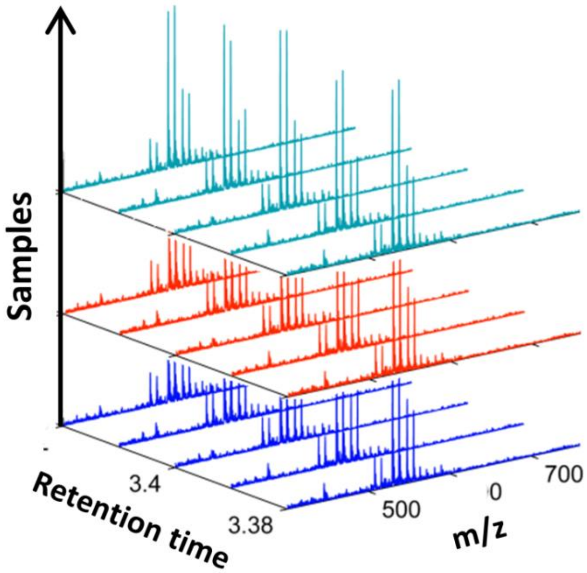
LC-MS DATA PREPROCESSING

Gözde Gürdeniz



Outline

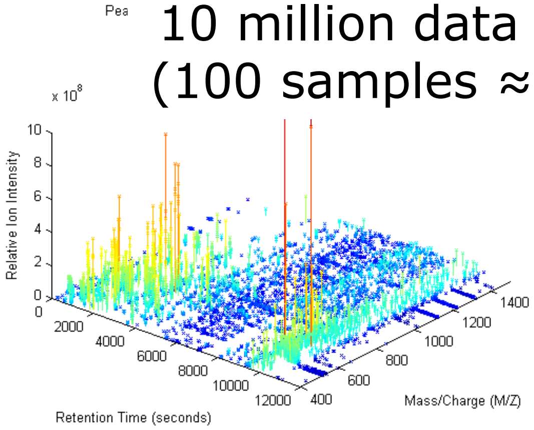
✓ LC-MS Data preprocessing pipeline (MZmine)

1. Peak detection
2. Deisotoping
3. Alignment
4. Gap filling

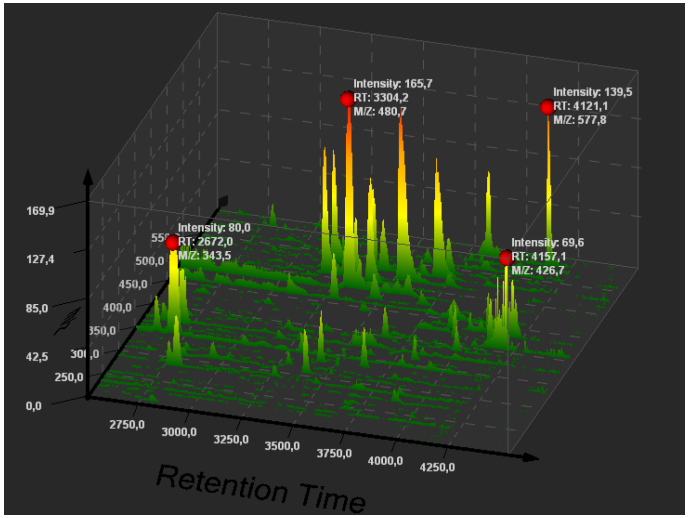
✓ Conclusions

Data Preprocessing : Data Reduction

10 million data points
(100 samples ≈ 4 GB)



Feature detection and Alignment



- **Identification**
  - **Annotation**

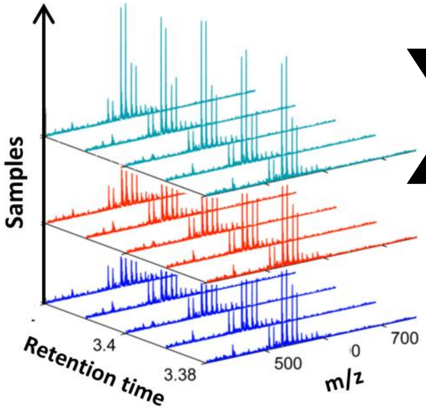
500-1500 compounds

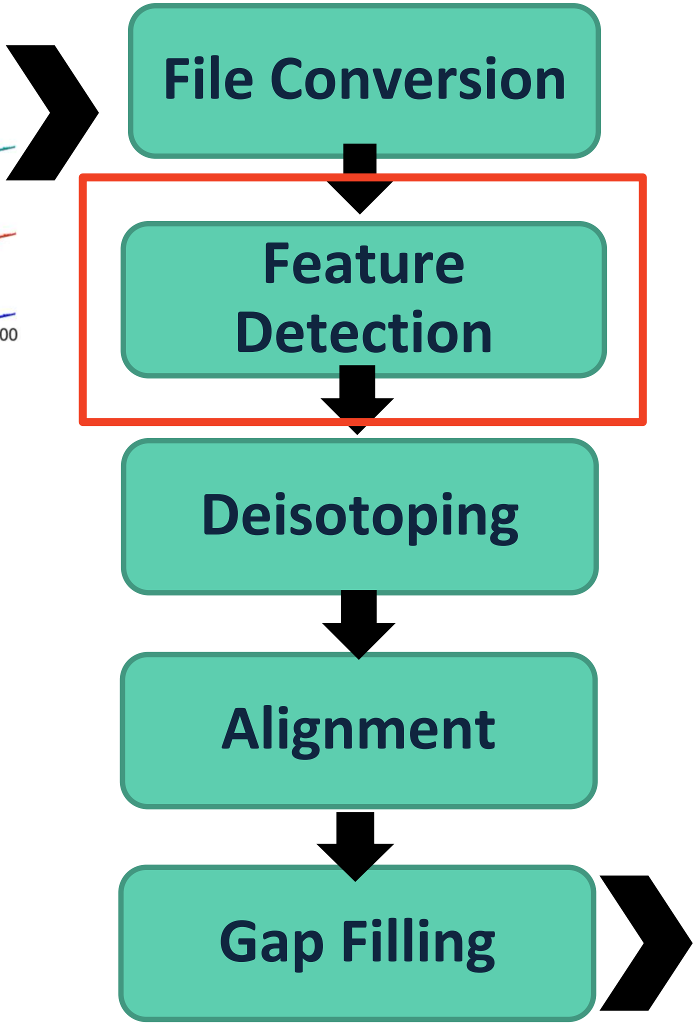
id	mz	rt	isotopes	adduct	pc
65	176.04	280.09			
76	136.05	280.43	[14][M+1]1+		5
77	135.05	280.43	[14][M]1+		5
74	153.06	280.43		[M+H]+ 152.05437	5
75	175.04	280.43		[M+Na]+ 152.05437	5
73	197.02	280.76		[M+2Na-H]+ 152.05437	5
78	377.74	286.15			
79	732.5	286.49			
83	488.32	286.82		[M+Na]+ 465.33205	7
82	466.34	286.82		[M+H]+ 465.33205	7
...					

