COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

12. Quantitative Metabolomics

LEARNING UNIT 12A
NON-TARGETED QUANTIFICATION

- Differences and similarities to label-free proteomics
- Feature finding for metabolites
- XCMS
- OpenMS feature-finding algorithm

Metabolite Quantification

- **Label-free proteomics** is similar to **non-targeted metabolomics**
- Overall workflow is identical
  - Feature finding
  - Map alignment
  - Feature linking
- Feature-finding approaches are algorithmically **similar** to those used in proteomics
  - Mass traces usually at the heart of the algorithm
  - Assembly into features can be done similarly
- However, there are some **differences**
  - Isotopic patterns differ from proteomics (no averagine!)
  - Mass range and charge states are different

Feature Finding – Terms

Map:

Two-dimensional data set (RT, m/z) containing the MS signal from one LC-MS run.

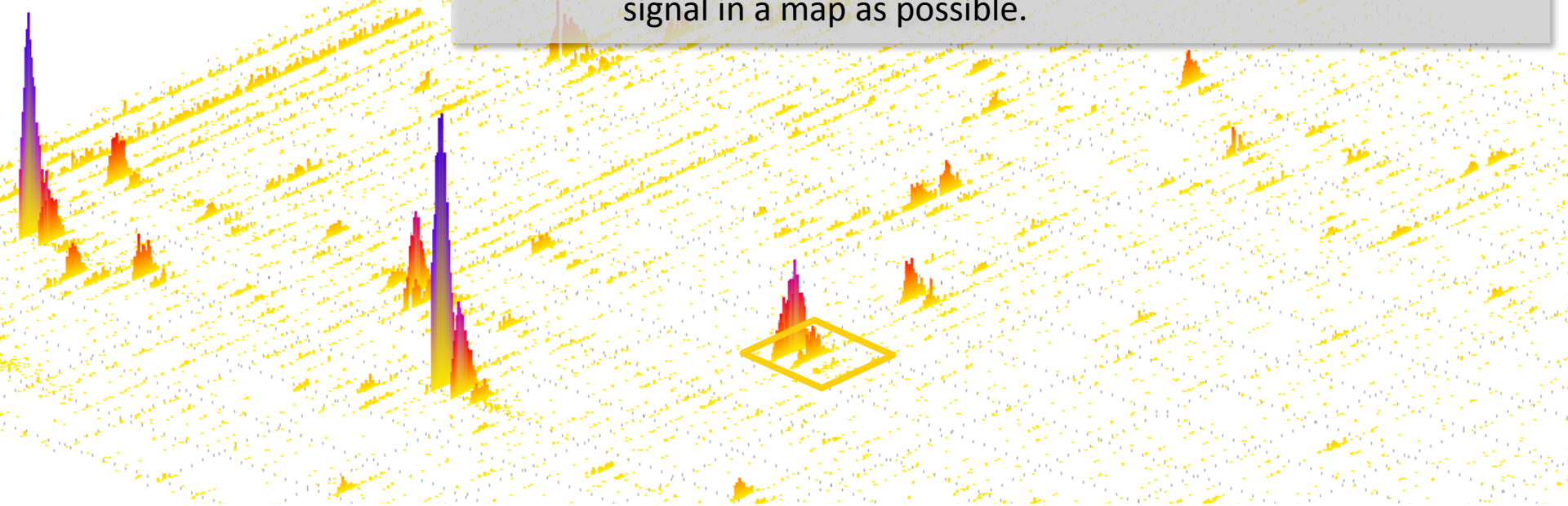
Feature:

The sum of all the MS signals caused by the same analyte in a specific charge state.

Different adducts will result in distinct features. Primarily characterized by RT, m/z, charge, intensity.

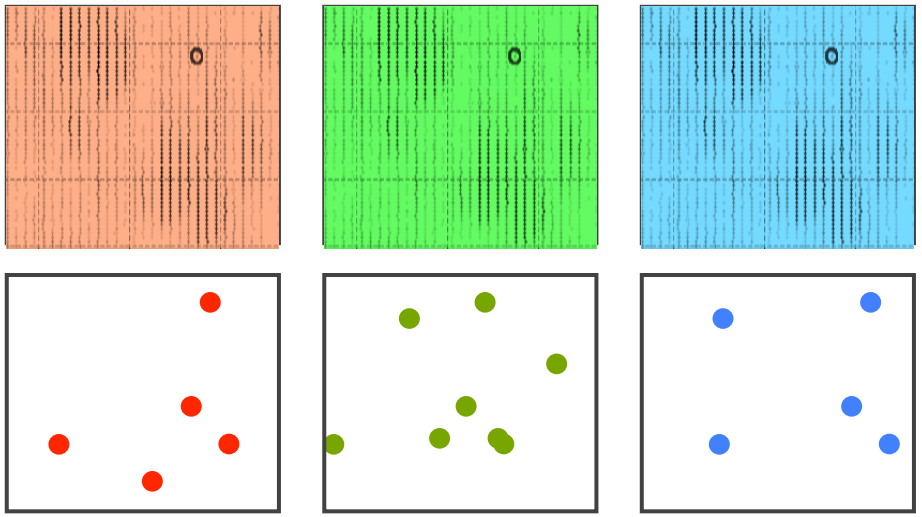
Feature finding:

Finding the set of features explaining as much of the signal in a map as possible.



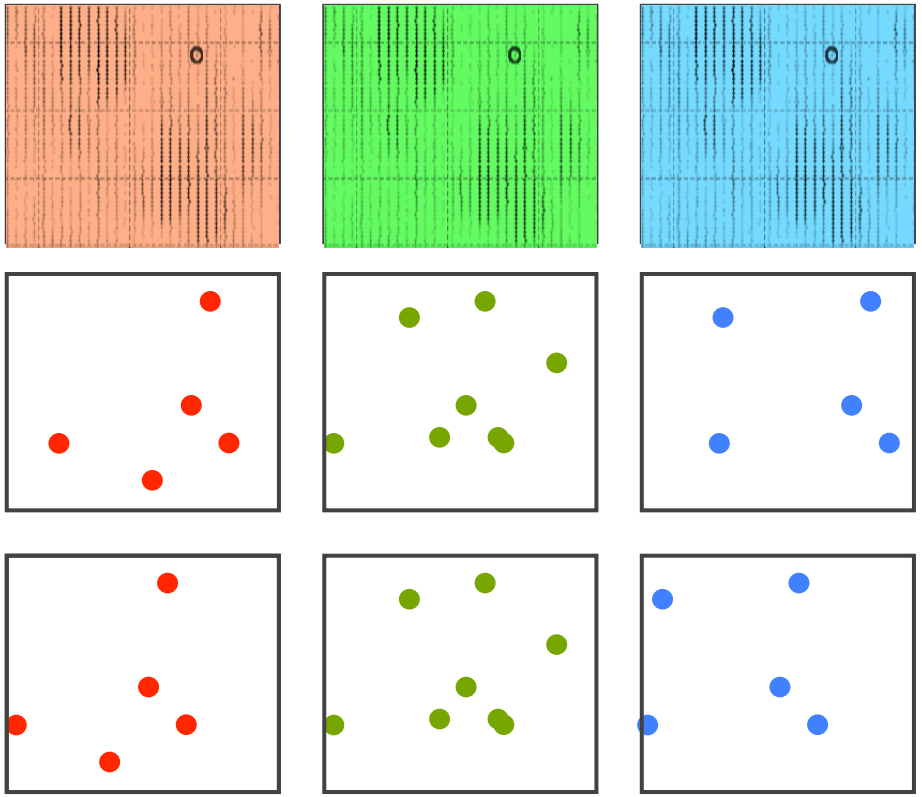
Metabolic Profiling

1. **Find** features in all maps



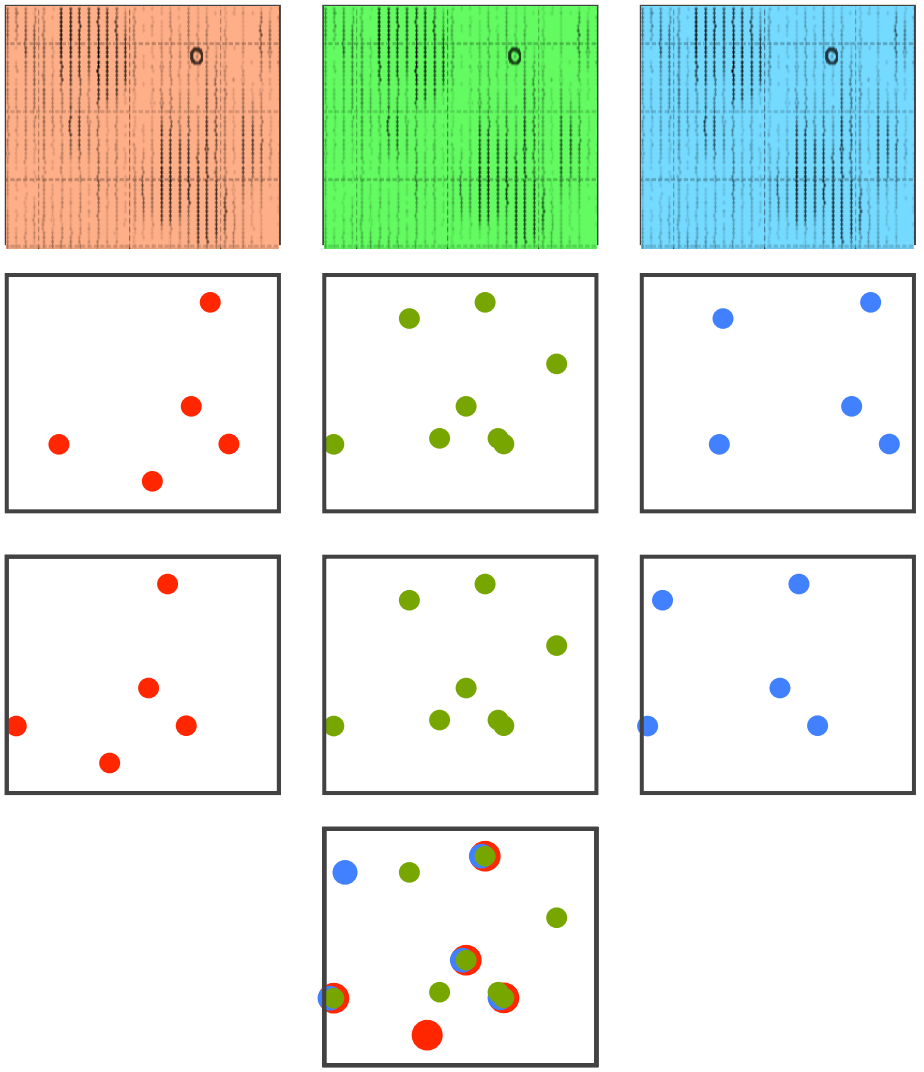
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps



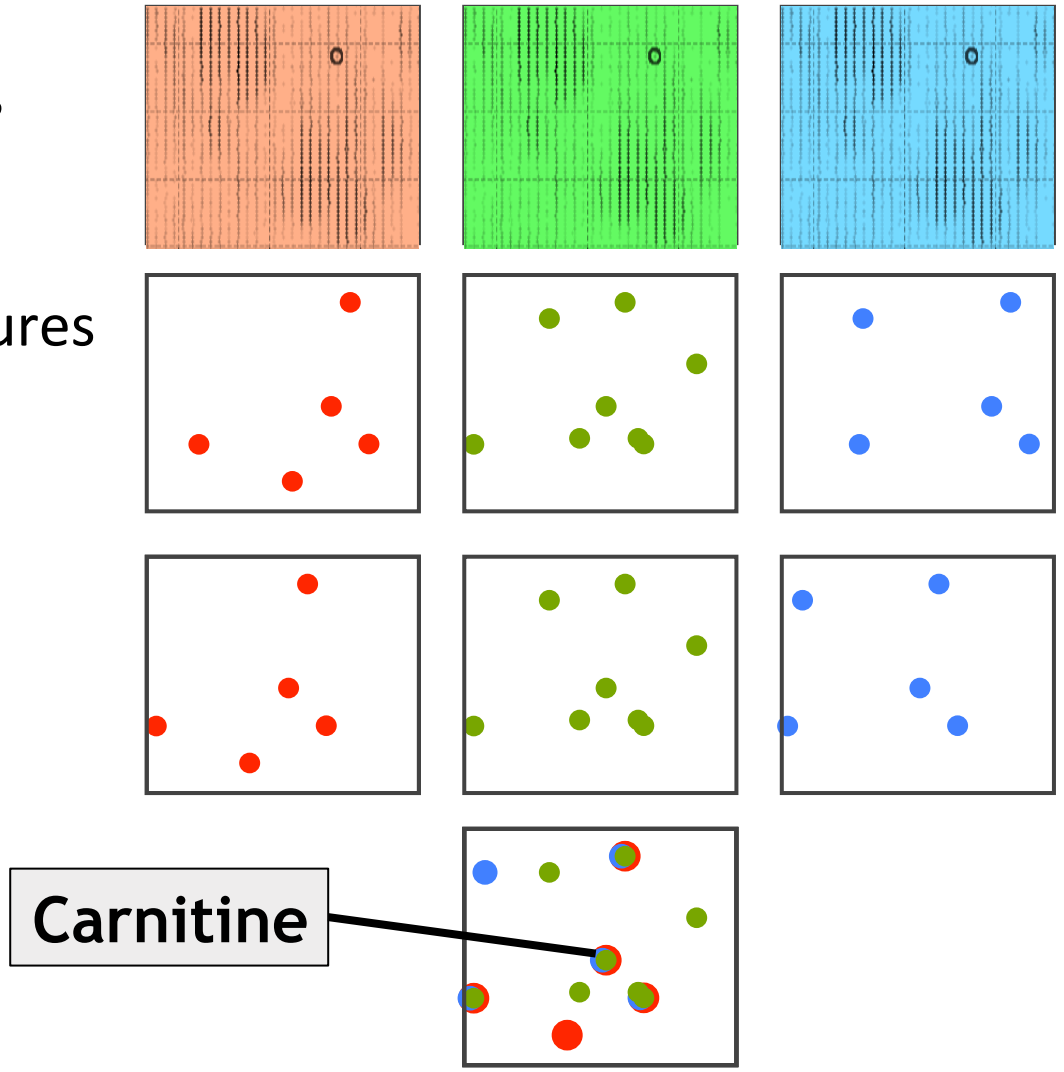
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features



