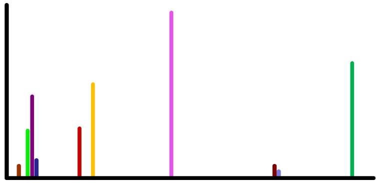
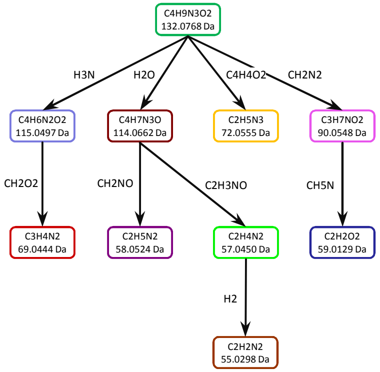
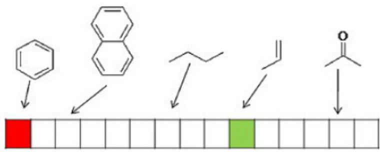
Elementary, my dear Watson: Fingerprint search in molecular structure databases

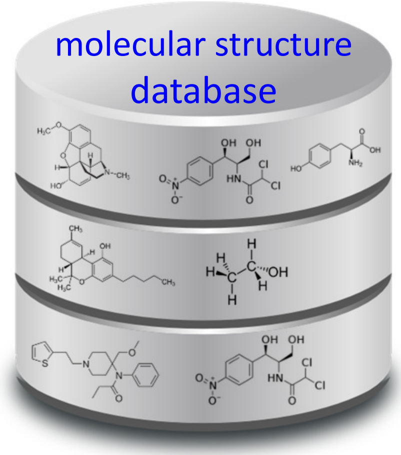
