netherlands eScience center

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

Florian Huber[1], Justin van der Hooft[2], Simon Rogers[3], Marnix Medema[2], Lars Ridder[1]

[1] Netherlands eScience Center

[2] Bioinformatics Group, Wageningen University

[3] School of Computing Science, University of Glasgow

de-RSE conference, Potsdam 05/06/2019

netherlands eScience center

Breaking down scientific mono-cultures by cross-disciplinary software development

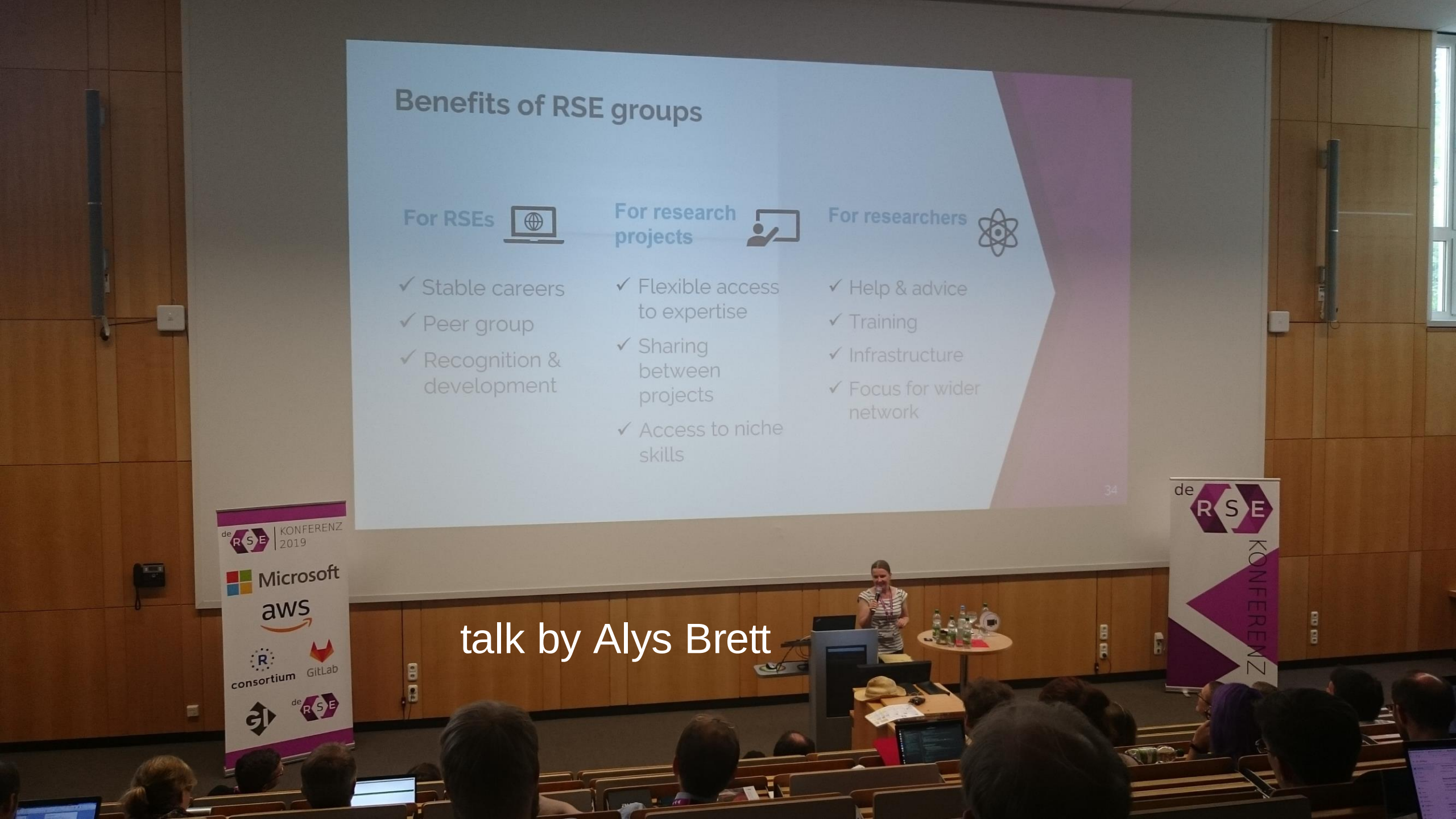
Florian Huber[1], Justin van der Hooft[2], Simon Rogers[3], Marnix Medema[2], Lars Ridder[1]

[1] Netherlands eScience Center

[2] Bioinformatics Group, Wageningen University

[3] School of Computing Science, University of Glasgow

de-RSE conference, Potsdam 05/06/2019

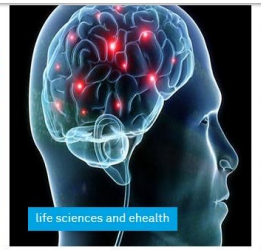

Benefits of RSE groups
For RSEs
✓ Stable careers
✓ Peer group
✓ Recognition & development
For research projects
✓ Flexible access to expertise
✓ Sharing between projects
✓ Access to niche skills
For researchers
✓ Help & advice
✓ Training
✓ Infrastructure
✓ Focus for wider network
34
de RSE KONFERENZ 2019
Microsoft
aws
consortium
GitLab
de RSE
de RSE KONFERENZ
talk by Alys Brett


netherlands eScience center

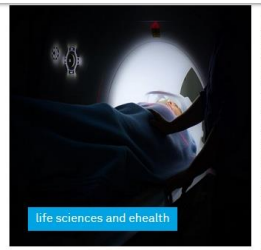
We signal challenges and opportunities at the intersection of software and academic research

Photography: Elodie Burrillon

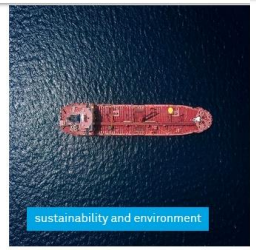
IS-ENES3
Providing the infrastructure to better understand and project climate variability and change



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

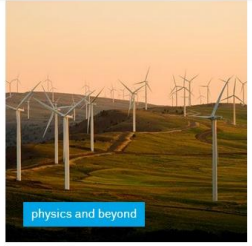


Digital twins: monitoring ships' state in real-time
Advanced data science to assist the design of cleaner, safer and smarter ships

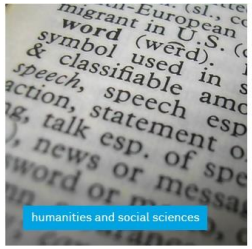
FAIR is as FAIR does
Integrating data publishing principles in scientific workflows

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

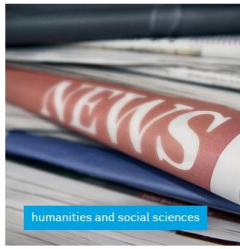
Computation of the Optical Properties of nano structures
Accurate and Efficient Computation of the Optical Properties of Nanostructures for Improved Photovol



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



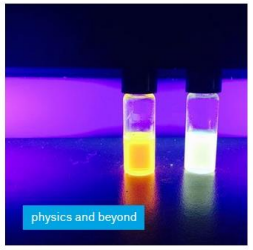
NEWSGAC
Advancing Media History by Transparent Automatic Genre Classification

EUropean Climate Prediction system

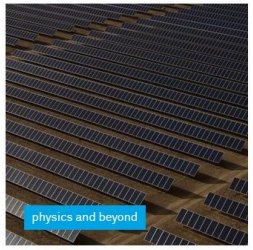
ePODIUM
Early Prediction of Dyslexia in Infants Using Machine learning

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

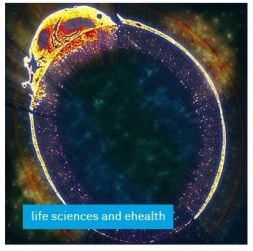
Automated Analysis of Online Behaviour on Social Media
Gaining insights in the use of Twitter by politicians and journalists

MOSAIC
MOdelling Sea level And Inundation for Cyclones

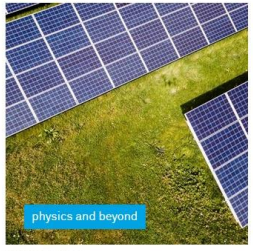
PROCESS
PROviding Computing solutions for ExaScale ChallengeS

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



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

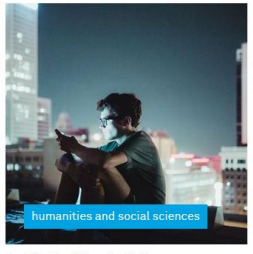
Stochastic Multiscale Climate Models
Coupling an implicit low-resolution model to an explicit high-resolution ocean model

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

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



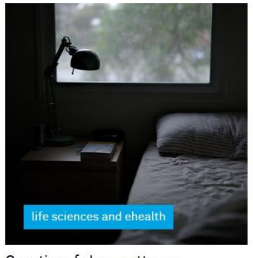
Inside the filter bubble
A framework for deep semantic analysis of mobile news consumption traces

SecConNet
Smart, secure container networks for trusted Big Data Sharing

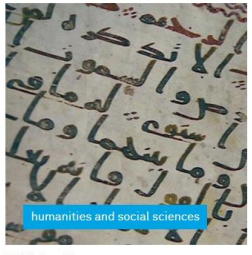
FEDMix
Fusible Evolutionary Deep Neural Network Mixture Learning from Distributed Data for Robust Medical

GlamMap
Visual Analytics for the World's Library Data

Deep learning OCR post-correction
Evaluation and post-correction of OCR of digitised historical newspapers



Genetics of sleep patterns
Detecting human sleep from wearable accelerometer data without the aid of sleep diaries

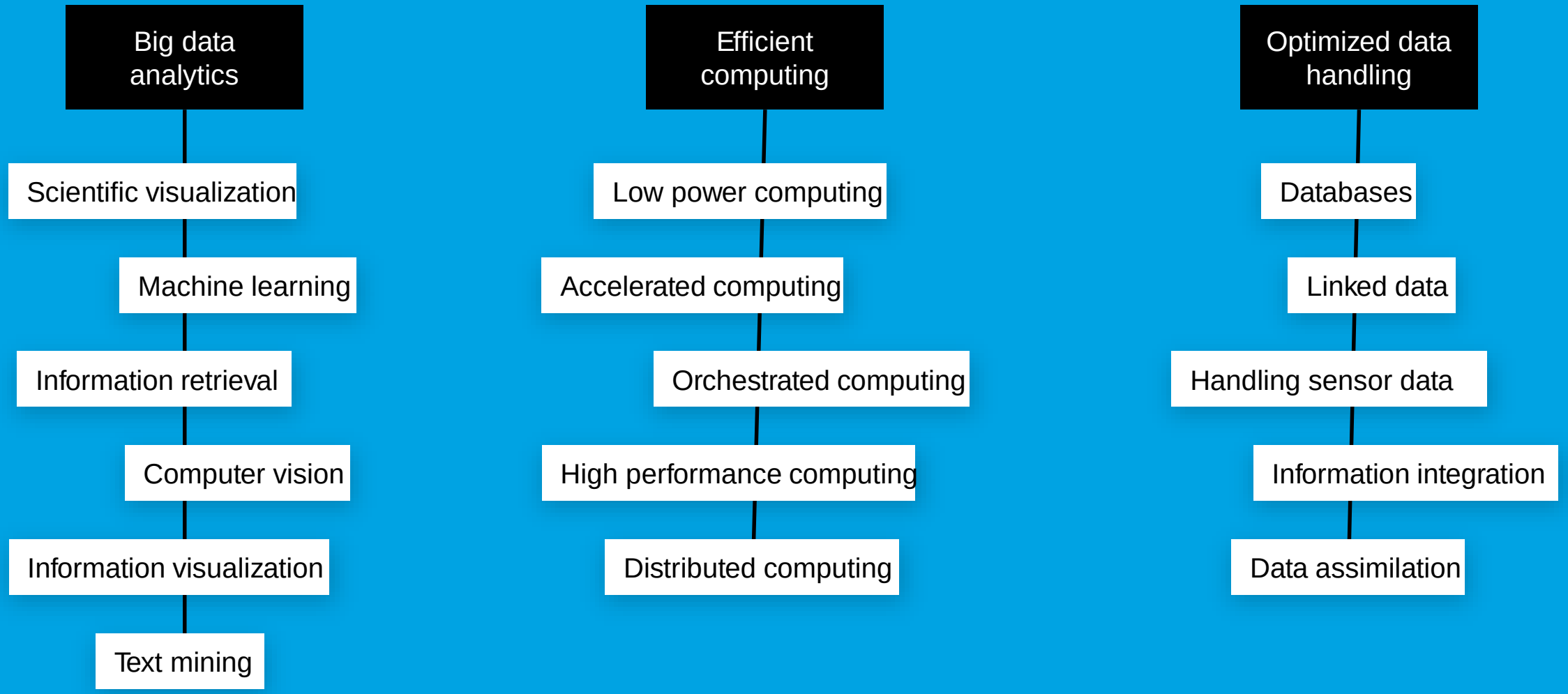


Bridging the gap
Digital Humanities and the Arabic-Islamic corpus

eEcoLiDAR
eScience infrastructure for Ecological applications of LiDAR point clouds

Emotion Recognition in Dementia
Advancing technology for multimodal analysis of emotion expression in

Our technological expertise areas



What do we do?