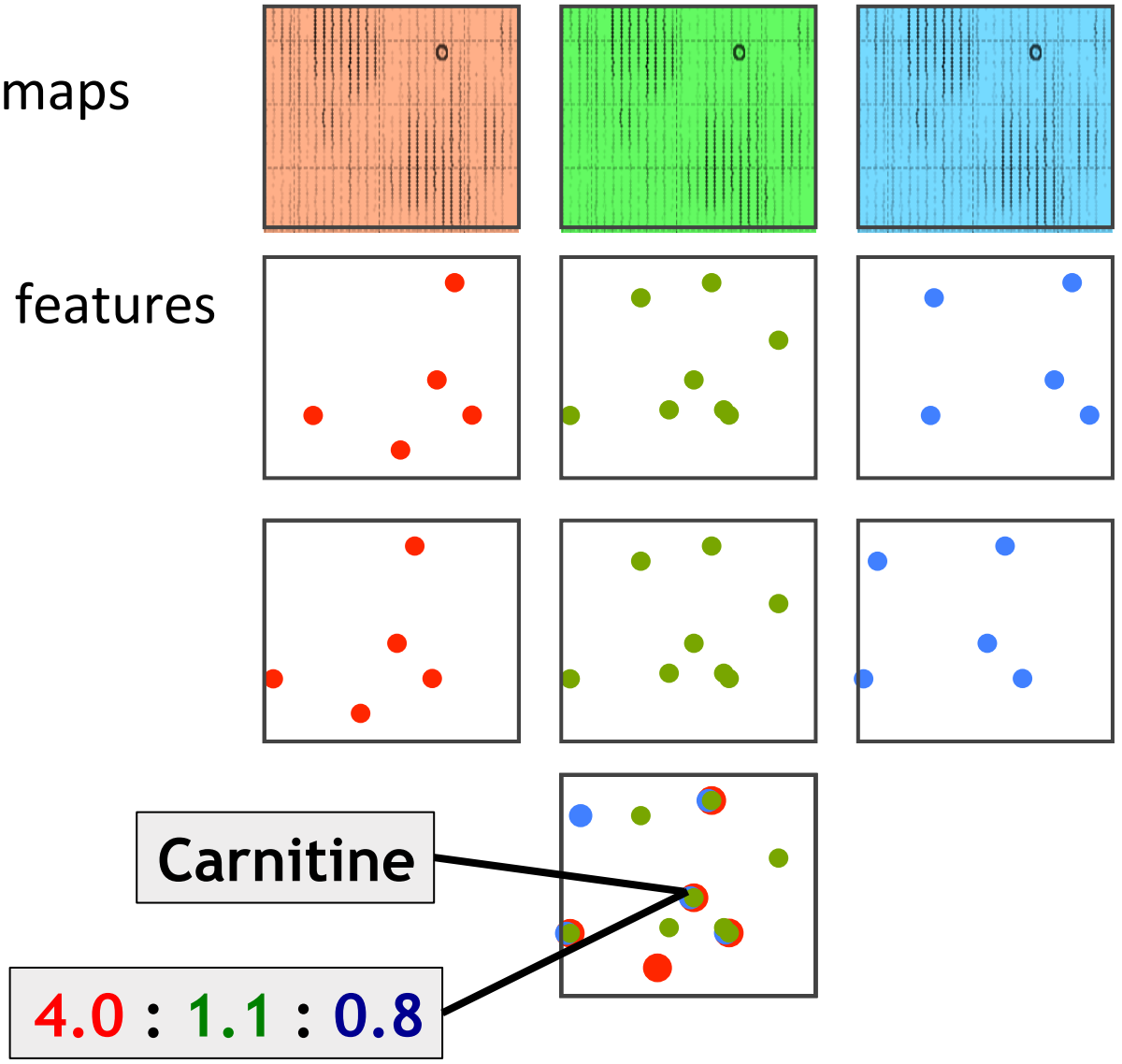
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features



Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**



Feature Finding in MTX – Issues

- Proteomics feature finding algorithms make extensive use of the **averagine** hypothesis: peptides have a well-defined average composition
- Metabolites are chemically much more diverse than peptides

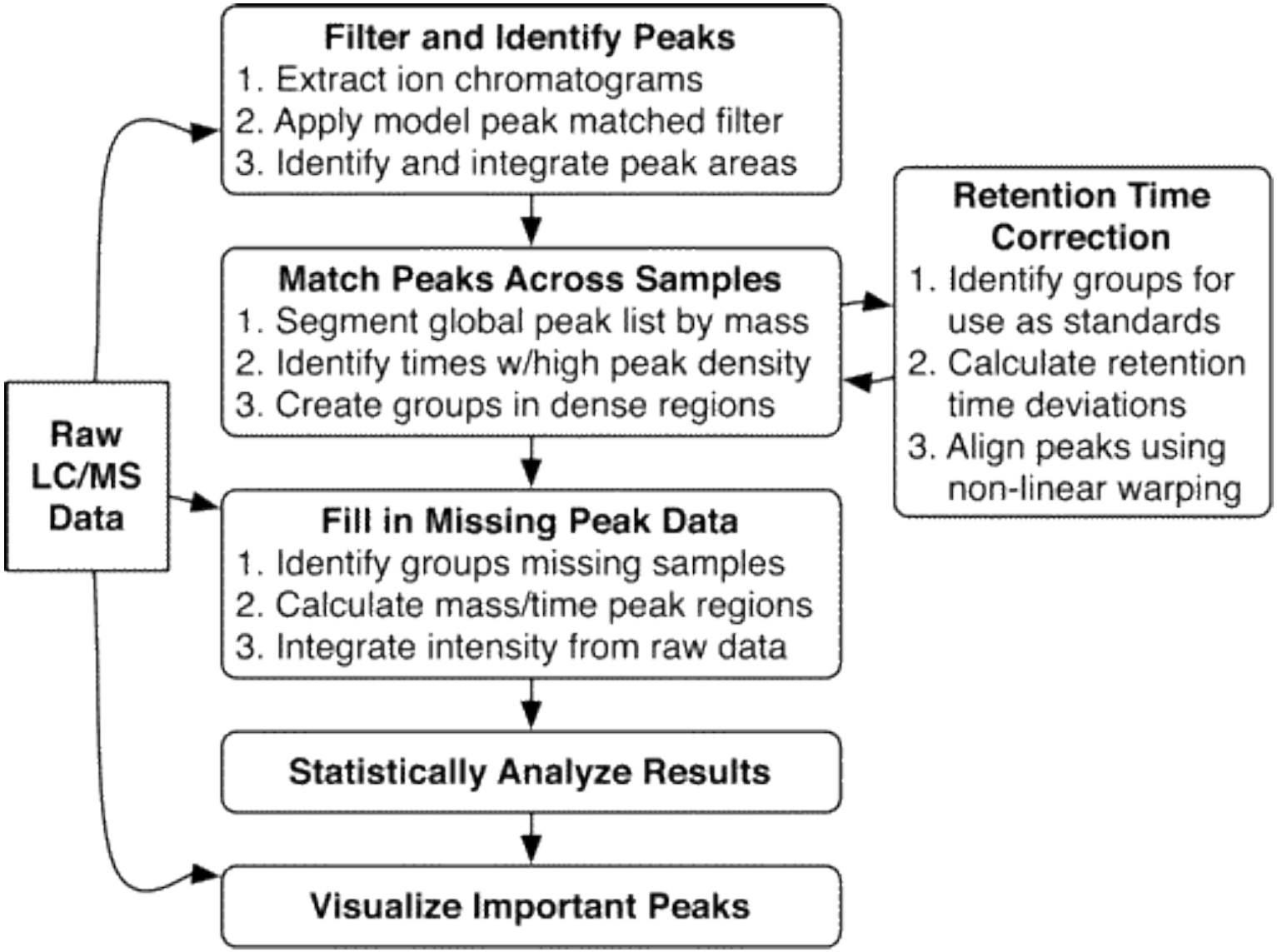
- Feature finding algorithms are often very sensitive to the choice of **parameters**
- Tuning these parameters can be a challenge

- **Sensitivity** is often an issue in feature finding: distinguishing signal from noise can be a challenge
- Lack of sensitivity is often a problem for large-scale studies – missing values

XCMS

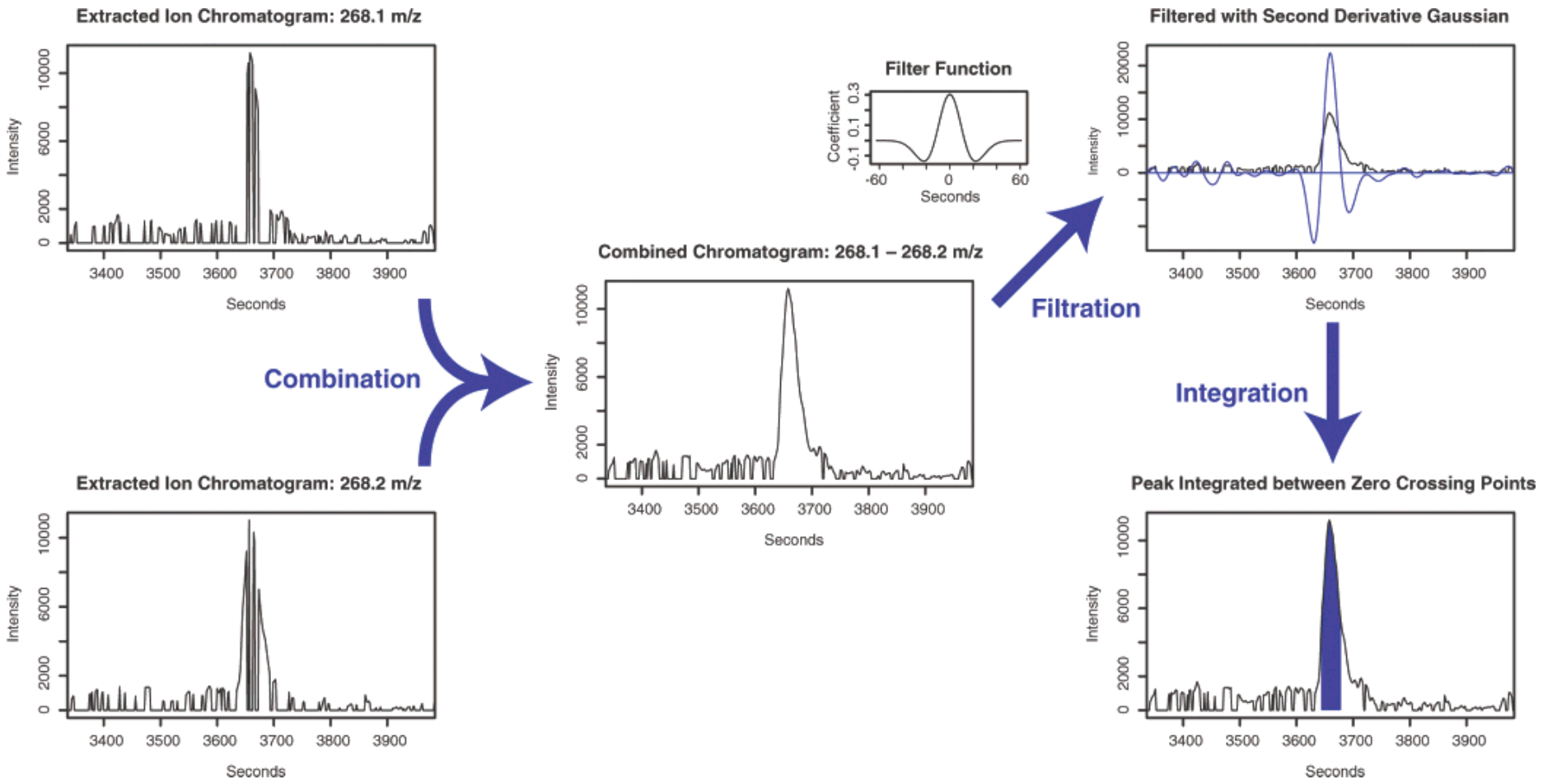
- XCMS is a Bioconductor package, written in R
- **Key ideas**
  - Extract mass traces by binning peaks w.r.t. m/z
  - Treat mass bins as distinct mass traces
  - Detect peaks in these mass traces using standard methods from signal processing
  - Align detected mass traces in the RT dimension across maps using nonlinear de-warping

