# Metabolomics and metabolite identification – where are we now and the route ahead?


## Dr Warwick (Rick) Dunn

School of Biomedicine and Centre for Advanced Discovery & Experimental Therapeutics (CADET), Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre and University of Manchester, UK.

warwick.dunn@manchester.ac.uk
http://www.manchester.ac.uk/research/warwick.dunn/
http://www.manchesterbrc.org/OurFacilities/CADET.php



A partnership between Central Manchester University Hospitals NHS Foundation Trust and The University of Manchester

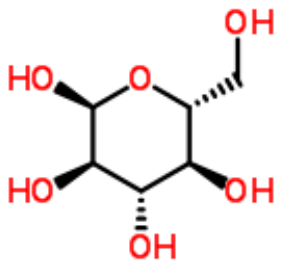
Supported by MAHSC, Northwest Regional Development Agency, NHS National Institute for Health Research

MANCHESTER BIOMEDICAL RESEARCH CENTRE

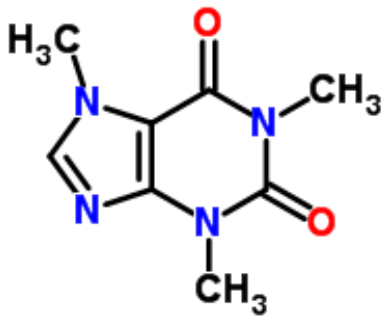
MANCHESTER 1824

The University of Manchester

# A tutorial – my views

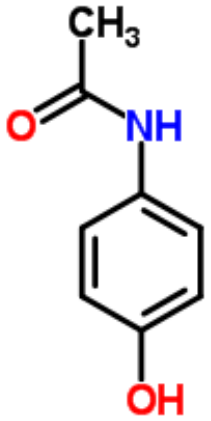



- Introduce our abilities and common difficulties/ limitations of metabolite identification in complex metabolomic samples
  - describe common workflows for GC-MS and UPLC-MS




- Describe the reporting standards for metabolite identification

- Discuss what innovative tools are required or are being developed



# Component characterisation of simple solutions

- Relatively easy for analytical chemists to characterise a single component solution!

- Many tools available for characterisation of unknowns
  - mass spectrometry (MS)
  - nuclear magnetic resonance (NMR) spectroscopy
  - ultraviolet spectroscopy (UV)
  - infrared spectroscopy (IR)
  - elemental analysis

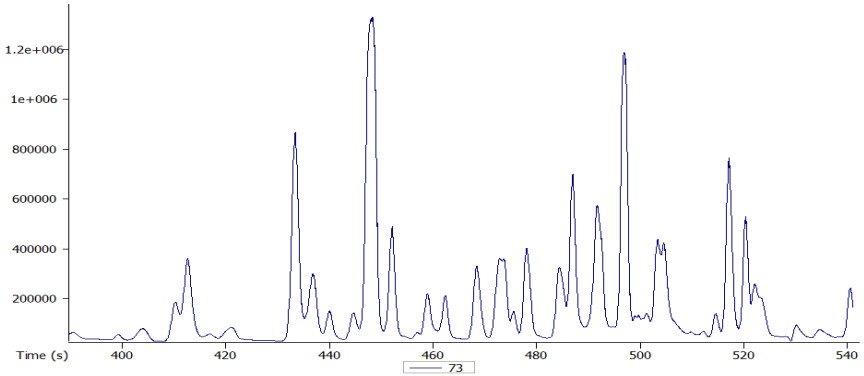

- Many of these are not appropriate for complex multi-component solutions
  - metabolomic samples are very complex (contain 100-1000s of metabolites)
  - mass spectrometry and NMR are two tools commonly applied in the analysis of complex metabolomic samples

MANCHESTER 1824

The University of Manchester

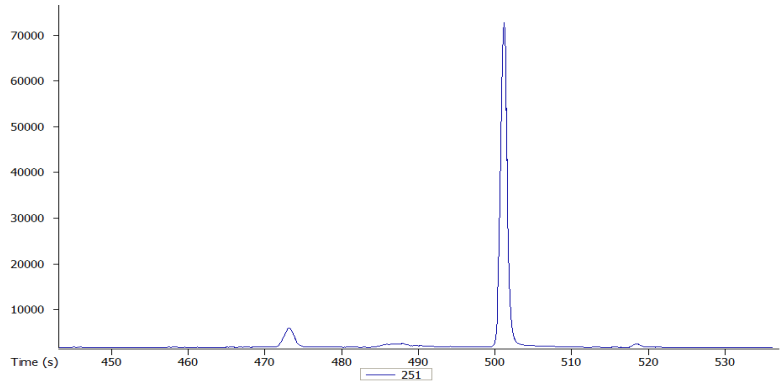
# Untargeted *vs.* (semi-)targeted metabolomic studies

## METABOLIC PROFILING or UNTARGETED ANALYSIS




- (semi)-quantitative (global) detection of a wide range of metabolites
- Orbitrap, TOF, Q-TOF, IT, Q, FTICR
- data acquisition without *a priori* knowledge of biologically interesting metabolites
- metabolite identification required post data acquisition

## TARGETED OR SEMI-TARGETED ANALYSIS




- quantification of a smaller number of (related) metabolites for
  - targeted = generally less than 20
  - semi-targeted = low 100s
- QQQ
- metabolite identity already known
  - no further metabolite identification required

**This seminar discusses metabolite identification in data acquired applying metabolic profiling strategies (i.e. complex samples containing 100-1000s of metabolites)**

MANCHESTER 1824

The University of Manchester

# Metabolite identification – A BOTTLENECK IN METABOLOMICS

For metabolomics to be successful it is essential to derive biological knowledge from analytical data - a view emphasised by a recent Metabolomics ASMS Workshop Survey 2009 which found that the biggest bottlenecks in metabolomics were thought to be identification of metabolites (35%) and assignment of biological interest (22%).

