

Linking biological data using data science and cross-disciplinary software development

Florian Huber¹, Justin van der Hooft², Simon Rogers³, Marnix Medema², Lars Ridder¹

¹ Netherlands eScience Center

² Bioinformatics Group, Wageningen University

³ School of Computing Science, University of Glasgow



Breaking down scientific monocultures by cross-disciplinary software development

Florian Huber¹, Justin van der Hooft², Simon Rogers³, Marnix Medema², Lars Ridder¹

¹ Netherlands eScience Center

² Bioinformatics Group, Wageningen University

³ School of Computing Science, University of Glasgow

Benefits of RSE groups

For RSEs



- ✓ Stable careers
- ✓ Peer group
- ✓ Recognition & development

For research projects



- ✓ Flexible access to expertise
- ✓ Sharing between projects
- ✓ Access to niche skills

For researchers



- ✓ Help & advice
- ✓ Training
- ✓ Infrastructure
- ✓ Focus for wider network

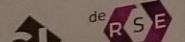
34

talk by Alys Brett



Microsoft
aws

R
consortium
GI



netherlands eScience center

We signal challenges and opportunities at the intersection of software and academic research



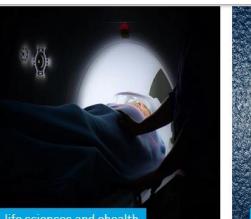
Photography: Elodie Burrillon



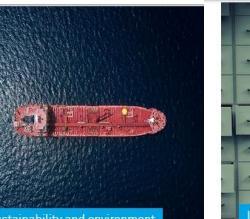
sustainability and environment



life sciences and eHealth



life sciences and eHealth



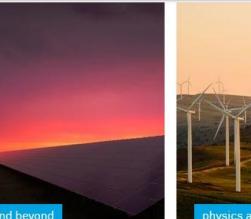
sustainability and environment



life sciences and eHealth



physics and beyond



physics and beyond



physics and beyond



humanities and social sciences



humanities and social sciences

IS-ENES3

Providing the infrastructure to better understand and project climate variability and change

TADPOLE-SHARE

SHaring TADPOLE's Algorithms for Reuse and Evaluation

DTL Semantic Analysis of radiology Reports utilizing Lexicon

Unlocking large volumes of knowledge locked in natural text

Digital twins: monitoring ships' state in real-time

Advanced data science to assist the design of cleaner, safer and smarter ships

FAIR is as FAIR does

Integrating data publishing principles in scientific workflows

Scalable high-fidelity simulations of reacting multiphase flows at transcritical pressure

Accurate and Efficient Computation of the Optical Properties of Nanostructures for Improved Photovoltaic

Computation of the Optical Properties of nano structures

Accurate and Efficient Computation of the Optical Properties of Nanostructures for Improved Photovoltaic

Parallel-in-time methods for the propagation of uncertainties in wind farm simulations

Studying uncertainties in large eddy simulations of wind farms

TICCLAT

Text-Induced Corpus Correction and Lexical Assessment Tool

NEWGAC

Advancing Media History by Transparent Automatic Genre Classification



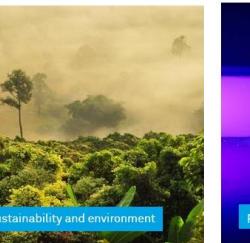
sustainability and environment



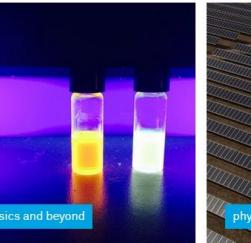
humanities and social sciences



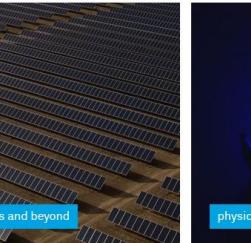
humanities and social sciences



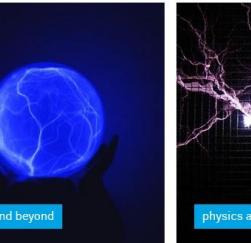
sustainability and environment



physics and beyond



physics and beyond



physics and beyond



physics and beyond



sustainability and environment



humanities and social sciences

EUropean Climate Prediction system

Early Prediction of Dyslexia in Infants Using Machine learning

Understanding visually grounded spoken language via multi-tasking

An alternative approach for intelligent systems to understand human speech

Monitoring tropical forest recovery capacity using RADAR Sentinel satellite data

Demonstrating the potential of European Sentinel satellite data

eScience Technology to Boost Quantum Dot Energy Conversion

More efficient lighting and solar energy conversion devices

A light in the dark

Quantum Monte Carlo meets solar energy conversion

Passing XSAMS

New tools for researchers in plasma, combustion and chemical reactor science

A phase field model to guide the development and design of next generation solid-state-batteries

Safer batteries with higher energy densities

Data mining tools for abrupt climate change

Updating our knowledge on abrupt climate change

Automated Analysis of Online Behaviour on Social Media

Gaining insights in the use of Twitter by politicians and journalists



sustainability and environment



eScience methodology



eScience methodology



humanities and social sciences



life sciences and eHealth



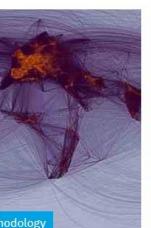
sustainability and environment



sustainability and environment



eScience methodology



eScience methodology



eScience methodology

MOSAIC

Modelling Sea level And Inundation for Cyclones

PROCESS

PROviding Computing solutions for ExaScale ChallengeS

Enhance Your Research Alliance (EYRA) Benchmark Platform

Uncovering Networks of Corporate Control

An interactive web-based platform to investigate the dynamics of global corporate networks

Integrated omics analysis for small molecule-mediated host-microbiome interactions

Advancing our understanding of molecular mechanisms of health and disease

MULTIXMAS

Multiscale simulations of excitation dynamics in molecular materials for sustainable energy applications

Stochastic Multiscale Climate Models

Coupling an implicit low-resolution model to an explicit high-resolution ocean model

MAGIC

Metrics and Access to Global Indices for Climate Projections

IMPACT

Software Analytics for the monitoring and assessment of the global impact of eScience Software on eStep

High spatial resolution phenological modelling at continental scales

Understanding phenological variability



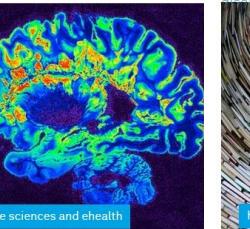
sustainability and environment



humanities and social sciences



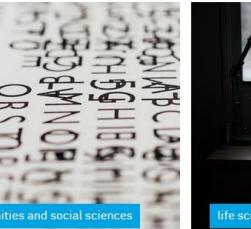
eScience methodology



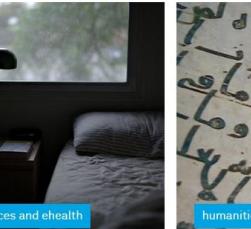
life sciences and eHealth



humanities and social sciences



humanities and social sciences



life sciences and eHealth



humanities and social sciences



sustainability and environment



humanities and social sciences

eWaterCycle II

Overcoming the challenge of locality using a Community Multi-Model Environment

Inside the filter bubble

A framework for deep semantic analysis of mobile news consumption traces

SecConNet

Smart, secure container networks for trusted Big Data Sharing

FEDMix

Fusible Evolutionary Deep Neural Network Mixture Learning from Distributed Data for Robust Medical

GlamMap

Visual Analytics for the World's Library Data

Deep learning OCR post-correction

Evaluation and post-correction of OCR of digitised historical documents

Genetics of sleep patterns

Detecting human sleep from wearable accelerometers data without the aid of a smartwatch

Bridging the gap

Digital Humanities and the Arabic-Islamic corpus

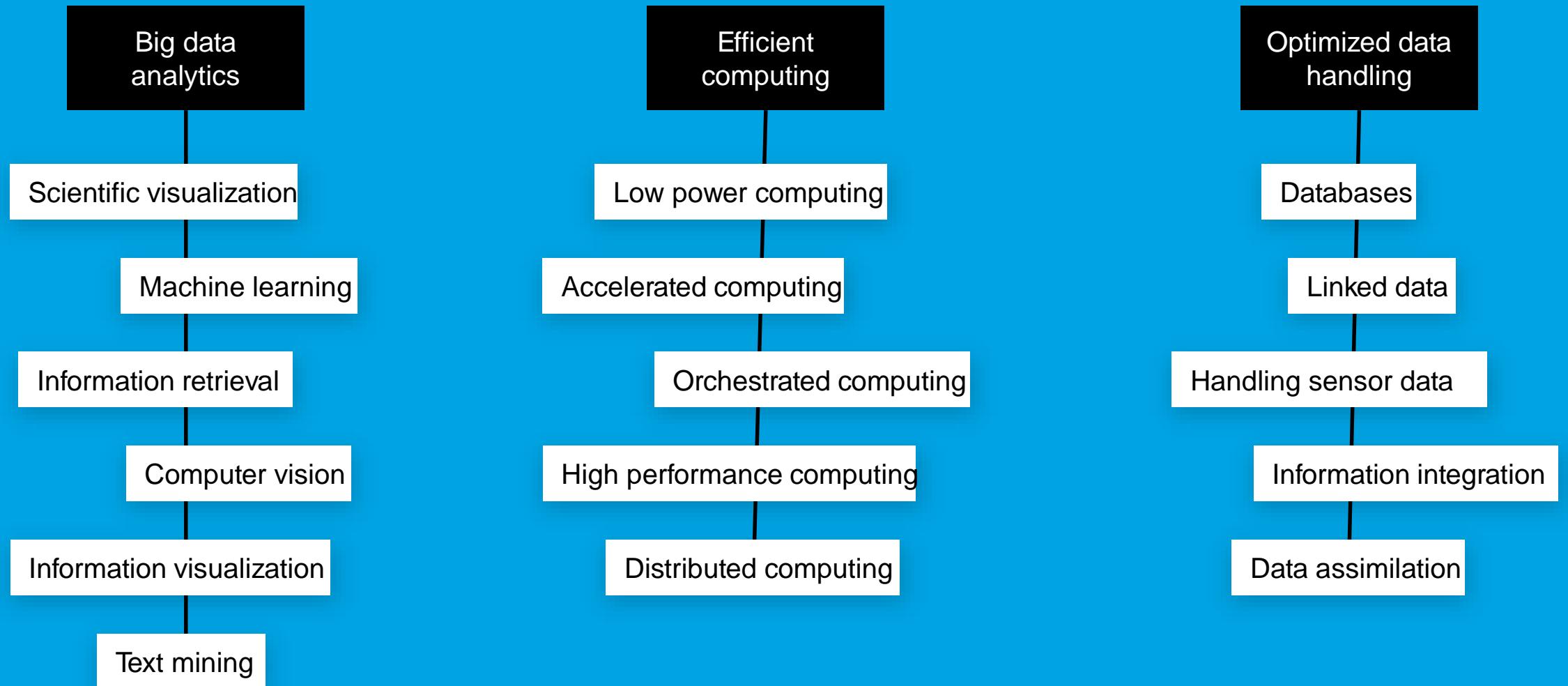
eEcoLiDAR

eScience infrastructure for Ecological applications of LiDAR point clouds

Emotion Recognition in Dementia

Advancing technology for multimodal analysis of emotion expression in dementia

Our technological expertise areas



What do we do?