Sebastian Böcker

Friedrich-Schiller-Universität Jena

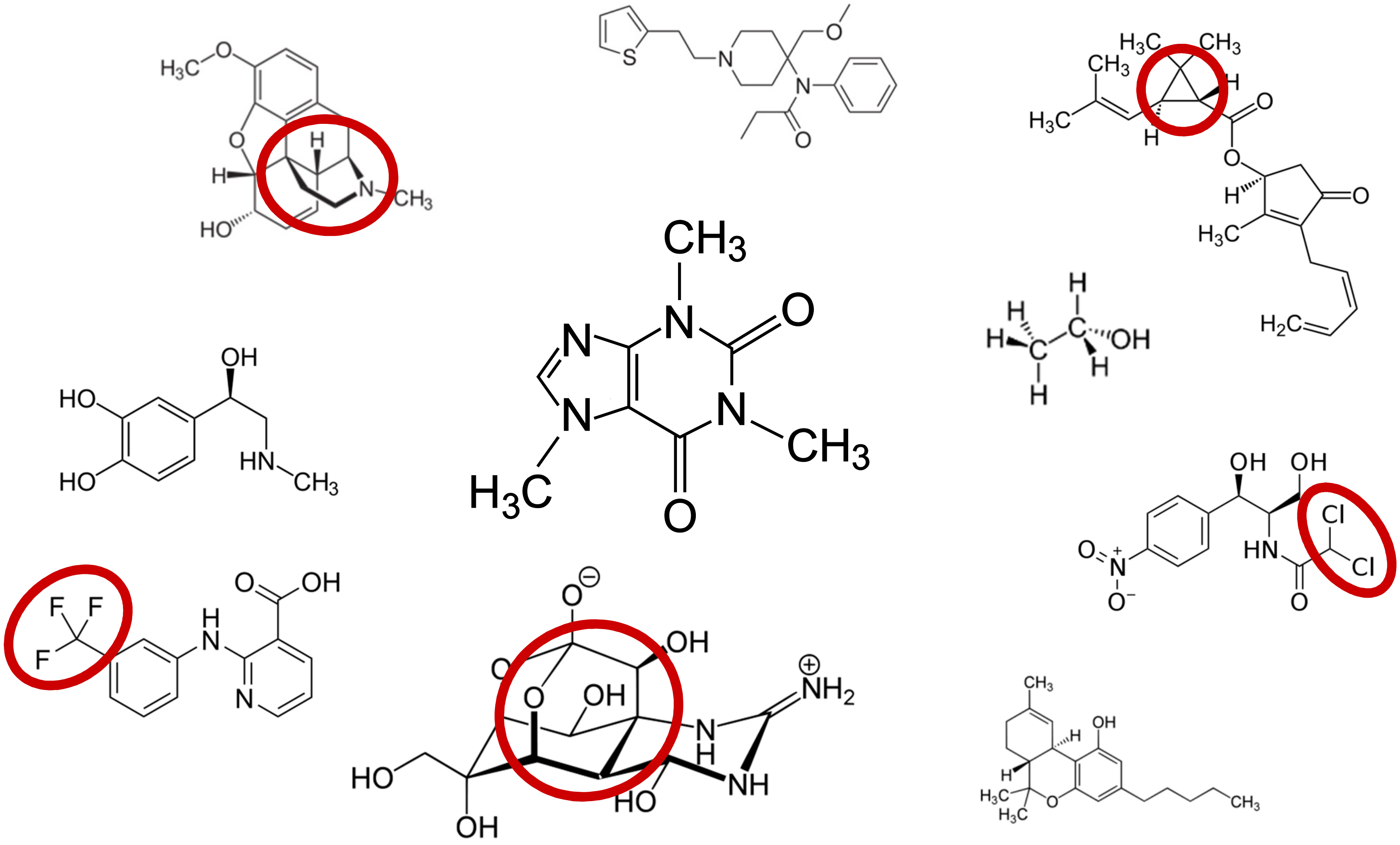








We are talking about small molecules



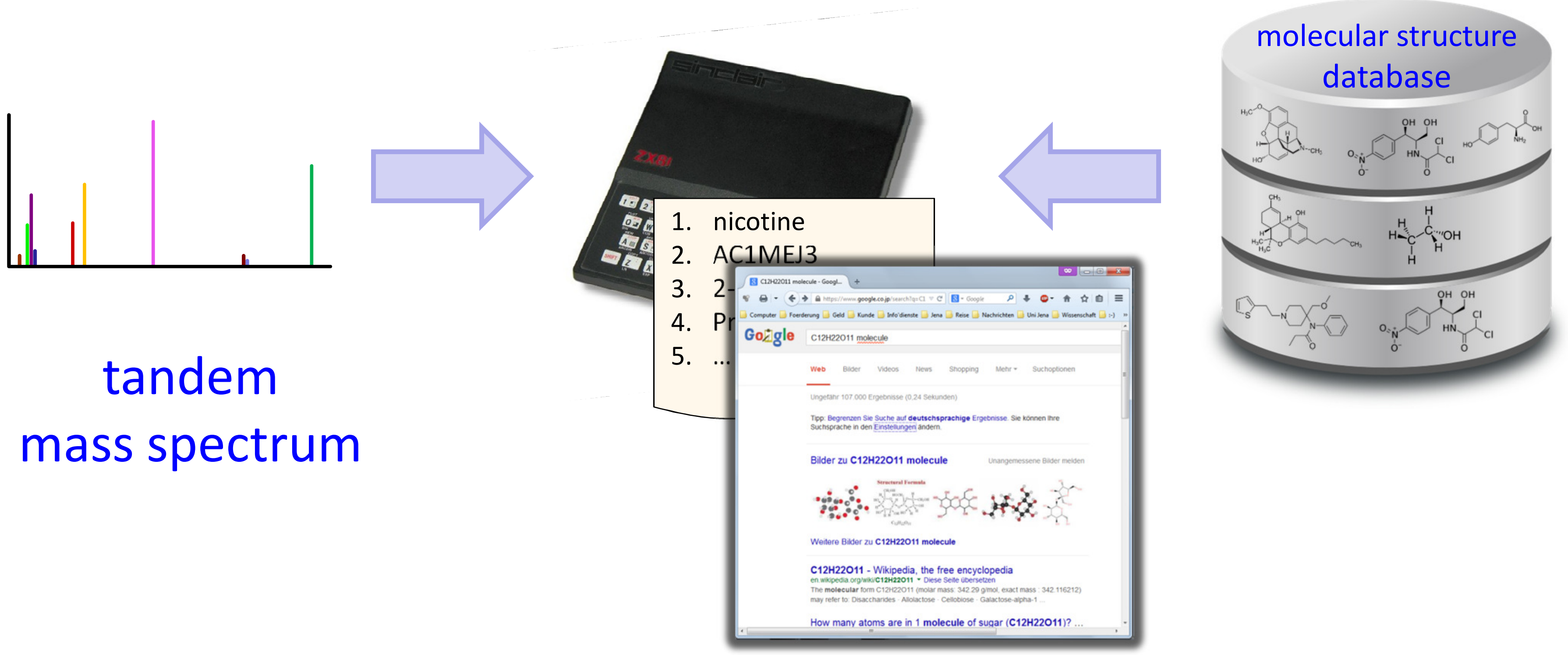
Metabolites: Why care?