XCMS



Smith et al., Anal. Chem., 2006, 78, 779-787

XCMS


Extracted Ion Chromatogram: 268.1 m/z
Intensity
Seconds
Extracted Ion Chromatogram: 268.2 m/z
Combination
Combined Chromatogram: 268.1 – 268.2 m/z
Filter Function
Coefficient
Filtration
Filtered with Second Derivative Gaussian
Integration
Peak Integrated between Zero Crossing Points

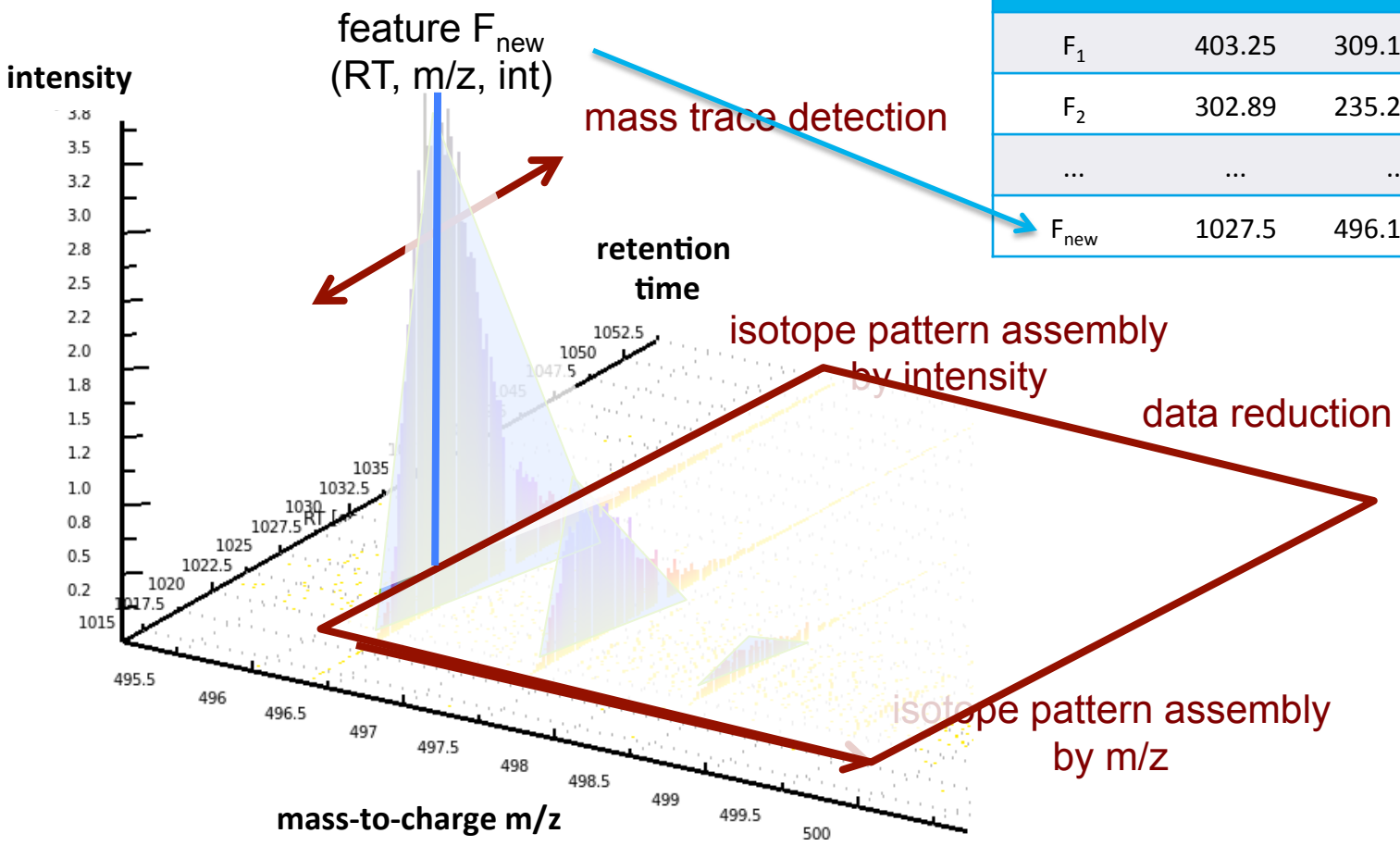
XCMS

- XCMS has become the quasi standard for LC-MS metabolomics data analysis
- Recent versions include more advanced methods, including wavelet peak detection
- For many tasks (e.g., biomarker detection), the identification of differential mass traces is sufficient (lower complexity of metabolomics data sets)
- Other software packages also assemble mass traces back to features (e.g., OpenMS FeatureFinderMetabo)
- Advantages here:
  - Profit from additional information, increase specificity
  - Reduced number of signals (multiple mass traces per feature)

OpenMS - Metabolite Feature Finding

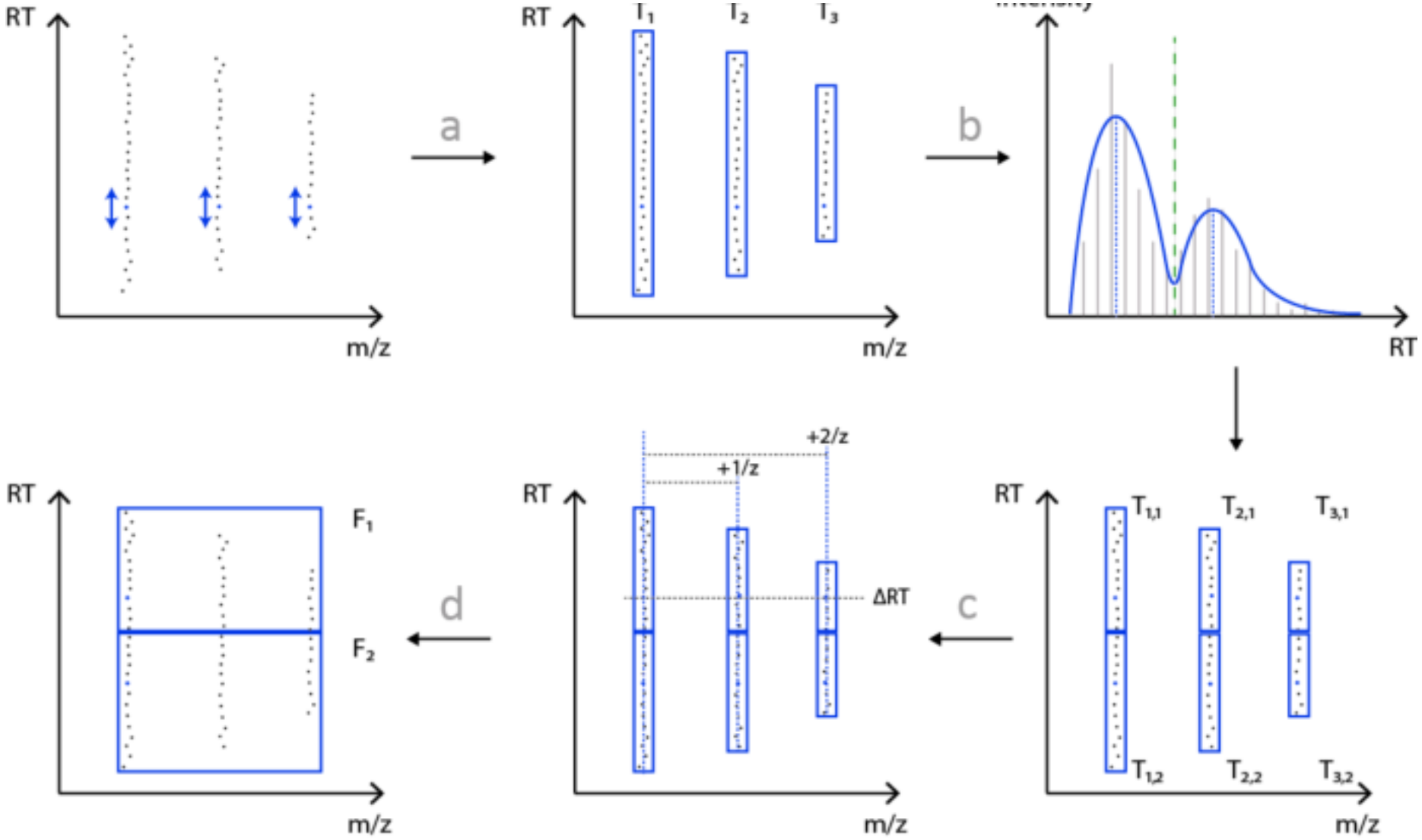
MS data condensed to feature list:

Feature ID	RT	m/z	intensity
F_1	403.25	309.13455	345923.1
F_2	302.89	235.20503	8109.5
...	...	...	...
F_{new}	1027.5	496.11304	45209.8



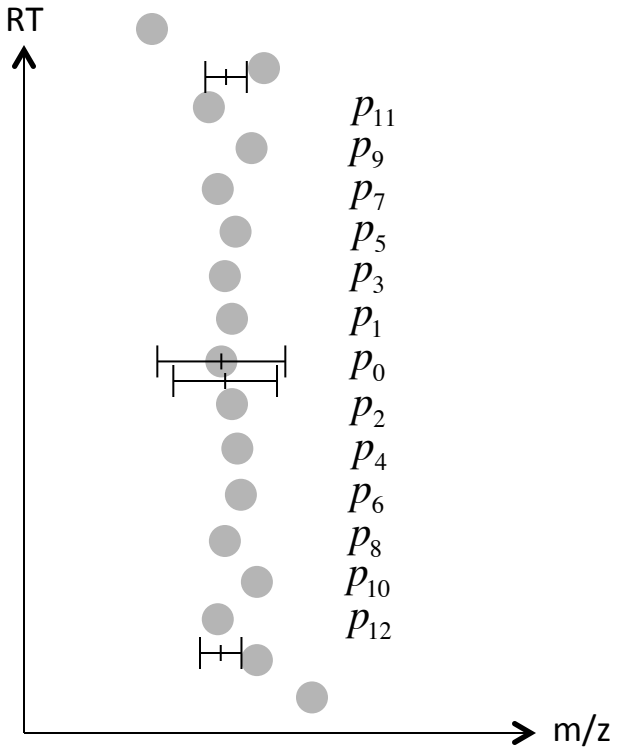
Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Algorithmic Overview



Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Mass Trace Detection



$\mu_{02} = 521.42388$

$\sigma^2_{02} \propto$ mass_error_param

$T = (p_{12}, p_{10}, p_8, p_6, p_4, p_2, p_0, p_1, p_3, \ldots, p_{11})$

- A mass spectrometric peak *p* is given by

$$p = (t, m, i)$$

t : retention time, m : mass-to-charge ratio, i : intensity

- A mass trace *T* is a list of peaks:

$$T = (p_1, p_2, \ldots, p_k, p_l, \ldots, p_n) \qquad t_k < t_l \, \forall k < l$$

- m/z error model is adaptive
- Online Gaussian density estimation

$$\mu_{n+1} = \frac{w_n \cdot \mu_n + i_{n+1} \cdot m_{n+1}}{w_n + i_{n+1}} \qquad \sigma^2_{n+1} = \frac{w_n \cdot \sigma^2_n + i_{n+1} \cdot (m_{n+1} - \mu_{n+1})^2}{w_n + i_{n+1}}$$