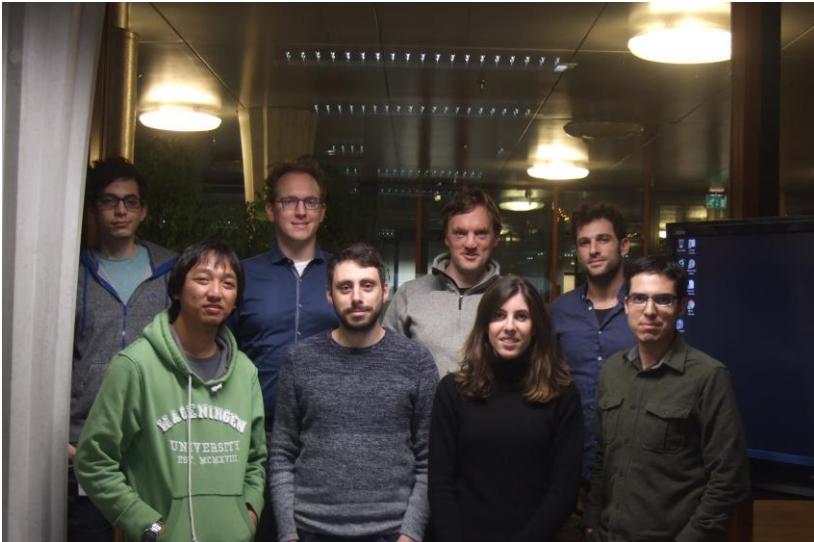
Research software

Link between researchers and IT infrastructure

Data stewards/data scientists

Cross-disciplinary transfer

Example project: Integrated ‘omics’ analysis



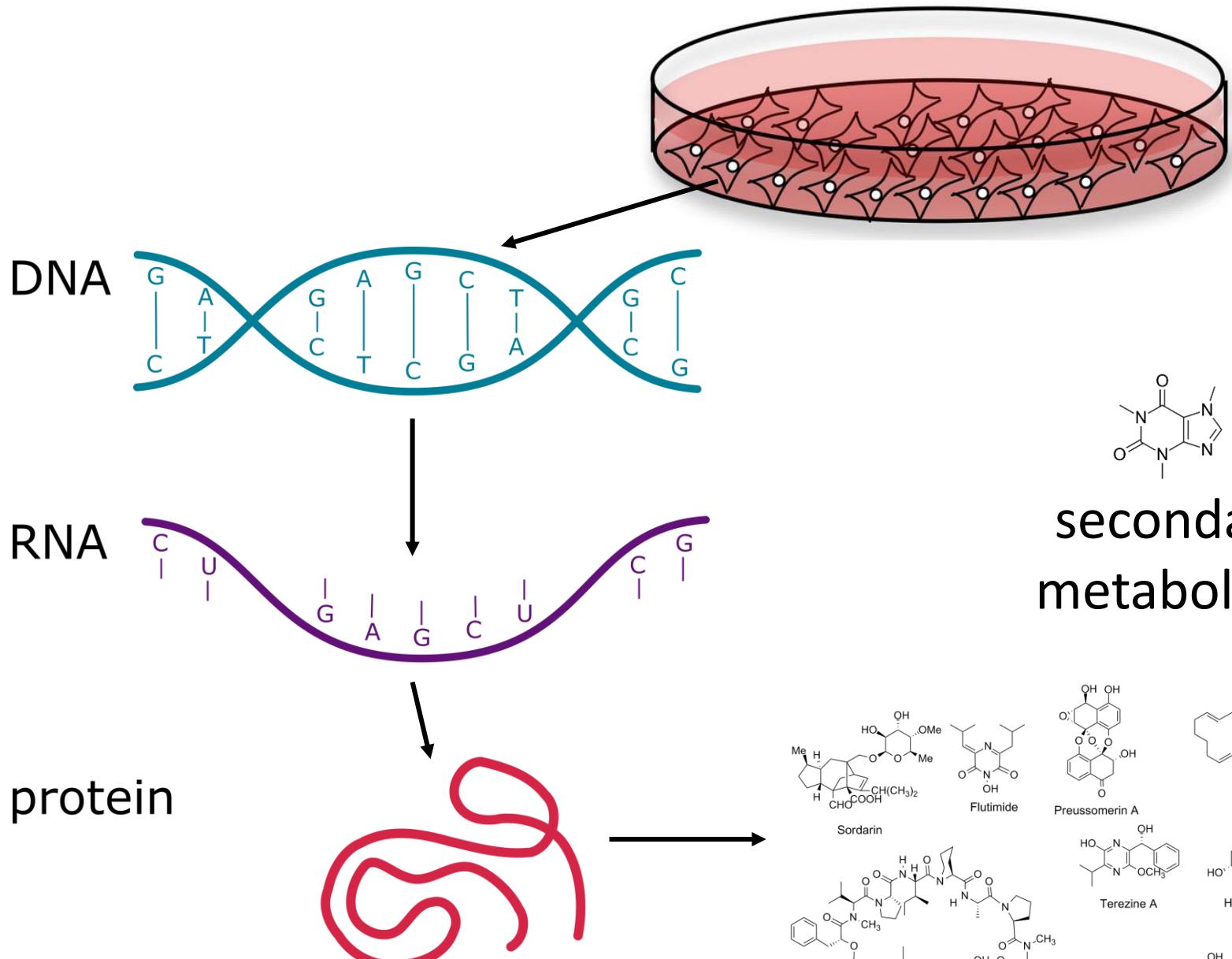
Medema lab - Wageningen UR, NL



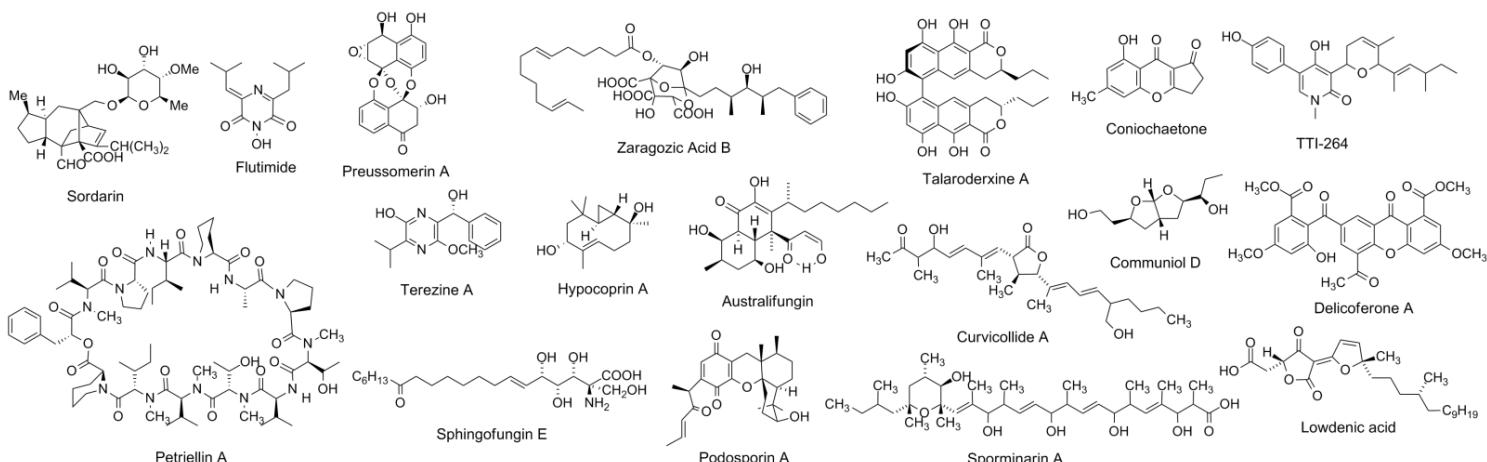
NL eScience Center

Glasgow University:
Simon Rogers,
Andrew Ramsay,
Grimur Hjorleifsson Eldjar

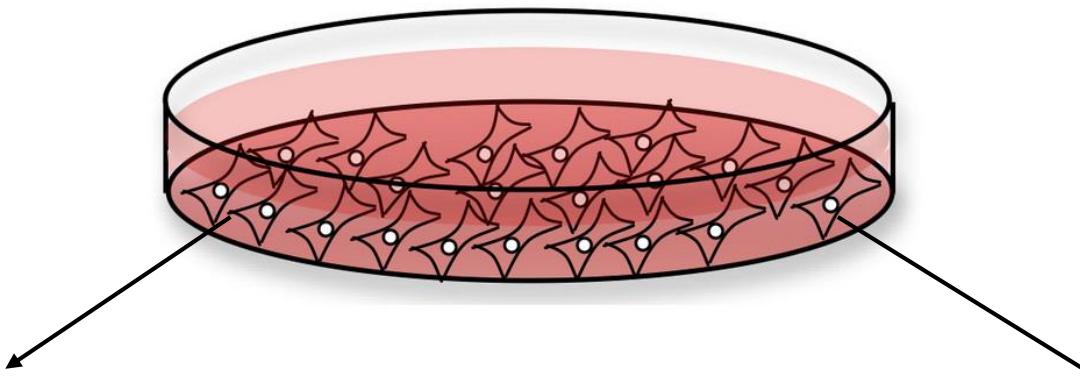
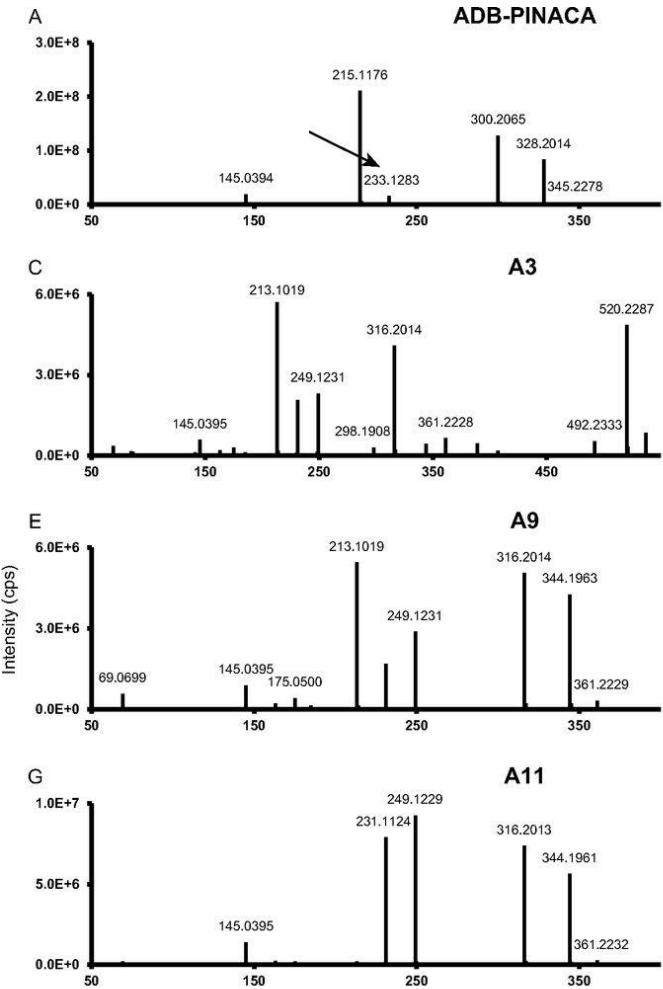
UCSD:
Madeleine Ernst
Pieter Dorrestein



