Metabolomics

Kim Kultima, Assoc. Prof.
Uppsala University
Uppsala University Hospital
Karolinska Institutet

Uppsala University Hospital, Sweden

CARAMBA

CARAMBA- **C**linical **A**nalysis & **R**esearch **A**pplying **M**ass spectrometry & **B**ioinformatics at **A**kademiska

- Joint venture Uppsala University and Uppsala University Hospital
- The facility is certified (ISO 15189) enabling us to deliver results for clinical care

Clinical metabolomics and proteomics

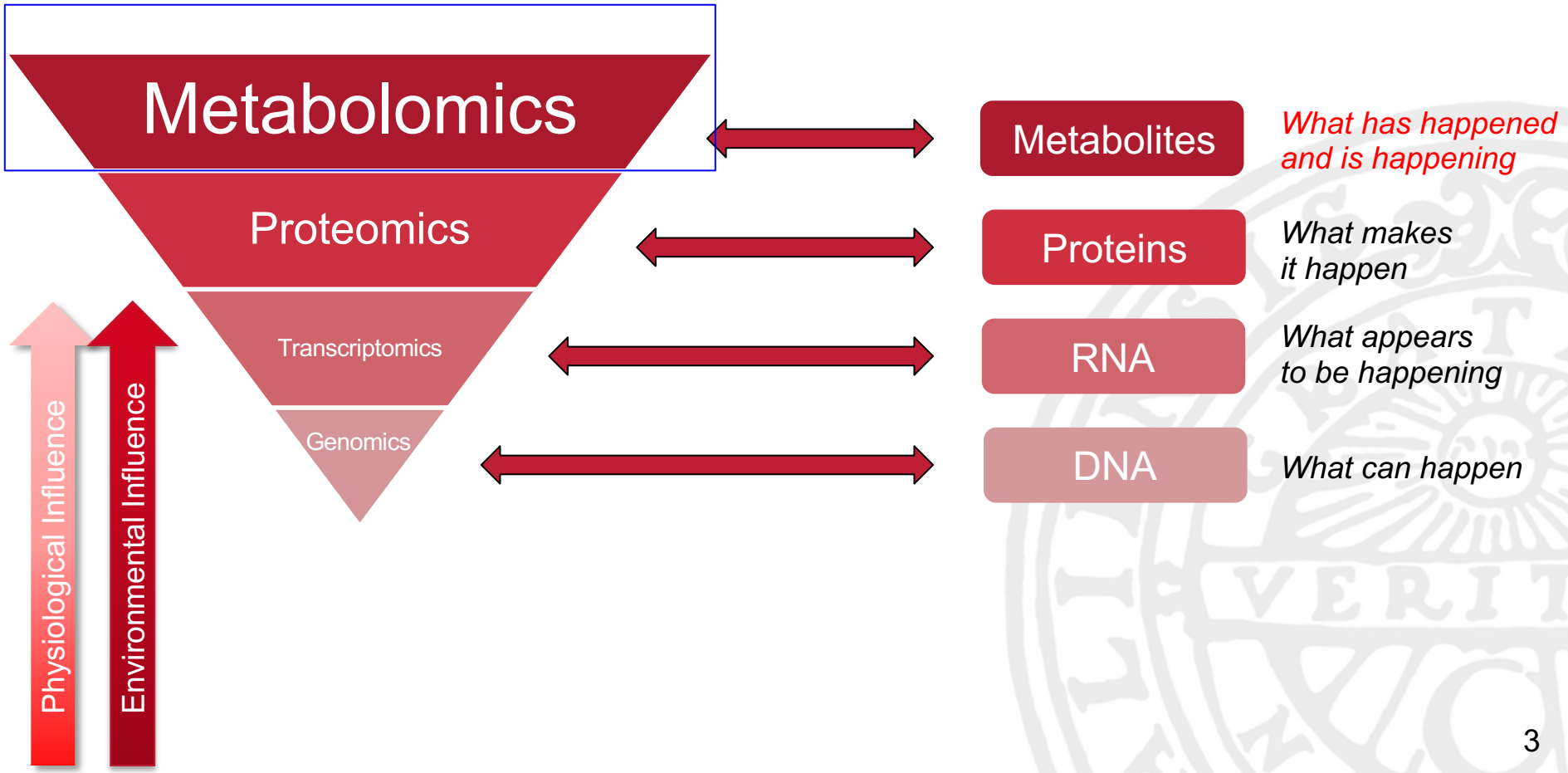
- ***Multiple sclerosis*****,** Huntington's disease, Alzheimer's disease and ***chronic pain diseases***

Uppsala University Hospital, Sweden

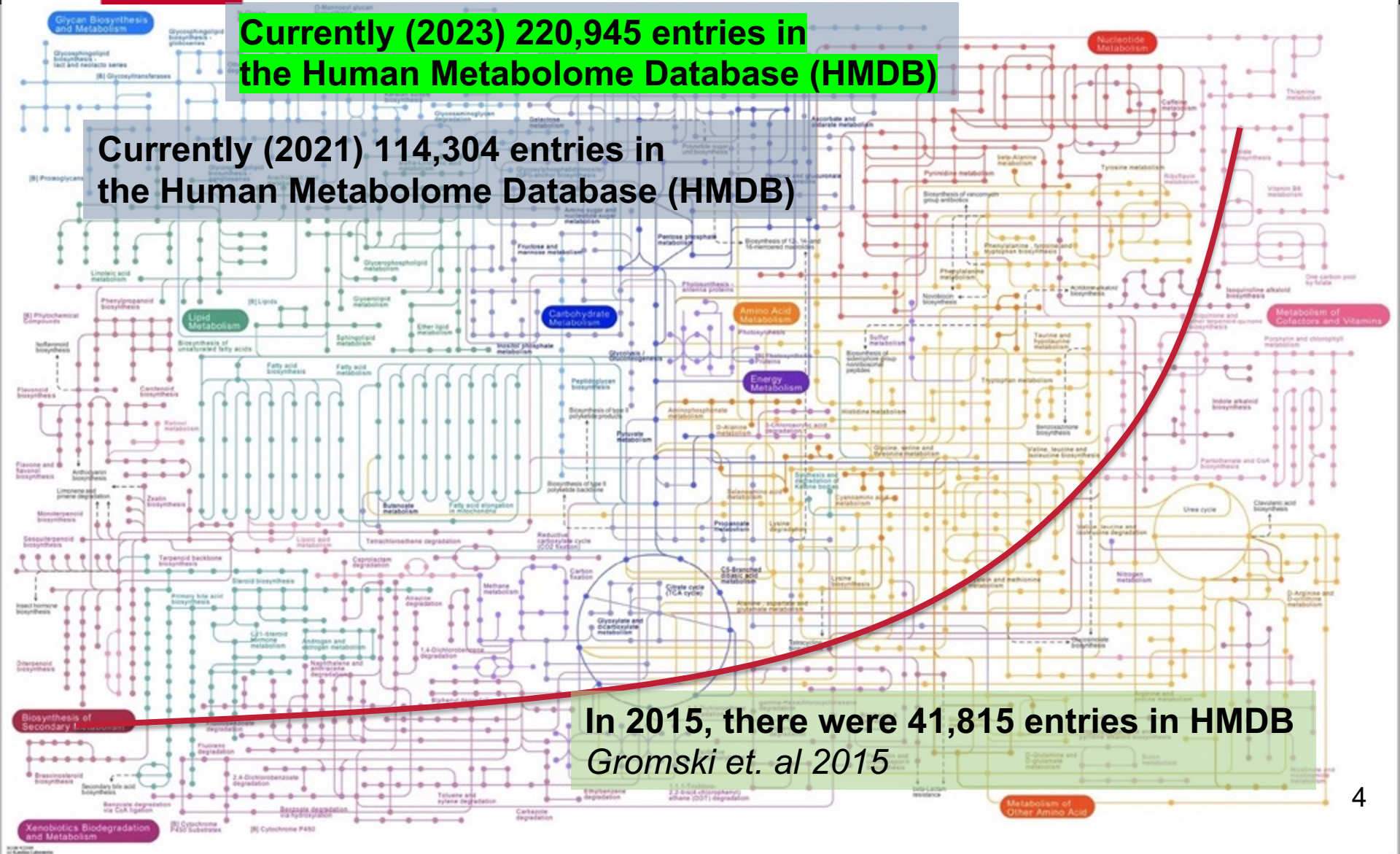
Shared laboratory for clinical- and research use only using HR-MS