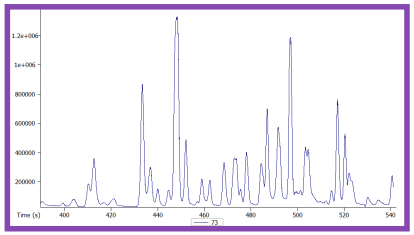
*http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics-Survey-2009*

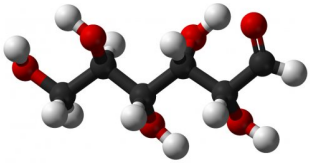
MANCHESTER 1824

The University of Manchester

# Why is metabolite identification a bottleneck?

- Tools available for identification of a limited number of metabolites in a semi-automated process (traditional analytical chemistry)
  - these are being applied for identification of 100-1000s of metabolites
  - limited number of these tools which have been developed and experimentally validated for high-throughput metabolite identification of all metabolites
- Metabolomes and raw data are complex
- 7800+ metabolites in human body (not including gut microflora –derived, drug-derived and many lipids)
- Qualitative description of all metabolomes is not complete (and not electronically available)
- Different physicochemical properties (diversity is greater than proteome for example)





MANCHESTER 1824

The University of Manchester

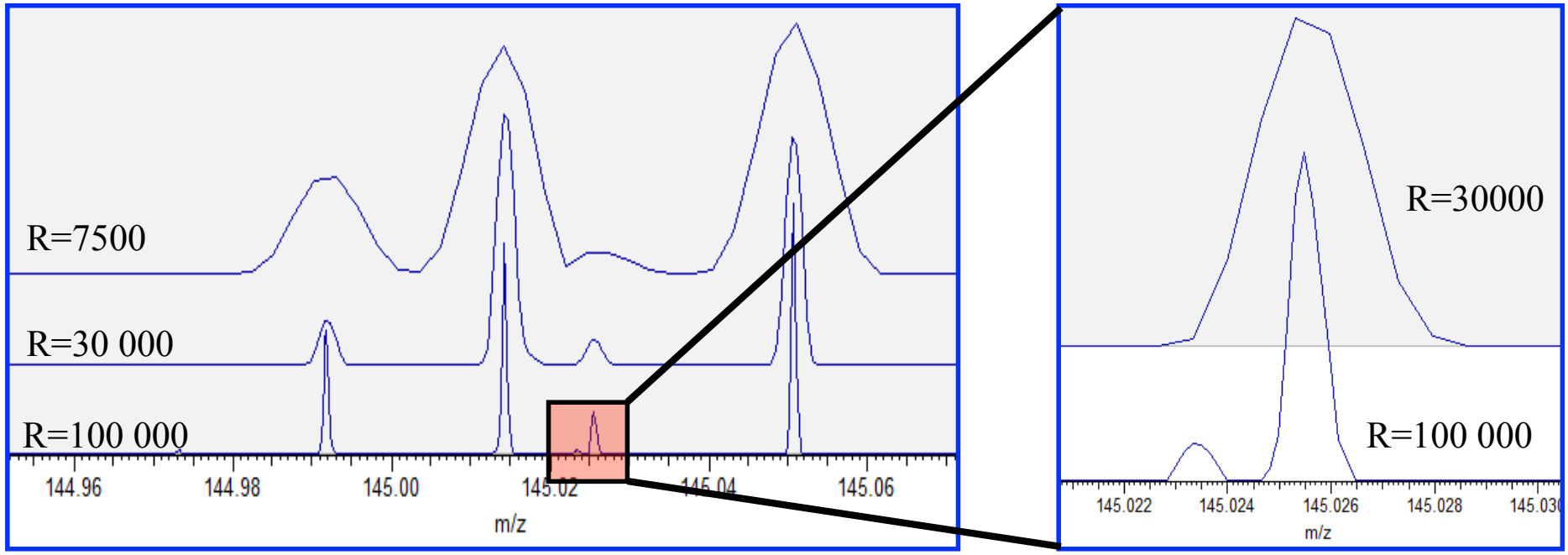
# Mass spectrometers provide many advantages for metabolite identification in metabolomics

- **Sensitive detection** (sub-micromoles.$L^{-1}$ to millimoles.$L^{-1}$)
  - detection of 100-1000s of metabolic features/metabolites
- **High mass resolution** (5000 to >200 000+ FWHM)
  - ability to separate features of similar but not identical monoisotopic mass
- **High mass accuracy**(< 5ppm)
  - ability to accurately determine the mass of detected metabolic features
  - molecular formula determination
- **Gas phase ion fragmentation**
  - GC-MS using EI sources (or CI or QQQ)
  - LC-MS using QQQ, Q-TOF or LIT for MS/MS
  - structural determination
- **Isotope patterns and relative isotope abundance (RIA)**
- **New developments in instruments and computational tools**

MANCHESTER 1824

The University of Manchester

# Technological advances during the last decade!




***Typically we detect 1.5 to 3 times more mass peaks in direct infusion experiments when applying a mass resolution of 100 000 compared to 7500***

MANCHESTER 1824

The University of Manchester

# Levels of metabolite identification

- Sumner et al. Proposed minimum reporting standards for chemical analysis, Metabolomics, 2007, 3(3), 211-221
- Currently, four levels of metabolite identifications can be reported
- Not defining how to perform metabolite identification but defining how to report it

| Level | Confidence of Identity | Level of Evidence |
|---|---|---|
| 1 | Confidently identified compounds. | Comparison of two or more orthogonal properties with an authentic chemical standard analysed under identical analytical conditions. |
| 2 | Putatively annotated compounds | Based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries, without reference to authentic chemical standards. |
| 3 | Putatively annotated compound classes | Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class. |
| 4 | Unknown compounds | Although unidentified and unclassified, these metabolites can still be differentiated and quantified based upon spectral data. |

