



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

Florian Huber<sup>1</sup>, Justin van der Hooft<sup>2</sup>, Simon Rogers<sup>3</sup>, Marnix Medema<sup>2</sup>, Lars Ridder<sup>1</sup>

<sup>1</sup> Netherlands eScience Center

<sup>2</sup> Bioinformatics Group, Wageningen University

<sup>3</sup> School of Computing Science, University of Glasgow



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

Florian Huber<sup>1</sup>, Justin van der Hooft<sup>2</sup>, Simon Rogers<sup>3</sup>, Marnix Medema<sup>2</sup>, Lars Ridder<sup>1</sup>

<sup>1</sup> Netherlands eScience Center

<sup>2</sup> Bioinformatics Group, Wageningen University

<sup>3</sup> 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



netherlands eScience center

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

---

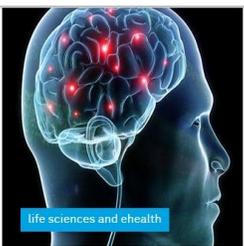


Photography: Elodie Burrillon



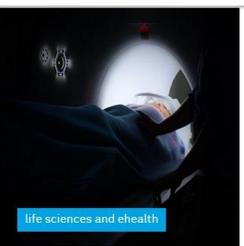
sustainability and environment

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



life sciences and ehealth

**TADPOLE-SHARE**  
SHaring TADPOLE's Algorithms for Reuse and Evaluation



life sciences and ehealth

**DTL Semantic Analysis of radiology Reports utilizing Lexicon**  
Unlocking large volumes of knowledge locked in natural text



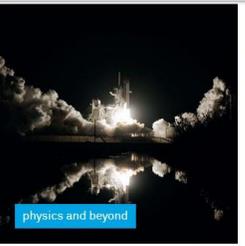
sustainability and environment

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



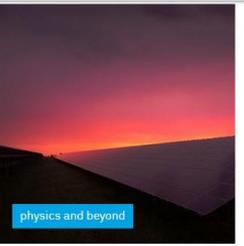
life sciences and ehealth

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



physics and beyond

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



physics and beyond

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



physics and beyond

**Parallel-in-time methods for the propagation of uncertainties in wind-farm simulations**  
Studying uncertainties in large eddy simulations of wind farms



humanities and social sciences

**TICLCLAT**  
Text-Induced Corpus Correction and Lexical Assessment Tool



humanities and social sciences

**NEWSGAC**  
Advancing Media History by Transparent Automatic Genre Classification



sustainability and environment

**European Climate Prediction system**



humanities and social sciences

**ePODIUM**  
Early Prediction of Dyslexia in Infants Using Machine learning



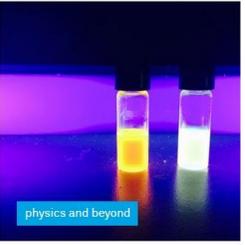
humanities and social sciences

**Understanding visually grounded spoken language via multi-tasking**  
An alternative approach for intelligent systems to understand human speech



sustainability and environment

**Monitoring tropical forest recovery capacity using RADAR Sentinel satellite data**  
Demonstrating the potential of European Sentinel satellite data



physics and beyond

**eScience Technology to Boost Quantum Dot Energy Conversion**  
More efficient lighting and solar energy conversion devices



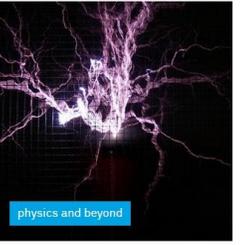
physics and beyond

**A light in the dark**  
Quantum Monte Carlo meets solar energy conversion



physics and beyond

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



physics and beyond

**A phase field model to guide the development and design of next generation solid-state-batteries**  
Safer batteries with higher energy densities



sustainability and environment

**Data mining tools for abrupt climate change**  
Updating our knowledge on abrupt climate change



humanities and social sciences

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



sustainability and environment

**MOSAIC**  
MOdelling Sea level And Inundation for Cyclones



eScience methodology

**PROCESS**  
PROviding Computing solutions for ExaScale ChallengeS



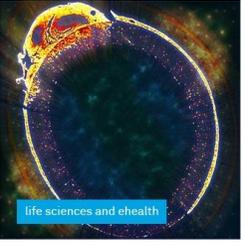
eScience methodology

**Enhance Your Research Alliance (EYRA) Benchmark Platform**



humanities and social sciences

**Uncovering Networks of Corporate Control**  
An interactive web-based platform to investigate the dynamics of global corporate networks



life sciences and ehealth

**Integrated omics analysis for small molecule-mediated host-microbiome interactions**  
Advancing our understanding of molecular mechanisms of health and



