19th Annual Conference of the Metabolomics Society
METABOLOMICS 2023
NIAGARA FALLS, CANADA JUNE 18-22

Spectra processing, statistical analysis and functional integration using MetaboAnalyst 5.0

TA: Zhiqiang Pang

Zhiqiang.pang@xialab.ca | www.xialab.ca

McGill University, Montreal, QC Canada

McGill UNIVERSITY

Acknowledgements

GenomeCanada

GenomeQuébec

NIH

The Common Fund

NSERC CRSNG

Canada Research Chairs

www.chairs-chaires.gc.ca

TMIC

The Metabolomics Innovation Centre

Schedule

Part I: 8:15 AM – 10:15 PM

- **8:15 – 8:30**: Opening lecture (Jeff)
- **8:30 – 8:50**: **Section 1: LC-MS spectral processing (Qiang)**
- **8:50 – 9:15**: **Section 1: Hands-on**
- **9:20 – 9:45**: Section 2: Stats I – simple experimental design (Jessica)
- **9:45 – 10:15:** Section 2 Hands - on

Part II: 10:30AM – 12:30PM

- **10:30 – 10:50**: Section 3: Functional analysis (Yao)
- **10:50 – 11:10**: Section 3: Hands on
- **11:15 – 11:40**: Section 4: Stats II - complex experimental design (Jessica)
- **11:40 – 12:15**: Section 4: Hands-on
- **12:15 – 12:30**: Summary (Jeff)

How are molecules different from each other?

Molecules are made of atoms.

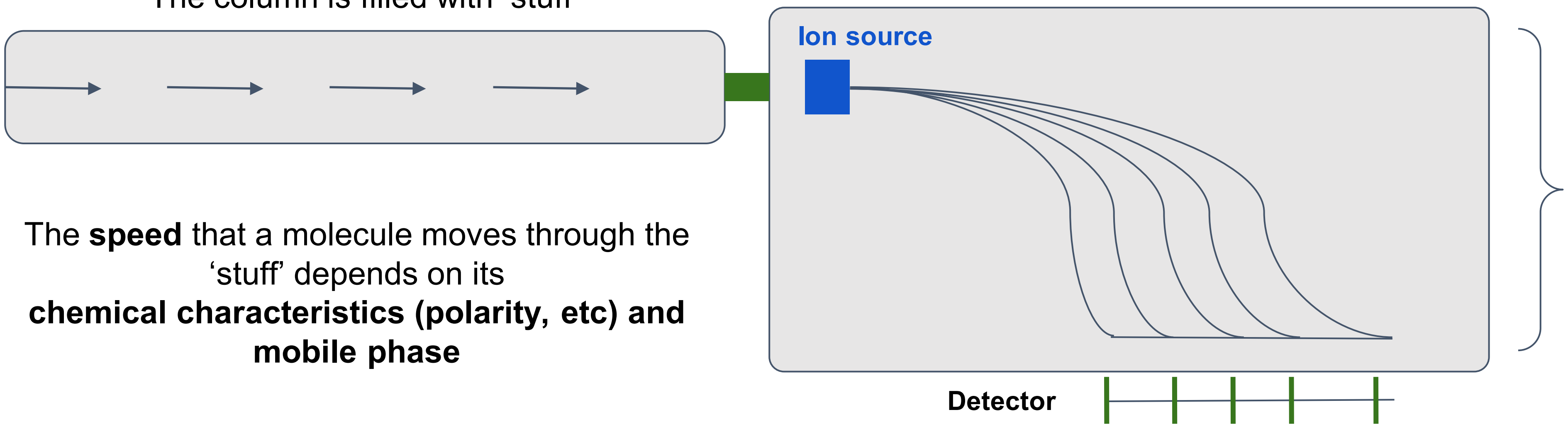
They can have different **mass** based on the **types of their atoms**.

They can have different **shapes and polarity** based on the **arrangement of their atoms**.

How can we measure these things with LC-MS?

Liquid Chromatography (LC) + Mass Spectrometry (MS) = LC-MS

The column is filled with ‘stuff’

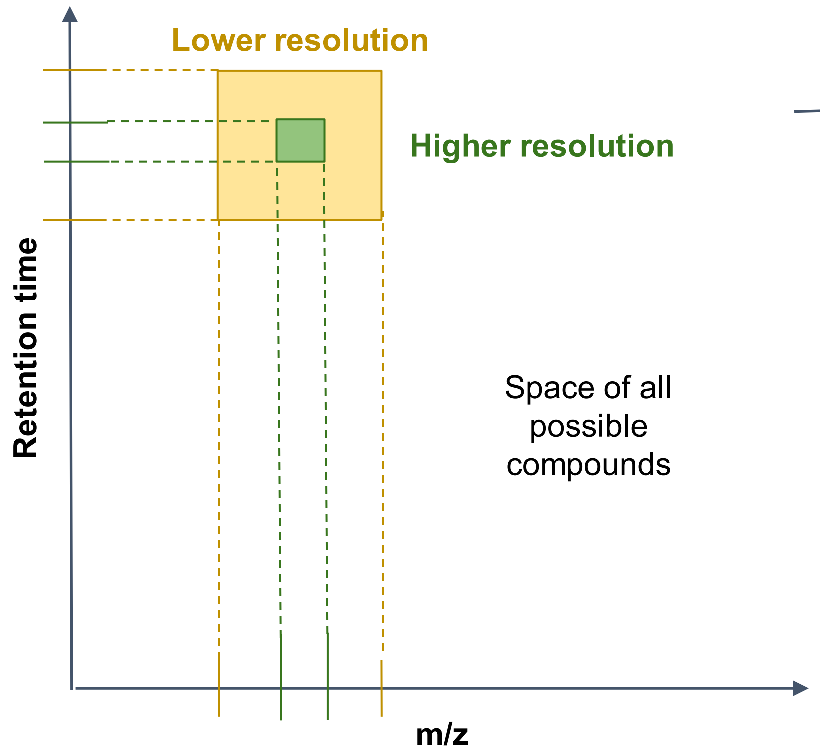


The **speed** that a molecule moves through the ‘stuff’ depends on its **chemical characteristics (polarity, etc) and mobile phase**

The **path** of the ionized molecule depends on its **mass:charge ratio** (*m/z*)

Therefore, the **location** that the ionized molecule collides with the detector is associated with the **mass** of the molecule

Can we uniquely identify compounds?



Narrow down the potential identity of the detected compounds

The higher the machine's resolution, the more we can narrow down the possibilities

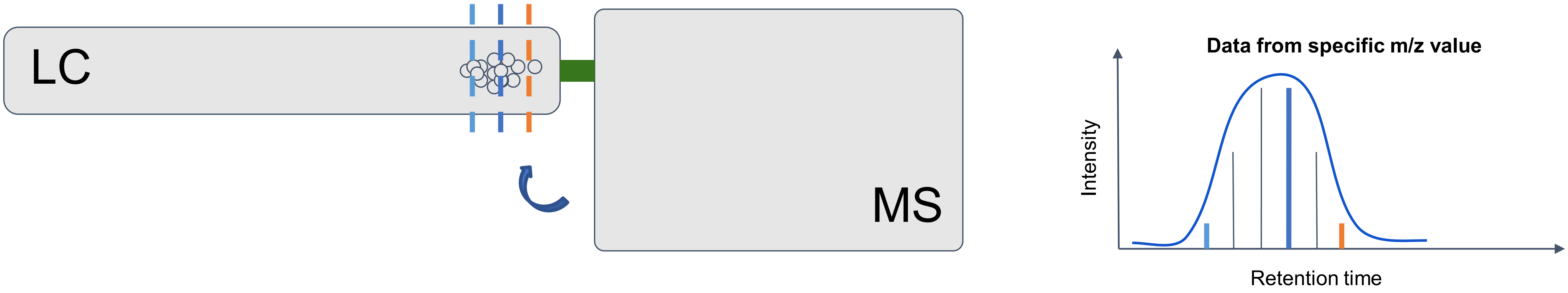
e.g. m/z = 157.0318

$C_7H_8O_2S$;
$C_4H_4N_4O_3$;