MANCHESTER 1824

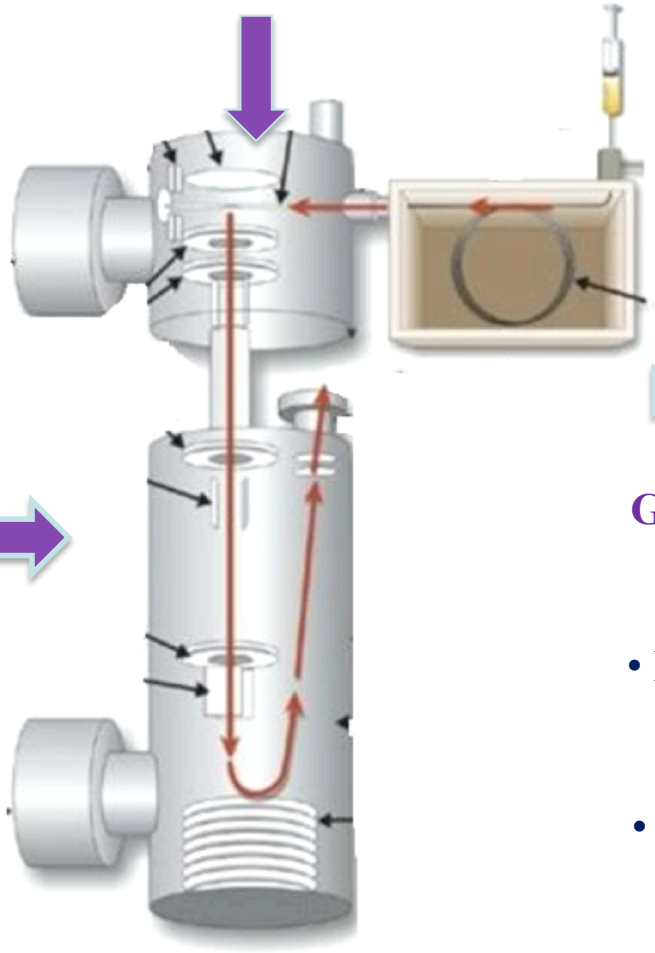
The University of Manchester

# Gas Chromatography-Mass Spectrometry (GC-MS)

## IONISATION SOURCES PROVIDE:

• **ELECTRON IMPACT** - PROVIDES REPRODUCIBLE GAS-PHASE FRAGMENTATION FOR STRUCTURE ELUCIDATION

• **CHEMICAL IONISATION** - NO FRAGMENTATION, ACCURATE MASS MEASUREMENT OF MOLECULAR ION



## MASS SPECTROMETRY PROVIDES:

• HIGH MASS RESOLUTION

• HIGH MASS ACCURACY (SOME NOT ALL INSTRUMENTS),

• HIGH SCAN SPEEDS / ACQUISITION RATES TO ACCURATELY DEFINE NARROW CHROMATOGRAPHIC PEAKS

• HIGH SENSITIVITY

## GAS CHROMATOGRAPHY PROVIDES:

• REPRODUCIBLE

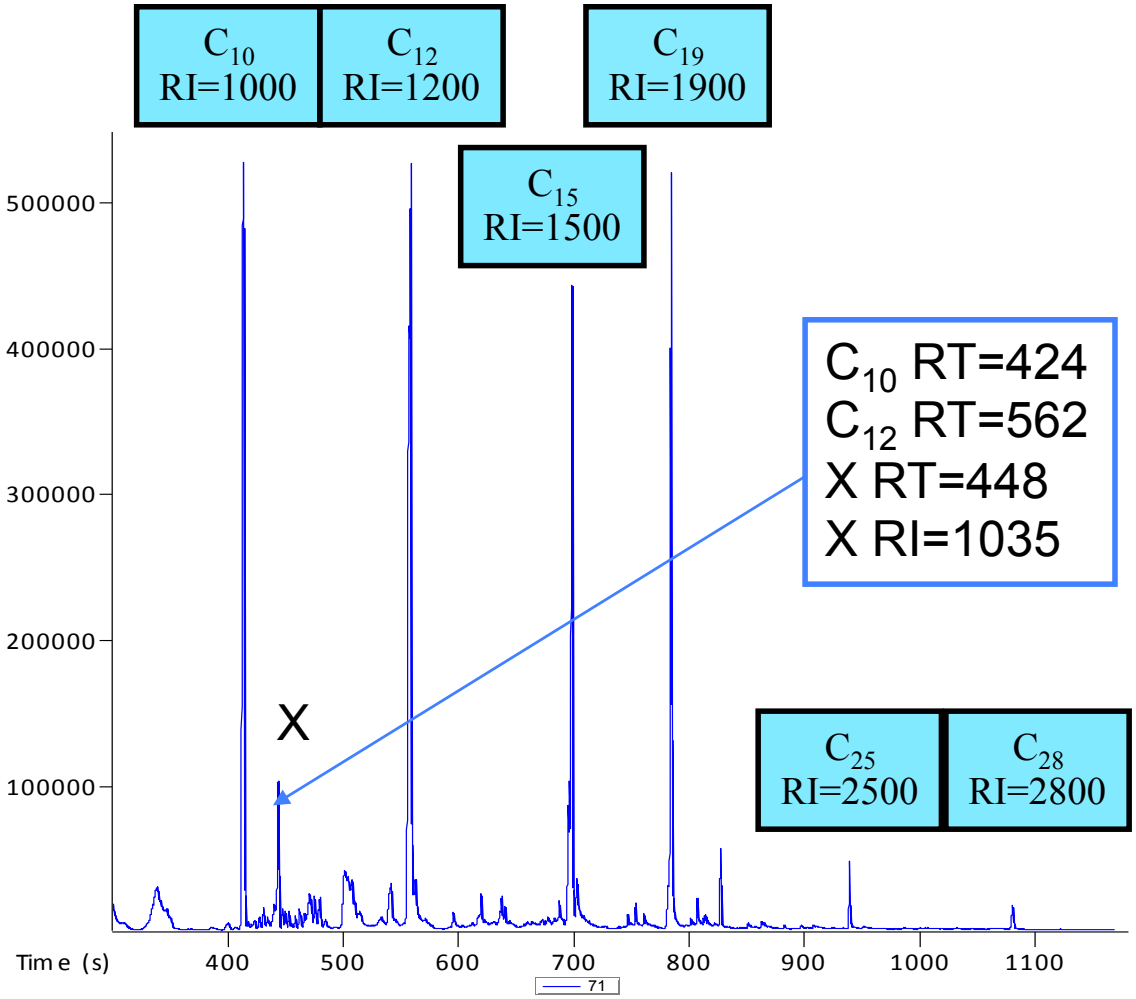
• HIGH RESOLUTION (PEAK WIDTHS OF A FEW SECONDS),

• RETENTION INDICES FOR METHOD AND LIBRARY TRANSFERABILITY

MANCHESTER 1824

The University of Manchester

# Retention indices




Applies a series of homologous compounds
e.g. n-alkanes
e.g. fatty acid methyl esters

- Normalisation of retention time range

- Minimises errors associated with drift in retention time

- Can be applied across different GC columns and instruments (method and mass spectral library transferability)

