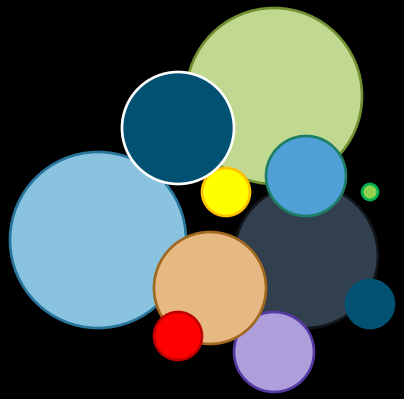
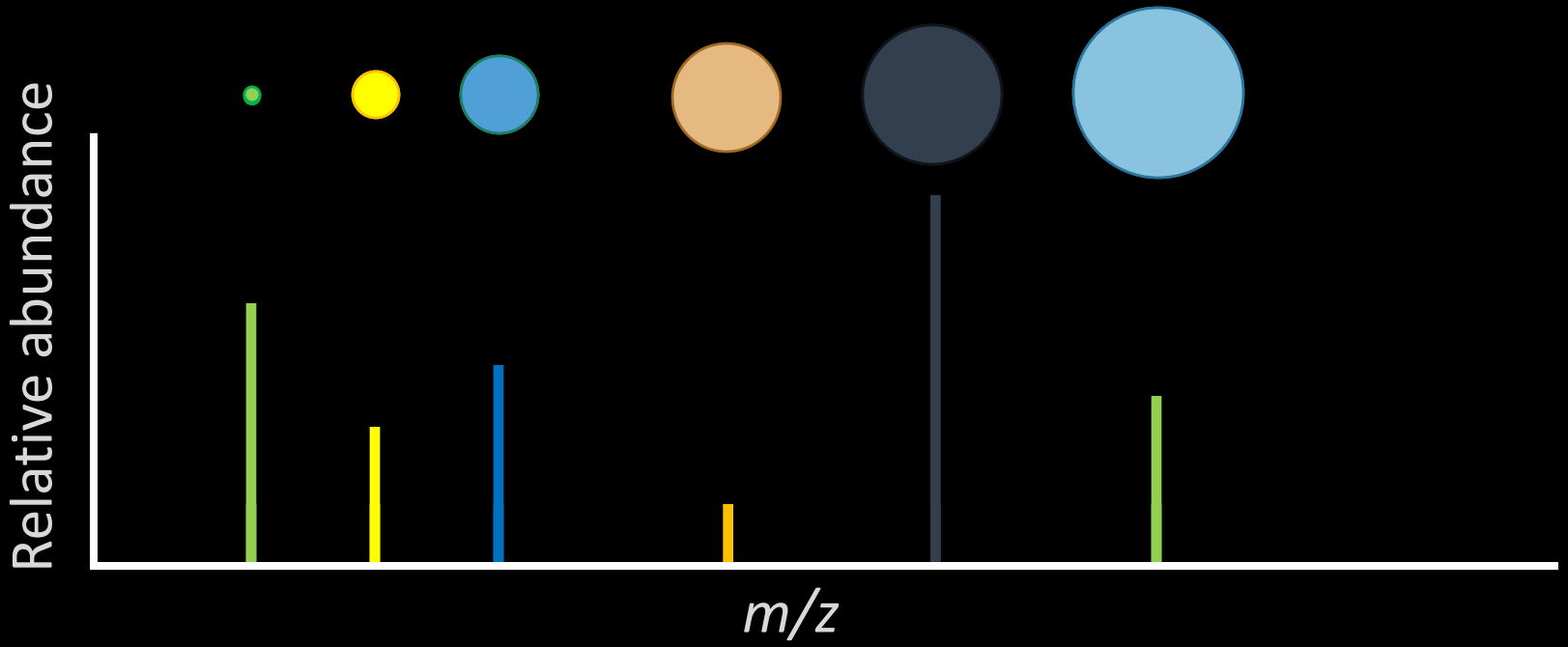
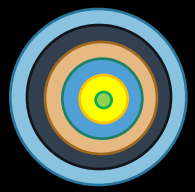


mass

mass


Relative abundance
m/z

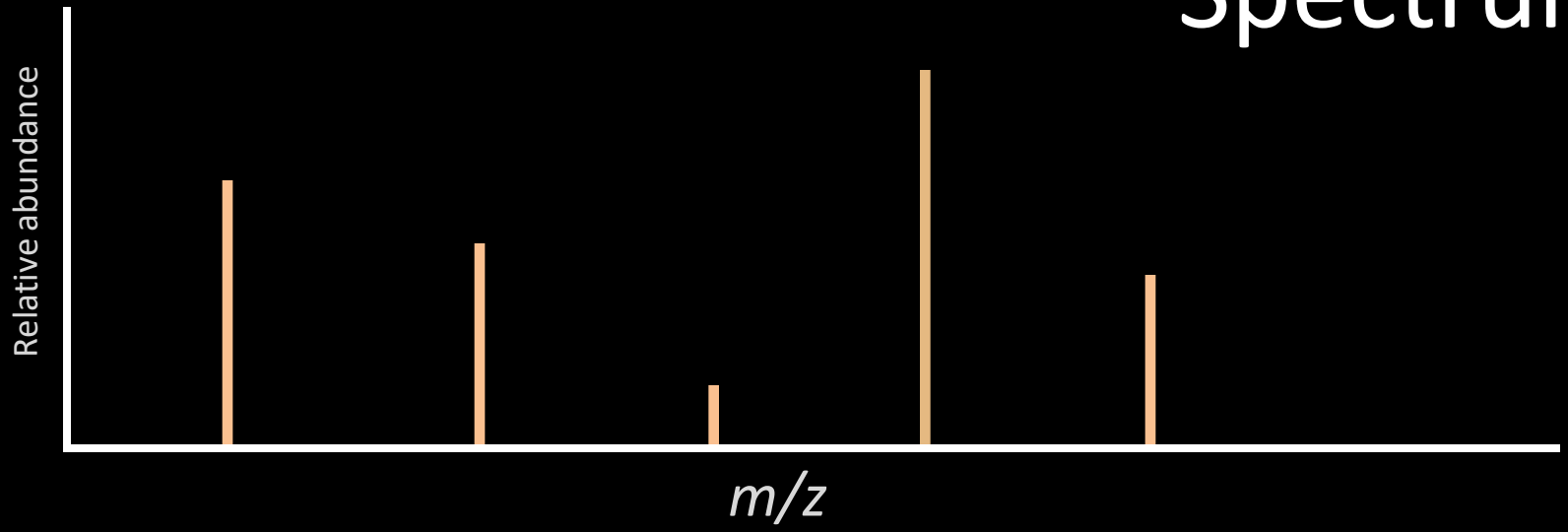

mass



MS^2

MS/MS

Tandem Mass Spectrum



Mining the Plant Specialized Metabolome with Mass Spectrometry:

Library Matching and Molecular Networking with GNPS

Justin J.J. van der Hooft et al.

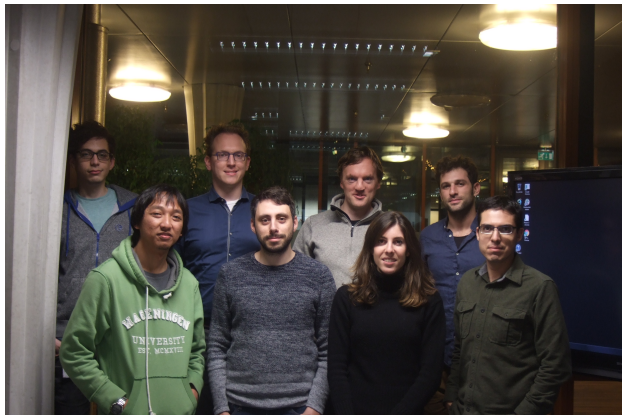
Bioinformatics Group – Wageningen University, NL

Online Workshop 10 March 2021

WAGENINGEN UNIVERSITY
WAGENINGEN UR

100years

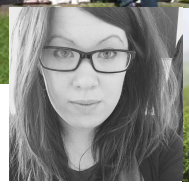
Team work! ☺

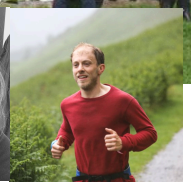


Medema lab
Wageningen University

NL eScience Center
€€ ASDI grant €€

Dorrestein lab – San Diego, USA





Glasgow Polyomics – University of Glasgow

WAGENINGEN UNIVERSITY
WAGENINGEN UR

University of Strathclyde Glasgow

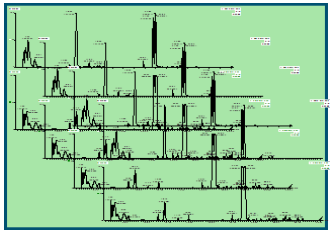
University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Typical metabolomics workflow

sample series

generating LCMS / GCMS profiles



mass chromatograms

Data pre-processing
- unbiased peak picking
  - >100.000 mass signals
- alignment, filtering, assembling
- > 100's – 1,000's of compounds

dedicated workflow

Data analyses
- multivariate analysis
- regression analysis
- Student T-tests
- .....

statistics

Identification of selected metabolites
- most differential by treatment
- best correlation with trait
- mapping on genome
- .....

data interpretation

Workshop objectives

Being able to:

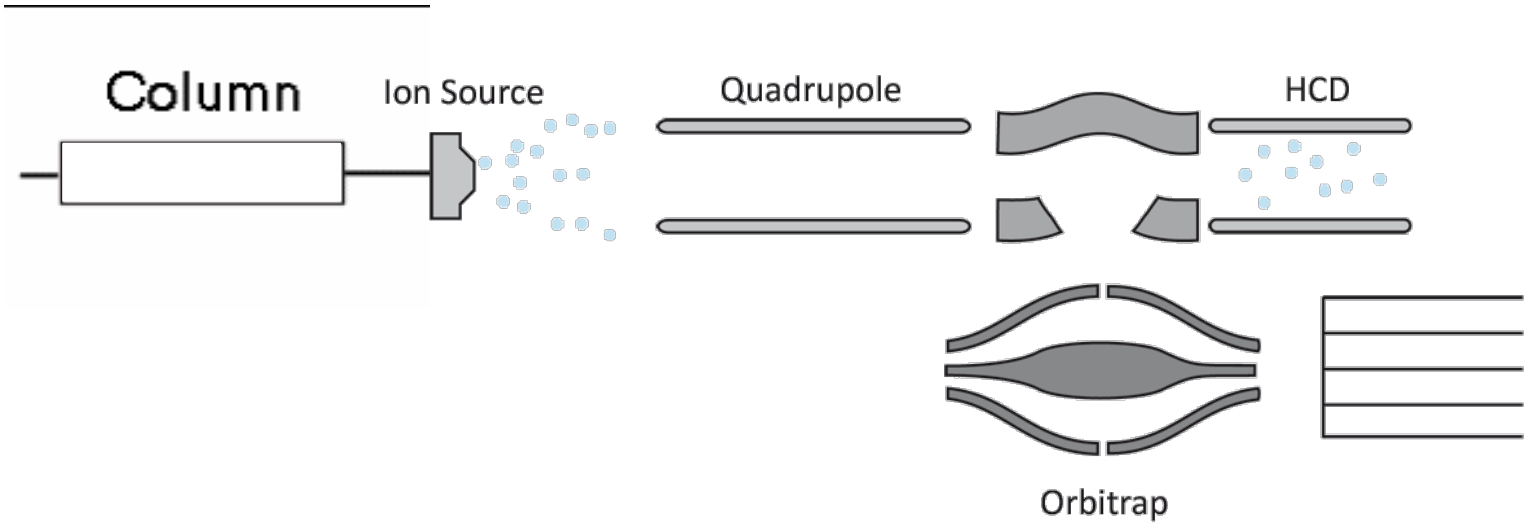
- Explain rationale behind metabolome mining tools
- Explore and assess GNPS Library Matches
- Explore and assess GNPS Molecular Families

Have:

- Fun

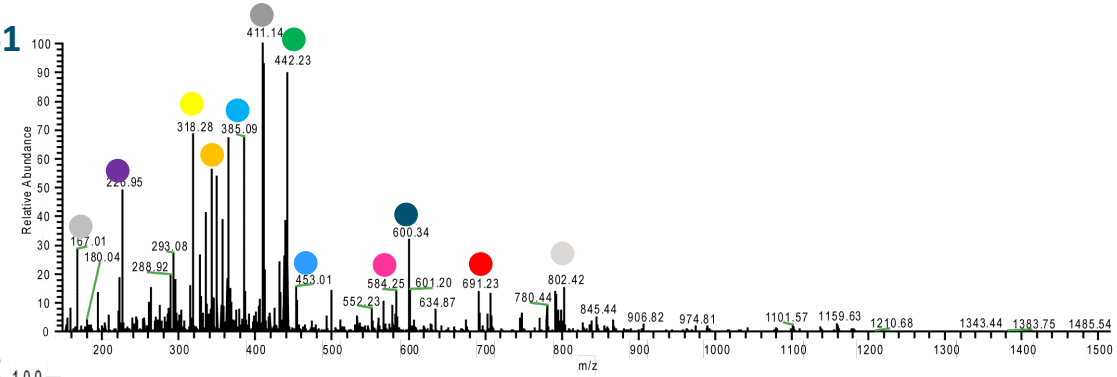
Why metabolome mining?