Furfuryl thioacetate;
Allantoxanamide;
Orotic acid;
etc…

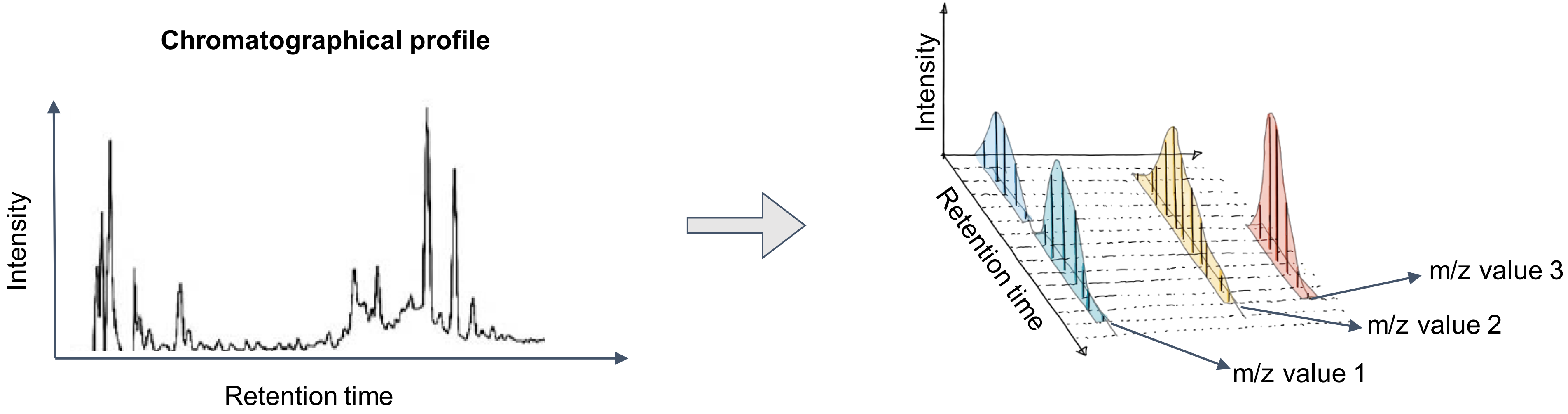
What does MS1 data look like?

The MS performs hundreds of ‘scans’ per second. Each scan measures the m/z values for all compounds that enter the MS during that time window (retention time).



Not all molecules from the same compound exit the column at *exactly* the same time (range). The mass spec will measure different intensities at a specific m/z value over different time points, creating a peak

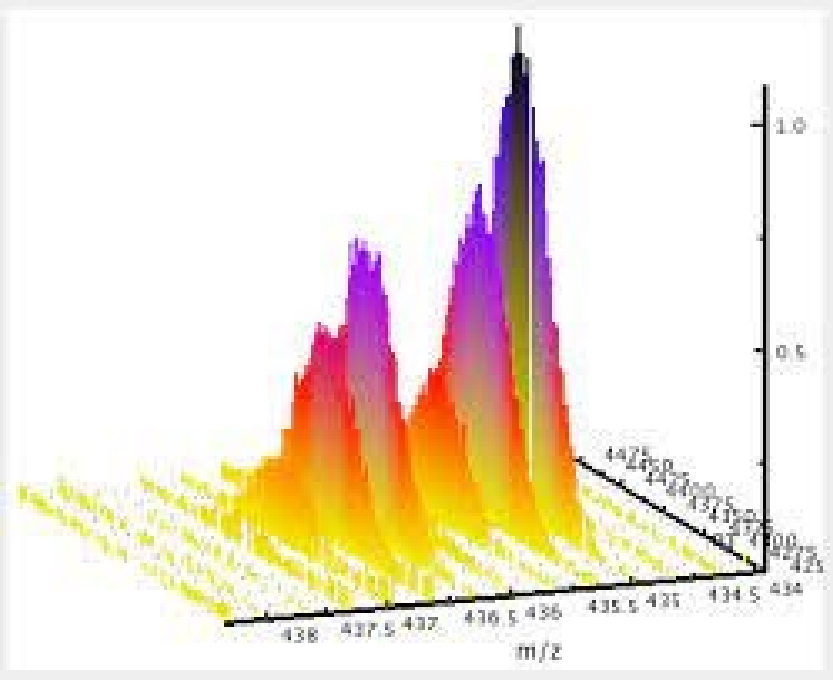
What does MS1 data look like?



This is what the raw data looks like: peaks of intensity values over time, for many different m/z values.
This is the input for the MS1 raw data processing algorithm in MetaboAnalyst.

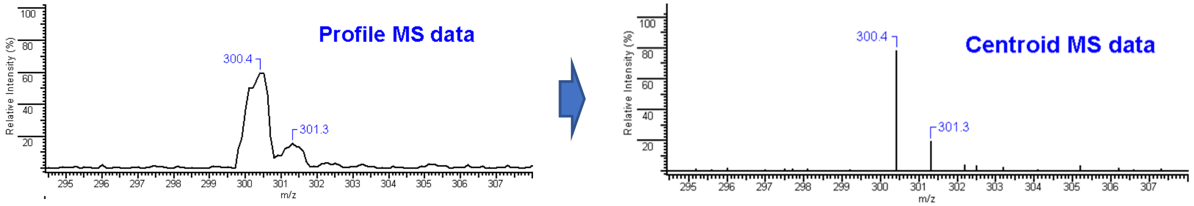
What is raw spectra (pre-)processing?

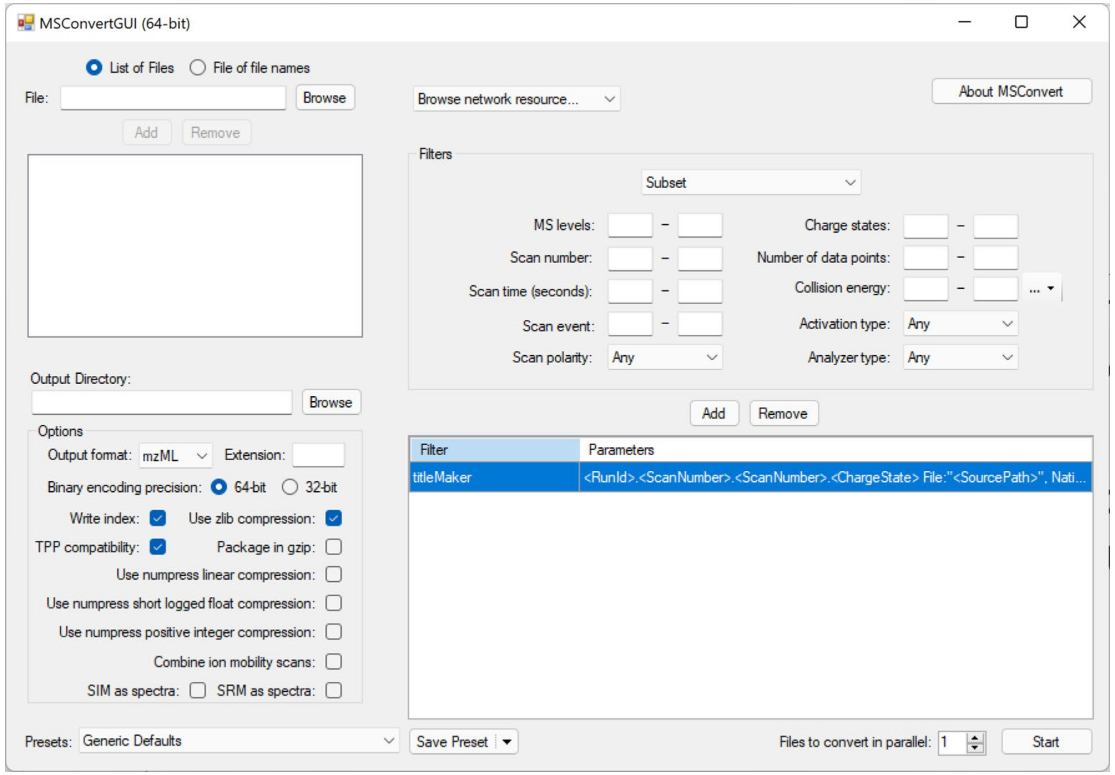
- Convert the raw spectra data from MS instrument into metabolic features (MS peaks);
- Usually contains 2~3 dimensions. For DI-MS, the raw spectra includes m/z and intensity (2D); while for the LC-MS, the raw spectra data includes m/z, retention time and intensity (3D);
- The most common vendor raw data file formats are .raw/.RAW/.wiff/.d/.D/ etc. They need to be converted into open-source format for further processing with open-source software.



Profile or Centroid?

- The vendor raw spectra data is usually in profile format, which is redundant for regular LC-MS based metabolomics analysis;
- We need to convert the MS data into centroid mode to condense the Gaussian Profile peaks into centroids.
- Open-source formats (.mzML/ etc.)..




MSConvertGUI (64-bit)

● List of Files ○ File of file names

File: Browse

Browse network resource...

About MSConvert

Add Remove

Output Directory: Browse

Options

Output format: mzML Extension:

Binary encoding precision: ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: □

Use numpress linear compression: □

Use numpress short logged float compression: □

Use numpress positive integer compression: □

Combine ion mobility scans: □

SIM as spectra: □ SRM as spectra: □

Filters

Subset

MS levels: –

Scan number: –

Scan time (seconds): –

Scan event: –

Scan polarity: Any

Charge states: –

Number of data points: –

Collision energy: –

Activation type: Any

Analyzer type: Any

Add Remove

Filter	Parameters
titleMaker	<RunId>.<ScanNumber>.<ScanNumber>.<ChargeState> File:"<SourcePath>", Nati...