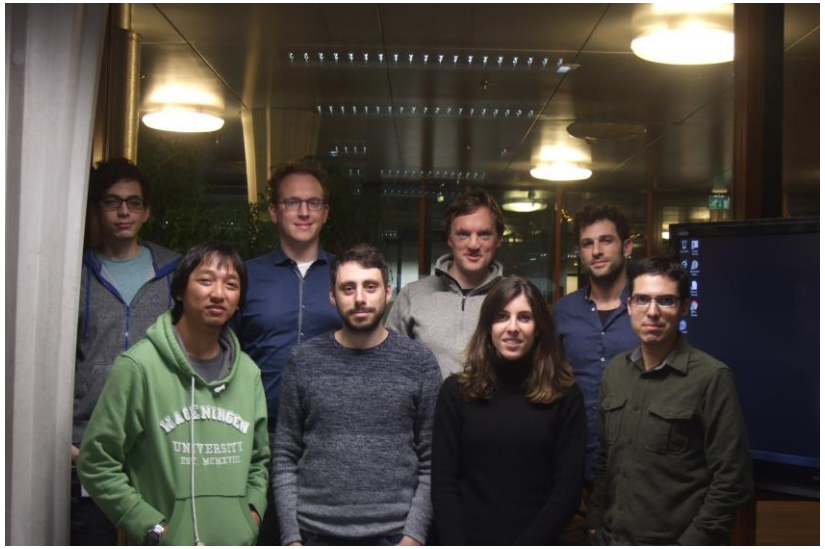
Research software

Link between researchers and IT infrastructure

Data stewards/data scientists

Cross-disciplinary transfer

Example project: Integrated ‘omics’ analysis

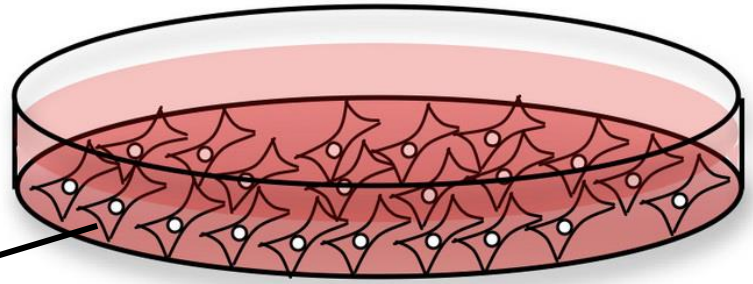


Medema lab - Wageningen UR, NL

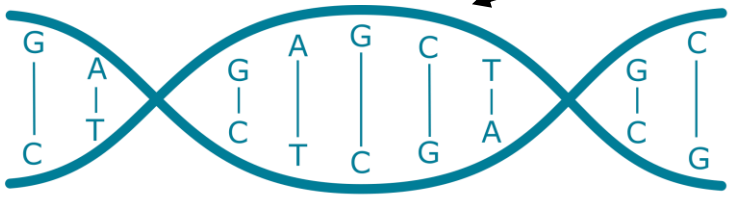
NL eScience Center

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

UCSD:
Madeleine Ernst
Pieter Dorrestein

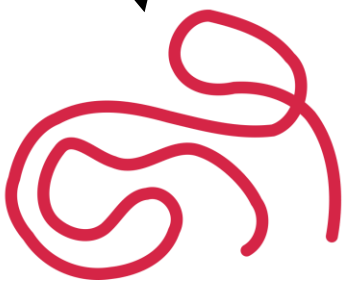


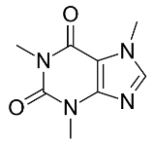
DNA



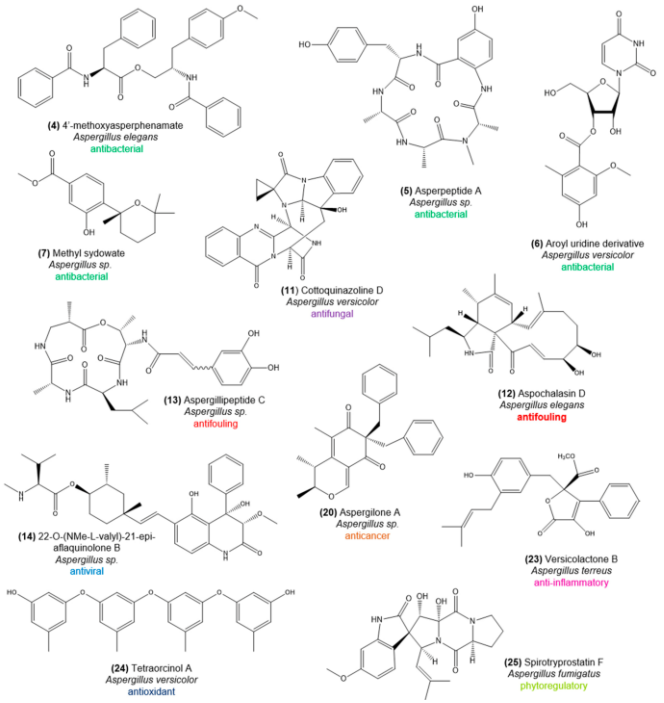
RNA

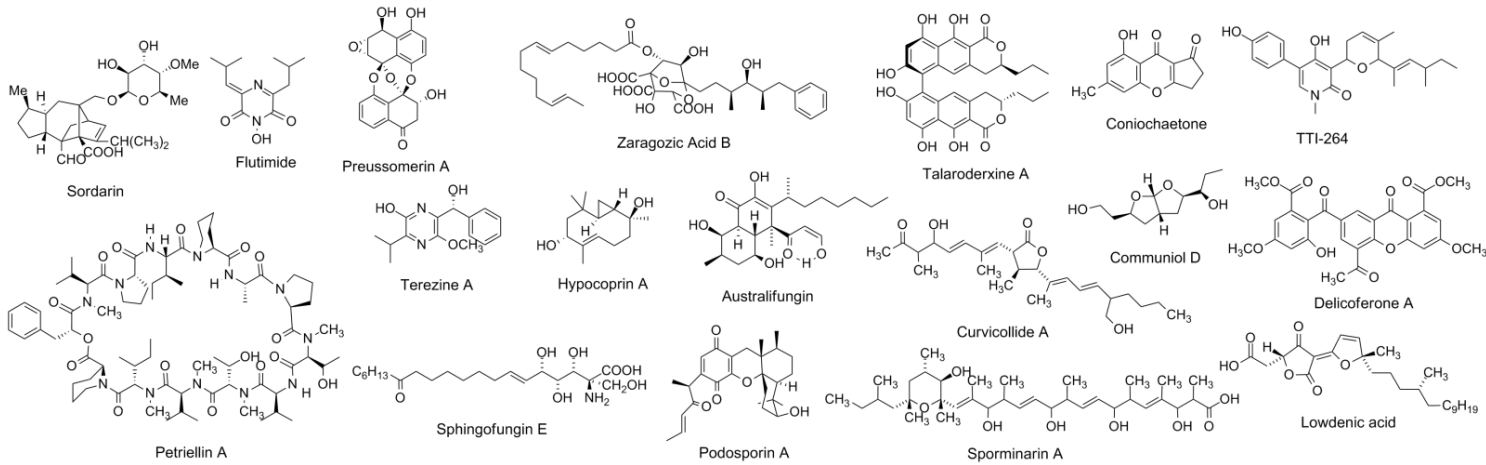
protein



