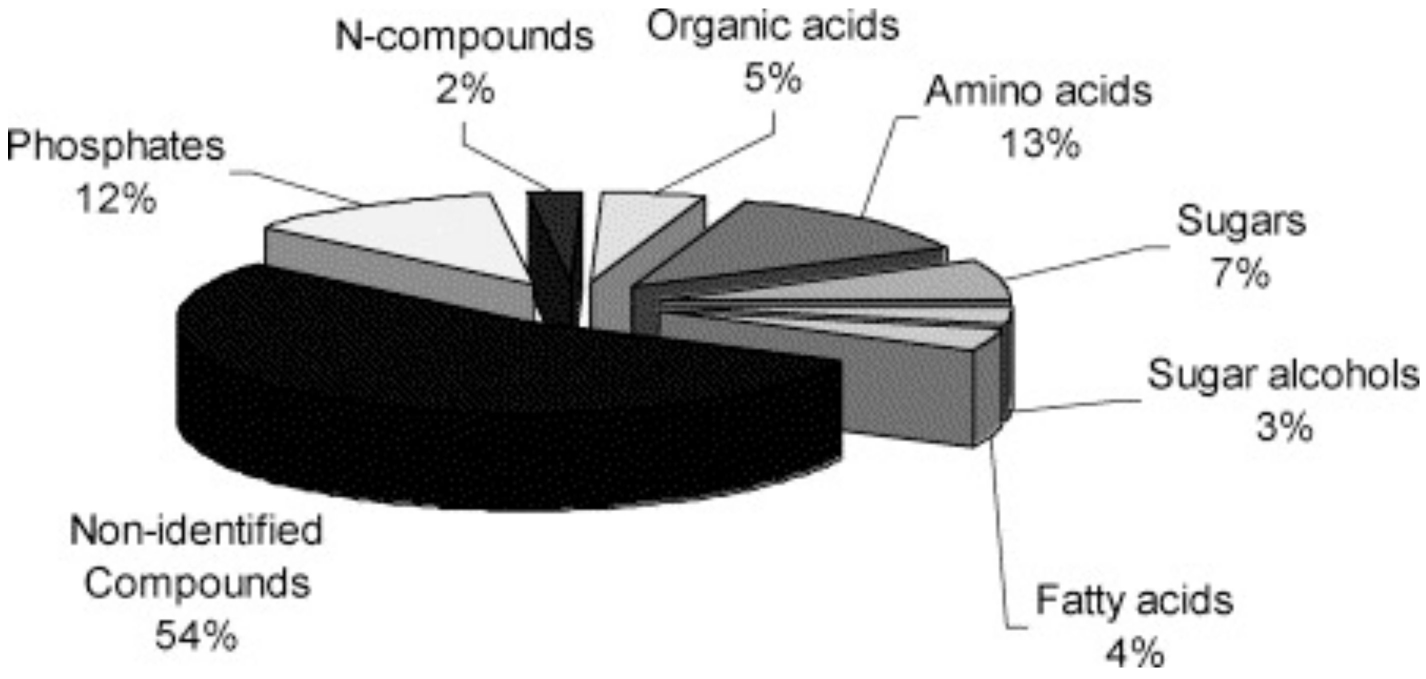
Glutamate



asparagine

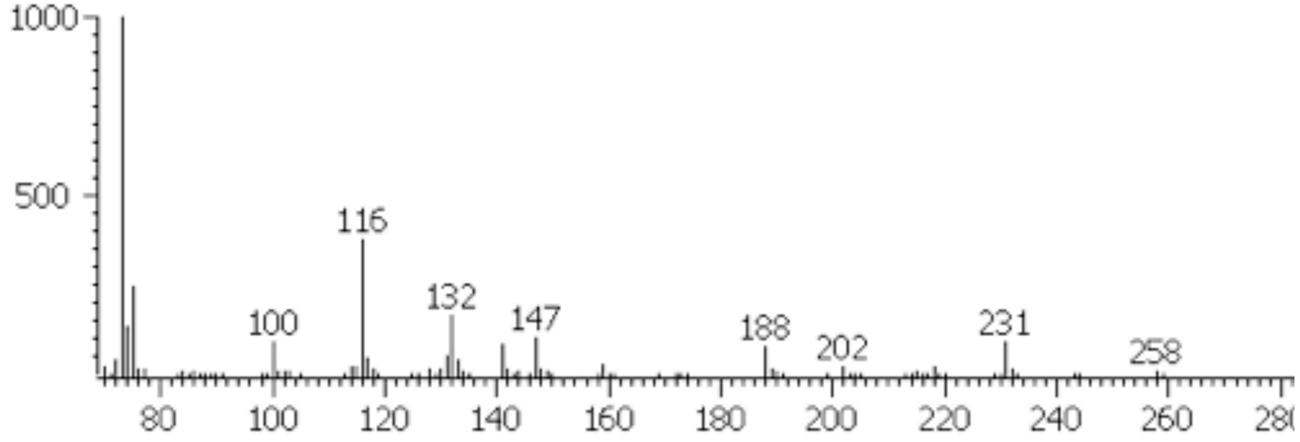
Metabolite Identification

- Reproducible fragmentation has generated libraries of known compounds
- Calculating similarity:
  - Retention indices are routinely used to validate or improve metabolite identification based on relative retention times. (Kovats index)
  - Using a dot-product based metric, analytes can be assigned an ID based on similarity to known compounds