secondary metabolites



mass spectra

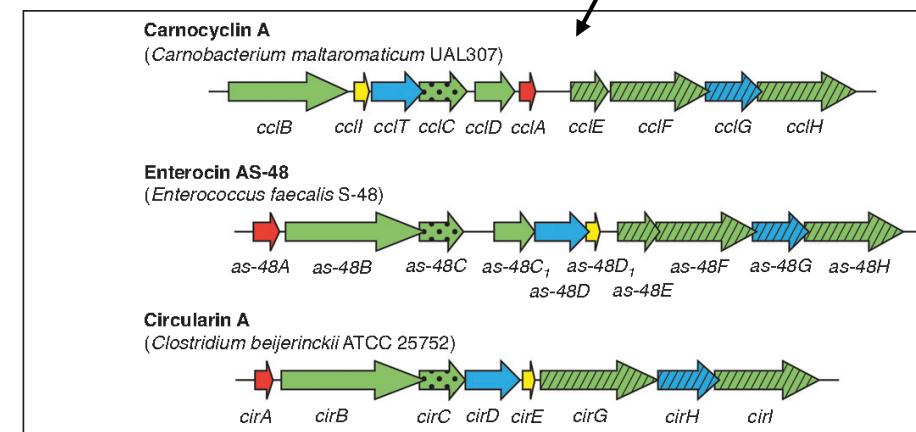


DNA

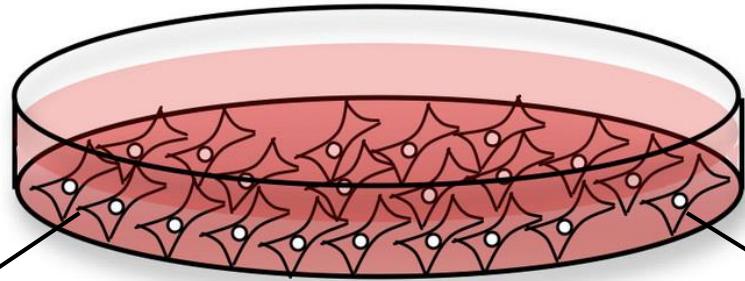
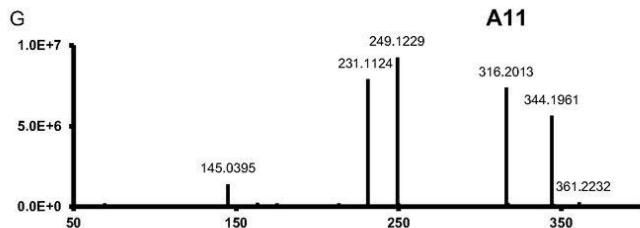
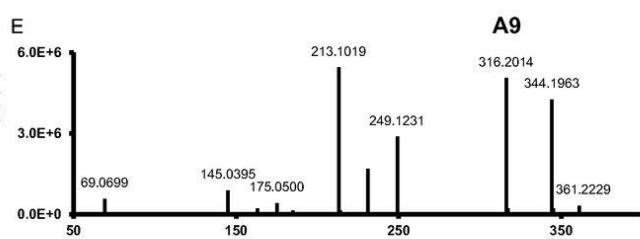
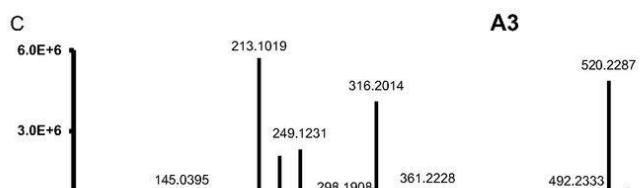
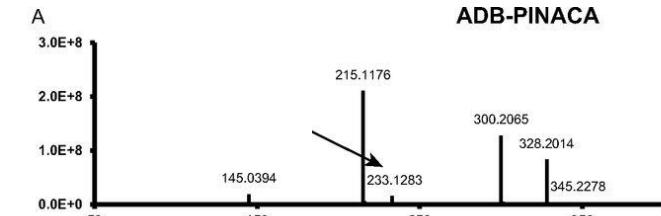
atggctatcg acgaaaacaa
caatttggta aaggctccat
atctctaccg gttcgcttgc
cgtatcgctcg aaatctacgg
atcgccgcag cgcaacgtga
gacccaatct acgcacgtaa
gacacccggcg agcaggcact
gttatcgctcg ttgactccgt
ggcgactctc acatgggcct
ggtaacctga agcaactccaa
ggtgtgtatg tcggtAACCC
tctgttgctc tcgacatccg
agcggaaatccc gcgtgaaagt
ttccagatcc tctacggcga
aaagagaagc tgatcgagaa
cagggttaaag cgaatcgac
gagaagaaaag tacgtgagtt
gatgatagcg aaggcgtac
aagggtcgca tqtgccggcc
acagaaaagcg ttggccggcag
catgcgcctg ggtgaagacc
actggatate ggcgttgggg
accggaatct tccggtaaaaa
ccacgctgac gctgcaggtg
aggtaaaacc tggcggtta tcgatgctga
actgggcgtc gatatcgaca acotgctgt
ggaaaatctgt gacgcctgg cgcttctgg
ggcggcactg acgccgaaag cgaaaatcga
tgccggcactg atgatgagcc aggccatgcg
cacgcgtctg atcttcatca accagatccg
ggaaaaccat acgggtggta acgcgcgtaa
tcgtatcgcc gggtgaaag agggcgaaaa
ggtgaaagaaacaaaatcgctg cgcgttta
aggtatcaa ttctacggcg aactgggttg
agcaggcgtc aactgggttgc acaaagggtga
tgccgttgc aggtatcabc cggaaaaccgc
gctgcttgc aaccggact caacgcggga
agaaaatacgatgcgttgc aactgggttgc
ttttttttt ttaagttgttgc aggtatcabc
atgacacgtatgcgttgc

HMM
(Hidden Markov Model)

HMM
(Hidden Markov Model)
+ manually written rules



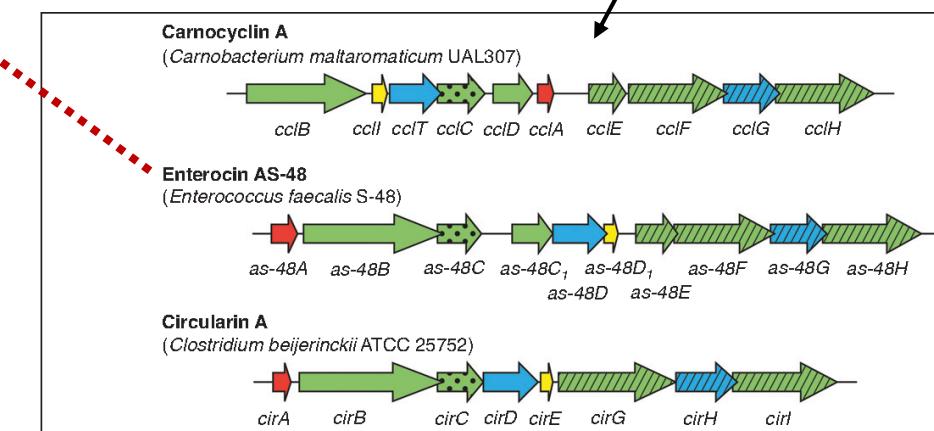
mass spectra



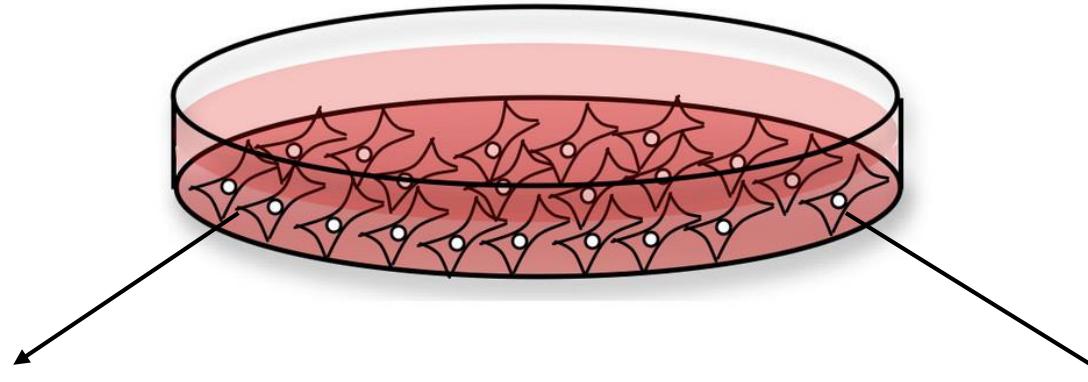
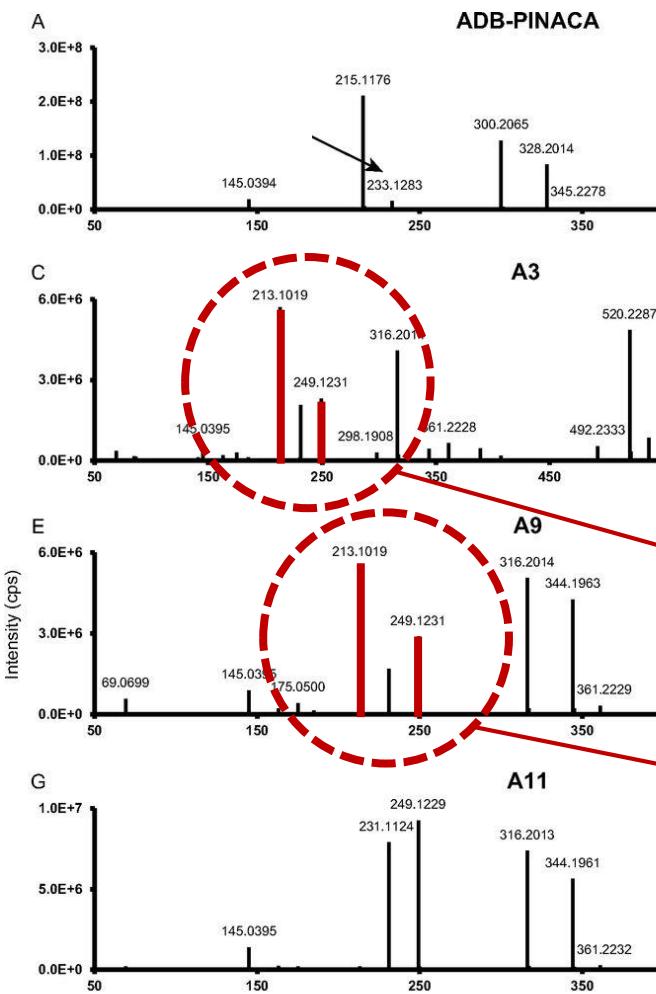
DNA

atggctatcg acgaaaacaa acagaaaagcg ttggccgcag cactggggcca gattgagaaaa
caattttgtta aaggctccat catgcgcctg ggtgaagacc gttccatggta tggggaaacc
atctctaccg gttcgcttc actggatatac ggcgttgggg cagggtgtct gccgatgggc
cgtatcgtcg aaatctacgg accggaatct tccggtaaaa ccacgctgac gctgcagggt
atcgccgcag cgacagcgtga aggtaaaacc ttgtgcgtta tcgatgctga acacgcgcgt
gacccaaatct acgcacgtaa actgggcgtc gatatcgaca acctgctgtg ctcccgaccc
gacaccggcg agcaggcact ggaaatctgt gacgcctgg cgccgtctgg cgccgttagac
gttatacgctcg ttgactccgt ggcggcactg acgcccggaaag cgaaaatcgaa aggccgaaatc
ggcgactctc acatgggcct tgccggcacgt atgatgagcc aggccatgcg taagctggcg
ggtaaacctga agcagtccaa cacgctgtg atcttcatca accagatccg tatgaaaatt
ggtgtgtatgt tcgtaaccc gcaaaccact accgggtgtta acgcgtgaa attctacgcc
tctgttgctc tcgacatccg tcgtatcgcc gccgtgaaag agggcgaaaa cgtgggtgggt
agcgaaaccc gcgtgaaagt ggtgagaacaaatcgctg cgccgtttaa acaggctgaa
ttccagatcc tcgacggcga aggtatcaat ttctacggcga aactgggtga cctgggcgt
aaagagaagg tgatcgagaa agcaggcgaaatct acaaagggtga gaagatcggt
cagggttaaa cgaatgcgc acgtccgttg aadatgtaaac cgaaaaccgc gaaagagatc
gagaagaaag tacgtgagtt gctgctggc aaccggact caacgcggaa tttctctgt
gatgatagcg aaggcgtago agaaaatac tttttttttt tttttttttt ttaatttgta aggatatcg
aagggtcgca tttttttttt tttttttttt ttaatttgta aggatatcg atgacacat