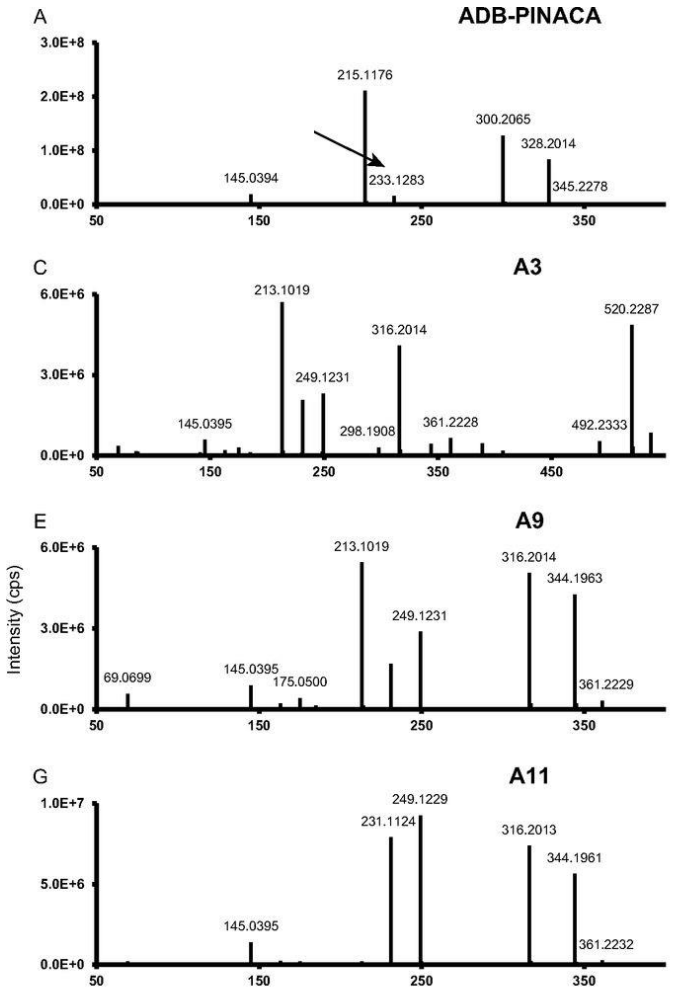


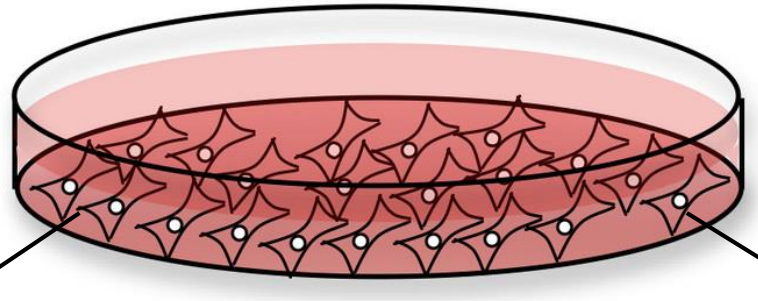
secondary metabolites



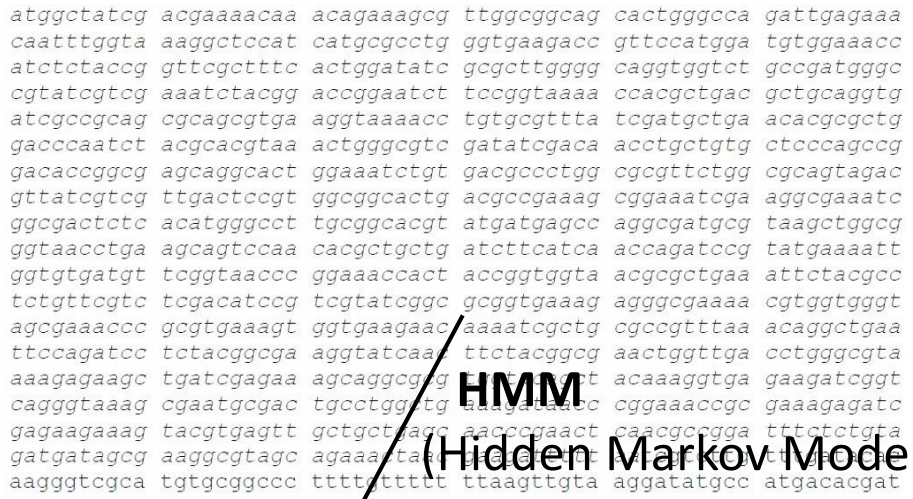


mass spectra

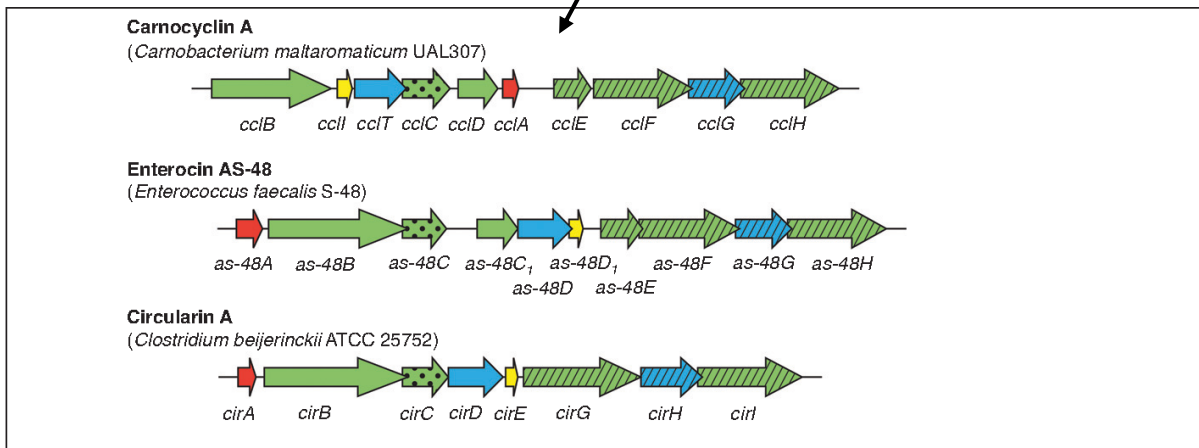


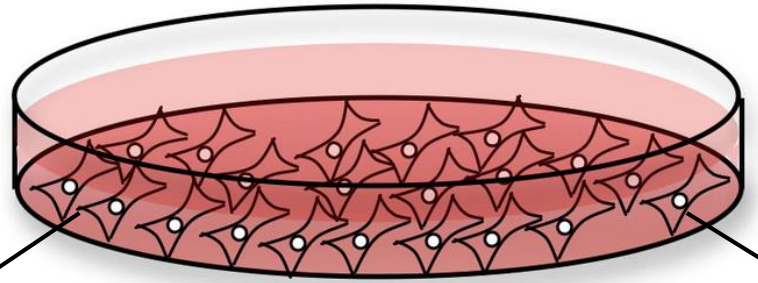


DNA



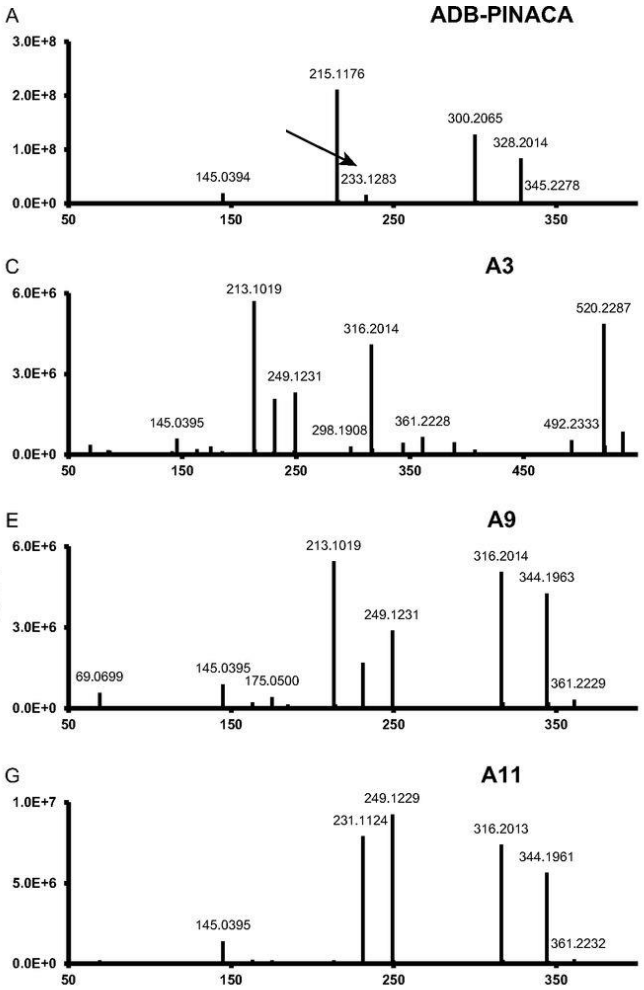
HMM
(Hidden Markov Model)
+ manually written rules

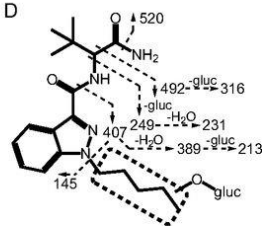


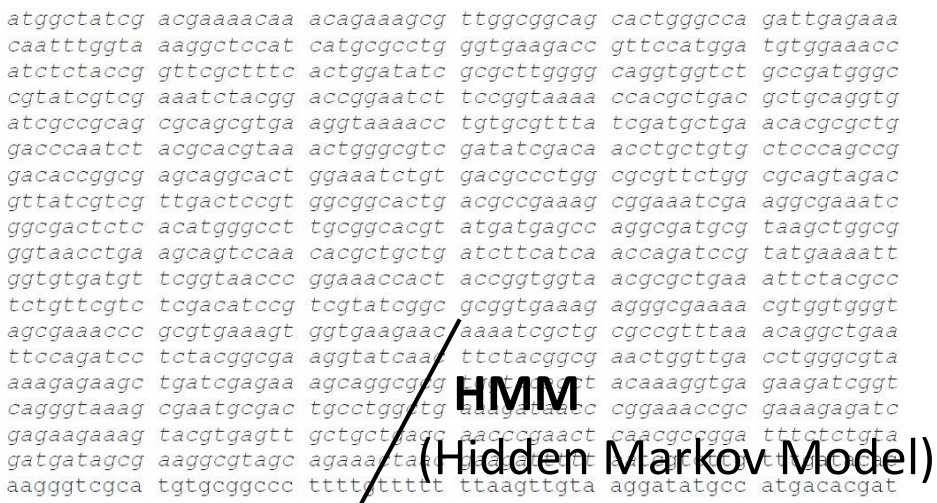


mass spectra

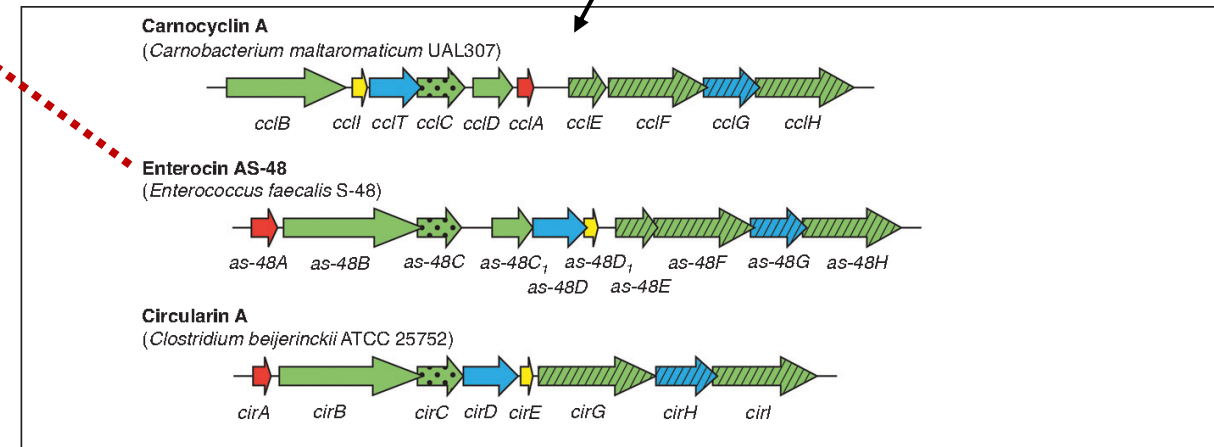
DNA

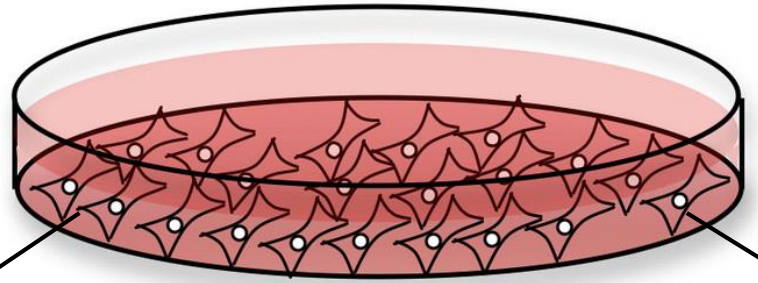






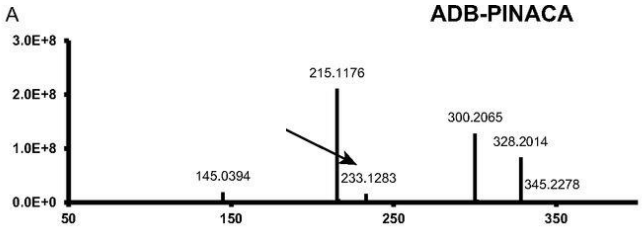
HMM
(Hidden Markov Model)
+ manually written rules