physics and beyond

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



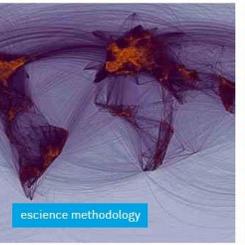
sustainability and environment

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



sustainability and environment

**MAGIC**  
Metrics and Access to Global Indices for Climate Projections



eScience methodology

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



eScience methodology

**High spatial resolution phenological modelling at continental scales**  
Understanding phenological variability



sustainability and environment

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



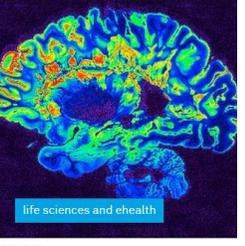
humanities and social sciences

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



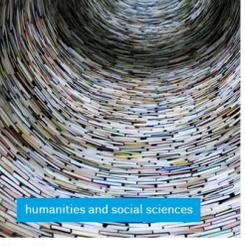
eScience methodology

**SecConNet**  
Smart, secure container networks for trusted Big Data Sharing



life sciences and ehealth

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



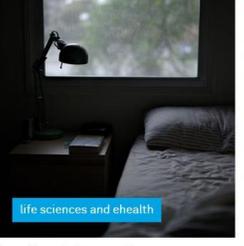
humanities and social sciences

**GlamMap**  
Visual Analytics for the World's Library Data



humanities and social sciences

**Deep learning OCR post-correction**  
Evaluation and post-correction of OCR of digitised historical manuscripts



life sciences and ehealth

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



humanities and social sciences

**Bridging the gap**  
Digital Humanities and the Arabic-Islamic corpus



sustainability and environment

**eEcoLiDAR**  
eScience infrastructure for Ecological applications of LiDAR point clouds

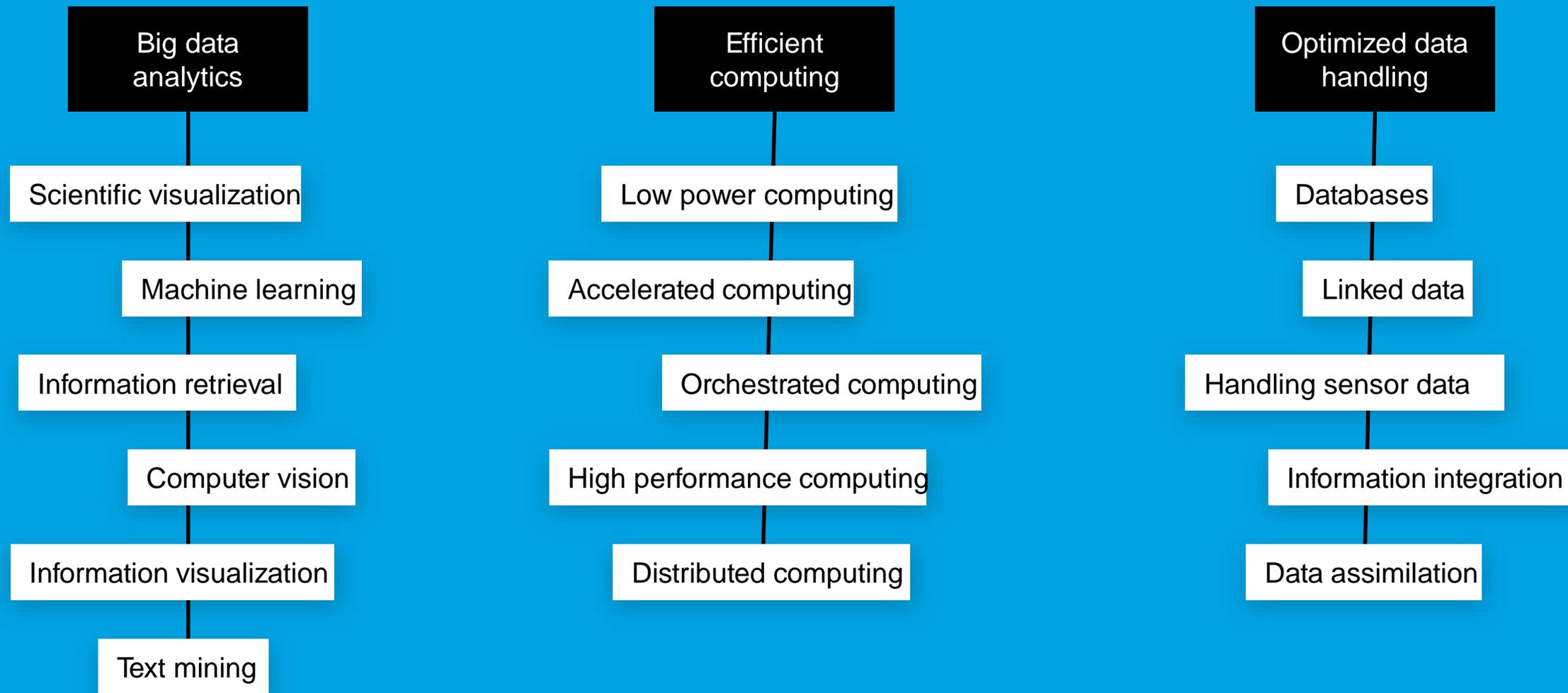


humanities and social sciences

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

# Our technological expertise areas

---



# What do we do?

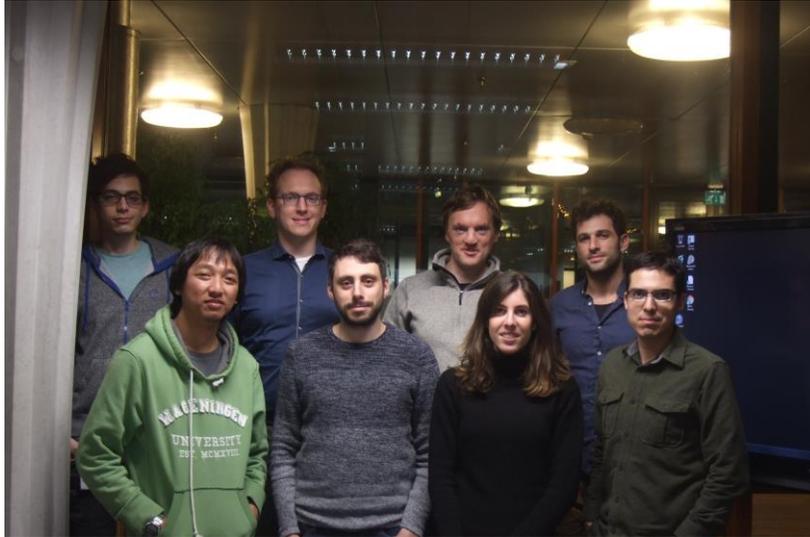
Research software

Link between researchers and IT infrastructure

Data stewards/data scientists

Cross-disciplinary transfer

# Example project: Integrated 'omics' analysis



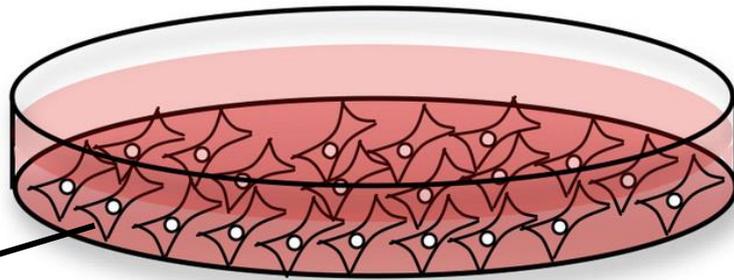
Medema lab - Wageningen UR, NL

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

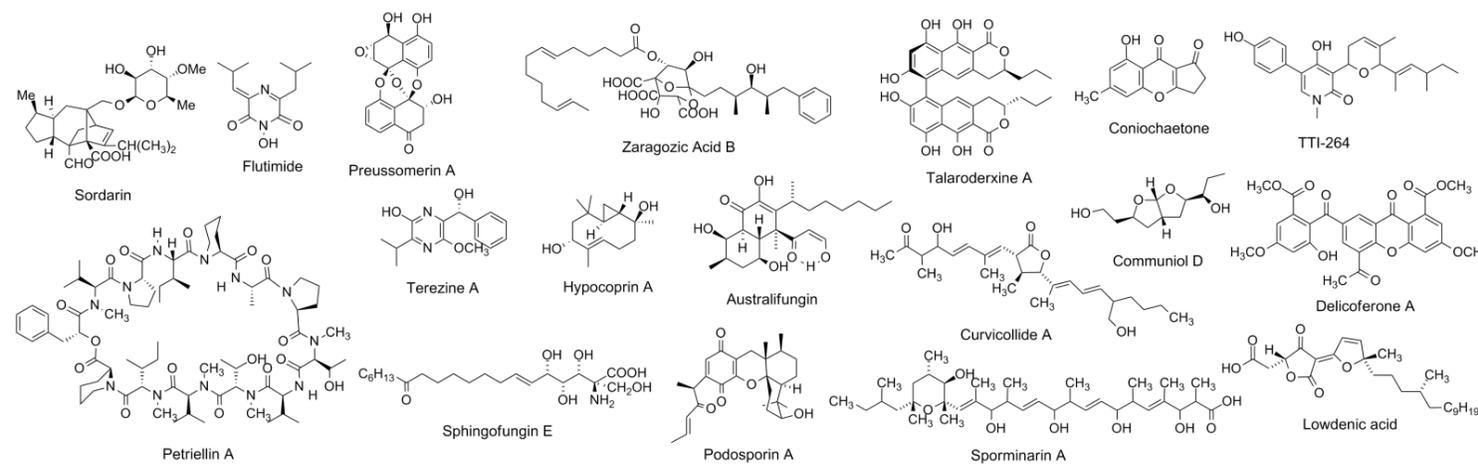
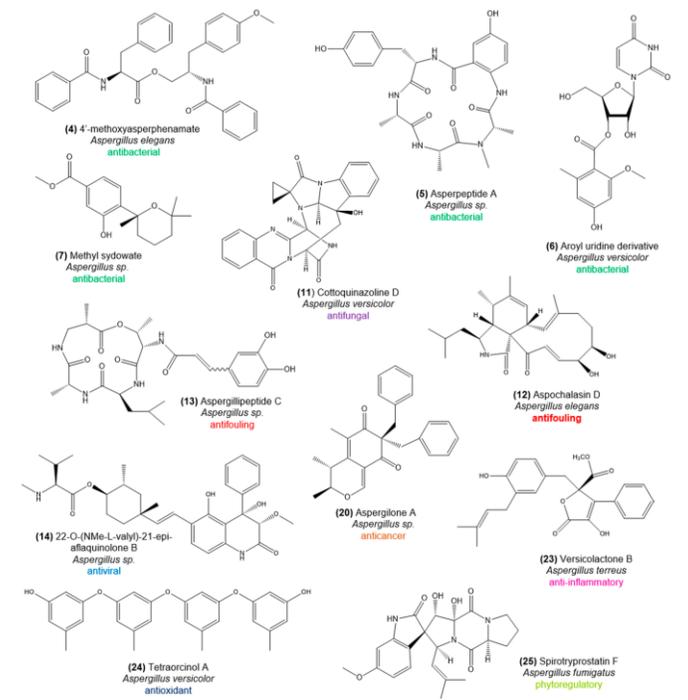
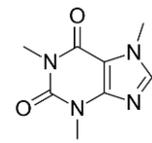


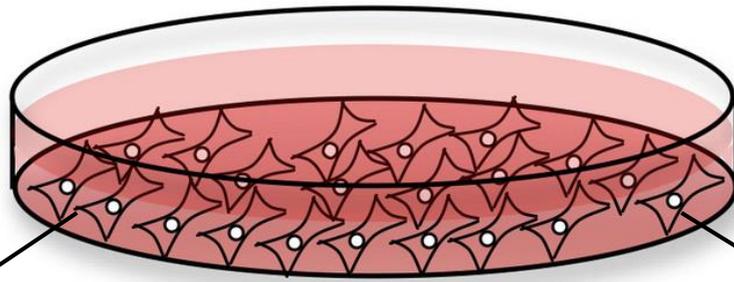
NL eScience Center

UCSD:  
Madeleine Ernst  
Pieter Dorrestein



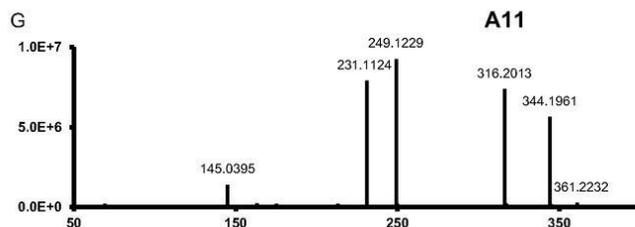
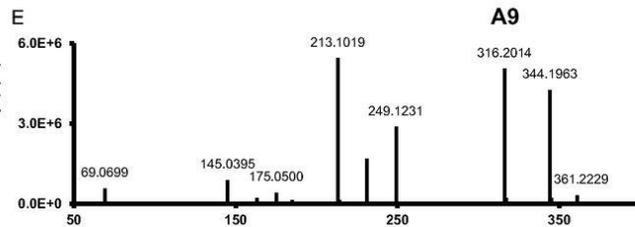
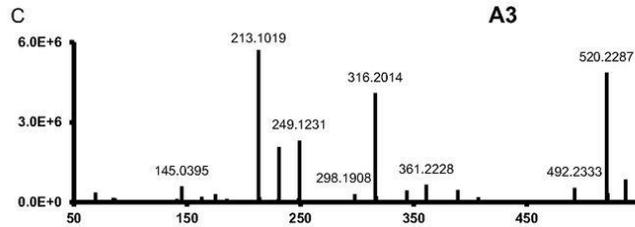
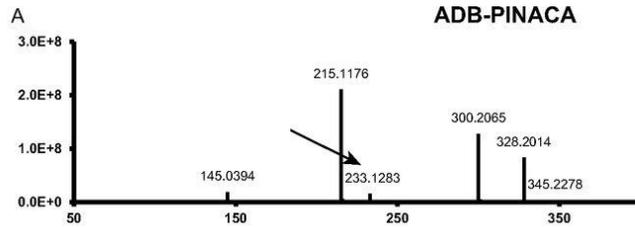
secondary metabolites





mass spectra

DNA

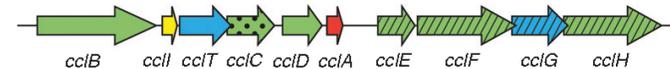


```

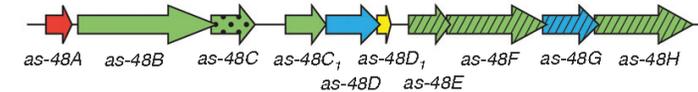
atggctatcg acgaaaaaaa acagaaagcg ttggcggcag cactgggcca gattgagaaa
caatttggtg aagggtccat catgcgctcg ggtgaagacc gttccatgga tgtggaaacc
atctctaccg gttcgtttc actggatabc gcgcttgggg cagggtggtc gccgatgggc
cgtatcgtcg aaatctacgg accggaatct tccggtaaaa ccacgctgac gctgcaggtg
atcgcgcgag cgcagcgtga aggtaaaaacc tgtgcgttta tccatgctga acacgcgctg
gacccaatct acgcacgtaa actgggcgtc gatatcgaca acctgctgtg ctcccagccg
gacaccggcg agcaggcact ggaaatctgt gacgcccctgg cgcgttctgg cgcagtagac
ggtatcgtcg ttgactcctg ggggcactg acgcccgaag cggaaatcga aggcgaaatc
ggcgactctc acatgggctt tgcggcacgt atgatgagcg aggcgatgcg taagtggcg
ggtaacctga agcagtccaa cacgctgctg atcttcatca accagatccg tatgaaaatt
ggtgtgatgt tccgtaaccc ggaaaccact accggtggta accgctgaa attctacgcc
totgttcgtc tccacatccg tccgatccgc cgcgtgaaag agggcgaaaa cgtggtgggt
agcgaacc ccggtgaaagt ggtgaagaac aaaaatcgtc cgcgctttaa acaggtgtaa
ttccagatcc tctacggcga aggtatcaa ttctacggcg aactggttga cctgggcgta
aaagagaagc tgatcgagaa agcaggcgtg tttttttttt acaaaggatga gaagatcgtt
cagggtaaag cgaatgogac tgactgctg taagatcctc cggaaaccgc gaaagagatc
gagaagaaag tacgtgagtt gctgctgagc aaccgaaact caaccgongga tttctctgta
gatgatagcg aaggcgtagc agaaaatac tttttttttt tttttttttt tttttttttt
aagggtcgcg tgtgcggccc tttttttttt ttaagttgta aggatatgcc atgacacgat
  
```

**HMM**  
(Hidden Markov Model)  
+ manually written rules

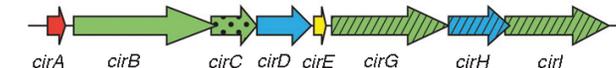
**Carnocyclin A**  
(*Carnobacterium maltaromaticum* UAL307)

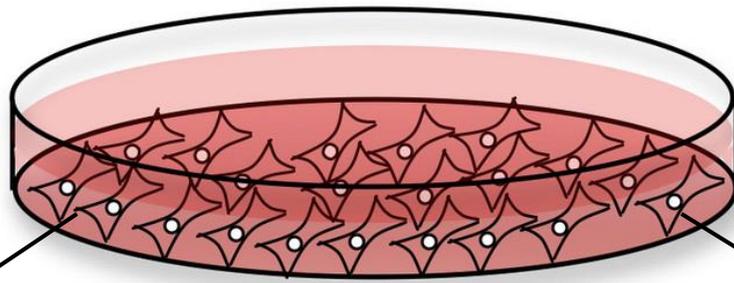


**Enterocin AS-48**  
(*Enterococcus faecalis* S-48)



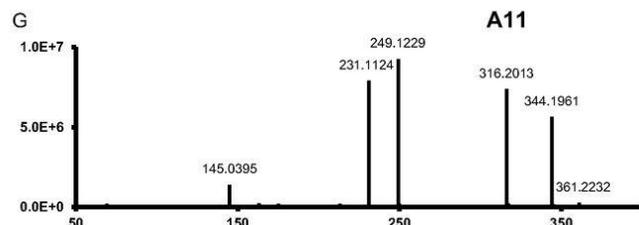
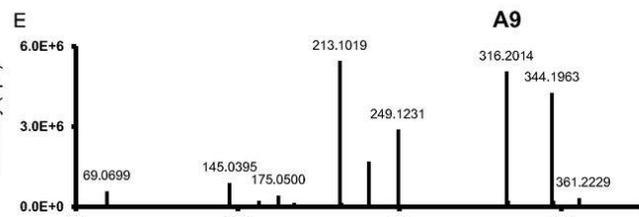
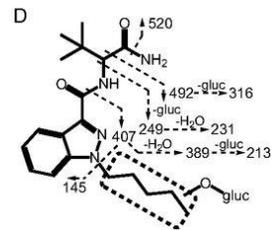
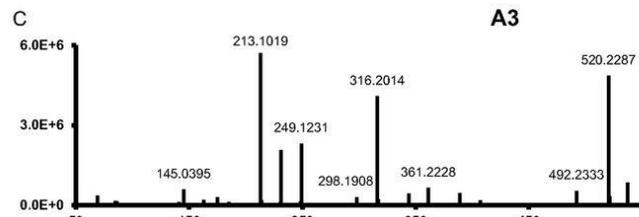
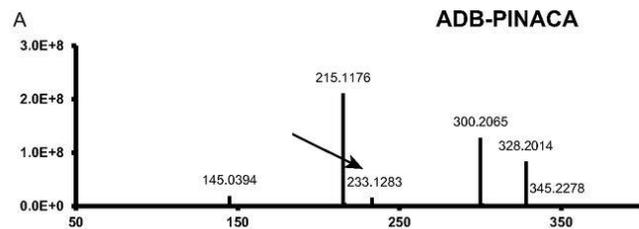
**Circularin A**  
(*Clostridium beijerinckii* ATCC 25752)