1000 - 10000 features
(100 samples ≈ 1 MB)

Data Preprocessing Pipeline

RAW DATA





File Conversion

Feature Detection

Deisotoping

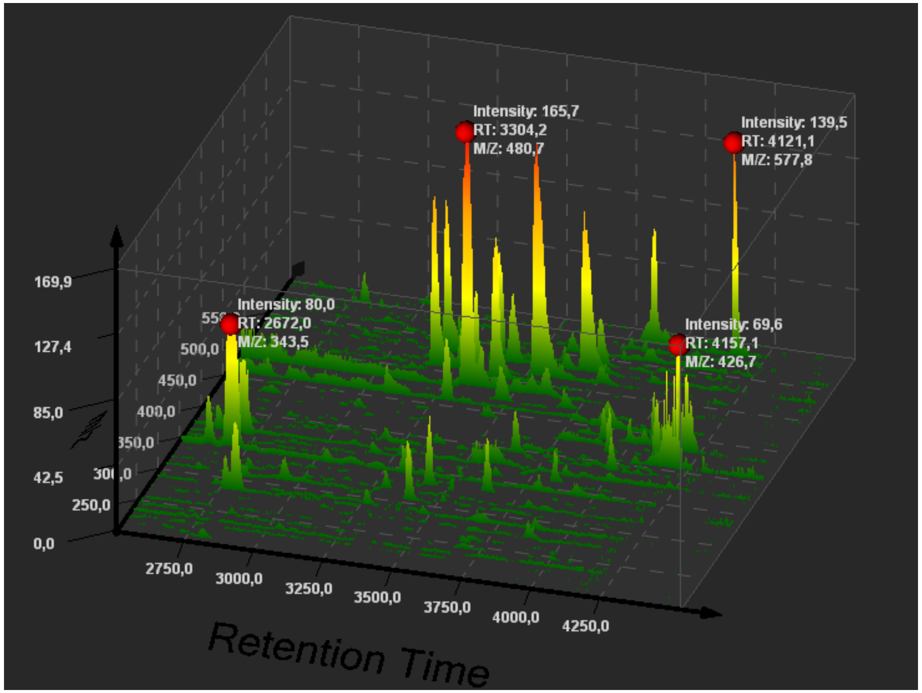
Alignment

Gap Filling

PREPROCESSED DATA

	Feature 1	Feature 2	Feature n
Ret.T	0.81	0.82	…
m/z	50.57	100.85	…
Sp. 1	45534	5445	…
Sp. 2	54	425	…
Sp. 3	561	538	…

Feature Detection



Aim

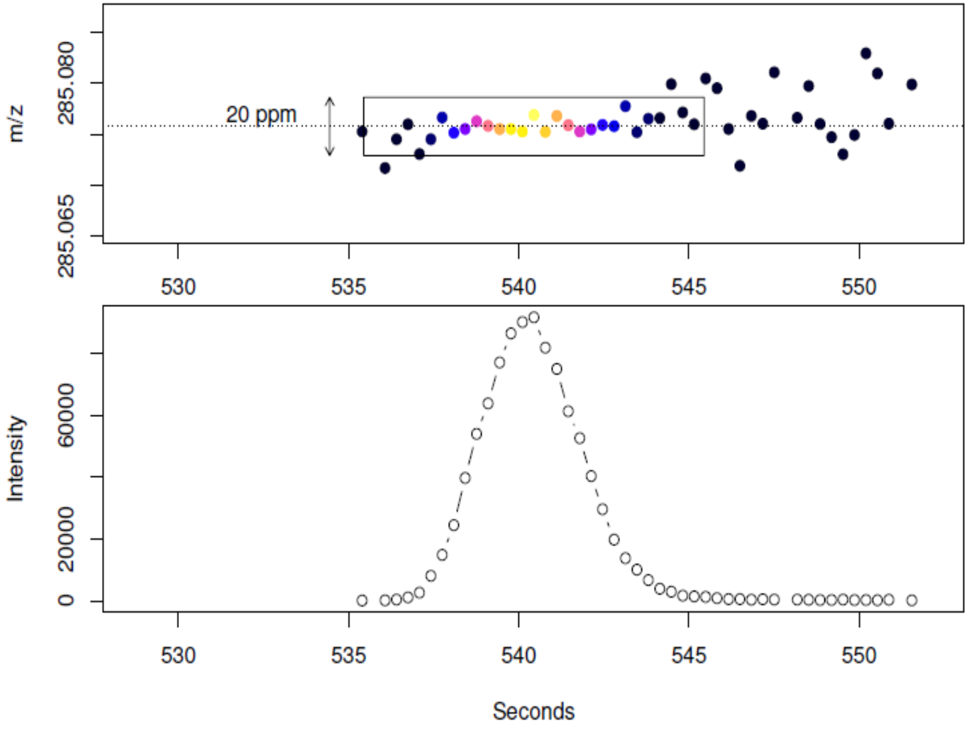
- ✓ Data reduction
- ✓ Identification and quantification of true signals
- ✓ Avoid noise-induced signals
- ✓ Precise quantification

Feature : 2D-signal induced by a single ion species of a compound (e.g. [M+H]+)