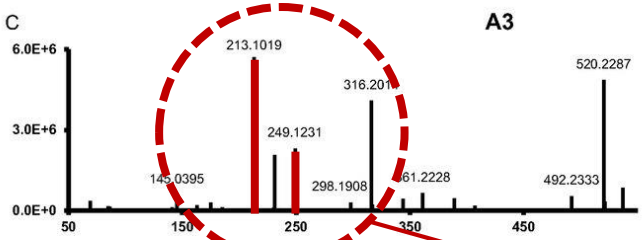


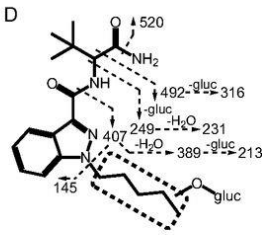


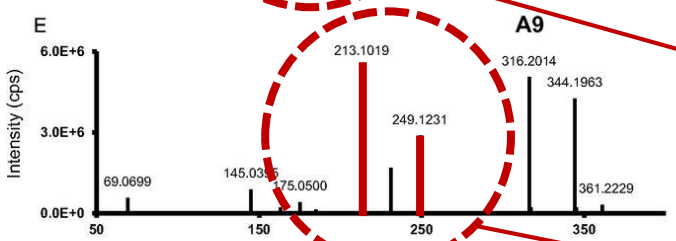
mass spectra

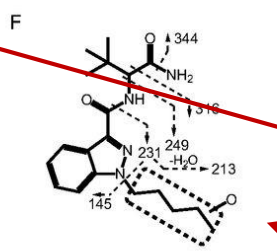
DNA

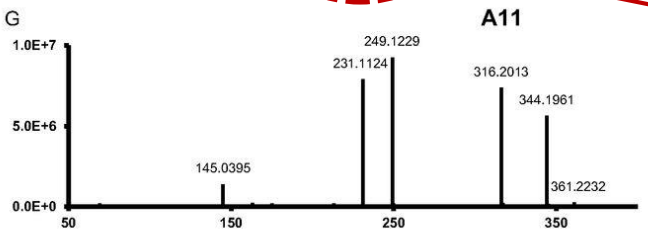




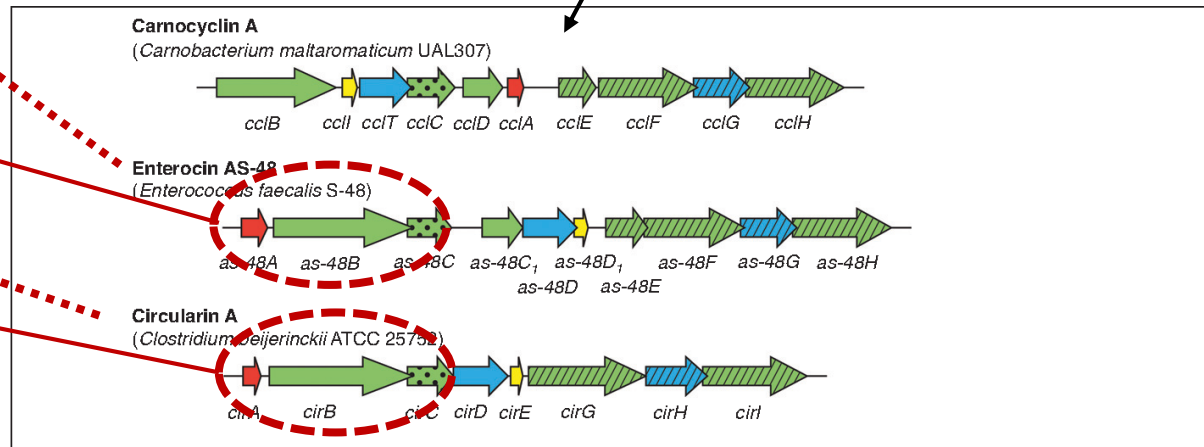




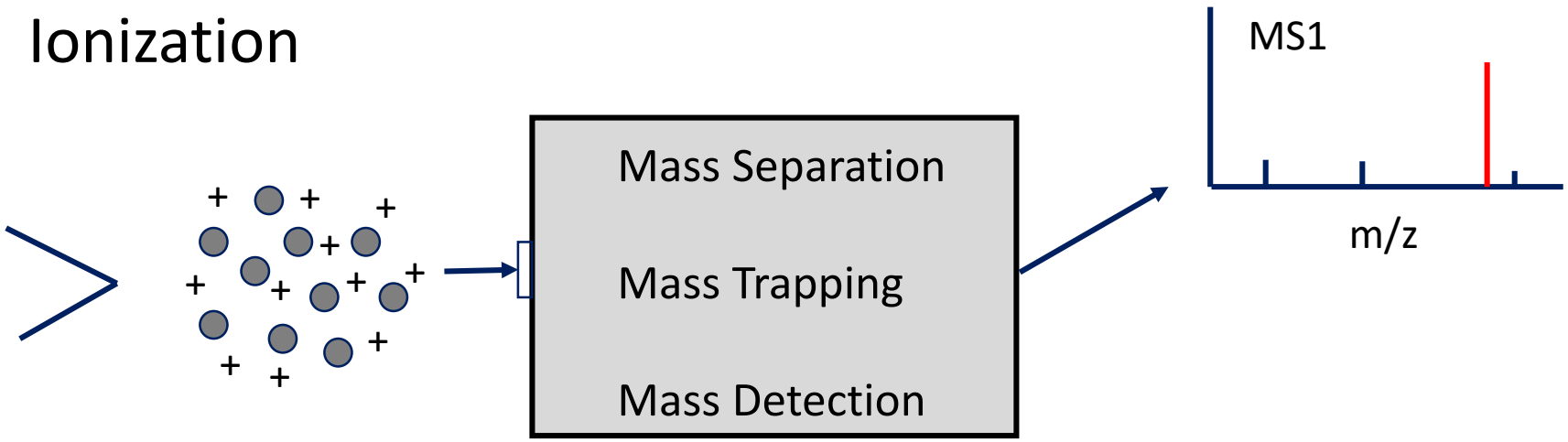




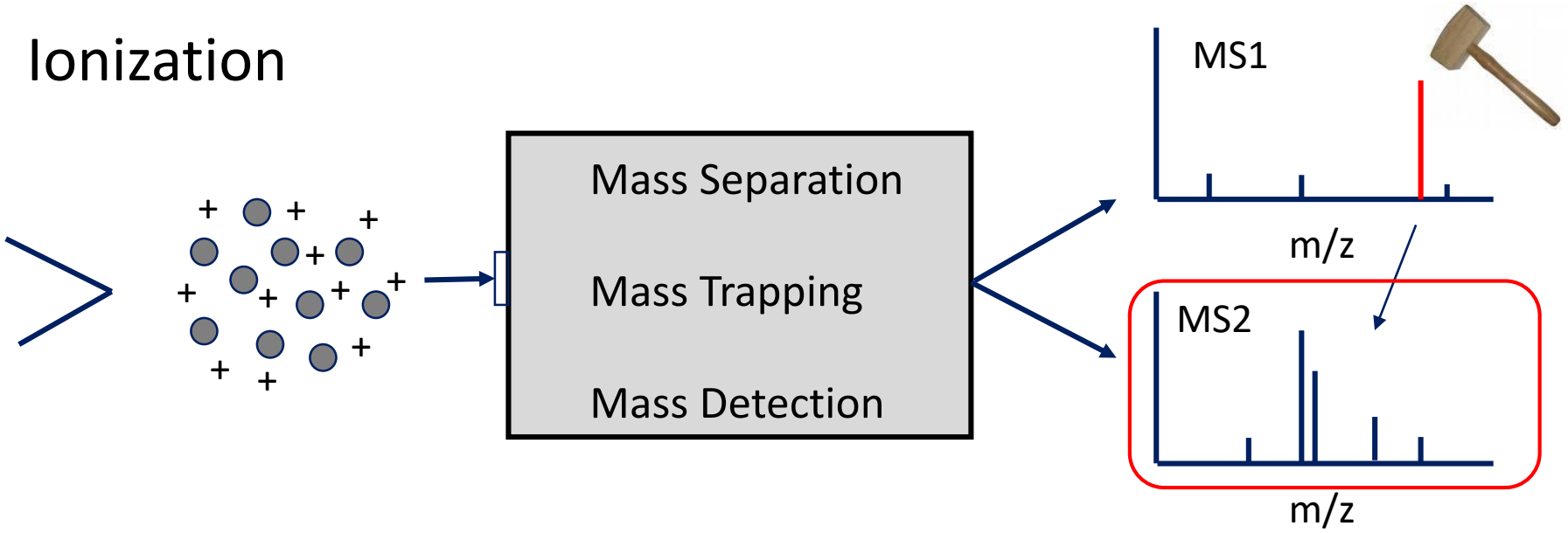
HMM
(Hidden Markov Model)
+ manually written rules



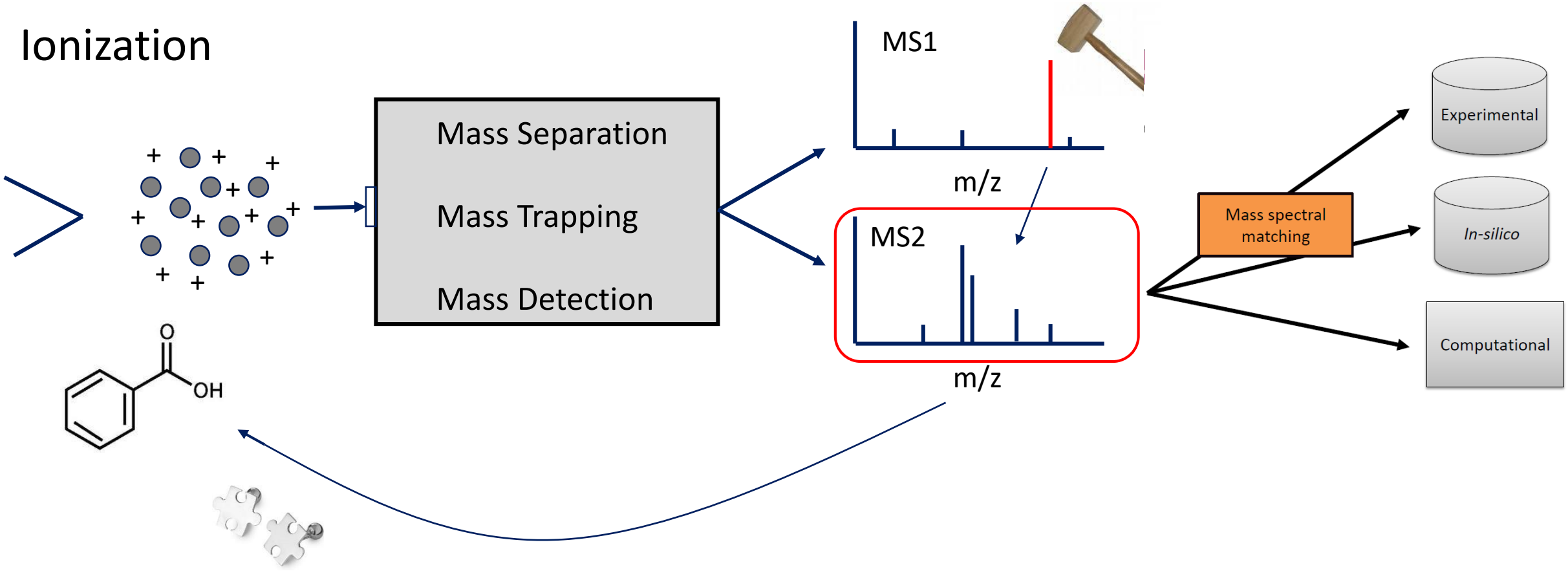
Mass spectrometry and fragmentation



Mass spectrometry and fragmentation

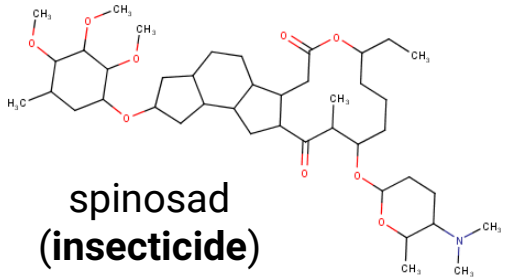


Mass spectrometry and fragmentation

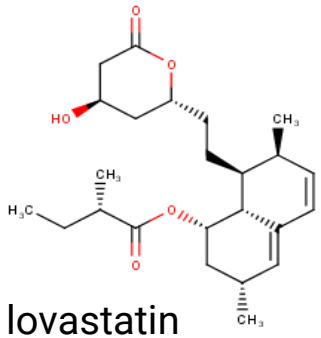


Fragments to puzzle the metabolite structure

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

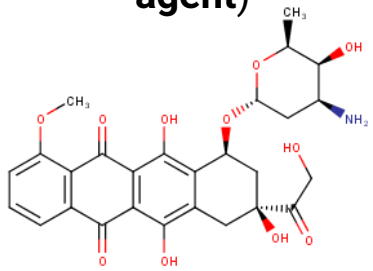


spinosad (**insecticide**)

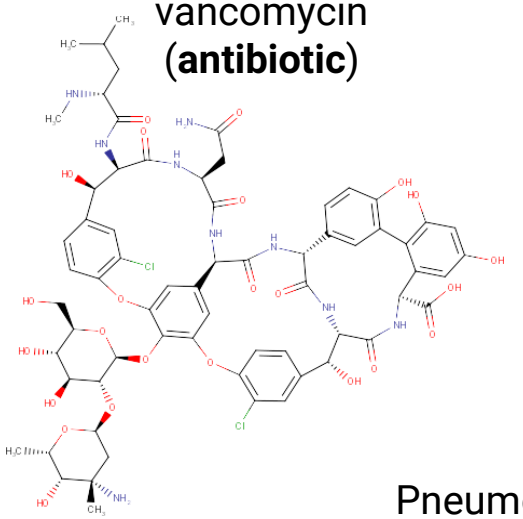


lovastatin (**cholesterol lowering agent**)

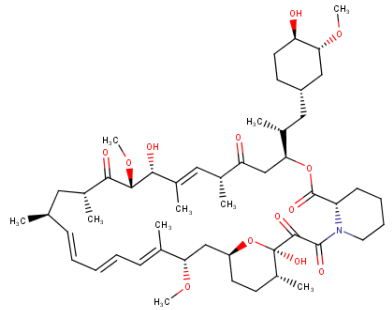
doxorubicin (**chemotherapeutic agent**)

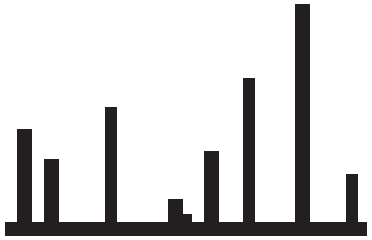


vancomycin (**antibiotic**)



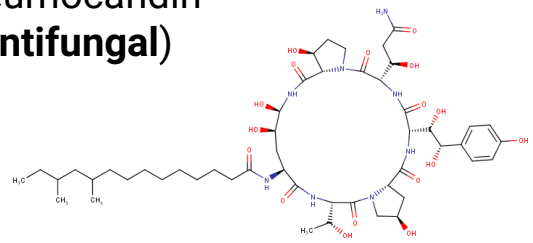
rapamycin (**immunosuppressant**)





Mass spectrometry fragmentation spectrum

Pneumocandin (**antifungal**)



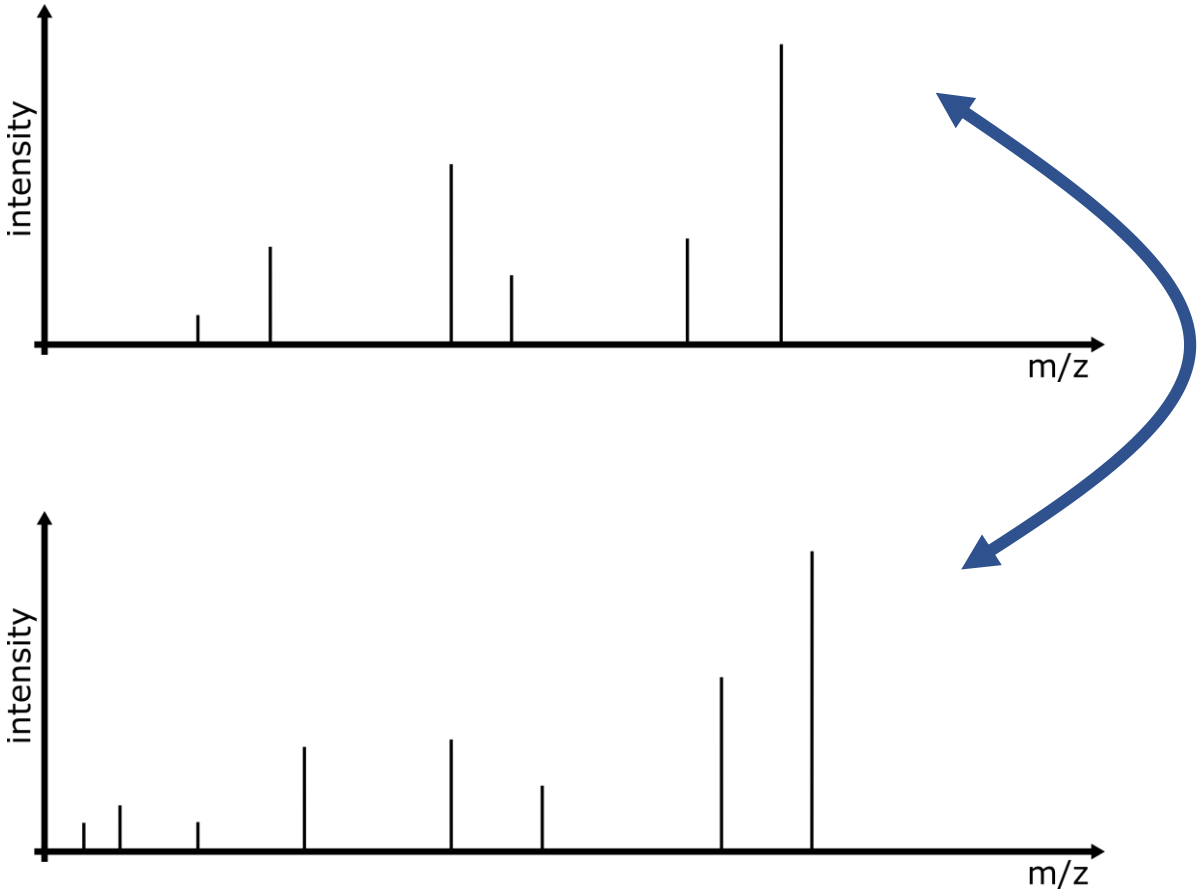
The challenge....