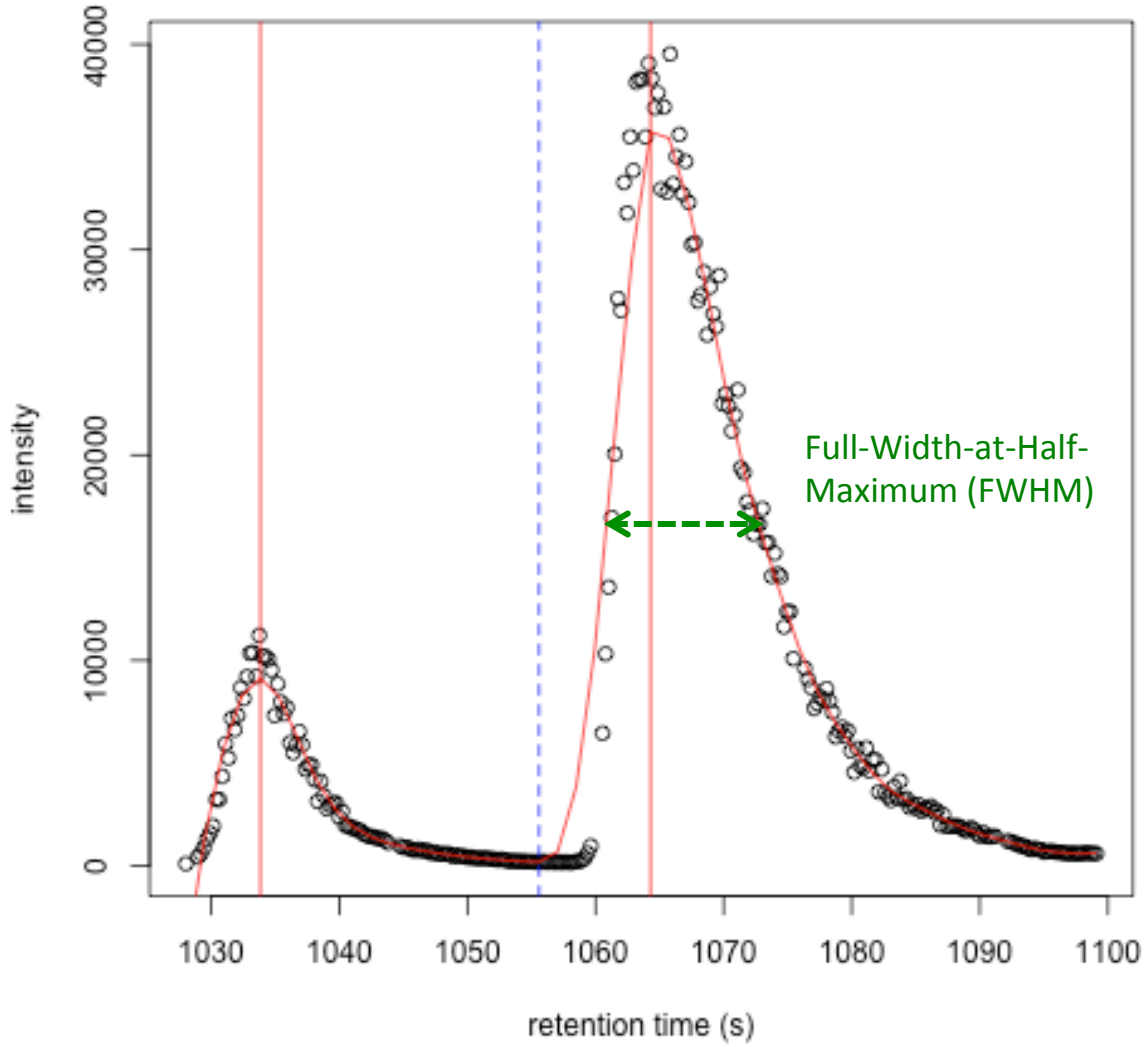
centroid m/z | m/z error

$$w_n = \sum_k^n i_k \qquad \mu_n - 3 \cdot \sigma_n \le m_{n+1} \le \mu_n + 3 \cdot \sigma_n$$

weight | m/z constraint

Peak Separation

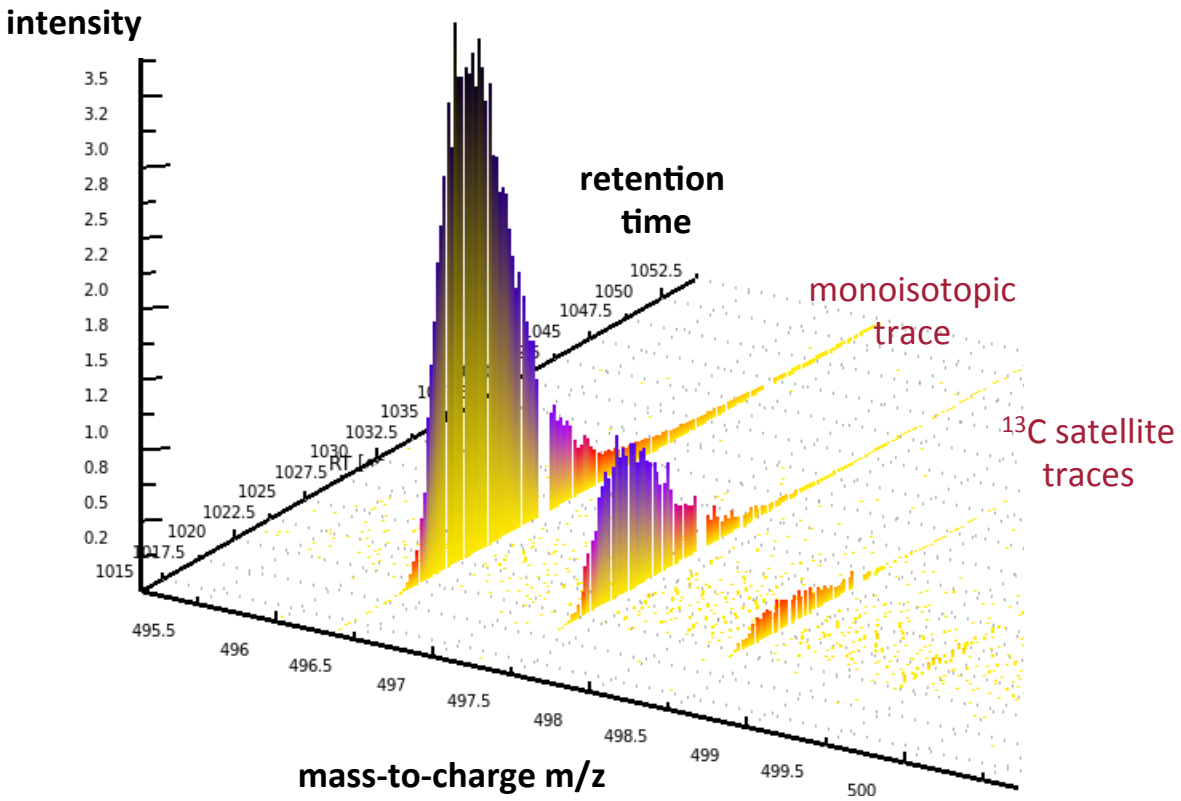
- Split chromatographic peaks overlapping in retention time



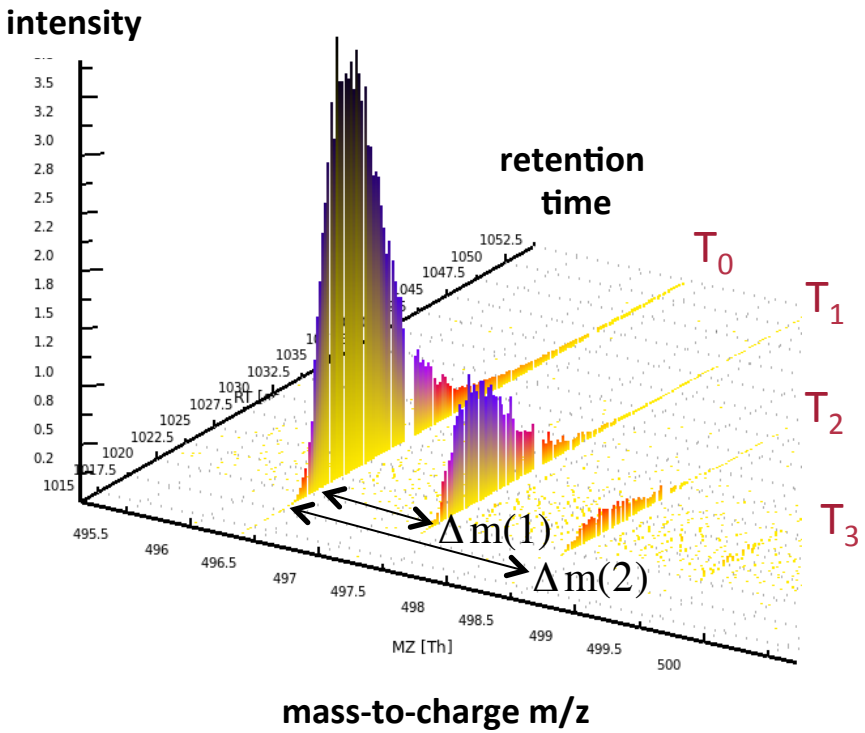
Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Feature Assembly

- Identify mass traces belonging to the same feature
- Multiple explanations are possible
- Create all potential hypotheses and score them



Feature Scoring – m/z



- m/z distances T_0 and T_j:

$$\Delta \mathrm{m(j)} = \left| \bar{m}_0 - \bar{m}_j \right|$$

- Theoretical m/z distances:

$$\mu(\mathrm{j}) = 1.0033\ \mathrm{Da} \cdot \frac{j}{z}$$

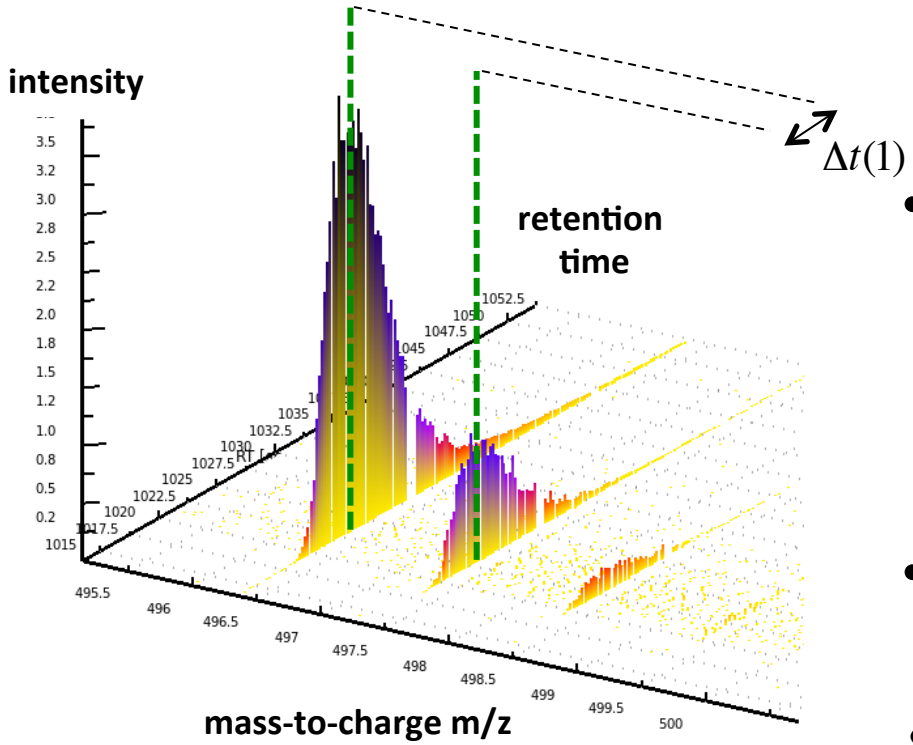
- Mass errors for T_0 and T_j:

$$\sigma^2(\mathrm{j}) = \sigma_0^2 + \sigma_j^2$$

- Pairwise scoring function:

$$S_{\Delta m}(\mathrm{j}) = \begin{cases} e^{-\frac{(\Delta \mathrm{m(j)} - \mu(\mathrm{j}))^2}{2\sigma^2(\mathrm{j})}}, & \text{if } \mu(\mathrm{j}) - 3 \cdot \sigma(\mathrm{j}) \le \Delta m(\mathrm{j}) \le \mu(\mathrm{j}) + 3 \cdot \sigma(\mathrm{j}) \\ 0 & \text{else.} \end{cases}$$

Feature Scoring – RT



- RT shifts between T_0 and T_j:

$$\Delta \mathrm{t(j)} = \left| \overline{t}_0 - \overline{t}_j \right|$$

- Gaussian error model with

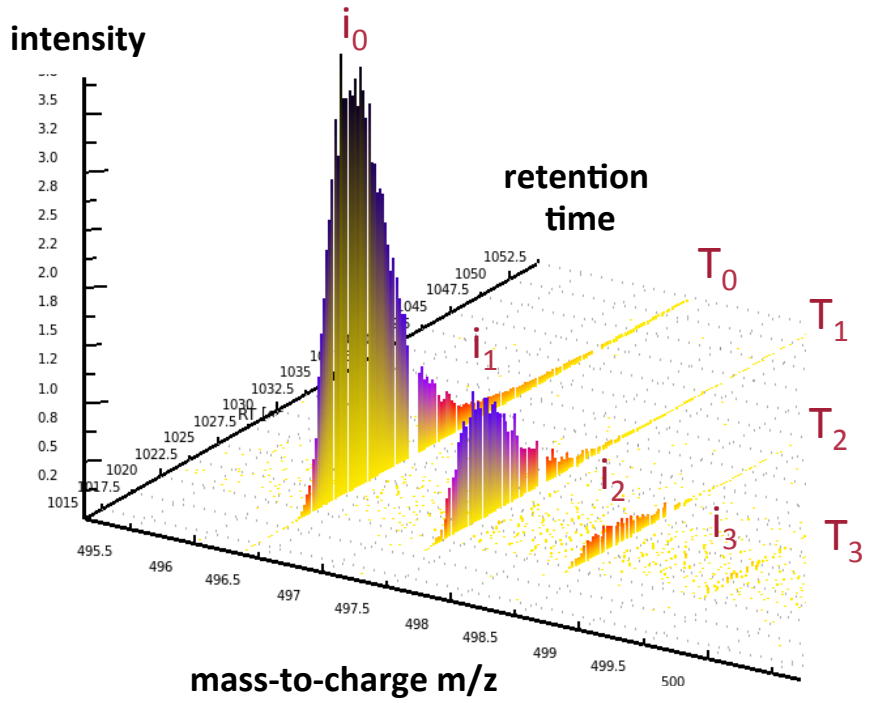
$$\mu_{\Delta RT} = 0 \qquad \sigma^2_{\Delta RT} = \left(\frac{\Delta t_{0.5}}{2\sqrt{2 \ln 2}} \right)^2$$