HMM
(Hidden Markov Model)
+ manually written rules

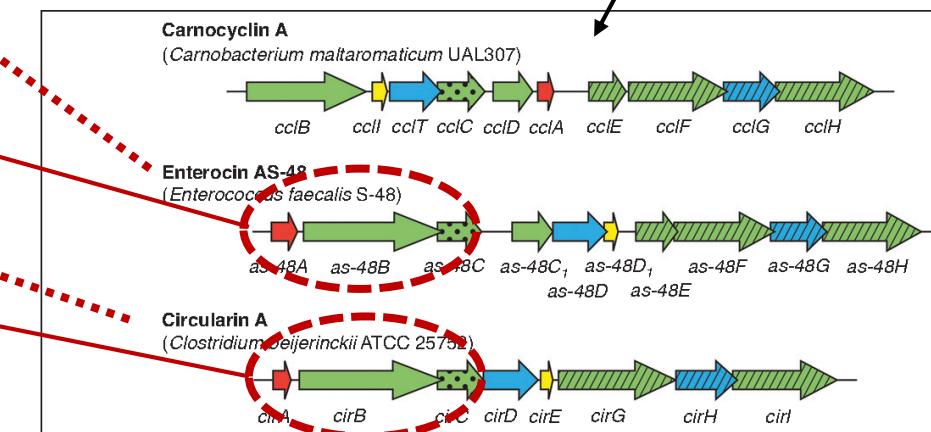


mass spectra



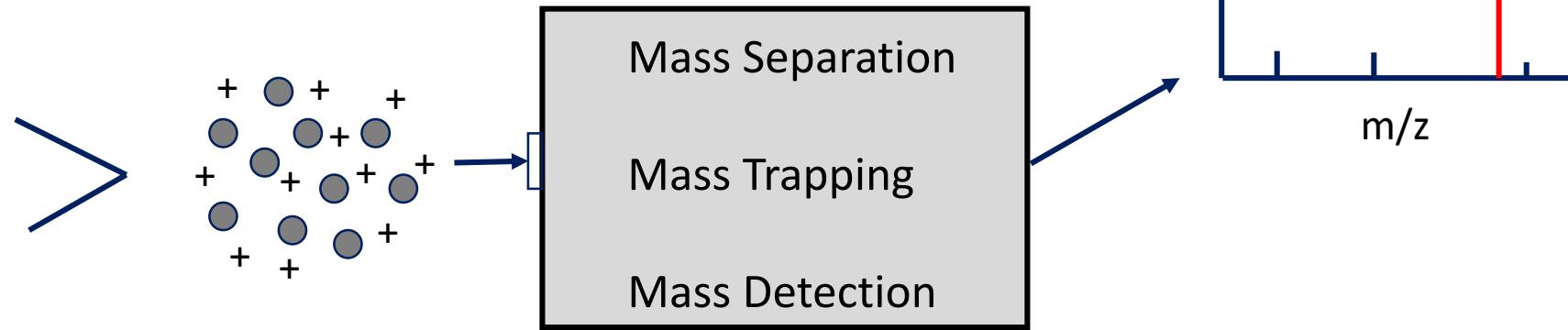
DNA

HMM
(Hidden Markov Model)
+ manually written rules



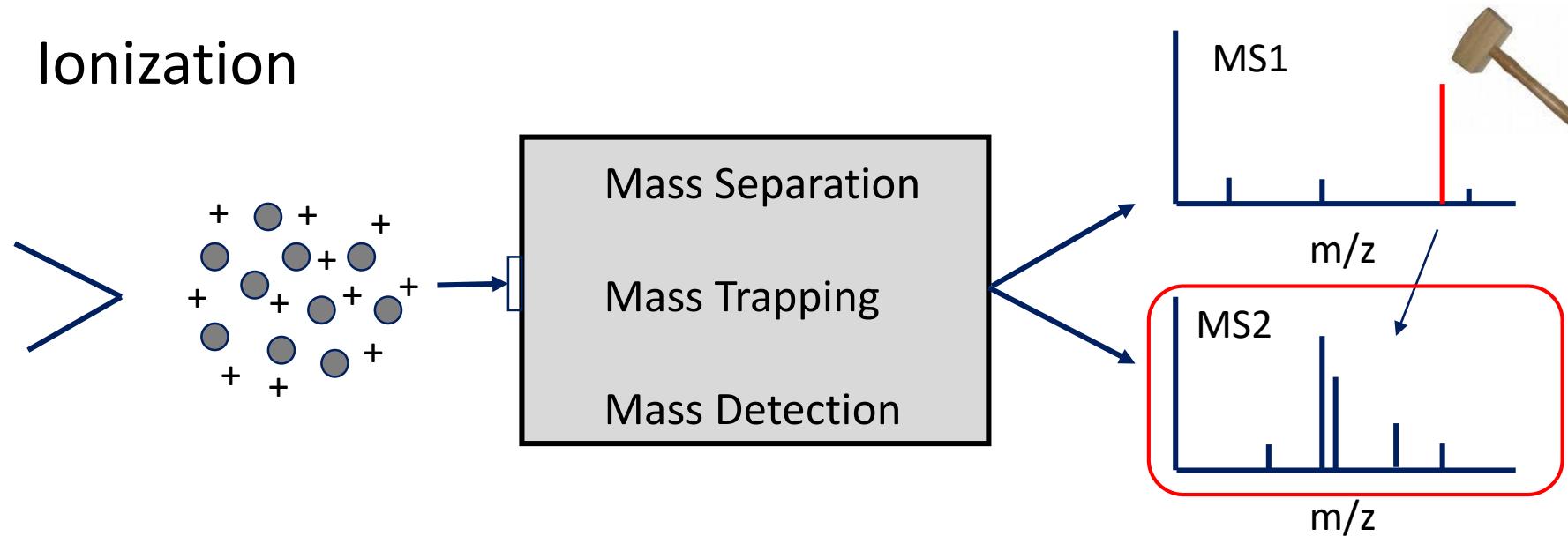
Mass spectrometry and fragmentation

Ionization

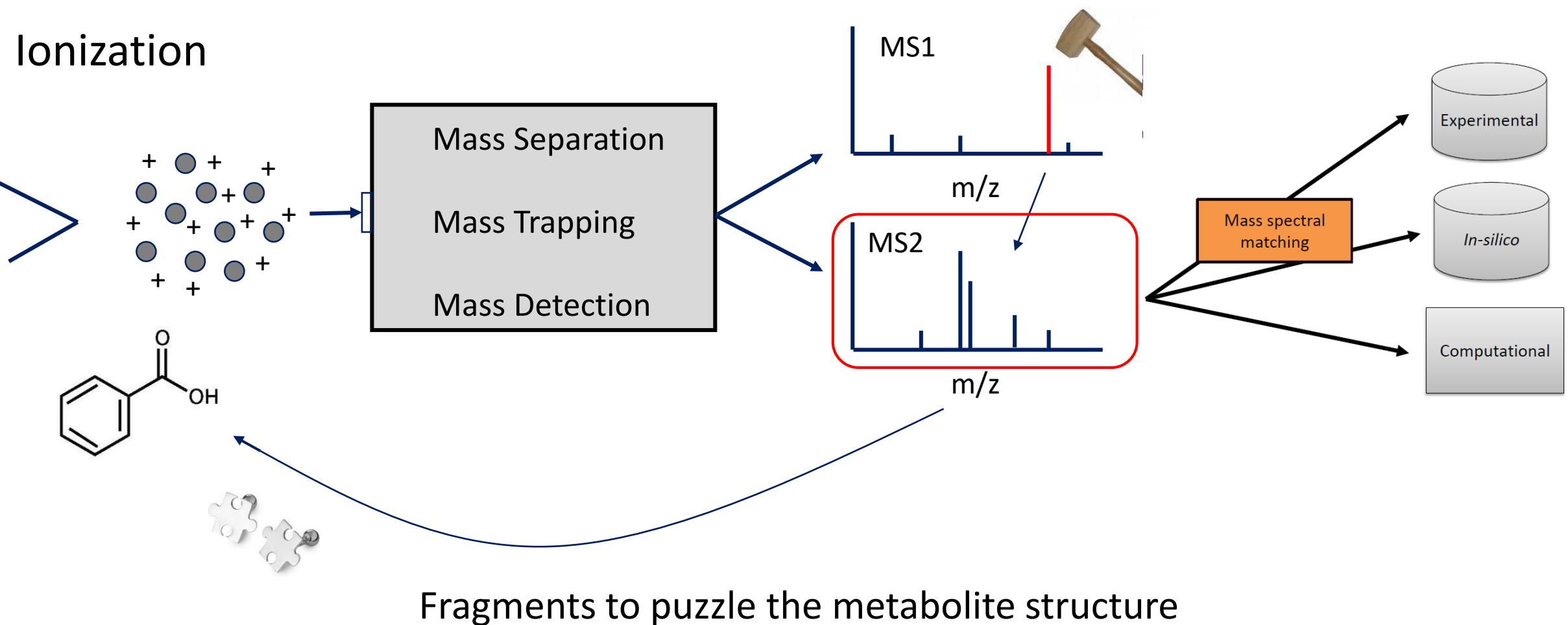


Mass spectrometry and fragmentation

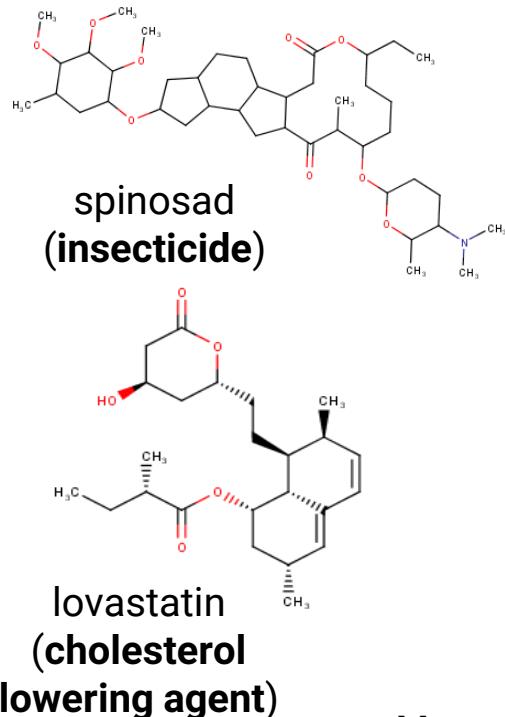
Ionization



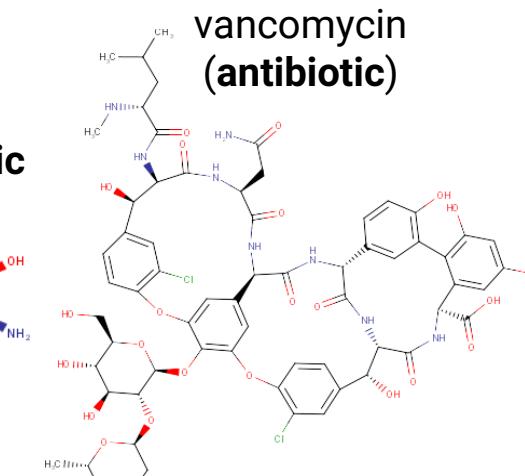
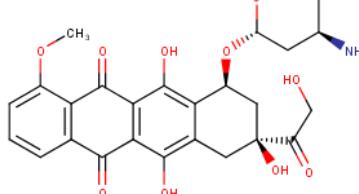
Mass spectrometry and fragmentation



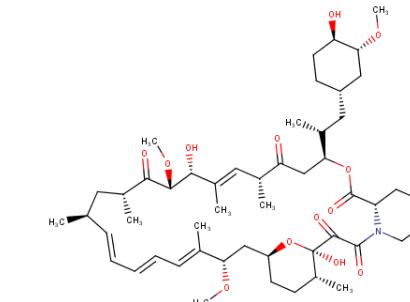
Bacteria, fungi, and plants produce a large & diverse arsenal of high-value molecules:



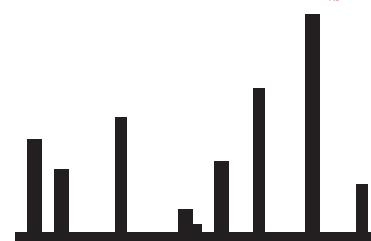
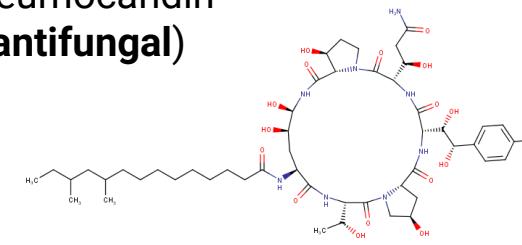
doxorubicin
(chemotherapeutic agent)



rapamycin
(immunosuppressant)



Pneumocandin
(antifungal)



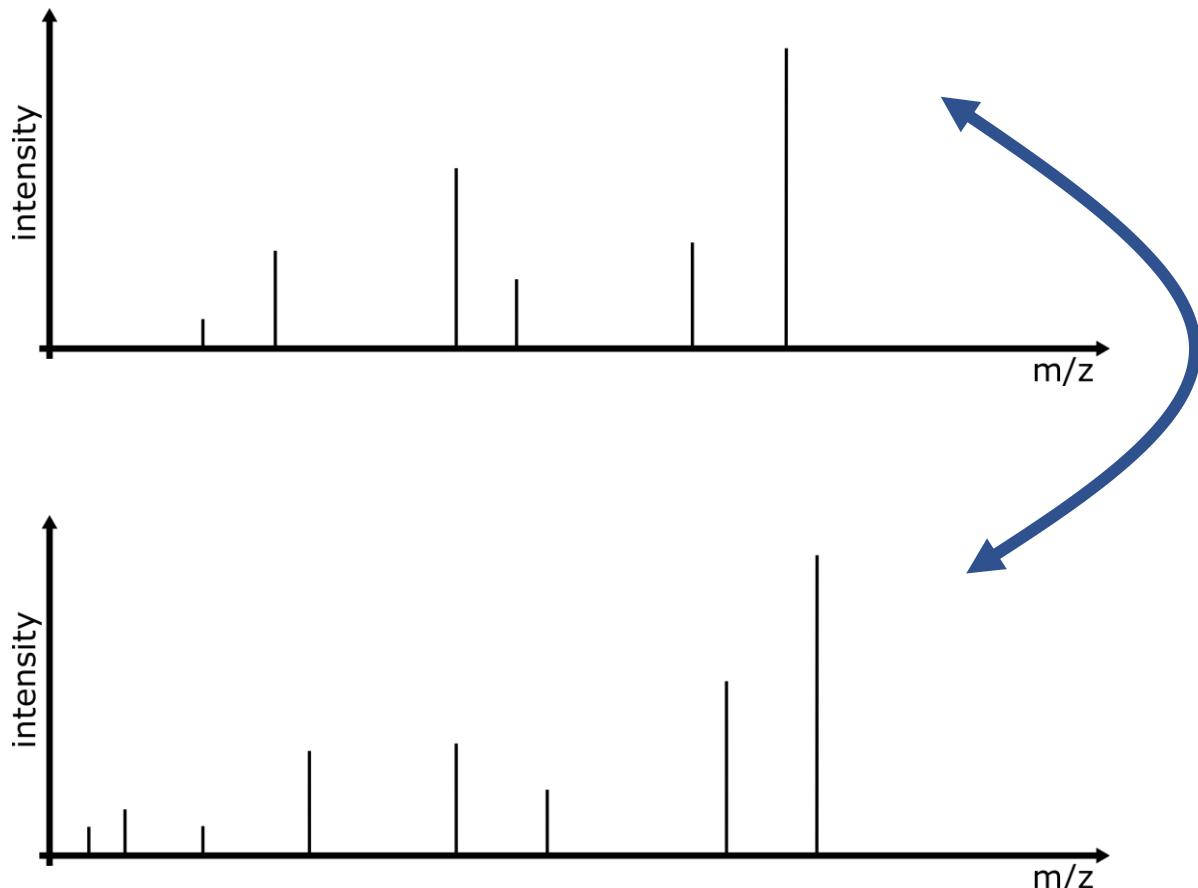
Mass spectrometry fragmentation spectrum

The challenge....