....is **large-scale coupling of spectral data to molecular structures**

of known & especially **novel** natural products molecules.

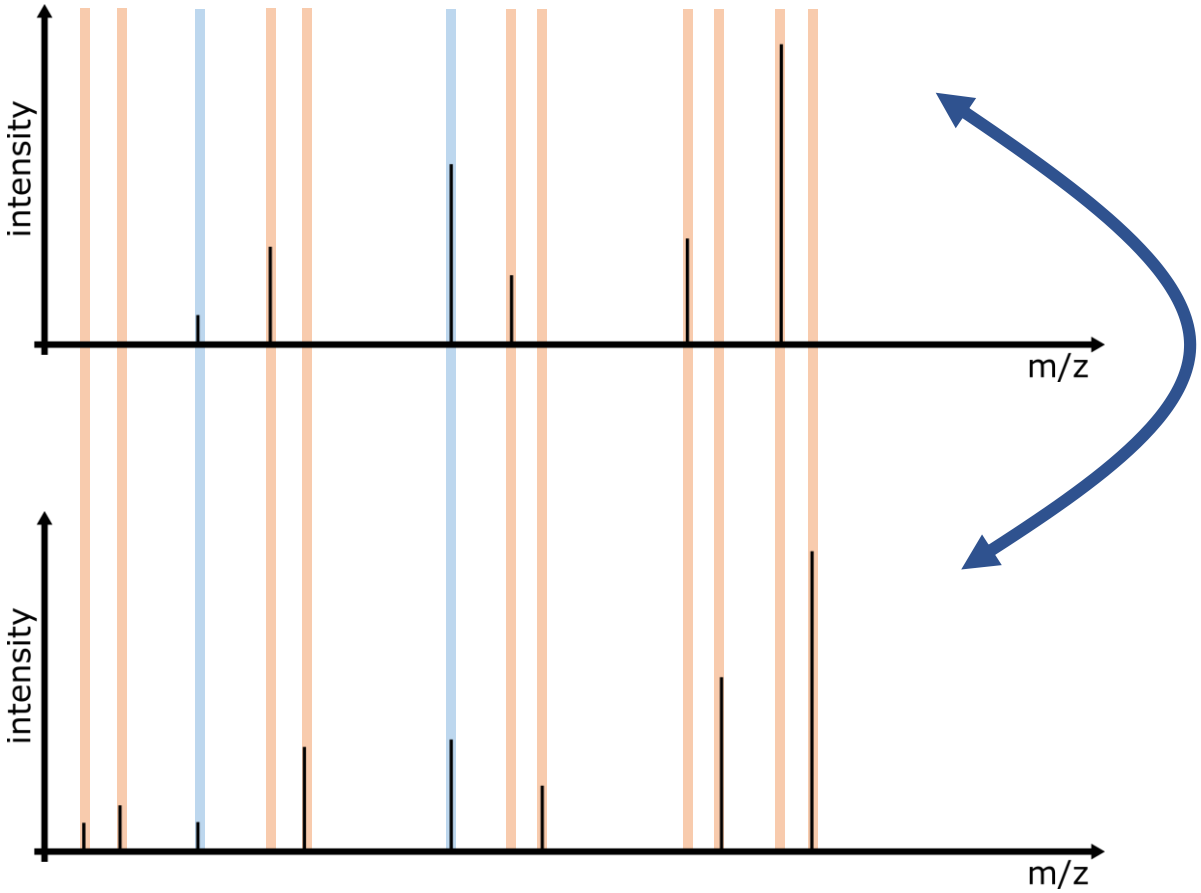
But…. How similar are they?

Spectral similarity



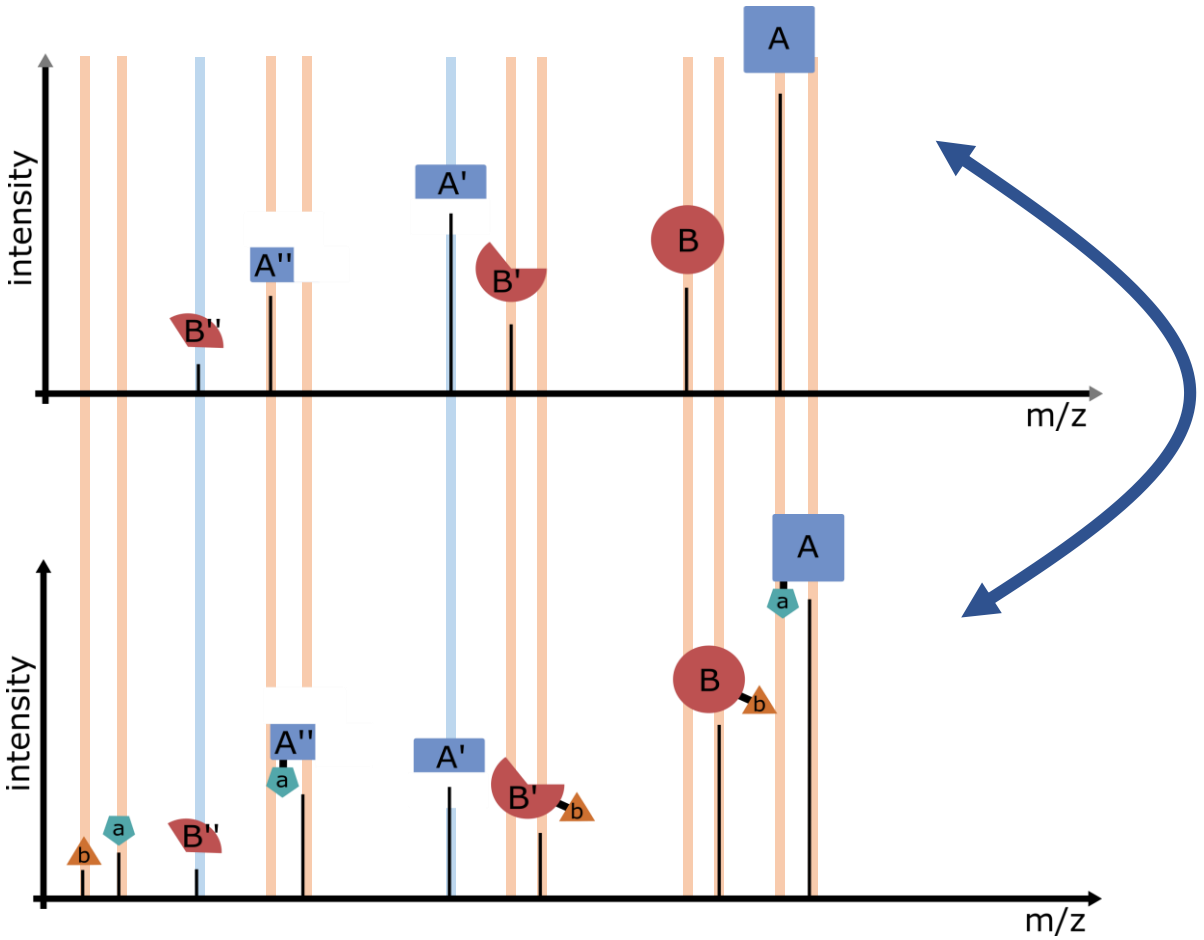
How similar are they?

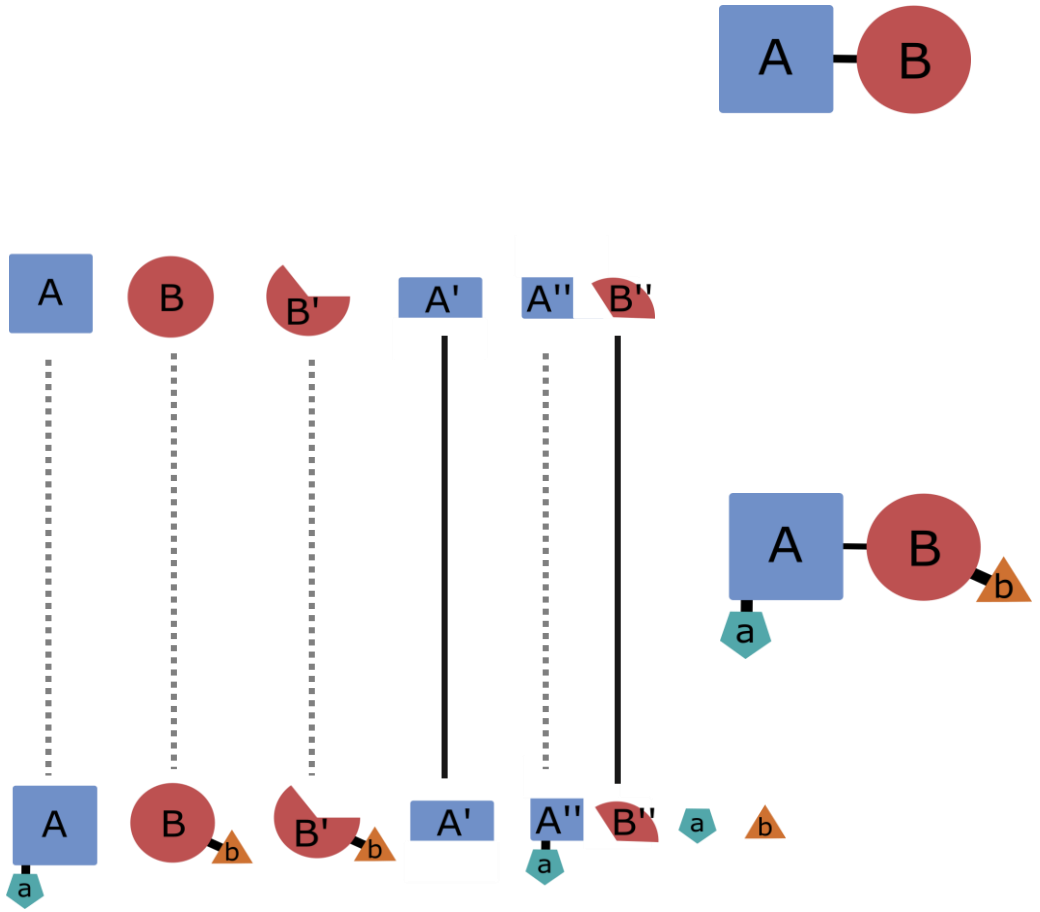
Spectral similarity



How similar are they?

Spectral similarity





How similar are they?

What does similar mean?

…likes cake with a cappuccino.

number of words?

number of characters?

…loves to have a cookie and a coffee.

grammatical structure?

topic?

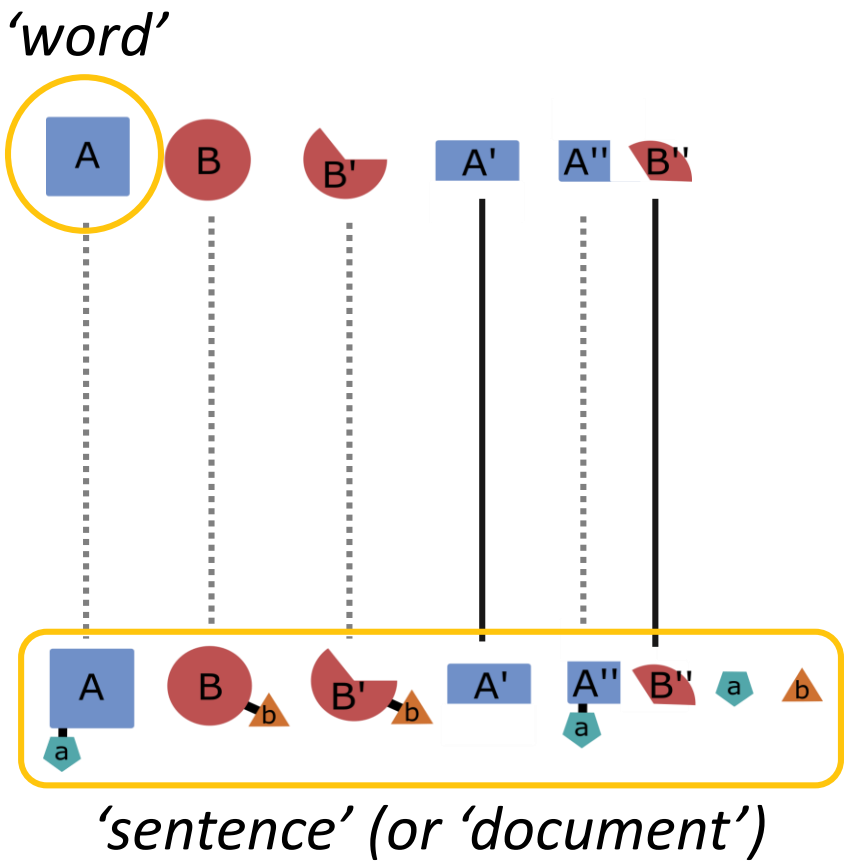
meaning?

style?