Metabolomics

The study of small molecules, typically <1500 Da in size



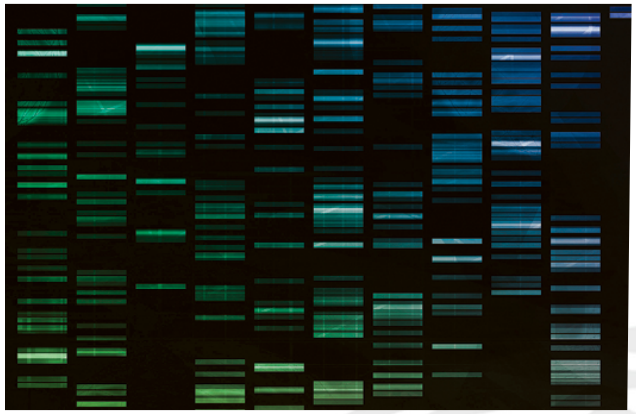
Metabolomics



Big Data in Metabolomics - Today

Big data, big picture: Metabolomics meets systems biology

Metabolomics—the study of the collection of an organism's metabolites—provides a molecular measurement of phenotype, or the characteristics resulting from the genotype's interaction with the environment. Using a range of analytical tools to scale the mountains of data collected, including molecular detection and bioinformatics, scientists use metabolomics to understand systems biology, which is the complete computational analysis and modeling of an organism and its well-being. **By Mike May**



"Metabolomics can't really function on its own without the genome sequence." — Jonas Korlach, chief scientific officer at Pacific Biosciences

Despite all of the advances in storing and analyzing data, scientists are still confronting significant obstacles in studying the metabolome. "One bottleneck in nontargeted workflows is the identification of unknown compounds," says Aiko Barsch, market manager for metabolomics at **Bruker Daltonics,** based in Bremen, Germany. "This is where MS and NMR [mass spectrometry and nuclear magnetic resonance] both have advantages."

UPPSALA UNIVERSITET

Big Data in Metabolomics - In health

Predicting human health from biofluid-based metabolomics using machine learning

Ethan D. Evans[1], Claire Duvallet[1,3], Nathaniel D. Chu[1], Michael K. Oberst[2], Michael A. Murphy[1,2], Isaac Rockafellow[1,4], David Sontag[2]✉ & Eric J. Alm[1]✉

Biofluid-based metabolomics has the potential to provide highly accurate, minimally invasive diagnostics. Metabolomics studies using mass spectrometry typically reduce the high-dimensional data to only a small number of statistically significant features, that are often chemically identified—where each feature corresponds to a mass-to-charge ratio, retention time, and intensity. This practice may remove a substantial amount of predictive signal. To test the utility of the complete feature set, we train machine learning models for health state-prediction in 35 human metabolomics studies, representing 148 individual data sets. Models trained with all features outperform those using only significant features and frequently provide high predictive performance across nine health state categories, despite disparate experimental and disease contexts. Using only non-significant features it is still often possible to train models and achieve high predictive performance, suggesting useful predictive signal. This work highlights the potential for health state diagnostics using all metabolomics features with data-driven analysis.

Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine

Dmitry Grapov,[1] Johannes Fahrmann,[2,*] Kwanjeera Wanichthanarak,[3,4,*] and Sakda Khoomrung[3,4]

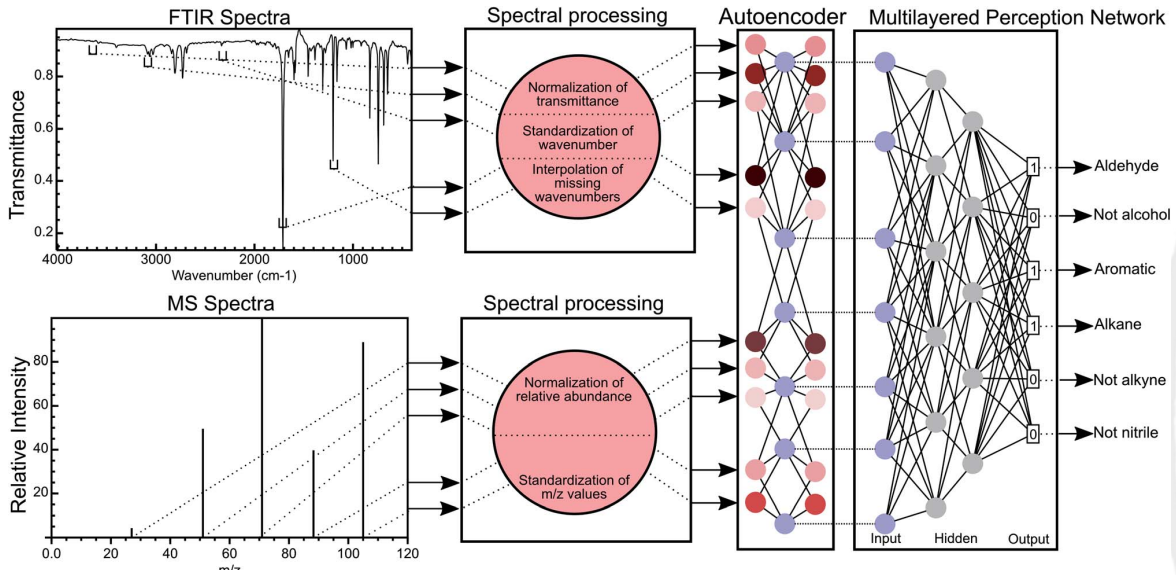
Big Data in Metabolomics - Out reach

Spectral deep learning for prediction and prospective validation of functional groups†

Jonathan A. Fine, ‡[a] Anand A. Rajasekar, ‡[b] Krupal P. Jethava [a] and Gaurav Chopra *[a]

Deep learning to generate *in silico* chemical property libraries and candidate molecules for small molecule identification in complex samples

Sean M. Colby, Jamie R. Nuñez, Nathan O. Hodas, Courtney D. Corley, Ryan R. Renslow*
Pacific Northwest National Laboratory, Richland, WA, USA.
* ryan.renslow@pnnl.gov



Analytical platforms for metabolomics

- Metabolites have a vast range of chemical structures, properties and concentrations
- No single platform provides complete comprehensive coverage
- No single extraction or analysis method works for all metabolites
- Selection of the platform is always a compromise between sensitivity, speed and chemical selectivity and coverage

Common techniques used in metabolite profiling studies:

- NMR
- MS

  Hyphenated techniques

  - GC-MS
  - CE-MS
  - LC-MS

Big Data in metabolomics- Challenges

- The rapid progress in field has resulted in a mosaic of independent, and sometimes incompatible, analysis methods that are difficult to connect into a useful and complete data analysis solution
- Configuring necessary software tools and chaining them together into a complete re-runnable analysis is challenging

NMR vs MS-based Methods

NMR (Nuclear Magnetic Resonance)

- Untargeted approach

MS (Mass spectrometry)

- Targeted or untargeted

Untargeted (global) approach

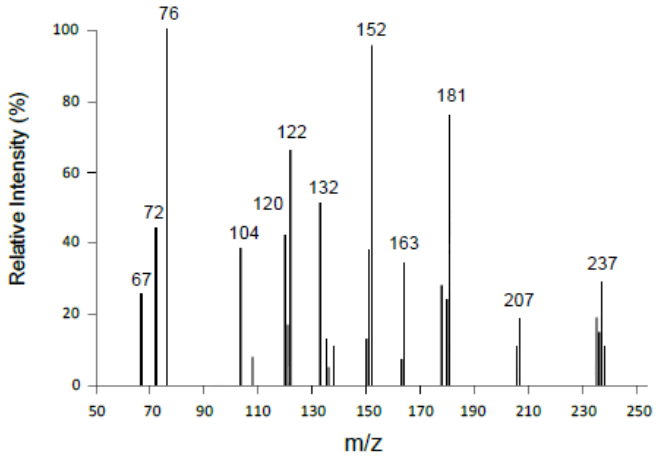
- Measures as many metabolites as possible from a range of biological samples

Targeted (specific) approach

- To measure sets of metabolites when you have a specific biochemical question you want to answer