....is large-scale coupling of spectral data to molecular structures
of known & especially novel natural products molecules.

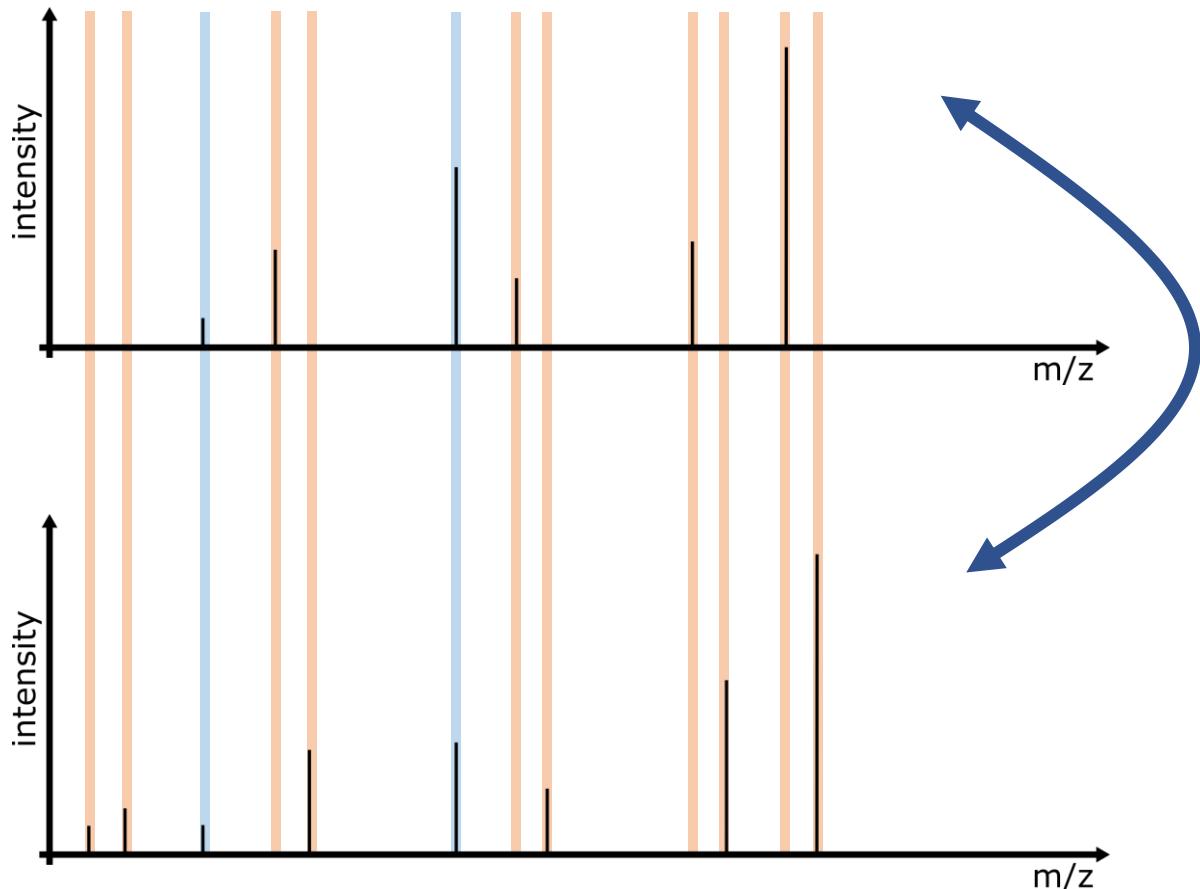
But.... How similar are they?

Spectral similarity



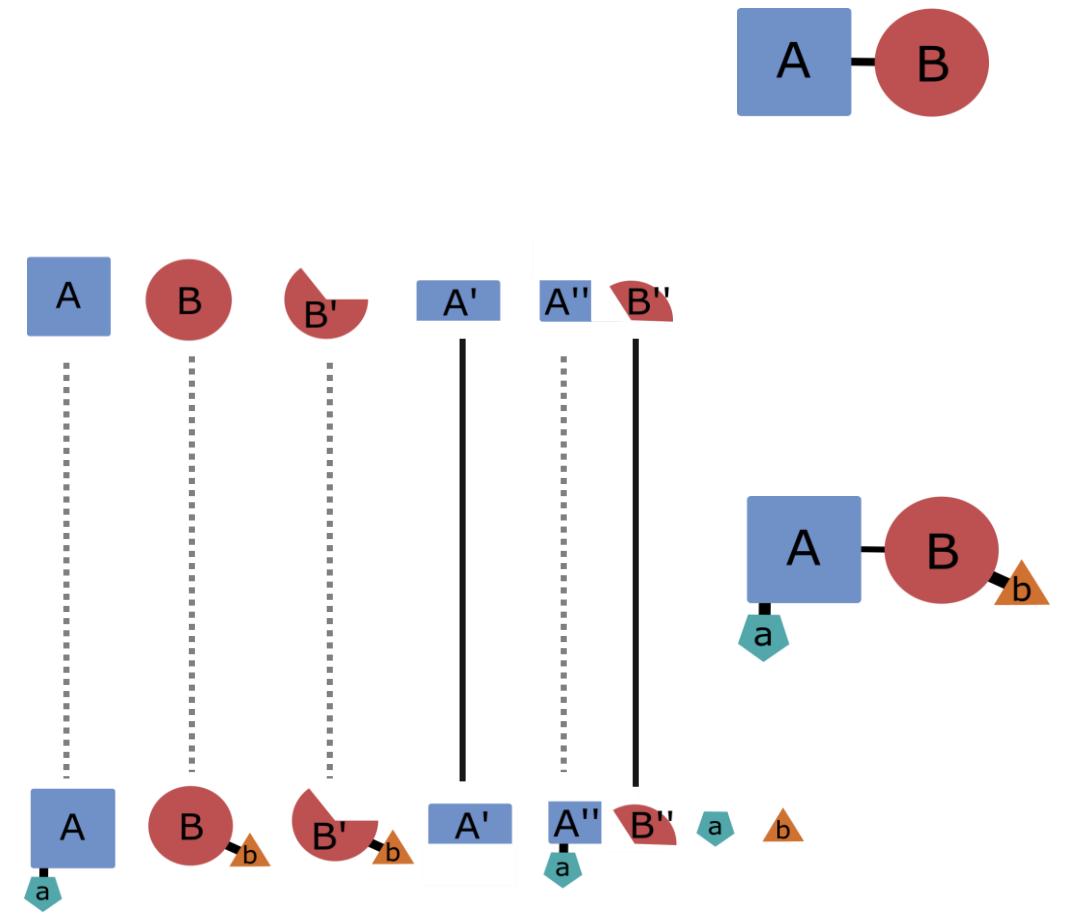
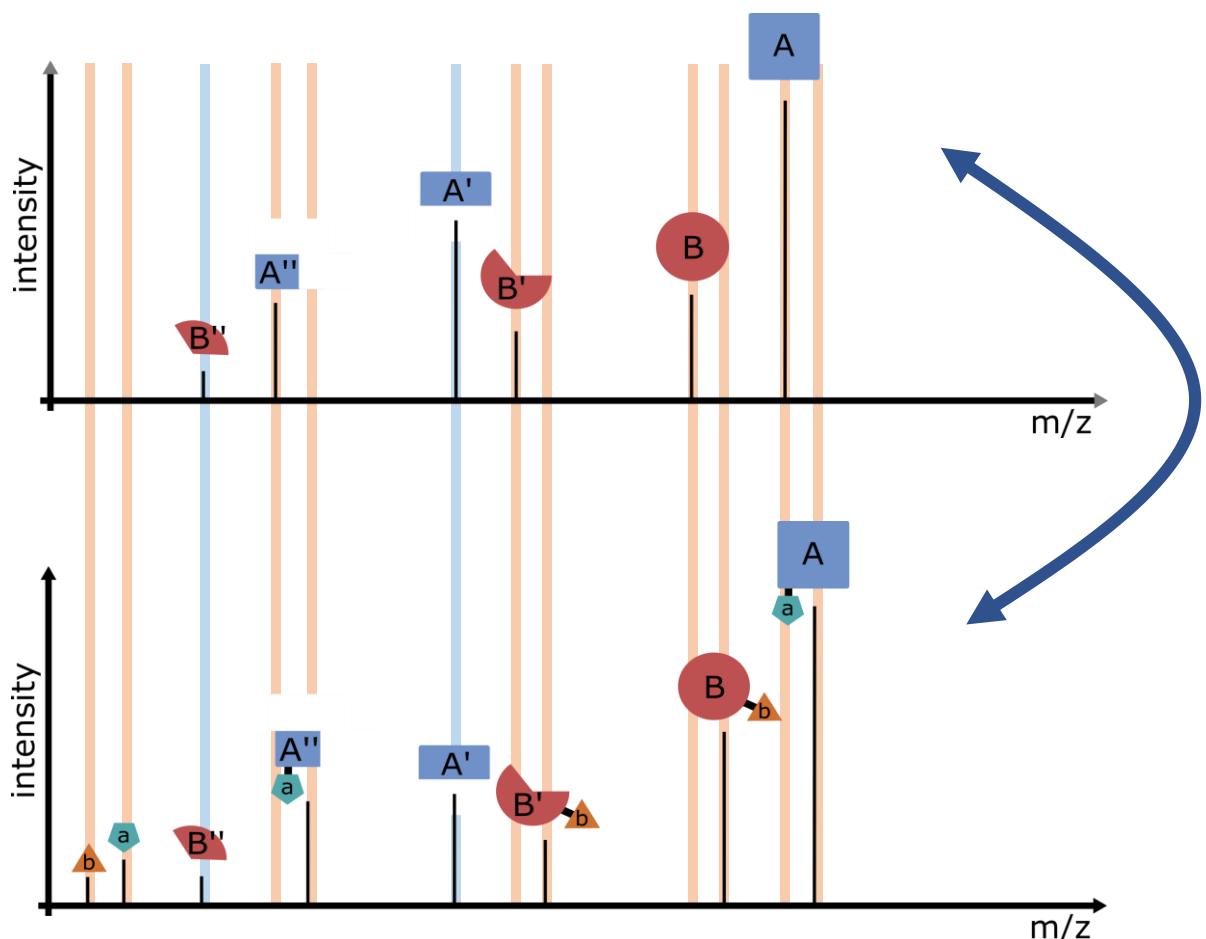
How similar are they?

Spectral similarity



How similar are they?

Spectral similarity



How similar are they?

...likes cake with a cappuccino.

...loves to have a cookie and a coffee.

What does similar mean?

number of words?

number of characters?

grammatical structure?

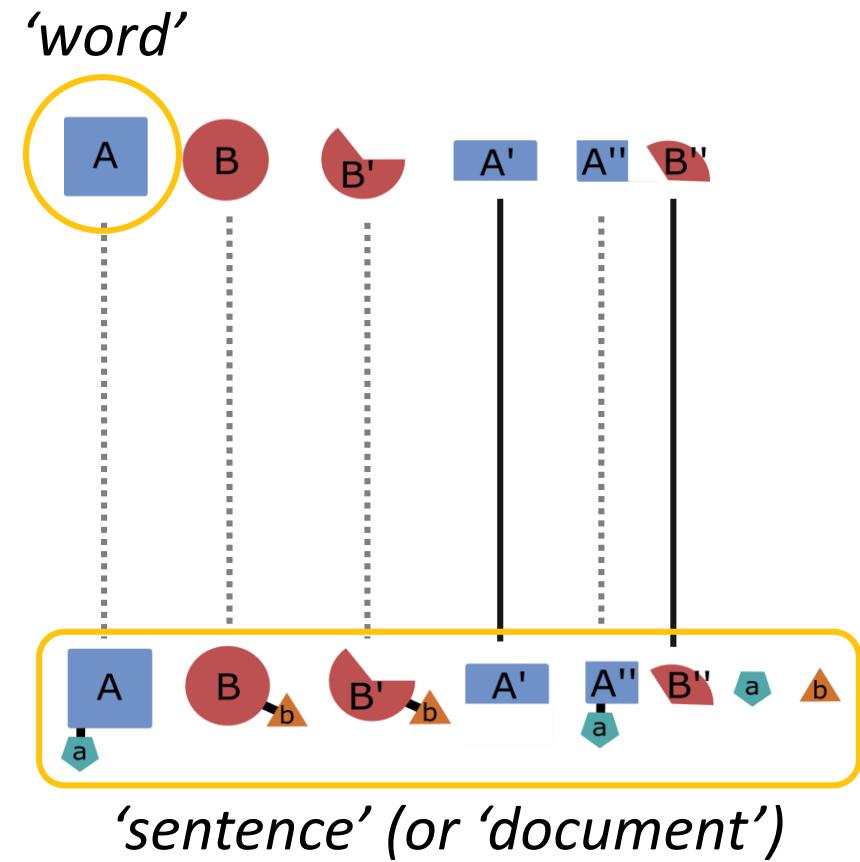
topic?

meaning?

style?

phonetic structure?