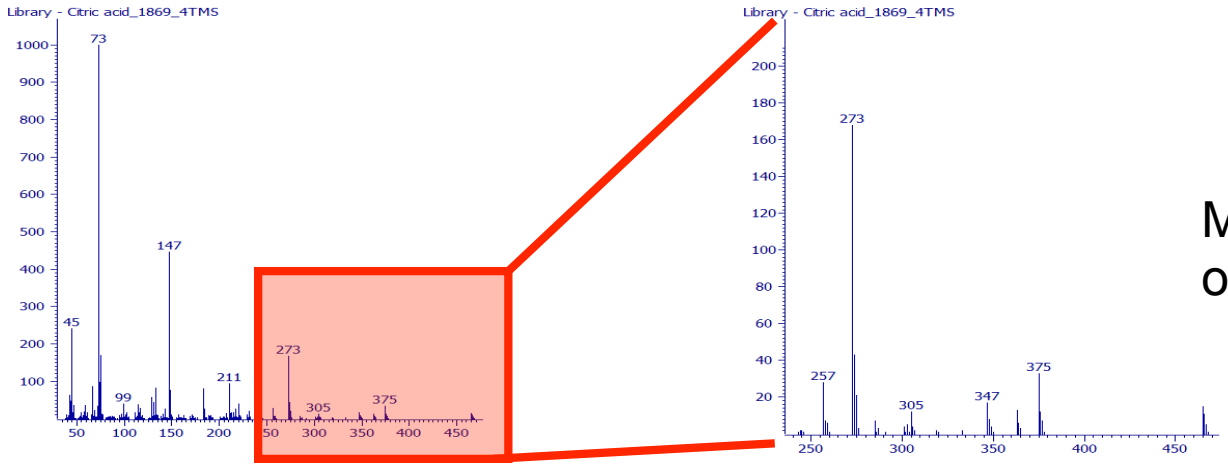
# Mass spectral libraries and library matching

- Mass spectral libraries
  - constructed with authentic chemical standards
  - multiple libraries available, none are comprehensive
  - we apply The Manchester Metabolomics Database (MMD) library

- Comparing the mass spectrum of an authentic chemical standard against the mass spectrum of an unknown metabolite
  - Compares and scores depending on number of matched ions and relative intensity of those ions

- Provides a confidence score on match (out of 1000 or as a %)

- Difficulty in trimethylsilyl spectra as m/z 73 and 147 are common in most TMS-metabolites at high intensity and so matching can be compromised
  - differences can be based on a limited number of high m/z ions with a low response



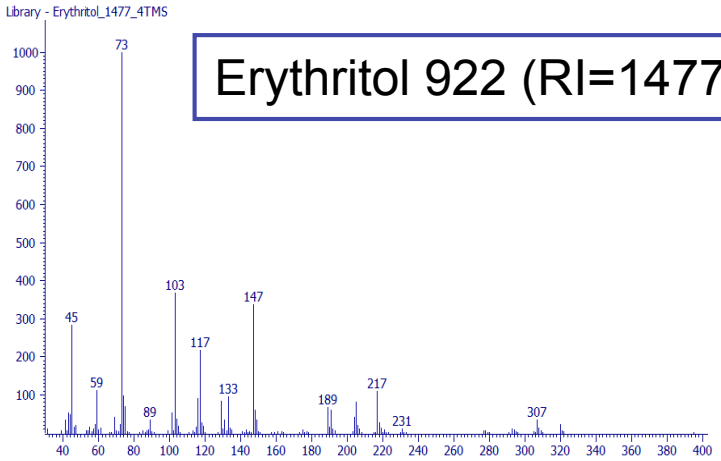

Mass spectrum of citric acid

MANCHESTER 1824

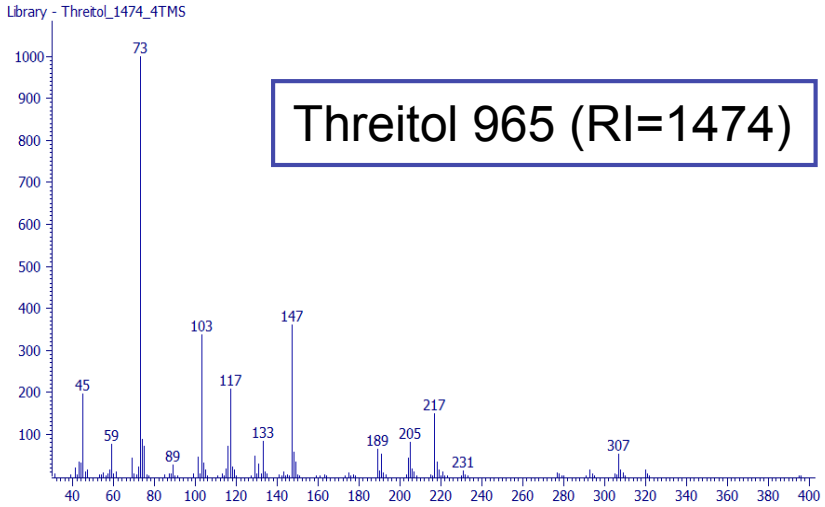
The University of Manchester

# Problems to consider

- Metabolites of similar chemical structures have a similar chemical structure and may have similar retention index and mass spectrum
- Targeted separation methods required
- Report as X and/or Y