- Pairwise scoring function:

$$S_{\Delta RT}(\mathrm{j}) = \begin{cases} e^{-\frac{(\Delta \mathrm{t(j)})^2}{2\sigma^2_{\Delta t}}}, & \text{if } -3 \cdot \sigma_{\Delta t} \le \Delta t(\mathrm{j}) \le 3 \cdot \sigma_{\Delta t} \\ 0 & \text{else.} \end{cases}$$

Feature Scoring – Intensity

- **Problem:** There is no ‘averagine’ for metabolites
- **Idea**
  - Enumerate metabolite compositions and learn intensities
  - ‘Golden rules’ describe likely chemistry *(Kind & Fiehn, BMC Bioinfo, 2007)*
  - Generate all compositions, remove unlikely ones based on heuristics



24 mio. compositions

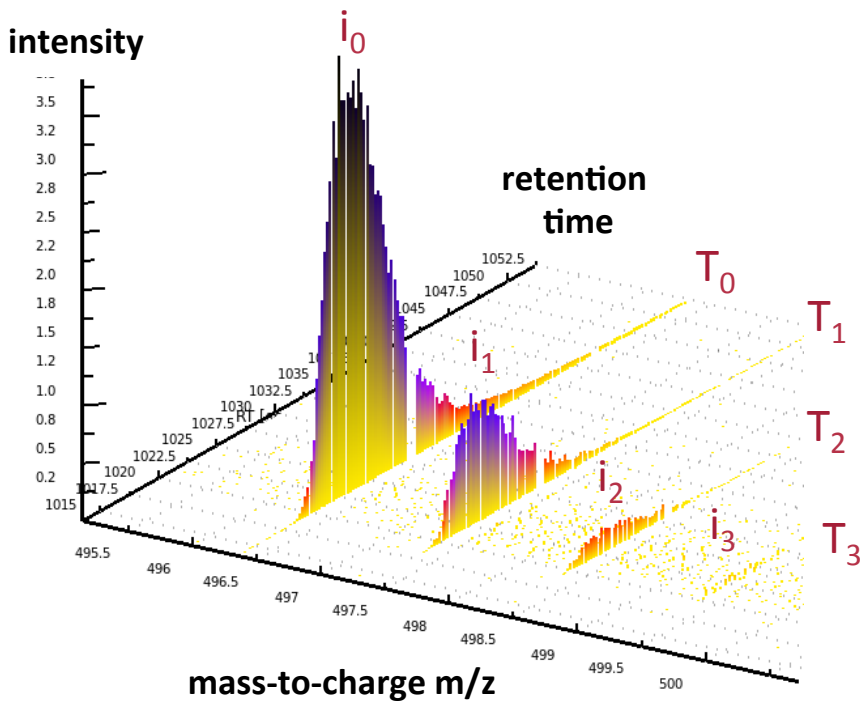
↓

Random subsample: 115 k compositions

↓

SVM (RBF kernel)

Feature Scoring – Intensity



- Intensity ratio of T_0 and T_j:

$$r(j) = \frac{i_j}{i_0}$$

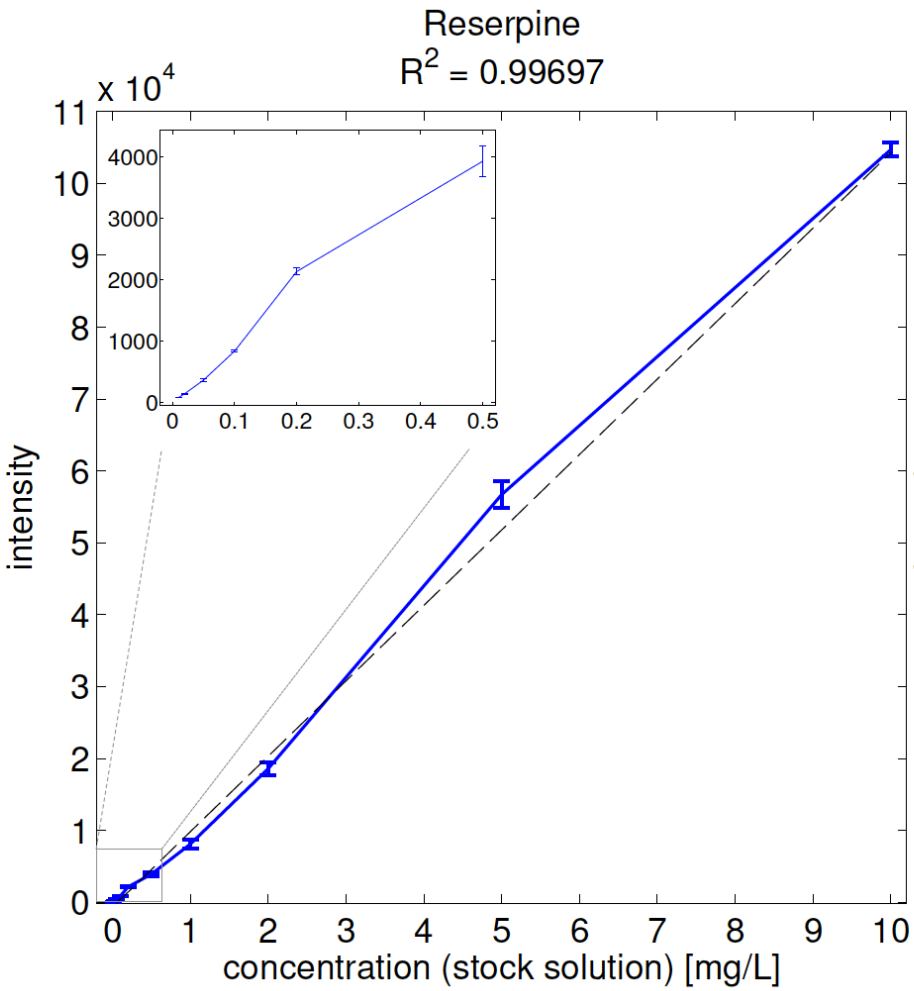
- Assess if valid isotope ratios:

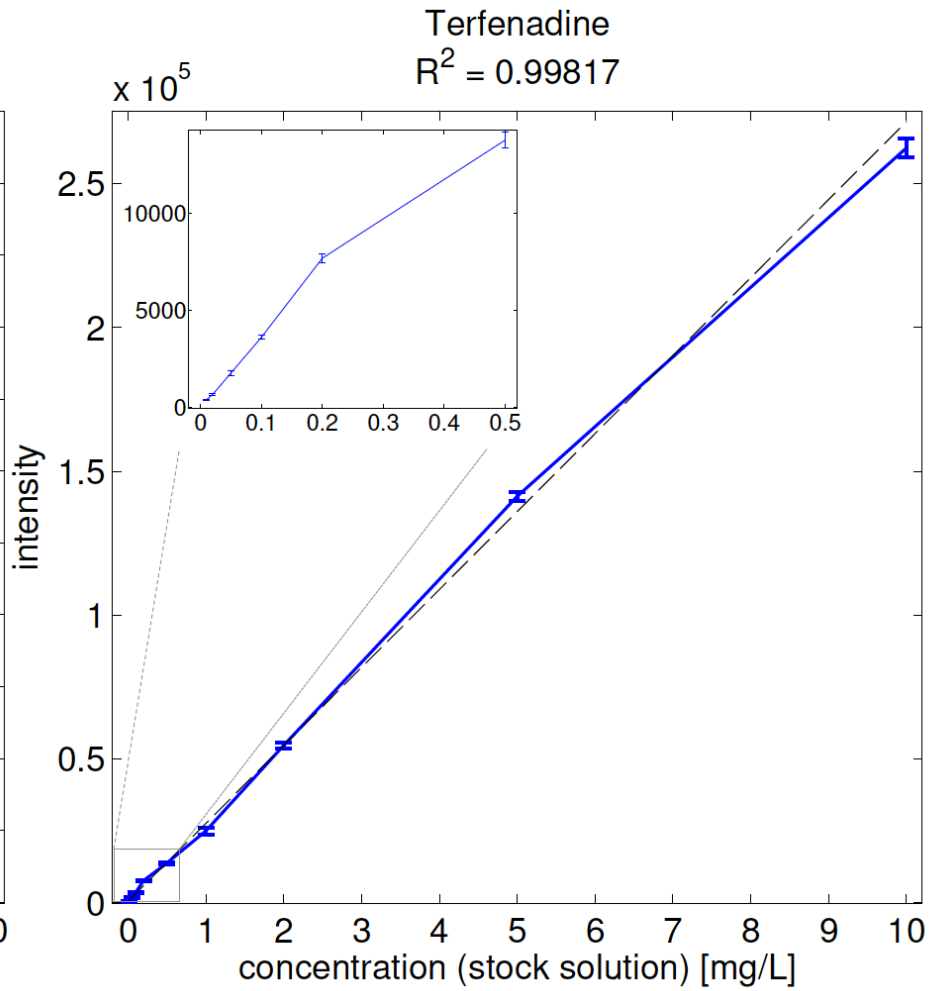
$$m(T_0), r(0), r(1), \ldots, r(5)$$

SVM
RBF kernel

Yes, it is a legal isotope pattern, ***keep it***
Or
No, it is not a legal isotope pattern,
discard it

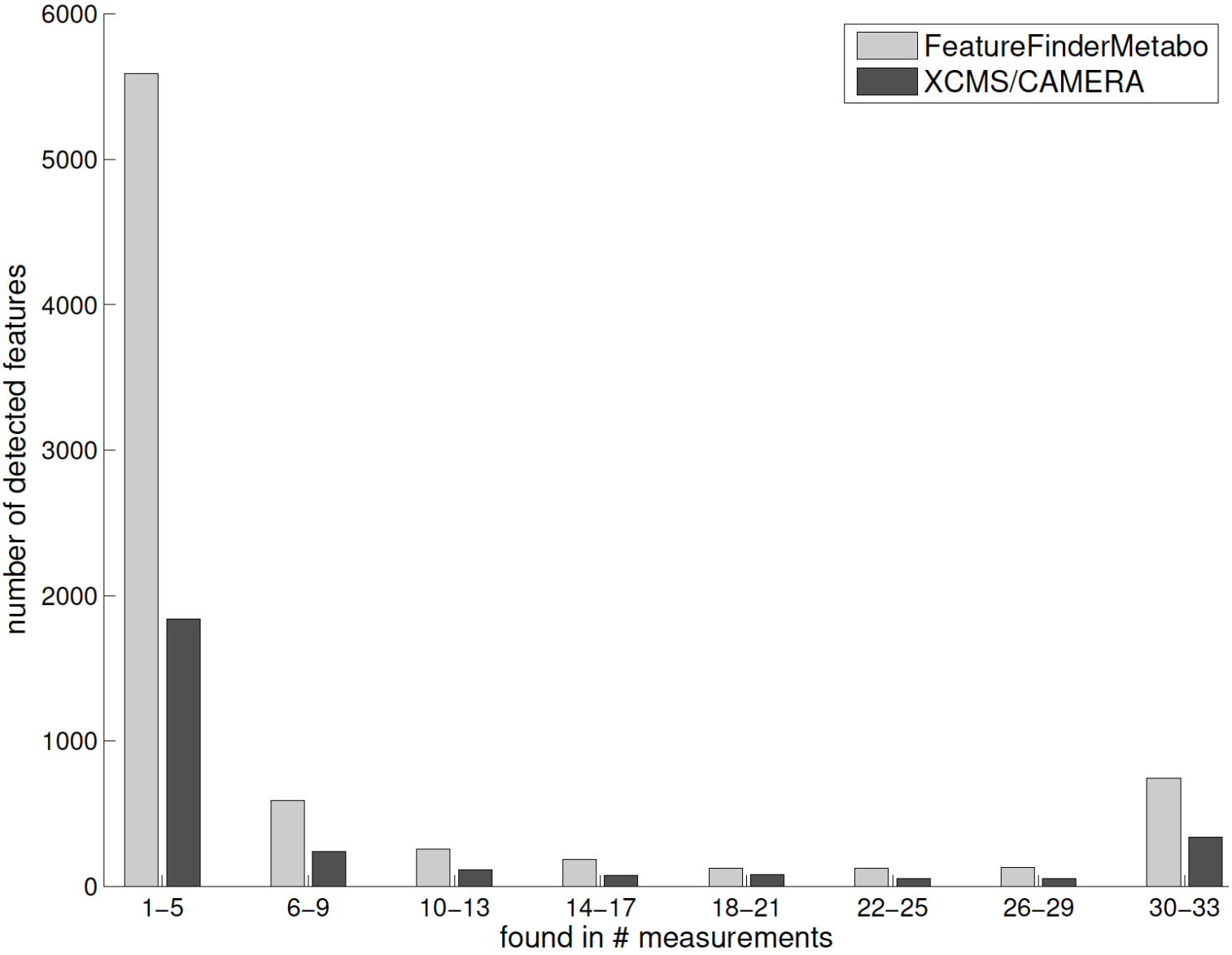
Quantification Linearity – Spike-In





Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Sensitivity – Human Plasma



Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Specificity – Synthetic Data