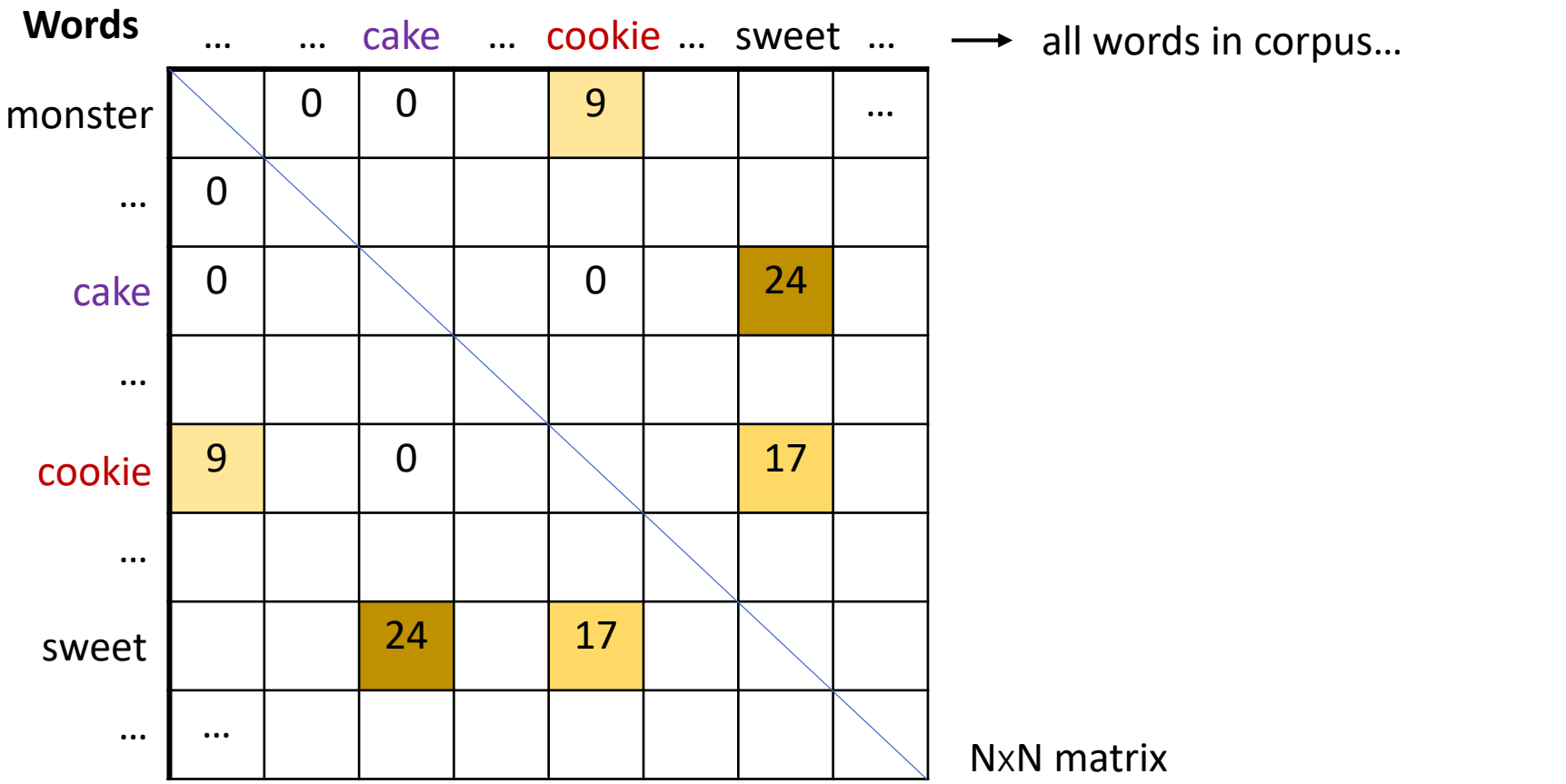
phonetic structure?

…likes cake with a cappuccino.

…loves to have a cookie and a coffee.



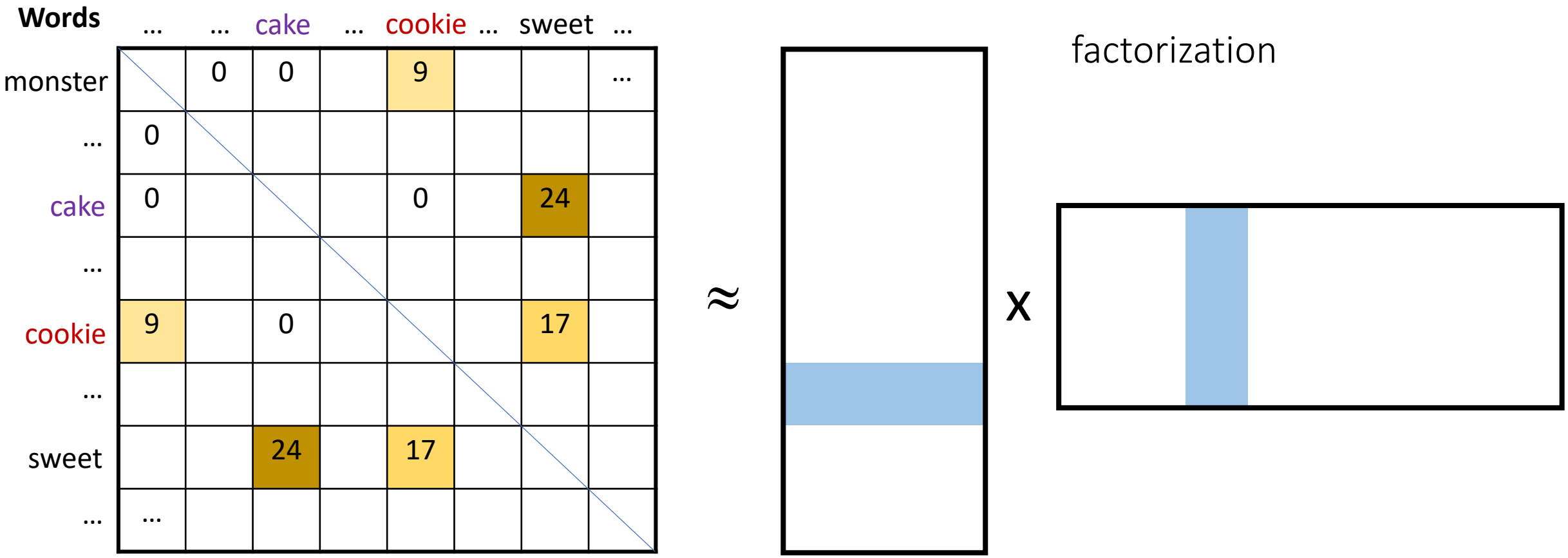
Count how often ‘words’ co-occur *(find word ‘context’)*



NxN matrix

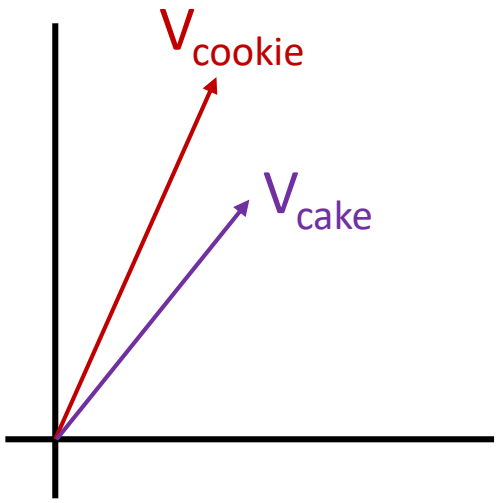
N: number of words in dictionary

‘Word2Vec’ → lower dimensional context vector



‘Word2Vec’ → lower dimensional context vector

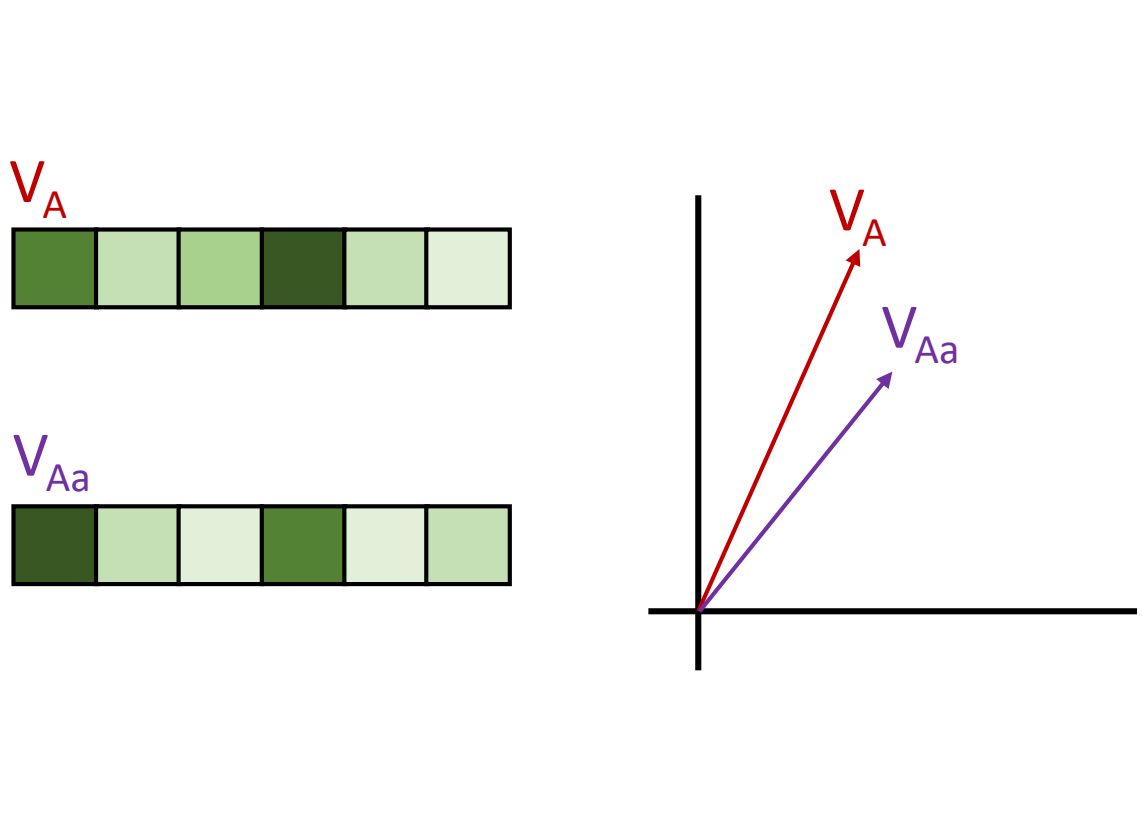
Words	…	…	cake	…	cookie	…	sweet	…
monster		0	0		9			…
…	0							
cake	0				0		24	
…								
cookie	9		0				17	
…								
sweet			24		17			
…	…							



NLP → metabolomics: use peaks as words

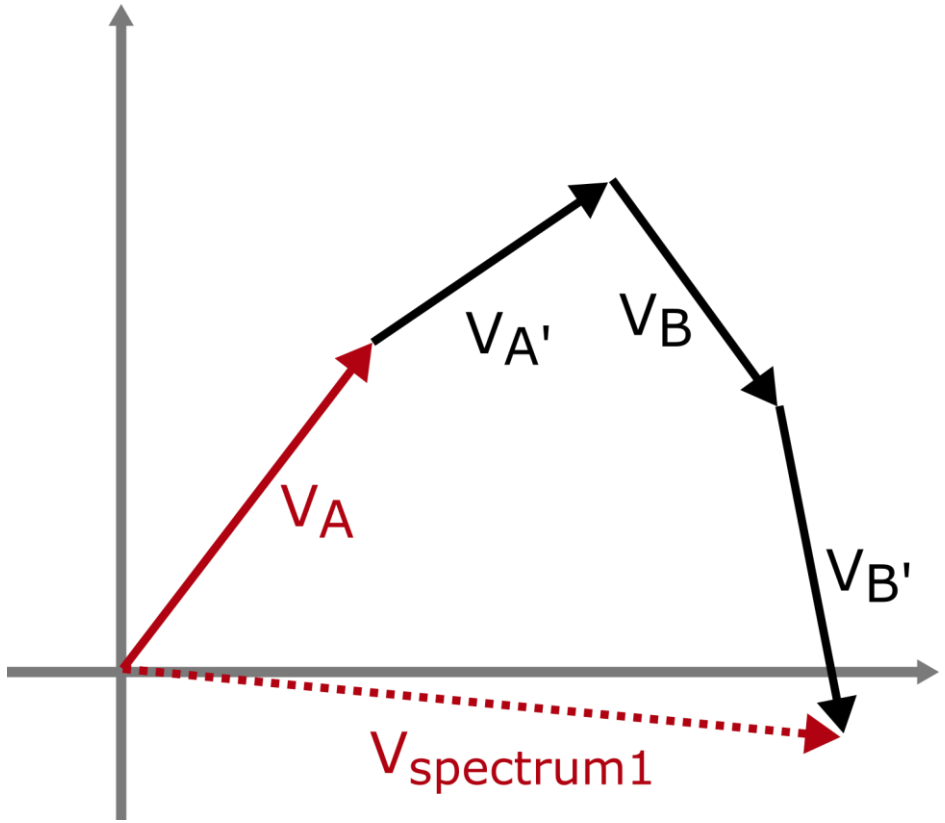
peak positions
='words'