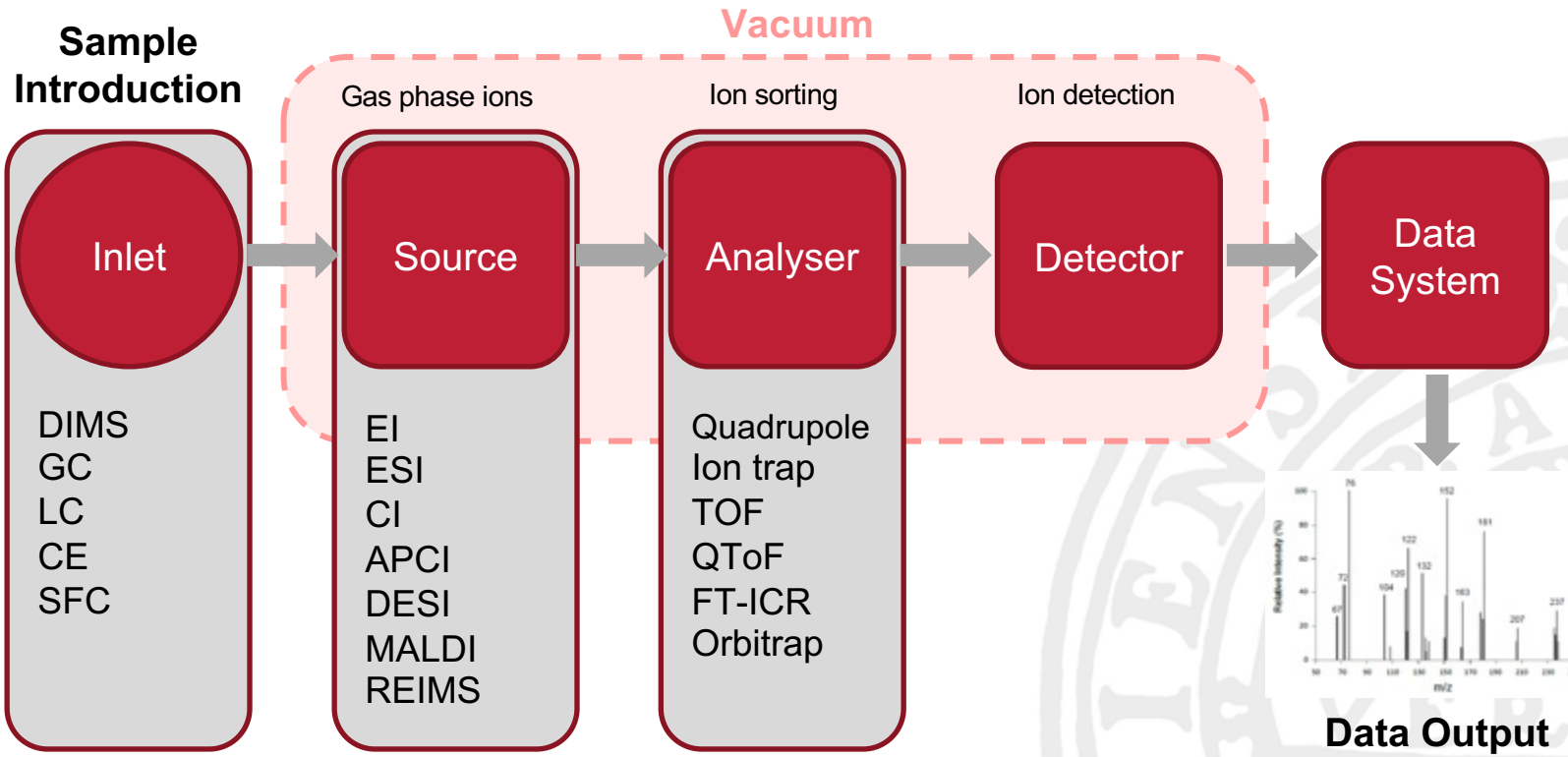
Mass spectrometry

- A tool for ionising molecules
- Sorts the ions based on their mass-to-charge (m/z) ratio
- In simple terms it measures the masses and abundance within a sample
- Qualitative/quantitative detection of molecular ions



Typical mass spectreum

Mass spectrometry



UPPSALA UNIVERSITET

What is Electrospray Ionisation?

- A technique used to produce ions
- A techniques used to transfer ions from the solution phase into the gas phase
- ESI is a soft ionisation process
- Generates a molecular ion
  - [M+H]+ or [M-H]-
  - In metabolomics it can also be [M+Na]+, [M+K]+, [M+NH4]+ or [M-Cl]-
- Both positive and negative ion data are typically collected

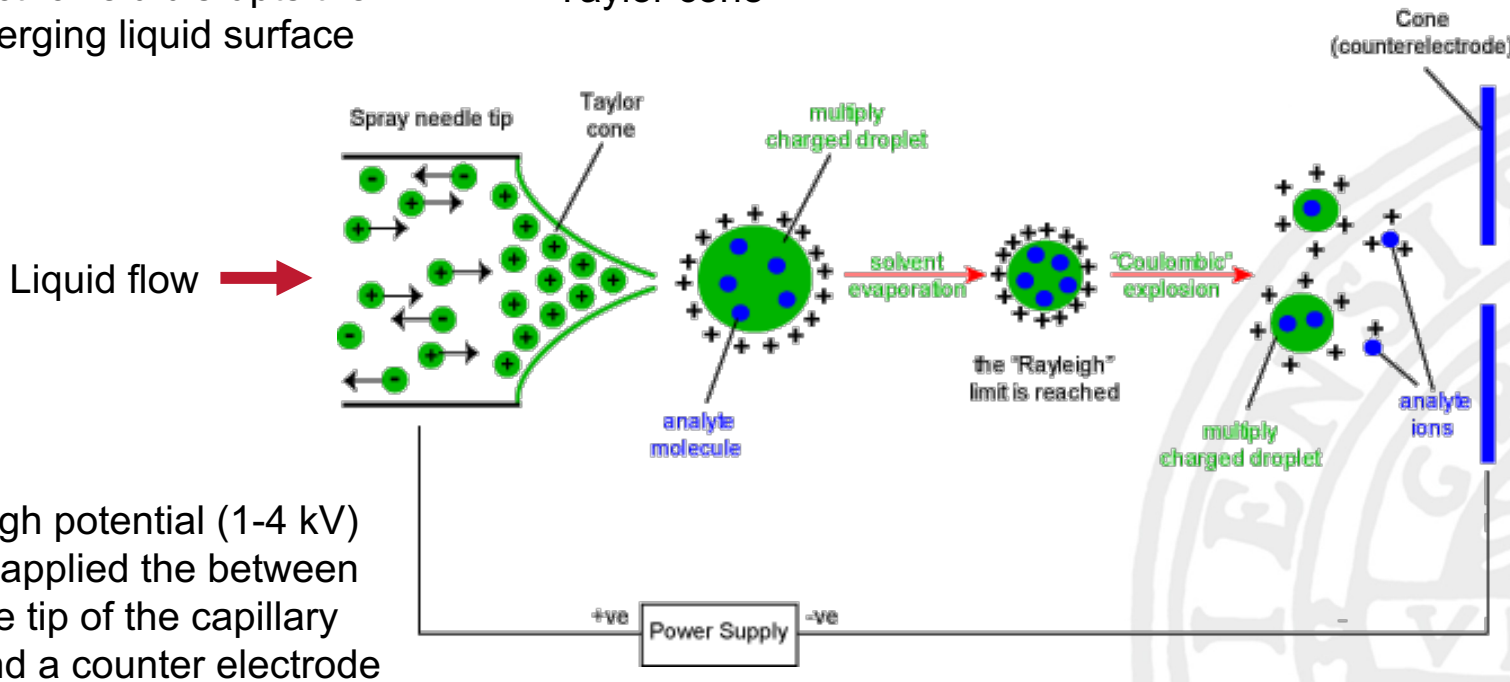
Electrospray Ionisation (ESI) - Theory

+ve potential on capillary → +ve ions

-ve potential on capillary → -ve ions

As the eluent passes through the capillary the electric field disrupts the emerging liquid surface

Eluent protrudes from the end of the capillary as a Taylor cone



High potential (1-4 kV) is applied the between the tip of the capillary and a counter electrode

The droplets move towards the orifice of the reduced pressure regions of the mass spectrometer

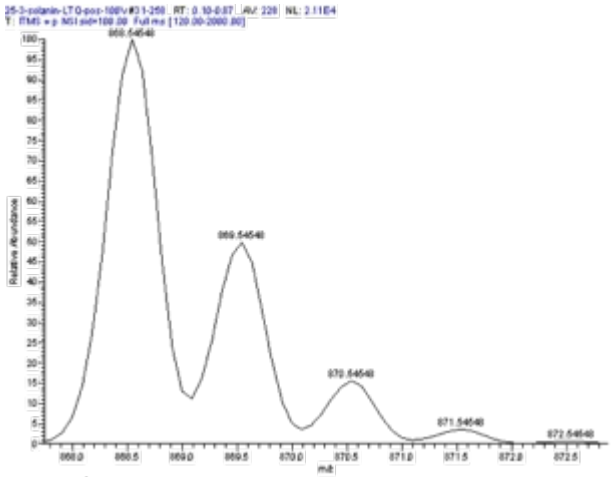
Mass analyser

Many different instrument types and configurations:

- Low mass resolution, e.g. quadrupole
- High mass resolution, e.g. quadrupole-time-of-flight (QToF), Orbitrap
- Untargeted or targeted approach?
  - Untargeted approaches are exploratory and rely on high resolution accurate mass instruments for the identification of features of interest
  - A targeted approach requires you to know what you want to measure
    - Principally based on nominal mass instruments using MS/MS methods
      - up till 2015, typically used in routine laboratories for assays
      - Now we also use high mass resolution instruments in the clinic