LC-MS/MS Data Structure

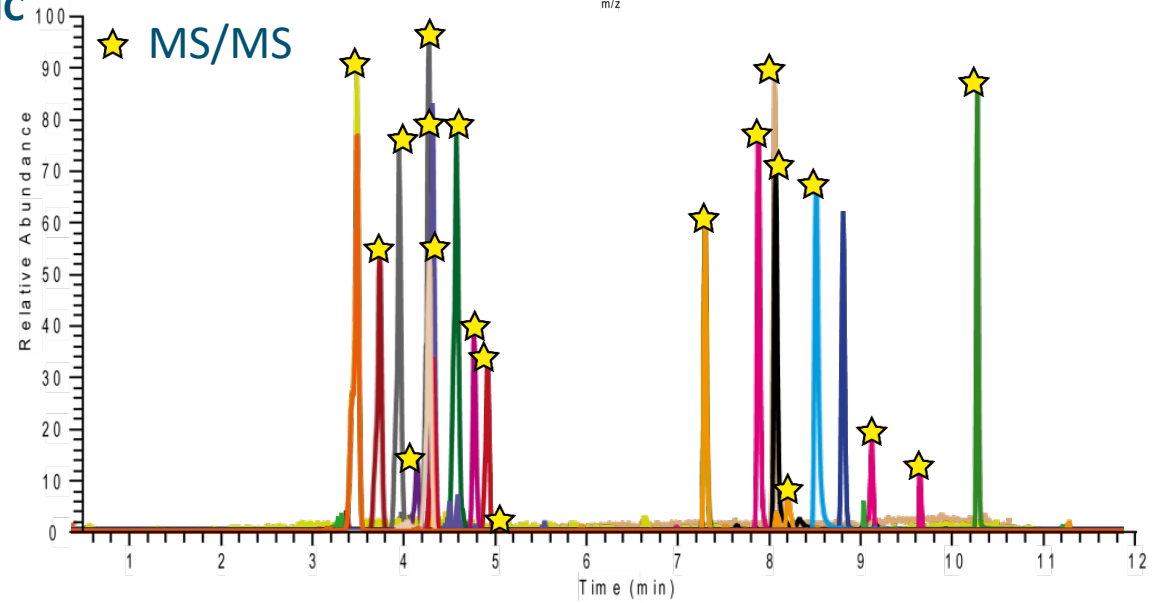


LC-MS/MS Data Structure

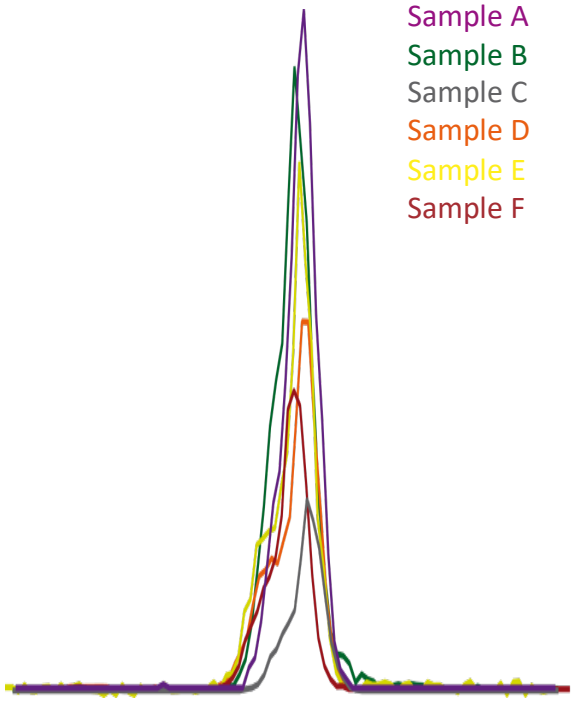
MS1



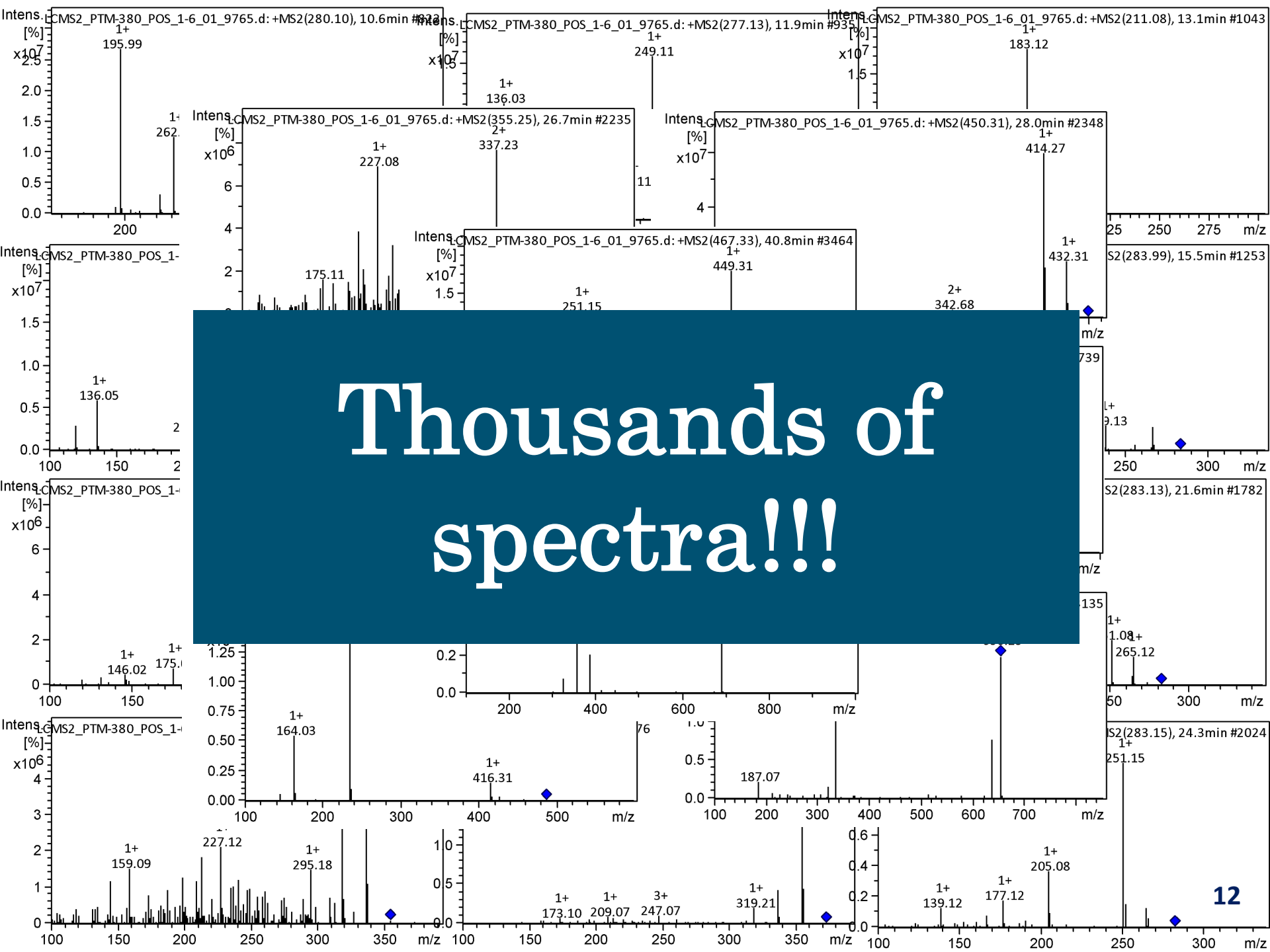
XIC

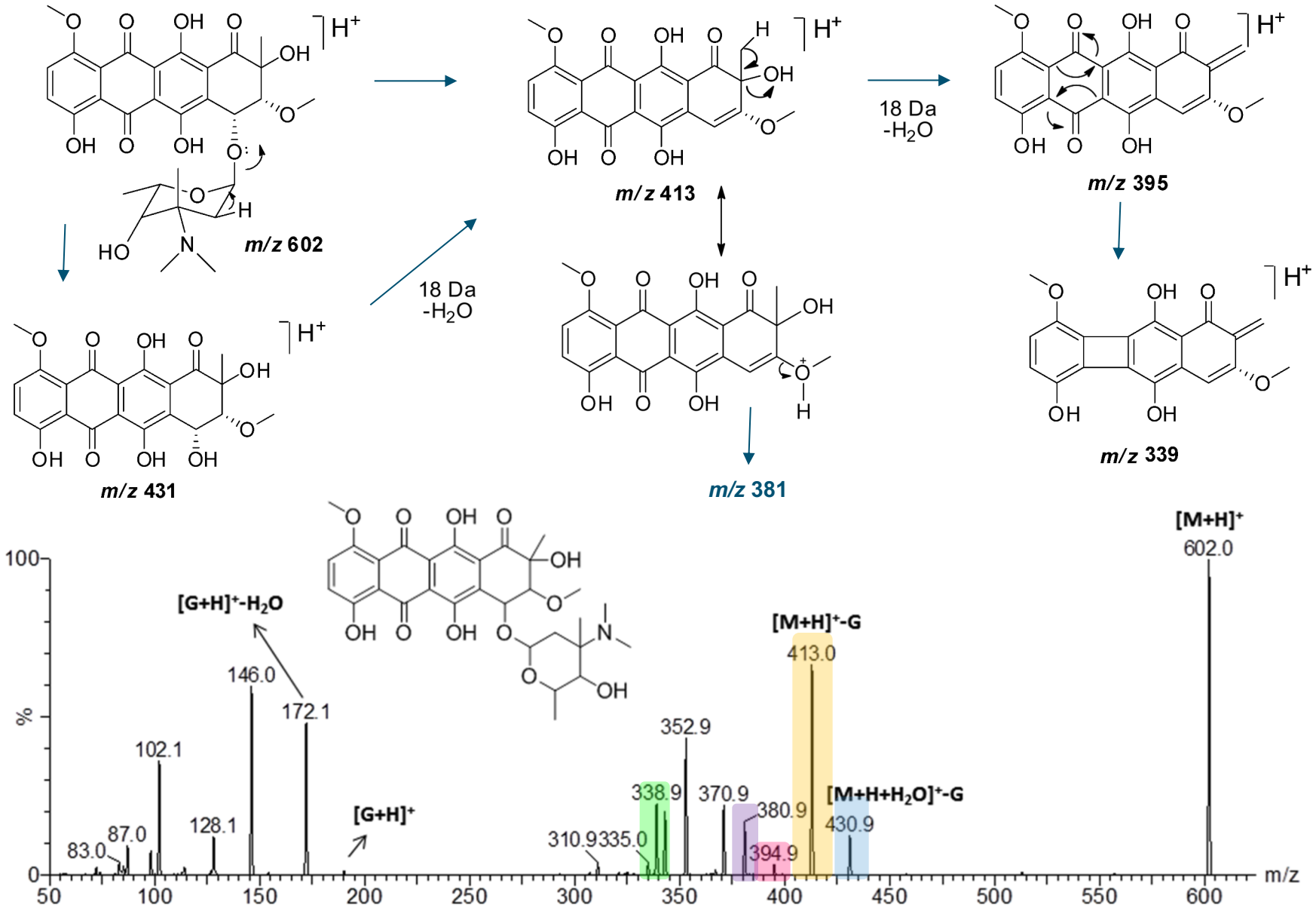


Aligned Chromatographic Peaks



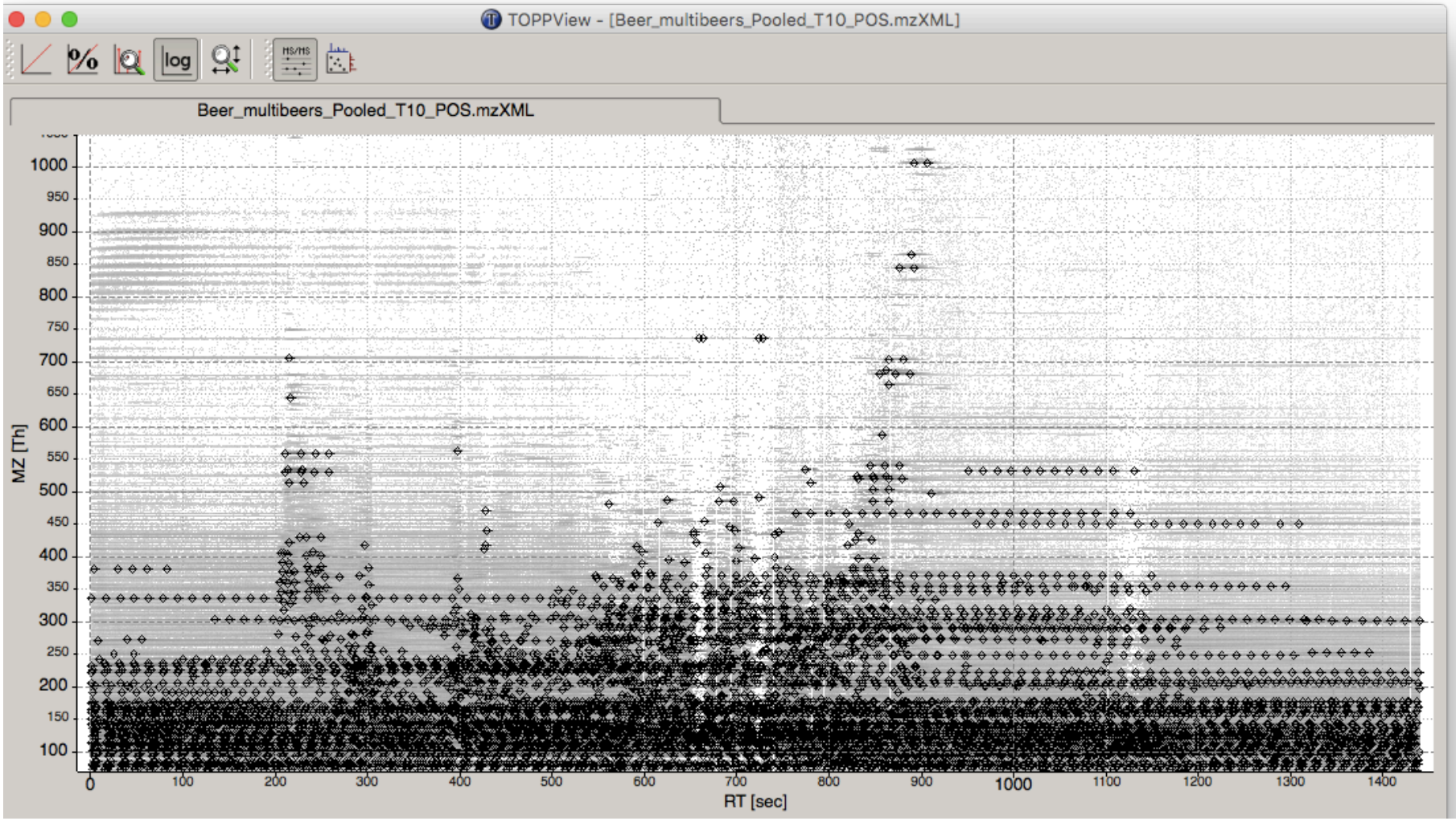
WAGENINGEN UNIVERSITY
WAGENINGEN UR


Thousands of spectra!!!



Bauermeister, A., Zucchi, T.D., & Moraes, L.A.B. Journal of Mass Spectrometry, 2016.

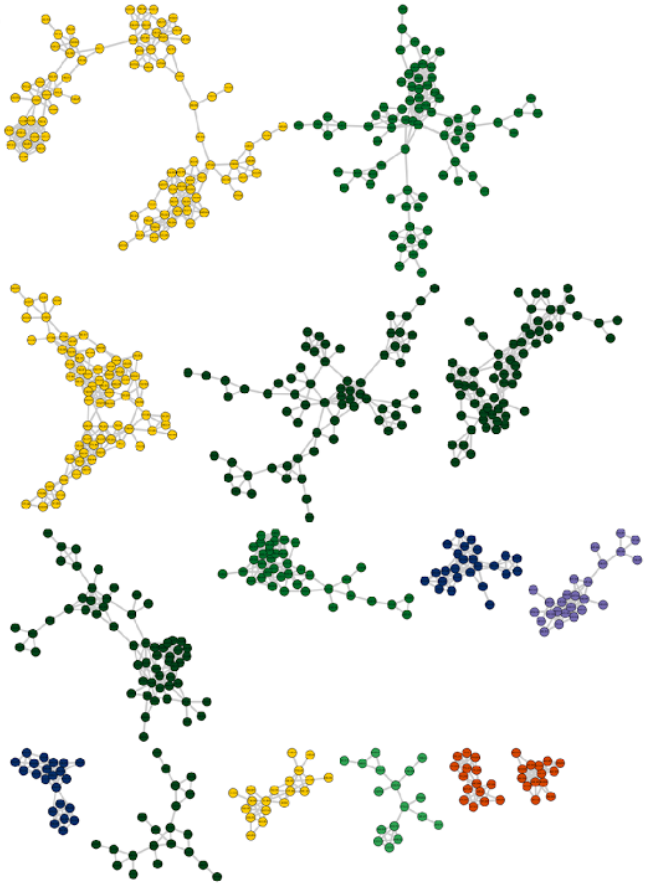
Why metabolome mining?



Why metabolome mining?

GNPS

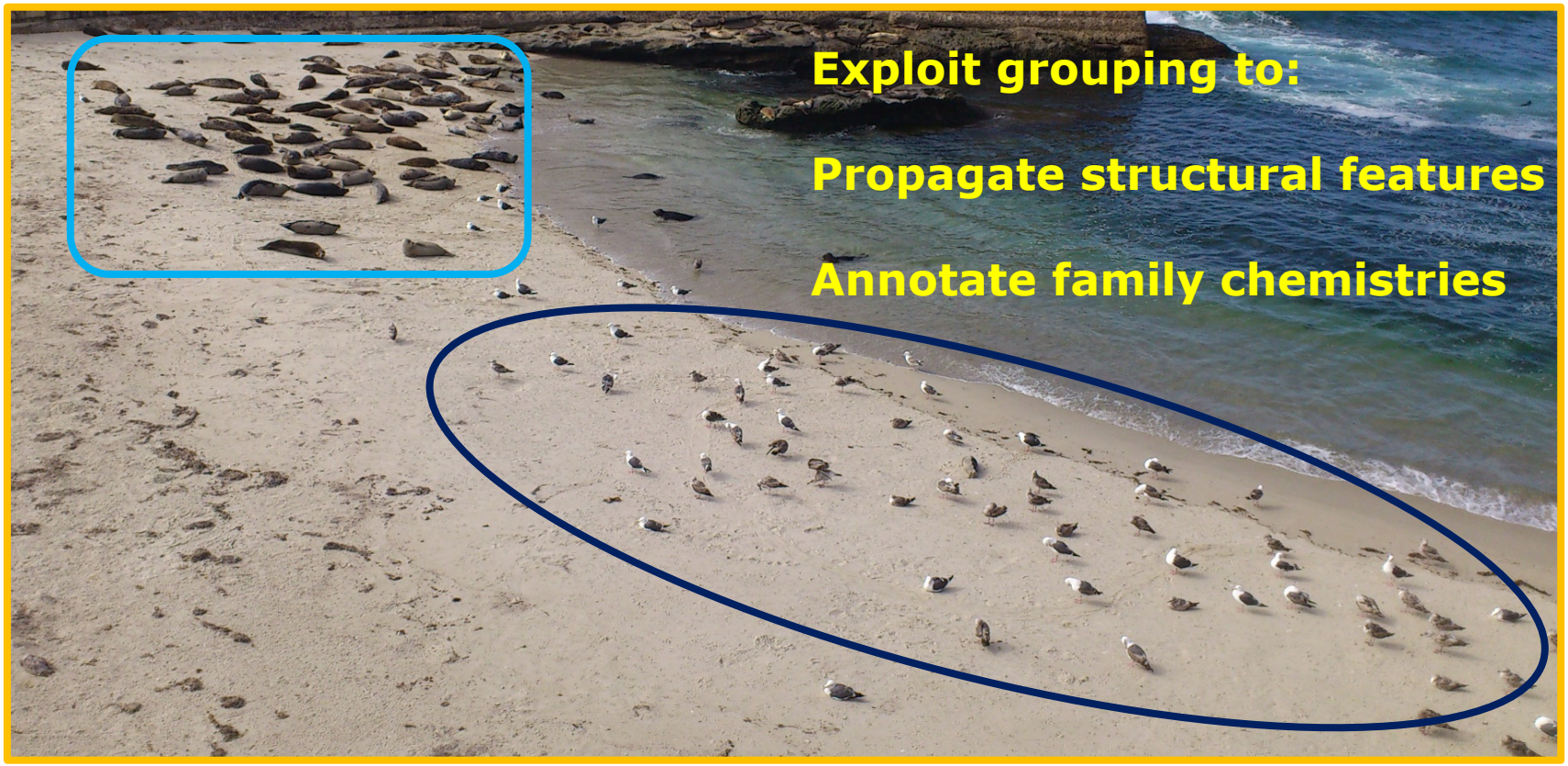
MS2LDA



Improved annotation power by pattern mining

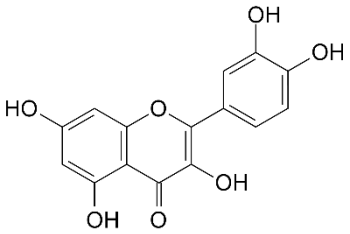
Finding **Molecular Families** by spectral similarity

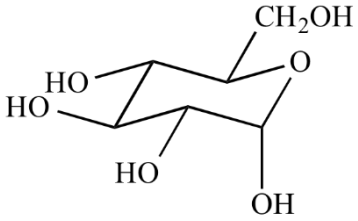
Extract "building blocks of metabolomics" = **substructures**



Building blocks of metabolomics?!

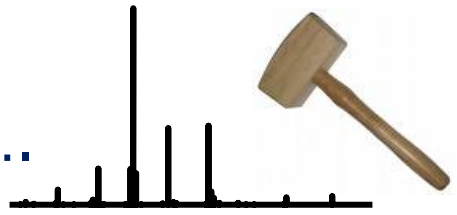
- Metabolites share biochemical substructures!





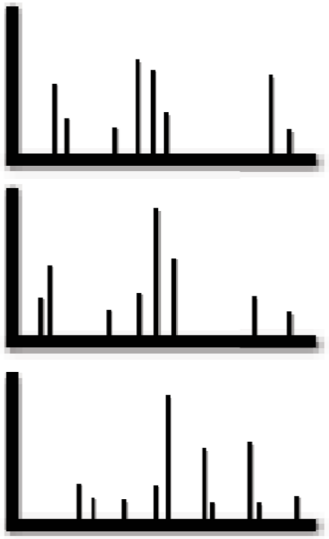
- Revealed by **mass spectrometry fragmentation** as:
  - Substructures often produce similar fragments or neutral losses….

- But remember….untargeted metabolomics….

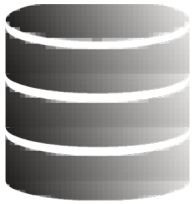


Large-scale Library Matching

MS/MS spectral data

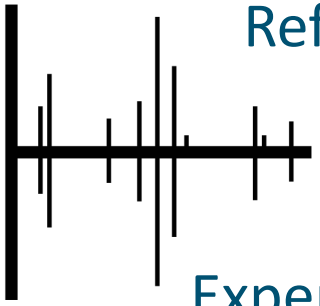


MS/MS library spectra



GNPS spectral libraries
MassBank
NIST
MoNA
METLIN

Spectral match

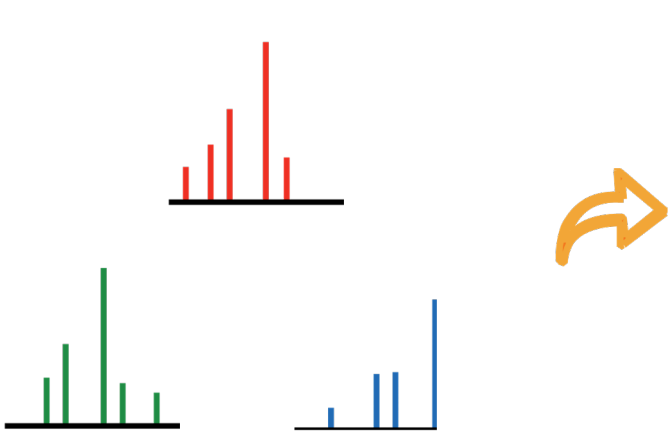


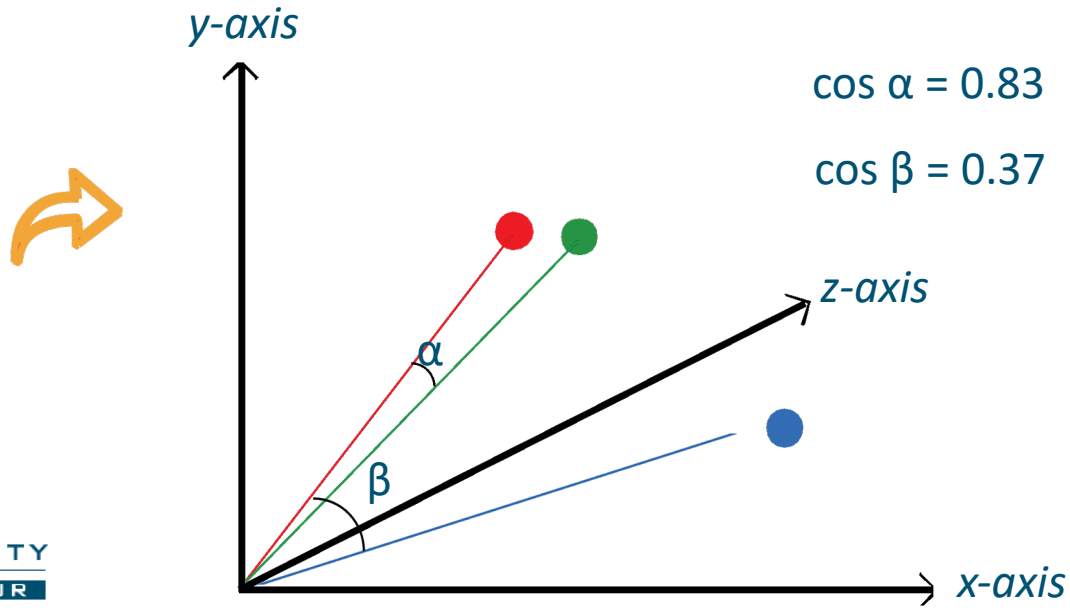
Mirror Plot

Spectral Similarity – cosine score

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Based on Mass Fragmental Overlap!