mass spectra

DNA

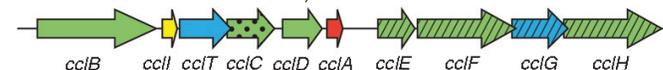


```

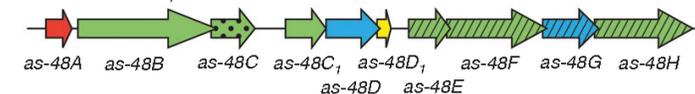
atggctatcg acgaaaaaaa acagaaagcg ttggcggcag cactgggcca gattgagaaa
caatttggtta aagggtccat catgcgcctg ggtgaagacc gttccatgga tgtggaaacc
atctctaccg gttcgtttc actggatac gcgcttgggg cagggtggtc gccgatgggc
cgtatcgtcg aaatctacgg accggaatct tccggtaaaa ccacgctgac gctgcaggtg
atcgcgcgag cgcagcgtga aggtaaaaacc tgtgcgttta tccatgctga acacgcgctg
gacccaatct acgcacgtaa actgggcgtc gatatcgaca acctgctgtg ctcccagccg
gacaccggcg agcaggcact ggaaatctgt gacgacctgg cgcgttctgg cgcagtagac
ggtatcgtcg ttgactcgt ggogggcactg acgocgaaag cggaaatcga aggcgaaatc
ggcgcacttc acatgggctc tgcggcacgt atgatgagc aggcgatgag taagtggcg
ggtaacctga agcagtccaa cacgctgctg atcttcatca accagatccg tatgaaaatt
ggtgtgatgt tcggtaaacc ggaaaccact accggtggta acgocgtgaa attctacgcc
totgttcgtc tcgacatccg tcgtatcggc gcggtgaaag agggcgaaaa cgtggtgggt
agcgaiaacc gcgtgaaagt ggtgaaagac aaaaatcgctg cgcgctttaa acaggtgtaa
ttccagatcc tctacggcga aggtatcaaa ttctacggcg aactggttga cctgggocgta
aaagagaagc tgatocgagaa agcaggcgtg taagatcctt acaaaggtga gaagatcggg
cagggtaaag cgaatgogac tgocgtggtg taagatcctt cggaaaccgc gaaagagatc
gagaagaaaag tacgtgagtt gctgctgagc aaccggaact caacgocgga tttctctgta
gatgatagcg aaggcgtagc agaaaatac gctgctgagc caacggaact caacgocgga
aagggtcgcg tgtgcggccc tttttttttt ttaagttgta aggatatgcc atgacacgat
  
```

**HMM**  
(Hidden Markov Model)  
+ manually written rules

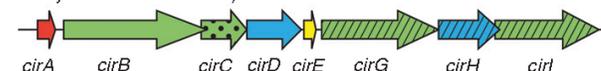
**Carnocyclin A**  
(*Carnobacterium maltaromaticum* UAL307)

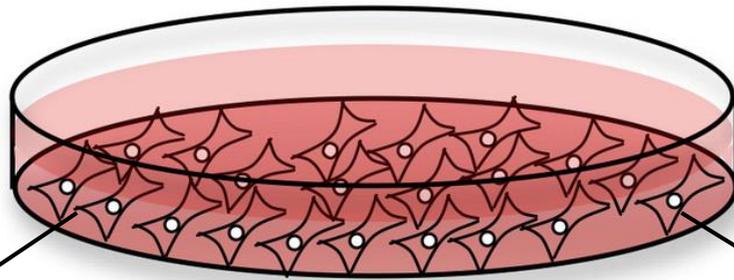


**Enterocin AS-48**  
(*Enterococcus faecalis* S-48)



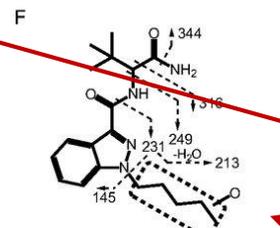
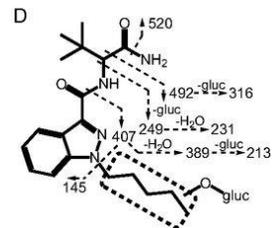
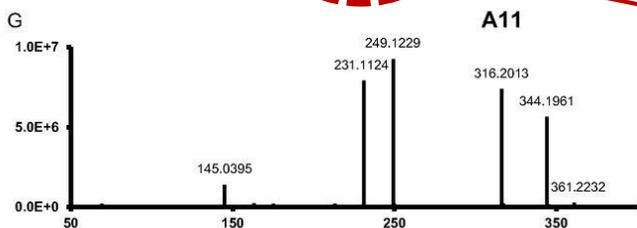
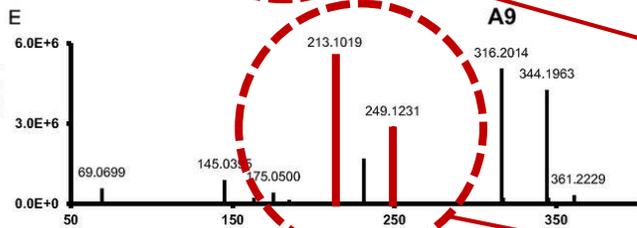
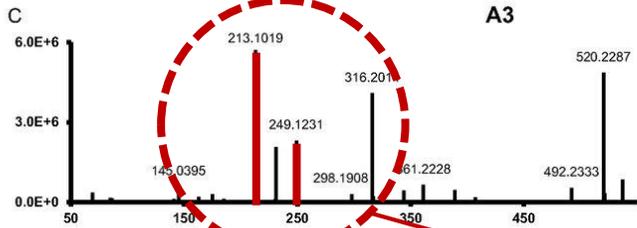
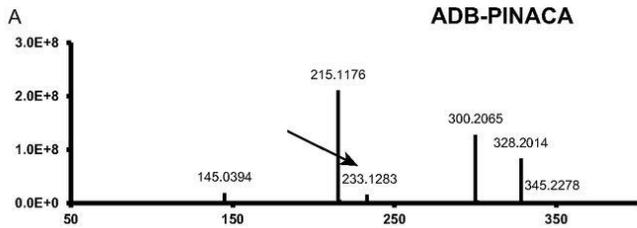
**Circularin A**  
(*Clostridium beijerinckii* ATCC 25752)





mass spectra

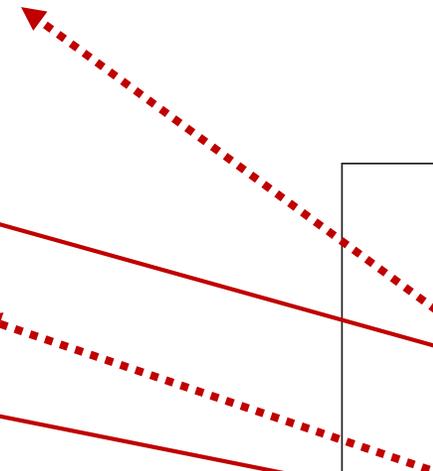
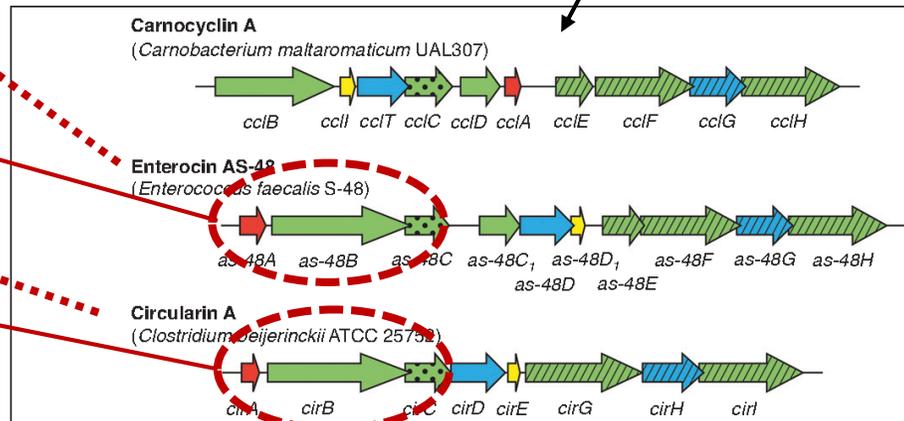
DNA



```

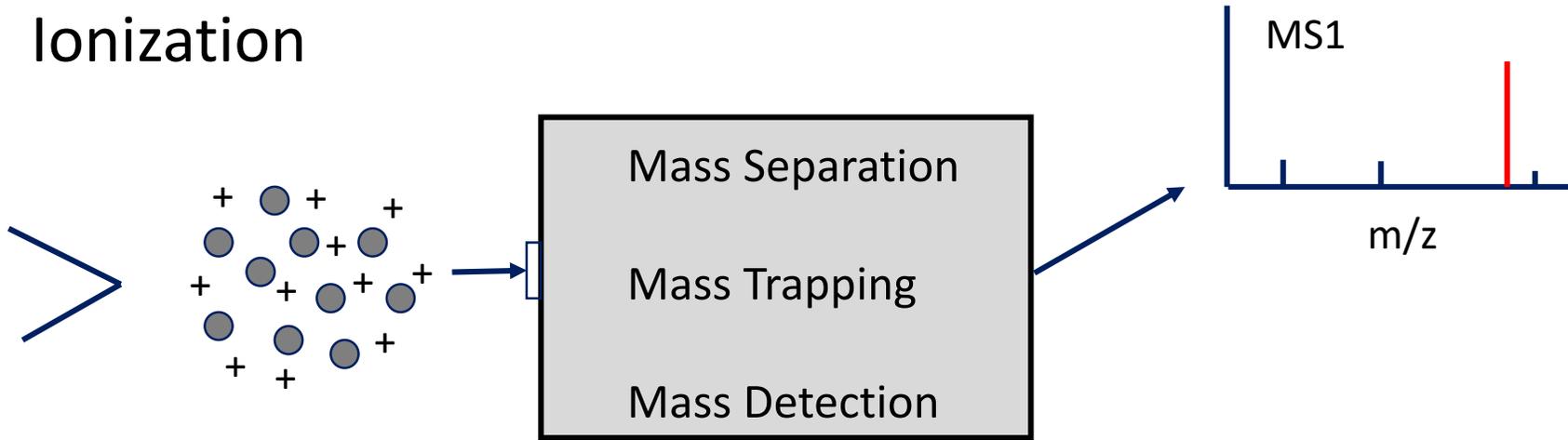
atggctatcg acgaaaaaaa acagaaagcg ttggcggcag cactgggcca gattgagaaa
caatttggtta aagggtccat catgcgcctg ggtgaagacc gttccatgga tgtggaaacc
atctctaccg gttcgtttc actggatac gcgcttgggg cagggtggtc gccgatgggc
cgtatcgtcg aaatctacgg accggaatct tccggtaaaa ccacgtgac gctgcaggtg
atcgcgcgag cgcagcgtga aggtaaaaacc tgtgcgttta tccatgctga acacgcgctg
gacccaatct acgcacgtaa actgggcgtc gatatcgaca acctgctgtg ctcccagccg
gacaccggcg agcaggcact ggaaatctgt gacgcccctg cgcgttctgg cgcagtagac
ggtatcgtcg ttgactcctg ggggcactg acgcccgaag cggaaatcga aggcgaaatc
ggcgactctc acatgggctt tgcggcacgt atgatgagc aggcgatgcg taagctggcg
ggtaacctga agcagtccaa cacgctgctg atcttcatca accagatccg tatgaaaatt
ggtgtgatgt tcggtaaccc ggaaaccact accggtggta acgcgctgaa attctacgcc
totgttctgc tcgacatccg tcgtatcggc gcggtgaaag agggcgaaaaa cgtggtgggt
agcgaiaacc gcgtgaaagt ggtgaaagac aaaaatcgctg cgcgctttaa acaggtgtaa
ttccagatcc tctacggcga aggtatcaa ttctacggcg aactggttga cctgggcgta
aaagagaagc tgatcgagaa agcaggcgtg tctctctctt acaaaggtga gaagatcggg
cagggtaaag cgaatgogac tgactggtg taagatcttc cggaaaccgc gaaagagatc
gagaagaaag tacgtgagtt gctgctgagc aaccgaaact caaccggga tttctctgta
gatgatagcg aaggcgtagc agaaaatac gctgctgagc caaccgaaact caaccggga
aagggtcgca tgtgcggccc tttttttttt ttaagttgta aggatatgcc atgacacgat
  