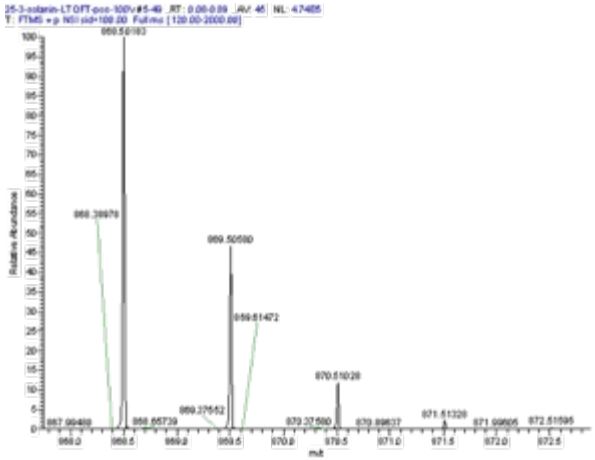
Mass resolution

Low resolution



LTQ, resolving power R = 1737 at m/z 868.5, peak width ~ 0.5 FWHM

High resolution



LTQ-FT, resolving power R = 48,250 at m/z 868.5, peak width ~ 0.018 FWHM

http://fiehnlab.ucdavis.edu/projects/seven-golden-rules/mass-resolution

Mass accuracy

- Mass accuracy = degree of conformity of a measured quantity to its actual value
  - Typically reported in ppm:

$$ppm = \left(\frac{m_{exp} - m_{obs}}{m_{obs}}\right) \text{ x } 10^6$$

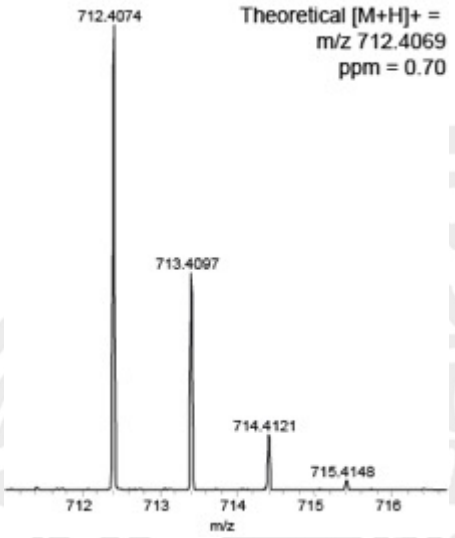
where m_{exp} = theoretical mass and m_{obs} is the measured mass

e.g. Theoretic mass = 500.0025
Measured mass = 500.0000

$$ppm = \left(\frac{500.0025 - 500.0000}{500.0000}\right) \text{ x } 10^6 = 5 \text{ ppm}$$

Mass accuracy

- Accurate mass measurements take advantage of the fact the the combination of elements contained in a molecule actually have a very specific, non-nominal molecular weight
  - Carbon has a mass of 12.0000
  - Hydrogen has mass of 1.0078
  - Oxygen has a mass of 15.9949
  - Nitrogen has a mass of 14.0031



- It is possible to have combination of atoms which have the same nominal (integer) mass but different accuracy mass

- If such compounds can be measured with sufficient mass accuracy it is possible to determine the elemental composition

High resolution mass analysers

- For untargeted analysis it is important to have high mass **resolution**, **accuracy** and **speed**

- When high resolving power is needed:
  - Time-of-Flight (ToF)
  - High resolution ion traps
    - Orbitrap
    - Fourier transform Ion Cyclotron Resonance (FT/ICR)

UPPSALA UNIVERSITET

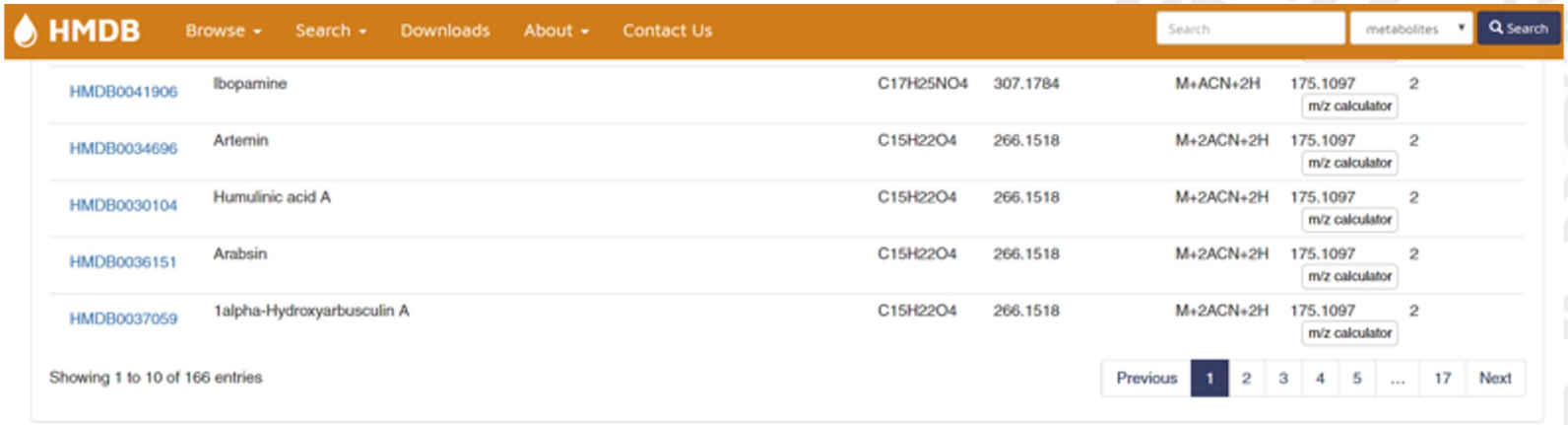
Resolution & Mass Accuracy

Type	Resolving Power (FWHM)	Mass Accuracy (ppm)
FT-ICR-MS	1,000,000	0.1 - 1
Orbitrap	100,000	0.5 - 1
High-Res-TOF	60,000	3 - 5
ToF	10,000	3 - 5
Triple Quadrupole	1,000	3 - 5
Ion Trap	1,000	50 - 200

Uniqueness of molecular ions

- The molecular ion, even when measured with the highest accuracy, is not a unique descriptor
- There are many theoretical possible structures for a given mass and empirical formula