Feature detection : *(1) Detection of mass signals*

✓ Build continuous chromatograms by defining **m/z window**

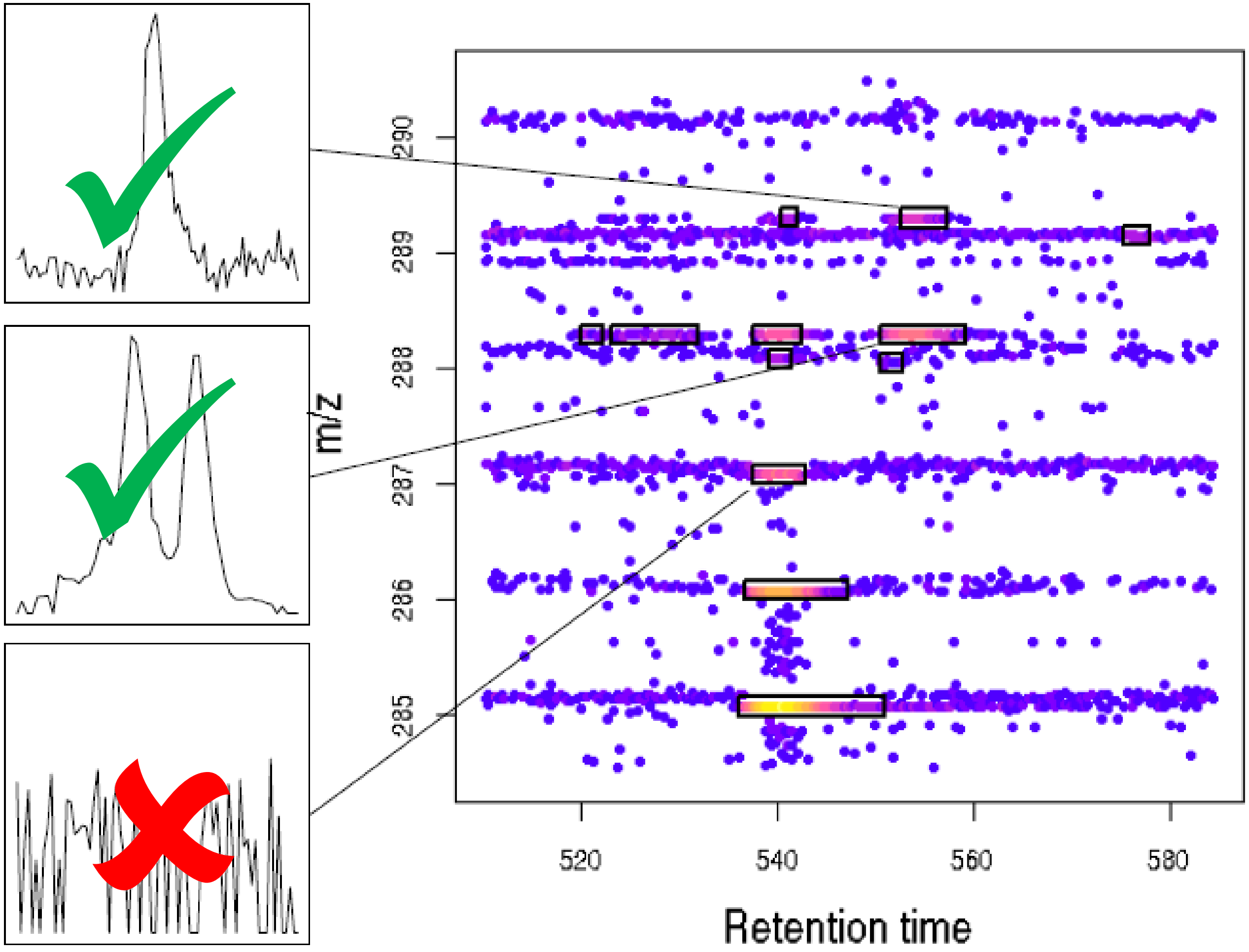
✓ Check its length (**time span**) and intensity (**height**)



Restrictions:

- min-max peak width
- minimum peak height
- signal/noise

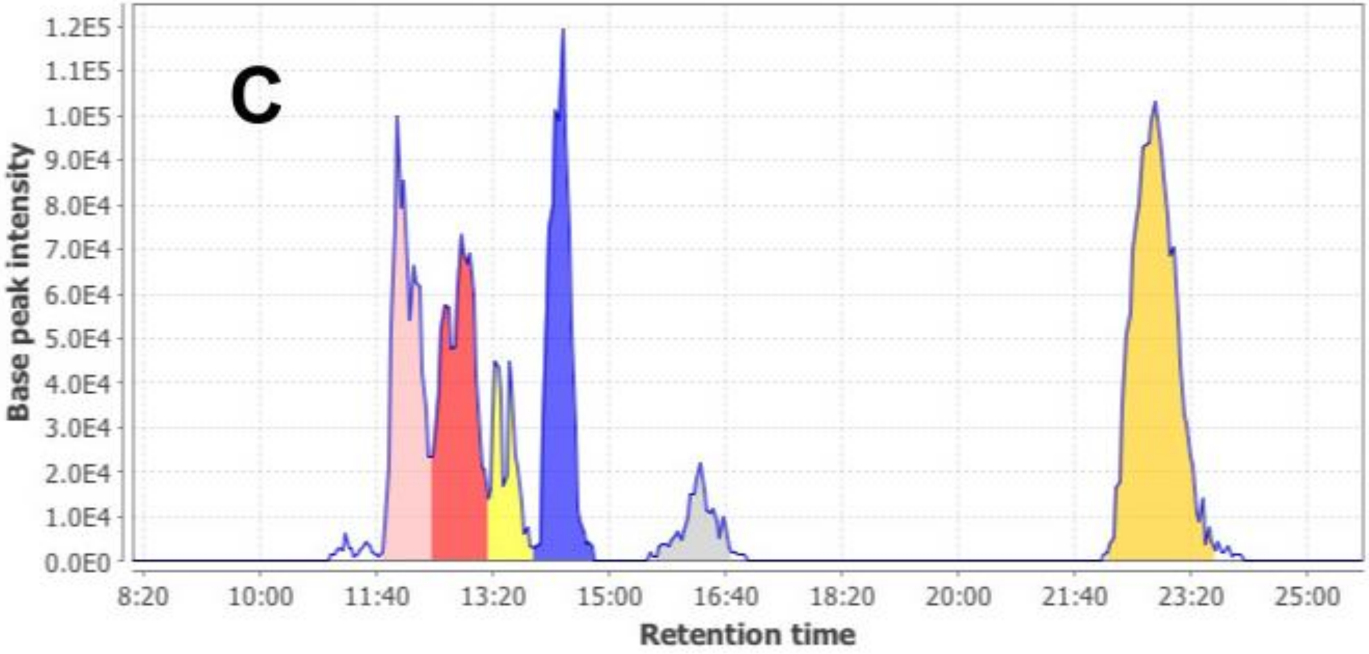
Feature detection : *(2) Detection of chromatographic peaks*