```

**HMM**  
(Hidden Markov Model)  
+ manually written rules



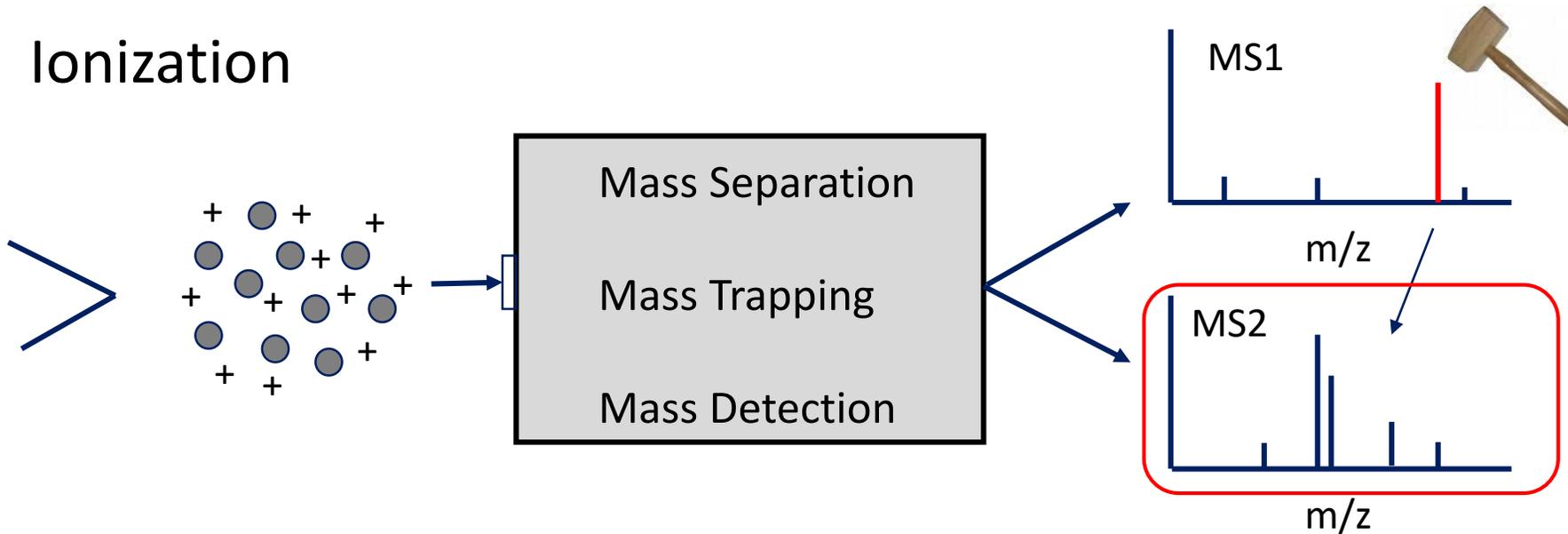
# Mass spectrometry and fragmentation

Ionization

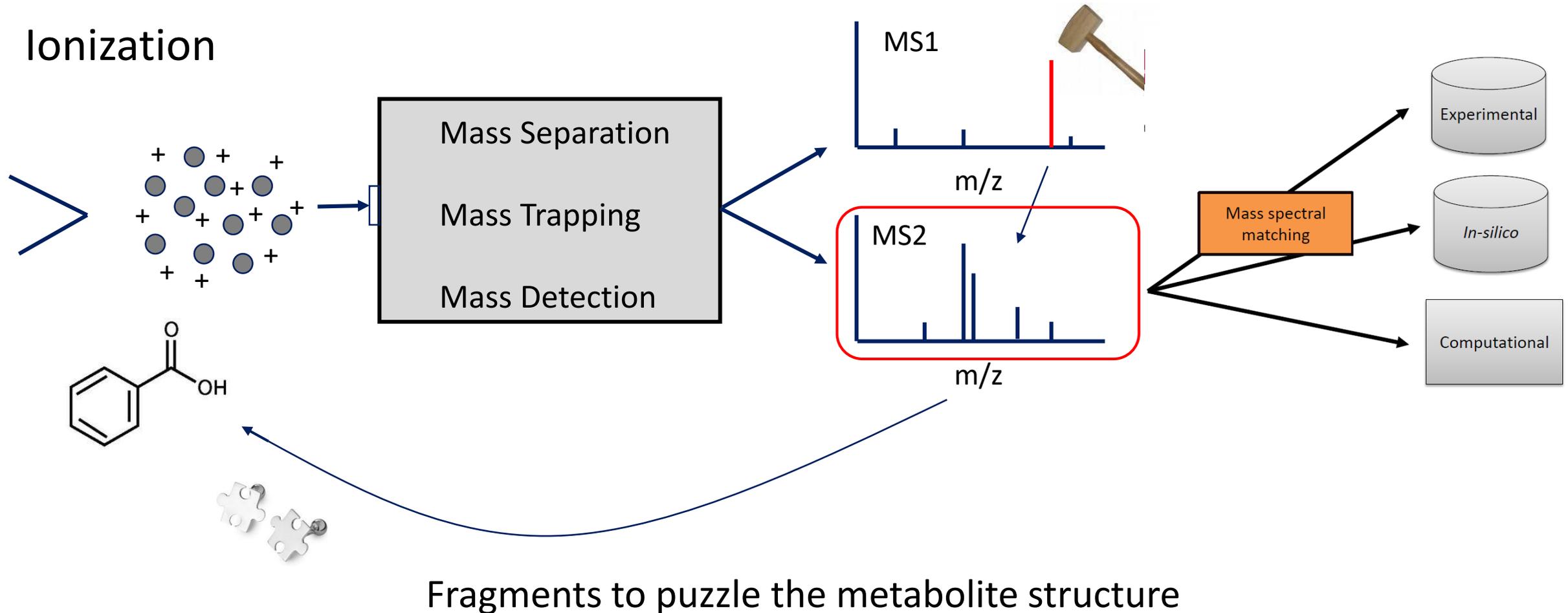


# Mass spectrometry and fragmentation

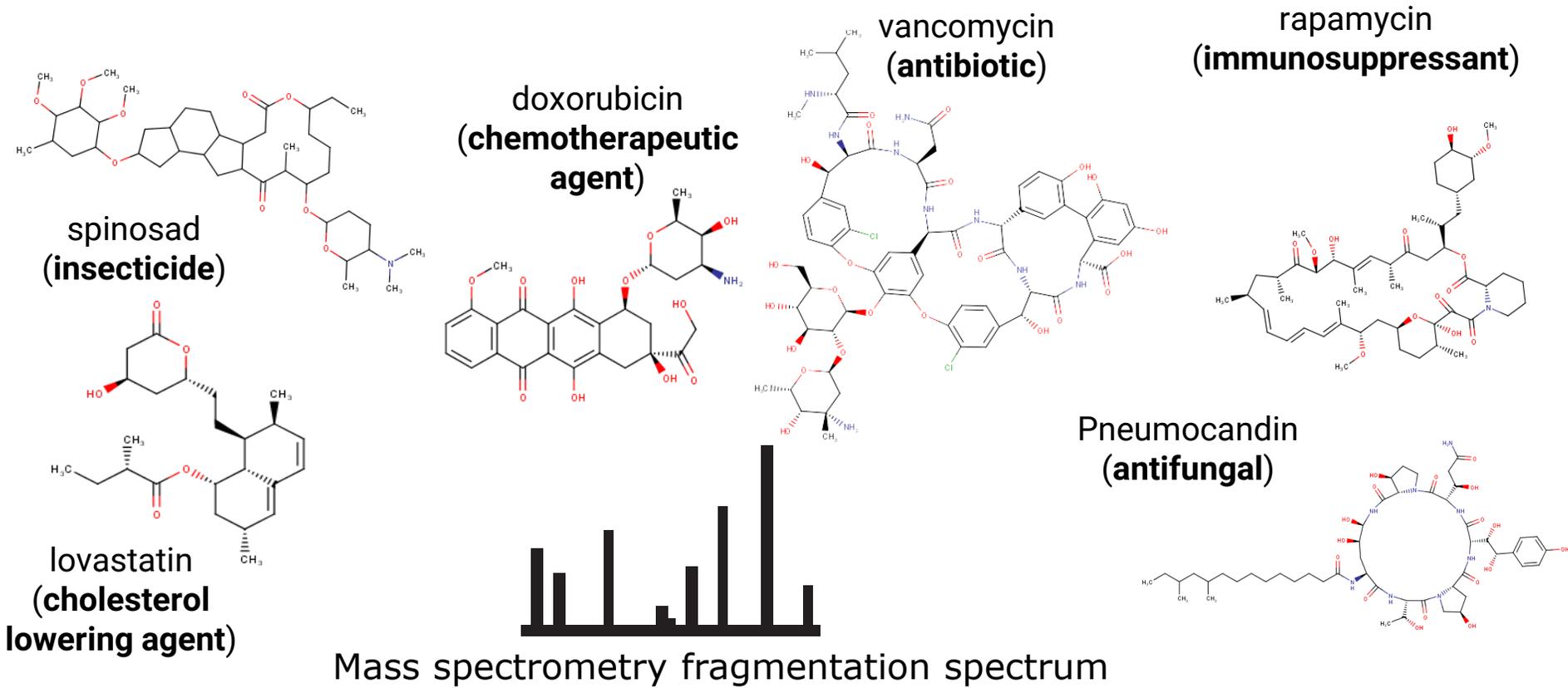
Ionization



# Mass spectrometry and fragmentation



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

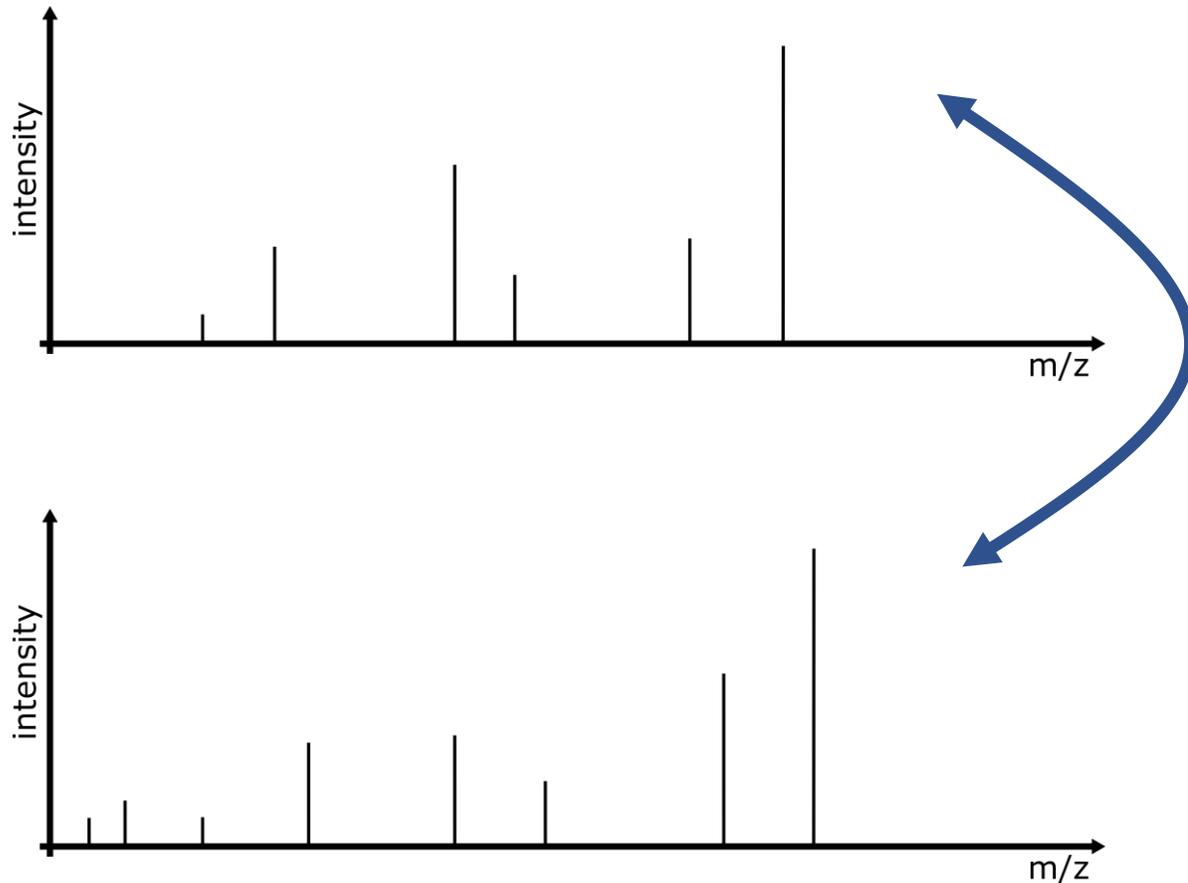


## The challenge....

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