...likes cake with a cappuccino.
...loves to have a cookie and a coffee.



Count how often ‘words’ co-occur (*find word ‘context’*)

Words cake ... cookie ... sweet ... → all words in corpus...

monster	...	0	0		9				...
...	0								
cake	0				0		24		
...									
cookie	9		0				17		
...									
sweet		24		17					
...									

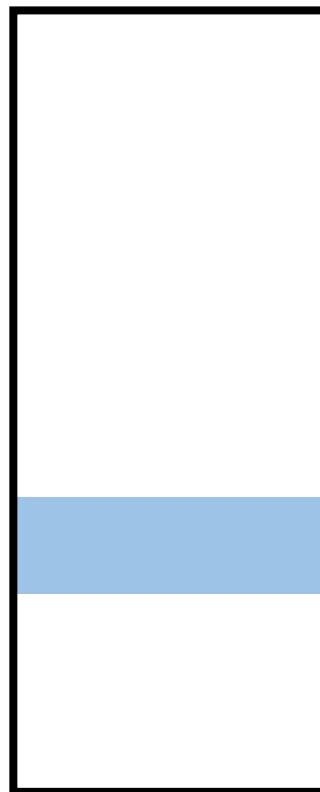
NxN matrix

N: number of words in dictionary

‘Word2Vec’ → lower dimensional context vector

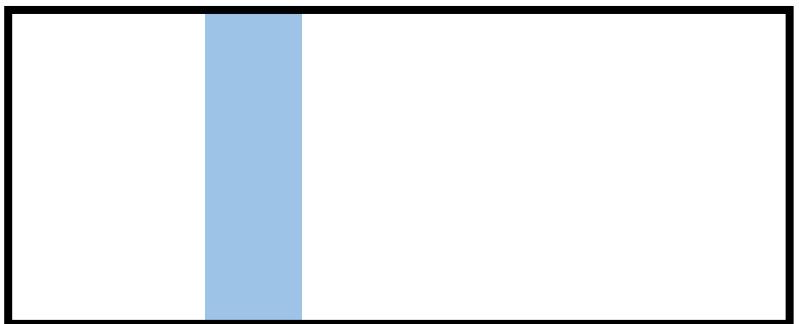
Words	cake	...	cookie	...	sweet	...
monster		0	0		9			...
...	0							
cake	0				0		24	
...								
cookie	9		0				17	
...								
sweet			24		17			
...	...							

2



factorization

X



‘Word2Vec’ → lower dimensional context vector

Words	cake	...	cookie	...	sweet	...
monster	...	0	0		9			...
...	0							
cake	0				0		24	
...								
cookie	9		0				17	
...								
sweet			24		17			
...	...							



v_{cookie}

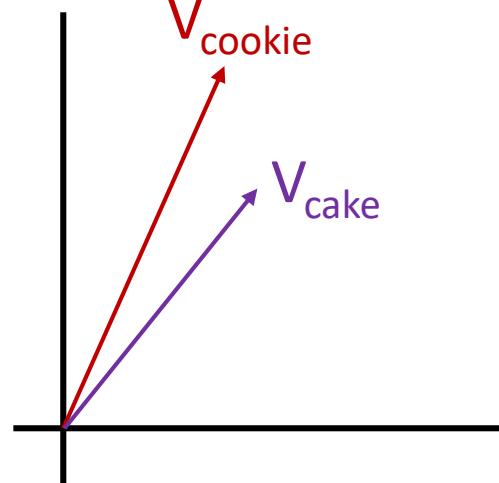


v_{cake}



v_{cookie}

v_{cake}

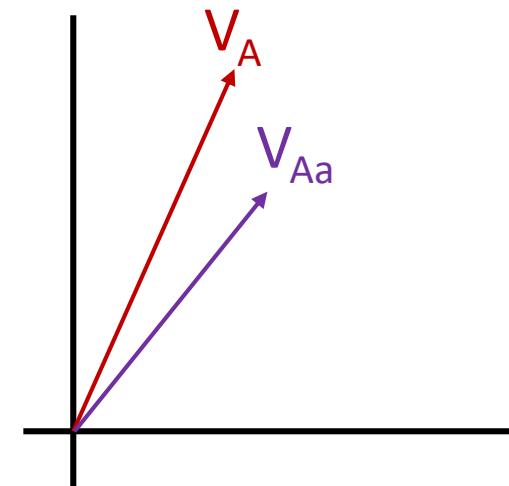
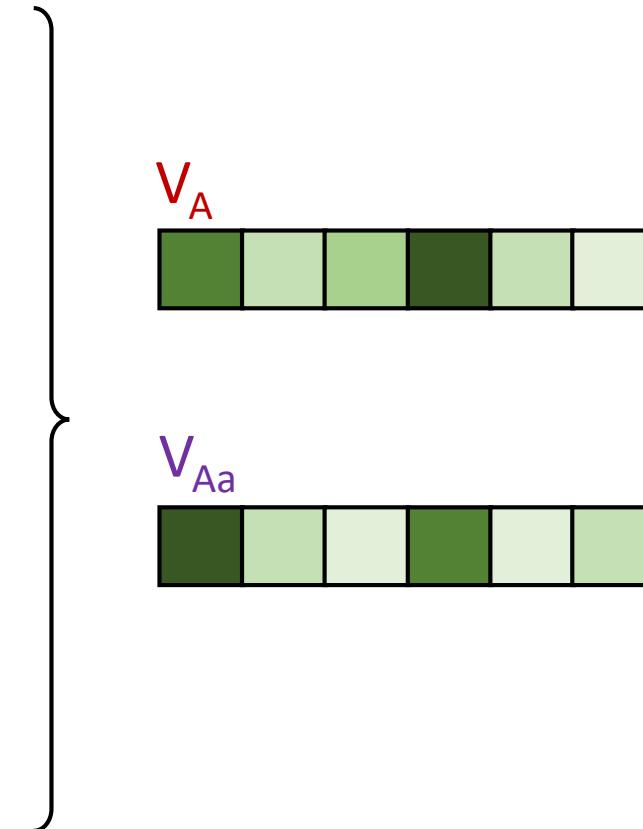


NLP → metabolomics: use peaks as words

peak positions

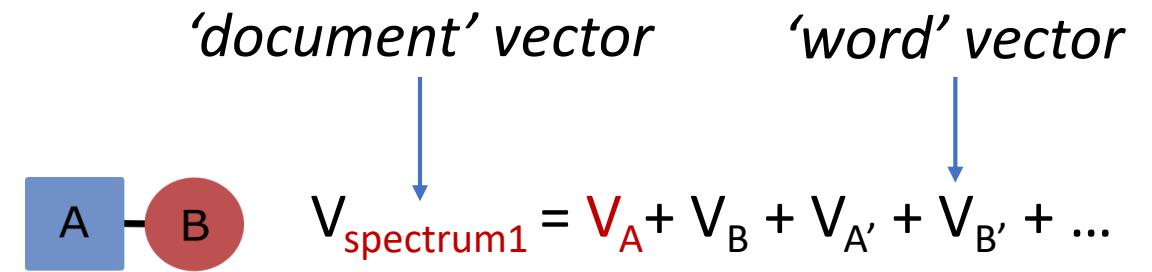
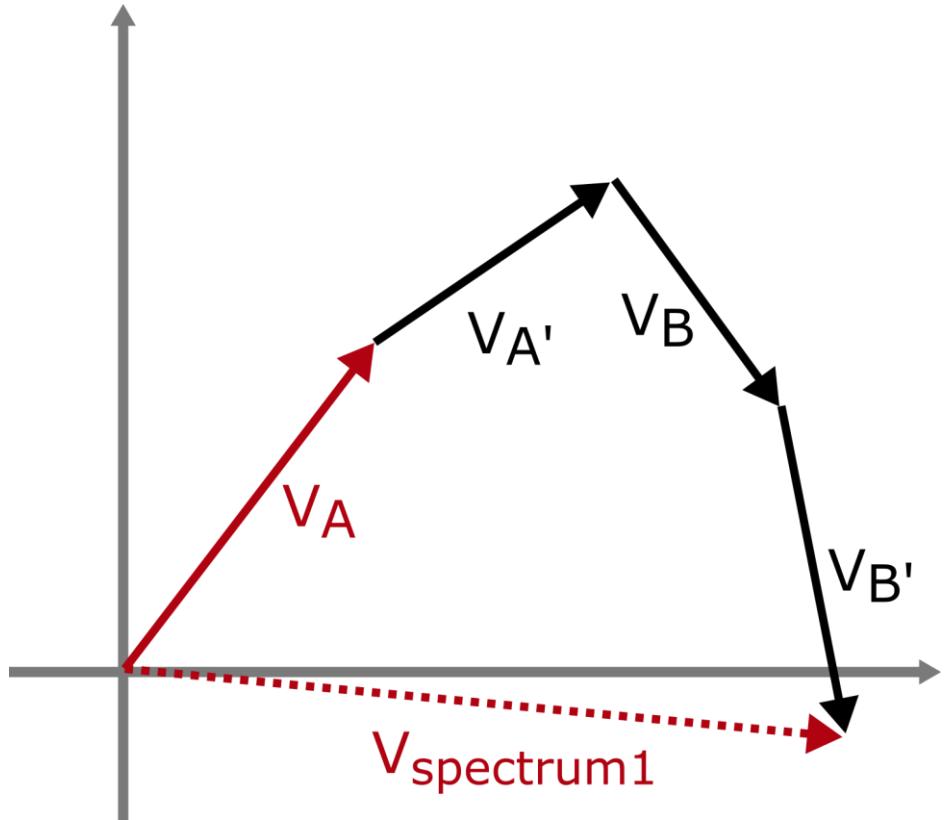
='words'

	$m(Aa)$...	$m(A)$...	$m(A'')$...
...	0	0		9				...
0								
$m(Aa)$	0				0		24	
...								
$m(A)$	9	0				17		
...								
$m(A'')$		24	17					
...	...							