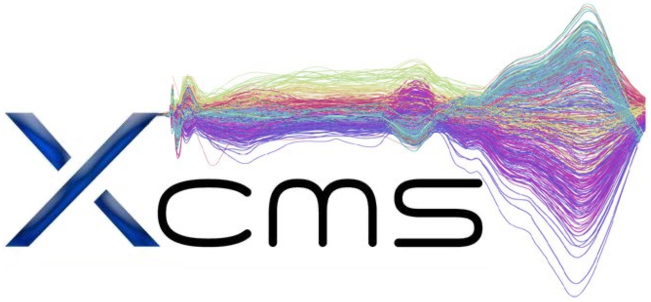
Presets: Generic Defaults

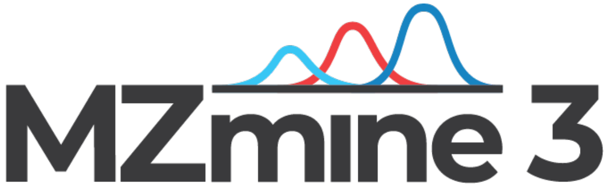
Save Preset

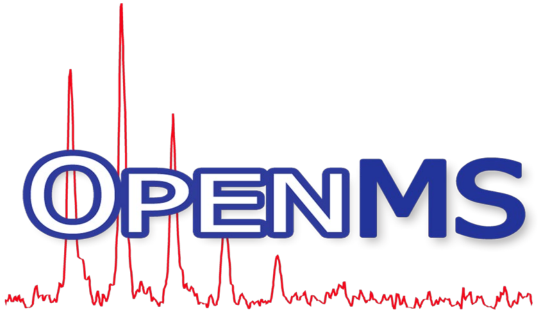
Files to convert in parallel: 1

Start

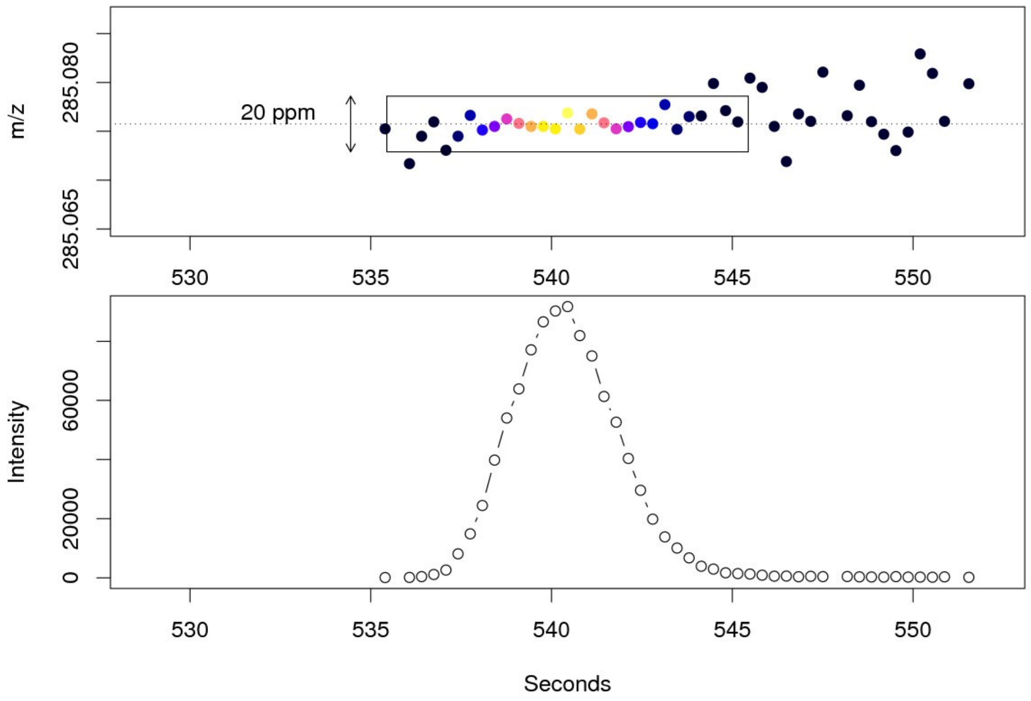
Open-source Software for raw spectra processing..

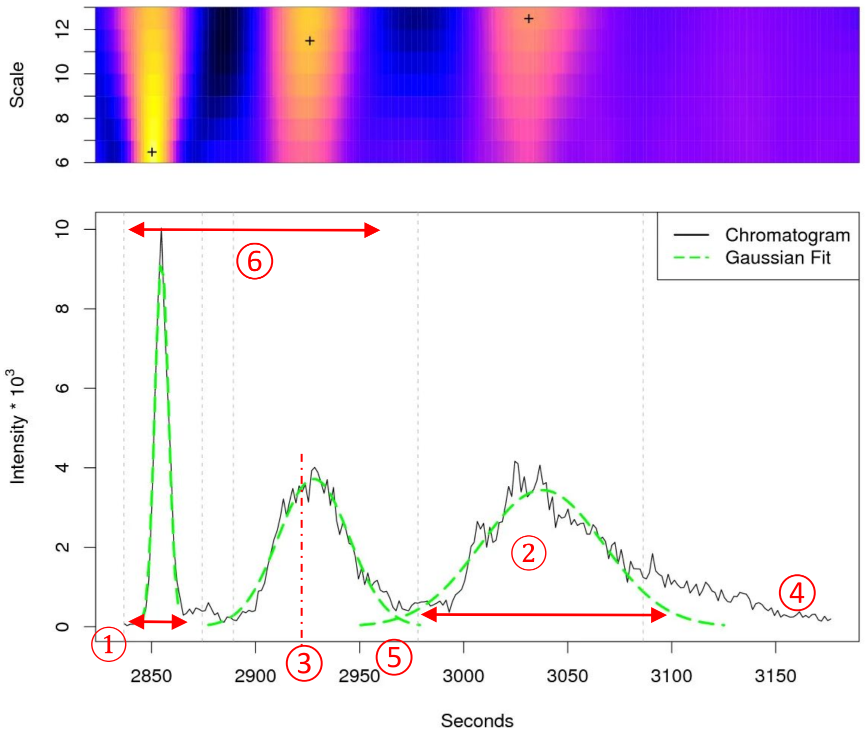




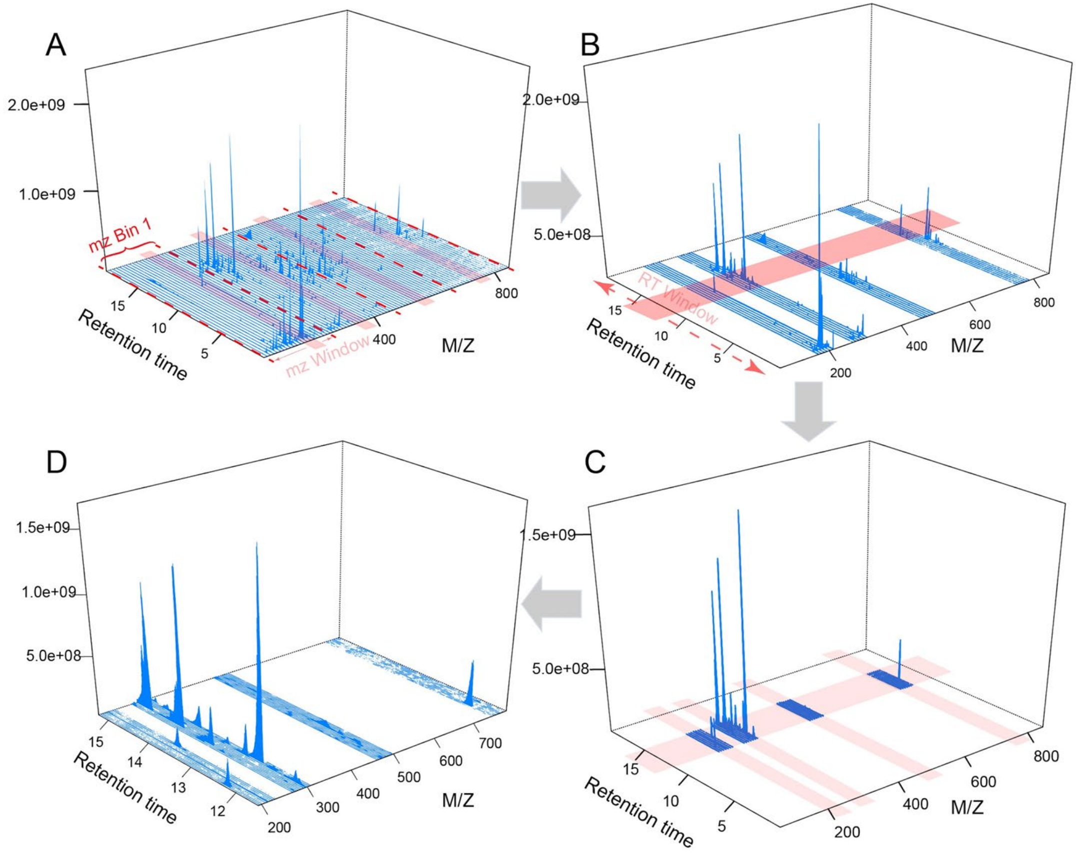


centWave


m/z
285.065
285.080
20 ppm
530
535
540
545
550
Intensity
0
20000
60000
530
535
540
545
550
Seconds