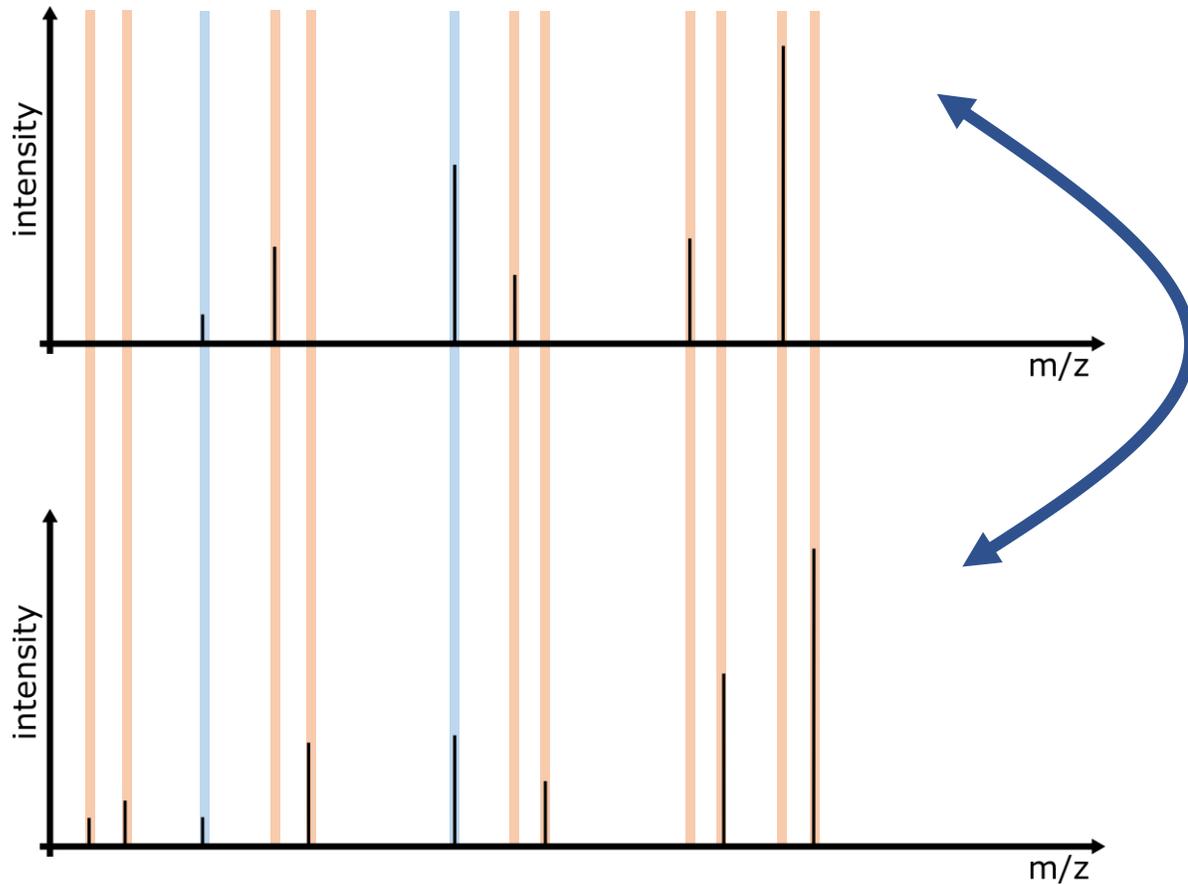
# But.... How similar are they?

## Spectral similarity



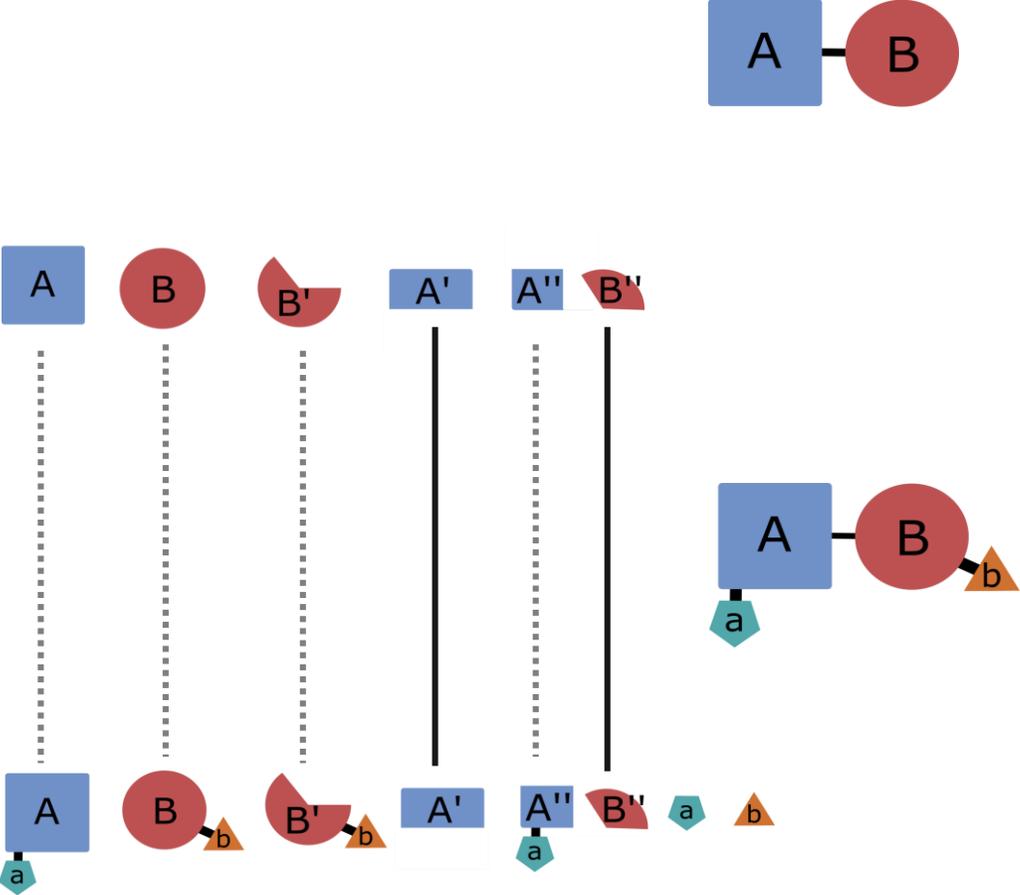
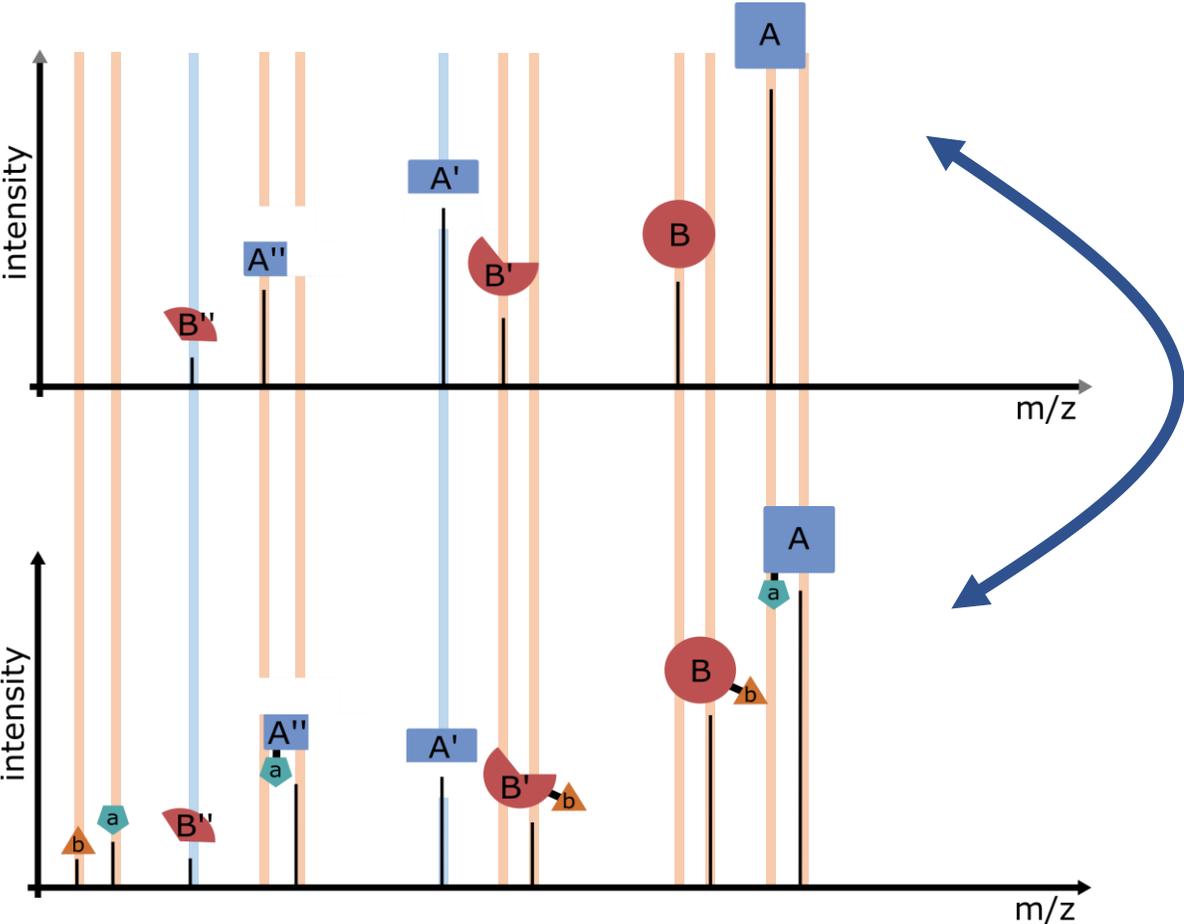
# How similar are they?

## Spectral similarity



# How similar are they?

## Spectral similarity



# How similar are they?

What does similar mean?

...likes cake with a cappuccino.

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

number of words?

number of characters?

grammatical structure?