	…	…	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

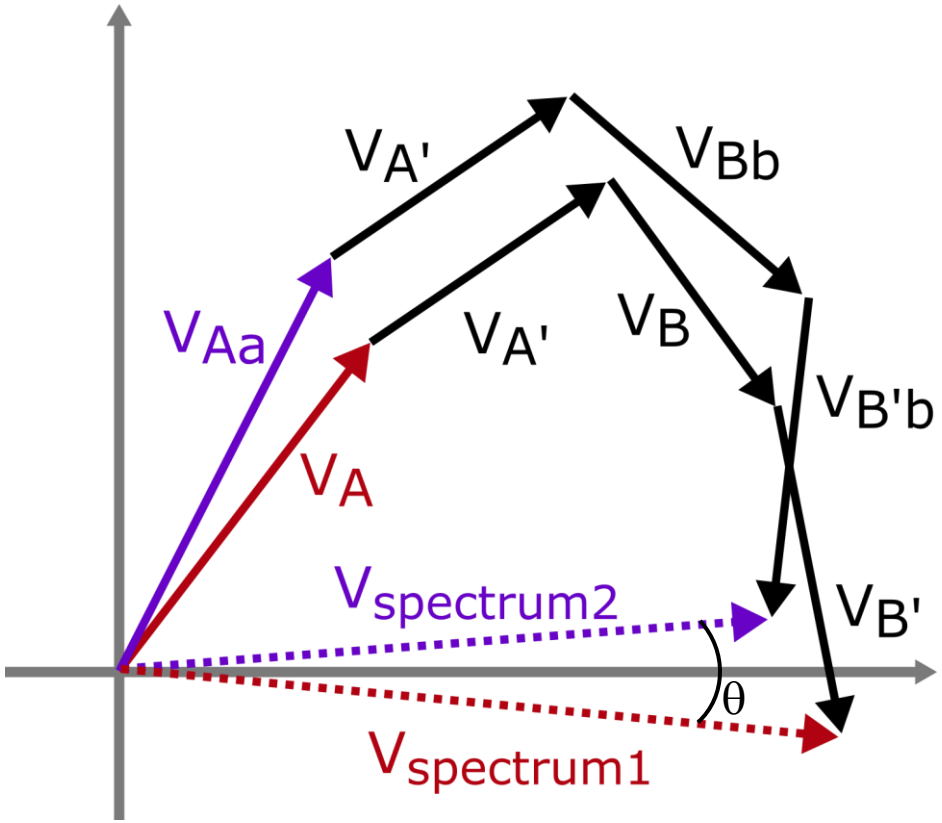


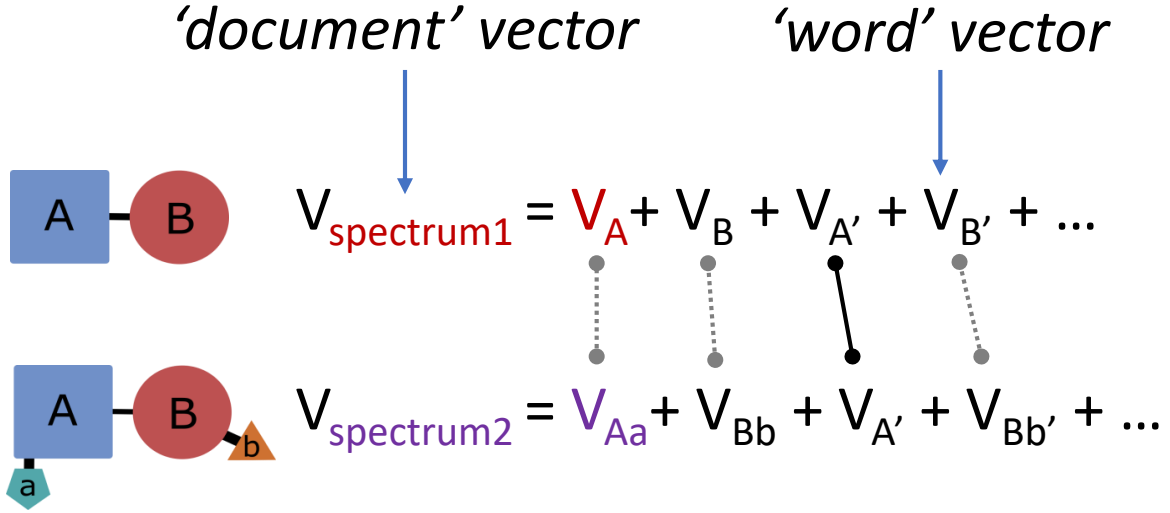
'document' vector *'word' vector*

$V_{spectrum1} = V_A + V_B + V_{A'} + V_{B'} + \ldots$

Spectral similarity measures.

NLP/word2vec based method

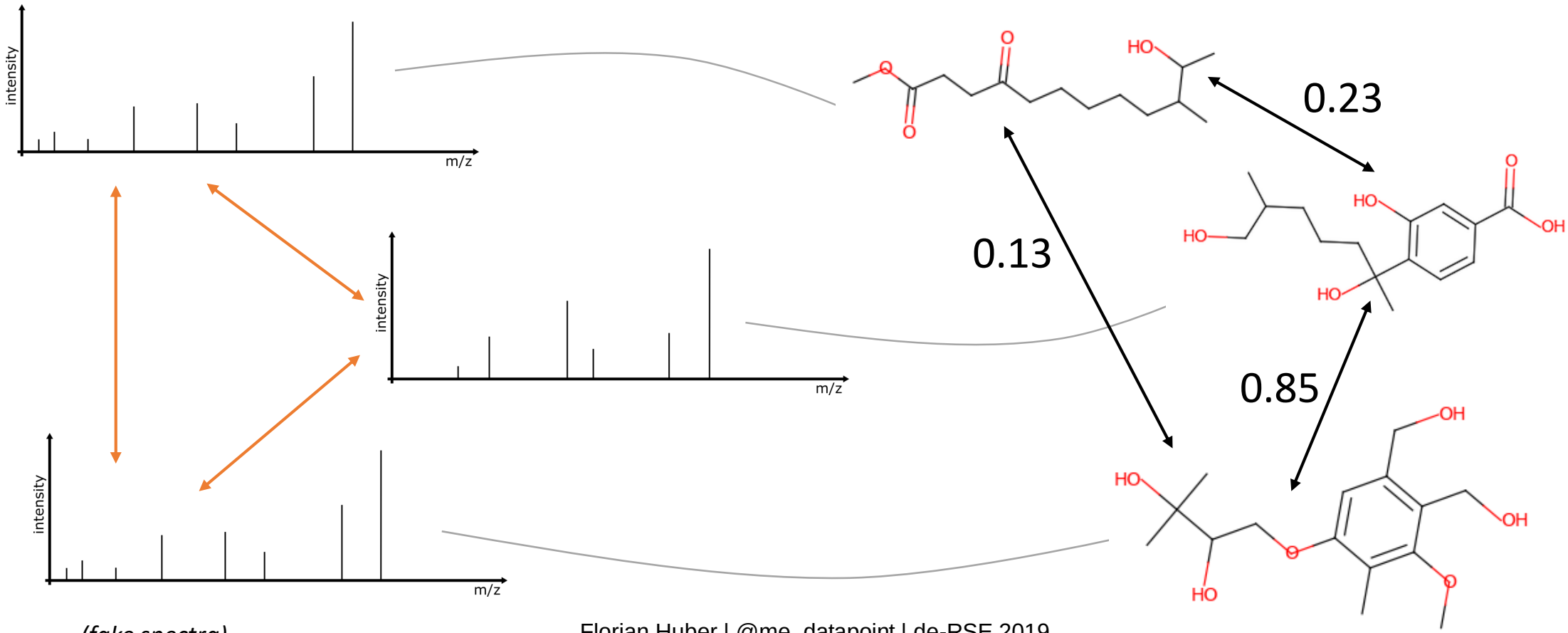




$$\text{Similarity} = \cos(\theta) = \frac{V_{spectrum1} \cdot V_{spectrum2}}{\|V_{spectrum1}\| \|V_{spectrum2}\|}$$

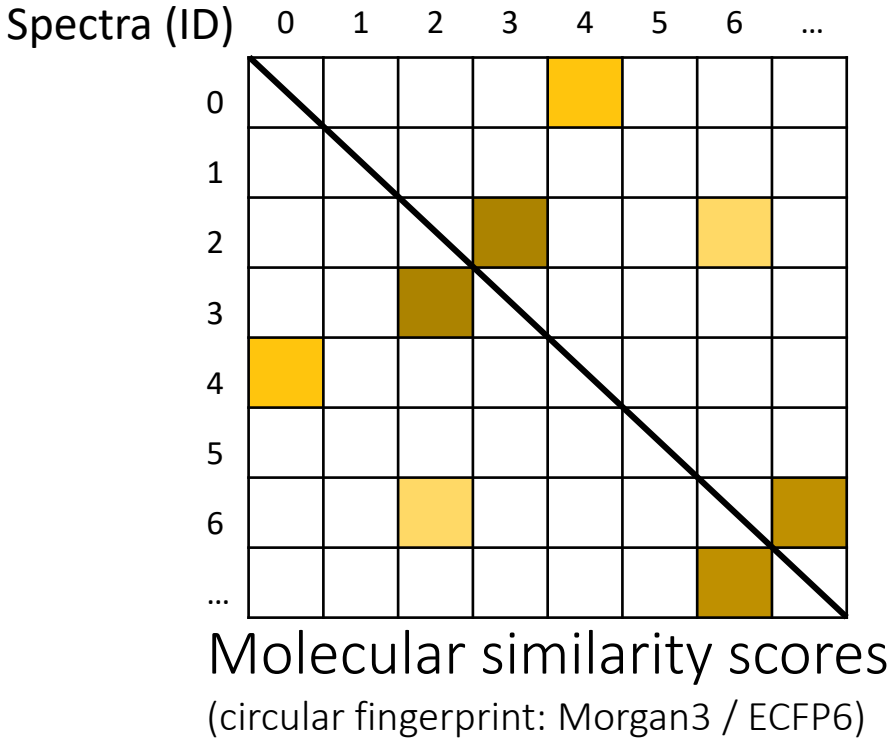
Spectral similarity measures: evaluation.

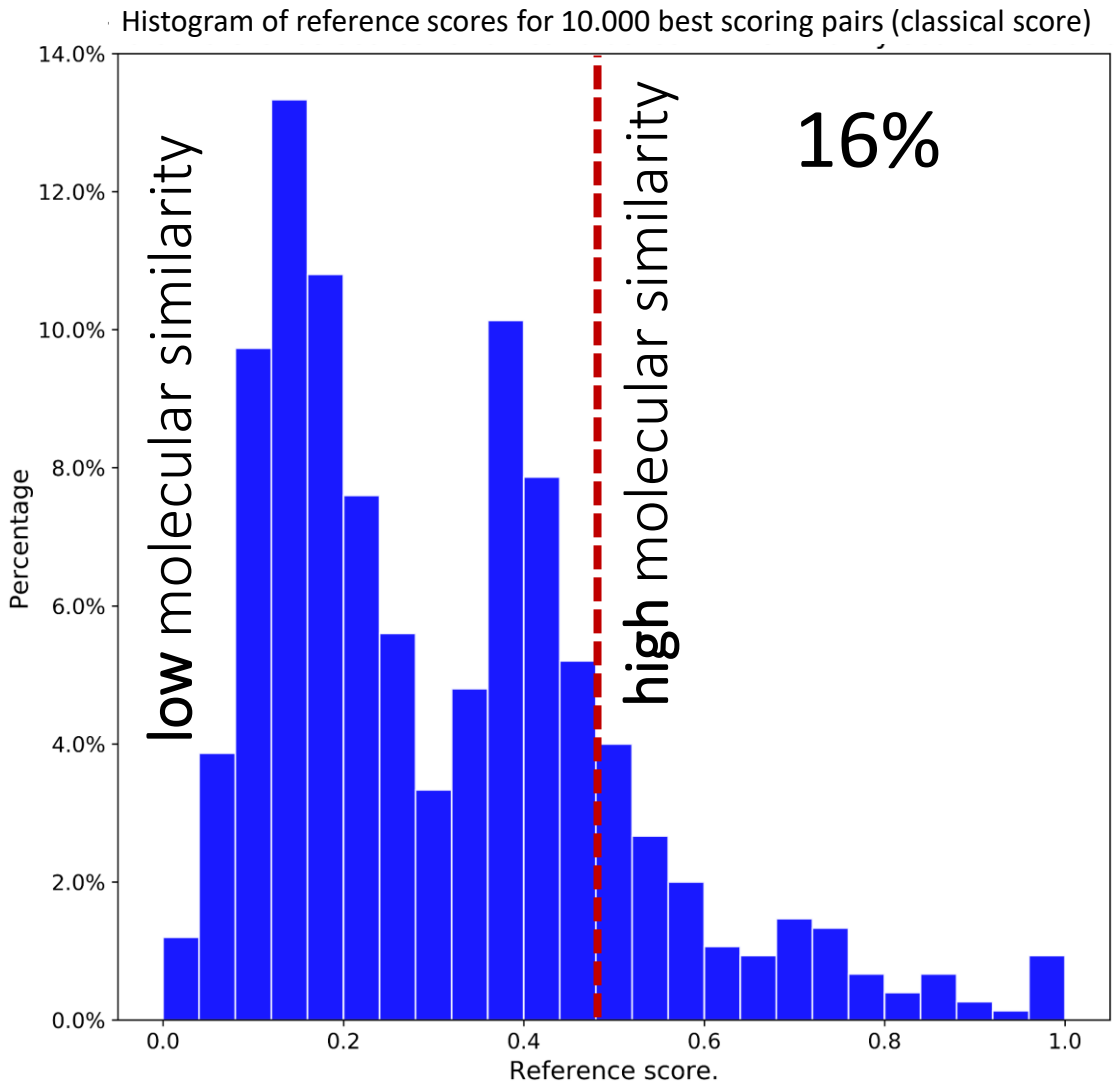
Dataset: 11.000 spectra with known molecular structures