Scale
6
8
10
12
Intensity * 10^3
0
2
4
6
8
10
Chromatogram
Gaussian Fit
①
②
③
④
⑤
⑥
2850
2900
2950
3000
3050
3100
3150
Seconds

ROI Extraction

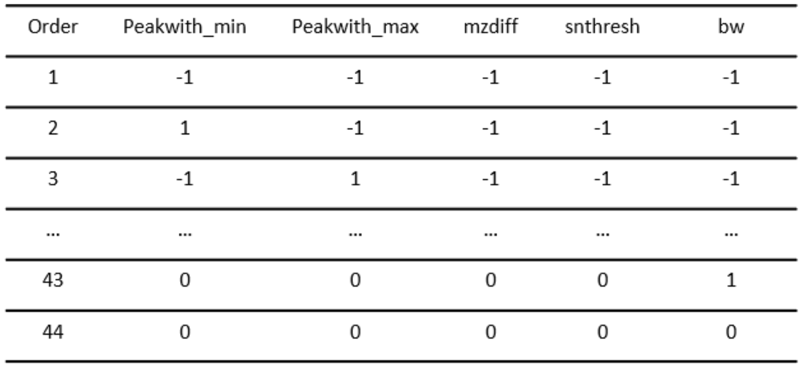


- Data-driven ROI extraction;
- Regions with high abundance of MS signals;
- Both low intensity peaks as well as high intensity peaks will be retained;

DoE-based Parameter Optimization

DoE -- Central composite design

44 runs

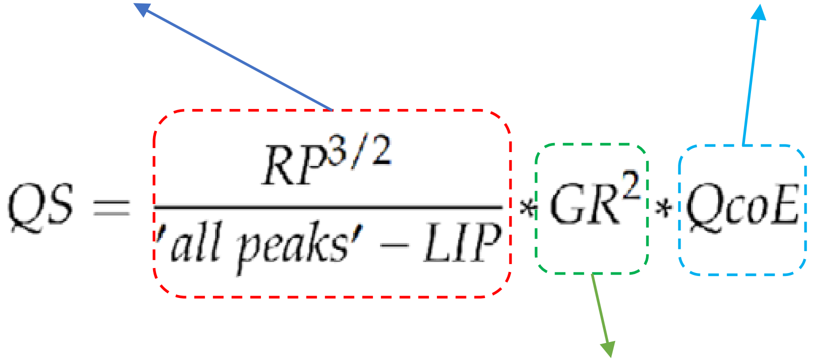


Order	Peakwith_min	Peakwith_max	mzdiff	snthresh	bw
1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1
3	-1	1	-1	-1	-1
...	...	...	...	...	...
43	0	0	0	0	1
44	0	0	0	0	0

3 level for every parameters (-1, 0, 1)

Relative reliable peaks ratio
(identified by their isotopes)

A co-efficient describing the stability
of a grouped feature

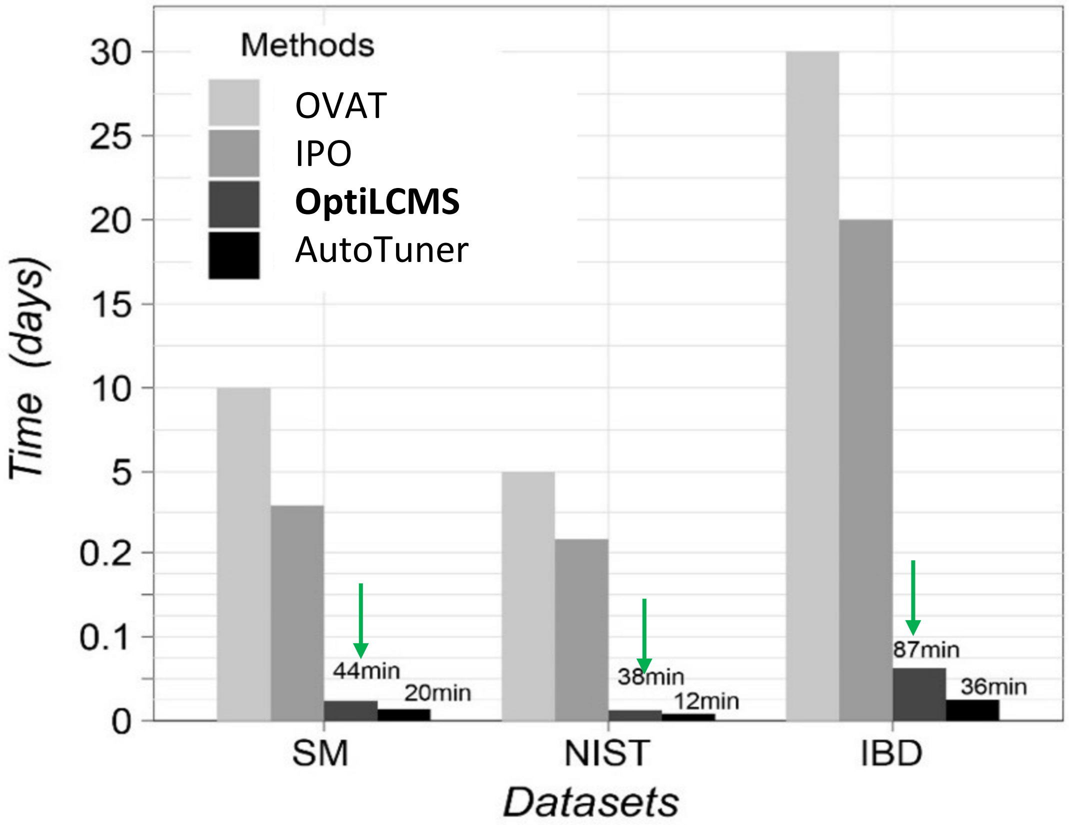


$$QS = \frac{RP^{3/2}}{'all\ peaks' - LIP} * GR^2 * QcoE$$

Gaussian peaks ratio

- The most important parameters are evaluated with 44 DoE runs
- Instead of 3^8 = 6561 one-variable-at-a-time runs.

Performance Evaluation - Speed



- 3 datasets: Standard Mixture (SM), NIST-SRM 1950 and IBD data from iHMP2.

Pang, Z.; Chong, J.; Li, S.; Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **2020**, *10*, 186

Benchmark Studies - 1

Table 1
Qualitative peak picking results of the different tools using different settings.

Methods	Total Peaks	True Peaks	Quantified Consensus	Gaussian Peak Ratio
Default	16,896	382	350	47.8%
IPO	24,346	744	663	52.0%
AutoTuner	25,517	664	603	40.5%
MetaboAnalyst	**18,044**	**799**	**754**	**64.4%**

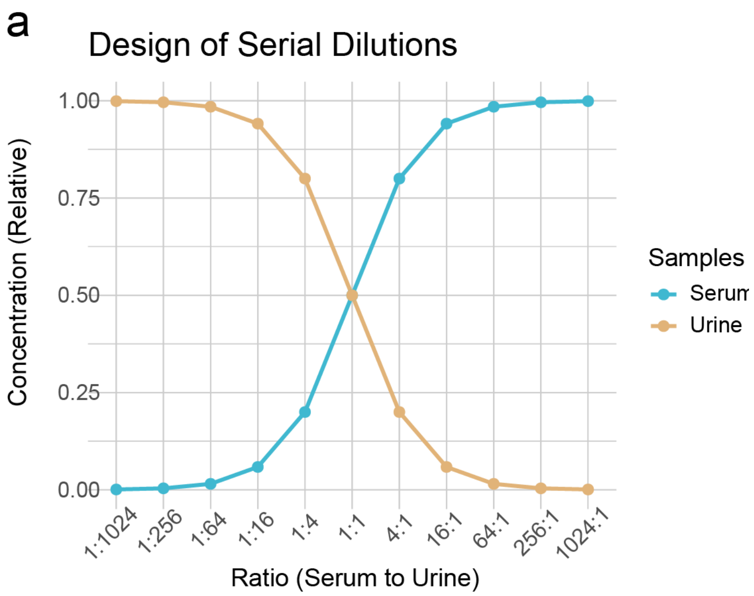
Pang Z, Chong J, Li S, Xia J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. Metabolites. 2020 May 7;10(5):186. doi: 10.3390/metabo10050186. PMID: 32392884; PMCID: PMC7281575.