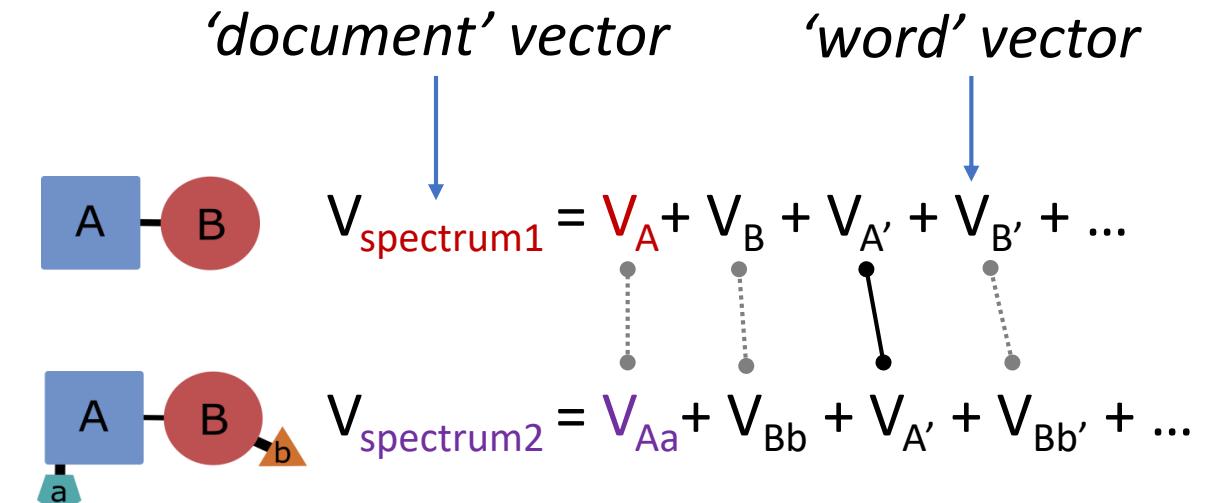
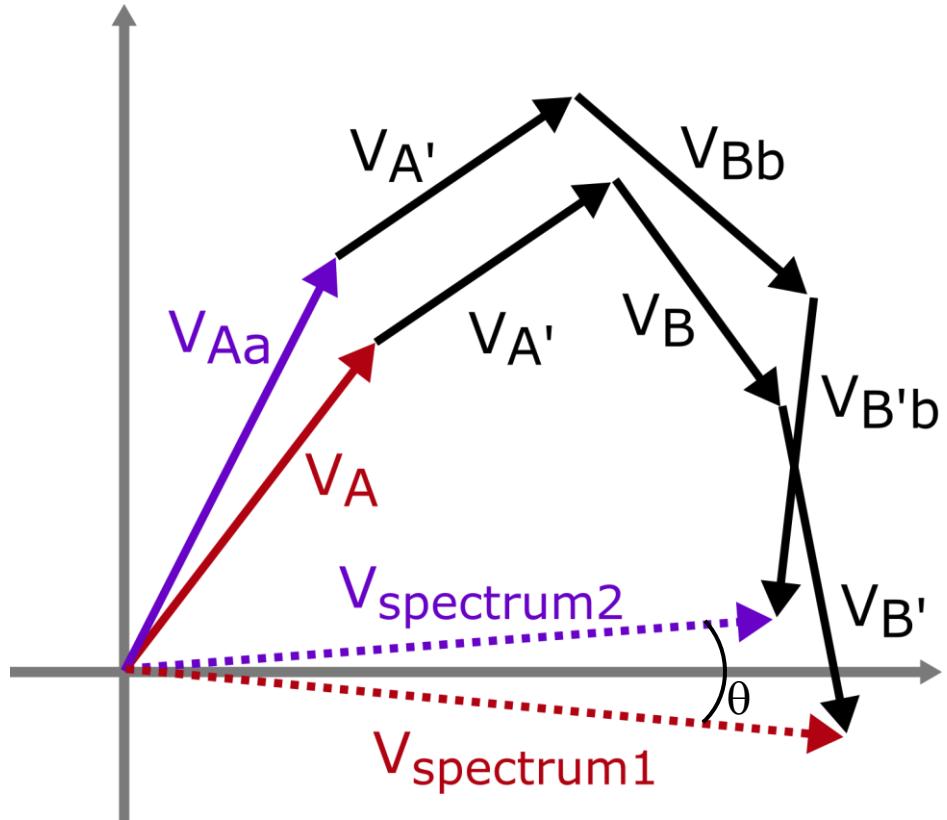
Spectral similarity measures.

NLP/word2vec based method



Spectral similarity measures.

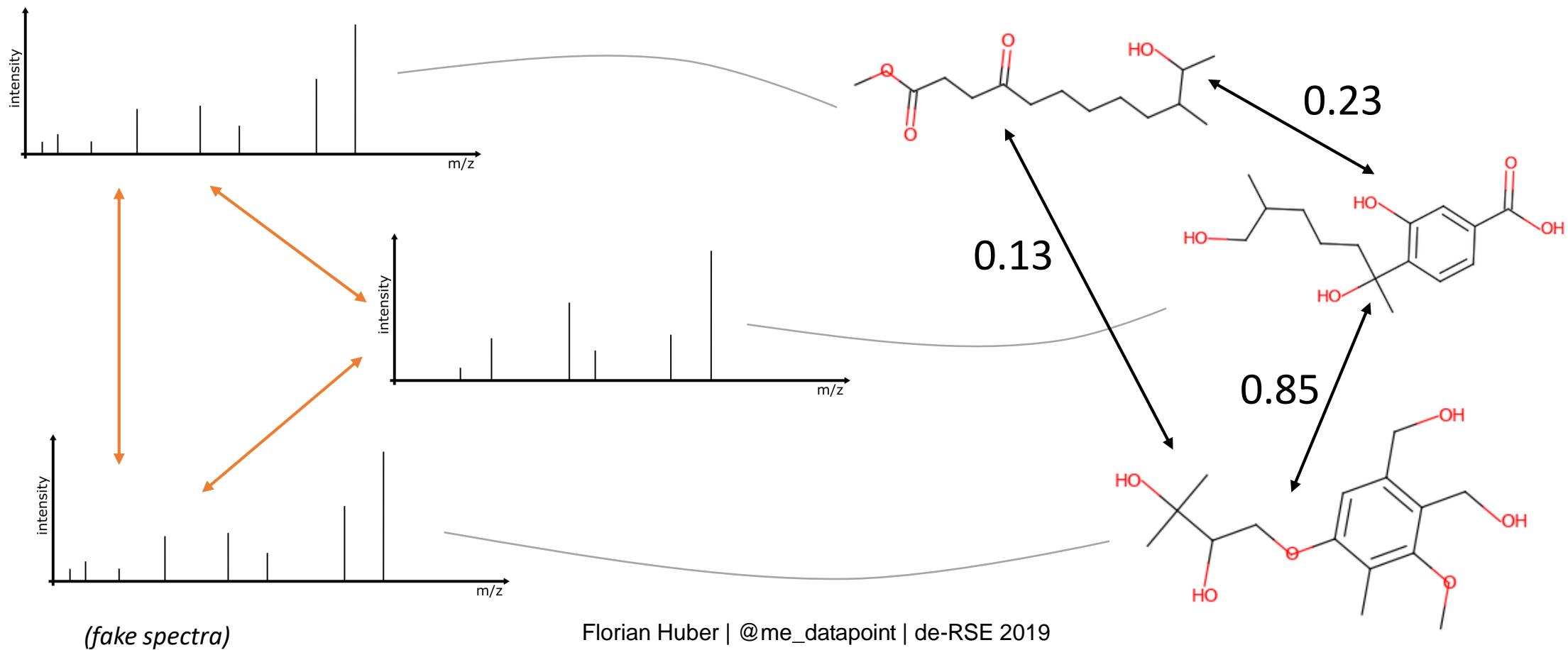
NLP/word2vec based method



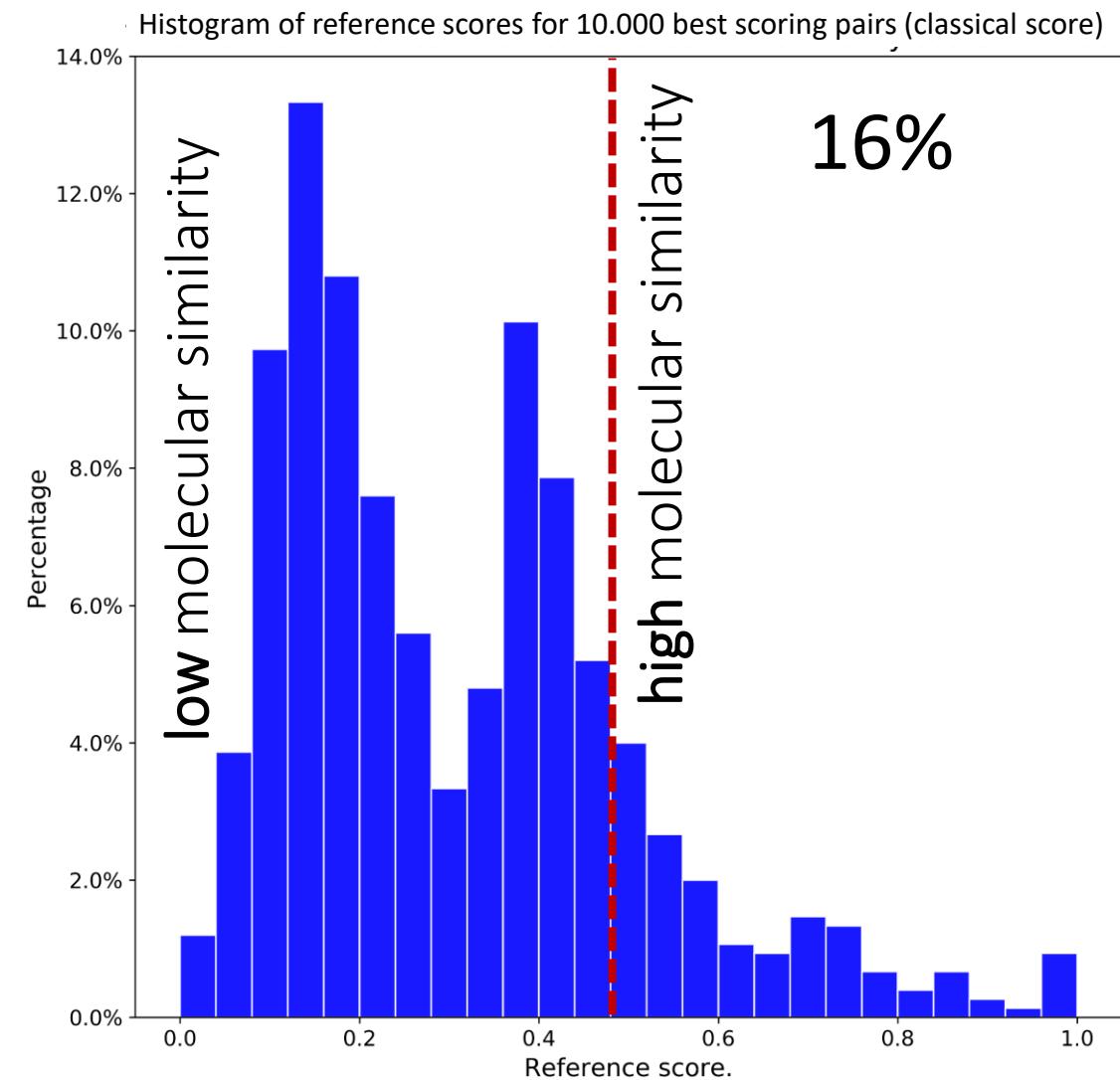
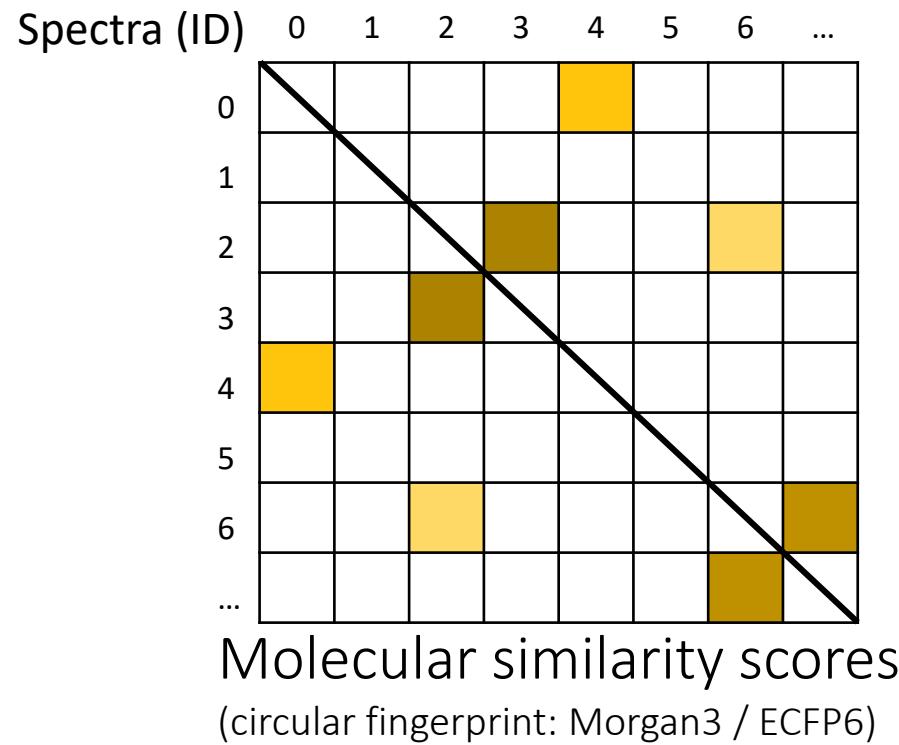
$$\text{Similarity} = \cos(\theta) = \frac{V_{\text{spectrum}_1} \cdot V_{\text{spectrum}_2}}{\|V_{\text{spectrum}_1}\| \|V_{\text{spectrum}_2}\|}$$

Spectral similarity measures: evaluation.