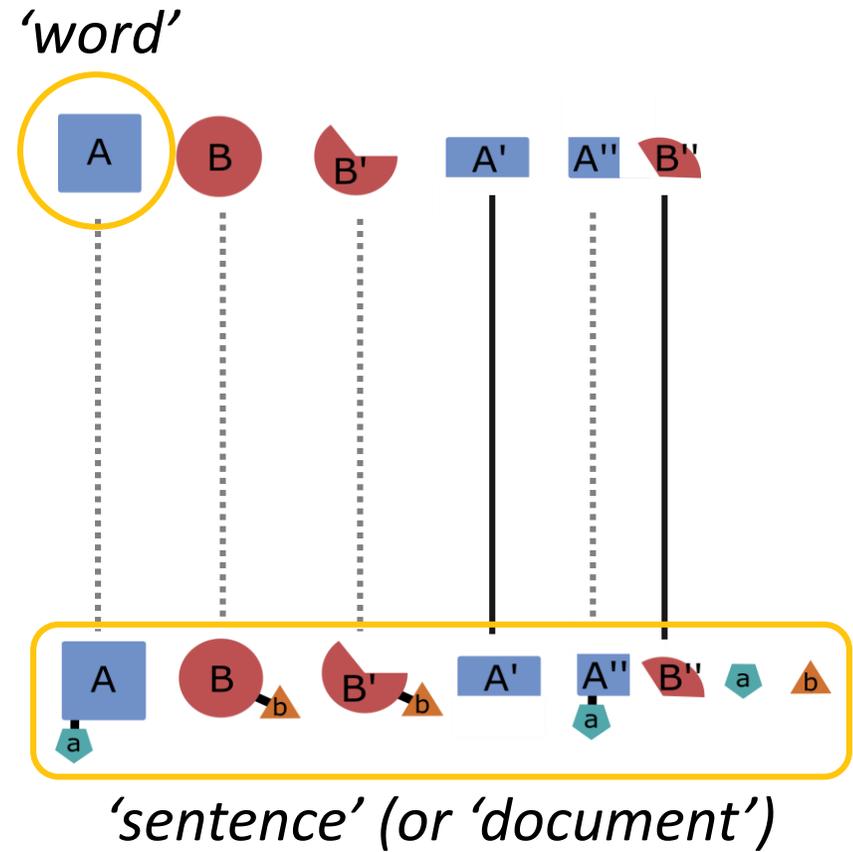
topic?

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   |    |    |   |     |     |     |
| <b>cake</b>   | 0   |    | 0  |   | 24  |     |     |
| ...           |     |    |    |   |     |     |     |
| <b>cookie</b> | 9   | 0  |    |   | 17  |     |     |
| ...           |     |    |    |   |     |     |     |
| sweet         |     | 24 | 17 |   |     |     |     |
| ...           | ... |    |    |   |     |     |     |

NxN matrix

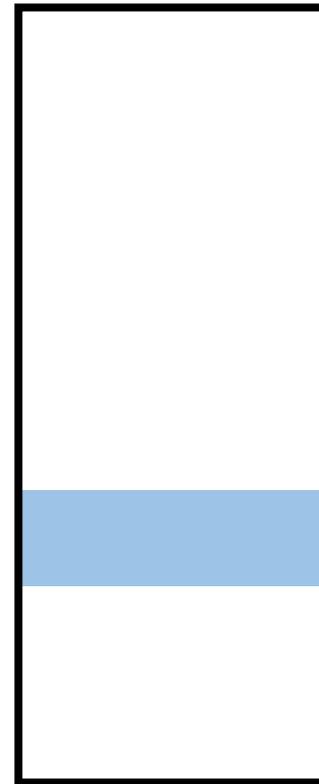
N: number of words in dictionary

'Word2Vec' → lower dimensional context vector

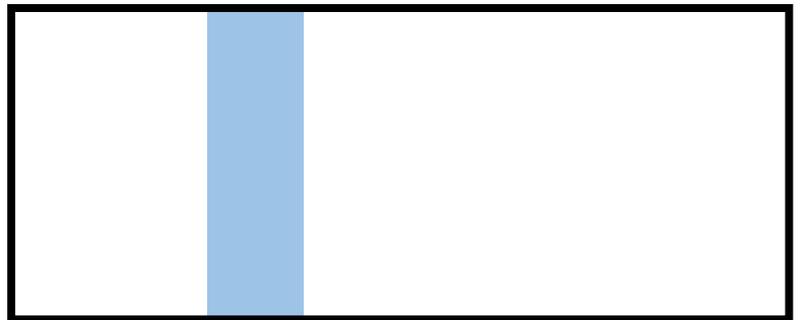
Words

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

≈

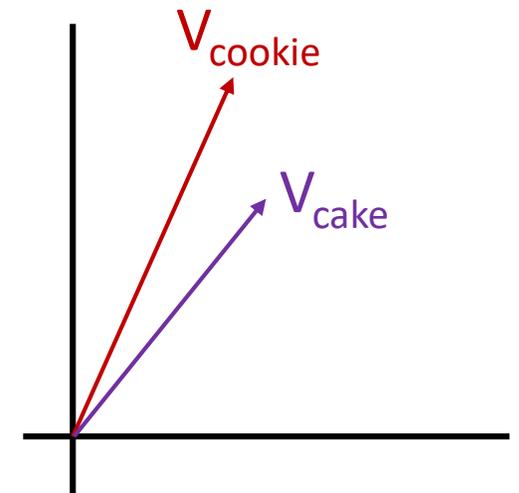
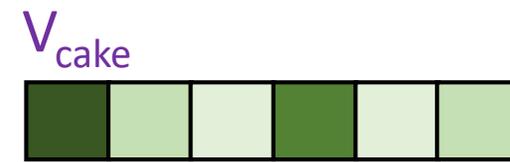
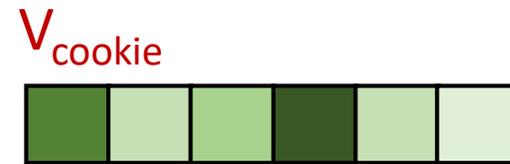


X



'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     |     |       |     |
| ...     | ... |     |      |     |        |     |       |     |

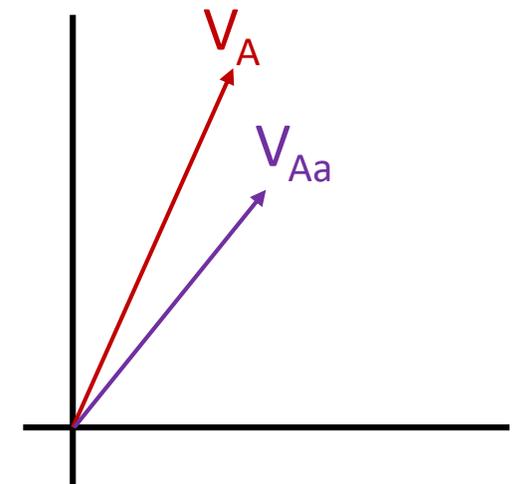
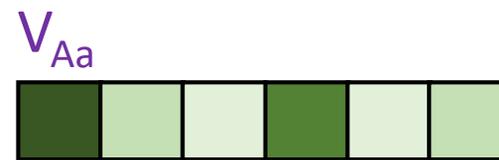
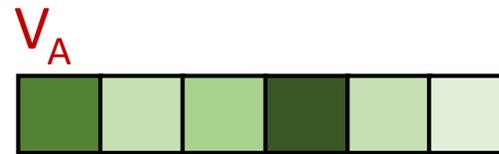


# NLP → metabolomics: use peaks as words

peak positions

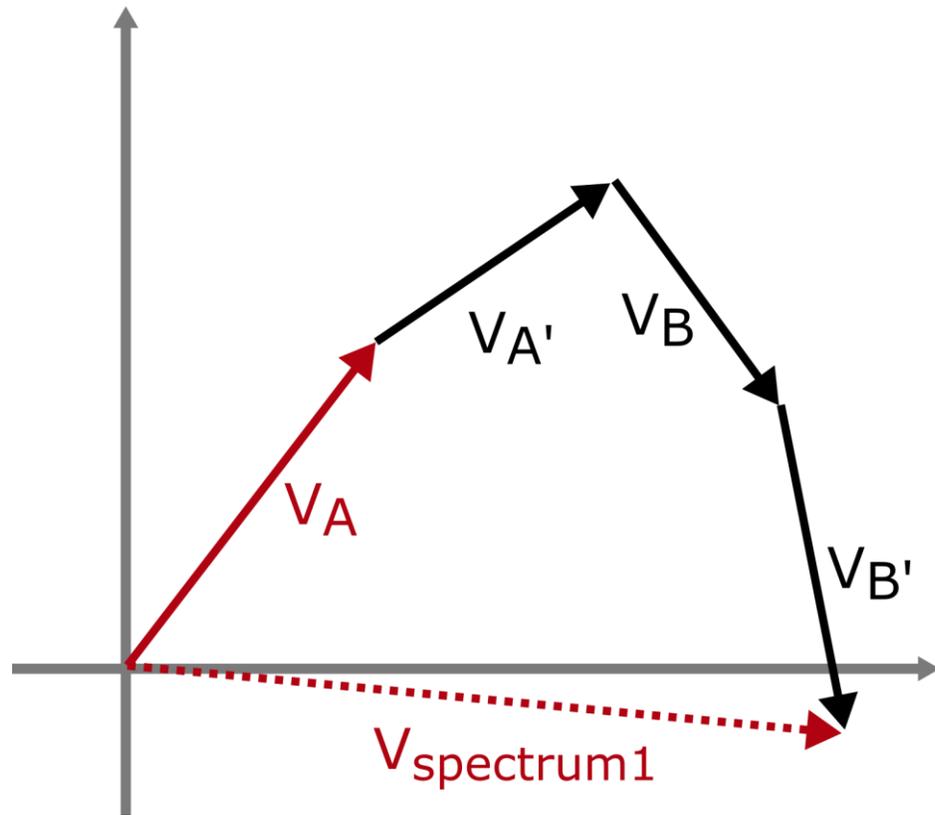
= 'words' ... m(Aa) ... m(A) ... m(A'') ...

|        |     |    |    |    |     |
|--------|-----|----|----|----|-----|
| ...    | ... | 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



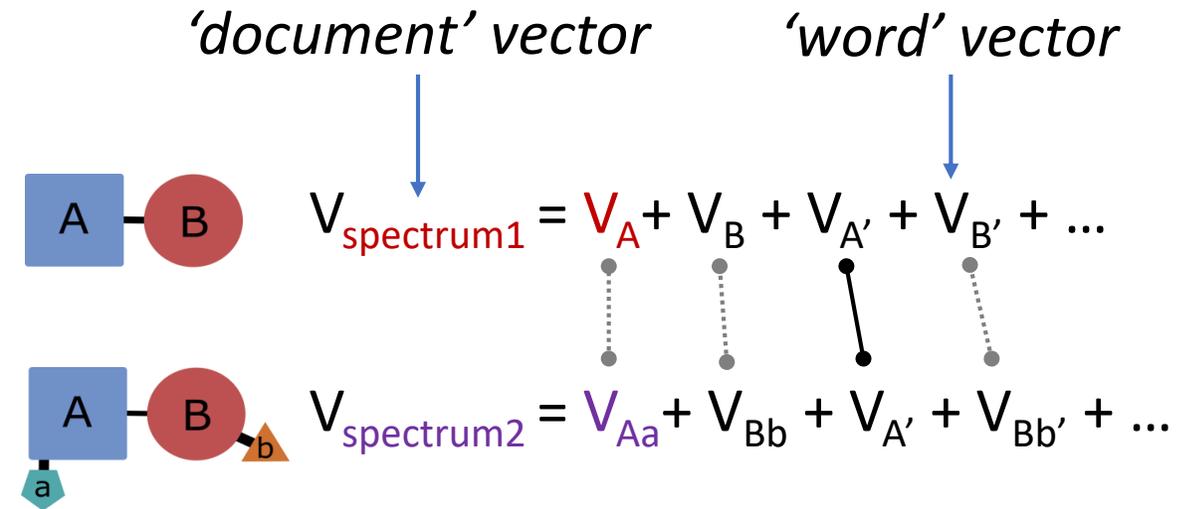
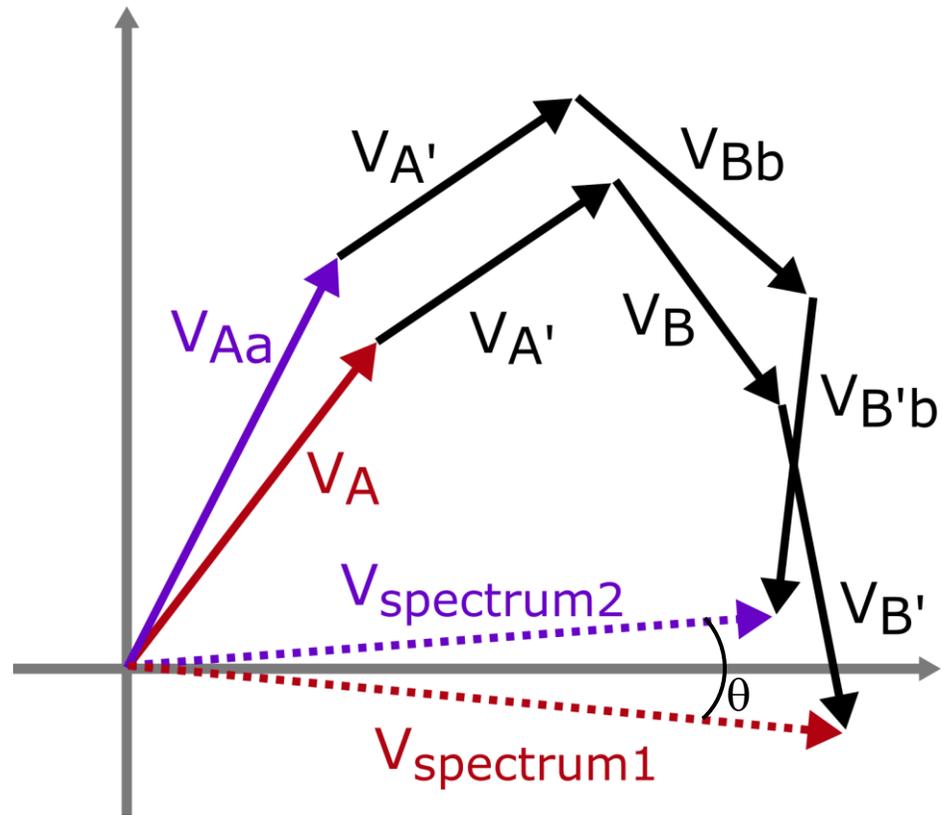
*'document' vector*      *'word' vector*