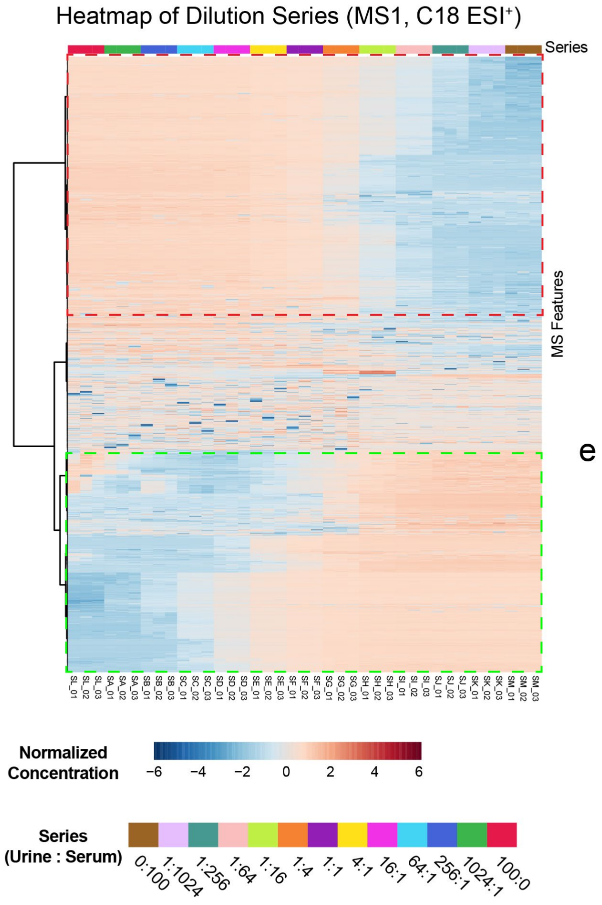
Benchmark Studies - 2

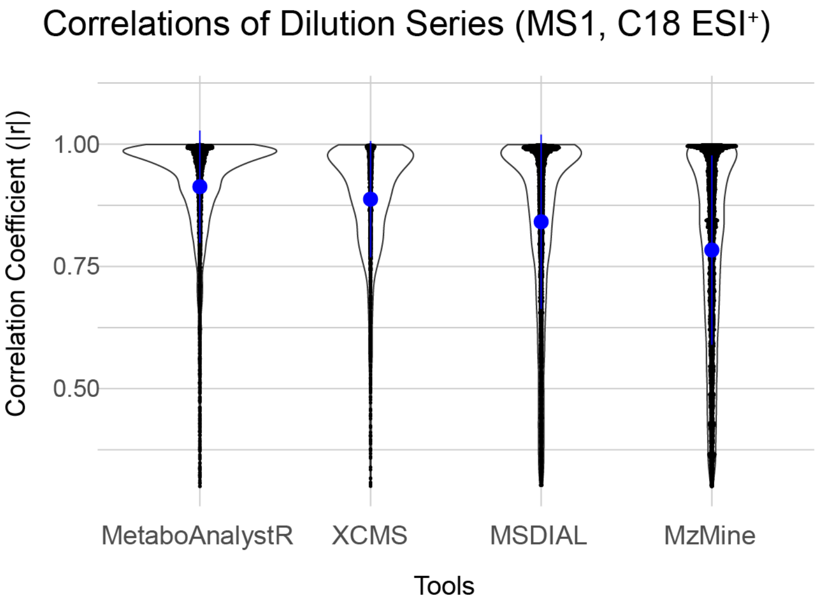
	Default	Optimized
Total peaks	2,492	2,423
Isotopes / Adducts	667 (26.8%)	1,112 (45.9%)
Formula Assigned	663	762
Potential compounds	1,085	1,692
Variance (PC1 + PC2)	37%	50%
Significant peaks	855	1,091

	Default	Optimized
Total peaks	4,344	5,113 (+ 17.7%)
Isotopes	760	1,274 (+ 67.6%)
Adducts	927	1,132 (+ 22.1%)
Formulas assigned	632	687 (+ 8.7%)
Potential compound matches	1,587	1,803 (+ 13.6%)
Variance explained (PC1 + PC2)	76.5%	81.3% (+ 4.8%)