Dataset: 11.000 spectra with known molecular structures

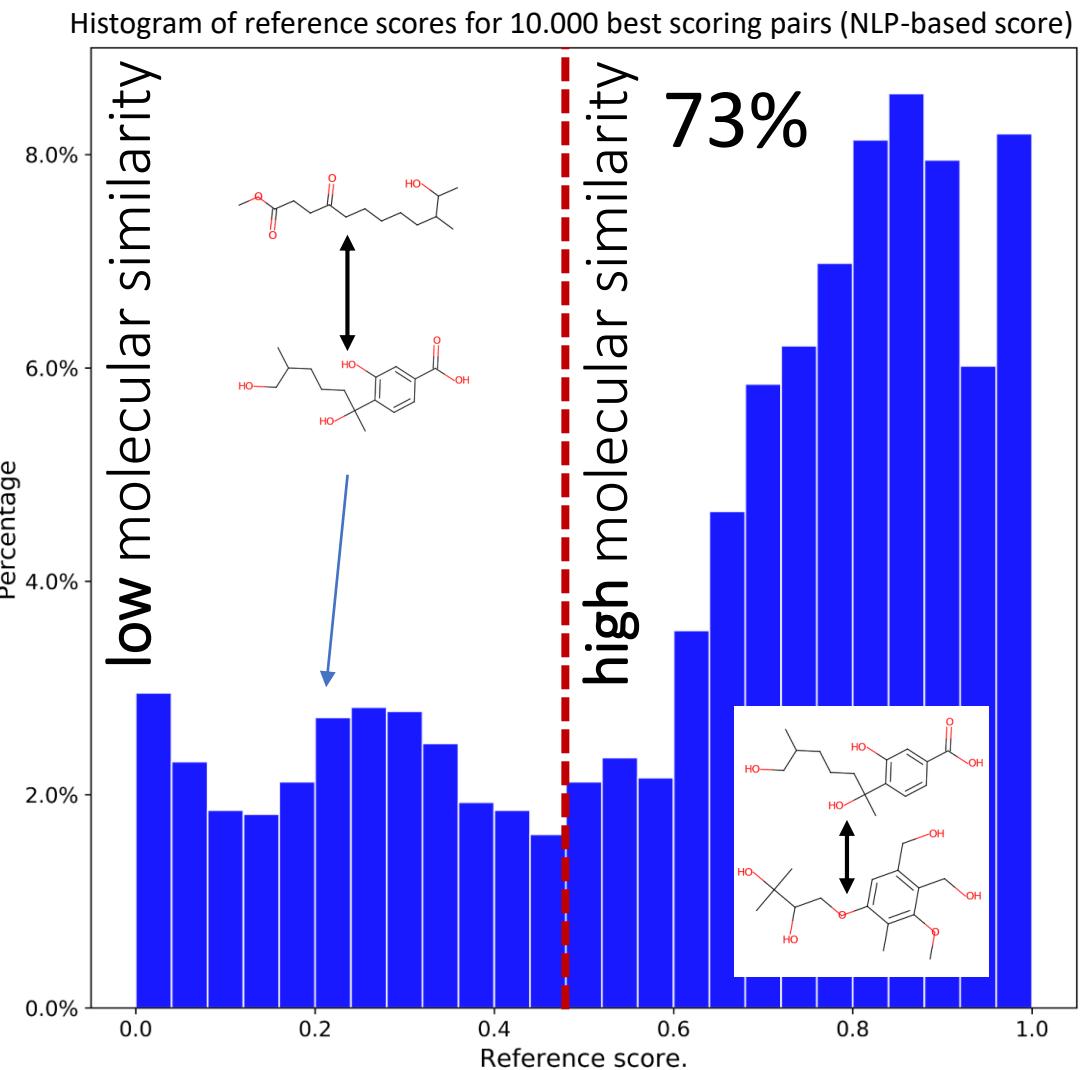
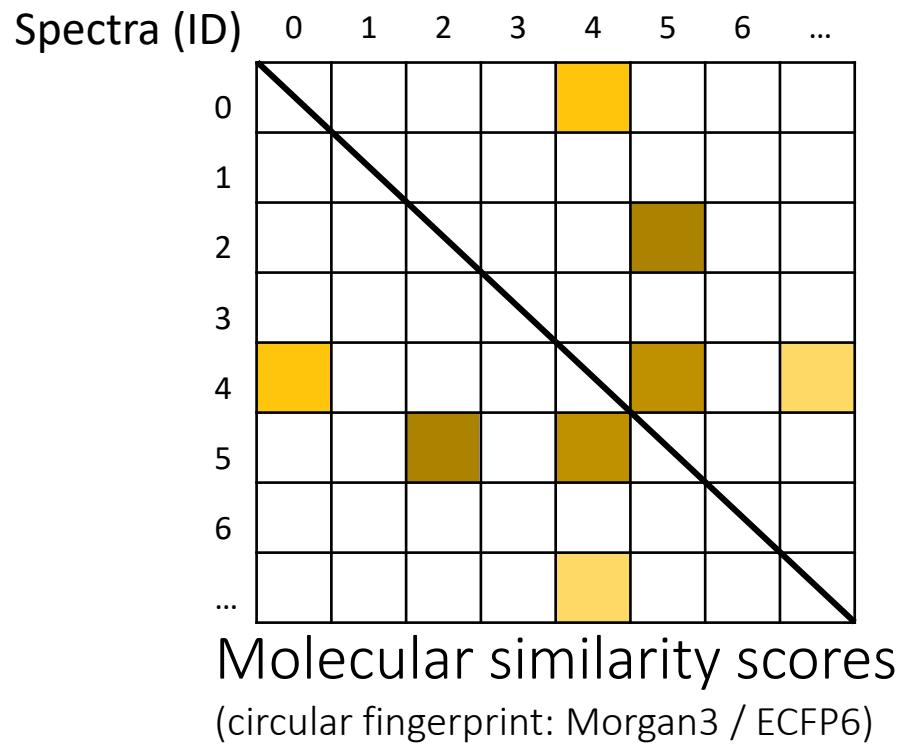


Molecular similarity scores: 10.000 highest 'classical' scores*



* = scores > 0.998

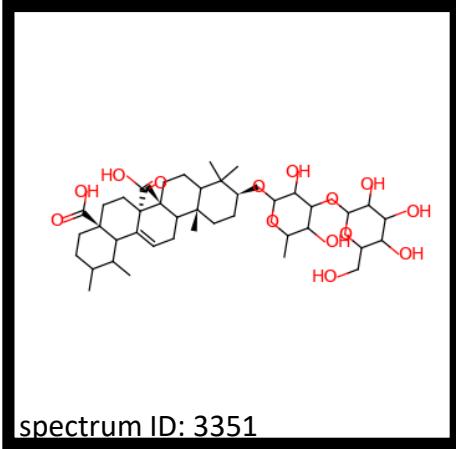
Molecular similarity scores: 10.000 highest NLP-based scores*



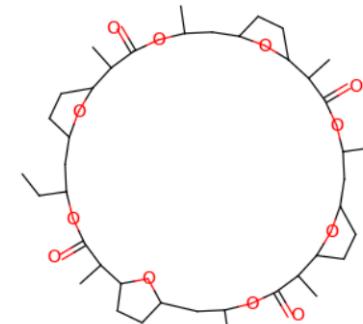
* = scores > 0.84

Spectral similarity measures: examples.

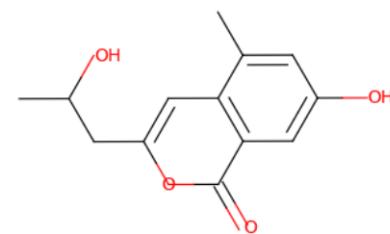
query molecule



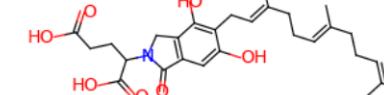
9 closest candidates (according to molecular networking similarity)



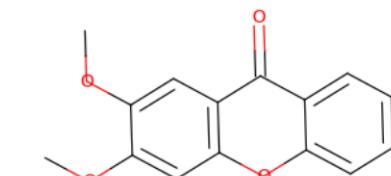
1



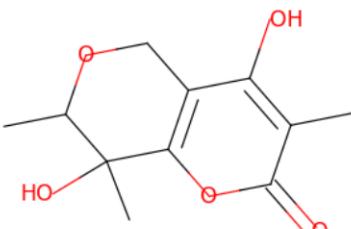
2



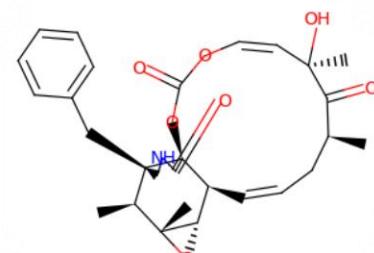
3



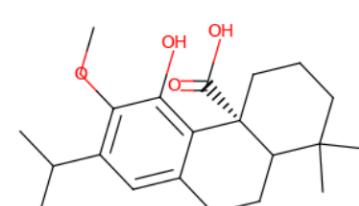
4



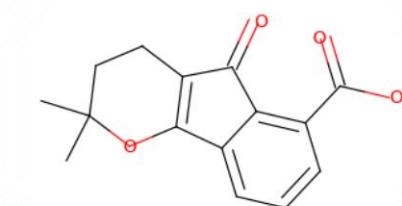
5



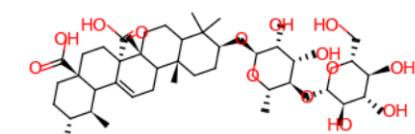
6



7



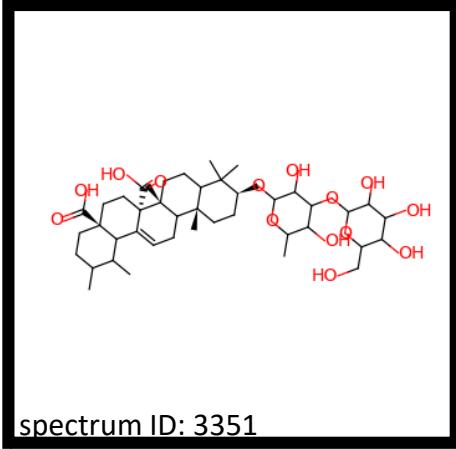
8



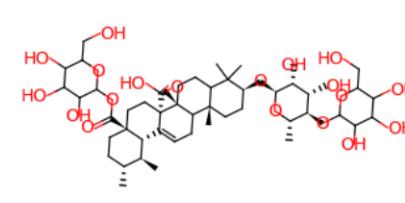
9

Spectral similarity measures: examples.

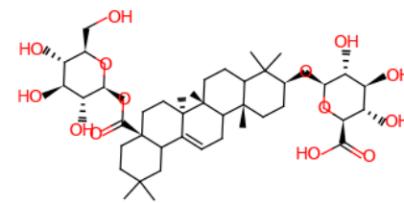
query molecule



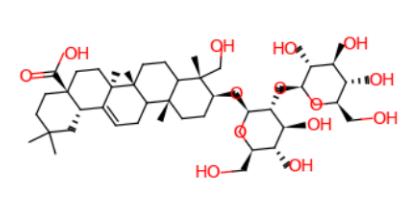
9 closest candidates (according to Word2vec-based spectral similarity)



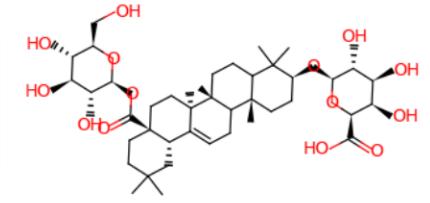
1



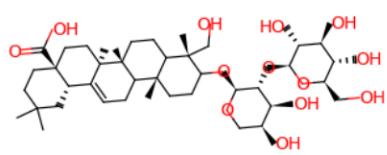
2



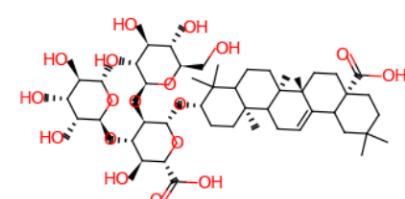
3



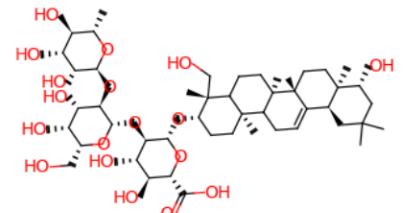
4



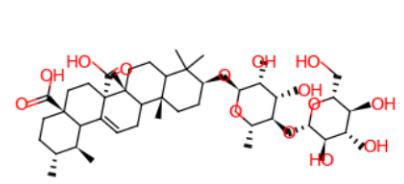
5



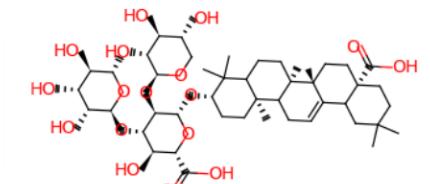
6



7



8



9

RSE's creating unique links

- RSE's – working in teams with broad range of expertise and backgrounds.
- RSE's – working on projects of different scientific domains.

→Creating opportunities unlike anywhere else in the academic setting!

- Transfer methods/techniques between domains.
- Spot potential synergies between (sub-)fields.

Interested in Research Software ?



The Netherlands eScience Center is the Dutch national center of excellence for the development and application of research software to advance academic research.

Join the team !



n.renaud@esciencecenter.nl



netherlands eScience center



+31 (0)20 460 4770



www.esciencecenter.nl



blog.esciencecenter.nl

Florian Huber



@me_datapoint



Carlos Martinez-Ortiz



@neocarlitos