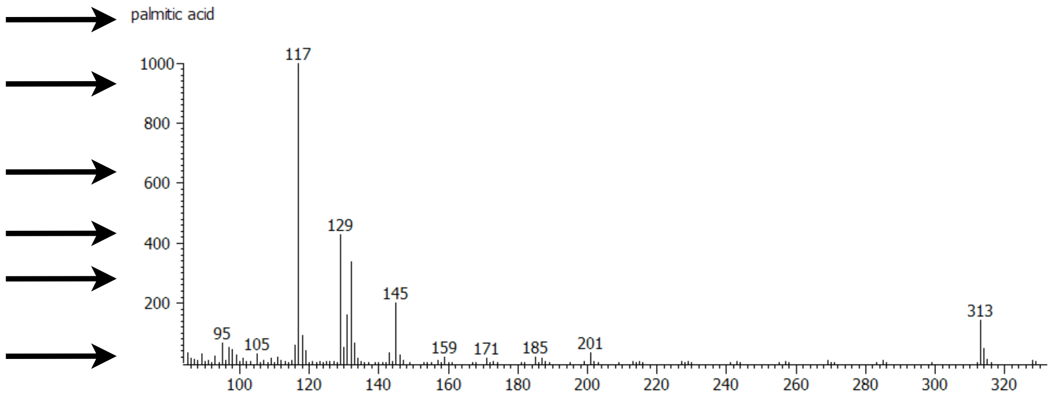


source: Schauer et al 2005

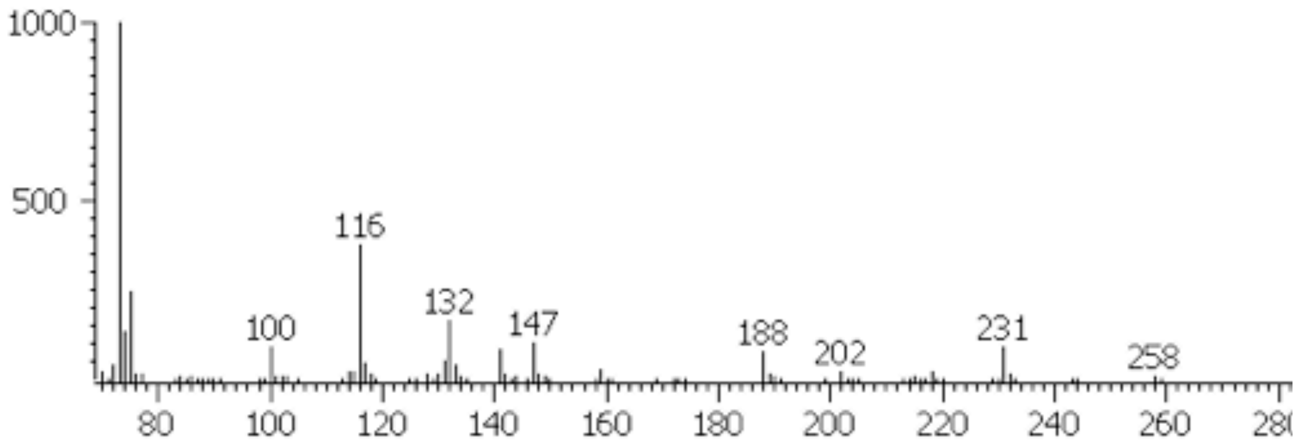
Library matching



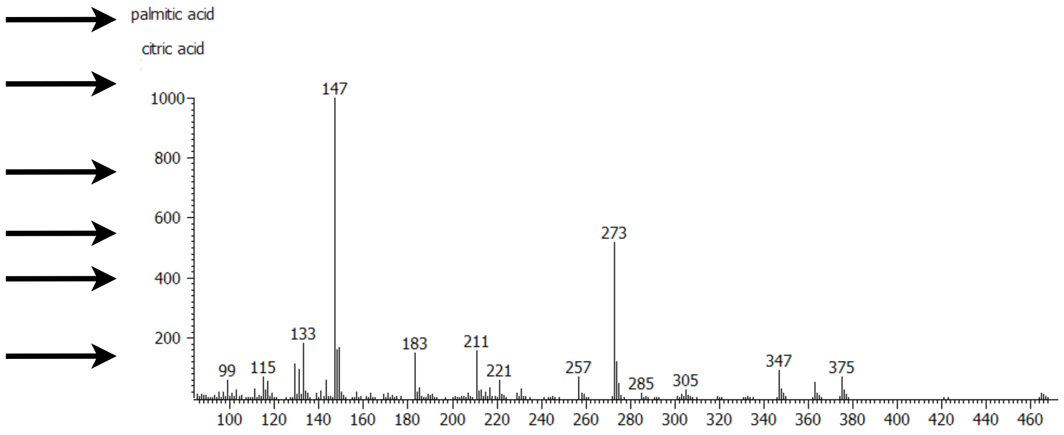
Unknown spectrum



Library matching



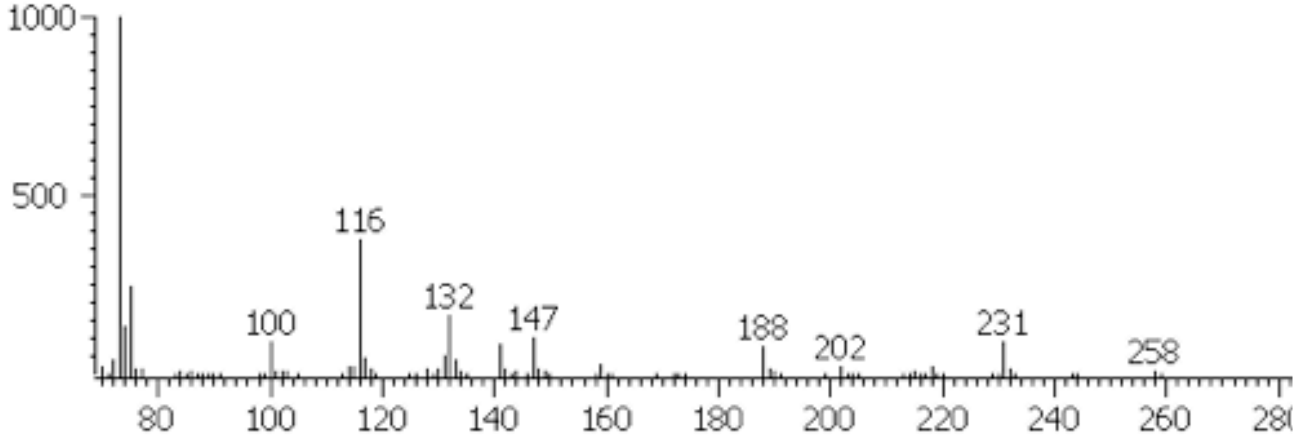
Unknown spectrum



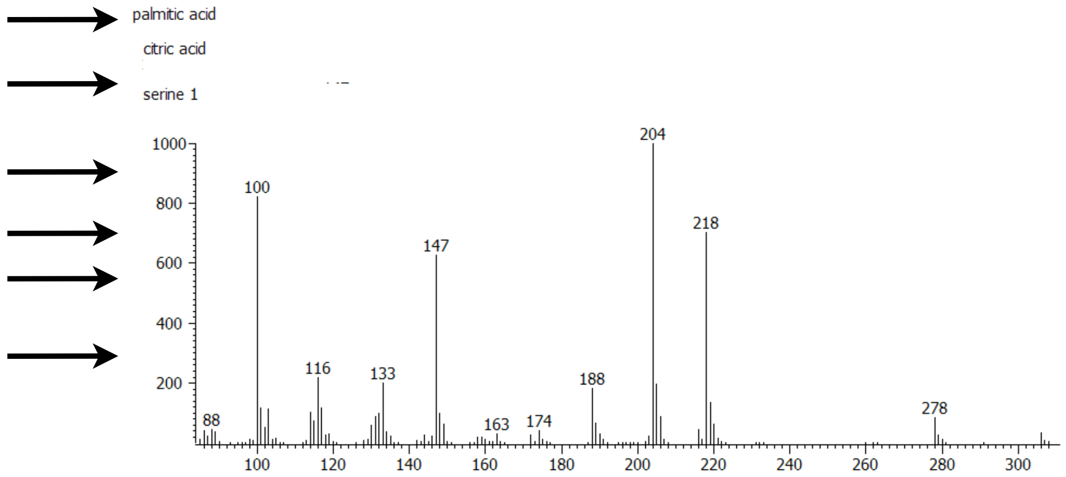
palmitic acid

citric acid

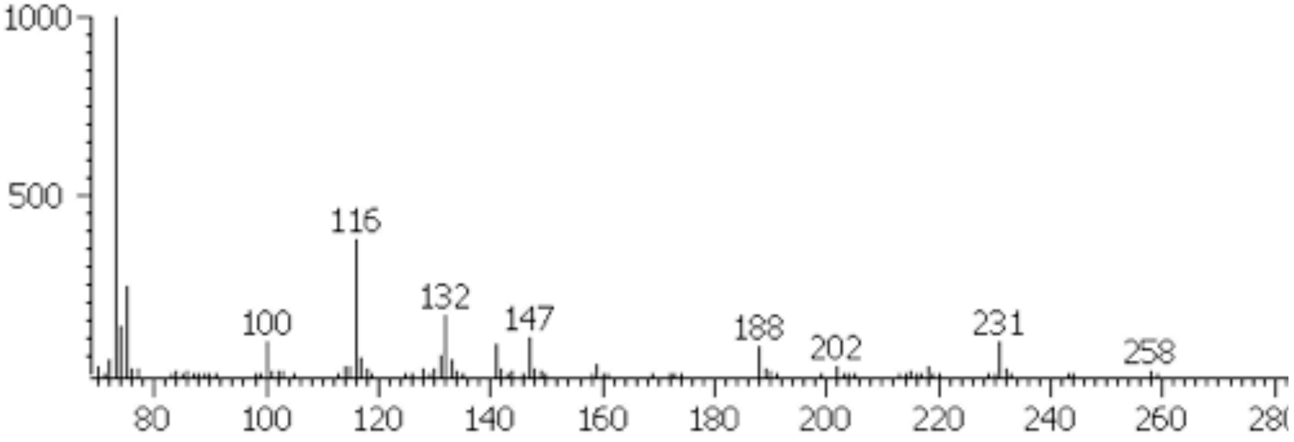
Library matching



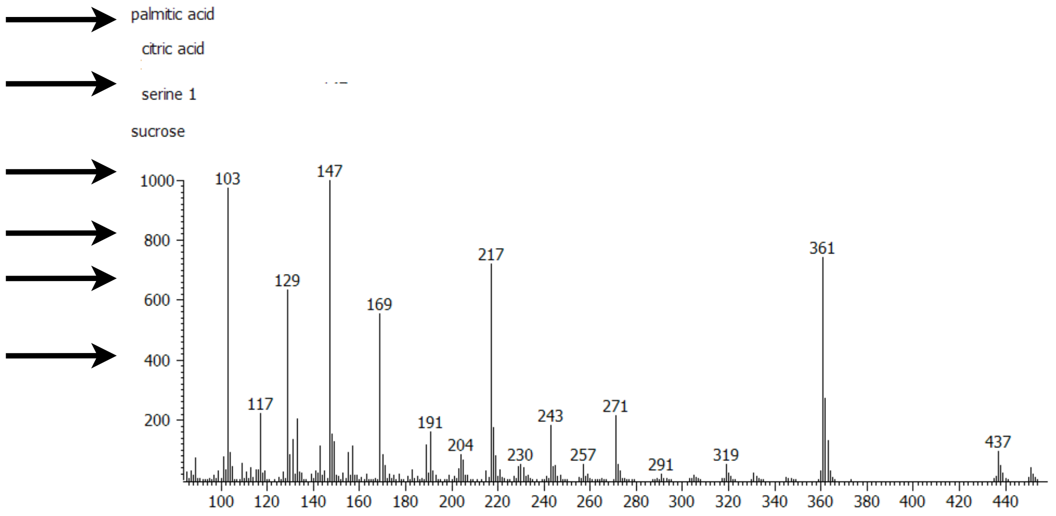
Unknown spectrum



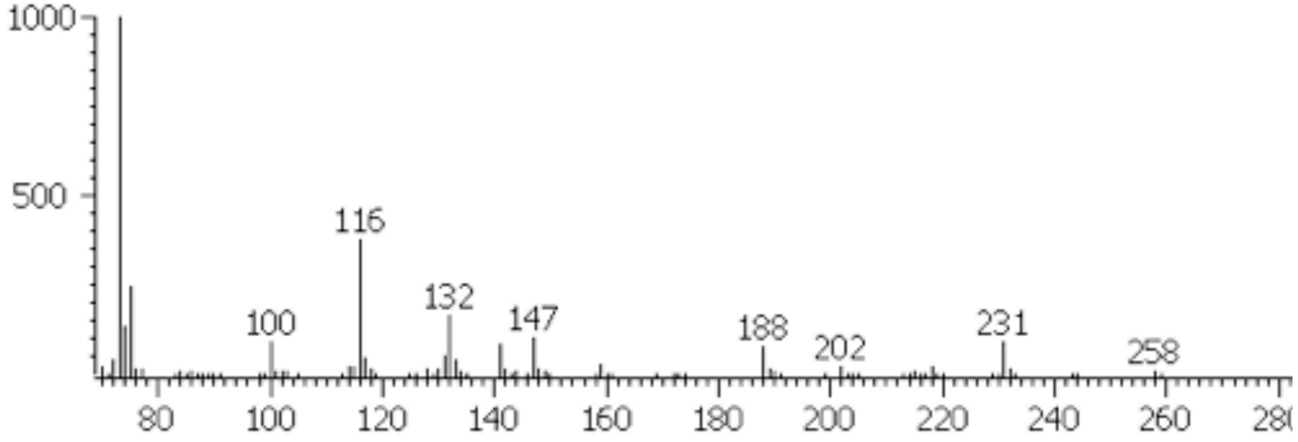
Library matching



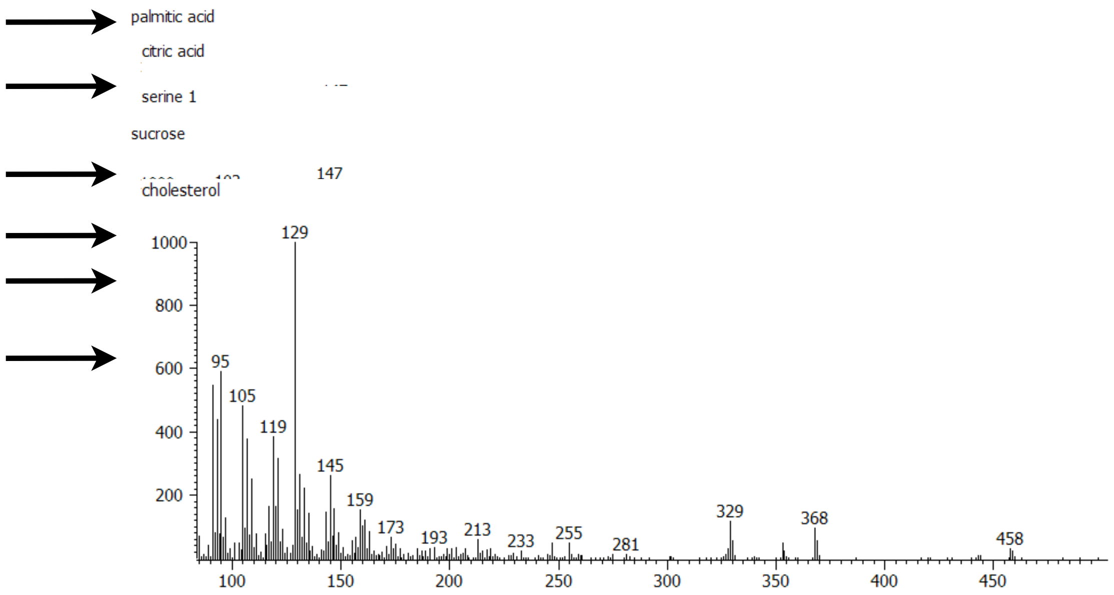
Unknown spectrum



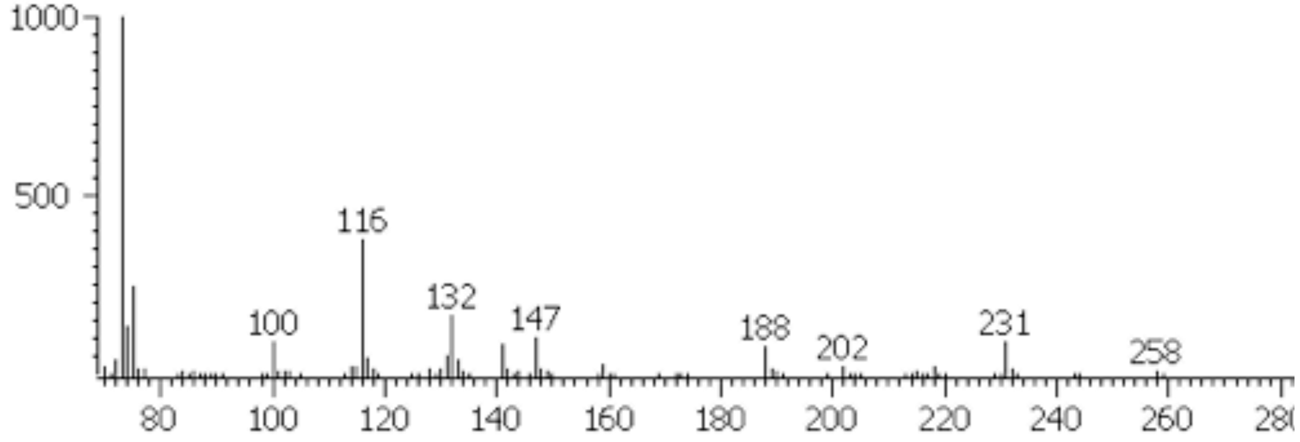
Library matching



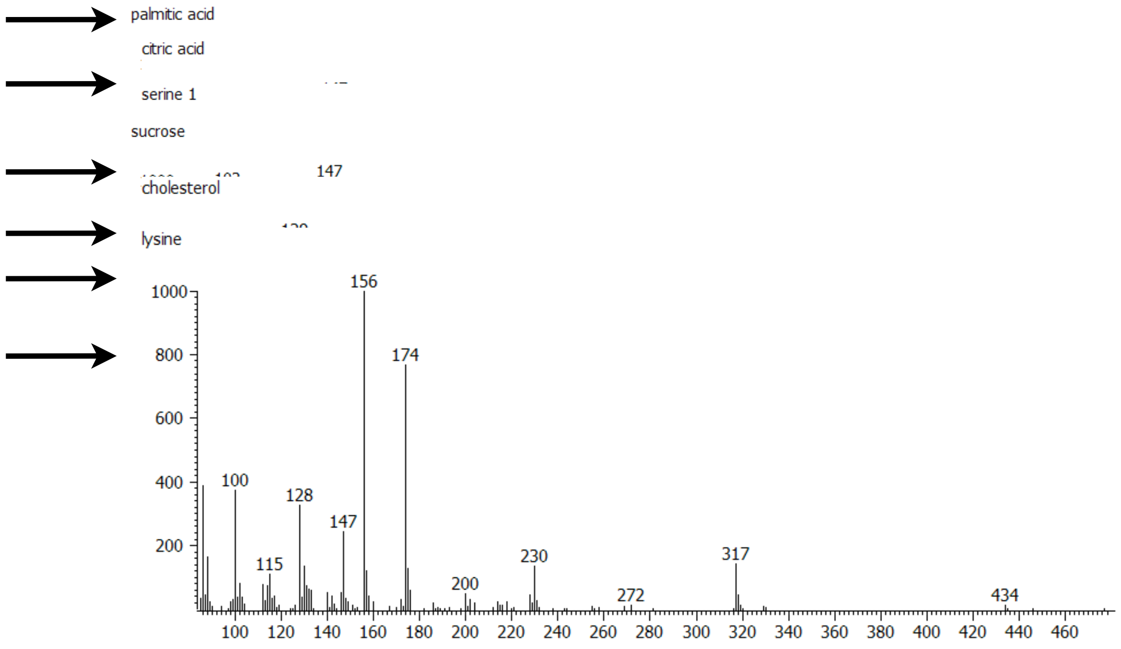
Unknown spectrum



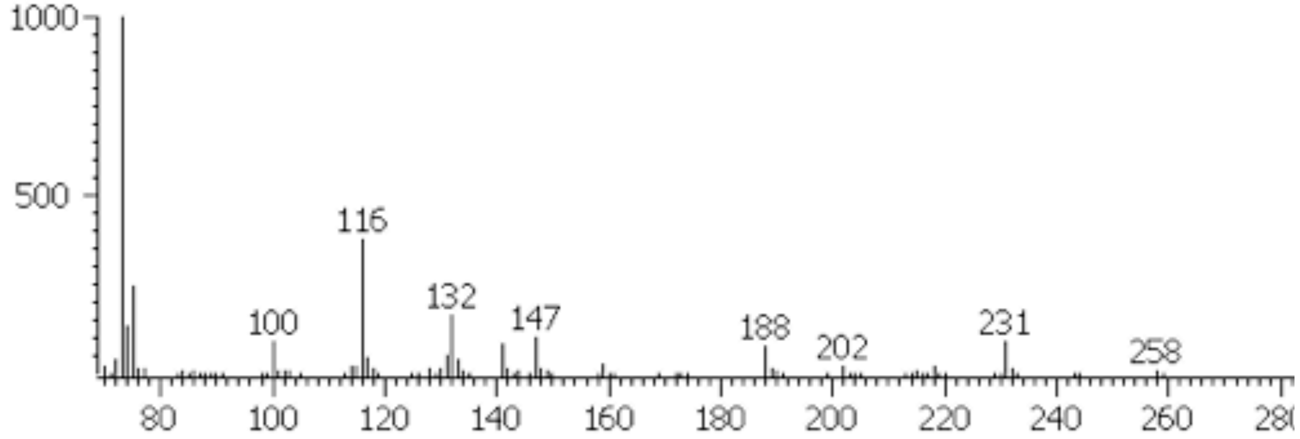
Library matching



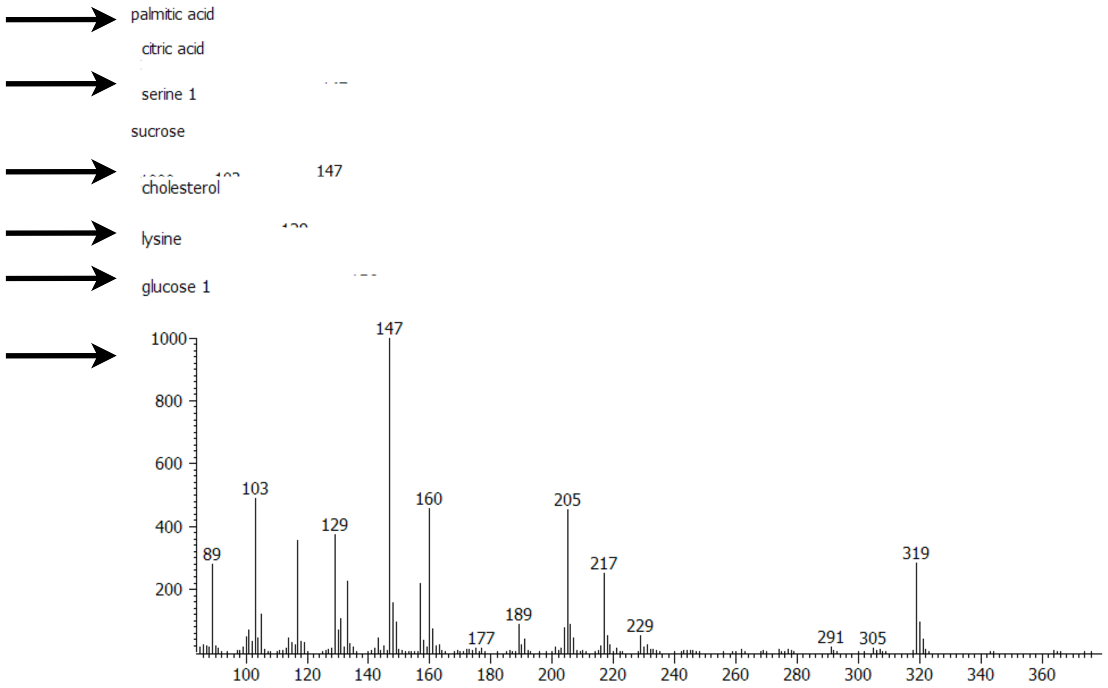
Unknown spectrum



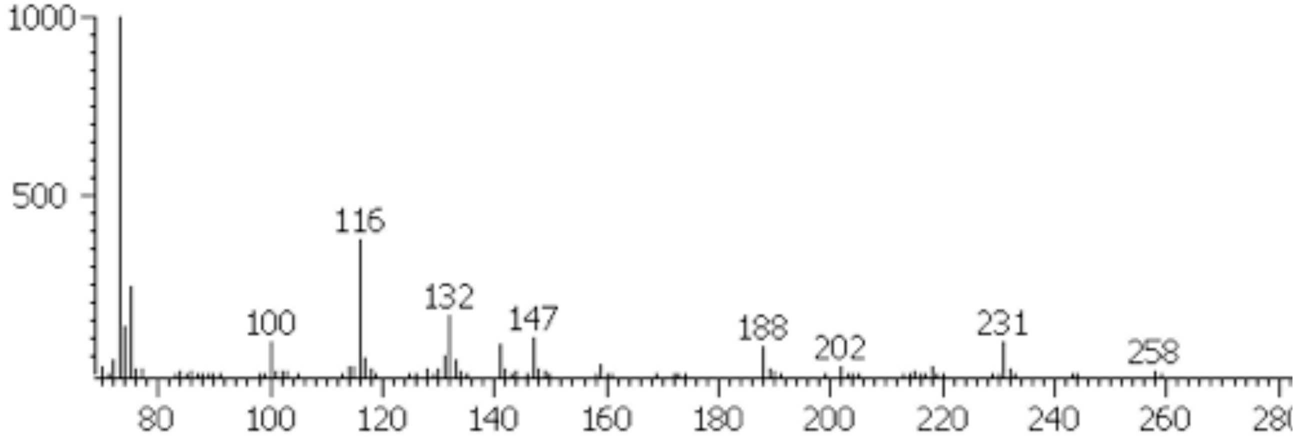
Library matching



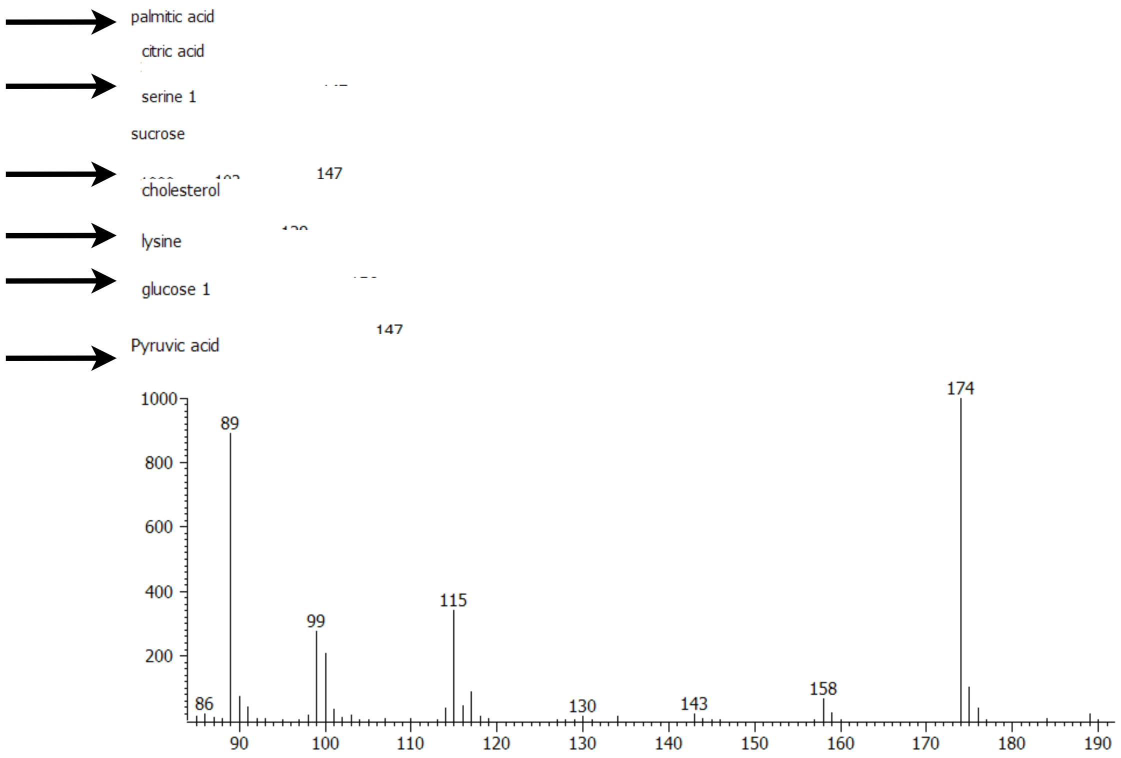