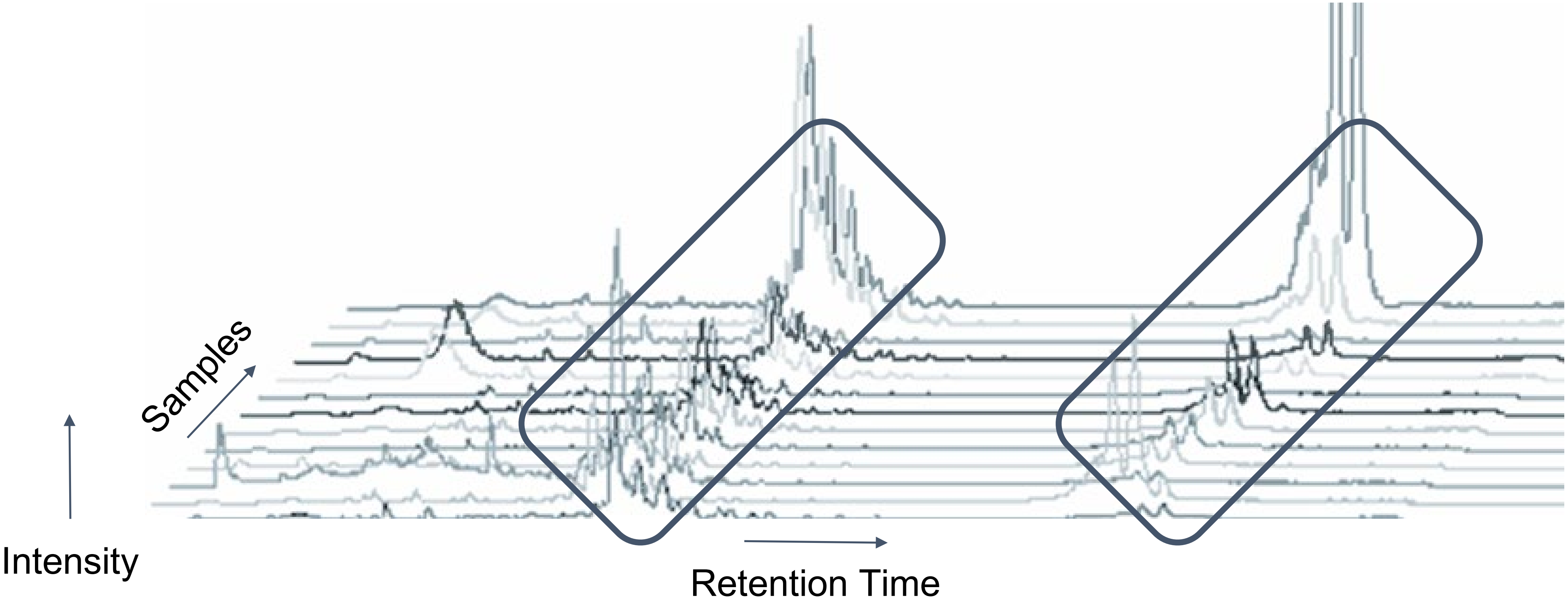
Benchmark Studies - 3





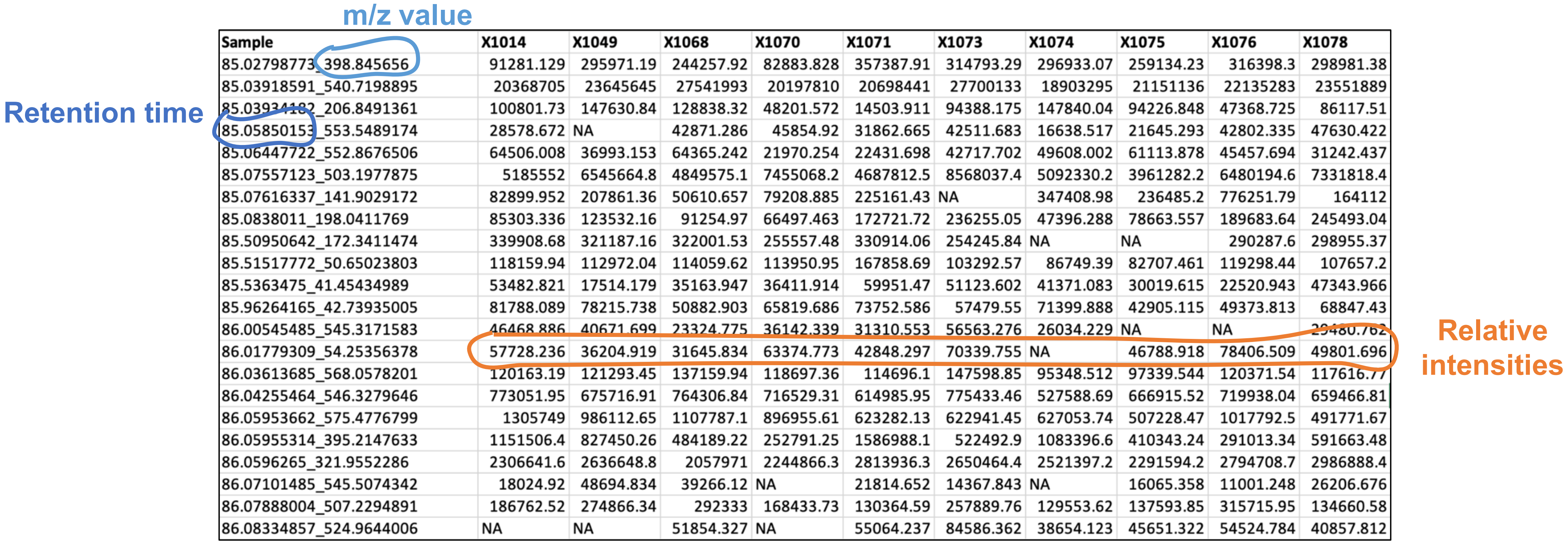


Preprocessing of MS data



Summarize peaks

- Uniquely identify each peak: retention time and m/z value
- Calculate the relative intensity in each sample



m/z value

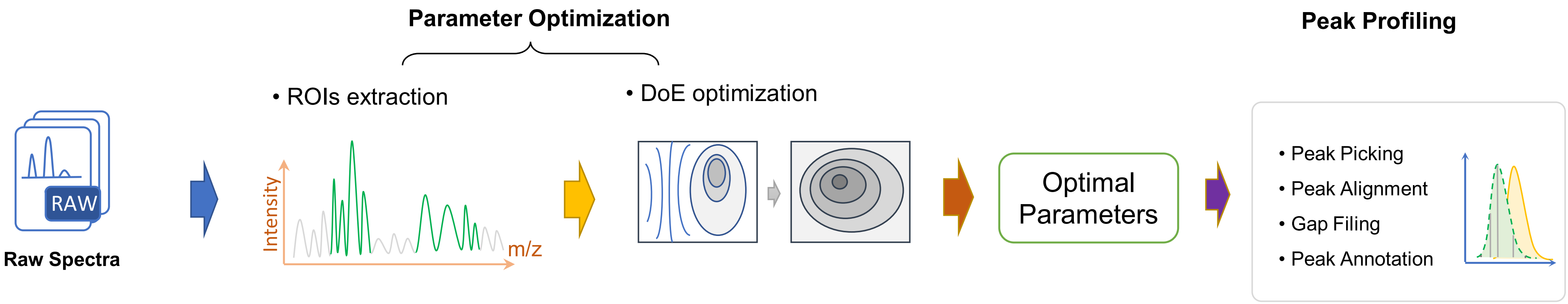
Retention time

Relative intensities

Sample	X1014	X1049	X1068	X1070	X1071	X1073	X1074	X1075	X1076	X1078
85.02798773_398.845656	91281.129	295971.19	244257.92	82883.828	357387.91	314793.29	296933.07	259134.23	316398.3	298981.38
85.03918591_540.7198895	20368705	23645645	27541993	20197810	20698441	27700133	18903295	21151136	22135283	23551889
85.03934182_206.8491361	100801.73	147630.84	128838.32	48201.572	14503.911	94388.175	147840.04	94226.848	47368.725	86117.51
85.05850153_553.5489174	28578.672	NA	42871.286	45854.92	31862.665	42511.683	16638.517	21645.293	42802.335	47630.422
85.06447722_552.8676506	64506.008	36993.153	64365.242	21970.254	22431.698	42717.702	49608.002	61113.878	45457.694	31242.437
85.07557123_503.1977875	5185552	6545664.8	4849575.1	7455068.2	4687812.5	8568037.4	5092330.2	3961282.2	6480194.6	7331818.4
85.07616337_141.9029172	82899.952	207861.36	50610.657	79208.885	225161.43	NA	347408.98	236485.2	776251.79	164112
85.0838011_198.0411769	85303.336	123532.16	91254.97	66497.463	172721.72	236255.05	47396.288	78663.557	189683.64	245493.04
85.50950642_172.3411474	339908.68	321187.16	322001.53	255557.48	330914.06	254245.84	NA	NA	290287.6	298955.37
85.51517772_50.65023803	118159.94	112972.04	114059.62	113950.95	167858.69	103292.57	86749.39	82707.461	119298.44	107657.2
85.5363475_41.45434989	53482.821	17514.179	35163.947	36411.914	59951.47	51123.602	41371.083	30019.615	22520.943	47343.966
85.96264165_42.73935005	81788.089	78215.738	50882.903	65819.686	73752.586	57479.55	71399.888	42905.115	49373.813	68847.43
86.00545485_545.3171583	46468.886	40671.699	23324.775	36142.339	31310.553	56563.276	26034.229	NA	NA	29480.762
86.01779309_54.25356378	57728.236	36204.919	31645.834	63374.773	42848.297	70339.755	NA	46788.918	78406.509	49801.696
86.03613685_568.0578201	120163.19	121293.45	137159.94	118697.36	114696.1	147598.85	95348.512	97339.544	120371.54	117616.77
86.04255464_546.3279646	773051.95	675716.91	764306.84	716529.31	614985.95	775433.46	527588.69	666915.52	719938.04	659466.81
86.05953662_575.4776799	1305749	986112.65	1107787.1	896955.61	623282.13	622941.45	627053.74	507228.47	1017792.5	491771.67
86.05955314_395.2147633	1151506.4	827450.26	484189.22	252791.25	1586988.1	522492.9	1083396.6	410343.24	291013.34	591663.48
86.0596265_321.9552286	2306641.6	2636648.8	2057971	2244866.3	2813936.3	2650464.4	2521397.2	2291594.2	2794708.7	2986888.4
86.07101485_545.5074342	18024.92	48694.834	39266.12	NA	21814.652	14367.843	NA	16065.358	11001.248	26206.676
86.07888004_507.2294891	186762.52	274866.34	292333	168433.73	130364.59	257889.76	129553.62	137593.85	315715.95	134660.58
86.08334857_524.9644006	NA	NA	51854.327	NA	55064.237	84586.362	38654.123	45651.322	54524.784	40857.812

Workflow Overview

- Our optimization approach is designed to extract a region abundant with MS signals for a design of experiment (DoE)-based optimization.