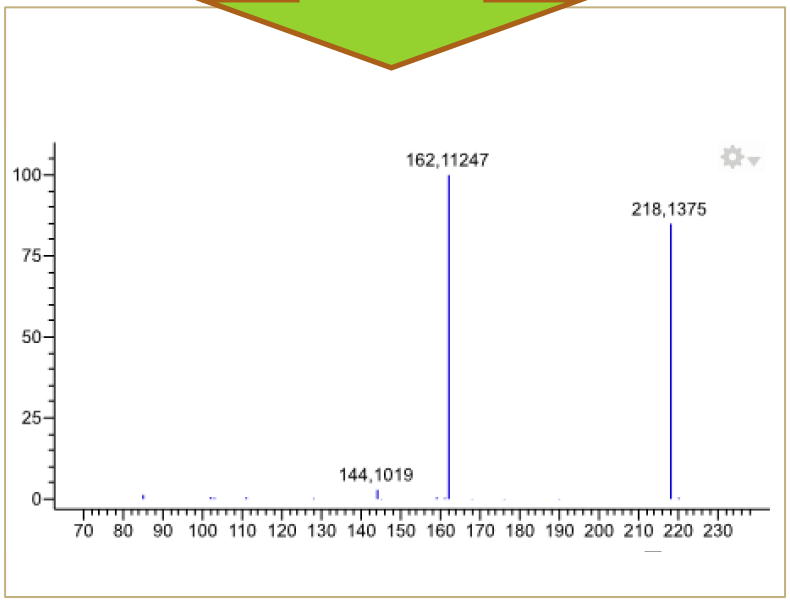
- **Benchmarking feature detection algorithms is HARD**
  - Multiple metrics are required: linearity, sensitivity, specificity
  - Sensitivity needs to be balanced with specificity
  - Experimental data does not come with a well-defined ground truth
- **Idea**
  - Simulated LC-MS data with known composition
  - Take a well-defined experimental dataset (identification lists from a metabolomics study, plant metabolites)
  - OpenMS LC-MS simulator was expanded to generate metabolite data

Method	Recall	Precision	F-score
OpenMS	96%	97%	0.97
XCMS/Camera	88%	37%	0.52

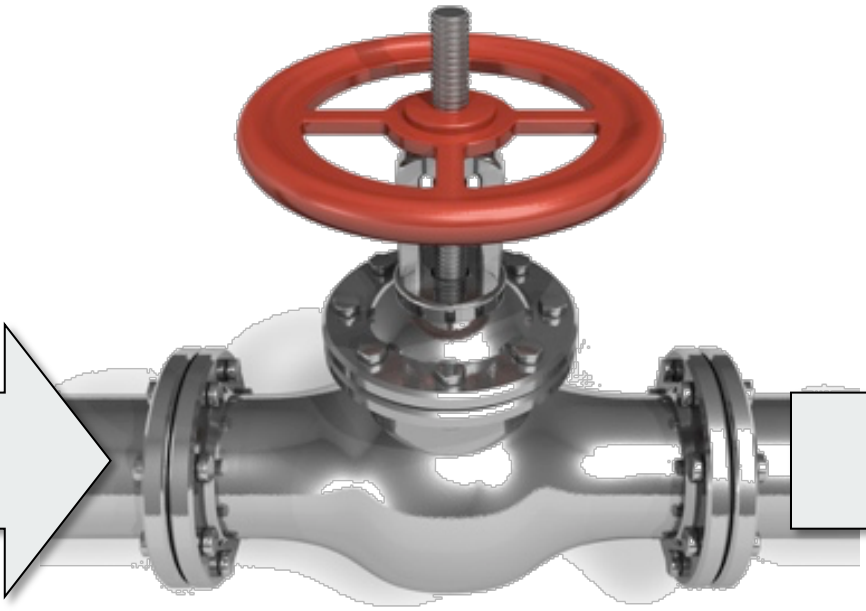
LU 12B
NON-TARGETD METABOLOMICS WITH OPENMS

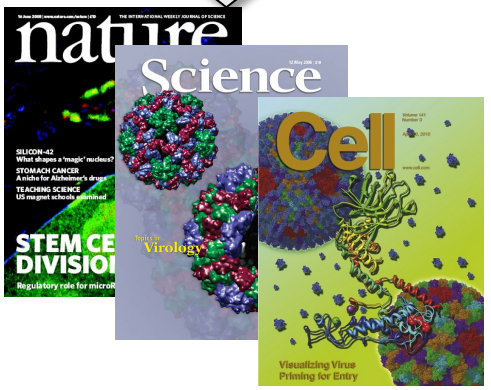
- Workflows for non-targeted metabolomics
- Metabolomics workflows with OpenMS in KNIME
- Integration into Compound Discoverer

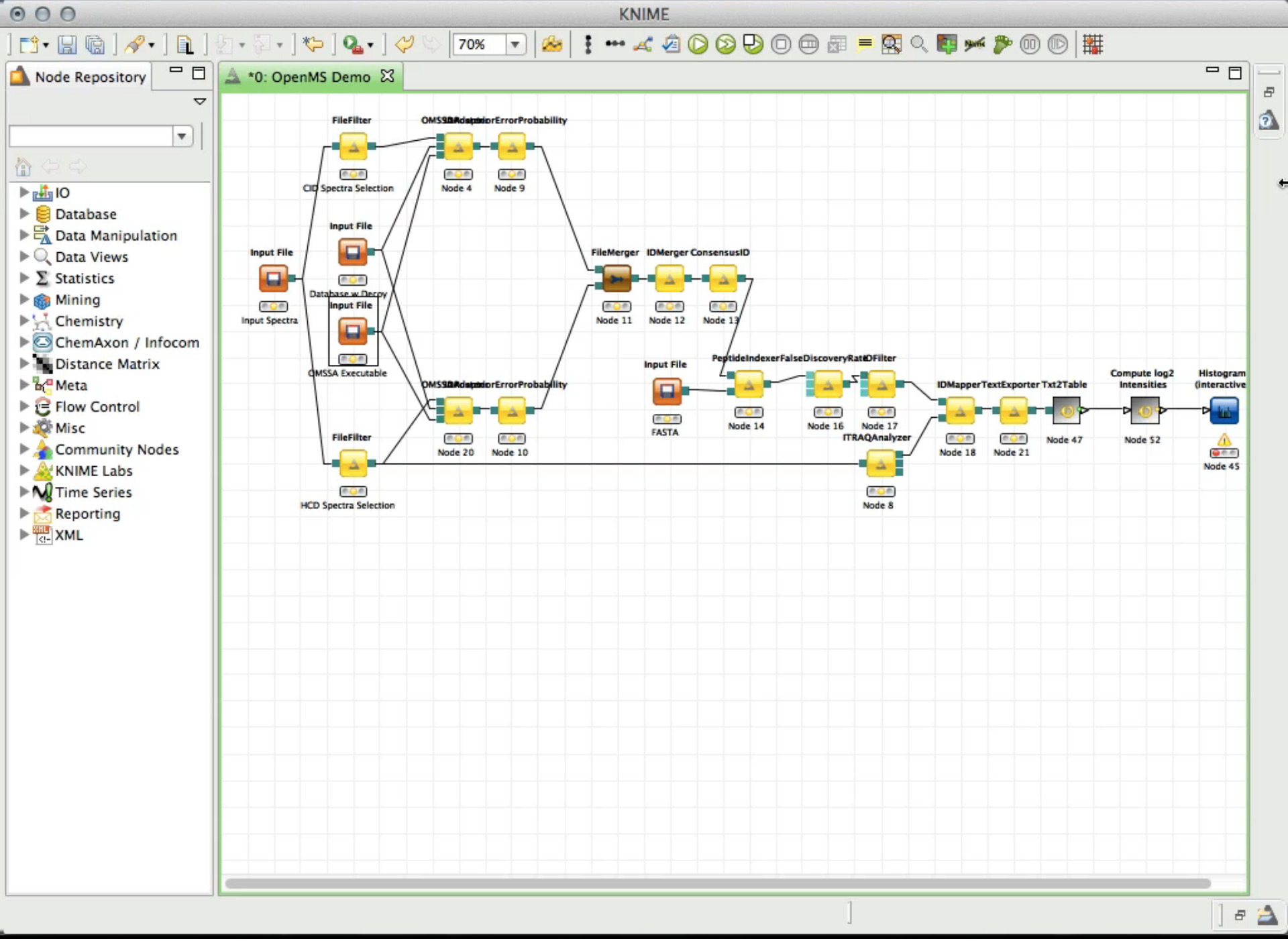
High Throughput vs. Low Throughput



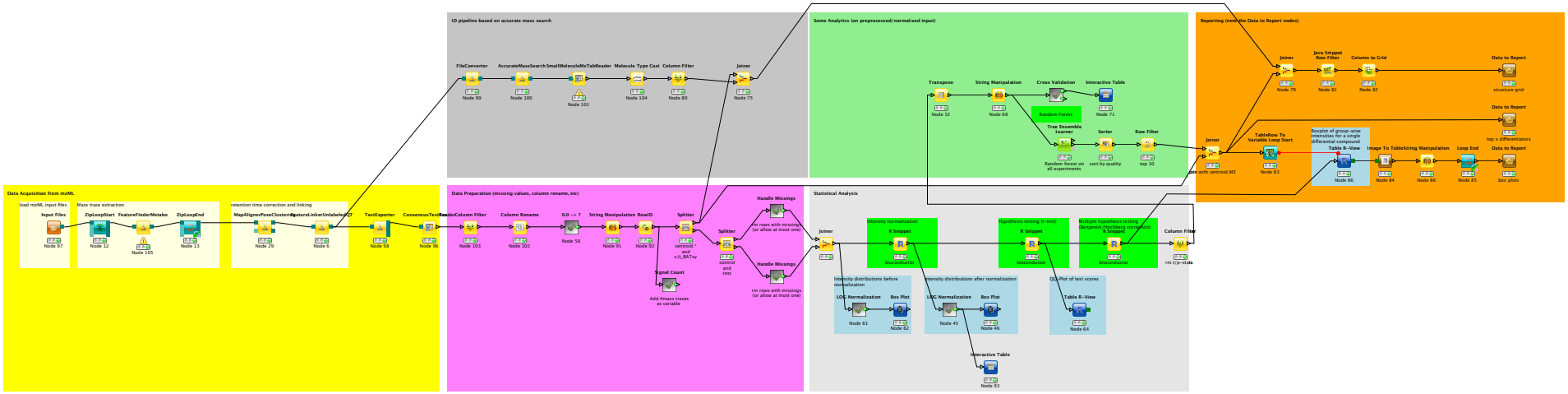
The Holy Grail





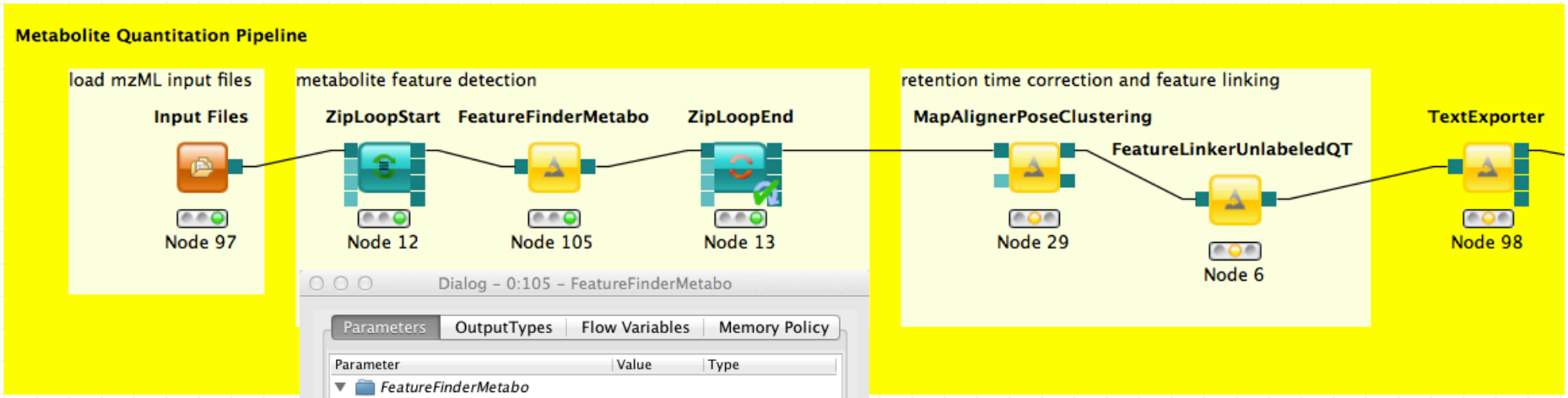

KNIME
Node Repository
IO
Database
Data Manipulation
Data Views
Statistics
Mining
Chemistry
ChemAxon / Infocom
Distance Matrix
Meta
Flow Control
Misc
Community Nodes
KNIME Labs
Time Series
Reporting
XML
*0: OpenMS Demo
FileFilter
CID Spectra Selection
Node 4
Node 9
Input File
Input Spectra
Database w/ Decoy
OMSSA Executable
FileFilter
HCD Spectra Selection
Node 20
Node 10
FileMerger
Node 11
IDMerger
Node 12
ConsensusID
Node 13
Input File
FASTA
Node 14
Node 16
Node 17
ITRAQAnalyzer
Node 8
Node 18
Node 21
Node 47
Compute log2 Intensities
Node 52
Histogram (interactive
Node 45