Searched arginine m/z: 174.1116 with 10 ppm tolerance



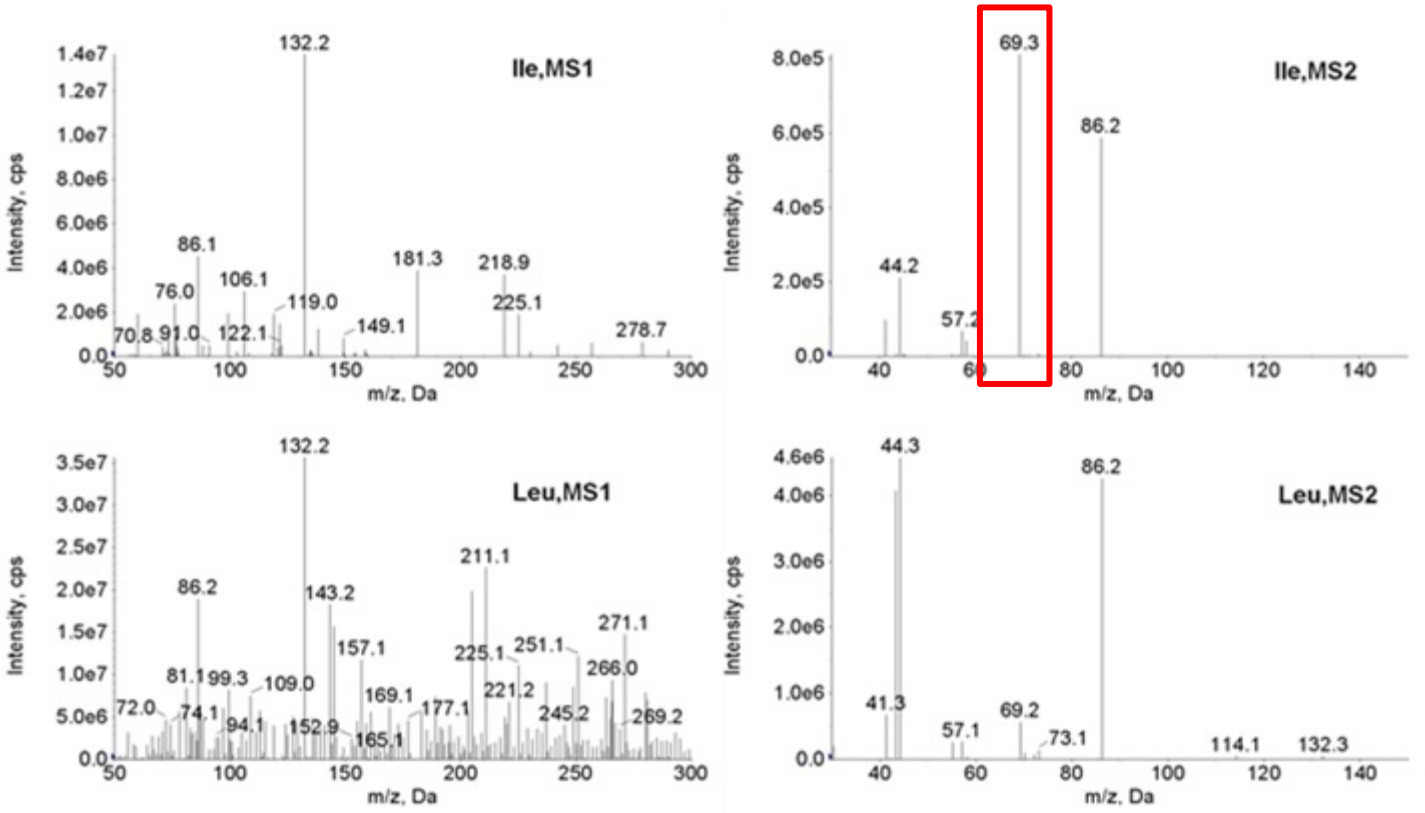
Tandem MS/MS

- A second mass spectrum (MS/MS) that is informative arises from isolating the molecular ion
- The molecular ion is heated, either by collision with neutral gas (collision induced dissociation (CID) – quadrupole, ion traps) or Higher-energy collisional dissociation (HCD) (Orbitraps)
- The extra energy increases the stretching of critical bonds, leading to dissociation of the molecular precursors ion into charged product ions
  - These generate the MS/MS spectrum for a metabolite
  - Ion traps can also isolate a product ion and create MSn spectra

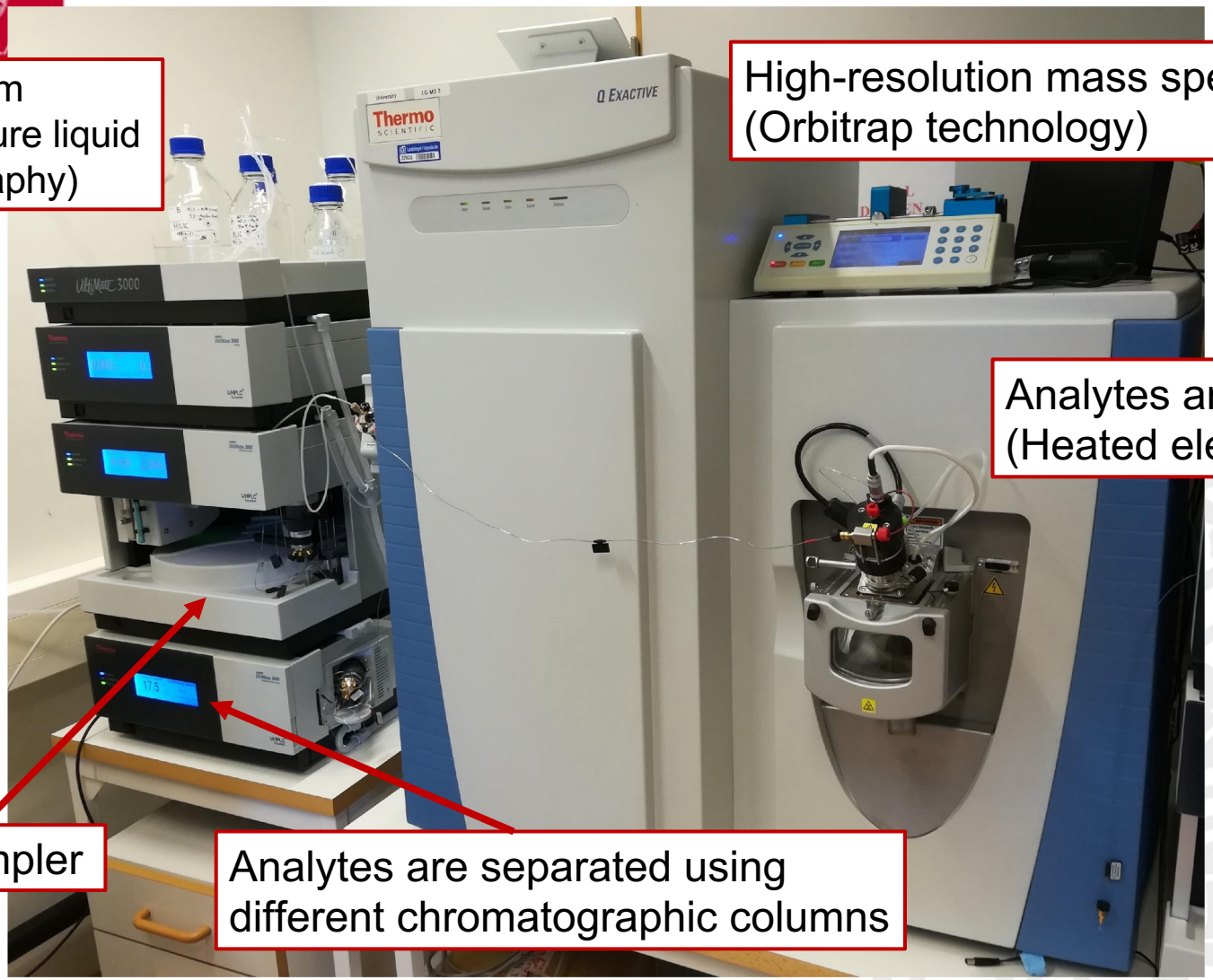
Fragmentation (MS/MS)

leucine and isoleucine are isobaric compounds
- Exact same mass

Unique fragment pattern enables discrimination between leucine and isoleucine



Yang R et al., PLoS One 8(12), 2013



HPLC system (High-pressure liquid chromatography)

High-resolution mass spectrometry (Orbitrap technology)

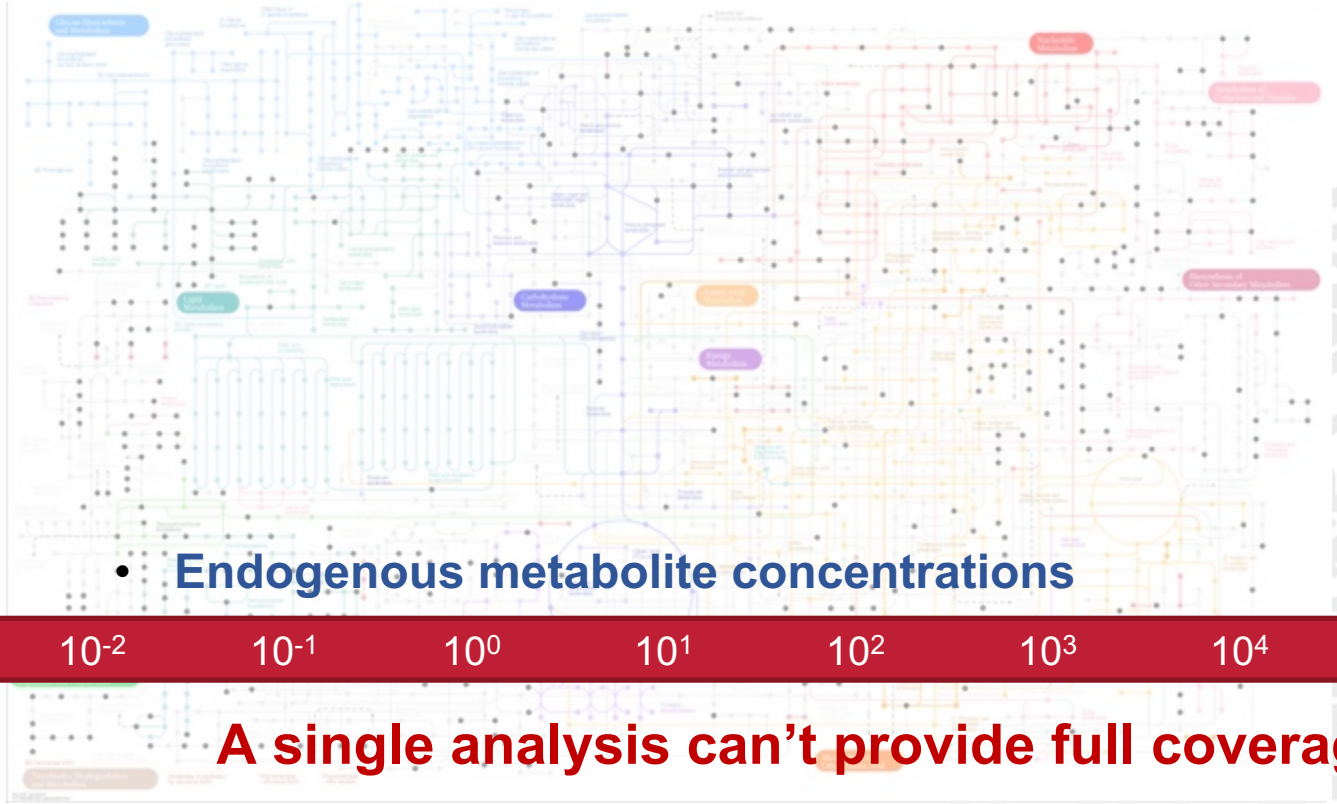
Analytes are ionized (Heated electrospray)