Raw Spectra

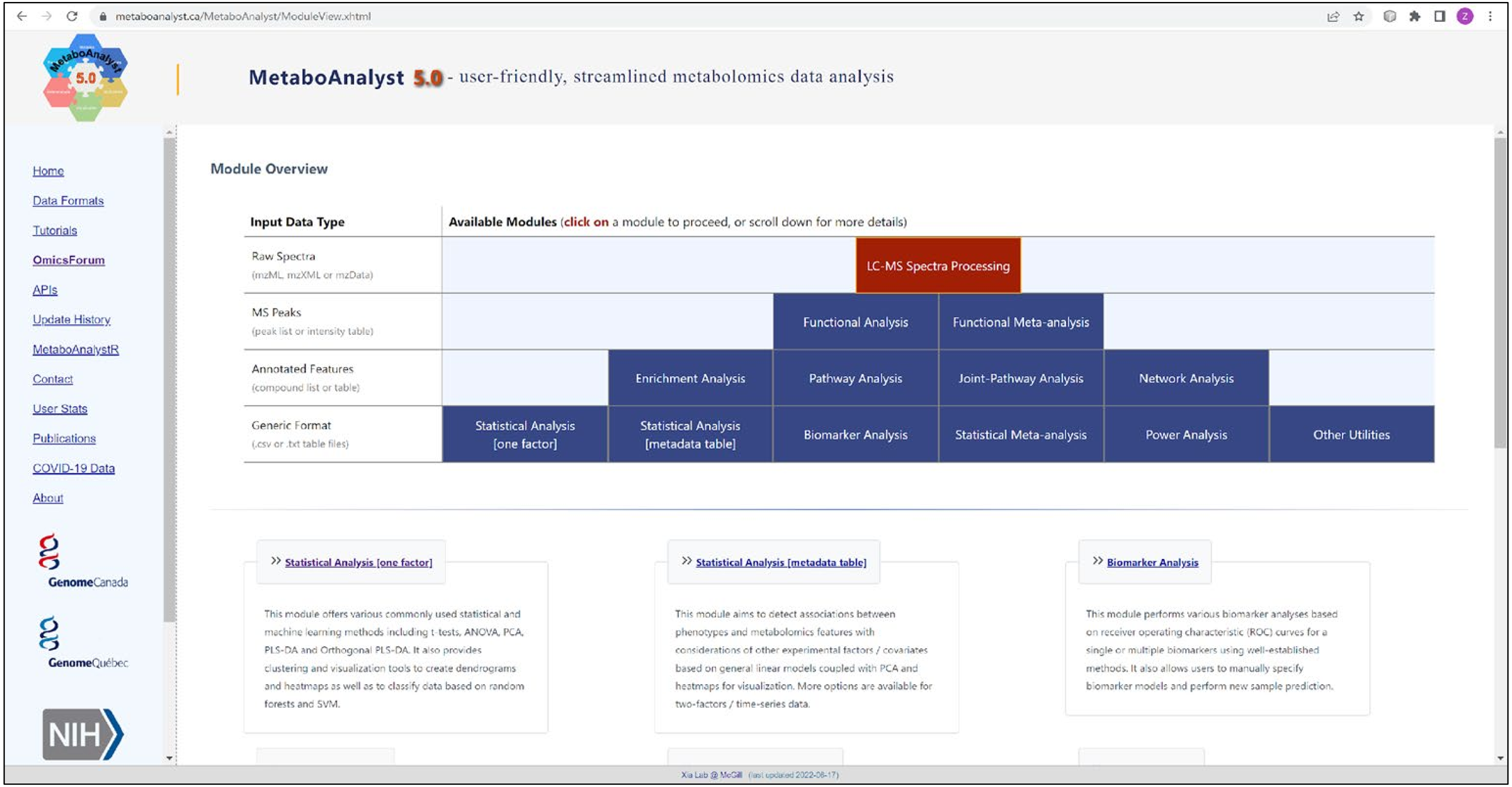
Parameter Optimization

- ROIs extraction
- DoE optimization

Optimal Parameters

Peak Profiling

- Peak Picking
- Peak Alignment
- Gap Filing
- Peak Annotation



MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

- Home
- Data Formats
- Tutorials
- OmicsForum
- APIs
- Update History
- MetaboAnalystR
- Contact
- User Stats
- Publications
- COVID-19 Data
- About

Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)			LC-MS Spectra Processing			
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

>> Statistical Analysis [one factor]

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-DA and Orthogonal PLS-DA. It also provides clustering and visualization tools to create dendrograms and heatmaps as well as to classify data based on random forests and SVM.

>> Statistical Analysis [metadata table]

This module aims to detect associations between phenotypes and metabolomics features with considerations of other experimental factors / covariates based on general linear models coupled with PCA and heatmaps for visualization. More options are available for two-factors / time-series data.

>> Biomarker Analysis

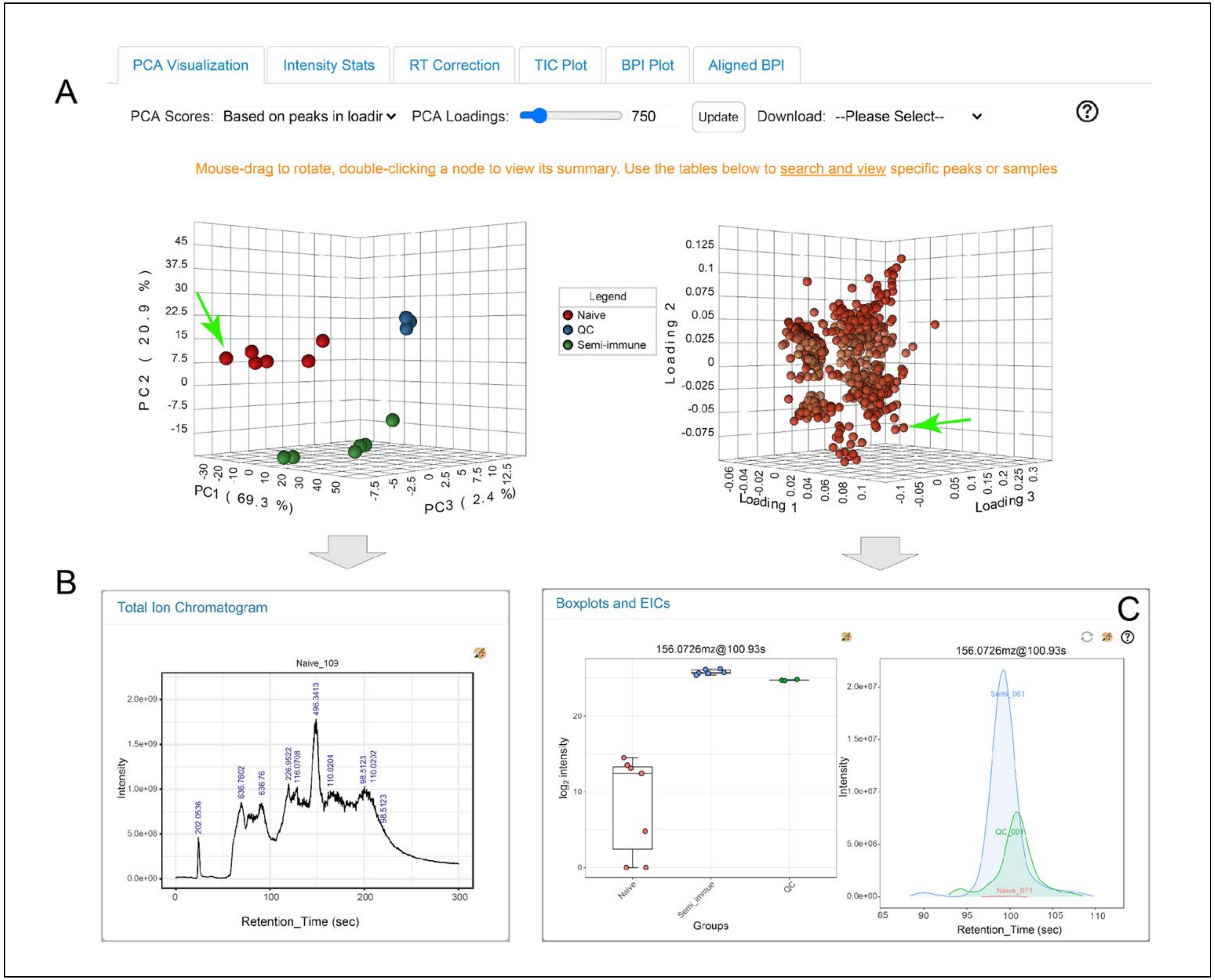
This module performs various biomarker analyses based on receiver operating characteristic (ROC) curves for a single or multiple biomarkers using well-established methods. It also allows users to manually specify biomarker models and perform new sample prediction.

Xia Lab @ McGill (last updated 2022-06-17)