A — B

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

# Spectral similarity measures.

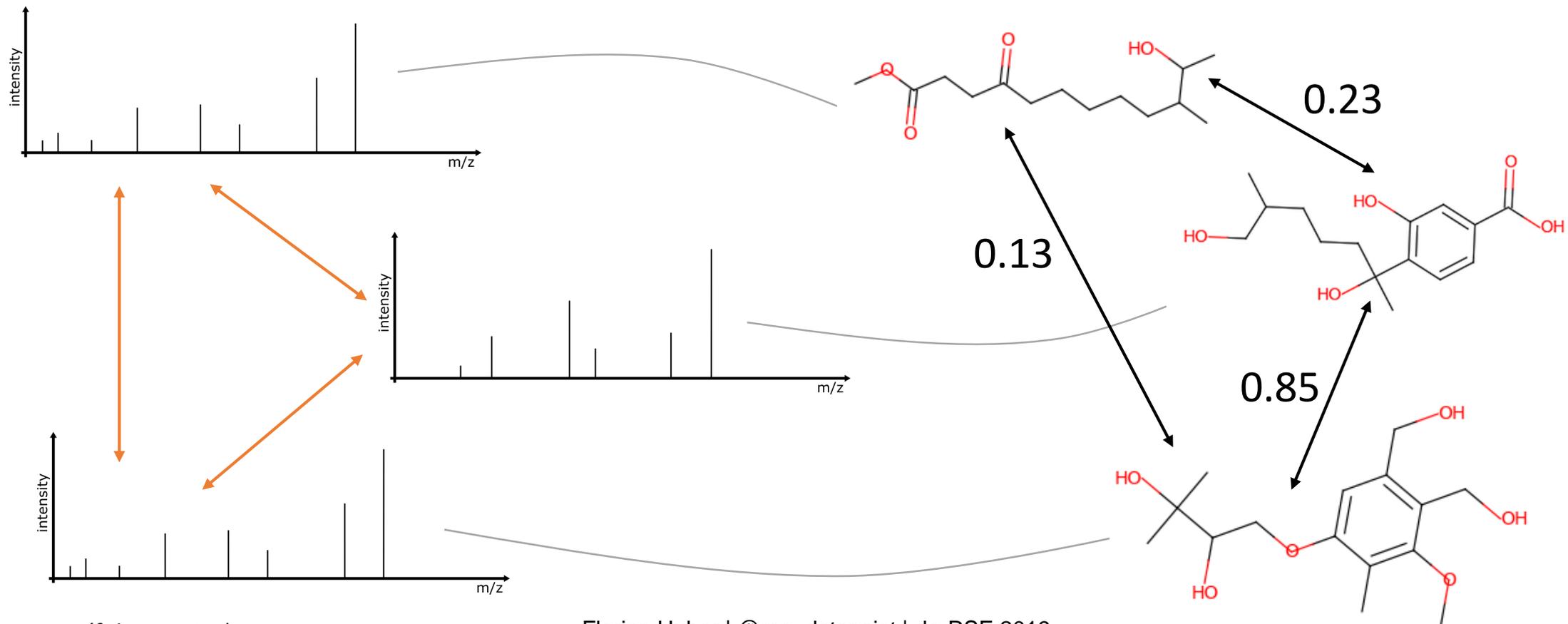
NLP/word2vec based method



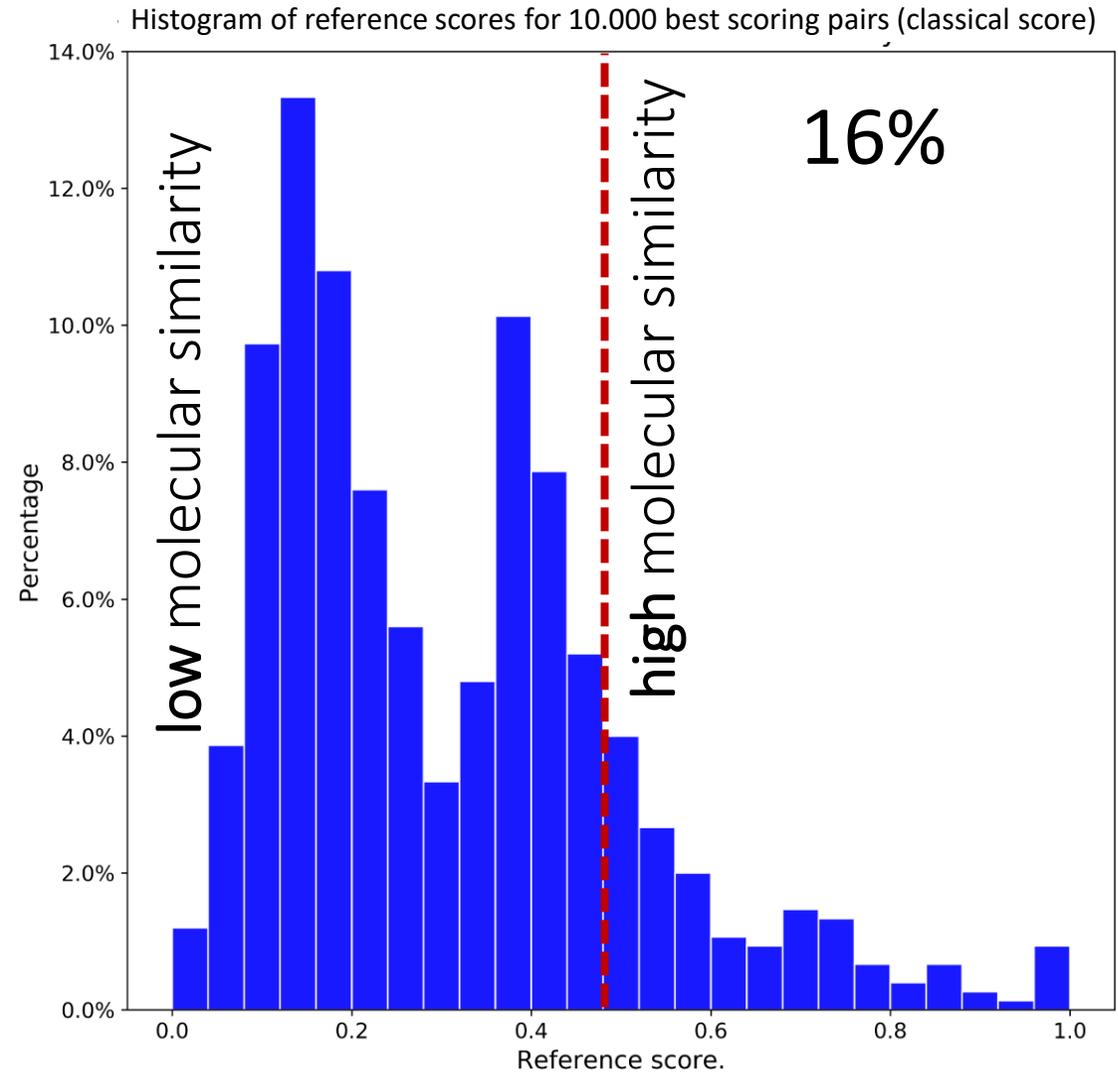
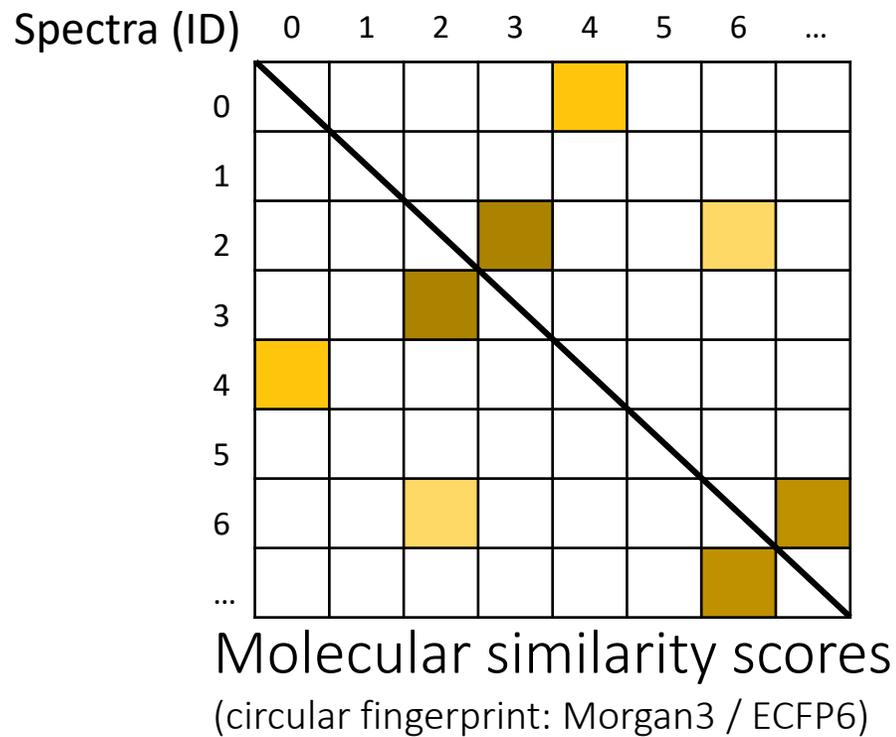
$$\text{Similarity} = \cos(\theta) = \frac{V_{\text{spectrum1}} \cdot V_{\text{spectrum2}}}{\|V_{\text{spectrum1}}\| \|V_{\text{spectrum2}}\|}$$