WAGENINGEN UNIVERSITY
WAGENINGEN UR

Validating matches

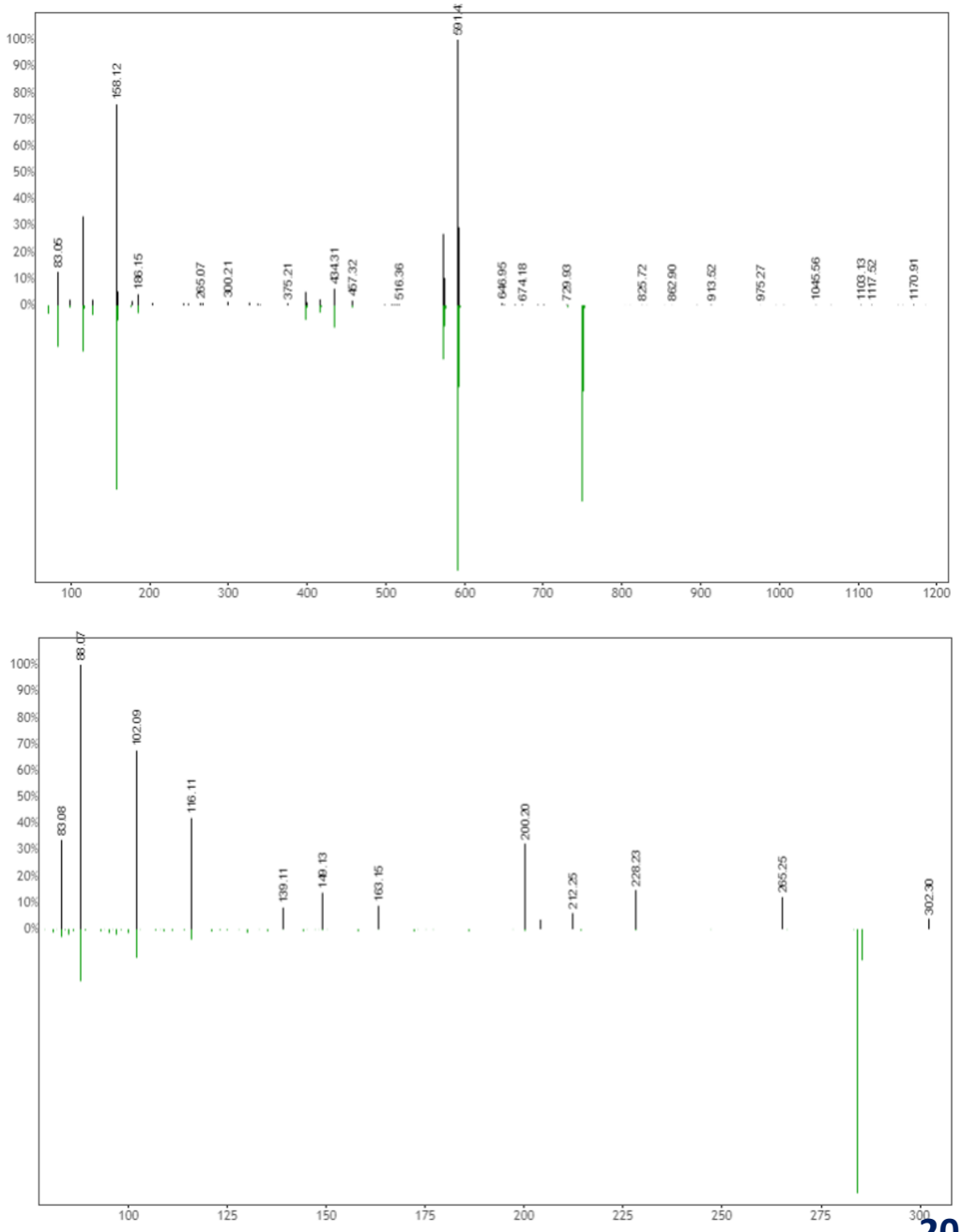
Mirror plot

Parent mass differences

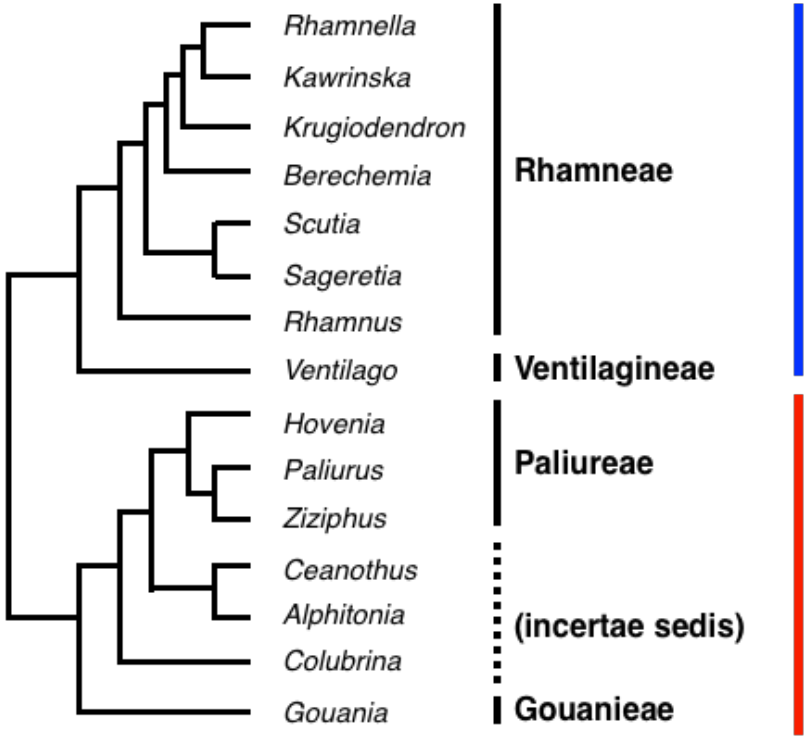
Number of ions that match

Retention time

Metadata



Application: Rhamnaceae plant family



Dr Kyo Bin Kang

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Sun et al. (2016), J. Syst. Evol. 54, 363-391.

Kang, Ernst, van der Hooft et al., The Plant Journal, 2019

Practice time! (15 min)

- Check a number of library matches:
  - What kind of metabolites are matches to the Rhamnaceae plant mass spectral data?
  - Do they make sense to you?
  - Are they trustworthy in your opinion?
    - Tip: study the mirror plots!
- Make a screenshot of two reliable and two non-reliable matches:
  - Share these screenshots in the Zoom chat
  - Paste these screenshots in a ppt presentation

Analyze Molecular Networking results

Browse to:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9eeb9b1cebf74305aa2b2f99e167f8cf

Job Status

Workflow	FEATURE-BASED-MOLECULAR-NETWORKING (version release_27)
Status	DONE [Clone] [Clone to Latest Version] [Restart][Delete] **Default Molecular Networking Results Views** [View All Library Hits | View Unique Library Compounds | View All Analog Library Hits | View All Spectra With IDs | Feature Quant Details List | File Summaries] **Network Visualizations** [View Spectral Families (In Browser Network Visualizer)] **Methods and Citation for Manuscripts** [Networking Parameters and Written Description] **Export/Download Network Files** [Download Cytoscape Data] **Advanced Views - Misc Views** [View Network Pairs | Networking Statistics] **Advanced Views - Networking Graphs/Histograms** [Edges, MZ Delta Histogram] **Advanced Views - External Visualization** [Direct Cytoscape Preview/Download | Direct Cytoscape IIN Collapsed Preview/Download | Global Comparison with ReDU PCA (Beta)] **Advanced Views - External Tools** [View Dereplicator Results] **Advanced Views - Experimental Views** [Analyze with MS2LDA | Enhance with MolNetEnhancer | Visualize with Qemistree | Network with Spec2vec] **Advanced Views - qiime2 Views** [View qiime2 Emperor Plots | View qiime2 Emperor Bi-Plots | Download qiime2 Emperor qzv | Download qiime2 features biom qza] **Advanced Views - Stats Views (Experimental)** [View All Column Plots | View Select Column Plots | Data Exploration with Interactive Plotting | API Data for Plotting] **Advanced Views - Metadata Views** [View Metadata]

Different Levels of Metabolite Annotation

MSI levels: A formal definition of metabolite annotation and identification of the Metabolomics Standard Initiative. It comprises four levels:

Level 1 - Identified metabolites;

Level 2 - Putatively annotated compounds;

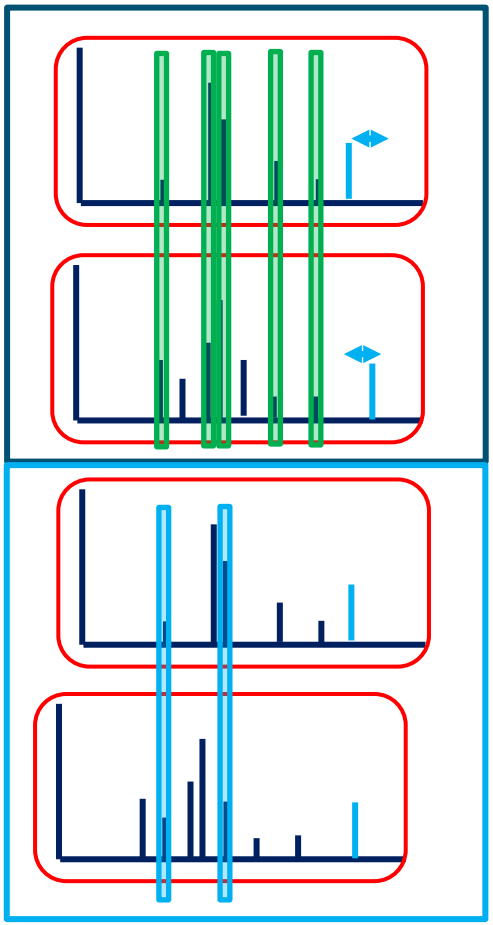
Level 3 - Putatively characterized chemical classes;

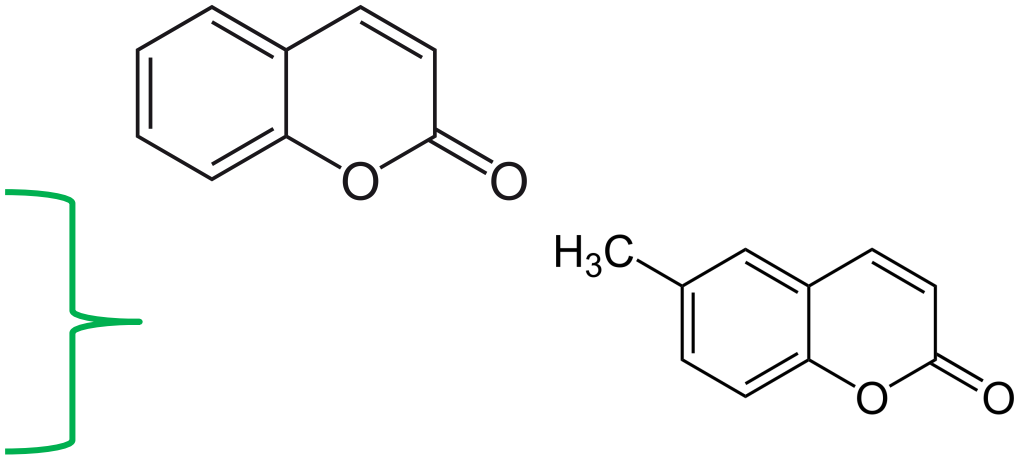
Level 4 – Unknown

Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics, 2007.

The modified cosine score....

Molecular Networking

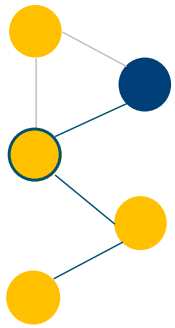




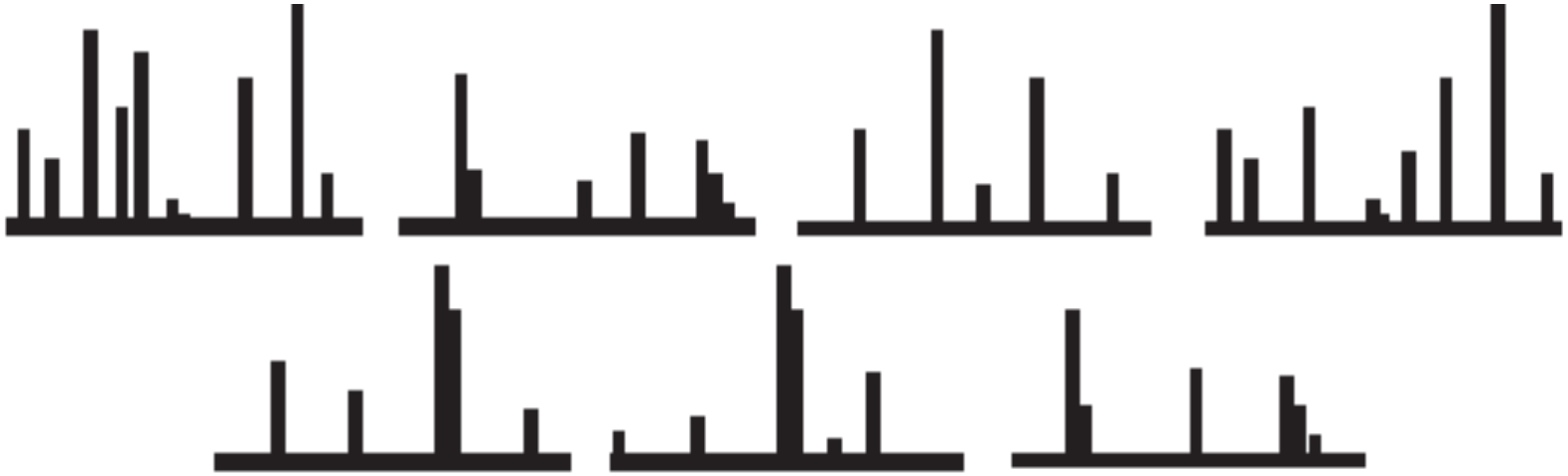
Spectral similarity = structural similarity

Spectral comparison → Score (modified cosine)

Threshold for network



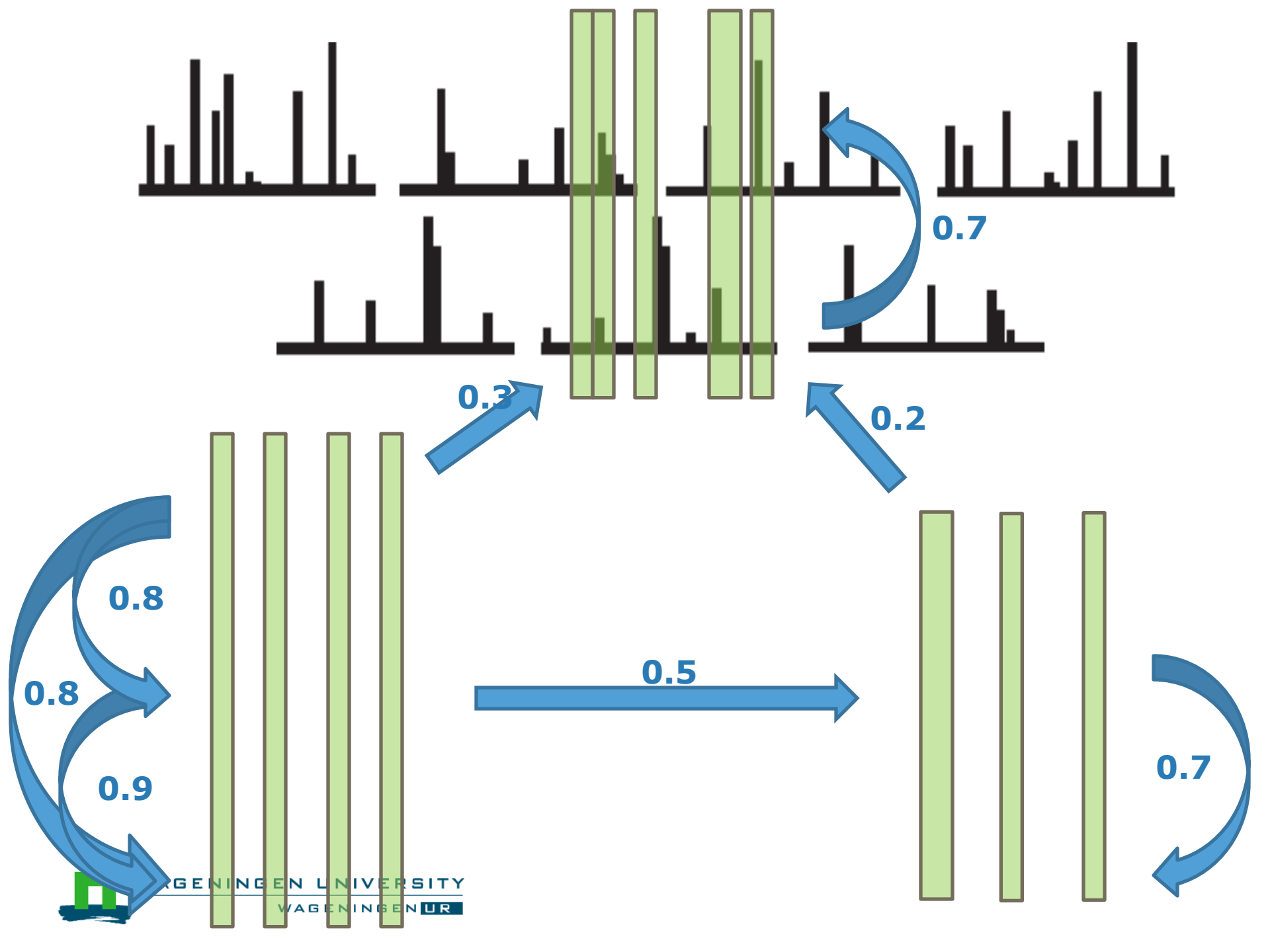
Wolfender et al., Anal. Chem., 2018



Molecular Networking

Very similar MS/MS spectra are grouped into Molecular Families

Wang, M., et al., *"Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking"*
Nat Biotech (2016)
Watrous, JD *et al*. "Mass spectral molecular networking of living microbial colonies" *PNAS* (2012)