Unknown spectrum



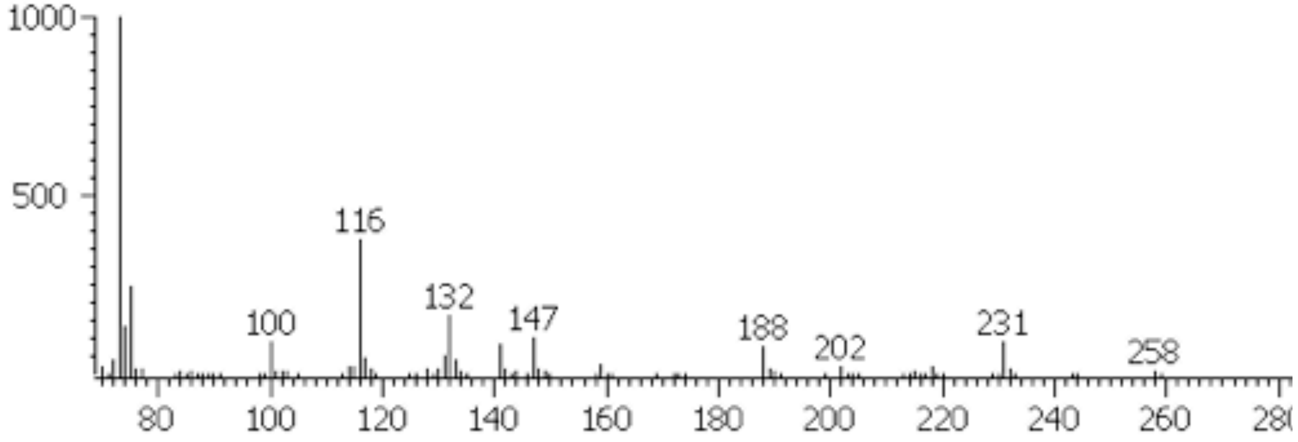
Library matching



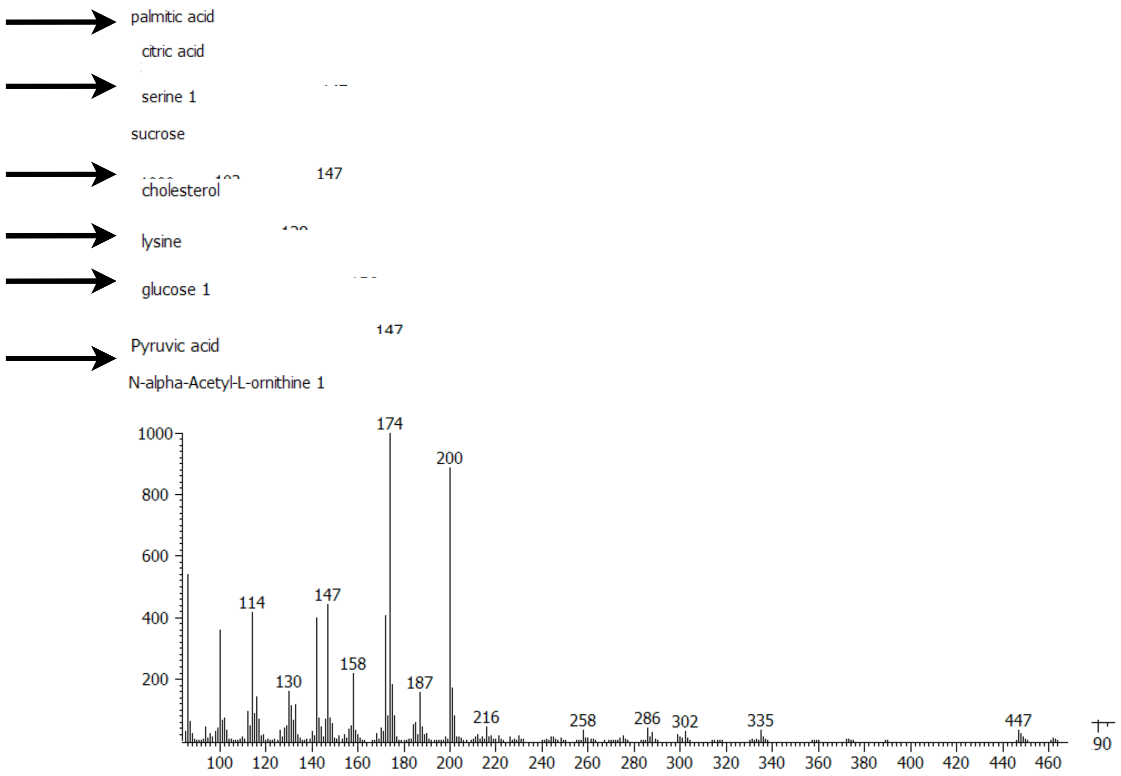
Unknown spectrum



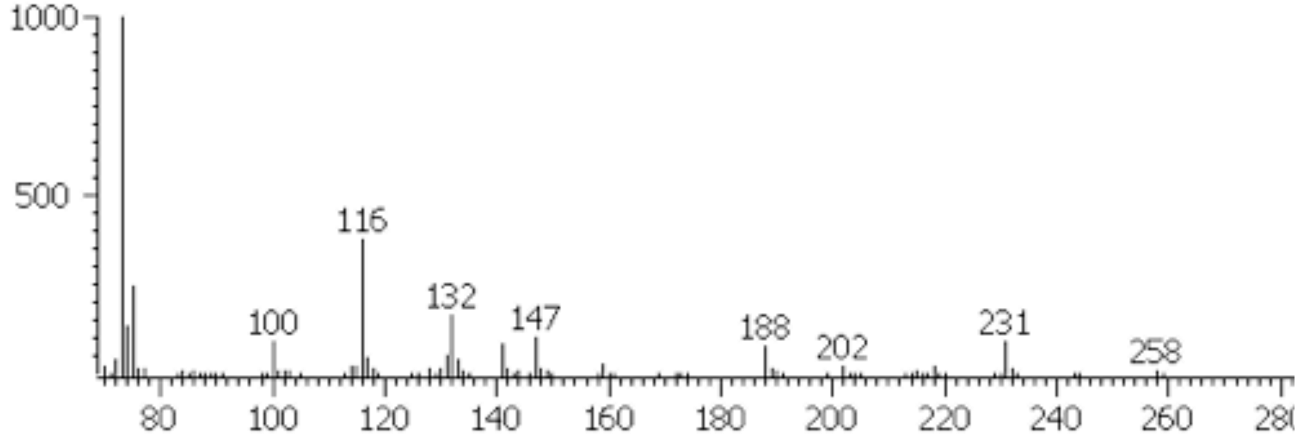
Library matching



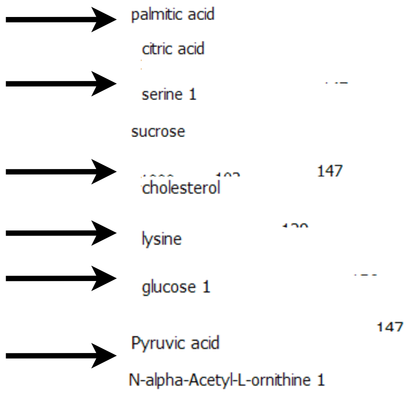
Unknown spectrum



Library matching



Unknown spectrum



palmitic acid

citric acid

serine 1

sucrose

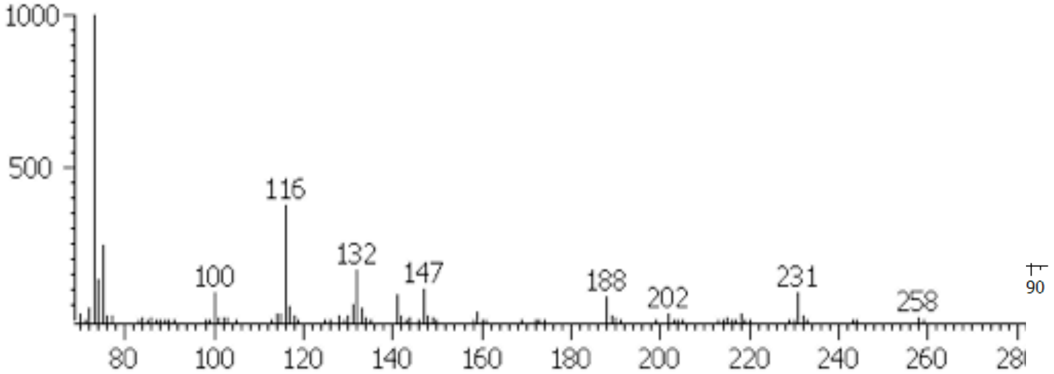
cholesterol

lysine

glucose 1

Pyruvic acid

N-alpha-Acetyl-L-ornithine 1



Asparagine

Metabolite ID advances

- Generation of publicly or commercially available databases
  - NIST
  - Golm
  - Fiehn ($)
- Metabolite structure prediction algorithms
  - Using clustering, modeling
- Improved algorithms for database searches

Why do GC-MS?

	GC	LC
Size	Small	Medium to Large
Polarity	Requires derivitization to reduce polarity	Better for polar
Metabolites	a.a., organic acids fatty acids (short-medium)	nucleotides, lipids (including large)