# Spectral similarity measures: evaluation.

**Dataset:** 11.000 spectra with known molecular structures

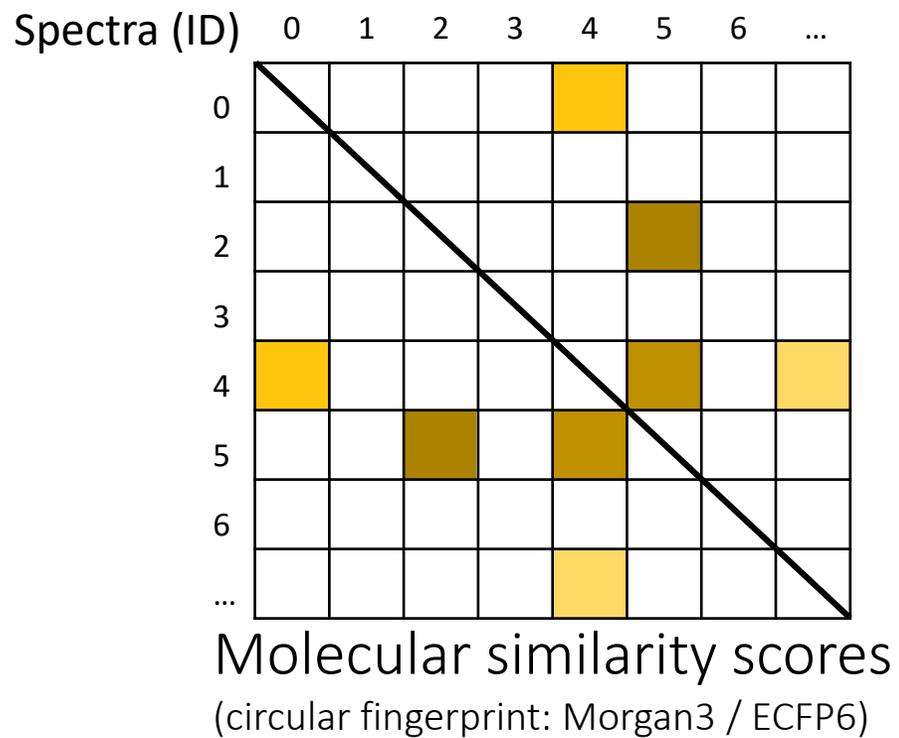


# Molecular similarity scores: 10.000 highest 'classical' scores\*

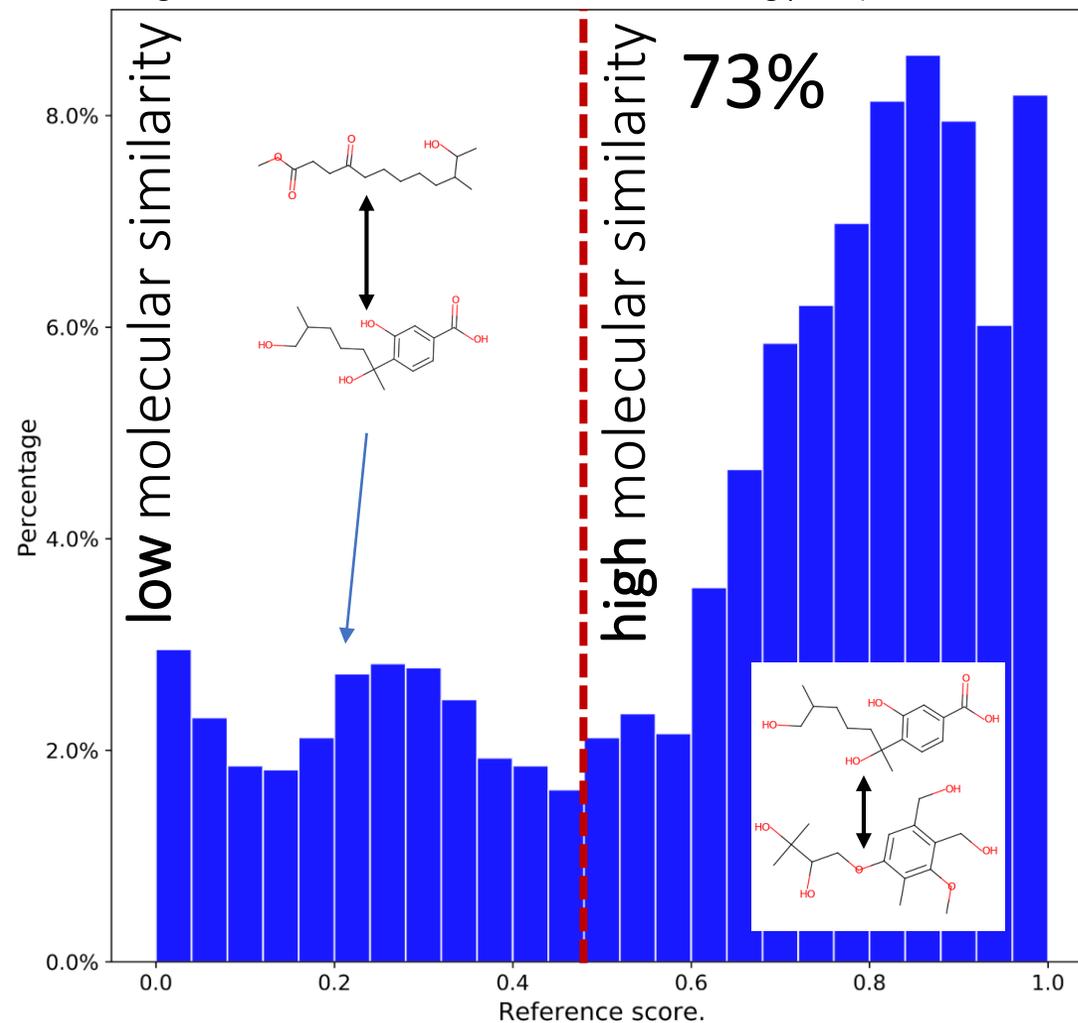


\* = scores > 0.998

# Molecular similarity scores: 10.000 highest NLP-based scores\*



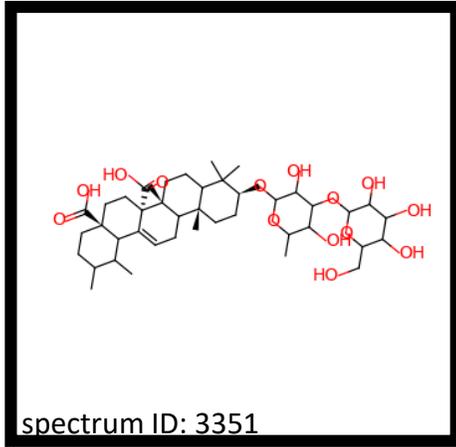
Histogram of reference scores for 10.000 best scoring pairs (NLP-based score)



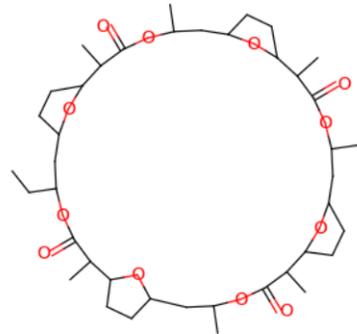
\* = scores > 0.84

# Spectral similarity measures: <sup>bad</sup> examples.

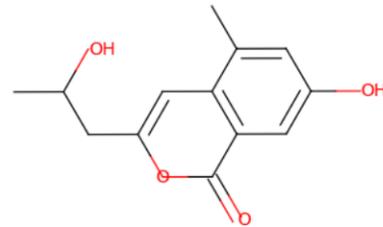
query molecule



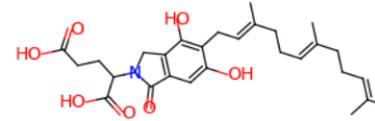
9 closest candidates (according to molecular networking similarity)



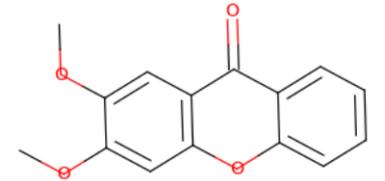
1



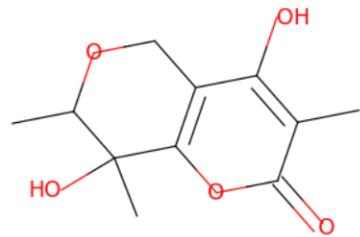
2



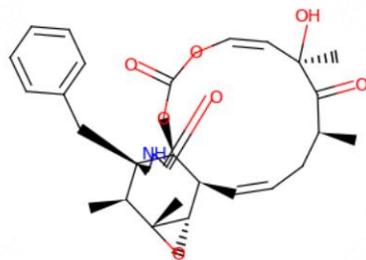
3



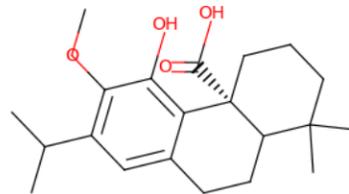
4



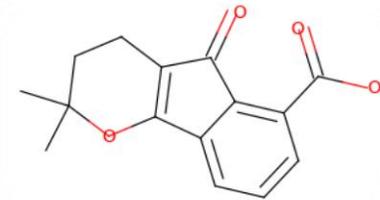
5



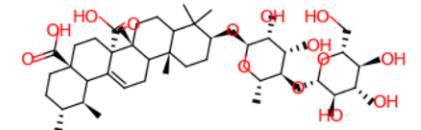
6



7



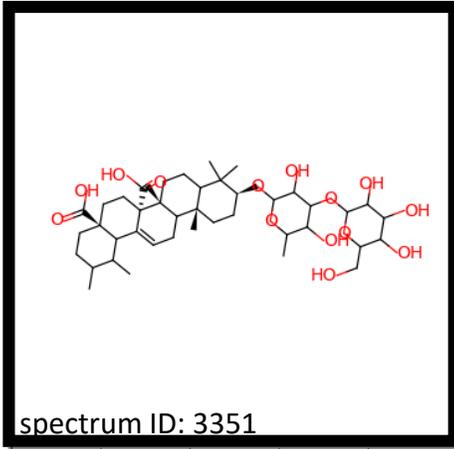
8



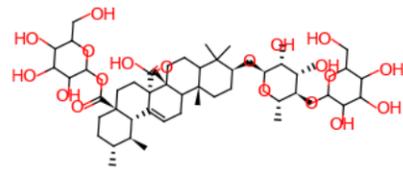
9

# Spectral similarity measures: examples.

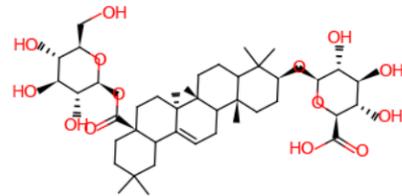
query molecule



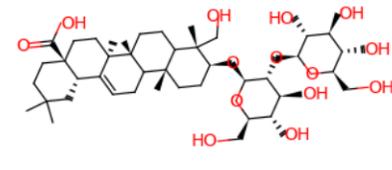
9 closest candidates (according to Word2vec-based spectral similarity)



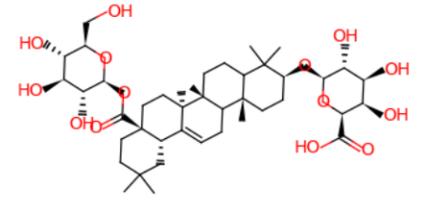
1



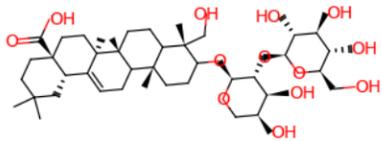
2



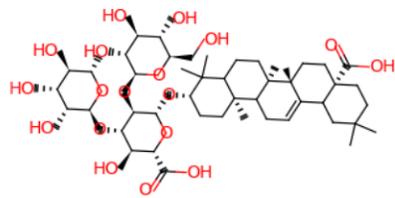
3



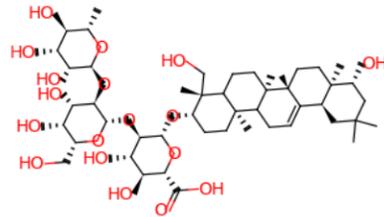
4



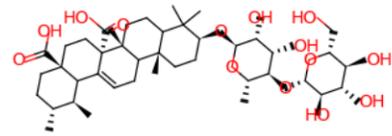
5



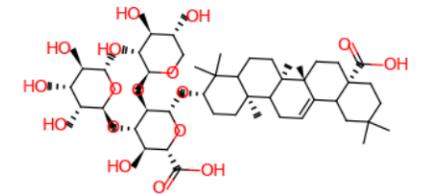
6



7



8

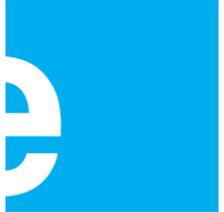


9

# 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](mailto:n.renaud@esciencecenter.nl)



netherlands **eScience** center

☎ +31 (0)20 460 4770

🌐 [www.esciencecenter.nl](http://www.esciencecenter.nl)

📄 [blog.esciencecenter.nl](http://blog.esciencecenter.nl)

Florian Huber

🐦 [@me\\_datapoint](https://twitter.com/me_datapoint)



Carlos Martinez-Ortiz

🐦 [@neocarlitos](https://twitter.com/neocarlitos)