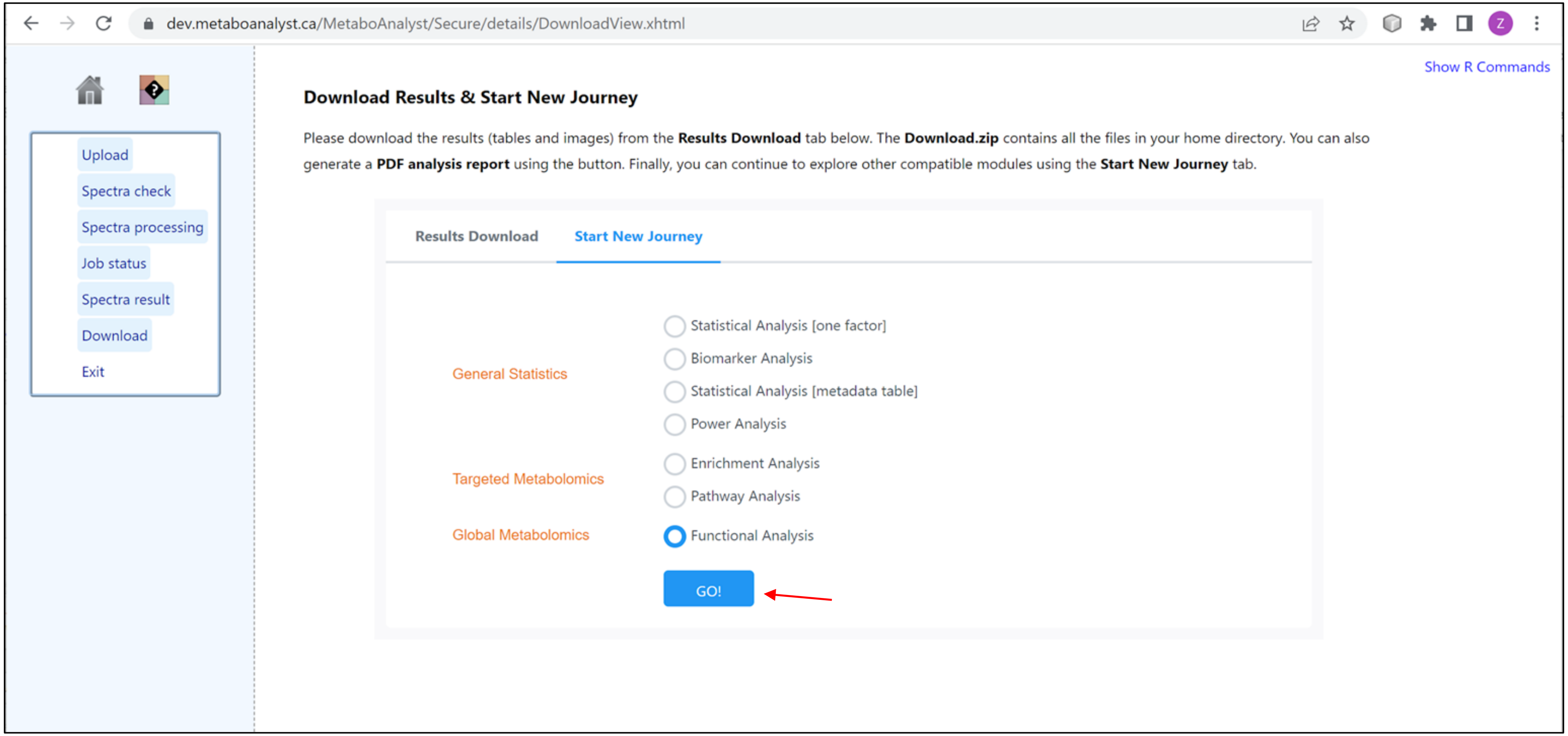
Functional Utilities

- Raw Data Uploading (.mzML/.mzXML/.mzData/.cdf);
- Centroiding on the fly;
- Parameters Optimization (automatically);
- Peak Profiling (Peak Picking/Alignment/Gap filling);
- Peak Annotation (Adducts + Isotopes);
- Putative Compound Mapping;
- Result visualization…

Result Demo..


PCA Visualization
Intensity Stats
RT Correction
TIC Plot
BPI Plot
Aligned BPI
A
PCA Scores: Based on peaks in loadir
PCA Loadings: 750
Update
Download: --Please Select--
Mouse-drag to rotate, double-clicking a node to view its summary. Use the tables below to search and view specific peaks or samples
PC2 (20.9 %)
PC1 (69.3 %)
PC3 (2.4 %)
Legend
Naive
QC
Semi-immune
Loading 1
Loading 2
Loading 3
B
Total Ion Chromatogram
Naive_109
Intensity
Retention_Time (sec)
202.0536
636.7602
636.76
226.9522
116.0708
496.3413
110.0204
98.5123
110.0202
98.5123
C
Boxplots and EICs
156.0726mz@100.93s
log2 intensity
Groups
Naive
Semi_immue
QC
156.0726mz@100.93s
Intensity
Retention_Time (sec)
Semi_061
QC_001
Naive_071

From raw spectra into biological insight



Download Results & Start New Journey

Please download the results (tables and images) from the **Results Download** tab below. The **Download.zip** contains all the files in your home directory. You can also generate a **PDF analysis report** using the button. Finally, you can continue to explore other compatible modules using the **Start New Journey** tab.

General Statistics
- ○ Statistical Analysis [one factor]
- ○ Biomarker Analysis
- ○ Statistical Analysis [metadata table]
- ○ Power Analysis

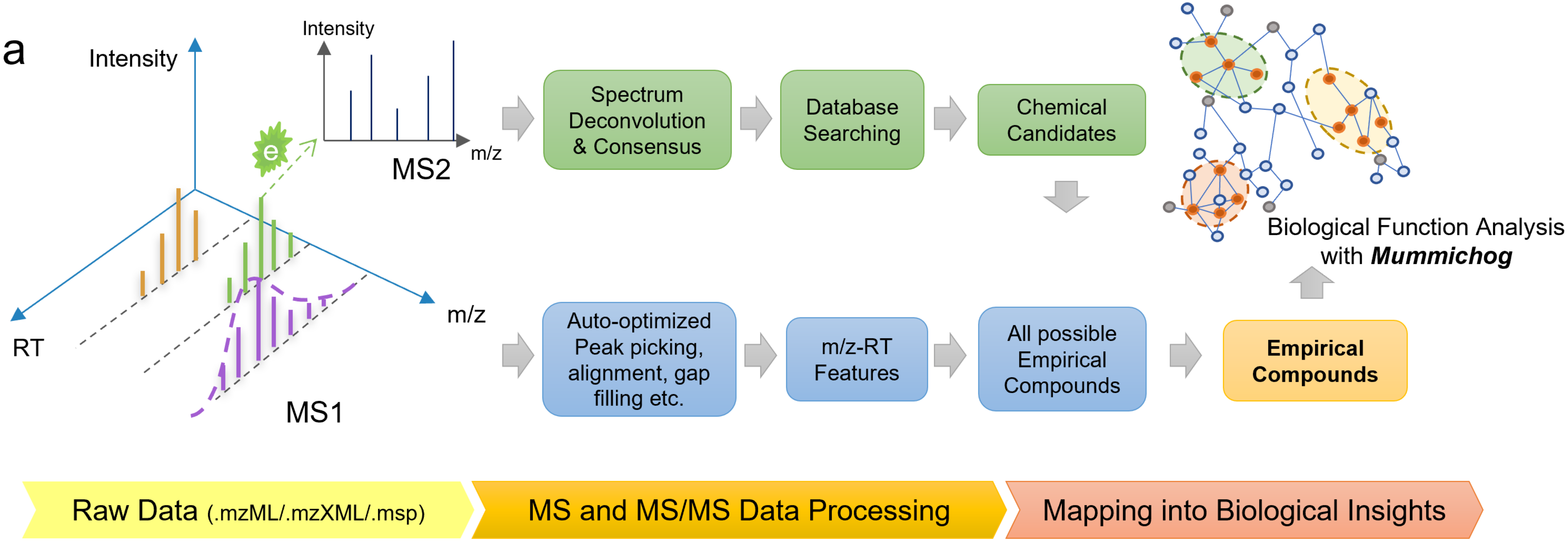
Targeted Metabolomics
- ○ Enrichment Analysis
- ○ Pathway Analysis

Global Metabolomics
- ● Functional Analysis

Demo Time

Raw data centroiding and conversion: 1 min;
Raw spectral data uploading and Parameters’ setting: 2min;
Results Exploration: 2min;

Next version: MetaboAnalyst 6



a
Intensity
RT
m/z
MS1
e
Intensity
m/z
MS2
Spectrum Deconvolution & Consensus
Database Searching
Chemical Candidates
Auto-optimized Peak picking, alignment, gap filling etc.
m/z-RT Features
All possible Empirical Compounds
Empirical Compounds
Biological Function Analysis with *Mummichog*
Raw Data (.mzML/.mzXML/.msp)
MS and MS/MS Data Processing
Mapping into Biological Insights

Learning Questions..

- Which software is encouraged to use for raw data centroiding?
- What is the algorithm MetaboAnalyst is using for MS data preprocessing? And why?
- Generate an EIC plot (overlay of at least one sample from the three groups) for the most significant peak?

Tutorials

- https://github.com/xia-lab/Metabolomics_2023
- Publication: https://www.nature.com/articles/s41596-022-00710-w
- Or our manuscript: https://www.dropbox.com/s/7184c4dheeiiz2p/NP-MetaboAnalyst-2022.pdf?dl=0
    - **Stage 1: LC–HRMS raw spectra processing;**
- Questions? → https://www.omicsforum.ca/
- If your question is not covered, please create a new topic – we will try to answer them in the coming days.

Caution:

1. For raw spectra processing, you are strongly encouraged to use 1st example rather than the 2nd one to avoid waiting in queue for learning purpose;
2. Avoid downloading and uploading any example raw spectra data due to the limited bandwidth.
3. Default MetaboAnalyst includes **www.metaboanalyst.ca , new.metaboanalyst.ca and genap.metaboanalyst.ca.** please use either of them.