Chromatography	Highly reproducible- Retention indices	Less critical
Metabolite ID	Libraries	Inferred composition by accurate mass

Applications for GC-MS

- Petroleum and Biodiesel
- Biofluids and tissues
- Breath
- Pesticides
- Pollutants in air, soil and water
- Yeast for brewing and wine-making

So you’ve decided to do GC...what to expect

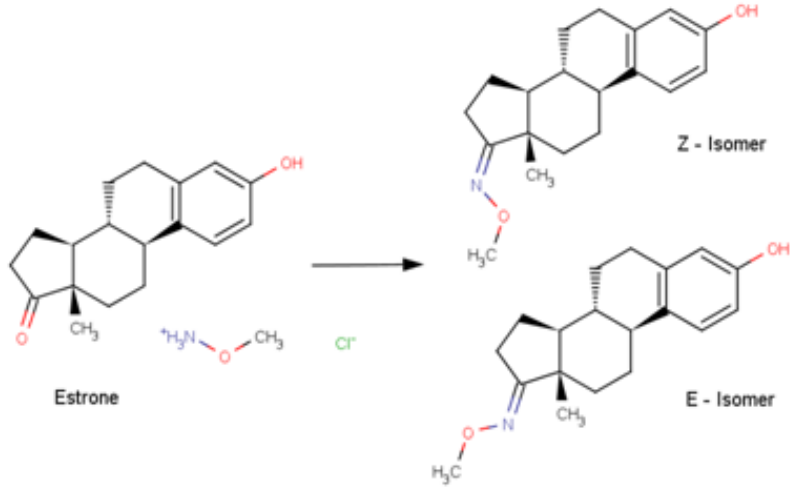
- Experimental Design!! What question(s) do you want to answer?
- Sample preparation
- Data collection
- Preliminary Data analysis
    - tools
- Metabolite identification

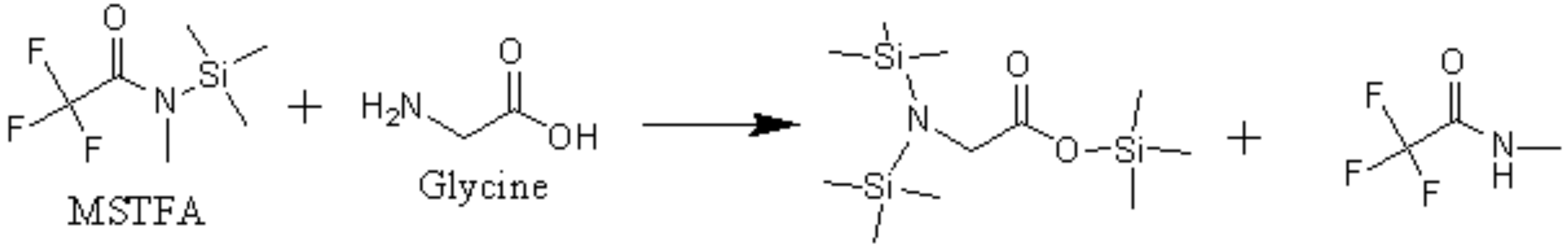
Sample procurement/preparation

- Samples should be snap frozen as quickly as possible after extraction and stored frozen until extraction
- Cultured cells should be grown in a minimal media if possible
    - Avoid conditions where there are media/solvent components are present at high concentration
        - e.g. Urine samples may be treated with urease
    - Aspiration or filteringis the best way to remove media efficiently before freezing
- Extraction should be done under cold conditions when possible

Gas Chromatography for Metabolomics

- Gas chromatography requires all analytes to be volatile
- Common procedure for biological samples is derivatization
- Most common method is methoximation + silylation
- Basic Protocol:
    - Dry all analytes by centrivap
    - Add methoxamine (stabilize ketones)
    - TMS reagent (generate volatile compounds)





Data collection

- You can expect anywhere from 500-5000 unfiltered peaks depending on extraction method, sample complexity and concentration

- Typical number of quantified metabolites found in the majority of samples (based on our typical 2D-GC protocol but it varies depending on column configuration and data collection speeds):
  - Yeast: 150-200
  - Serum: 200-250
  - Urine: 350-500
  - Tissue: 200-300