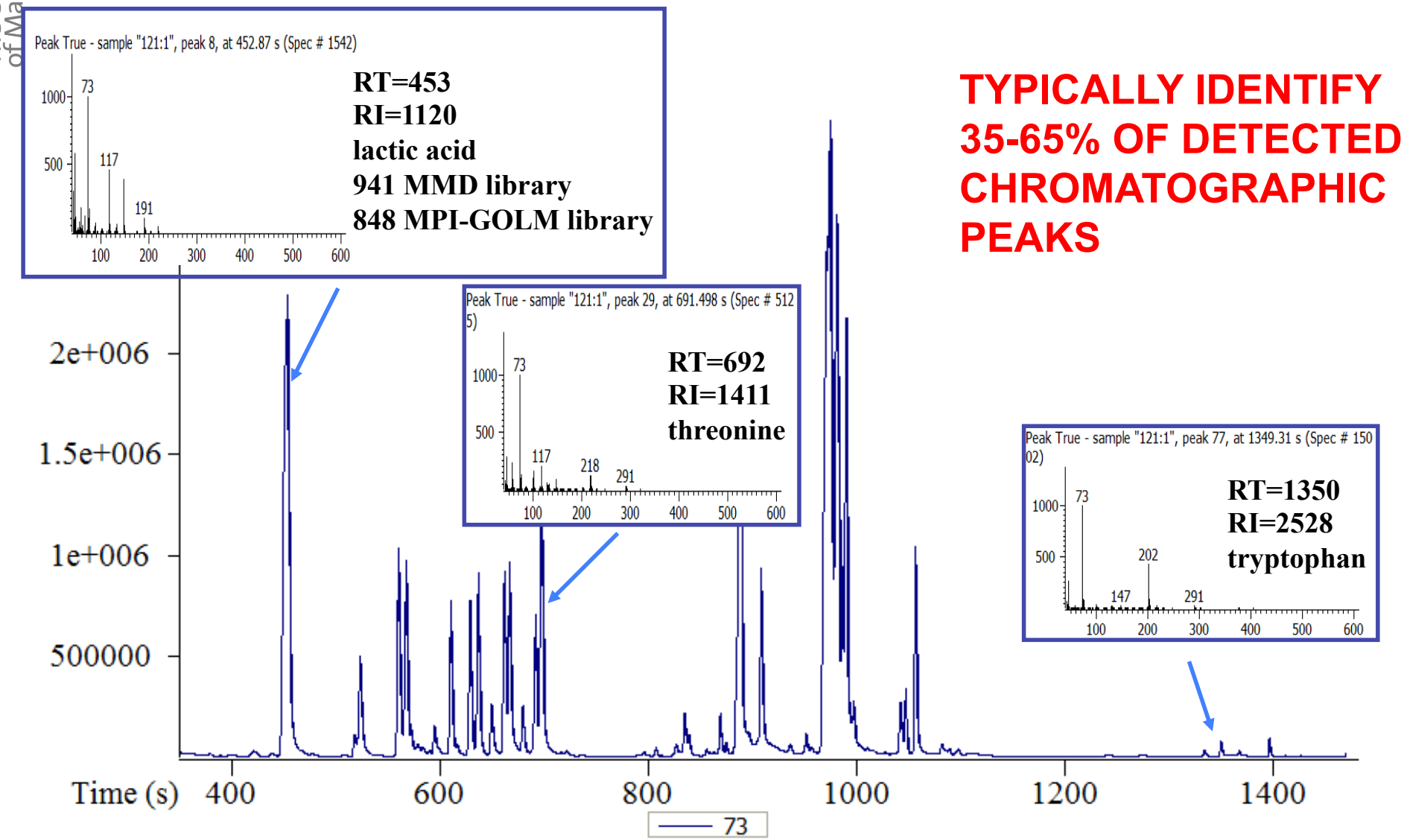

Erythritol 922 (RI=1477)




Threitol 965 (RI=1474)

Erythritol 922 (RI=1477) AND/OR
Threitol 965 (RI=1474)

# Mammalian cell footprint sample



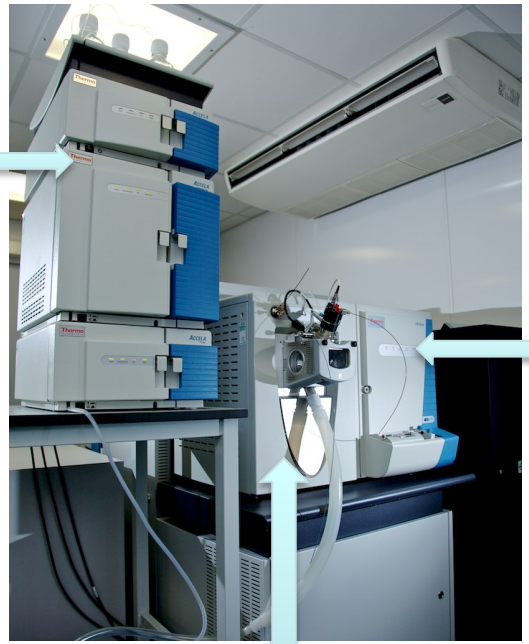
MANCHESTER 1824

The University of Manchester

# Electrospray Ionisation (ESI): DIMS, LC-MS, UPLC-MS and CE-MS

## UPLC/UHPLC PROVIDES:

- REPRODUCIBLILITY
- HIGH RESOLUTION (PEAK WIDTHS OF A FEW SECONDS),
- NO RETENTION INDICES FOR METHOD AND LIBRARY TRANSFERS



## MASS SPECTROMETRY PROVIDES:

- HIGH MASS RESOLUTION
- HIGH MASS ACCURACY (SOME NOT ALL INSTRUMENTS),
- HIGH SCAN SPEEDS / ACQUISITION RATES FOR TO ACCURATELY DEFINE NARROW CHROMATOGRAPHIC PEAKS
- HIGH SENSITIVITY
- MS/MS OR $MS^n$ CAPABILITIES FOR MOLECULAR ION FRAGMENTATION (ALL ION OR SELECTED ION)

## IONISATION SOURCES PROVIDE:

- ELECTROSPRAY OR APCI
- REPRODUCIBLE
- MINIMAL/NO ION FRAGMENTATION
- COMPLEX ADDUCT FORMATION (LIQUID AND GAS-PHASE

MANCHESTER 1824

The University of Manchester

# Routine workflow applied

- Accurate *m/z* data acquisition

1. Apply data related to ion type

- Molecular formula determination

1. Applications of rules to filter (e.g. relative isotope abundances)
2. Comparison of MF to metabolite databases

- Putative metabolite identification

- MS/MS or $MS^n$ fragmentation
  - experimental and authentic standard / *in-silico*