Feature detection : *(2) Peak detection and deconvolution*

- ✓ To detect and quantify individual peaks in chromatograms

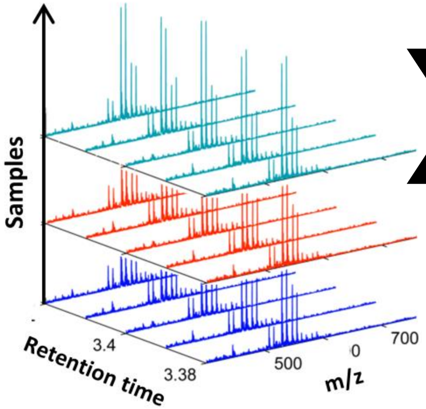
m/z = 356.585 +/- 0.01

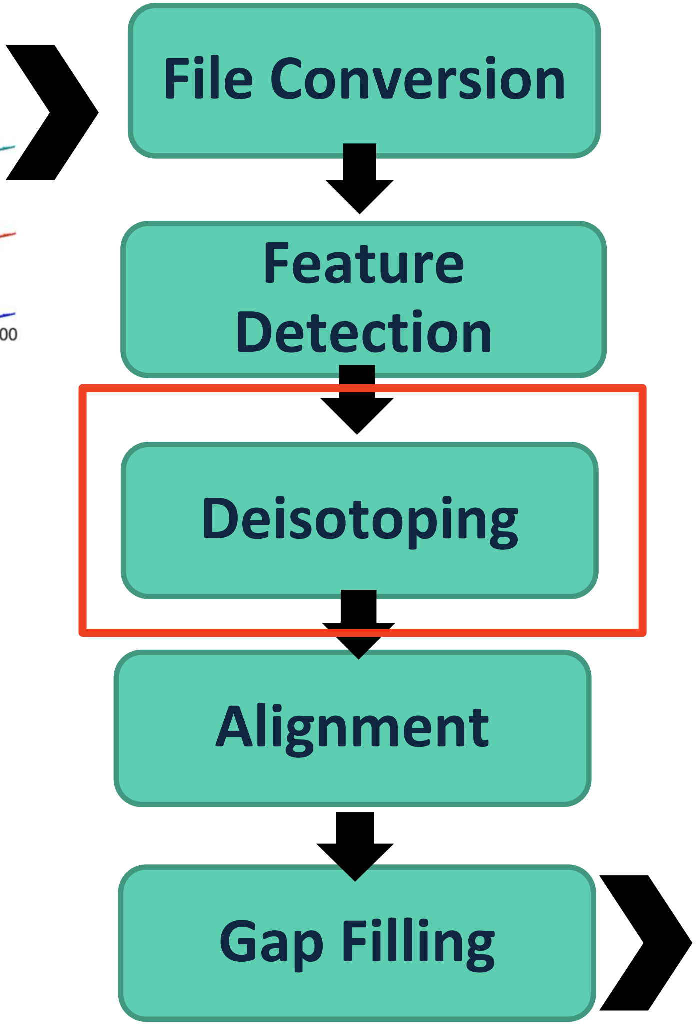


(MZmine) Local minimum search :
Define parameters such as min height, min peak width

Data Preprocessing Pipeline

RAW DATA




File Conversion

Feature Detection

Deisotoping

Alignment

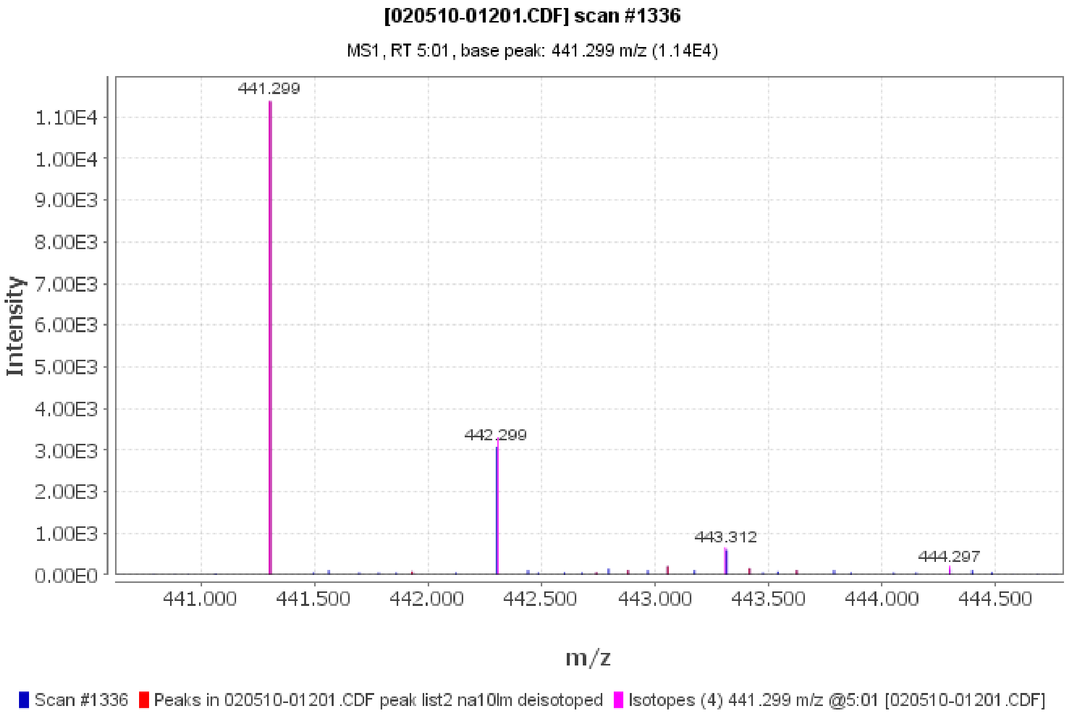
Gap Filling

PREPROCESSED DATA

	Feature 1	Feature 2	Feature n
Ret.T	0.81	0.82	…
m/z	50.57	100.85	…
Sp. 1	45534	5445	…
Sp. 2	54	425	…
Sp. 3	561	538	…

Deisotoping (optional)

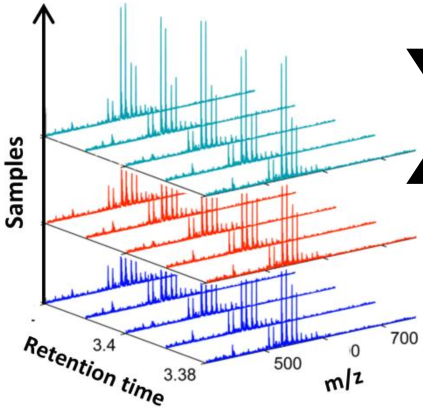
- ✓ Redundant info for data analysis
- ✓ Useful for identification