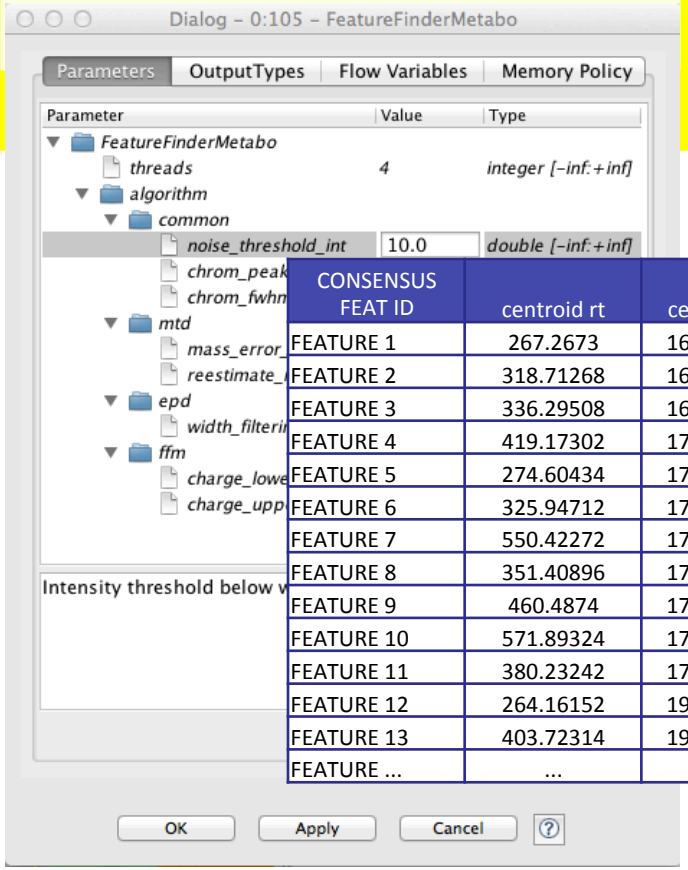
Metabolomics – Biomarker ID



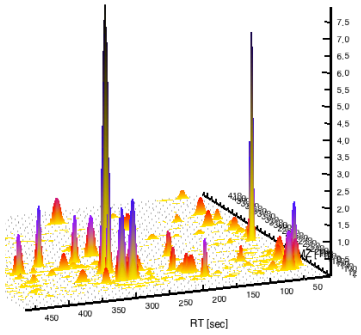
- Complex workflow analyzing a diabetes-related metabolomics biomarker study
  - Data preprocessing (yellow)
  - Quantification (purple)
  - Identification based on accurate mass/HMDB (gray)
  - Detection of distinctive features, statistics (green/gray)
  - Reporting of differential features and their structures (orange)

Metabolite Quantitation





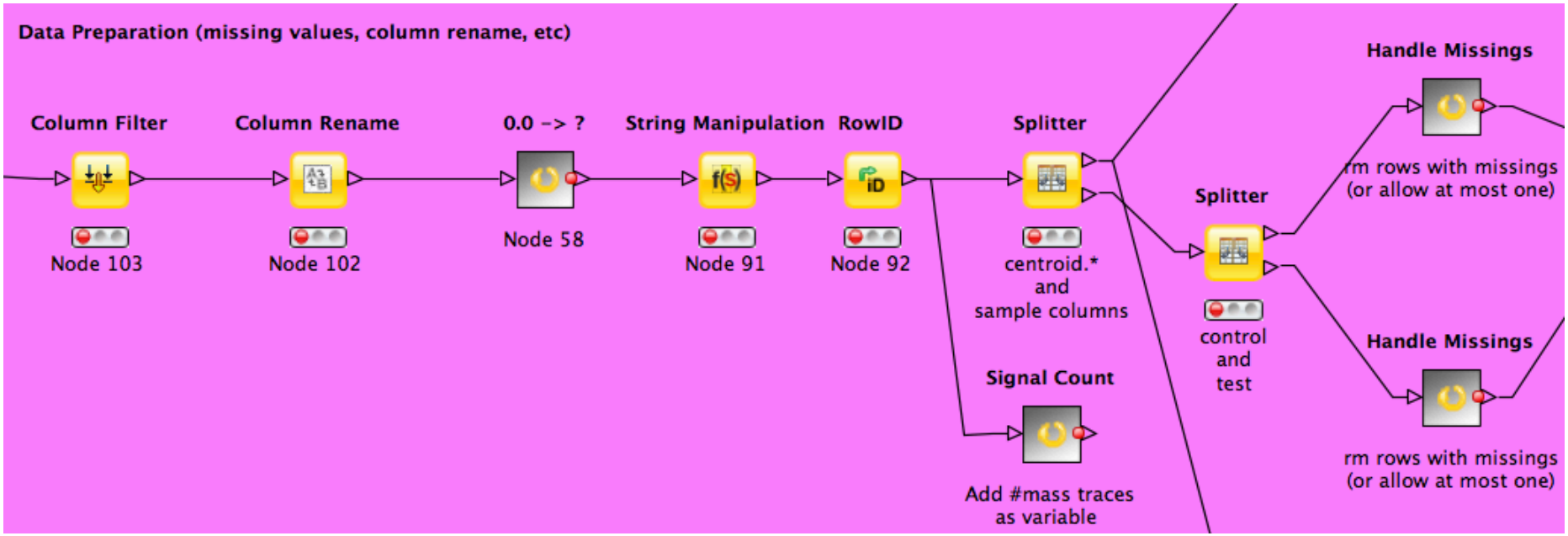
MSConvert



5 controls vs.
5 samples

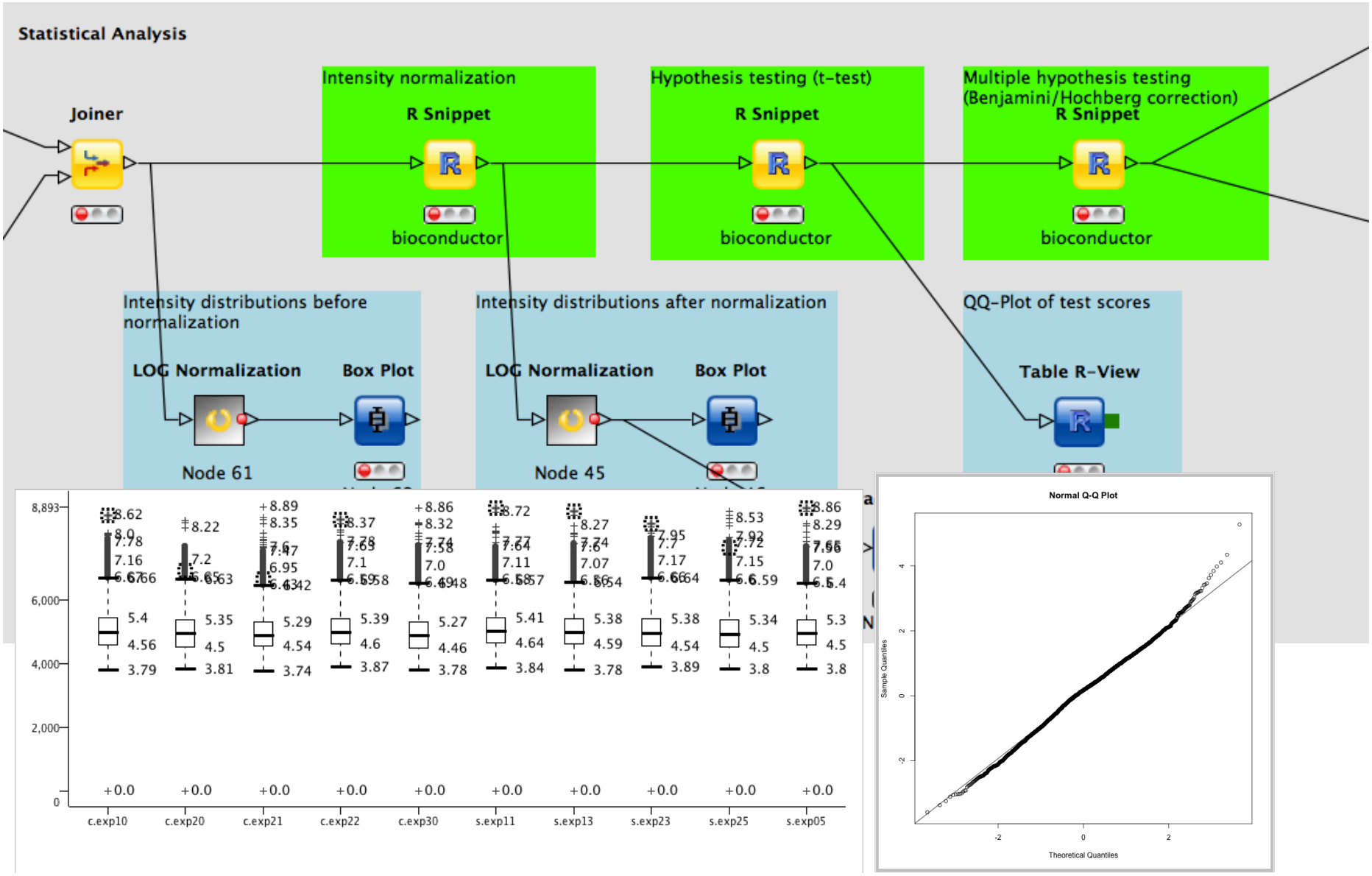
CONSENSUS FEAT ID	centroid rt	centroid m/z	...	charge	sample 1 intensity	sample 2 intensity
FEATURE 1	267.2673	163.0753568	...	1	5288099840	50020923440
FEATURE 2	318.71268	163.0753568	...	1	18835900	17835200
FEATURE 3	336.29508	163.0753568	...	1	7285210	6285210
FEATURE 4	419.17302	179.0702718	...	1	175022000	105022000
FEATURE 5	274.60434	179.0702718	...	1	44317400	33317400
FEATURE 6	325.94712	179.0702718	...	1	11875200	12879200
FEATURE 7	550.42272	179.0702718	...	1	4871360	5071360
FEATURE 8	351.40896	179.0702718	...	1	2919350	1019350
FEATURE 9	460.4874	179.0702718	...	1	2021340	3221340
FEATURE 10	571.89324	179.0702718	...	2	1546820	1446820
FEATURE 11	380.23242	179.0702718	...	2	1993120	1893120
FEATURE 12	264.16152	195.0651868	...	2	269592992	279592532
FEATURE 13	403.72314	195.0651868	...	2	21862600	20342600
FEATURE ...	...	...	...	...	...	...

Data Table Magic

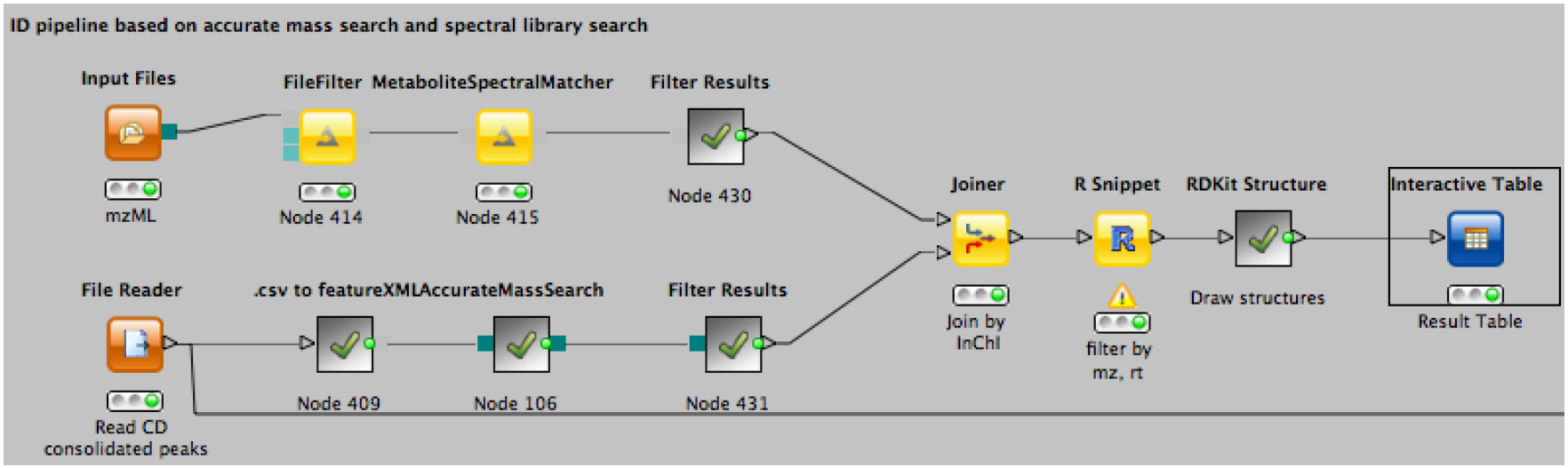


Row ID	centroid.rt	centroid.mz	...	charge	Control_1	Control_2	Sample_2	Sample_2
FEATURE 1	267.2673	163.0753568	...	1	5288099840	50020923440	5288099840	50020923440
FEATURE 2	318.71268	163.0753568	...	1	18835900	17835200	18835900	17835200
FEATURE 3	336.29508	163.0753568	...	1	7285210	6285210	7285210	6285210
FEATURE 4	419.17302	179.0702718	...	1	175022000	105022000	175022000	105022000
FEATURE 5	274.60434	179.0702718	...	1	44317400	33317400	44317400	33317400
FEATURE 6	325.94712	179.0702718	...	1	11875200	12879200	11875200	12879200
FEATURE 7	550.42272	179.0702718	...	1	4871360	5071360	4871360	5071360
FEATURE 8	351.40896	179.0702718	...	1	2919350	1019350	2919350	1019350
FEATURE 9	460.4874	179.0702718	...	1	2021340	3221340	2021340	3221340
FEATURE 10	571.89324	179.0702718	...	2	1546820	1446820	1546820	1446820
FEATURE 11	380.23242	179.0702718	...	2	1993120	1893120	1993120	1893120
FEATURE 12	264.16152	195.0651868	...	2	269592992	279592532	269592992	279592532
FEATURE 13	403.72314	195.0651868	...	2	21862600	20342600	21862600	20342600
FEATURE ...	...	...	...	...	...	...	...	...