*de novo* structural elucidation

comparison to mass spectral libraries

- Reduction of number of putative hits from accurate *m/z* data

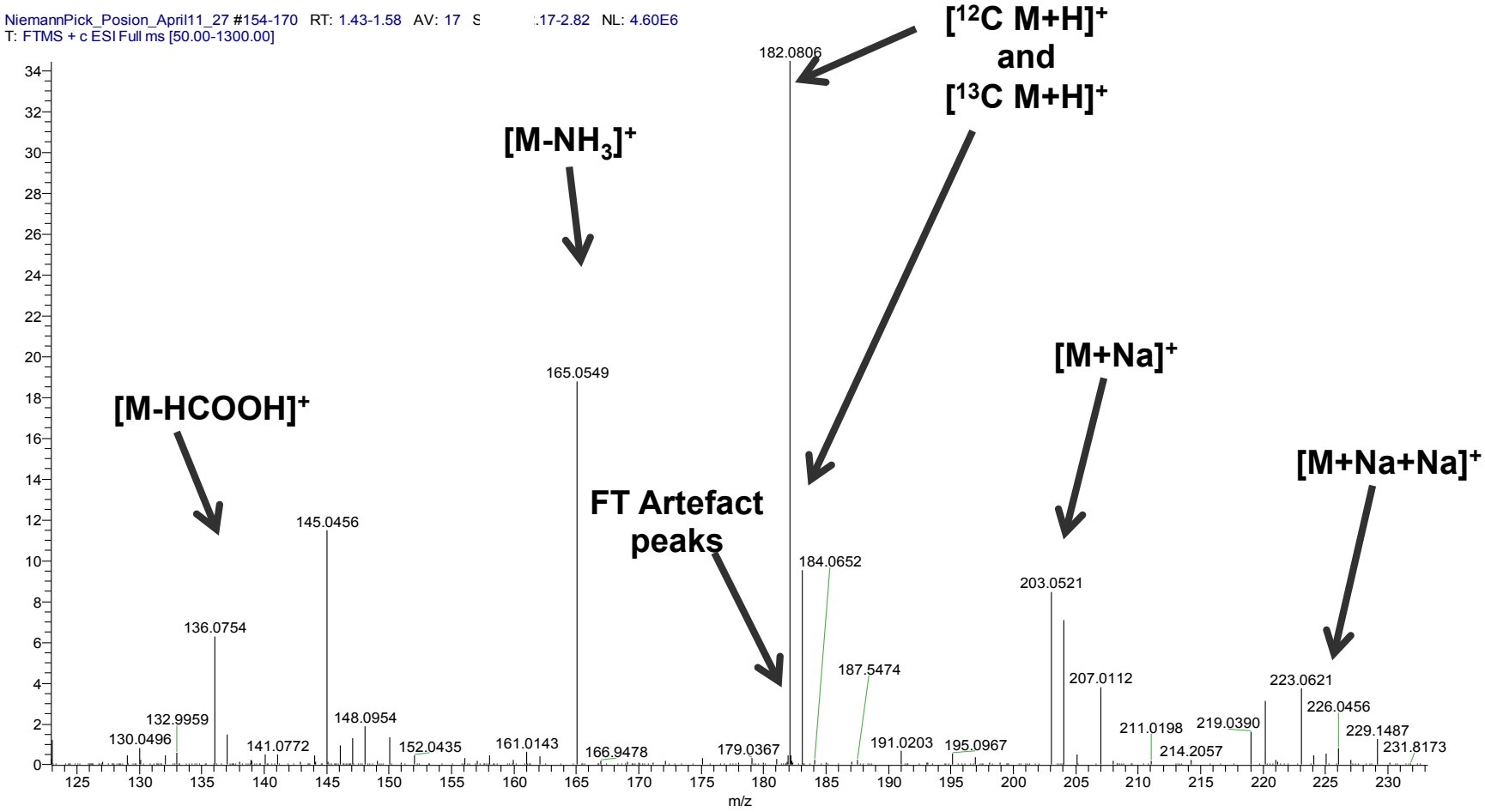
- Authentic standard purchase or synthesis and comparative analysis

MANCHESTER 1824

The University of Manchester

# The complexity of ESI data - tyrosine

## 17 IONS OF DIFFERENT MASS AND SAME RETENTION TIME CREATED FROM A SINGLE METABOLITE – NEED TO DEFINE THE ION TYPE FOR IDENTIFICATION OR HIGH PROBABILITY OF FALSE POSITIVE IDENTIFICATION



MANCHESTER 1824

The University of Manchester

# Accurate measurement of *m/z* as the first process applied

- However, high chance of false positives if type of ion is not determined before conversion to molecular formula
  - see Brown M, Dunn W.B., et al. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *The Analyst* 2009, 134, 1322-1332.
  - determine ion type first using accurate mass differences, RT and correlation analysis

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 8 2011, pages 1108–1112
doi:10.1093/bioinformatics/btr079

*Systems biology*

Advance Access publication February 16, 2011

**Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets**

Marie Brown[1], David C. Wedge[2], Royston Goodacre[2,3], Douglas B. Kell[2], Philip N. Baker[4], Louise C. Kenny[5], Mamas A. Mamas[1,6], Ludwig Neyses[1,6] and Warwick B. Dunn[1,2,3,7,*]

[1]School of Biomedicine, The University of Manchester, Manchester M13 9PT, [2]School of Chemistry, [3]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK, [4]Department of Obstetrics and Gynecology, Faculty of Medicine and Dentistry, University of Alberta, 2J2.01 WMC, Edmonton AB T6G 2R7, Canada, [5]The Anu Research Centre, Department of Obstetrics and Gynaecology, University College Cork, Cork University Maternity Hospital, Cork, Ireland, [6]Manchester Heart Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester Royal Infirmary and [7]Centre for Advanced Discovery and Experimental Therapeutics, York Place (off Oxford Road), Central Manchester University Hospitals NHS Foundation Trust, Manchester M13 9WL, UK