Spectral library matches from GNPS

Libraries from diverse sources

Seed node annotations for molecular families

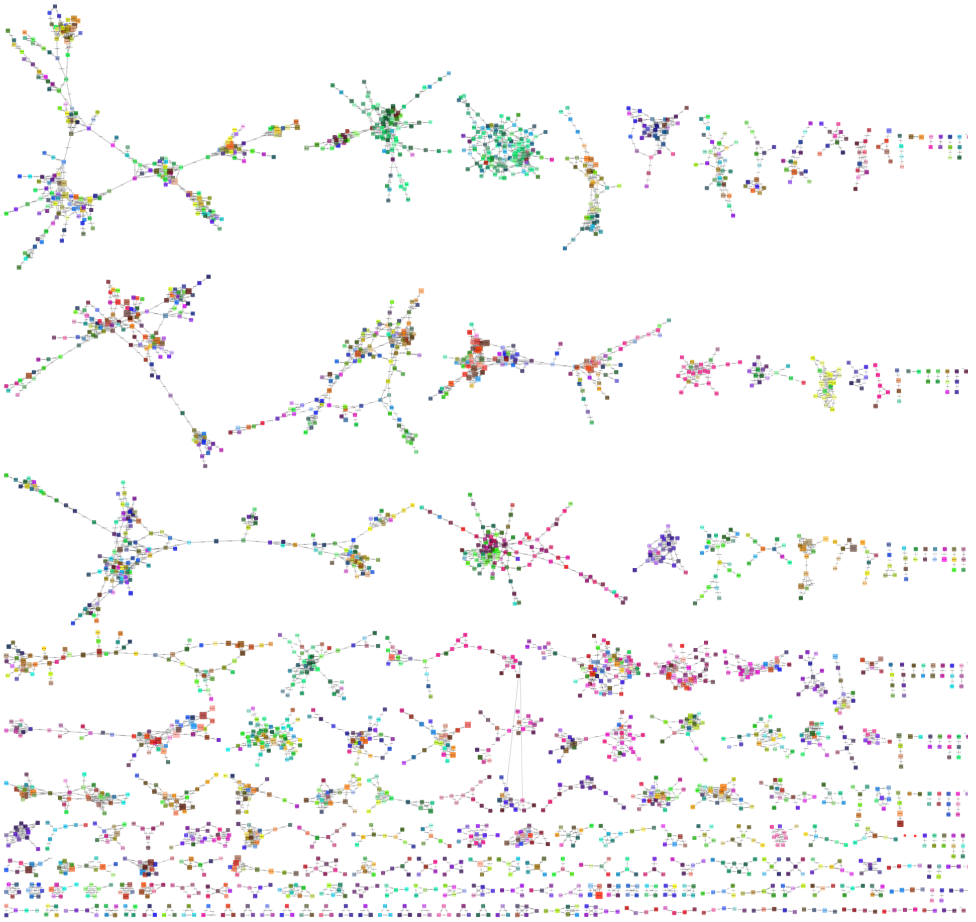
Library MS/MS Spectra

Diterpenoids

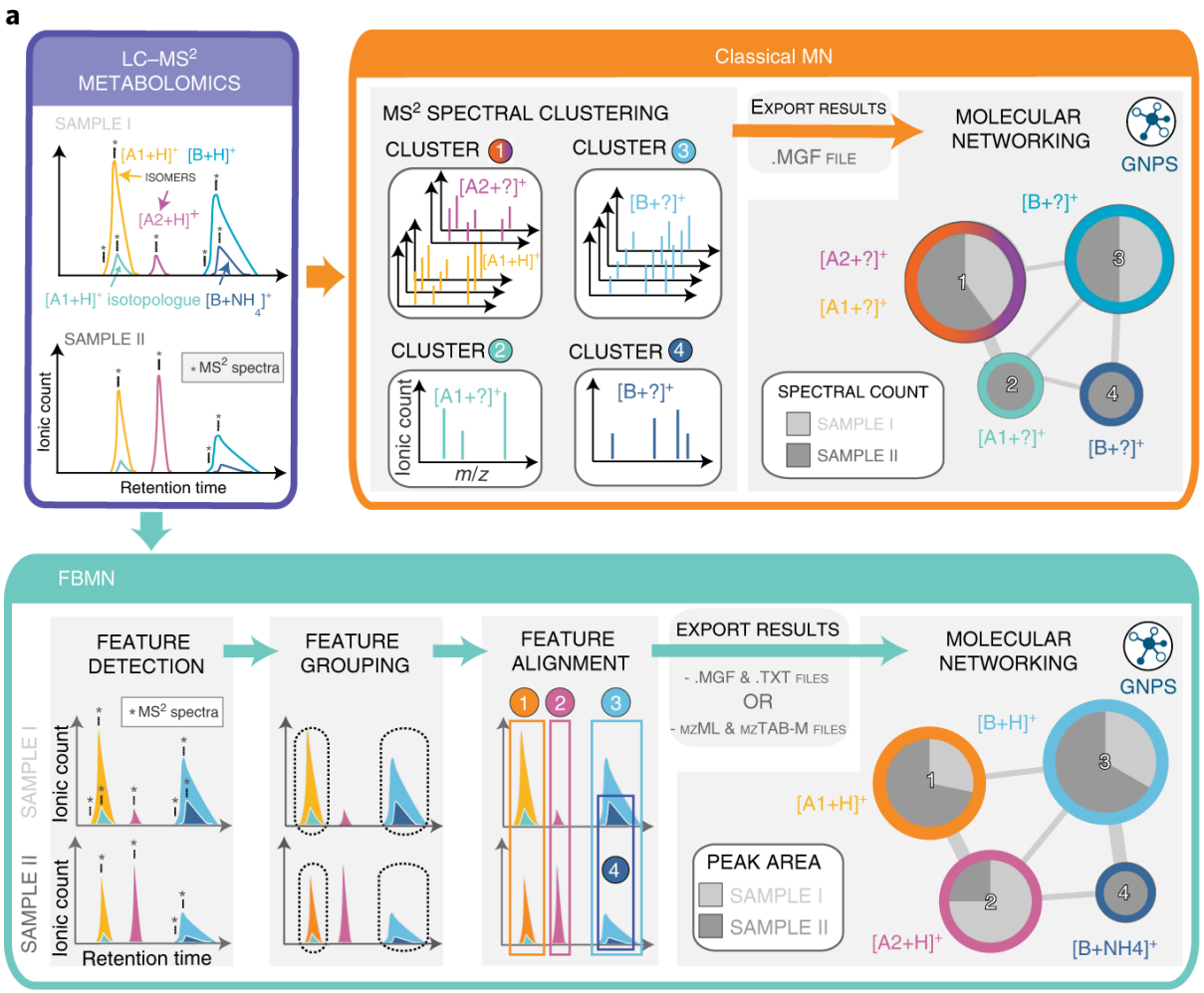
Flavonoids

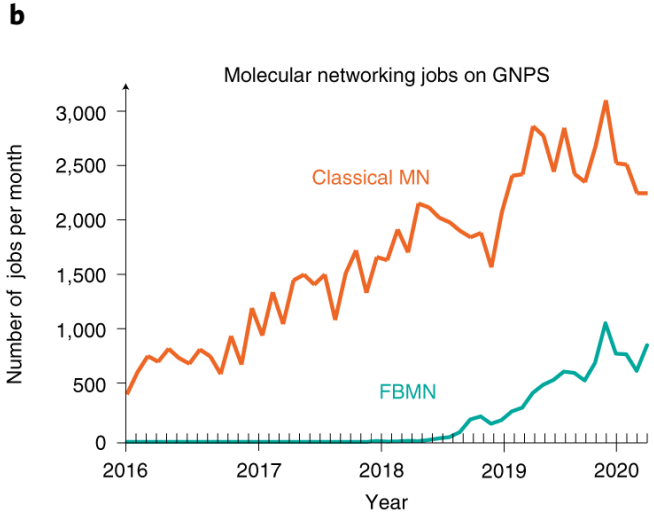
Pharmaceuticals

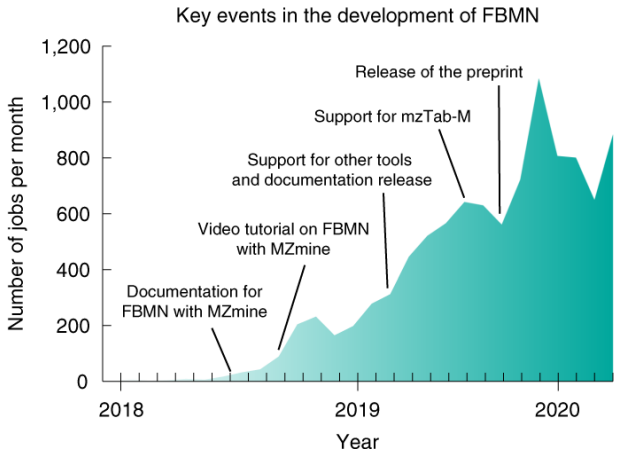
MS contaminants



Feature-based Molecular Networking





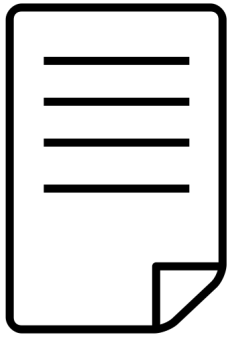


Nothias et al., Nature Methods, 2020

WAGENINGEN UNIVERSITY
WAGENINGEN UR

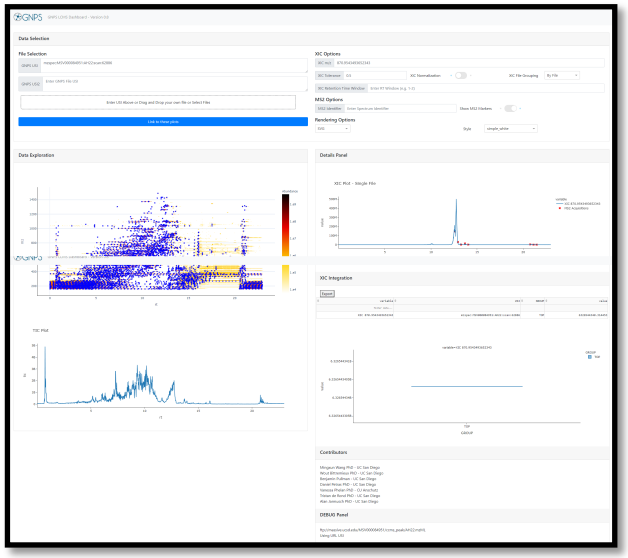
Interactively Exploring LCMS Data

Full LC/MS Mass Spec File

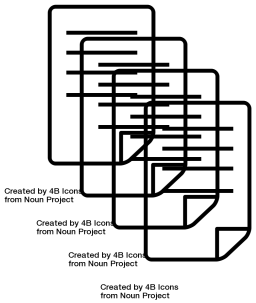


Created by 4B Icons
from Noun Project

Visualize



Full LC/MS Mass Spec Files



Analyze

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Practice time! (25 min)

- Find the molecular family from your breakout group name. What can you learn about this family?
  - Try out different visualization options in the browser
  - Make a screenshot of the nicest layout you had, share it in the Zoom chat, and put it in your ppt.
- Study the Rhamnaceae metadata shortly. Then go to GNPS Interactive Plotting and study the metabolite's behaviour and also of some of its connected spectra in the molecular family using their feature ids
  - Tip: try genus and clade as metadata to plot
  - The Feature IDs can be found as cluster index in the the GNPS Browser Network Visualizer.
  - Make a screenshot of the nicest layout you had, share it in the Zoom chat, and put it in your ppt.

Analyze Molecular Networking results

Browse to:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9eeb9b1cebf74305aa2b2f99e167f8cf

Job Status

Workflow FEATURE-BASED-MOLECULAR-NETWORKING (version release_27)

DONE
[Clone] [Clone to Latest Version] [Restart][Delete]

Default Molecular Networking Results Views
[View All Library Hits | View Unique Library Compounds | View All Analog Library Hits | View All Spectra With IDs | Feature Quant Details List | File Summaries]

Network Visualizations
[View Spectral Families (In Browser Network Visualizer)]

71 Rhamnaceae extracts - FBMN MzMine release 27 Hits 1 ~ 30 out of 143 Go to Go

Select columns

Filter	Visualize Network	View Network Nodes	NodeCount	%ID	#Spectra	AllIDs
			~	~	~	emodin

Status

[View Network Pairs | Networking Statistics]

Advanced Views - Networking Graphs/Histograms
[Edges, MZ Delta Histogram]

Advanced Views - External Visualization
[Direct Cytoscape Preview/Download | Direct Cytoscape IIN Collapsed Preview/Download | Global Comparison with ReDU PCA (Beta)]

Advanced Views - External Tools
[View Dereplicator Results]

Advanced Views - Experimental Views
[Analyze with MS2LDA | Enhance with MolNetEnhancer | Visualize with Qemistree | Network with Spec2vec]

Advanced Views - qiime2 Views
[View qiime2 Emperor Plots | View qiime2 Emperor Bi-Plots | Download qiime2 Emperor qzv | Download qiime2 features biom qza]

Advanced Views - Stats Views (Experimental)
[View All Column Plots | View Select Column Plots | Data Exploration with Interactive Plotting | API Data for Plotting]

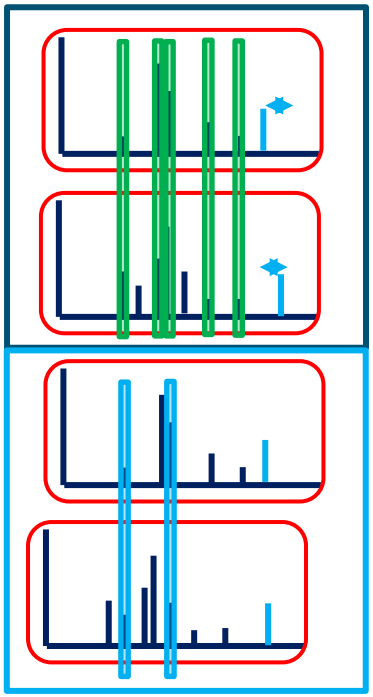
Advanced Views - Metadata Views
[View Metadata]

Limitations of Molecular Networking

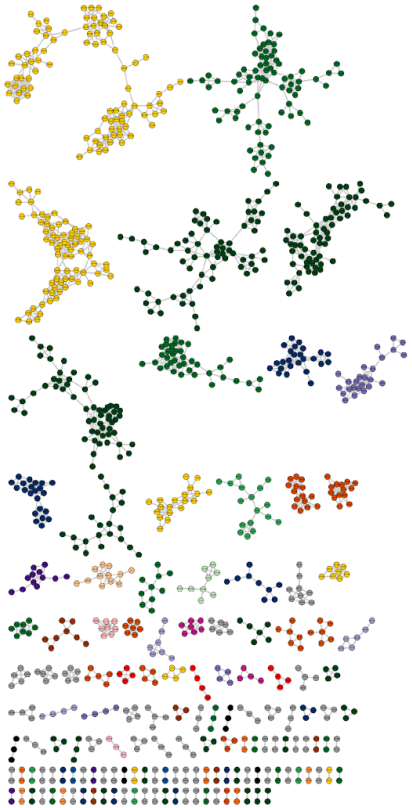
- No direct information on why spectra group into molecular families
- Each molecule can only go into one molecular family – even if it shares building blocks with two families
- Shape is (highly) dependent on parameter settings

Illuminating the Rhamnaceae chemistry

Molecular Networking



Wolfender et al., Anal. Chem., 2018



- triterpenoids
- triterpenoidal saponins
- steroidal saponins
- flavonoid 3-*O*-glycosides
- flavonoid 7-*O*-glycosides
- flavonoid *O*-glycosides
- flavones, flavonones, flavonols
- 8-methylated flavonoids
- (hydroxyl)anthraquinones
- xanthones
- hydrolyzable tannins
- phenolic glycosides
- coumaric acid and derivatives
- lignan glycosides
- iridoid glycosides
- peptide
- hybrid peptide
- oligosaccharides
- long-chain fatty acids
- others / no matches

MOLNET ENHANCER

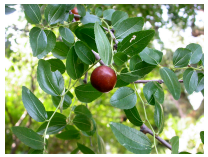
plant related classifications:

different flavonoids

phenolic glycosides

triterpenoids

Dr Kyo Bin Kang, UCSD



WAGENINGEN UNIVERSITY
WAGENINGEN UR

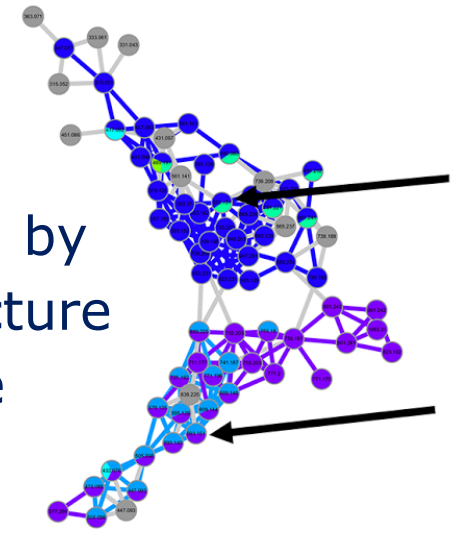
Kang, Ernst, van der Hooft et al., The Plant Journal, 2019
Ernst et al., Metabolites, 2019

Deeper insight into Rhamnaceae molecular families

Flavonoid-3-*O*-glycosides

Subfamilies!

Coloured by Substructure Presence



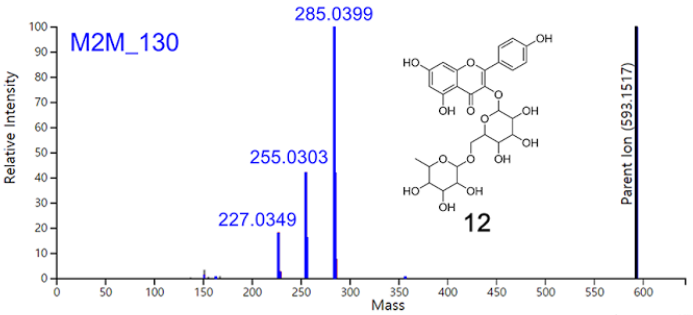
M2M_130 (kaempferol)

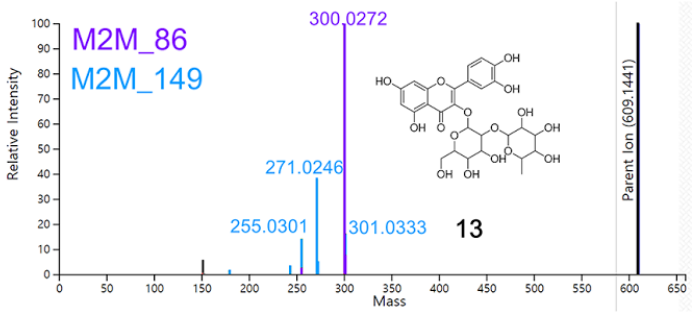
M2M_86

M2M_149

(quercetin and its fragments)