Multiple Hypothesis Testing



Metabolite ID

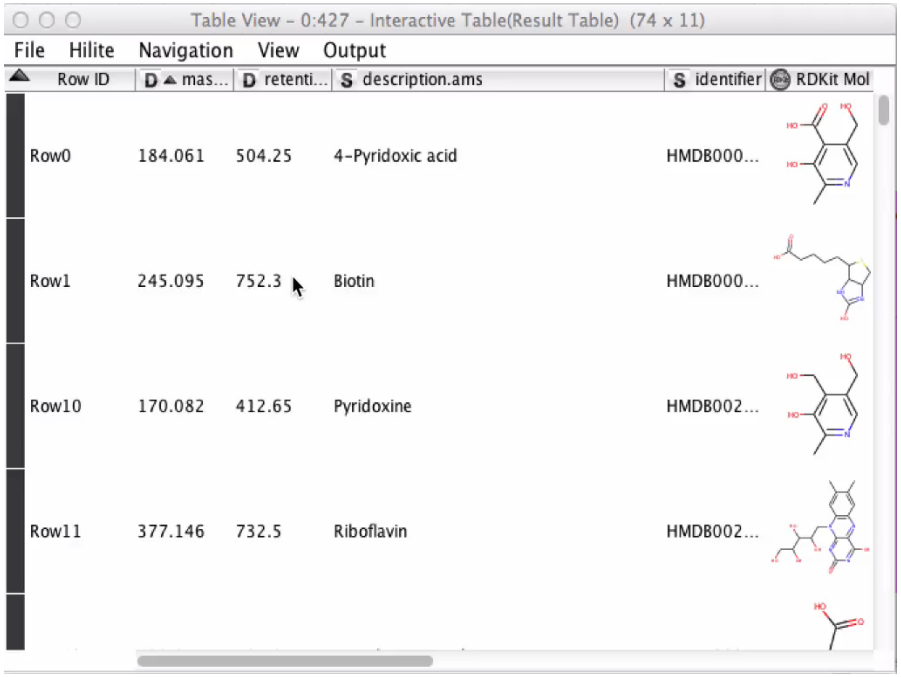


Multiple ID strategies

- Accurate mass
- Retention time database
- Retention time prediction
- Spectral matching

KNIME provides

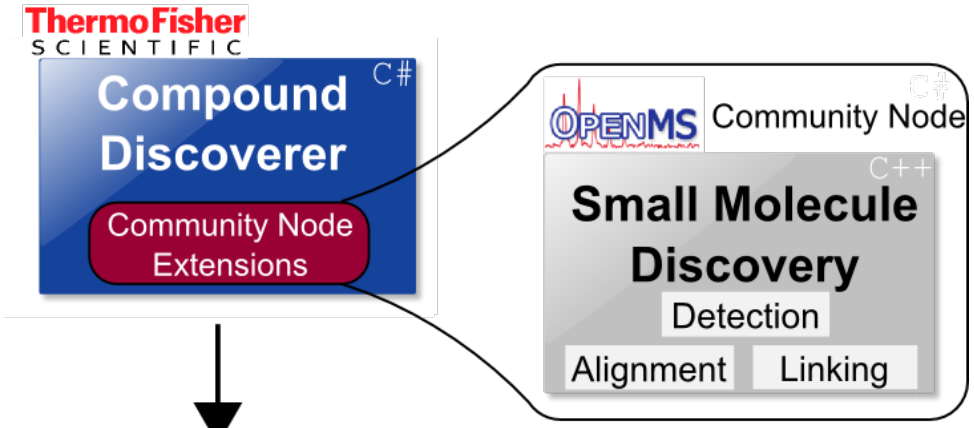
- Online access to structure databases
- Structure visualization
- Cheminformatics
  - Metabolization
  - Substructure search



Thermo Fisher Compound Discoverer

Community Node invokes OpenMS Tools

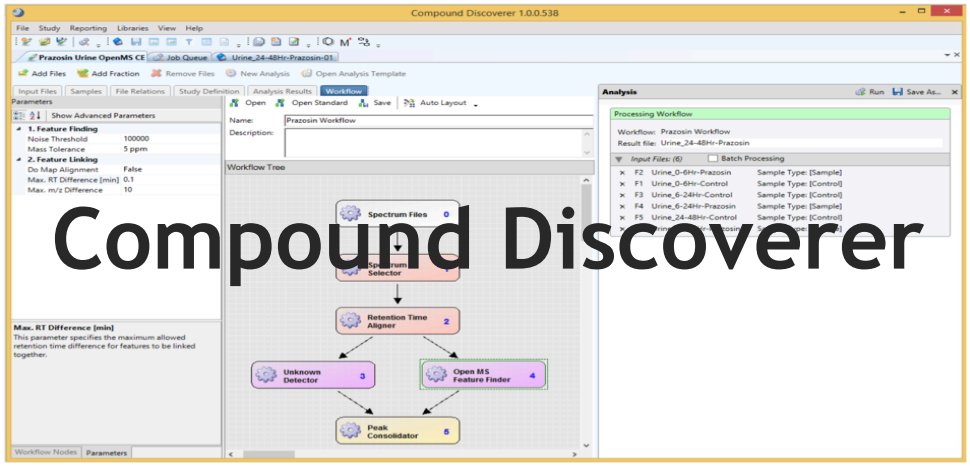
- **Workflows**
- **Analysis Tools**
- **Visualization**



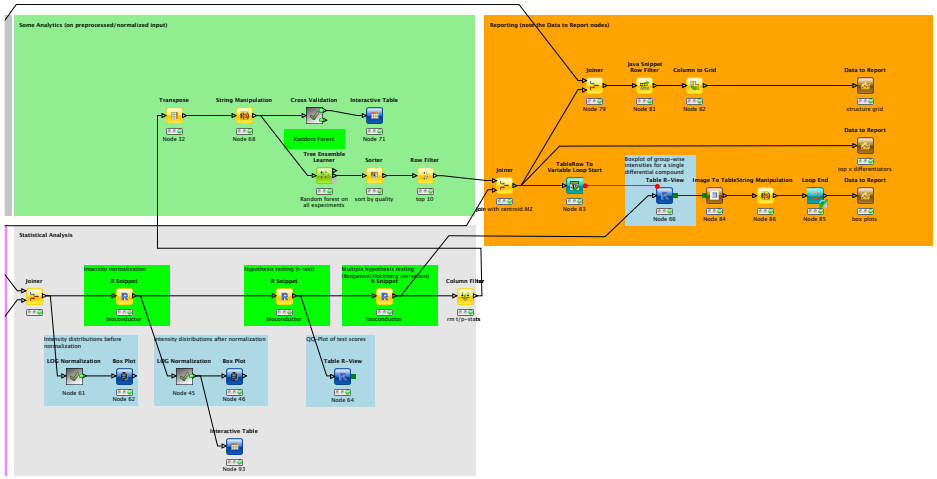
Results exported as Table

- **Quality Control**
- **Statistical Analysis**

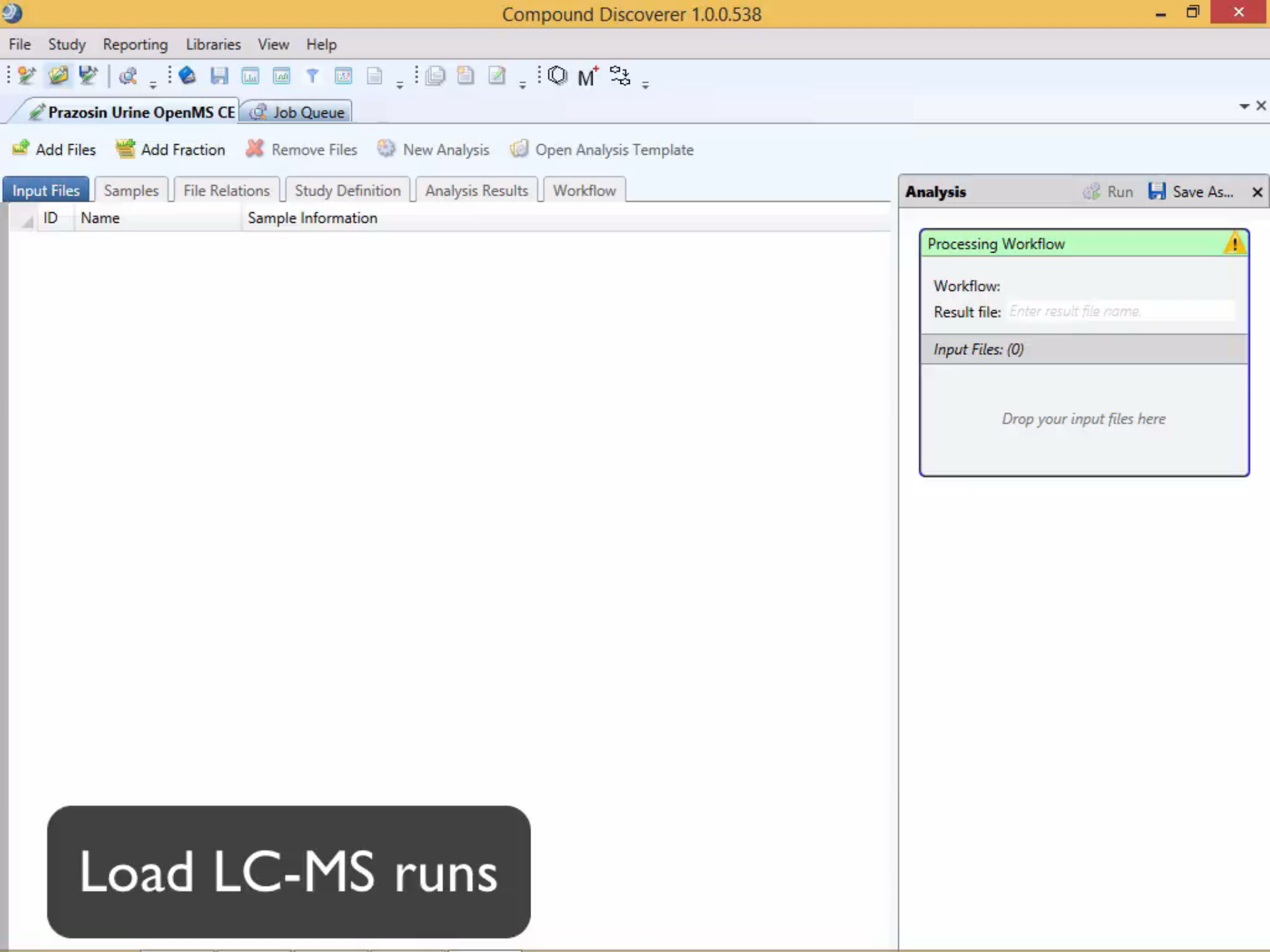
Metabolomics – Biomarker ID



Compound Discoverer



- Complex workflow for a diabetes-related metabolomics biomarker study (normal vs. disease)
  - Data preprocessing (yellow)
  - Quantification (purple)
  - Identification based on accurate mass/HMDB (gray)
  - Detection of distinctive features, statistics (green/gray)
  - Reporting of differential features and their structures (orange)


Compound Discoverer 1.0.0.538
File
Study
Reporting
Libraries
View
Help
Prazosin Urine OpenMS CE
Job Queue
Add Files
Add Fraction
Remove Files
New Analysis
Open Analysis Template
Input Files
Samples
File Relations
Study Definition
Analysis Results
Workflow
ID
Name
Sample Information
Analysis
Run
Save As...
Processing Workflow
Workflow:
Result file:
Enter result file name.
Input Files: (0)
Drop your input files here
Load LC-MS runs

References

- **XCMS**
  - C.A. Smith, E.J. Want, G.C. Tong, R. Abagyan, and G. Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal. Chem., 2006,
- **FeatureFinderMetabo**
  - Kenar, E, Franken, H, Forcisi, S, Wörmann, K, Häring, H, Lehmann, R, Schmitt-Kopplin, P, Zell, A, and Kohlbacher, O (2014). Automated Label-Free Quantification of Metabolites from LC-MS Data. Mol. Cell. Prot., 13(1):348-59. http://dx.doi.org/10.1074/mcp.M113.031278

39

Materials

- Learning Units 12A, 12B