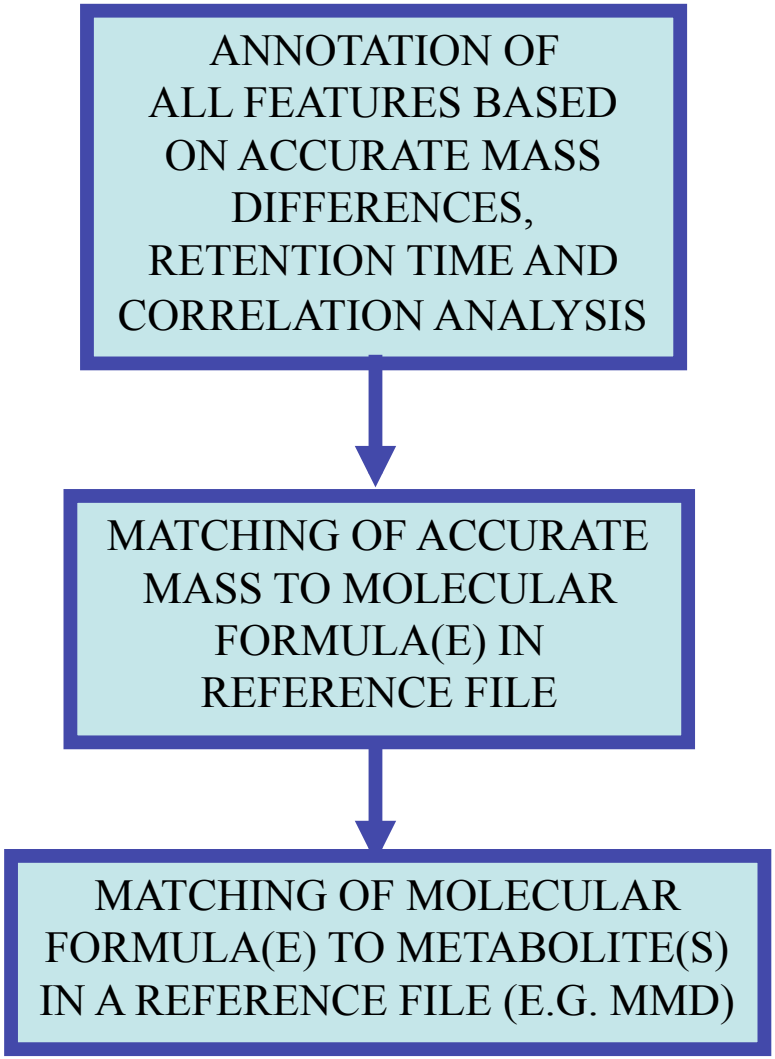
Associate Editor: John Quackenbush

MANCHESTER 1824

The University of Manchester

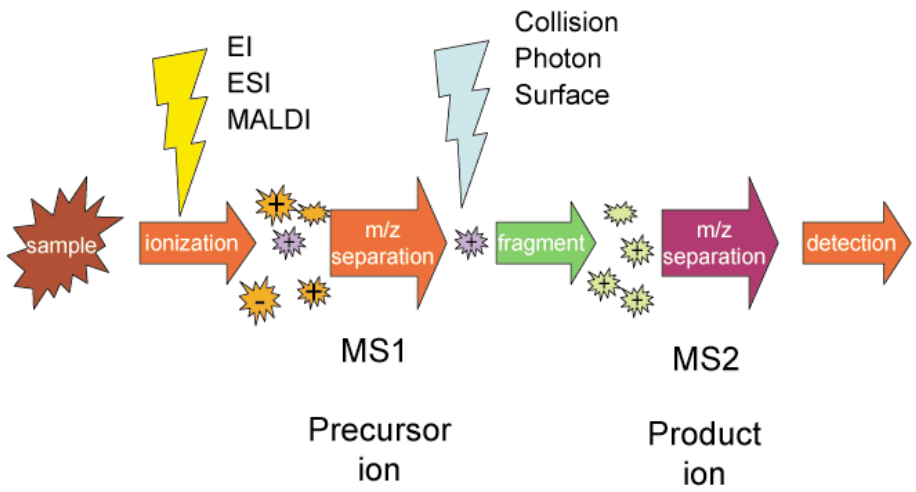
# PUTMEDID-LCMS

- We have developed an *in-silico* suite of workflows for metabolite identification
  - automated and high-throughput
  - for holistic identification of all features
  - easy to use and idiot-proof (i.e. I can use it!!)
  - fill a gap in currently available tools
  - apply information on ion type to reduce number of false positives
- Three separate Taverna workflows have been developed
  - flexibility built in
  - converts accurate mass data to molecular formula(e) and potential metabolite
  - applies reference files which can be developed by the user to be instrument/organism specific
  - developed for Windows not Macs



# MS/MS and MS$^n$

- Gas-phase fragmentation through ion activation in a vacuum



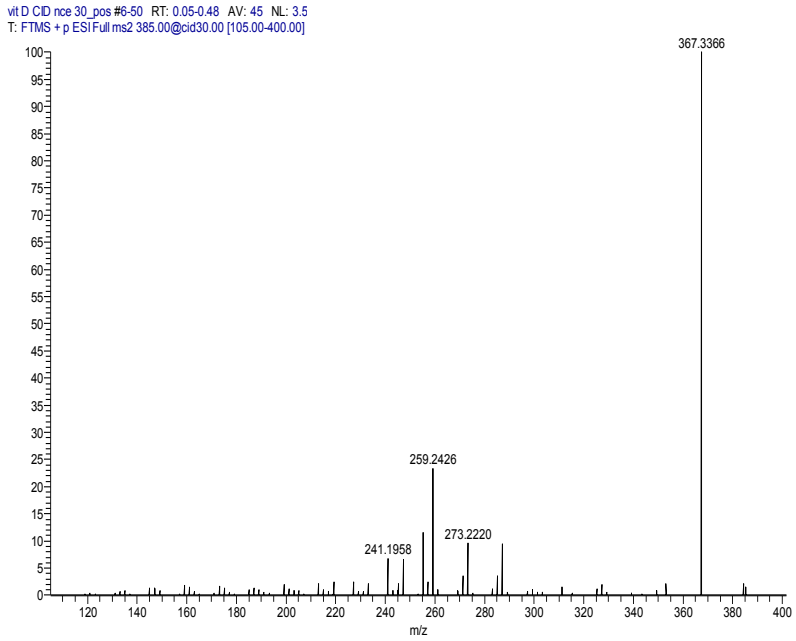

http://en.wikipedia.org/wiki/Tandem_mass_spectrometry

- Specific to a single mass/metabolite OR all-ion fragmentation
- Different ion activation mechanisms available
  - Collision Induced Dissociation (CID) in a Q-TOF or QQQ (MS/MS)
  - CID in an ion trap/linear ion trap (MS$^n$ whewre n can be greater than 2)
  - HCD in Orbitrap instruments (MS/MS)
- Advantages and limitations (e.g. IT/LIT (1/3$^{rd}$ rule))
- Provide structural information
- Apply to reduce number of potential molecular formula
  - like putting a jigsaw puzzle together