Autosampler

Analytes are separated using different chromatographic columns

Challenges in Metabolomics

Differences in physiochemical properties among metabolites



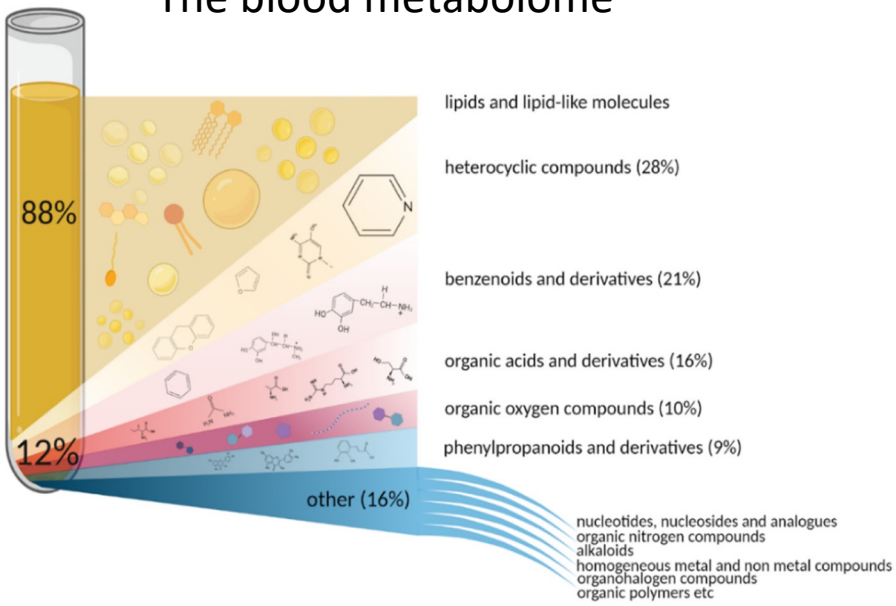
- **Endogenous metabolite concentrations**

10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5}

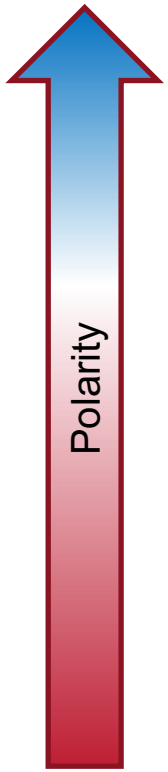
A single analysis can't provide full coverage

Complex sample – separation of metabolites needs different separation chemistry

The blood metabolome



- Lipid-soluble molecules account for 88% of the metabolome
- Non-lipid metabolites account for only about 12%
- 220,945 molecules in the Human Metabolome Database (HMDB)



HILIC

HILIC as a separation technique is the strong retention of polar, hydrophilic compounds

Reversed-phase C18

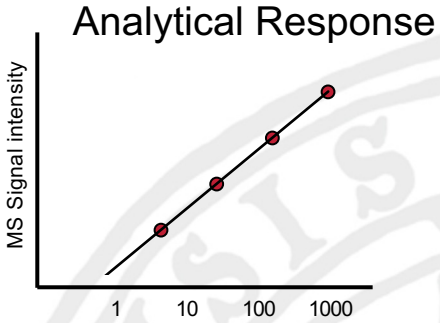
Reversed-phase C8

Reversed-phase as a separation technique is the strong retention of nonpolar, hydrophobic compounds

Mass spectrometry based metabolomics

Sensitivity dependent on analyte and ionisation

Untargeted profiling
100s -1000s
metabolites
Typical measurement over 4 orders of magnitude



Endogenous metabolite concentrations

10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5}

A single analysis can't provide full coverage

UPPSALA UNIVERSITET

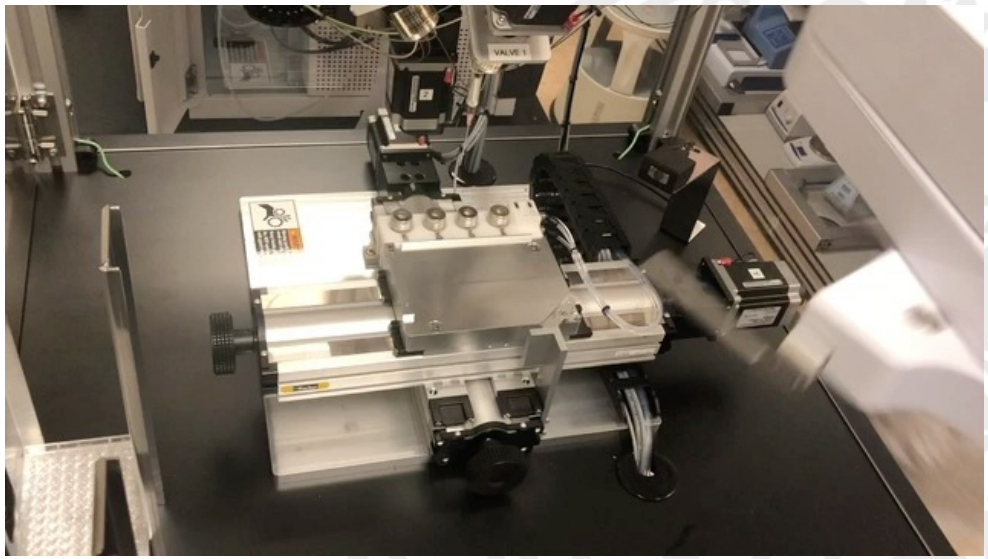
Direct infusion MS (DIMS)

- High-throughput screening
- Crude sample extracts injected or infused into the mass spectrometer
- Coverage depends on ability of the metabolite to be ionized
- Ion suppression
- Not quantitative
- Unable to distinguish isomers

Direct infusion MS (DIMS)

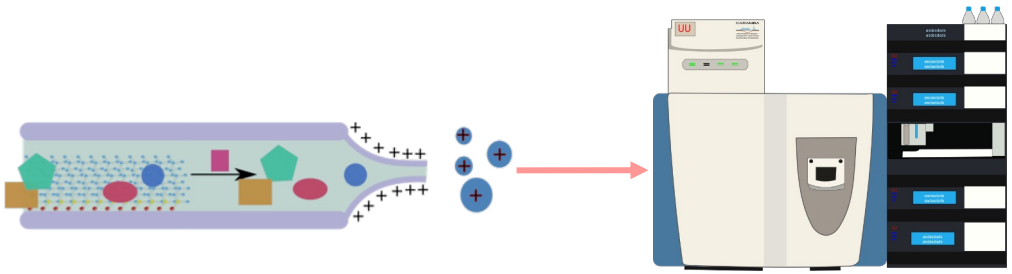
More and more vendors are inventing mass throughput instruments
- Solid Phase Extraction based- High-throughput sampling
- SPE-IMS-MS (Agilent RapidFire + 6560 Ion mobility QTOF)

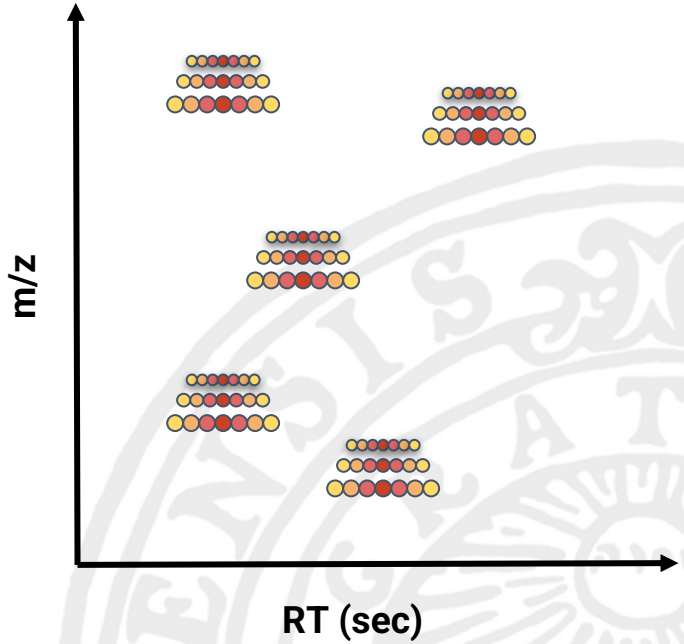
- 2TB of data in 10 minutes (96 samples)
- 4D-data in MS1 (ion mobility)



Background (what data is recorded?)