Kaempferol





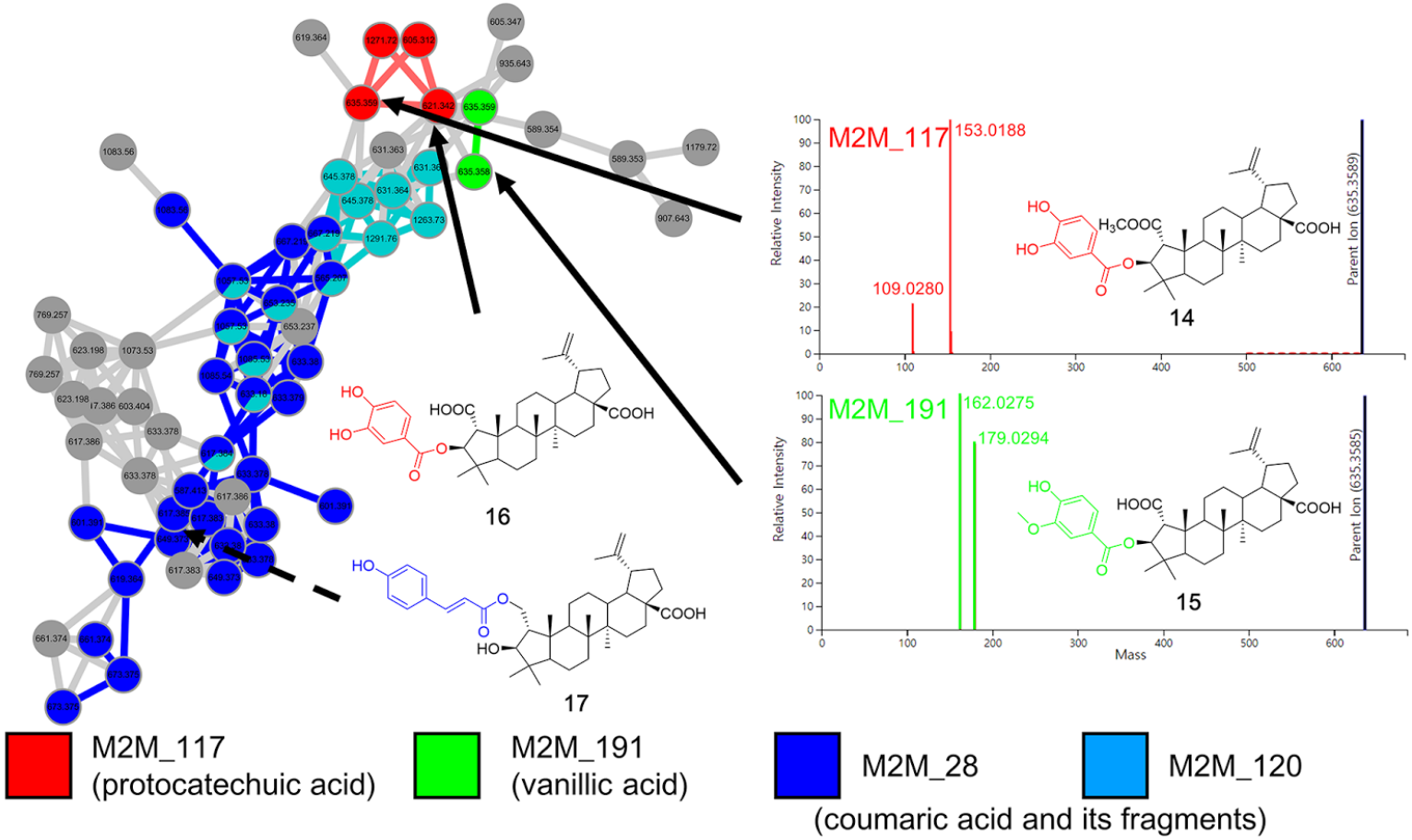
Quercetin

Kang, Ernst, van der Hooft et al., The Plant Journal, 2019

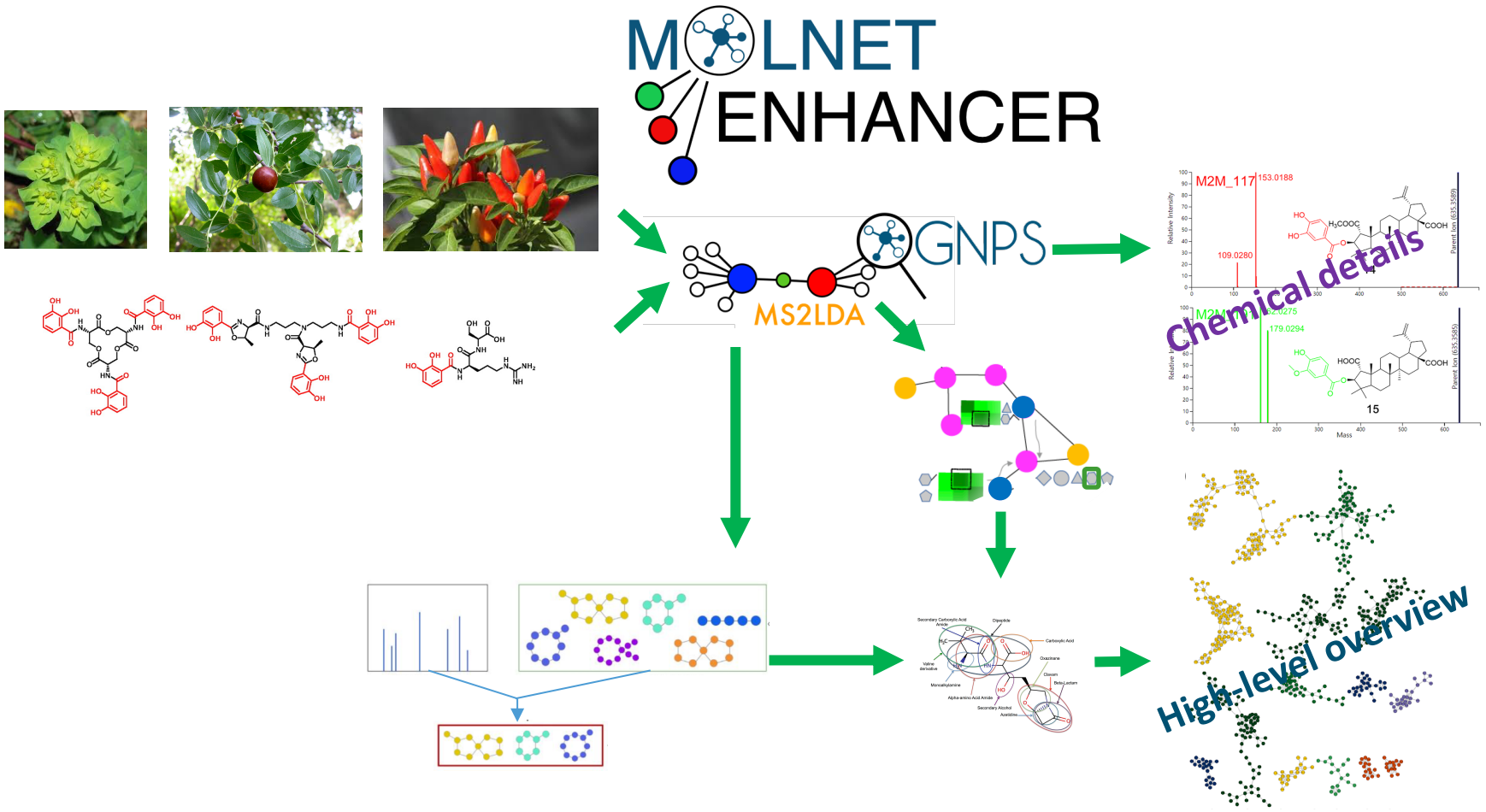
Triterpenoid family with benzoic acid conjugates

Triterpenoid Family: Differentiation of modifications

Protocatechuic acid and Vanillic acid based

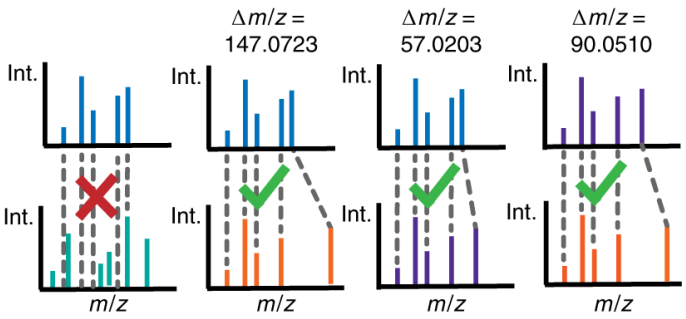


MolNetEnhancer Workflow Combining Outputs



Kang, Ernst, van der Hooft et al., The Plant Journal, 2019
Ernst et al., Metabolites, 2019

Spectral Similarity – Cosine Score

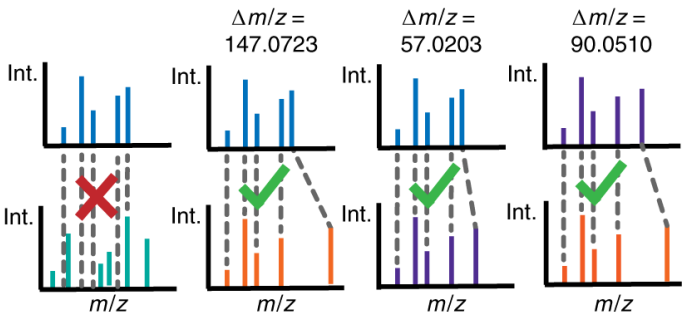


Aron et al., Nature Protocols, 2020

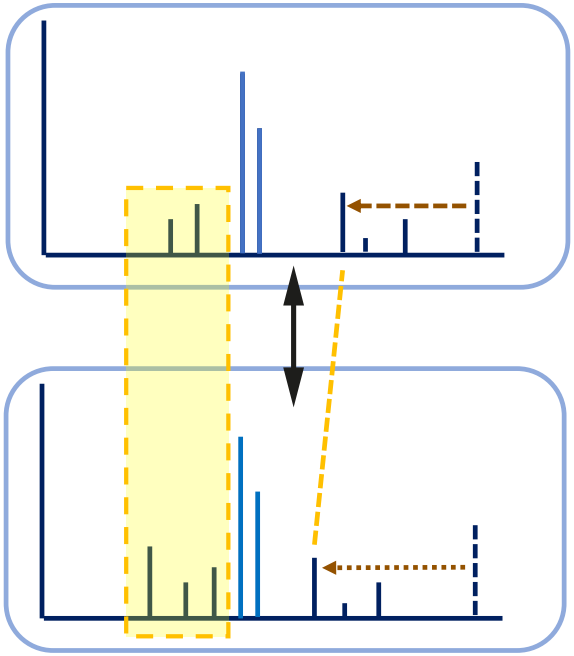
Cosine similarity-based scores:

+ do not need any training
+ work well for nearly identical spectra/molecules
+ work for one distinct modification
- often fail for multiple (subtle, local) modifications
- result in spurious hits at large-scale

Spectral Similarity – Cosine Score



Aron et al., Nature Protocols, 2020



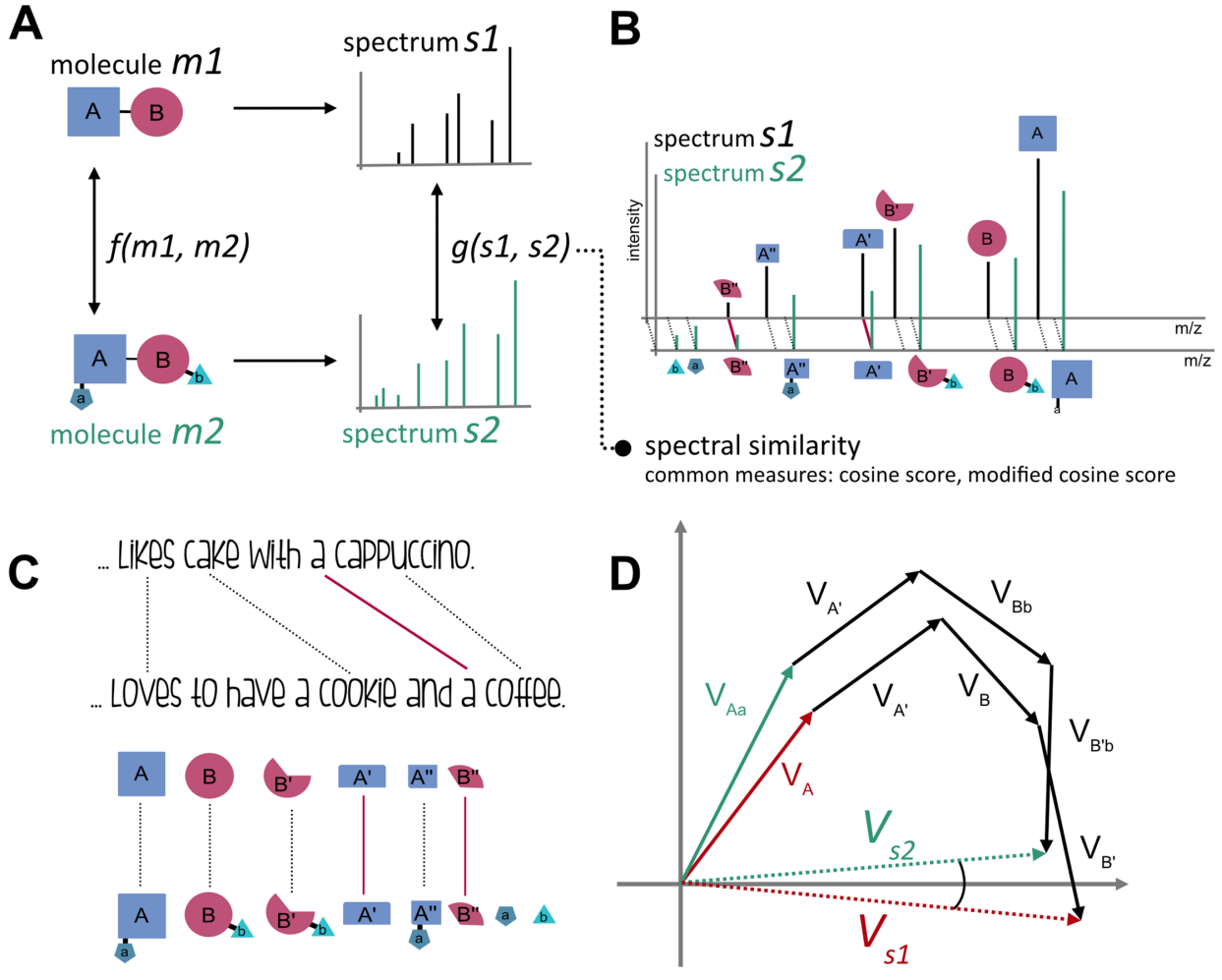
Ideal situation:

Similar fingerprints contribute to higher scores

Can we learn how mass fragments are related?

Spec2Vec:

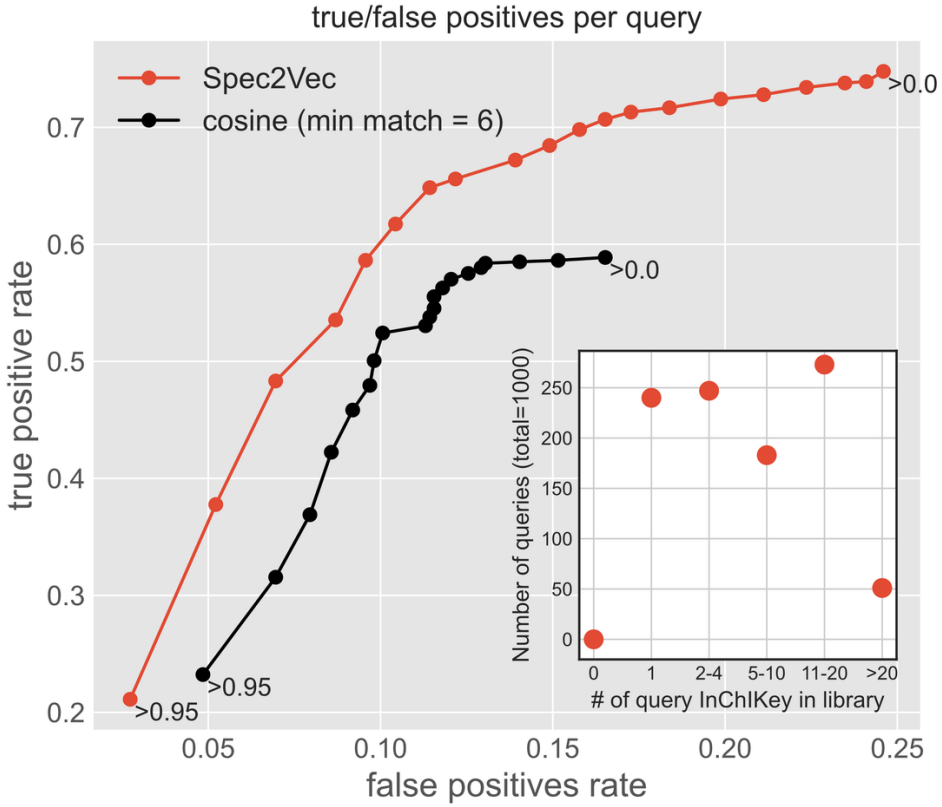
a novel alternative mass spectral similarity score

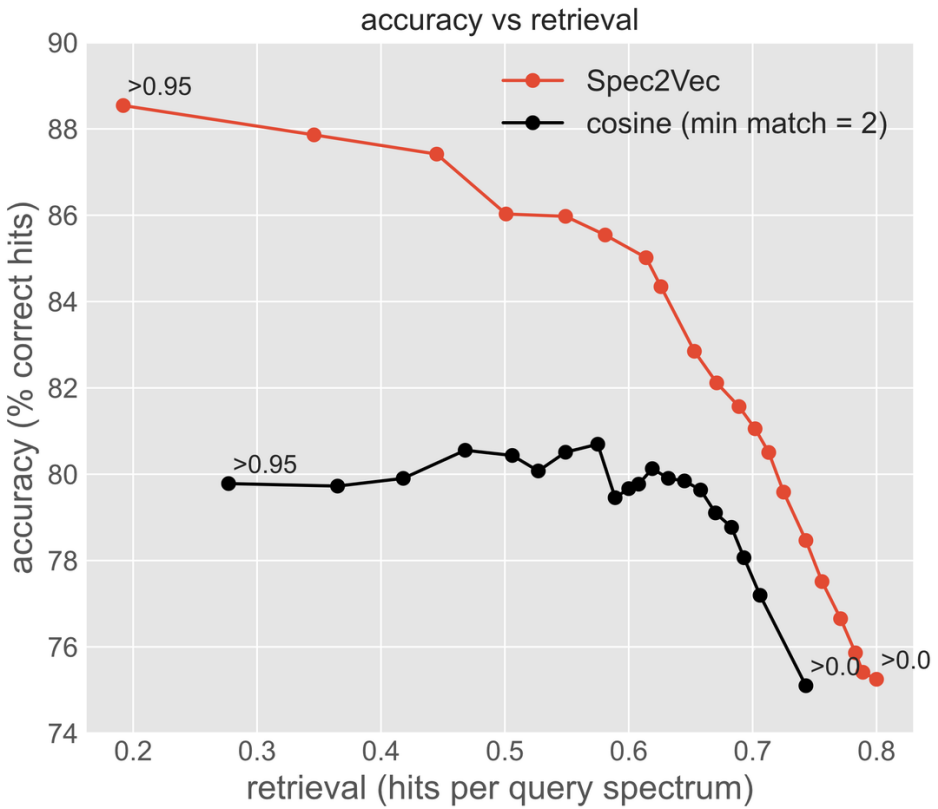


Huber et al., PLOS Computational Biology, 2021

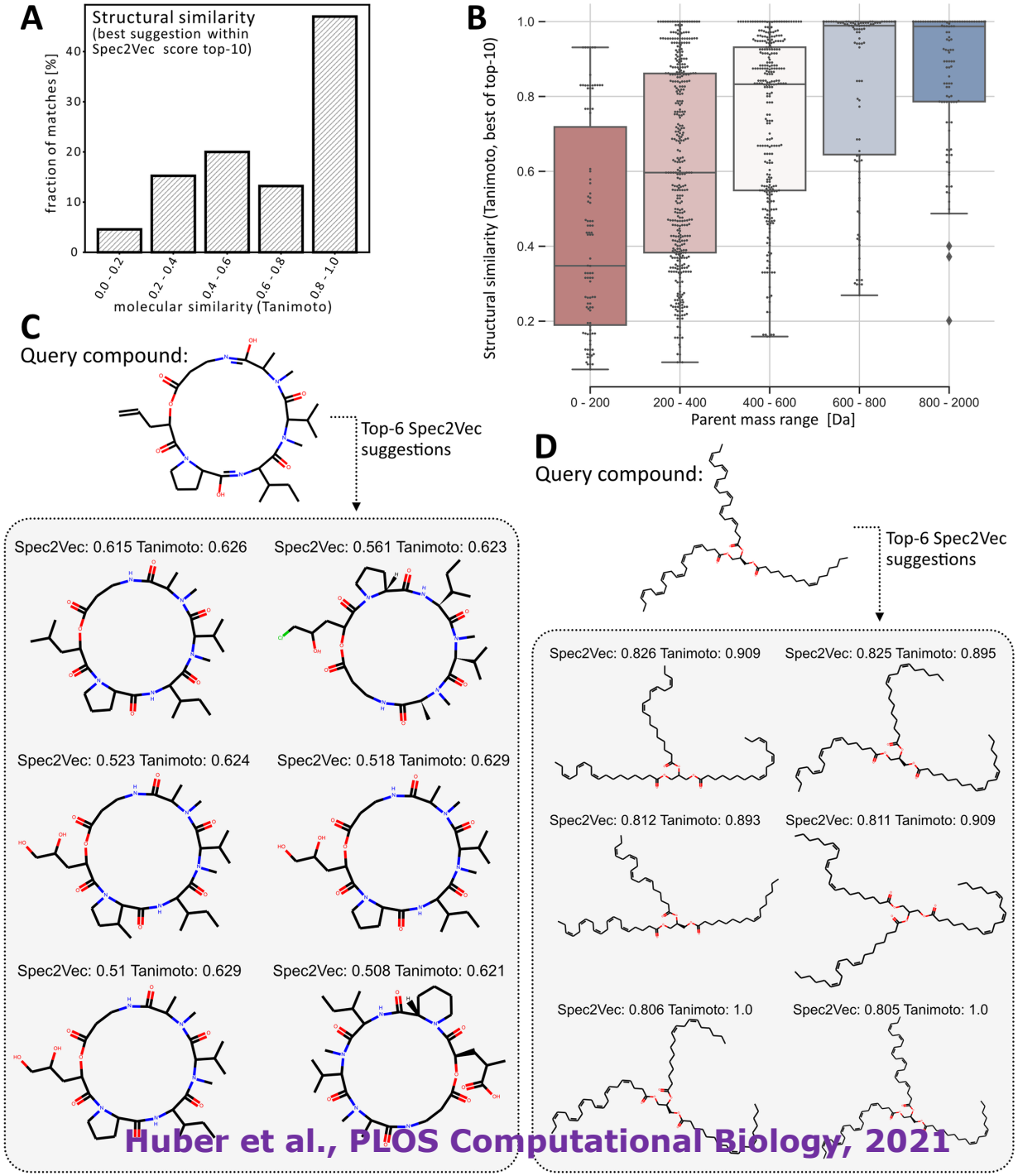
Spec2Vec:

a novel alternative mass spectral similarity score



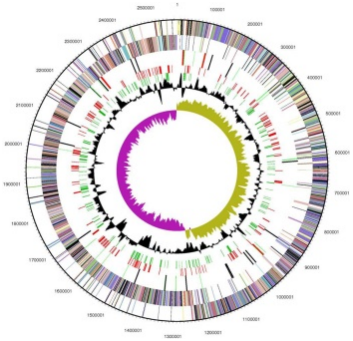


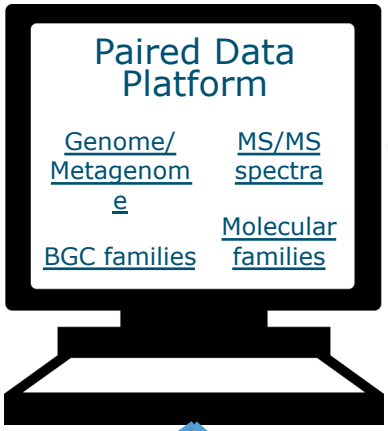
Huber et al., PLOS Computational Biology, 2021