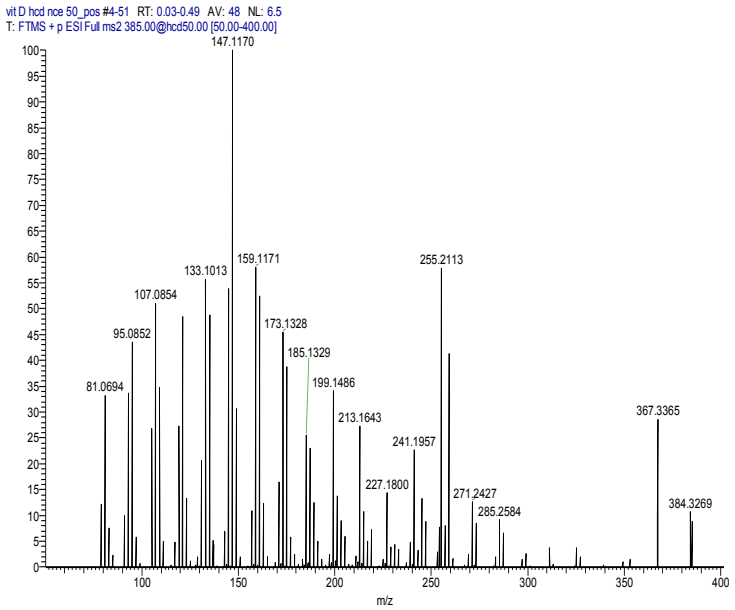
MANCHESTER 1824

The University of Manchester

# Complementary ion activation mechanisms are advisable – CID *vs* HCD




CID in a linear ion trap

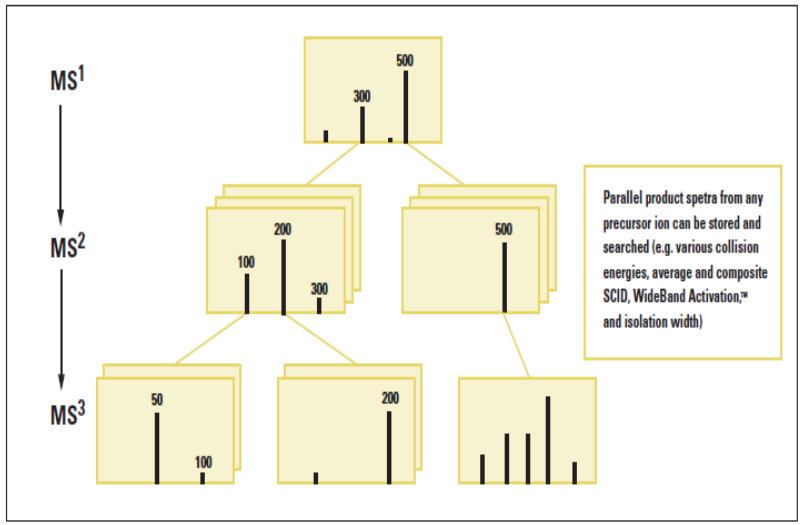



HCD

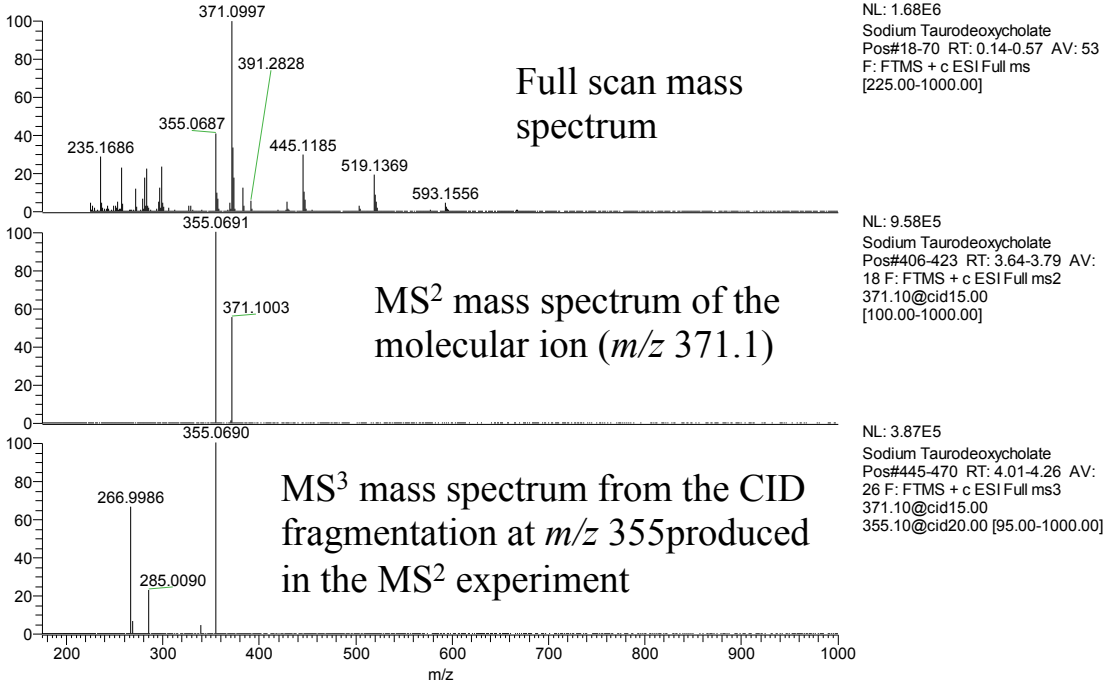
- CID and HCD can (but not always) provide complementary mass spectral data
- Comparable with CID and ECD in proteomics

Thanks to Dr Graham Mullard in CADET/School of Biomedicine for providing the data

# $MS^n$ – mass spectral trees and increased specificity




Parent to daughter to grand-daughter to great grand-daughter




**Courtesy of ThermoFisher Scientific**

MANCHESTER 1824

The University of Manchester

# Problems to consider

- Not all metabolites present in a diverse range of metabolomes are known and electronically tagged
  - *can only apply comparative data if metabolites are present in these databases/ libraries*

- Mass spectral libraries
  - *not all known metabolites are commercially available and so are present in mass spectral libraries*
  - *LC-MS/MS libraries are significantly less developed than for GC-MS*
  - *LC-MS/MS libraries – are they transferable? RT and mass spectra*

- Data are complex
  - *one metabolite = multiple features*
  - *false positives*

- No automated workflows employing multiple strategies
  - *alot of manual work still involved!!!!*

MANCHESTER 1824

The University of Manchester

# In-silico/computational tools in development

- Use data from ‘knowns’ or computational algorithms to predict ‘unknowns’
- *In-silico* fragmentation for LC-MS
  - MassFrontier (http://www.highchem.com/massfrontier/mass-frontier.html)
- Substructure prediction for GC-MS
  - Hummel J, et al., Decision tree supported substructure prediction of metabolites from GC-MS profiles. Metabolomics. 2010, 6(2), 322-333.
- Retention time/index prediction
  - Kumari S, et al., Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. Anal Chem. 2011, 83(15):5895-902
- Ionisation behaviour rules for ESI-MS
  - Draper J et al. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted **ionisation** behaviour 'rules’. BMC Bioinformatics. 2009, 10:227.
- Application of prior biological knowledge (biological samples are not random collections of chemicals but chemicals are linked by enzymatic reactions)
  - Weber RJM et al. MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. Chemometrics and Intelligent Laboratory Systems, 2010, 104, 75-82.
- Synthesis of novel metabolites
- In-vivo stable isotope labelling

- **If all else fails……isolation of metabolite and *de novo* structural elucidation**

MANCHESTER 1824

The University of Manchester

# Summary

- Metabolite identification is a highly complex process in metabolomics
- Mass spectrometry offers many tools for metabolite identification
  - accurate mass
  - MS/MS and $MS^n$
  - retention time and retention index
  - mass spectral libraries
  - computational tools
- Limited automated and high-throughput INTEGRATED workflows available as of yet (especially for ESI-MS)
- Unable to identify all metabolites in a sample currently and we are a long way off
- Require a slow cataloguing of metabolites present in a diverse range of metabolomes across many research groups and their database integration
- We are currently on an important developmental journey which is essential for metabolomics to be successful