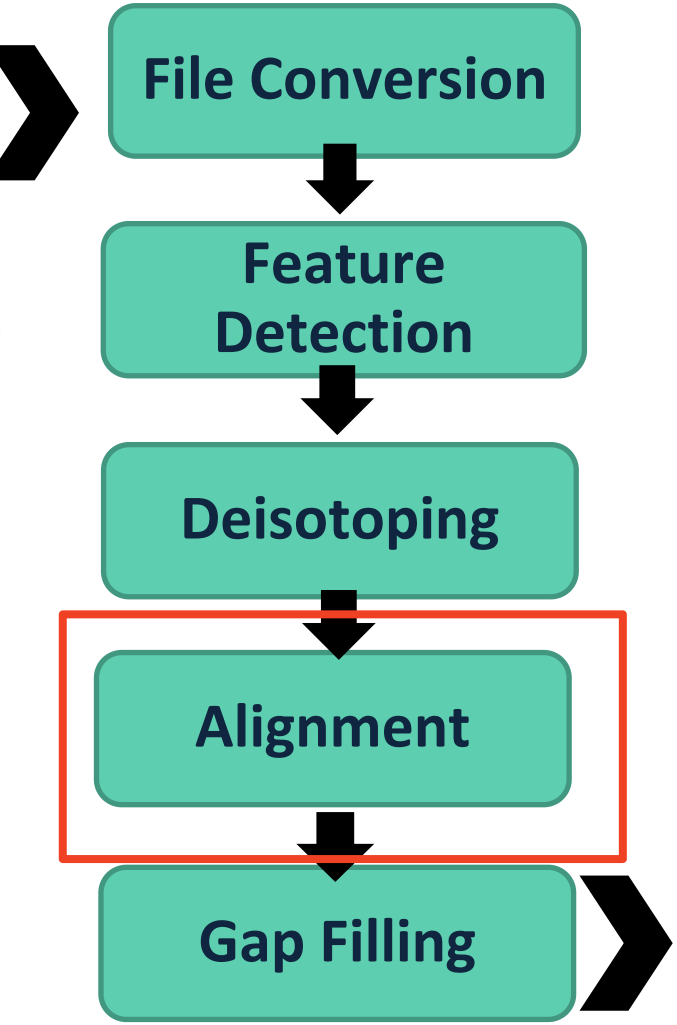


- ✓ **MZmine -** m/z and RT tolerance
- ✓ **XCMS** - CAMERA , m/z tolerance

Data Preprocessing Pipeline

RAW DATA




File Conversion

Feature Detection

Deisotoping

Alignment

Gap Filling

PREPROCESSED DATA

	Feature 1	Feature 2	Feature n
Ret.T	0.81	0.82	…
m/z	50.57	100.85	…
Sp. 1	45534	5445	…
Sp. 2	54	425	…
Sp. 3	561	538	…

Peak List Alignment

Sample 1

	Ret. Time	m/z	Height /Area
Feature 1	0.81	58.545	805.12
Feature 2	0.94	75.1685	240.52
-	-	-	-
Feature n	5.45	750.35	1052.45

Sample 2

	Ret. Time	m/z	Height /Area
Feature 1	0.82	58.585	500.12
Feature 2	0.98	75.161	40.59
-	-	-	-
Feature n	5.48	750.35	9152.55

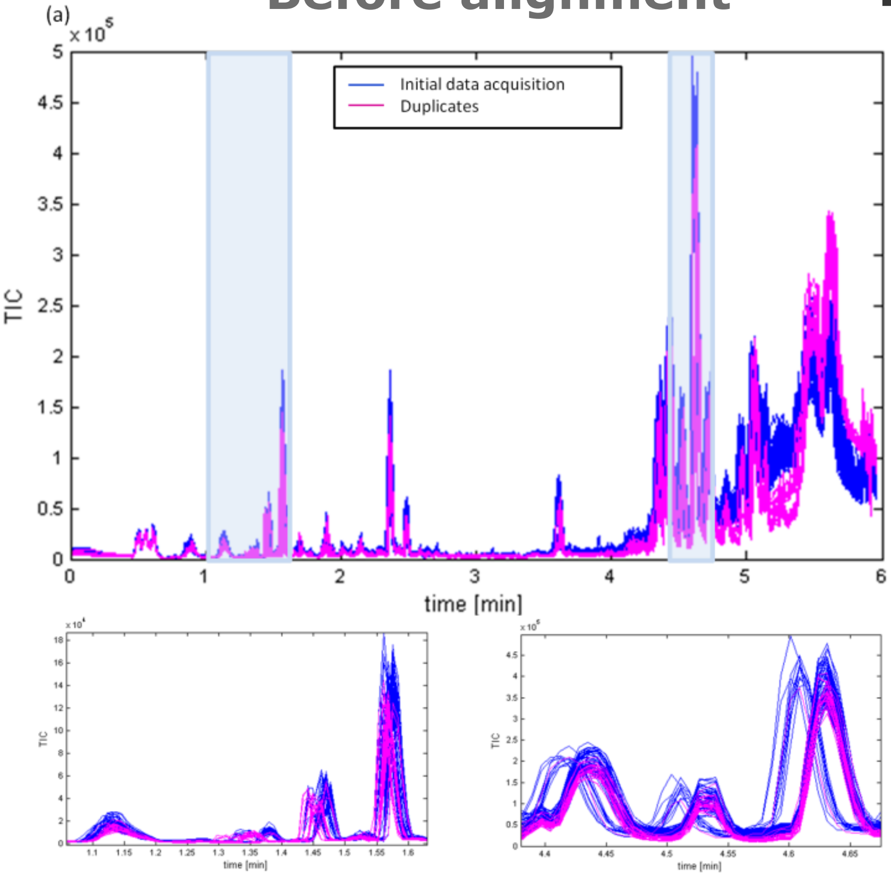
Matched Peak List

	Ret. Time	m/z	Samp1 Height/ Area	Samp2 Height/ Area
Feature 1	0.81	58.565	500.12	805.12
Feature 2	0.96	75.1668	40.59	240.52
-				
Feature 2	5.46	750.35	9152.55	1052.45