Analyzing the Data

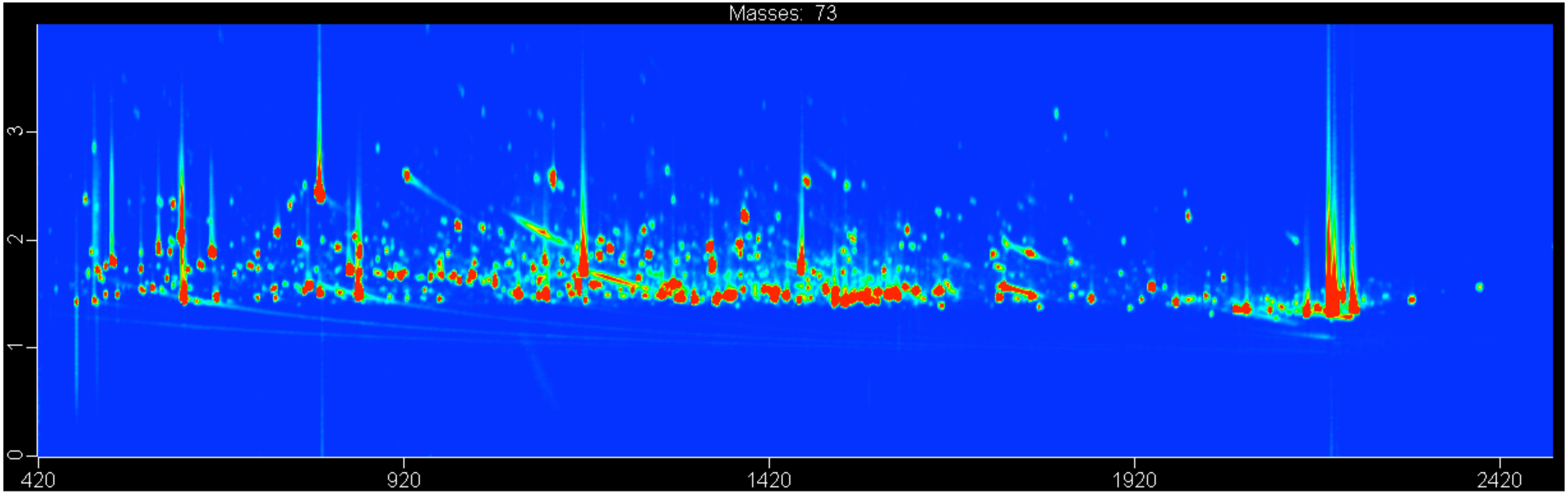
- Most instruments utilize proprietary software to do peak deconvolution
- Raw data can be analyzed as well and there are tools out there to analyze raw data (e.g. Metlin, XCMS)
- ChromaTOF (Leco’s peak calling and deconvolution software) Output:
  - List of peaks
  - Determination of Quant Mass for each peak (unique mass, typically)
  - Quantification of metabolite (either relative to reference or absolute)
  - Library Matches for Metabolite ID

Steps to analyzing Metabolomics Data

1. Filtering Peaks
2. Alignment
3. Missing Values
4. Normalization
5. Statistical Analysis

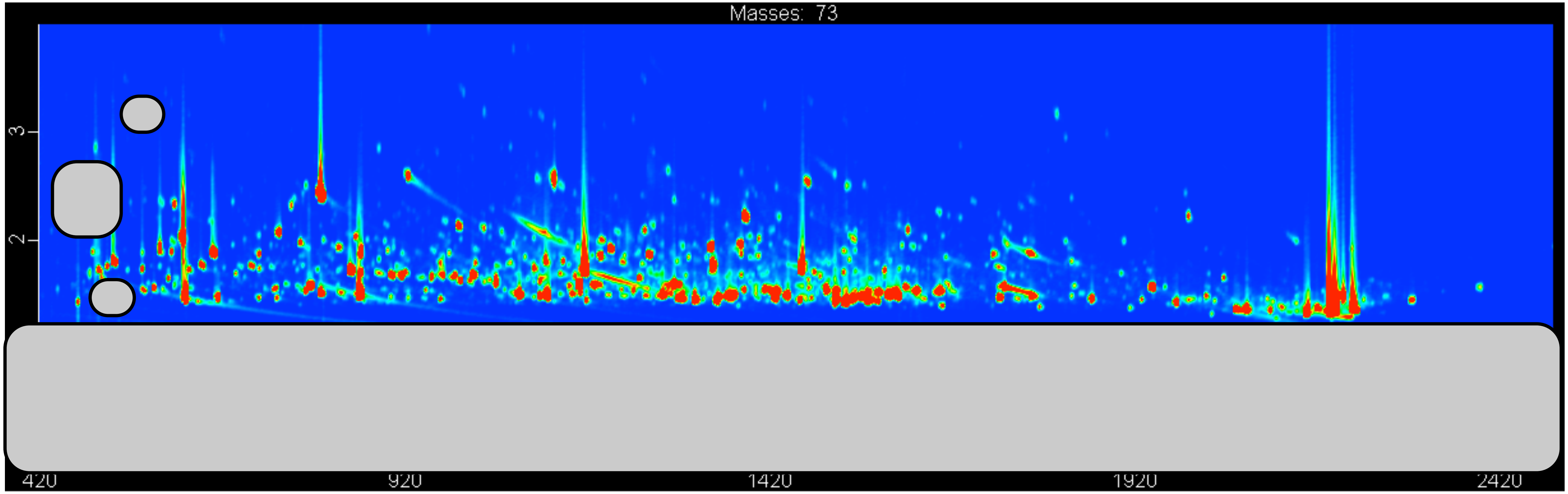
Data Analysis: Filtering

Filter peaks originating from derivitization reagents or from solvent



Data Analysis: Filtering

Filter peaks originating from derivitization reagents or from solvent



Data Analysis: Alignment

- For each sample, determine whether every measured metabolite (from every other sample) is present
- Complex, computationally intense problem
- Use all available information: Retention Index, (RT1 and RT2 for 2D-GC), and Spectral Match
    - MetPP, Guineu (2D GC) or MetAlign (e.g.) for GC
- Typical Result from high quality raw data: 200-400 peaks are present in ~80% of samples-Missing values 2-5% of data

Data Analysis: Missing Values

- Conservative Filter: only consider metabolites present in the VAST majority of the samples (~95%)

Limited to small number of metabolites (High Confidence)

- Assuming missing values are below detectable levels (0.5x lowest value for that metabolite)

Can skew results if there are a large number of missing values

- Assume missing values are present at an average or median level

Conservative, but can skew data

- K nearest neighbor estimation-characterizes what values are present in other samples with the most highly correlated values for other metabolites to estimate a likely concentration

Moderately conservative , but not possible if missing data is abundant

Data Analysis: Normalization

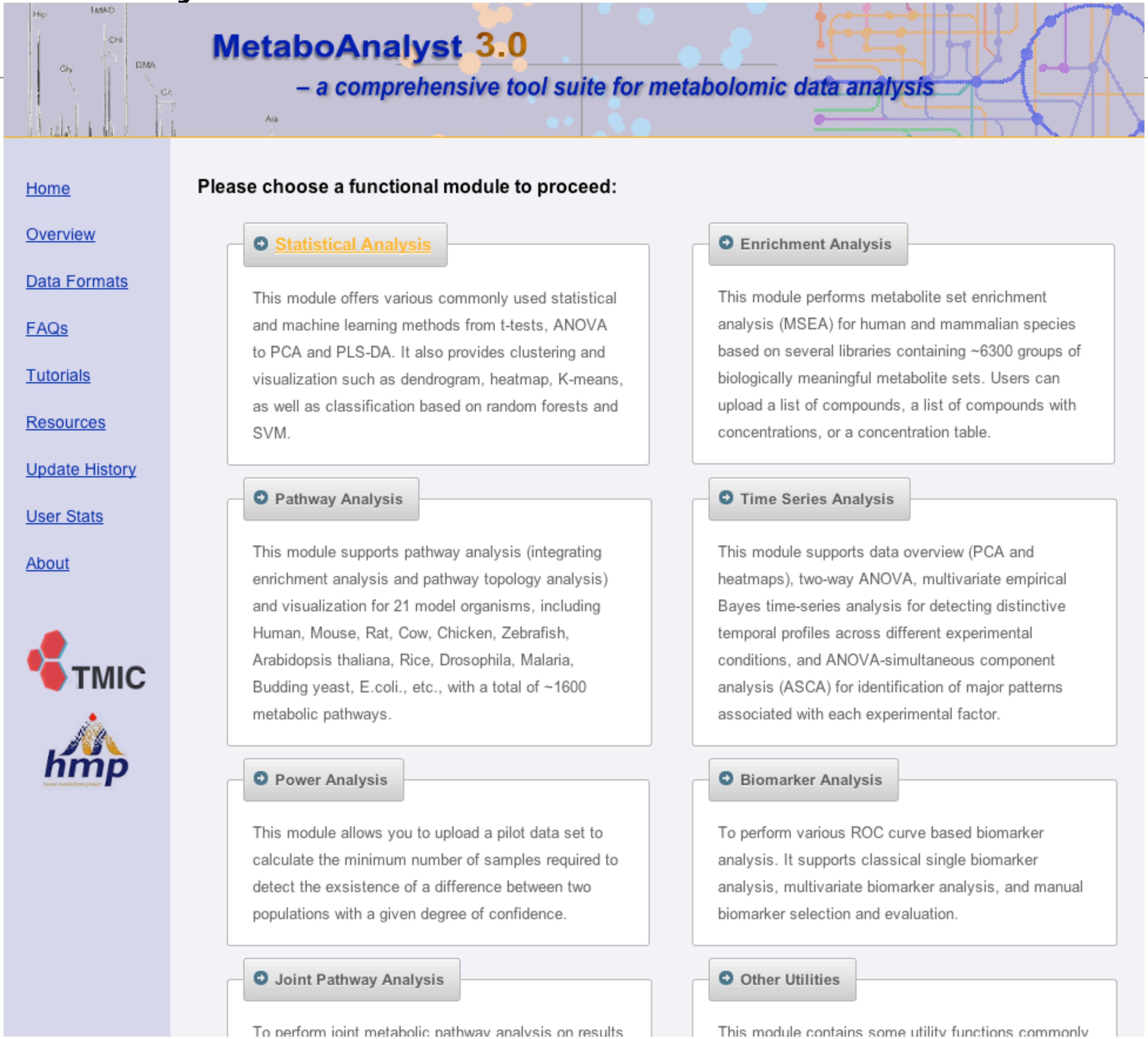
- Common Practice:
    - Injection Control (A known amount of substance is injected with each sample. Those peaks should have the same area each time)
    - Normalization by SUM (total area under the curve). Normalizes for overall sample concentration
    - Clinical samples: normalization by creatinine or other specific analytes (not ideal for research, but sometimes necessary depending on application)

Data Analysis: Statistical Analysis

- A wide variety of tools and packages available
- Metaboanalyst is a great place to start (R-package in web-based app)
  - Upload your aligned data in .csv or .txt format. It goes through the normalization, missing data and filtering steps and then allows a variety of analysis
  - Heatmaps, Clustering
  - PCA
  - PLS-DA
  - T-tests (paired and unpaired)
  - Some pathway analysis
  - etc.

www.metaboanalyst.ca

Metaboanalyst



MetaboAnalyst 3.0

– a comprehensive tool suite for metabolomic data analysis

Please choose a functional module to proceed:

Statistical Analysis

This module offers various commonly used statistical and machine learning methods from t-tests, ANOVA to PCA and PLS-DA. It also provides clustering and visualization such as dendrogram, heatmap, K-means, as well as classification based on random forests and SVM.