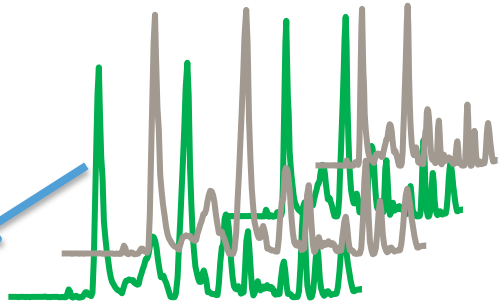


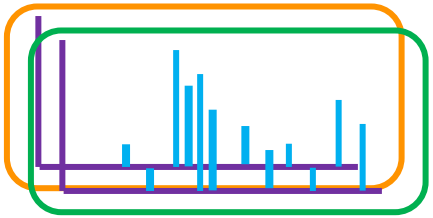
Huber et al., PLOS Computational Biology, 2021

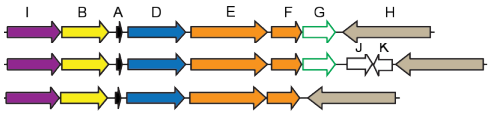
iOMEGA solutions at the data level: pairing omics data



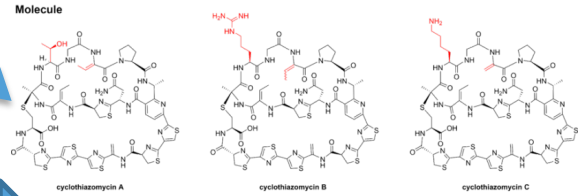








Gene clusters
Gene cluster families

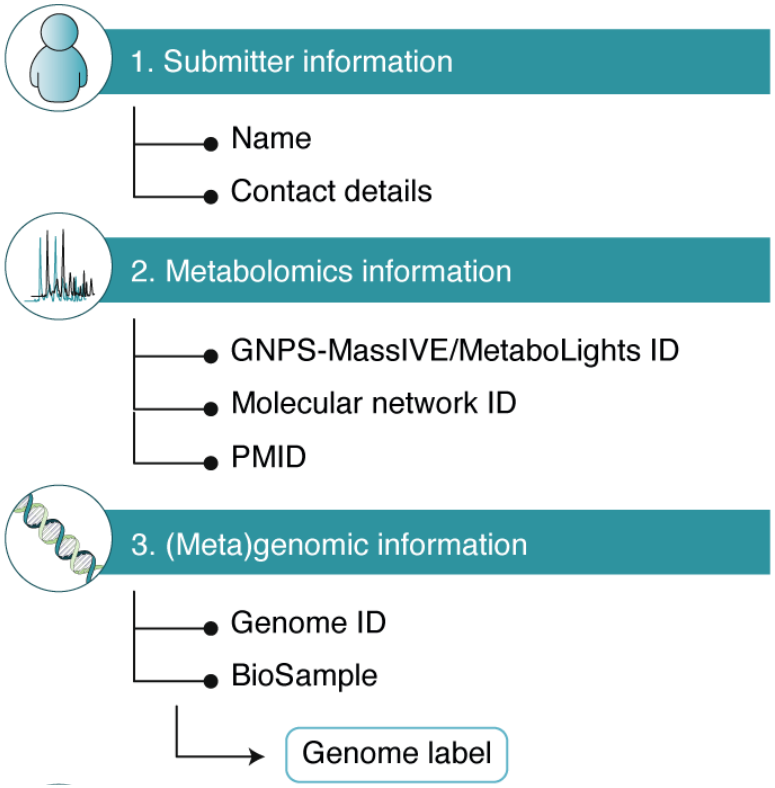


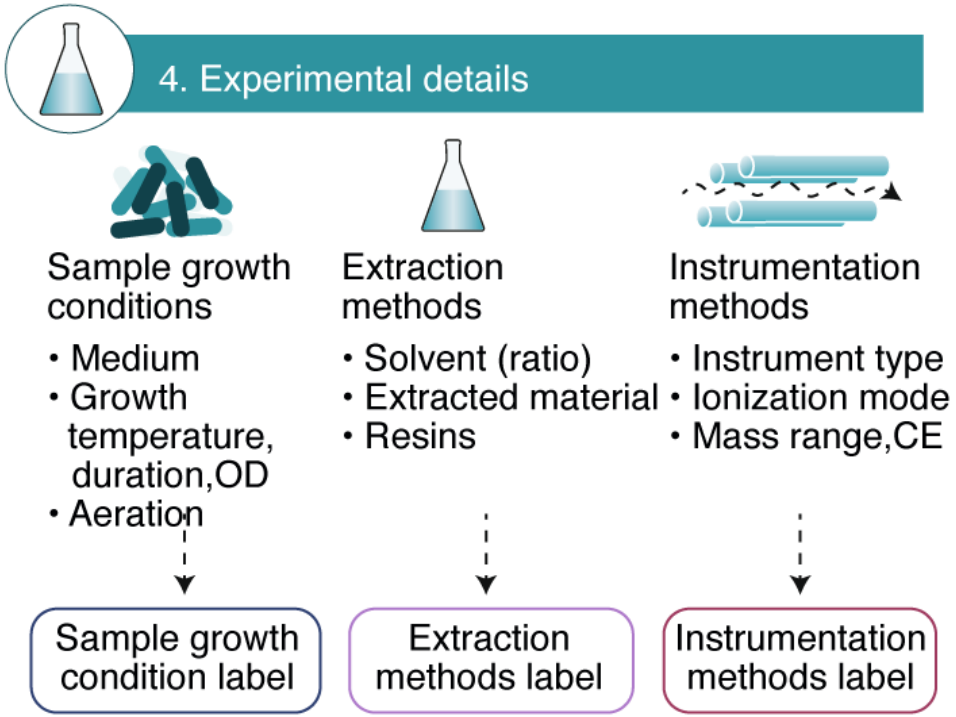
LINK!

Molecules
Molecular families

Schorn *et al.*, Nature Chemical Biology, 2021

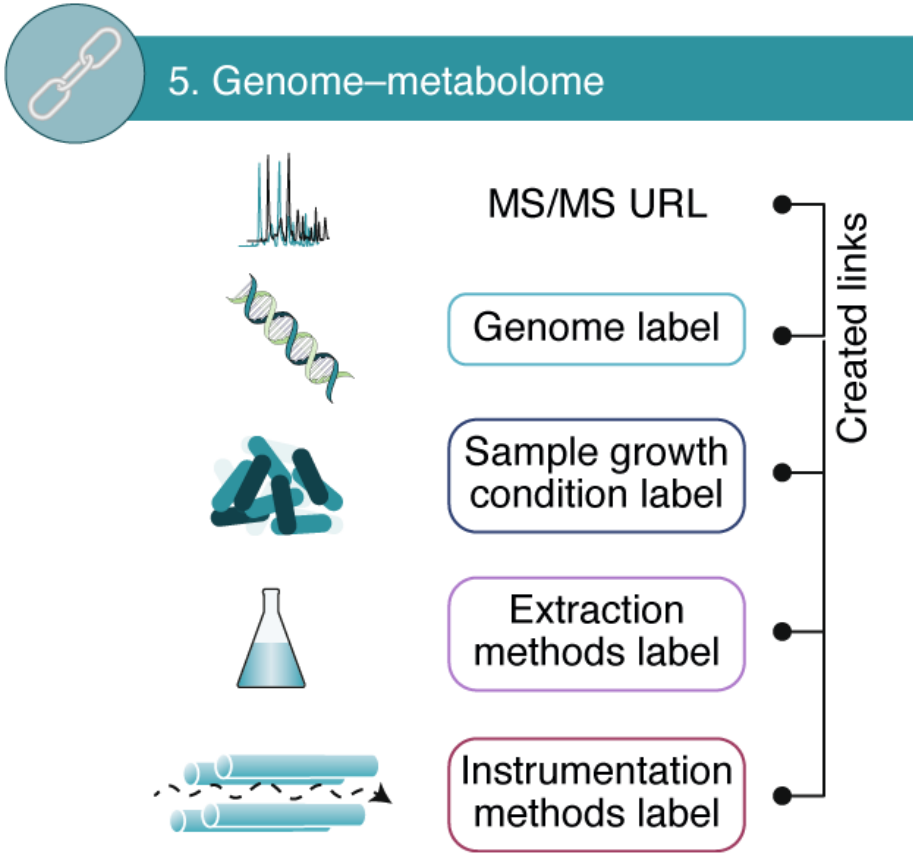
The Paired Omics Data Platform: recording minimal metadata

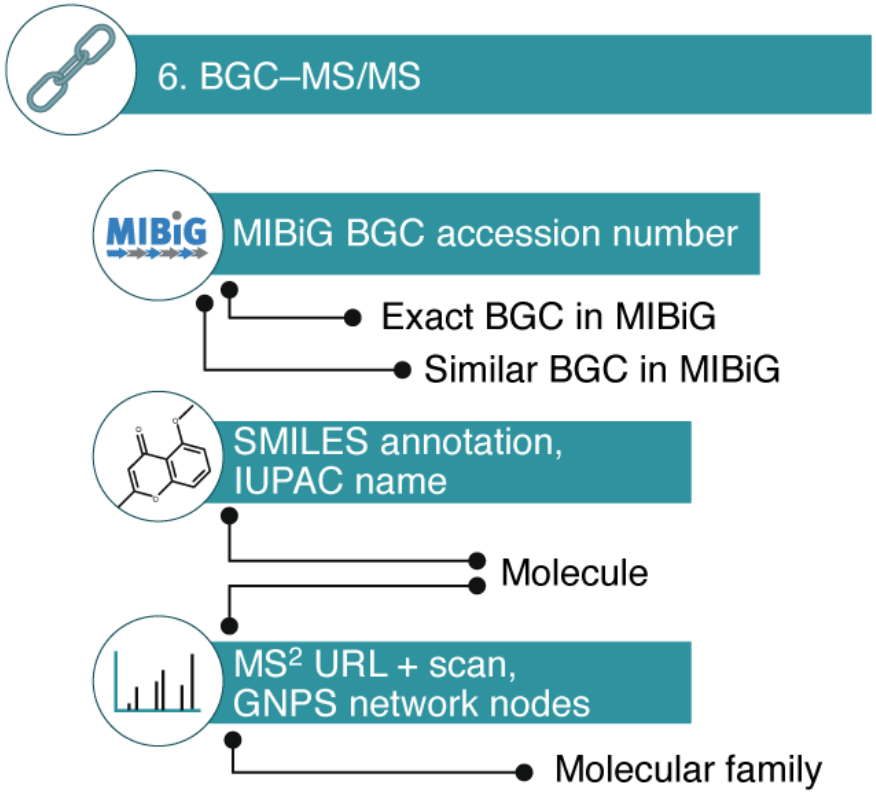




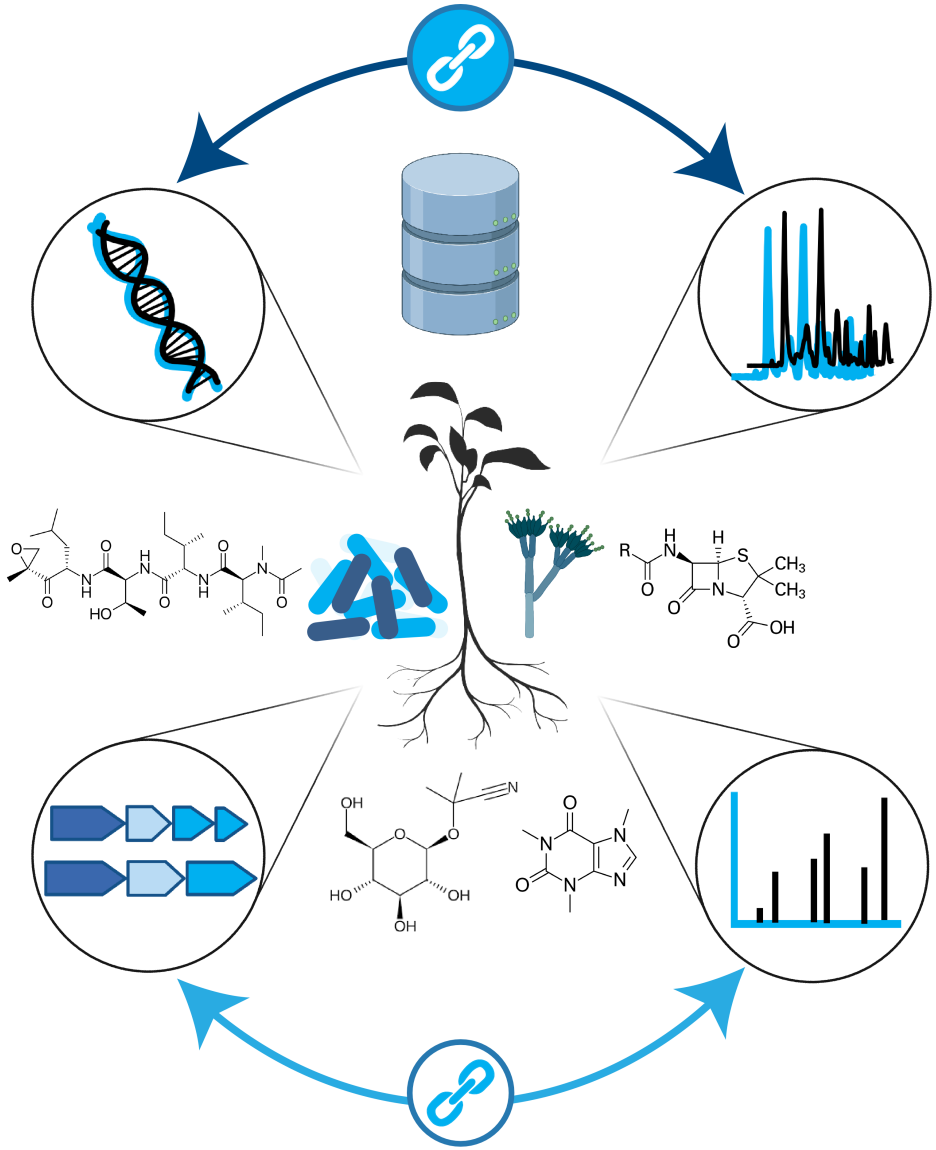
Schorn *et al.*, Nature Chemical Biology, 2021

The Paired Omics Data Platform: recording data links





Schorn *et al.*, Nature Chemical Biology, 2021



Big Shoutout to:

Public genome repositories

Public metabolome repositories

https://pairedomicsdata.bioinformatics.nl

Schorn *et al.*, Nature Chemical Biology, 2021

Look back at Workshop objectives

Being able to:

- Explain rationale behind metabolome mining tools
- Explore and assess GNPS Library Matches
- Explore and assess GNPS Molecular Families

Have:

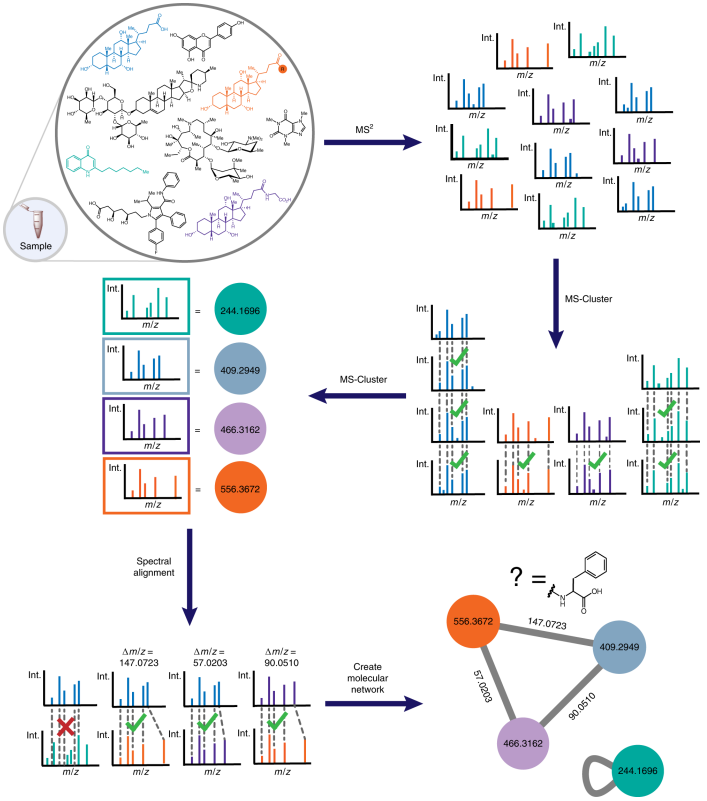
- Fun

Further Reading

- GNPS Documentation:
  - https://ccms-ucsd.github.io/GNPSDocumentation/

GNPS 50

Welcome to GNPS Documentation



- Nature Protocols Paper:
  - https://www.nature.com/articles/s41596-020-0317-5

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Mining the Plant Specialized Metabolome with Mass Spectrometry:

Library Matching and Molecular Networking with GNPS

THANKS FOR YOUR PARTICIPATION! ☺

WAGENINGEN UNIVERSITY
WAGENINGEN UR

100years

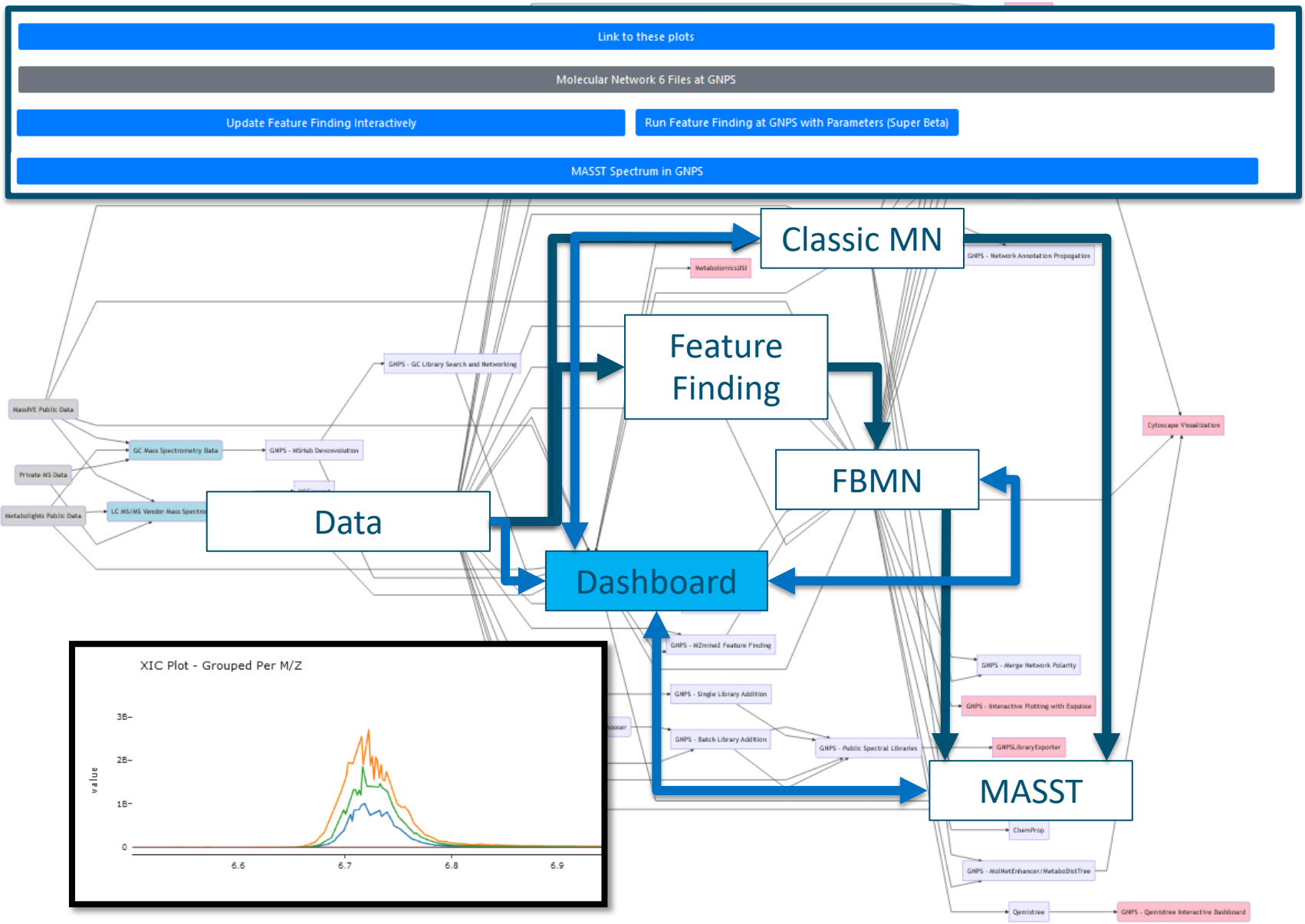
Further Reading (2) – MS2LDA

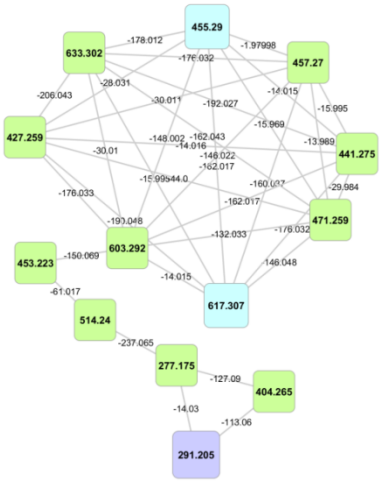
Van der Hooft et al., *PNAS* 2016, 113 (48), 13738-13743
Van der Hooft et al., Anal. Chem. 2017,
Rogers et al., Faraday Discussions 2019,
Ernst et al., *Metabolites* 2019, *9*(7), 144

Tutorials to get familiar with individual tools from which the output is combined with MolNetEnhancer can be found here:
GNPS molecular networking:
https://ccms-ucsd.github.io/GNPSDocumentation/networking
MS2LDA:
https://ccms-ucsd.github.io/GNPSDocumentation/ms2lda/
http://ms2lda.org/user_guide
MolNetEnhancer workflow tutorials in both R and Python can be found here:
https://github.com/madeleineernst/pyMolNetEnhancer
https://github.com/madeleineernst/RMolNetEnhancer

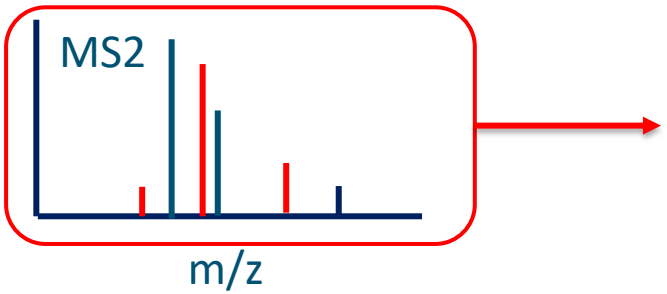

WAGENINGEN UNIVERSITY
WAGENINGEN UR

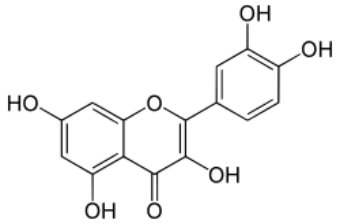
GNPS Workflows





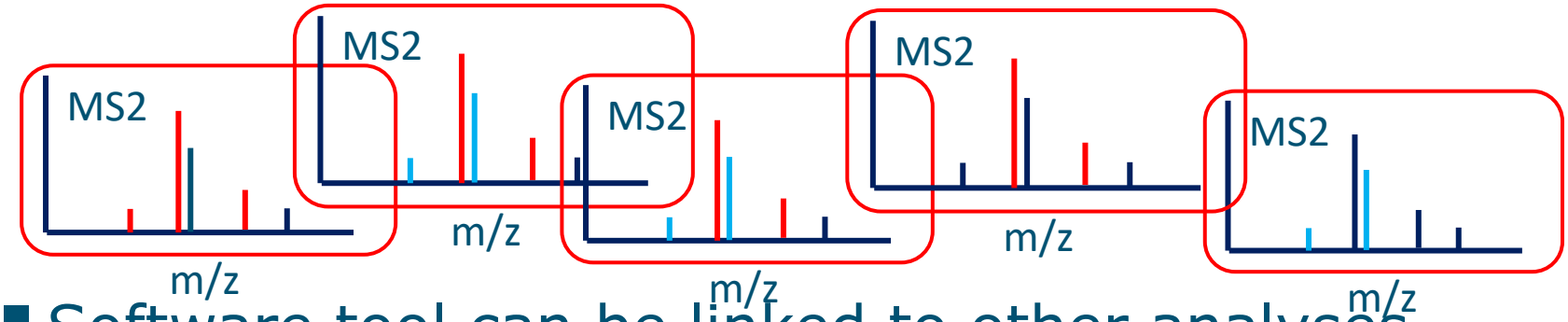
Motivation for MS2LDA





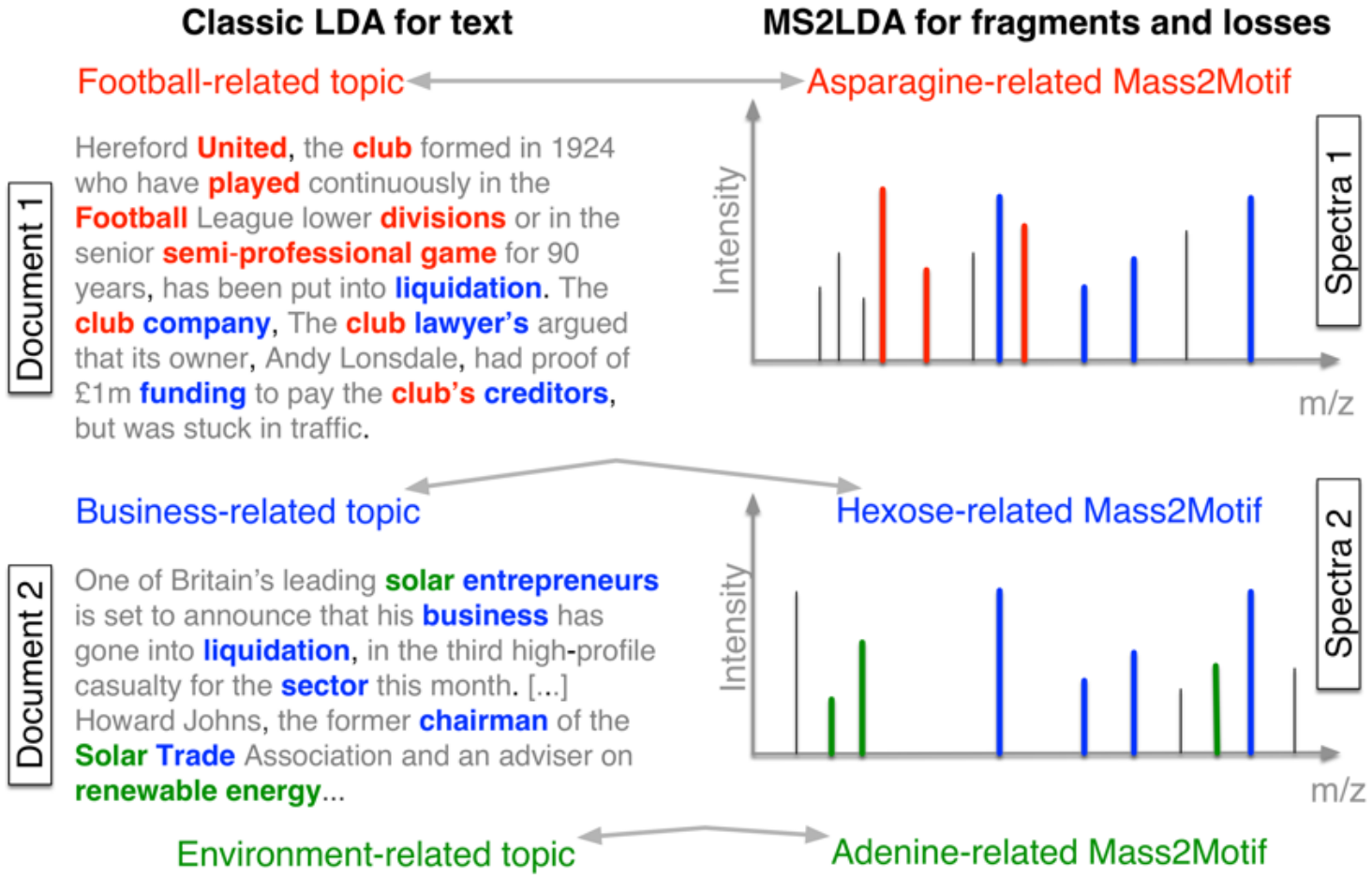
Quercetin aglycone

- Manual interpretation of fragmentation patterns guide in annotation and chemical classification
- Automated recognition of mass fragmental motifs speeds up analysis and enables it at large scale



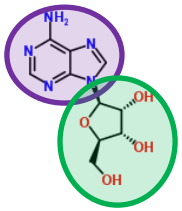
- Software tool can be linked to other analyses

Topic modelling: from text to molecules



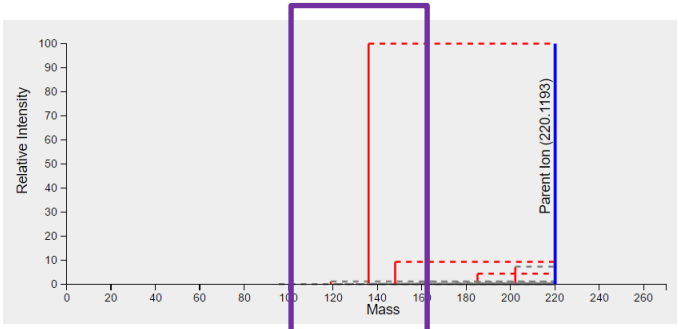
Documents <-> molecules

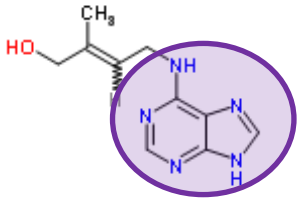
Words <-> fragments and neutral losses

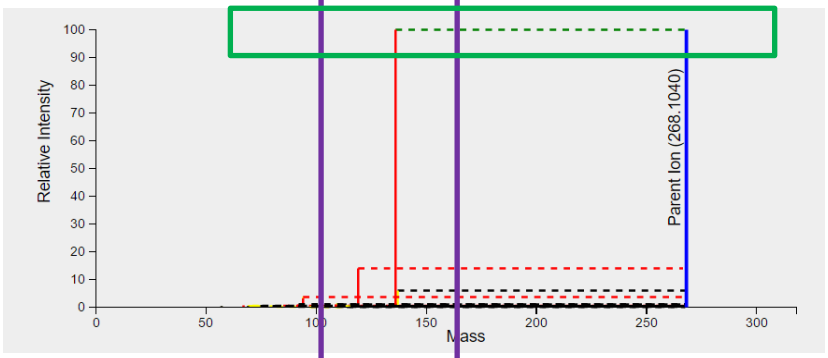


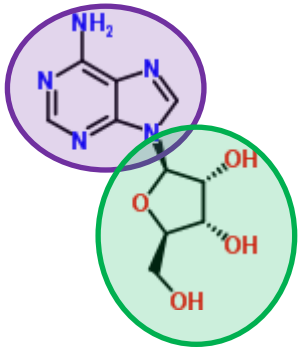
van der Hooft et al., PNAS, 2016

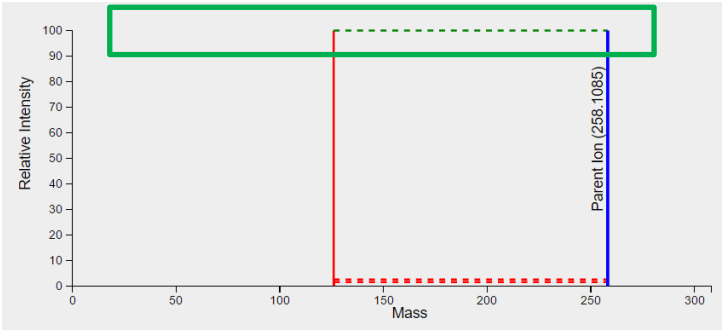
Does it work?

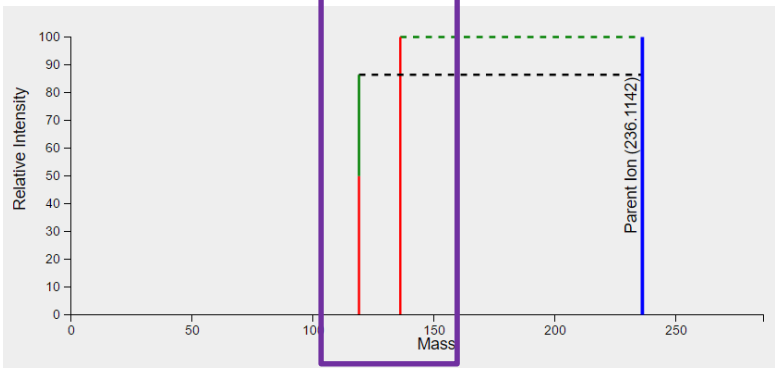

Relative Intensity
Mass
Parent Ion (220.1193)

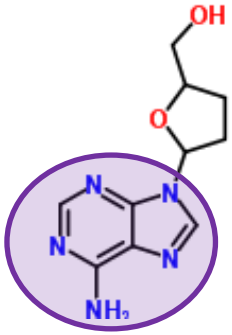

CH3
HO
NH
N
N
N
N
H

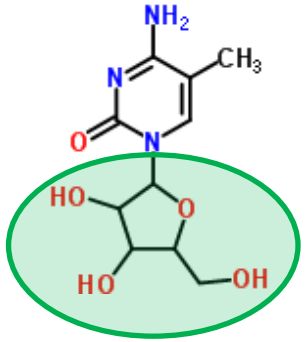

Relative Intensity
Mass
Parent Ion (268.1040)


NH2
N
N
N
N
OH
O
OH
OH


Relative Intensity
Mass
Parent Ion (258.1085)

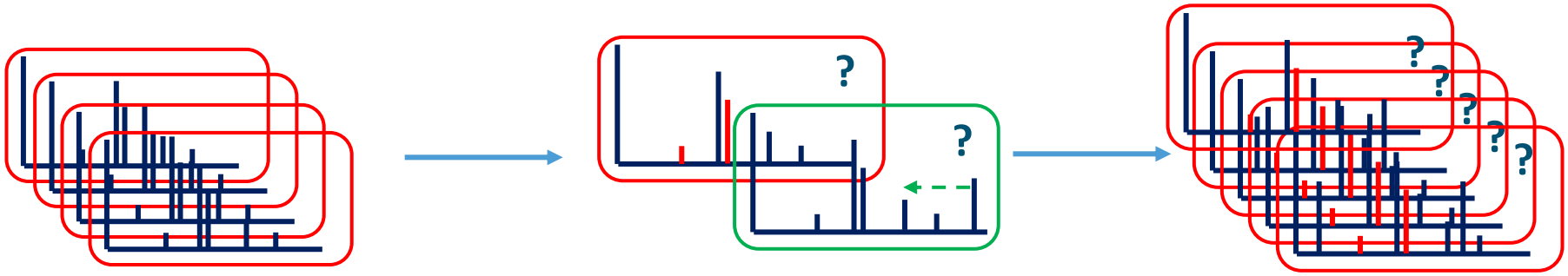

Relative Intensity
Mass
Parent Ion (236.1142)