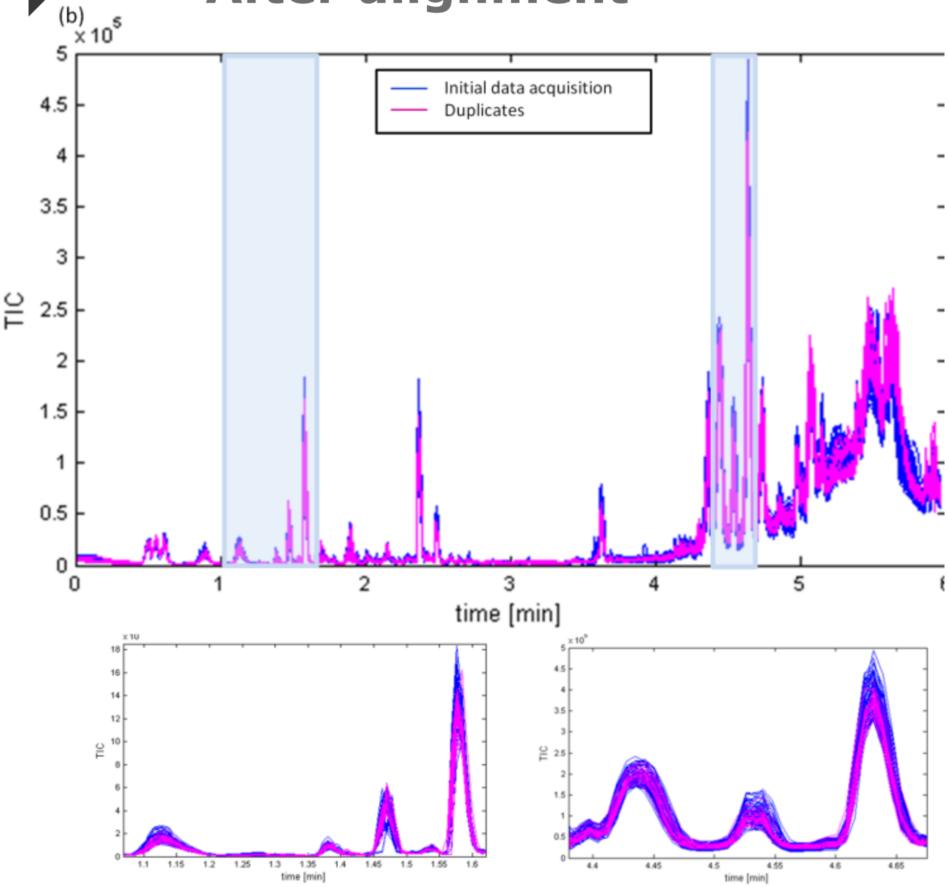
Retention time shifts:

- ✓ Pressure, temperature and flow rate fluctuations
- ✓ Matrix effects
- ✓ Stationary phase decomposition

Before alignment

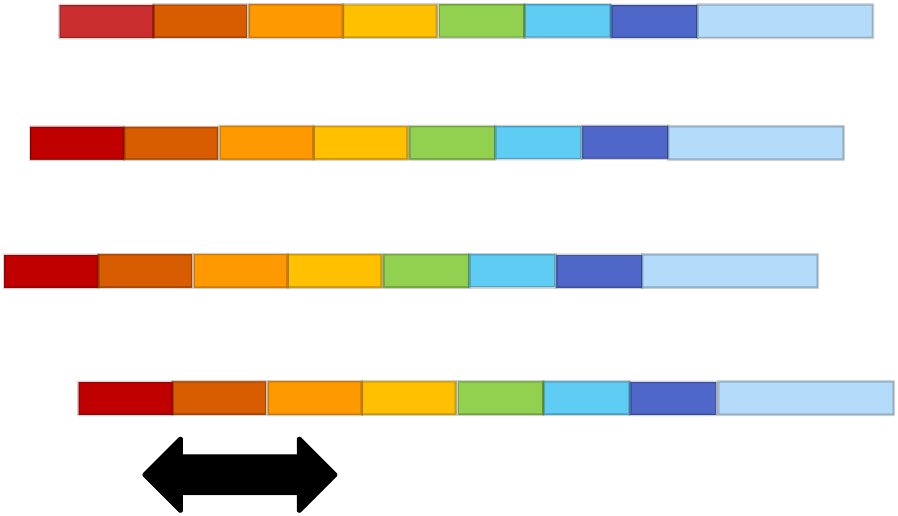


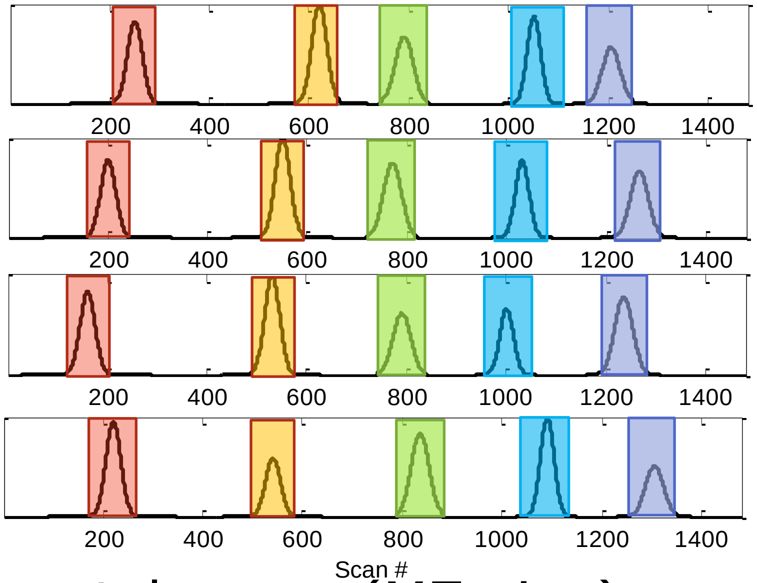
After alignment



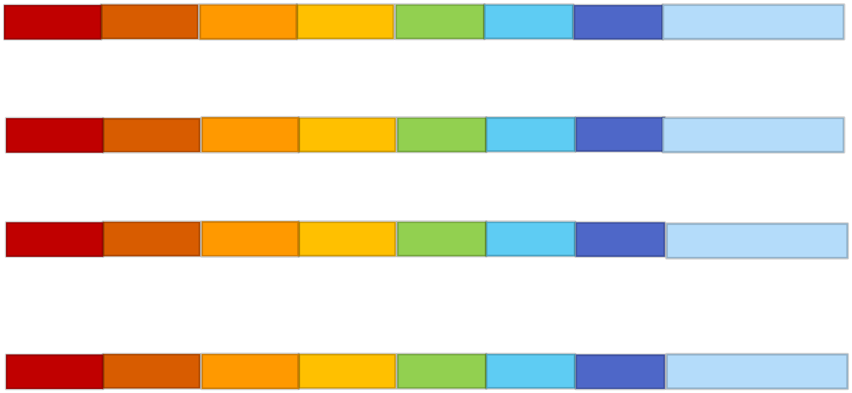
Peak List Alignment

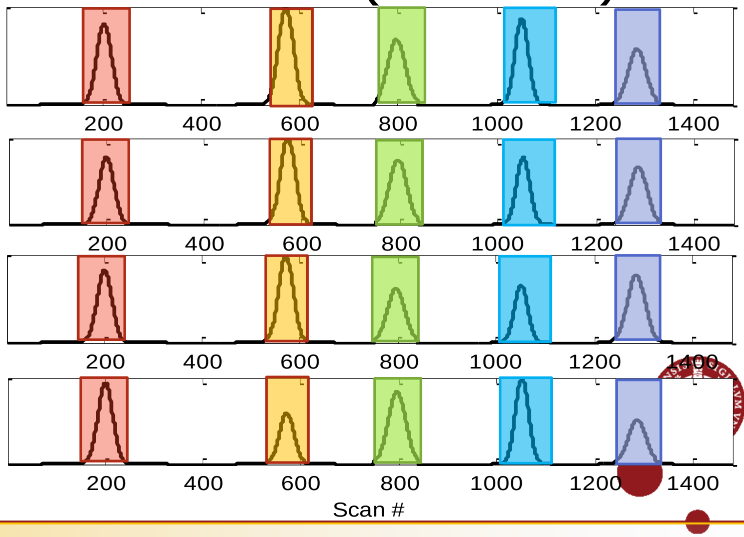
List of Integrated Peaks





Alignment – Ret. time tolerance and mass tolerance (MZmine)