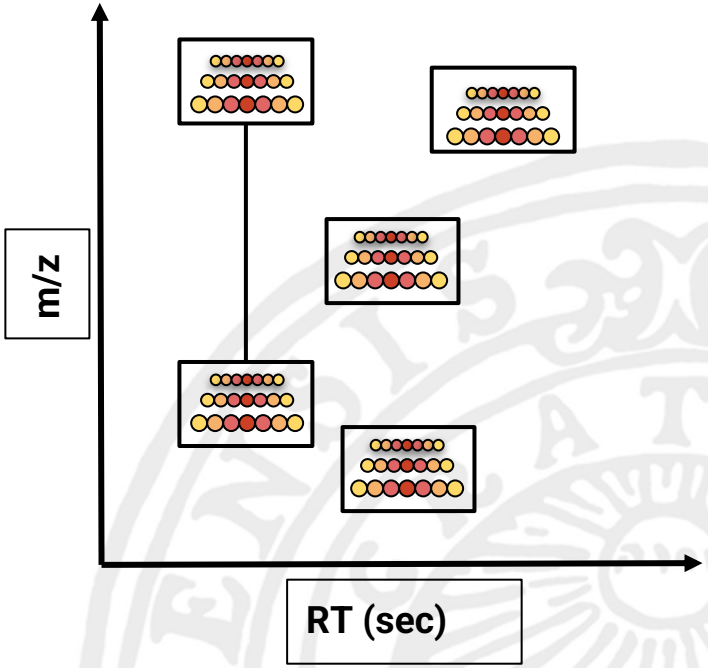
- MS1 data:
  - Retention time (RT): time taken for compound to elute from LC column
  - Mass to charge ratio (m/z): mass to charge as measured by mass analyzer
  - Measured signal (intensity): relative abundance of molecules





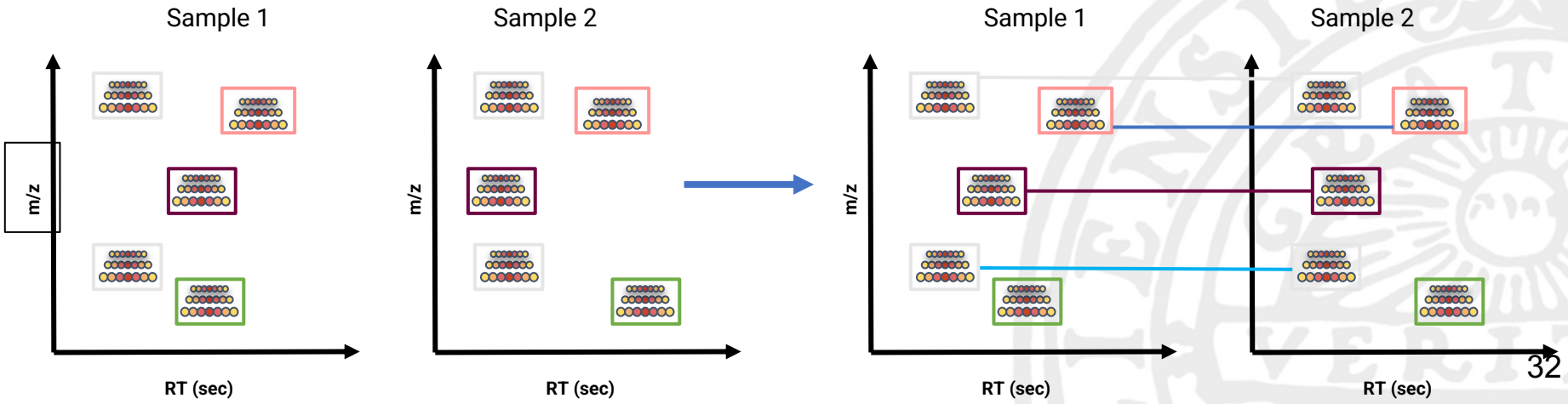
Background (pre-processing of MS1)

- A single molecule can show different signal patterns
- We cluster these signals for each of the molecules to calculate its neutral mass and abundance
- This is normally called feature detection + annotation
- Terminology:
  - Mass trace: signal recorded for a molecule over time with identical m/z: 
  - Feature: collection of mass trace for different isotopes: 



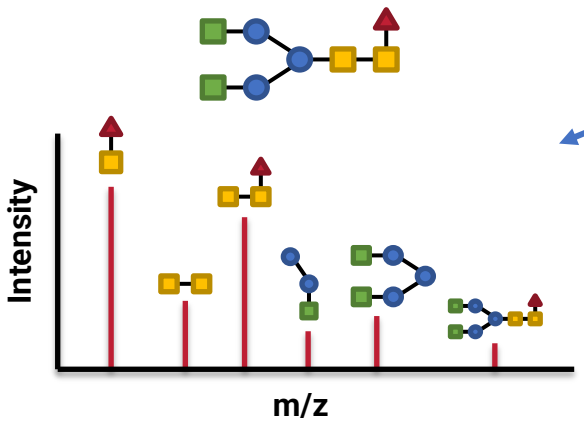
Background (pre-processing of MS1)

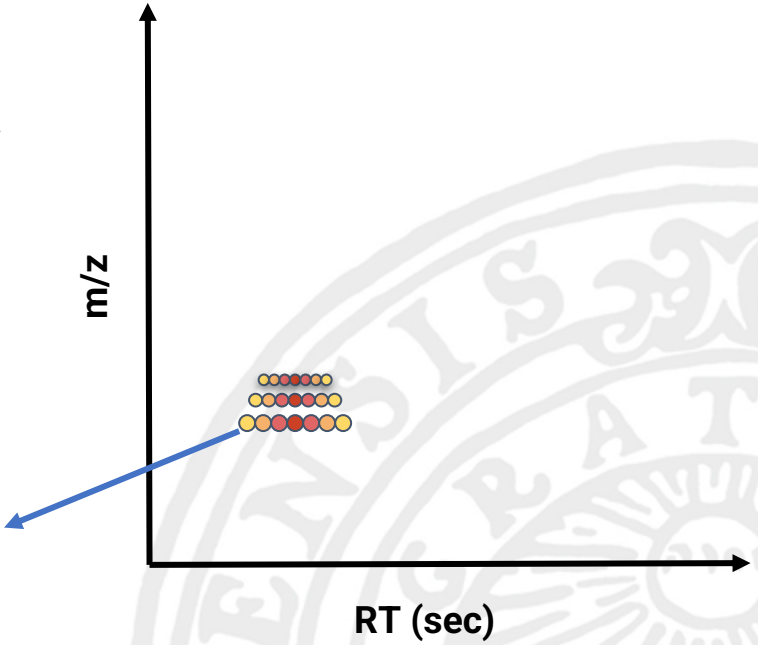
- We encounter LC time drift and also mass deviation between different samples
- Retention time alignment and feature grouping are performed to find corresponding features across different samples
- At this stage we can get the quantification results including estimated neutral mass etc



Background (identification)

- We can use the m/z of features to identify what molecule they however, in practice the accuracy of such MS1-based identification is low
- MS2 or fragmentation: selecting an ion, break it into the pieces and measure the m/z and abundance of the pieces