OH
O
N
N
N
N
NH2

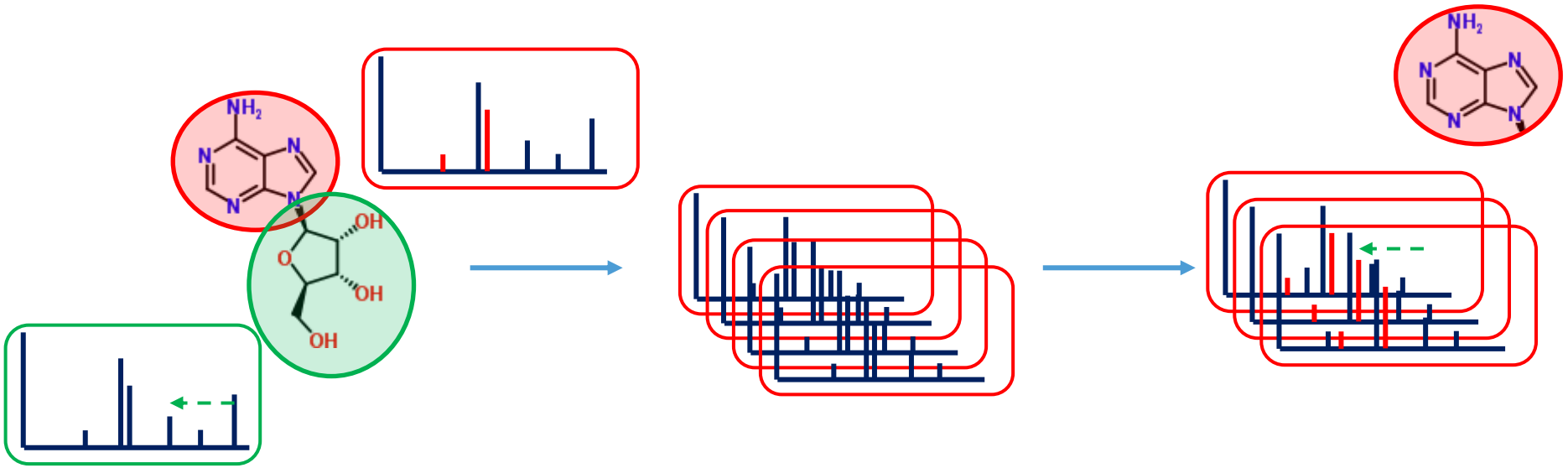

NH2
CH3
N
O
N
HO
O
HO
OH

LDA and Decomposition

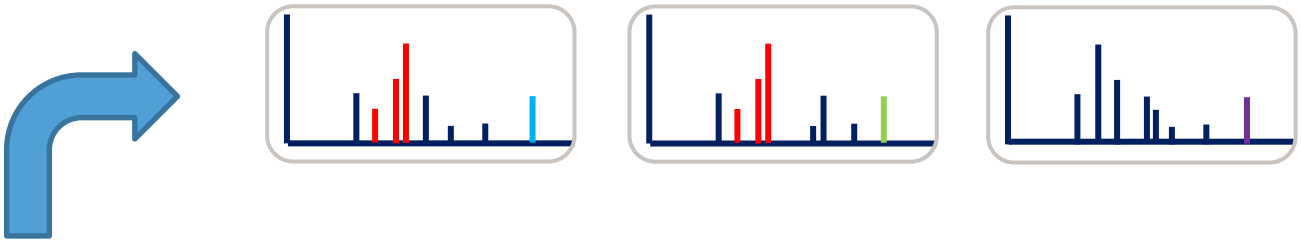
- Unsupervised discovery <--> LDA



- Predefined motif search <--> Decomposition



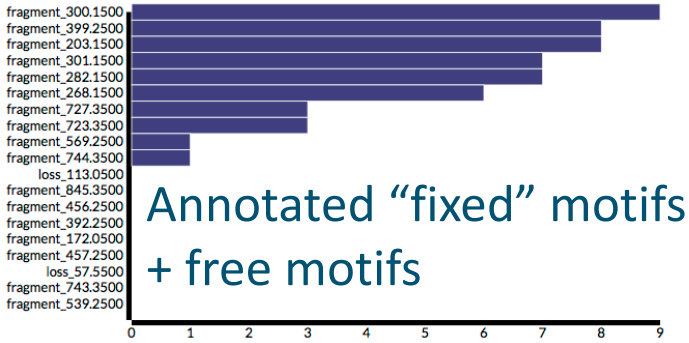
MotifDB – annotated motifset database



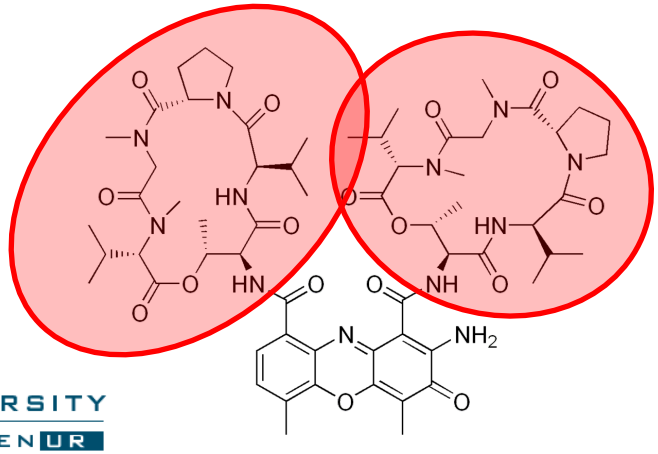
MotifDB

Annotated motif sets

StrepSalini_motif_10.m2m	Actinomycin related Mass2Motif (H-VPMeGMeV-OH peptide lactone sequence)
StrepSalini_motif_37.m2m	C11H24NO and C11H22N fragments - 186/168 - related Mass2Motif
StrepSalini_motif_233.m2m	C9H20N and C6H14N fragments (142/100) related Mass2Motif - Streptomyces sp related
StrepSalini_motif_221.m2m	Generic motif - generic fragment [70.05]
StrepSalini_motif_85.m2m	Generic motif - generic fragment [72.05]
StrepSalini_motif_108.m2m	Hydroxy-staurosporine related Mass2Motif

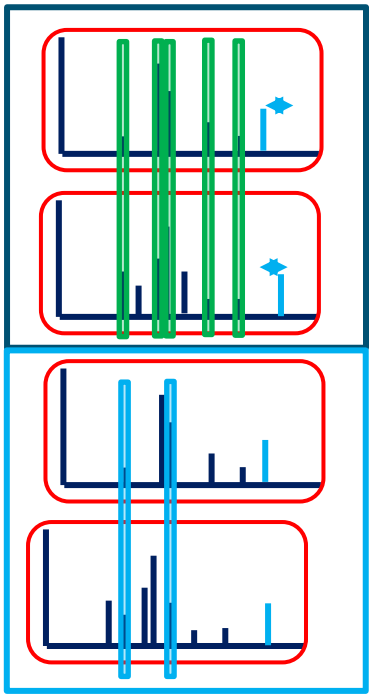


Annotated “fixed” motifs
+ free motifs

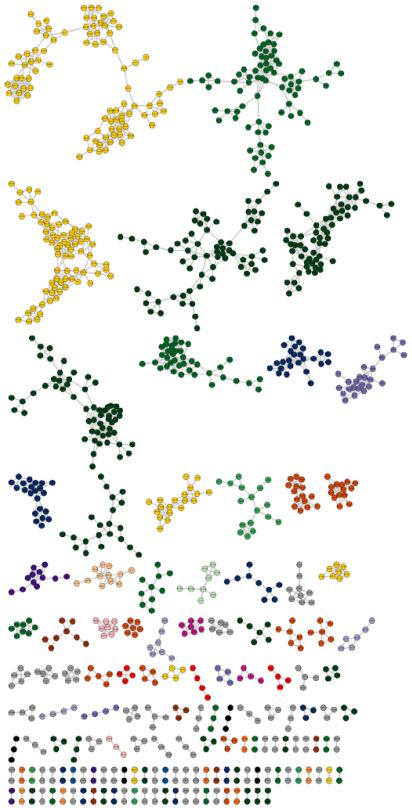


Illuminating the Rhamnaceae chemistry

Molecular Networking



Wolfender et al., Anal. Chem., 2018



- triterpenoids
- triterpenoidal saponins
- steroidal saponins
- flavonoid 3-*O*-glycosides
- flavonoid 7-*O*-glycosides
- flavonoid *O*-glycosides
- flavones, flavonones, flavonols
- 8-methylated flavonoids
- (hydroxyl)anthraquinones
- xanthones
- hydrolyzable tannins
- phenolic glycosides
- coumaric acid and derivatives
- lignan glycosides
- iridoid glycosides
- peptide
- hybrid peptide
- oligosaccharides
- long-chain fatty acids
- others / no matches

MOLNET ENHANCER

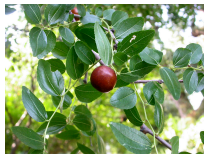
plant related classifications:

different flavonoids

phenolic glycosides

triterpenoids

Dr Kyo Bin Kang, UCSD



WAGENINGEN UNIVERSITY
WAGENINGEN UR

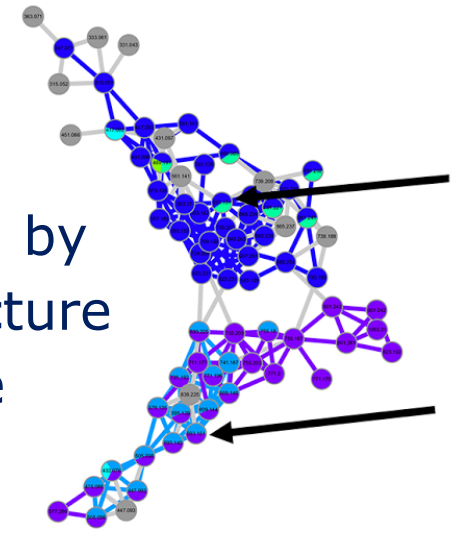
Kang, Ernst, van der Hooft et al., The Plant Journal, 2019
Ernst et al., Metabolites, 2019

Deeper insight into Rhamnaceae molecular families

Flavonoid-3-*O*-glycosides

Subfamilies!

Coloured by Substructure Presence



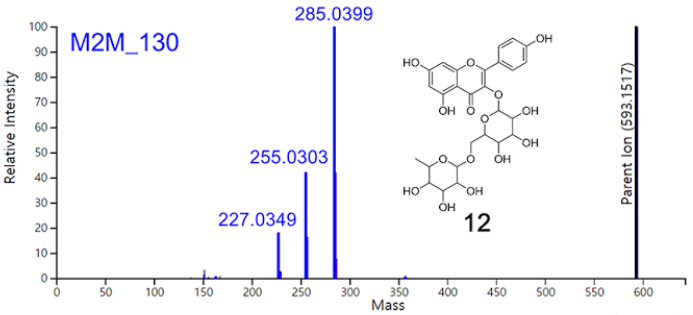
M2M_130 (kaempferol)

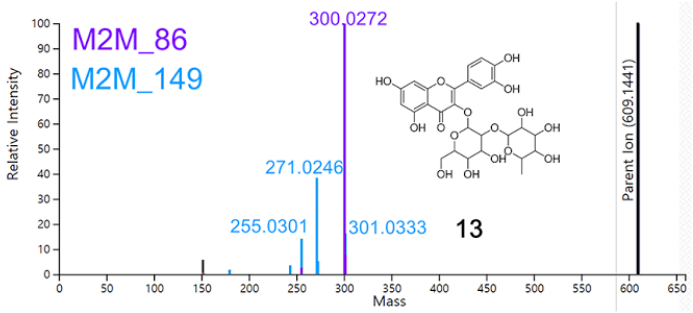
M2M_86

M2M_149

(quercetin and its fragments)

Kaempferol





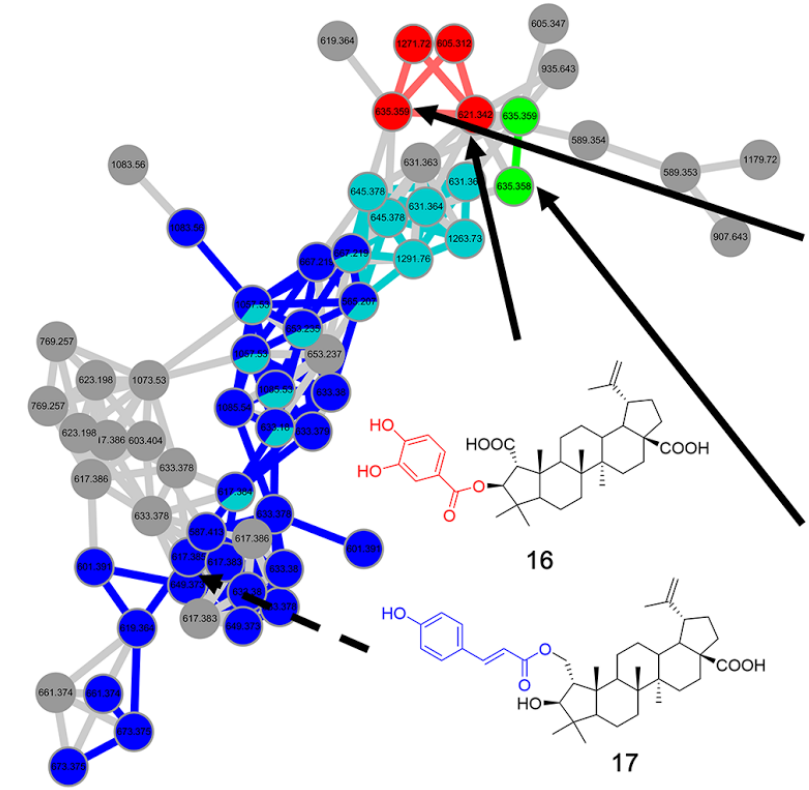
Quercetin

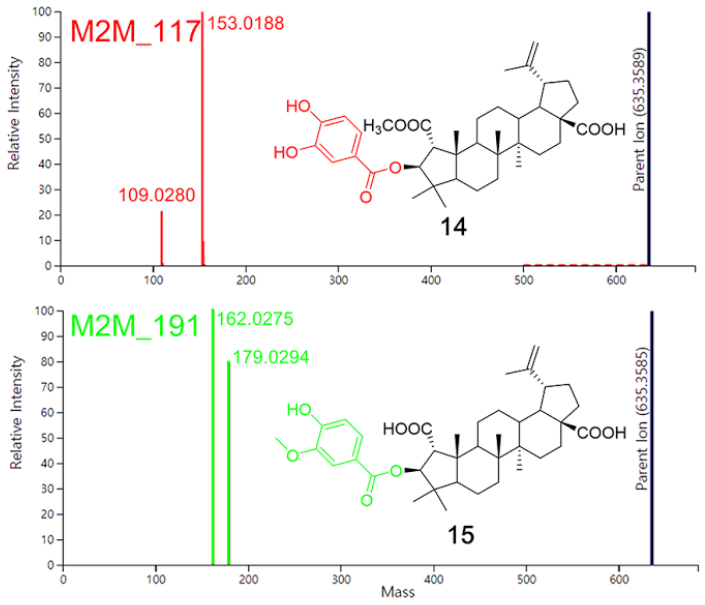
Kang, Ernst, van der Hooft et al., The Plant Journal, 2019

Triterpenoid family with benzoic acid conjugates

Triterpenoid Family: Differentiation of modifications

Protocatechuic acid and Vanillic acid based





M2M_117 (protocatechuic acid)

M2M_191 (vanillic acid)

M2M_28

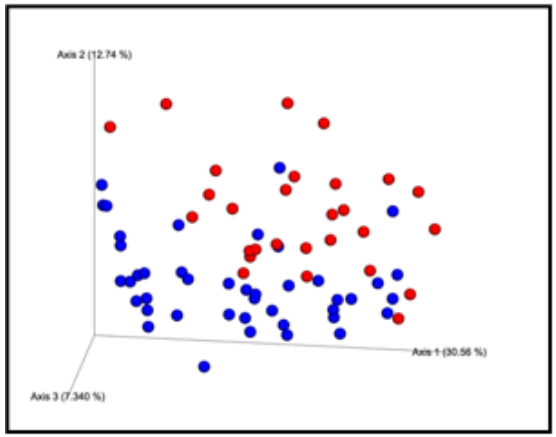
M2M_120

(coumaric acid and its fragments)

Improved clade separation by chemically informed similarity measures

Ziziphoid **Rhamnoid**

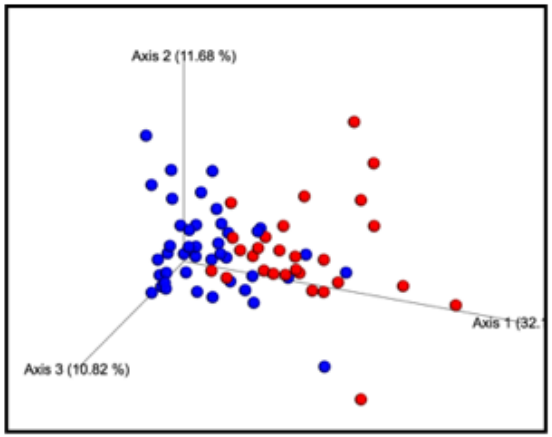
BRAY-CURTIS



adonis R^2 = 0.09 p = 0.001

<

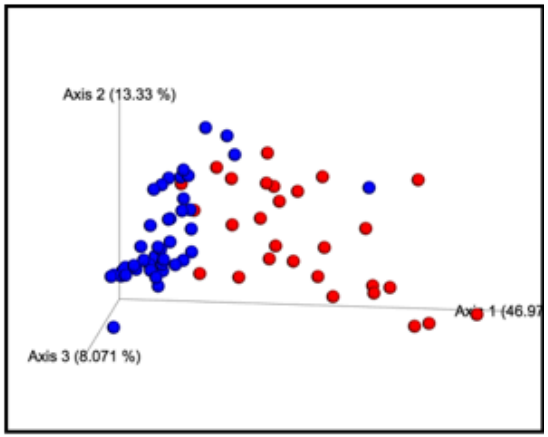
MOTIFTREE UNIFRAC



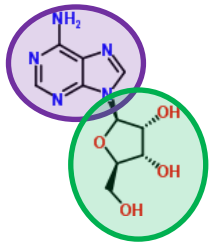
adonis R^2 = 0.19 p = 0.001

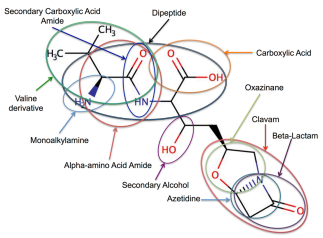
<

CLASSYTREE UNIFRAC



adonis R^2 = 0.30 p = 0.001

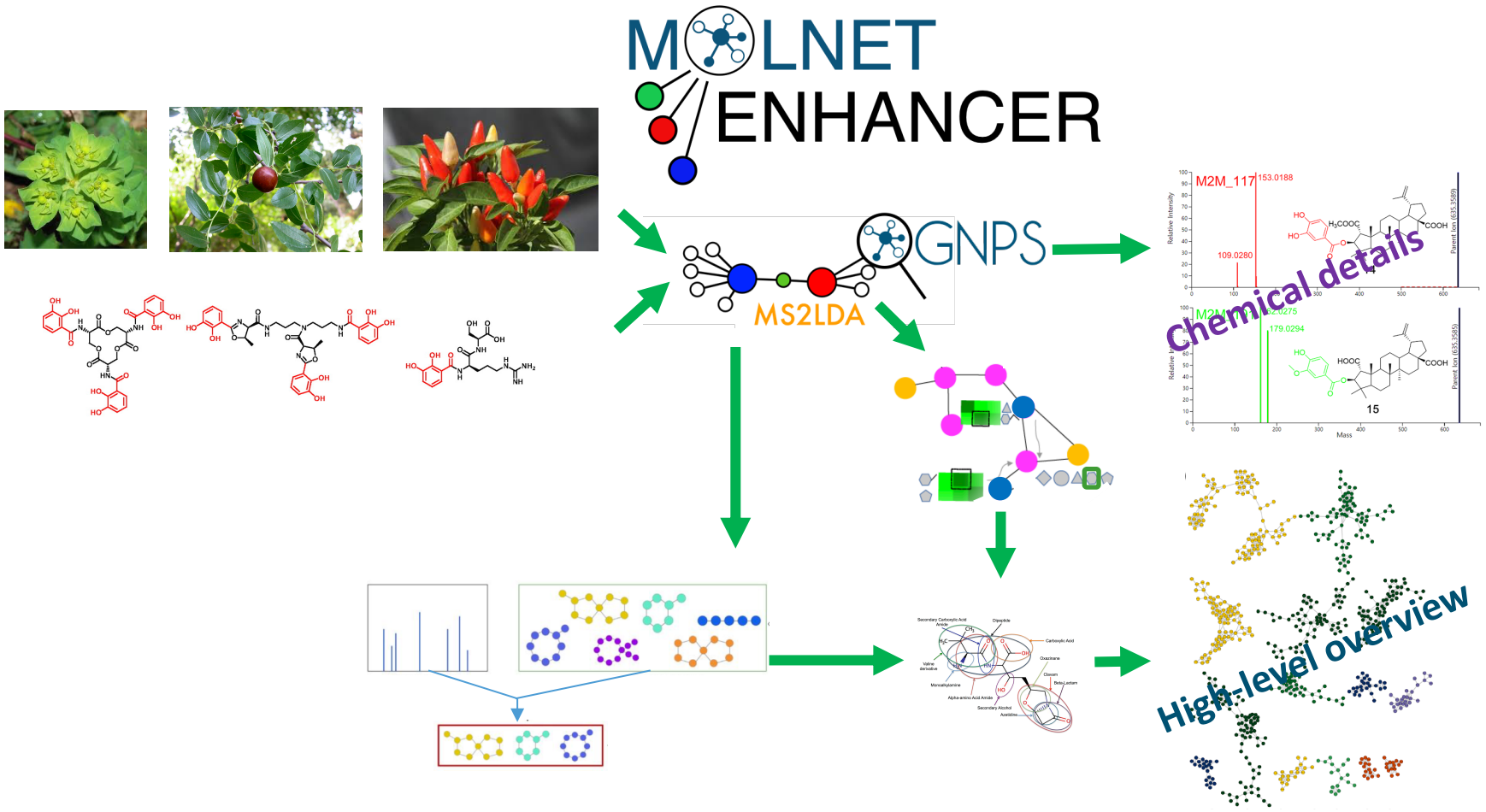




WAGENINGEN UNIVERSITY
WAGENINGEN UR

Inspired by: Tripathi et al., Nature Chemical Biology, 2020

MolNetEnhancer Workflow
Combining Outputs



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Kang, Ernst, van der Hooft et al., The Plant Journal, 2019
Ernst et al., Metabolites, 2019