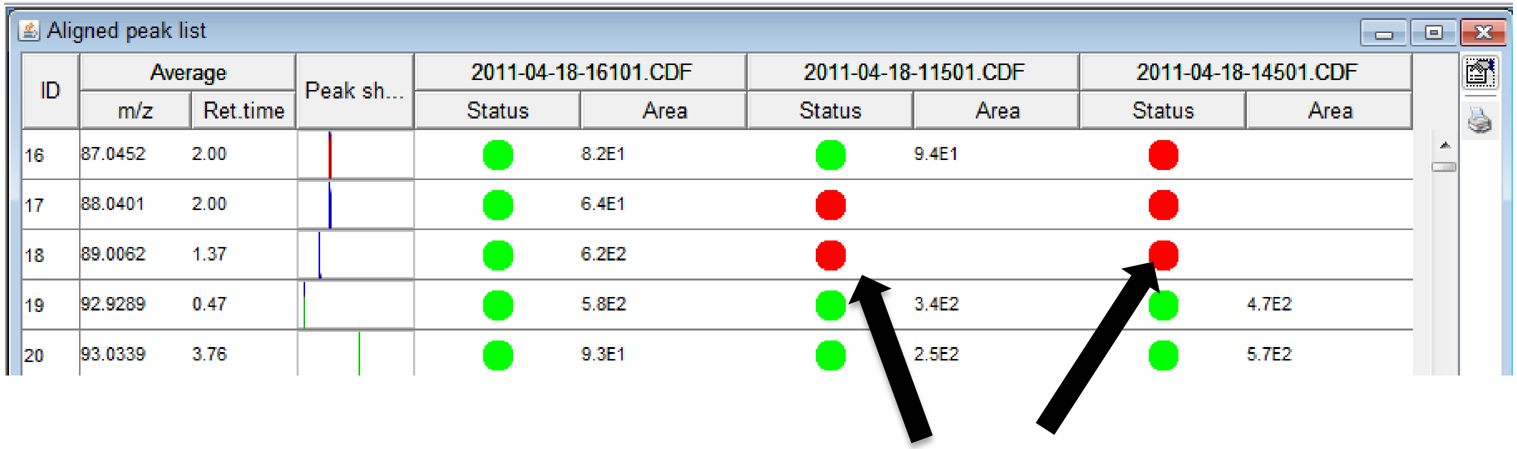


MZmine
Join Aligner

- Create a master peak list : concatenate all the features for all the samples
- Alignment window : **m/z and RT** bi-dimensional **window**.
- Score function : similarity of peaks between master peak list and each sample

Gap Filling

Gap filling refers to recovering the missing signals from raw data.



Aligned peak list

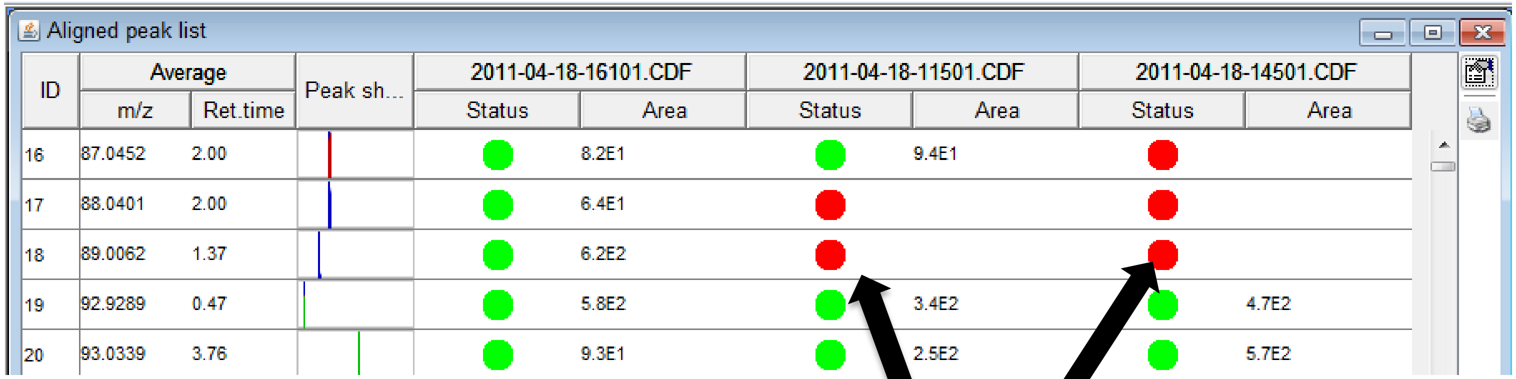
ID	Average		Peak sh...	2011-04-18-16101.CDF		2011-04-18-11501.CDF		2011-04-18-14501.CDF	
	m/z	Ret.time		Status	Area	Status	Area	Status	Area
16	87.0452	2.00		🟢	8.2E1	🟢	9.4E1	🔴	
17	88.0401	2.00		🟢	6.4E1	🔴		🔴	
18	89.0062	1.37		🟢	6.2E2	🔴		🔴	
19	92.9289	0.47		🟢	5.8E2	🟢	3.4E2	🟢	4.7E2
20	93.0339	3.76		🟢	9.3E1	🟢	2.5E2	🟢	5.7E2

Missing peaks

Missing peaks:

1. True zeros. They don’t appear in that sample.
2. False zeros. Low intensity, bad quality shape, or a mistake in peak detection.