Big Data in metabolomics- Challenges

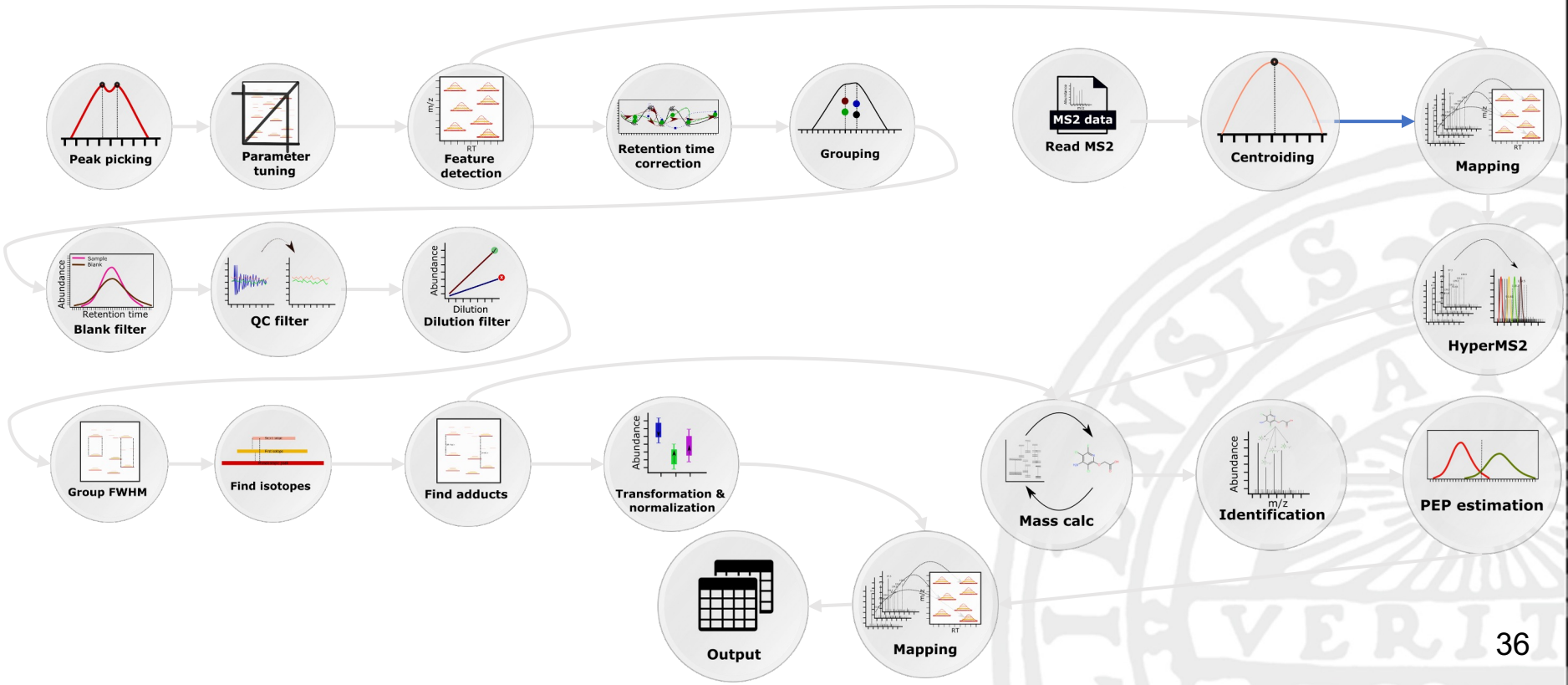
- The rapid progress in field has resulted in a mosaic of independent, and sometimes incompatible, analysis methods that are difficult to connect into a useful and complete data analysis solution
- Configuring necessary software tools and chaining them together into a complete re-runnable analysis

Big Data in metabolomics- Challenges

File formats, all vendors have their own…

The mzML format is an open, XML-based format for mass spectrometer output files, developed with the full participation of vendors and researchers in order to create a single open format that would be supported by all softwares

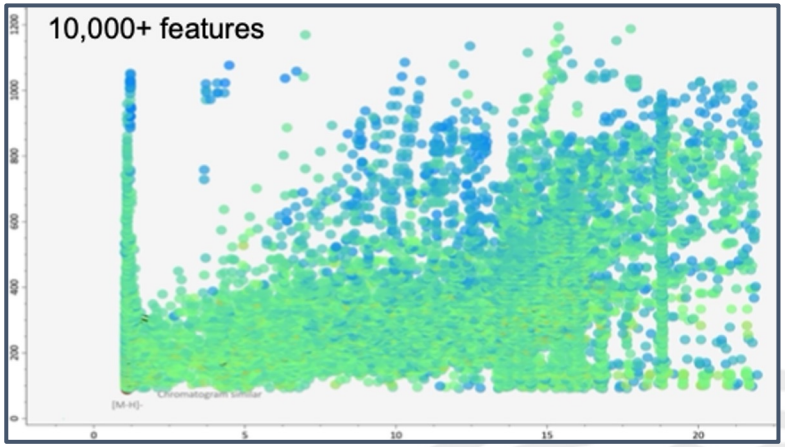
Complex pipelines needed to process complex data

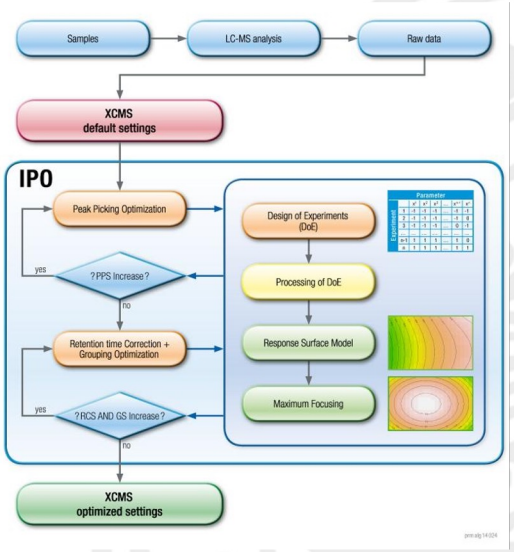


https://nf-co.re/metaboigniter/2.0.0

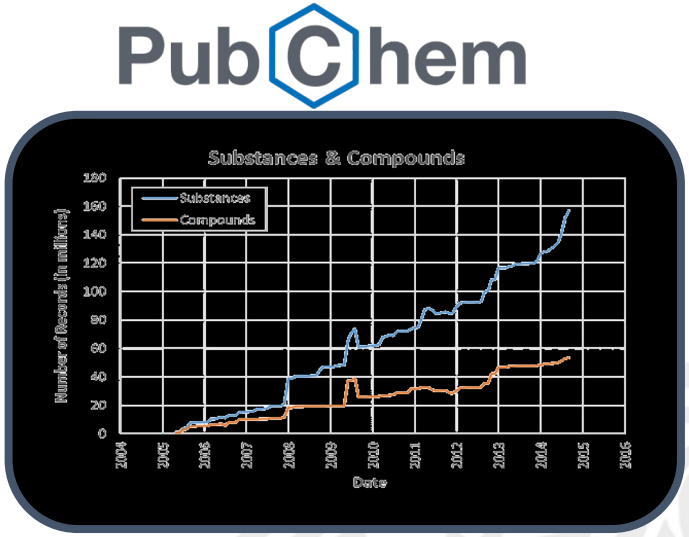
Is metabolomics data “BIG”?!

- Thousands of features can be detected in a single sample, often containing a considerable number of ions
  - For deep characterization, we need scalable platforms and efficient algorithms for parameter tuning

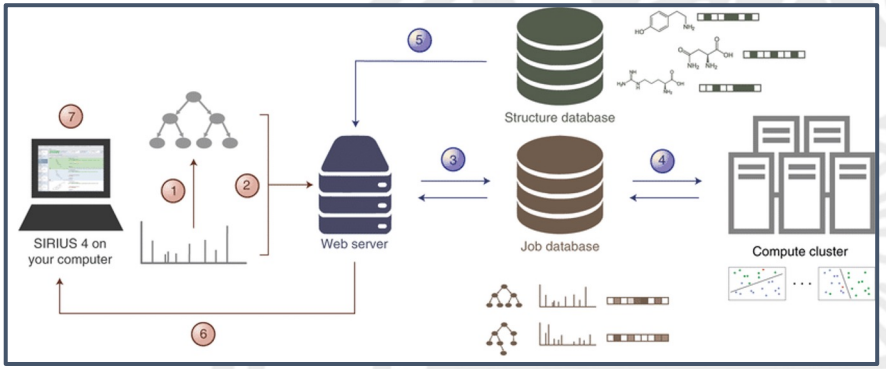




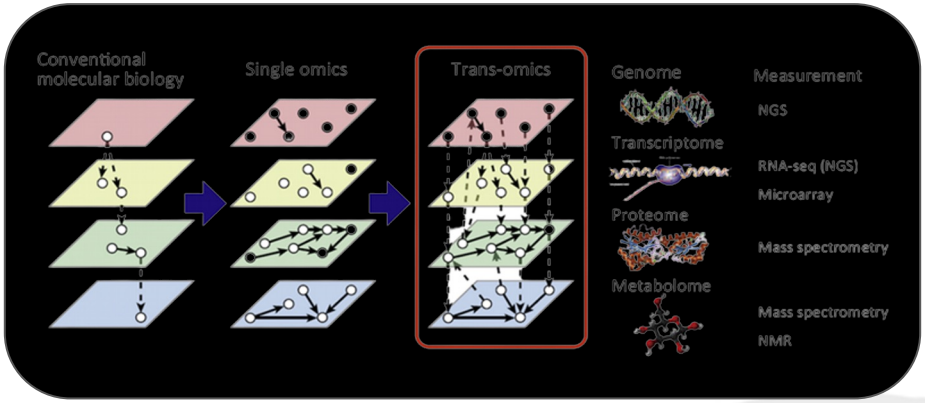
Is metabolomics data “BIG”?!



- Thousands of features can be detected in a single sample, often containing a considerable number of ions
  - For deep characterization, we need scalable platforms and efficient algorithms for parameter tuning
- Identification is a significant challenge in metabolomics data analysis!
  - More than 100 Million compounds (chemical structures)
  - It takes months to run!



Data integration



- Discover complex pattern not visible in single OMICS alone
- Often requires the same samples to be in all the OMICS data types
- But this is most often not the case due to cost or nature of the samples etc
- So not possible to do integration
- Solution: Analyze huge amount of data without correspondence but with common background
- Requires intensive computation
  - E.g., all studies in a repo!