Enrichment Analysis

This module performs metabolite set enrichment analysis (MSEA) for human and mammalian species based on several libraries containing ~6300 groups of biologically meaningful metabolite sets. Users can upload a list of compounds, a list of compounds with concentrations, or a concentration table.

Pathway Analysis

This module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 21 model organisms, including Human, Mouse, Rat, Cow, Chicken, Zebrafish, Arabidopsis thaliana, Rice, Drosophila, Malaria, Budding yeast, E.coli., etc., with a total of ~1600 metabolic pathways.

Time Series Analysis

This module supports data overview (PCA and heatmaps), two-way ANOVA, multivariate empirical Bayes time-series analysis for detecting distinctive temporal profiles across different experimental conditions, and ANOVA-simultaneous component analysis (ASCA) for identification of major patterns associated with each experimental factor.

Power Analysis

This module allows you to upload a pilot data set to calculate the minimum number of samples required to detect the exsistence of a difference between two populations with a given degree of confidence.

Biomarker Analysis

To perform various ROC curve based biomarker analysis. It supports classical single biomarker analysis, multivariate biomarker analysis, and manual biomarker selection and evaluation.

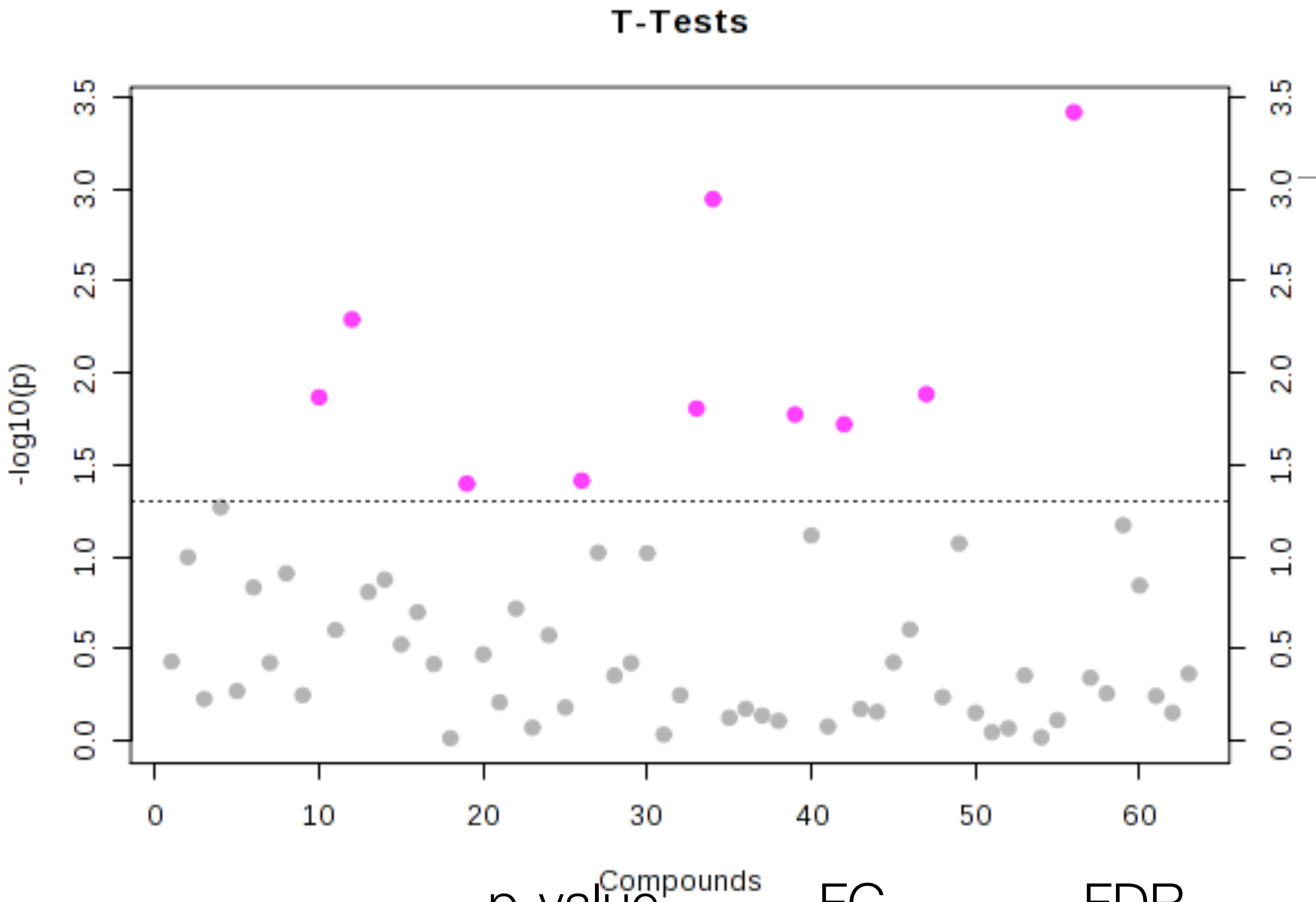
Joint Pathway Analysis

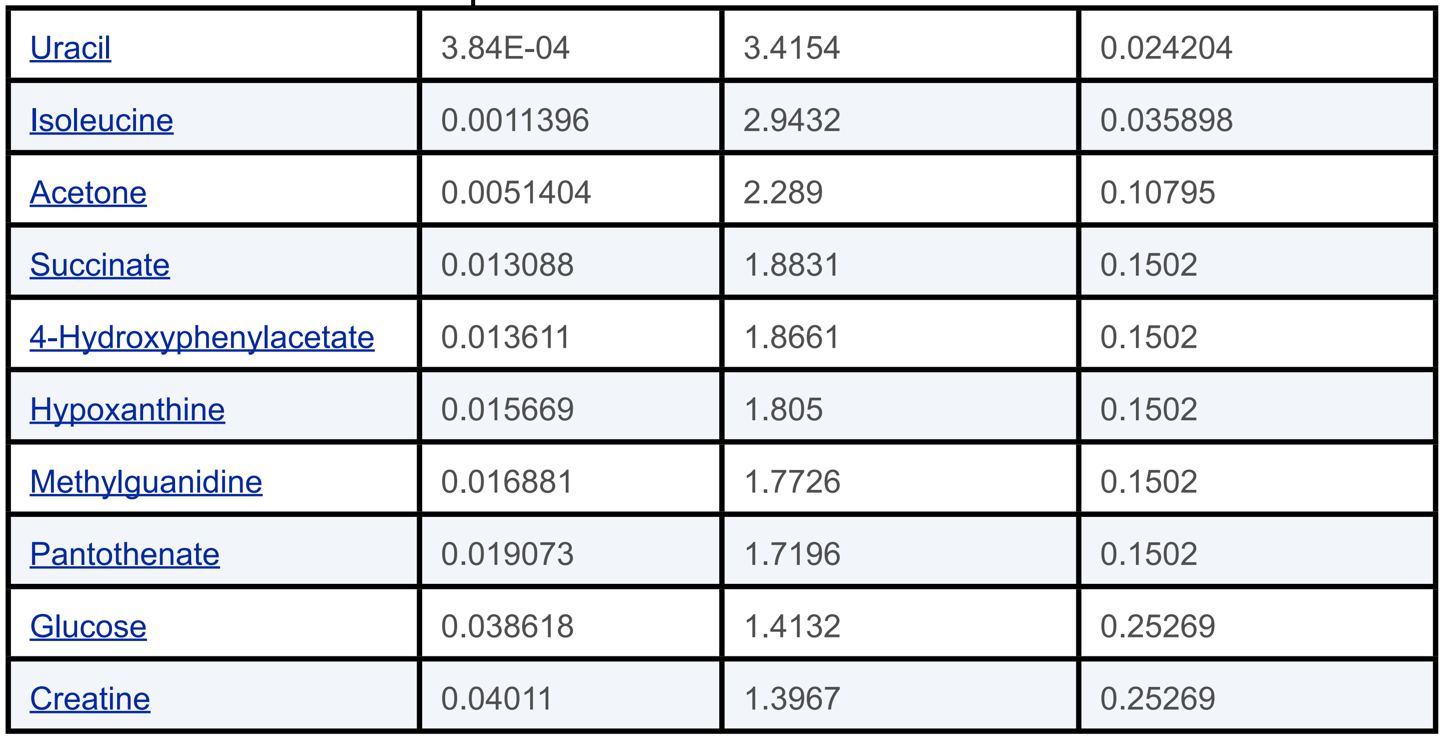
To perform joint metabolic pathway analysis on results