(fake spectra)

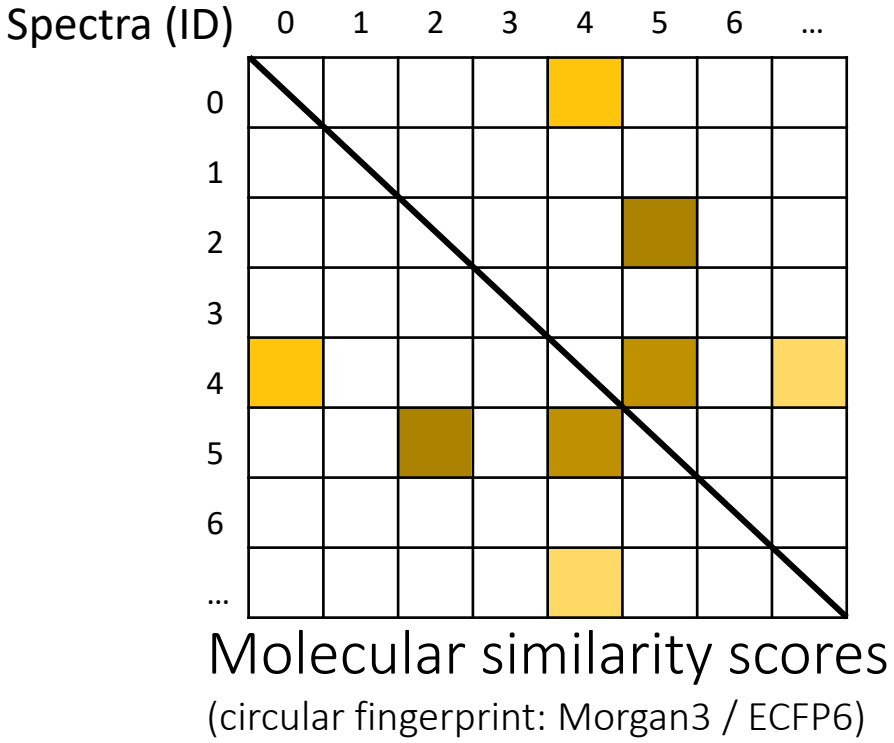
Molecular similarity scores:
10.000 highest ‘classical’ scores*

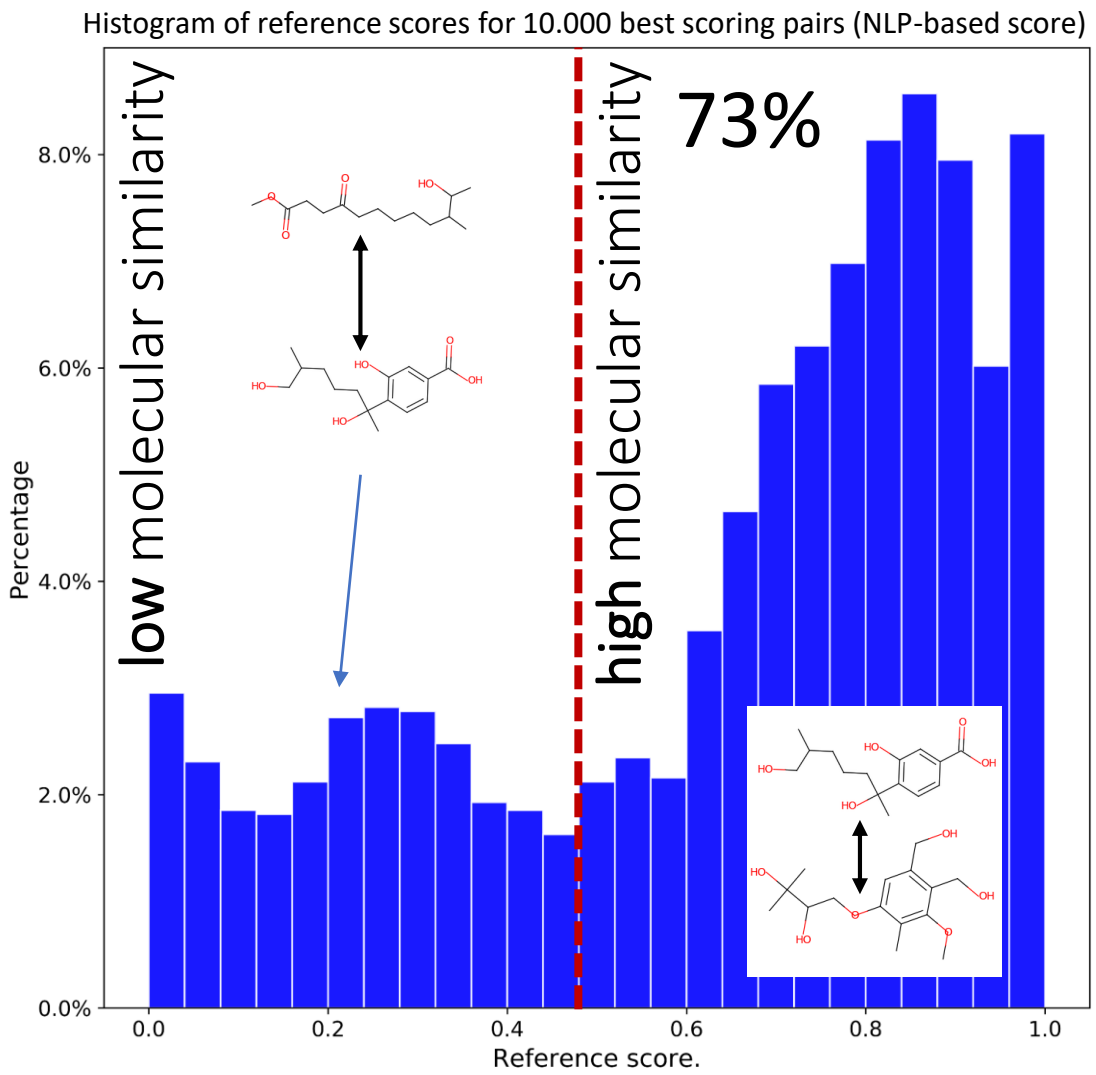




* = scores > 0.998

Molecular similarity scores:
10.000 highest NLP-based scores*

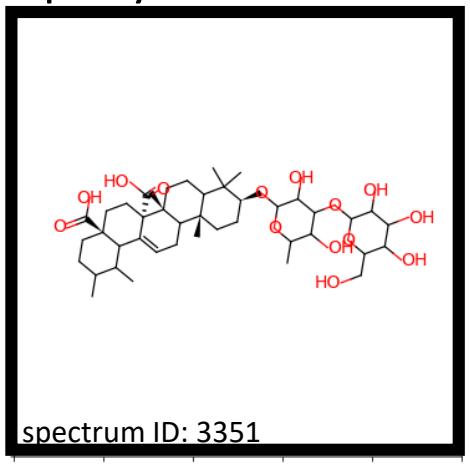




* = scores > 0.84

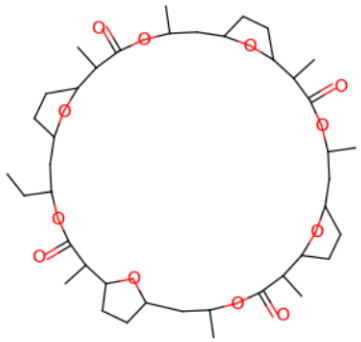
Spectral similarity measures: bad examples.

query molecule

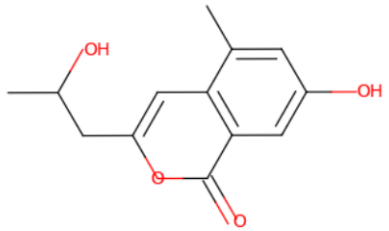


spectrum ID: 3351

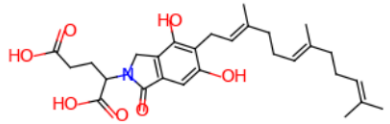
9 closest candidates (according to molecular networking similarity)



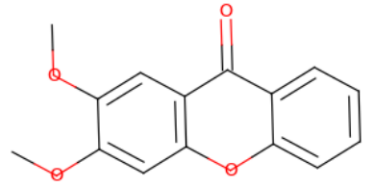
1



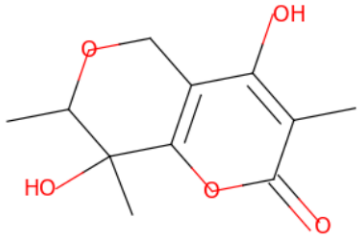
2



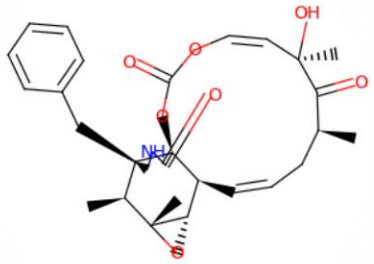
3



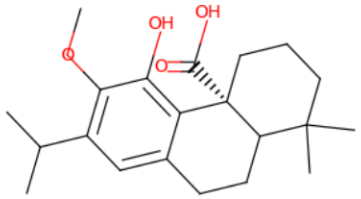
4



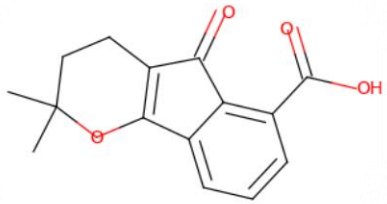
5



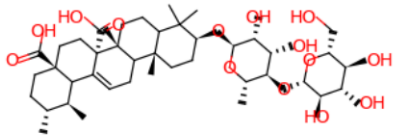
6



7



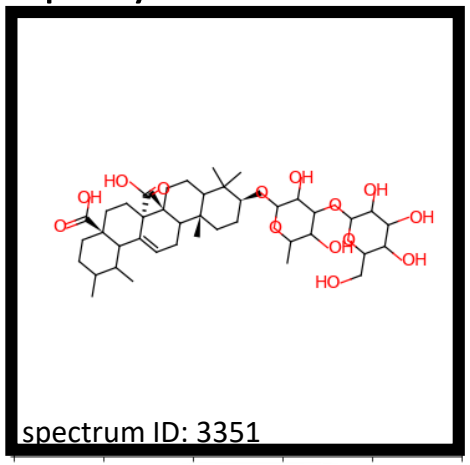
8



9

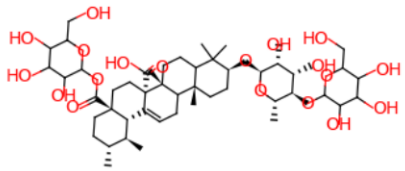
Spectral similarity measures: examples.

query molecule

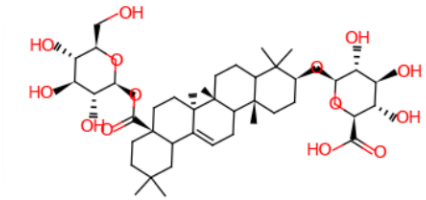


spectrum ID: 3351

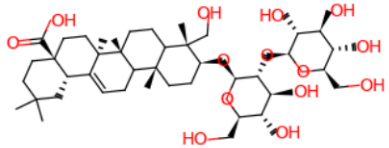
9 closest candidates (according to Word2vec-based spectral similarity)



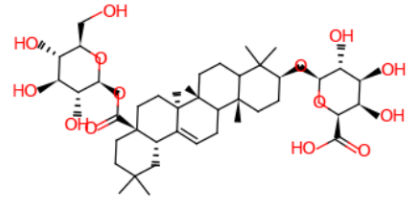
1



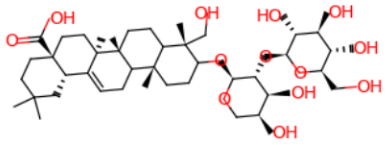
2



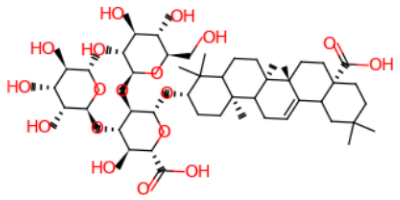
3



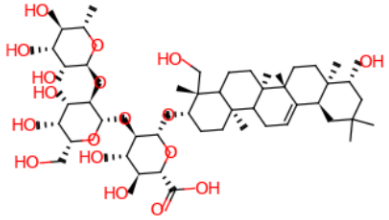
4



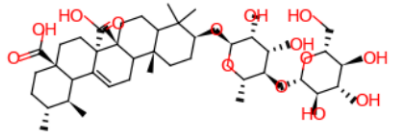
5



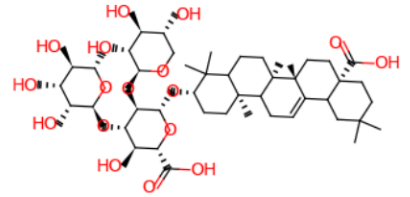
6



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