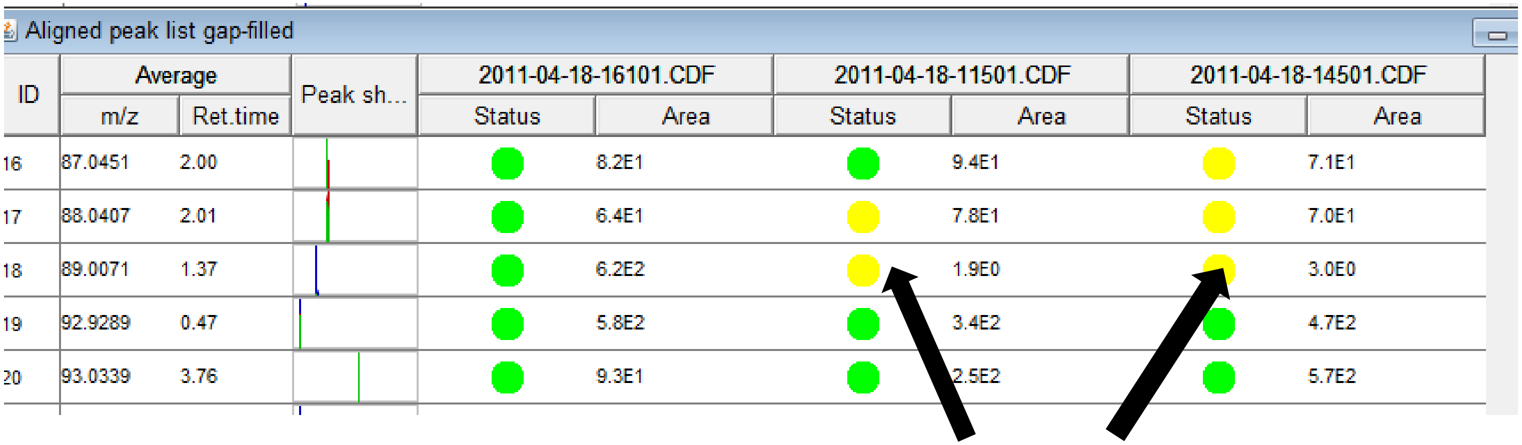
Gap Filling



Aligned peak list

ID	Average		Peak sh...	2011-04-18-16101.CDF		2011-04-18-11501.CDF		2011-04-18-14501.CDF	
	m/z	Ret.time		Status	Area	Status	Area	Status	Area
16	87.0452	2.00		🟢	8.2E1	🟢	9.4E1	🔴	
17	88.0401	2.00		🟢	6.4E1	🔴		🔴	
18	89.0062	1.37		🟢	6.2E2	🔴		🔴	
19	92.9289	0.47		🟢	5.8E2	🟢	3.4E2	🟢	4.7E2
20	93.0339	3.76		🟢	9.3E1	🟢	2.5E2	🟢	5.7E2

Missing peaks



Aligned peak list gap-filled

ID	Average		Peak sh...	2011-04-18-16101.CDF		2011-04-18-11501.CDF		2011-04-18-14501.CDF	
	m/z	Ret.time		Status	Area	Status	Area	Status	Area
16	87.0451	2.00		🟢	8.2E1	🟢	9.4E1	🟡	7.1E1
17	88.0407	2.01		🟢	6.4E1	🟡	7.8E1	🟡	7.0E1
18	89.0071	1.37		🟢	6.2E2	🟡	1.9E0	🟡	3.0E0
19	92.9289	0.47		🟢	5.8E2	🟢	3.4E2	🟢	4.7E2
20	93.0339	3.76		🟢	9.3E1	🟢	2.5E2	🟢	5.7E2

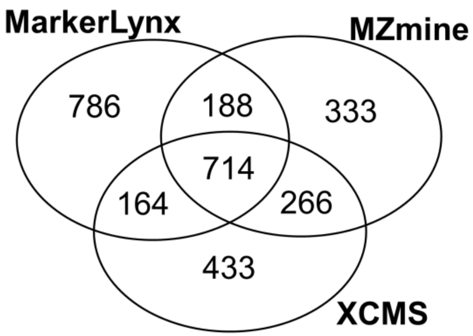
Gap-filled peaks

Gap Filling (MZmine)

- ✓ This algorithm fills the gaps in the peak list from raw data according with the parameters defined by the user.
- ✓ The most crucial parameters are **m/z tolerance** and **RT tolerance** which define the window where the algorithm should find the new peak.

Comparison of number of total features

	MarkerLynx	MZmine	XCMS
Number of features	1852	1501	1562

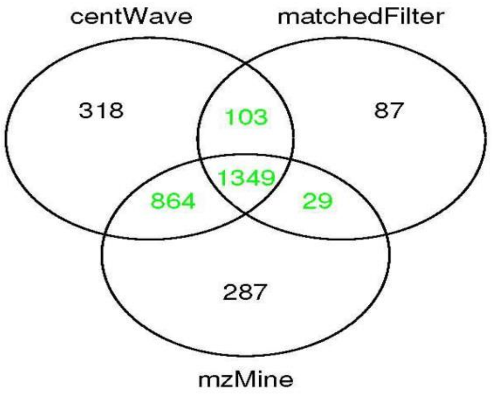


Common features → 25%

Seed extracts analyzed by HPLC-QTOF

Tautenhahn et al. (2008)

	XCMS (centWave)	XCMS (matchedFilter)	MZmine
Number of features	2634	1568	2529



Common features → 44%

Tautenhahn R, Bottcher C, Neumann S (2008) BMC Bioinformatics 9: 504.

Practical properties of MZmine, XCMS and MarkerLynx

	MZmine	XCMS	MarkerLynx
Availability	Free	Free	Commercial
User interface	• GUI* • No requirement of programming skills	• R software command line • Some programming skills is required	• GUI* • No requirement of programming skills
Memory usage	• Adjustable to maximum available memory in the PC. • Less efficient than XCMS e.g. 16 GB RAM = maximum ~2000 samples	• Adjustable to maximum available memory in the PC • e.g. 16 GB RAM = maximum ~5000 samples	• Fixed • e.g. maximum ~1000 samples
CPU usage	• Adjustable to maximum available CPU in the PC	• Adjustable to maximum available CPU in the PC	• Fixed
Identification	• Basic identification tools. • Automated advanced tool CAMERA is incorporated from XCMS	• Automated advanced identification tool CAMERA	• Basic identification tools
Coverage of preprocessing pipeline	• All steps	• Final feature table includes isotopic peaks	• Gap filling is missing
Visualization of the results	Yes	Yes	No

Conclusions – Comparison of data preprocessing methods

- ✓ Considering the large number of peaks with varying peak shapes, so far there is no common method to evaluate the preprocessing algorithms from different software.

- ✓ Parameter settings: Evaluate based on your instrument
  Try several settings

- ✓ None of the software tools was able to extract all metabolites.

- ✓ Use more than one software tool.

Thank you for your attention!!