Other Utilities

This module contains some utility functions commonly

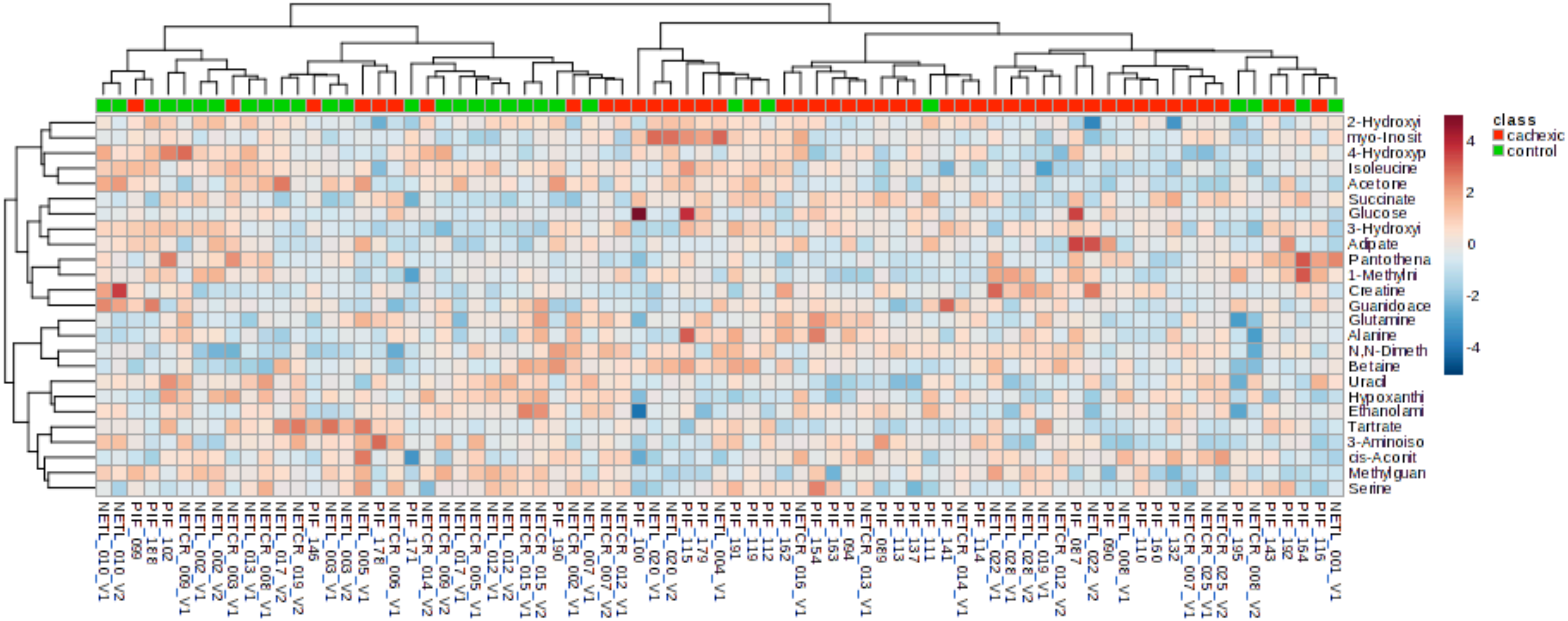
Input test dataset (Cancer patients Cachexic v. control)



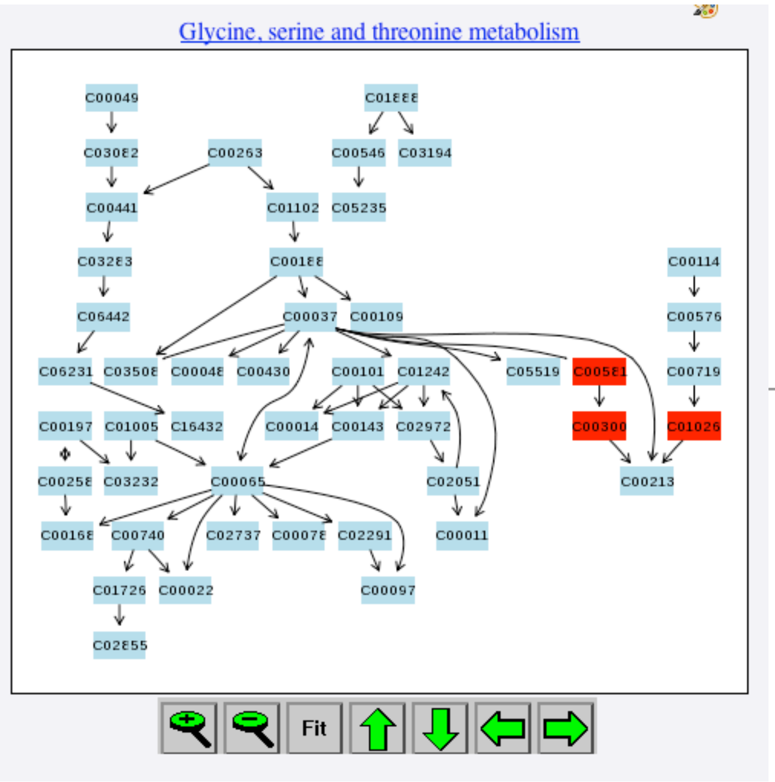


	p-value	FC	FDR
Uracil	3.84E-04	3.4154	0.024204
Isoleucine	0.0011396	2.9432	0.035898
Acetone	0.0051404	2.289	0.10795
Succinate	0.013088	1.8831	0.1502
4-Hydroxyphenylacetate	0.013611	1.8661	0.1502
Hypoxanthine	0.015669	1.805	0.1502
Methylguanidine	0.016881	1.7726	0.1502
Pantothenate	0.019073	1.7196	0.1502
Glucose	0.038618	1.4132	0.25269
Creatine	0.04011	1.3967	0.25269

Sample Data-top25 features by Ttest

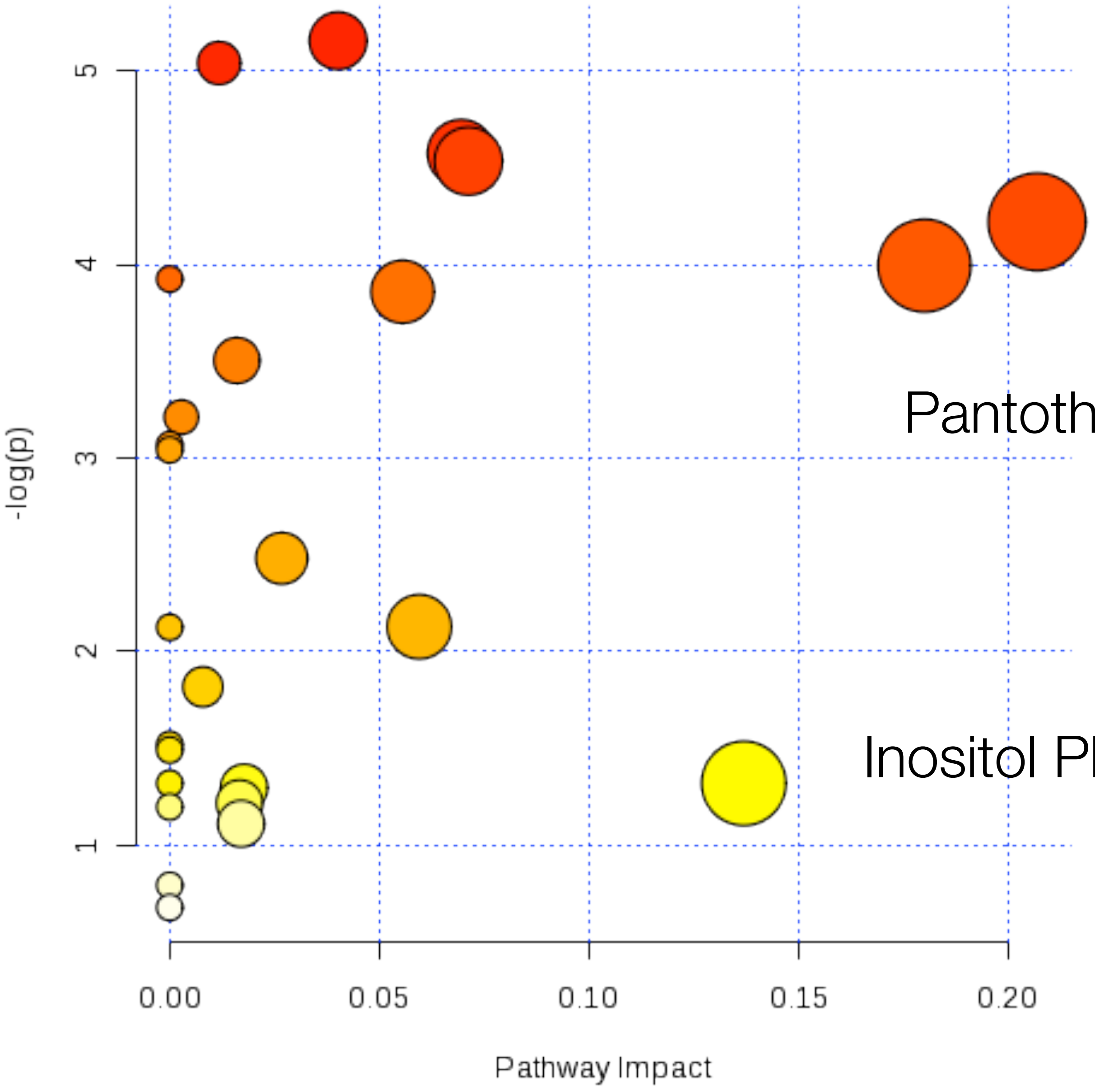


Pathway Analysis



Glycine, serine and threonine metabolism

Glycine, Serine, Threonine



Alanine/Aspartate

Pantothenate and CoA

Inositol Phosphate

Data Analysis: Biological Understanding

- Web-based tools for pathway analysis
  - KEGG (KEGGMapper) (all organisms)
  - HMDB (Human Metabolome Database)
    - Serum, urine, metabolome databases
  - Yeast- Biochemical Pathways at yeastgenome.org
    - ymdb (yeast metabolome database)
- Integrated analysis with genomic, proteomic data
  - IMPaLA (similar to GO enrichment but specific to metabolic pathways)
  - Ingenuity ($$$)
  - Metaboanalyst (new)

Resources for GC-MS

- Restek Column Selection guide www.restek.com/
  - http://www.restek.com/pdfs/GNBR1724-UNV.pdf
- Leco
- Agilent
- Sigma https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Aldrich/Bulletin/1/the-basics-of-gc.pdf
- Books,Chapters, Reviews:
  - *Metabolomics* by Wofram Weckwerth (Methods and Protocols)
  - “Mass Spectrometry based metabolomics” Dettmer 2007 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1904337/
- Analysis
  - Metaboanalyst.ca
  - impala.molgen.mpg.de
  - hmdb.ca
  - golm database: gmd.mpimp-golmmpg.de
  - metlin.scripps.edu
  - xcmsonline.scripps.edu

Questions???

Thank you