- metabolites closest to phenotype
- majority of drugs derived from natural products (that is, metabolites)
- vast majority of medical biomarker assays target metabolites
- vast majority of (plant, animal and human) diseases have a non-genetic cause

Stolen from a talk by David Wishart

A simple question

- Given the tandem mass spectrum of a compound, can we find it – in a molecular structure database?



1. nicotine
2. AC1MEJ3
3. 2-
4. Pr
5. ...

C12H22O11 molecule - Googl...

https://www.google.co.jp/search?q=C1

Computer Foerderung Geld Kunde Info'dienste Jena Reise Nachrichten Uni Jena Wissenschaft :-)

Google C12H22O11 molecule

Web Bilder Videos News Shopping Mehr Suchoptionen

Ungefähr 107.000 Ergebnisse (0,24 Sekunden)

Tipp: Begrenzen Sie Suche auf **deutschsprachige** Ergebnisse. Sie können Ihre Suchsprache in den Einstellungen ändern.

Bilder zu **C12H22O11 molecule** Unangemessene Bilder melden

Structural Formula

Weitere Bilder zu **C12H22O11 molecule**

C12H22O11 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**C12H22O11** Diese Seite übersetzen
The **molecular** form C12H22O11 (molar mass: 342.29 g/mol, exact mass : 342.116212) may refer to: Disaccharides · Allolactose · Cellobiose · Galactose-alpha-1 ...

How many atoms are in 1 **molecule** of sugar (**C12H22O11**)? ...

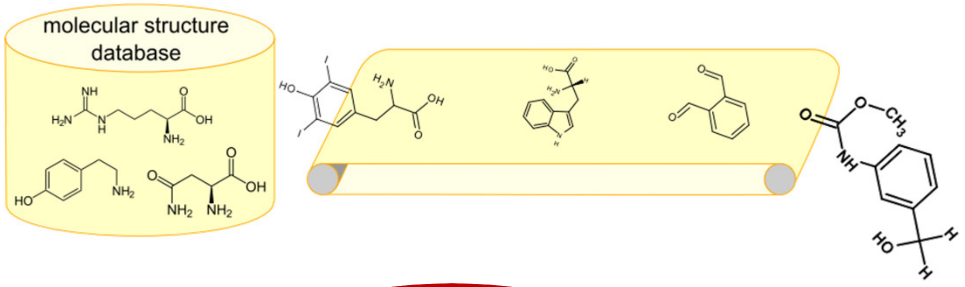
Why is this so complicated?

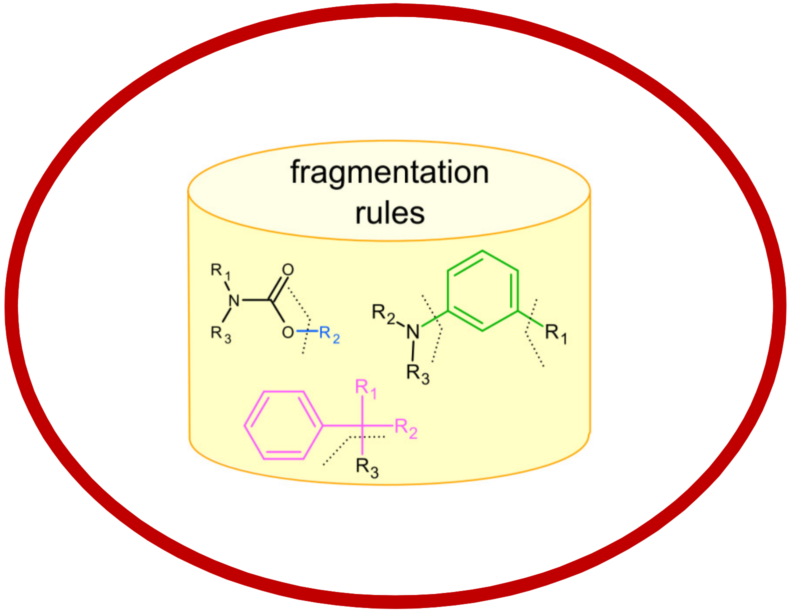
- **SEQUEST: Searching peptide sequence databases since 1994**
- but metabolites are different

	proteins/peptides	**metabolites**
molecules are…	structurally similar	highly diverse
genome information tells you…	everything but PTMs	(almost) nothing
molecules fragment…	at one fixed energy	some need 0 eV, some 80 eV
fragmentation is…	“easily” predictable	pretty involved

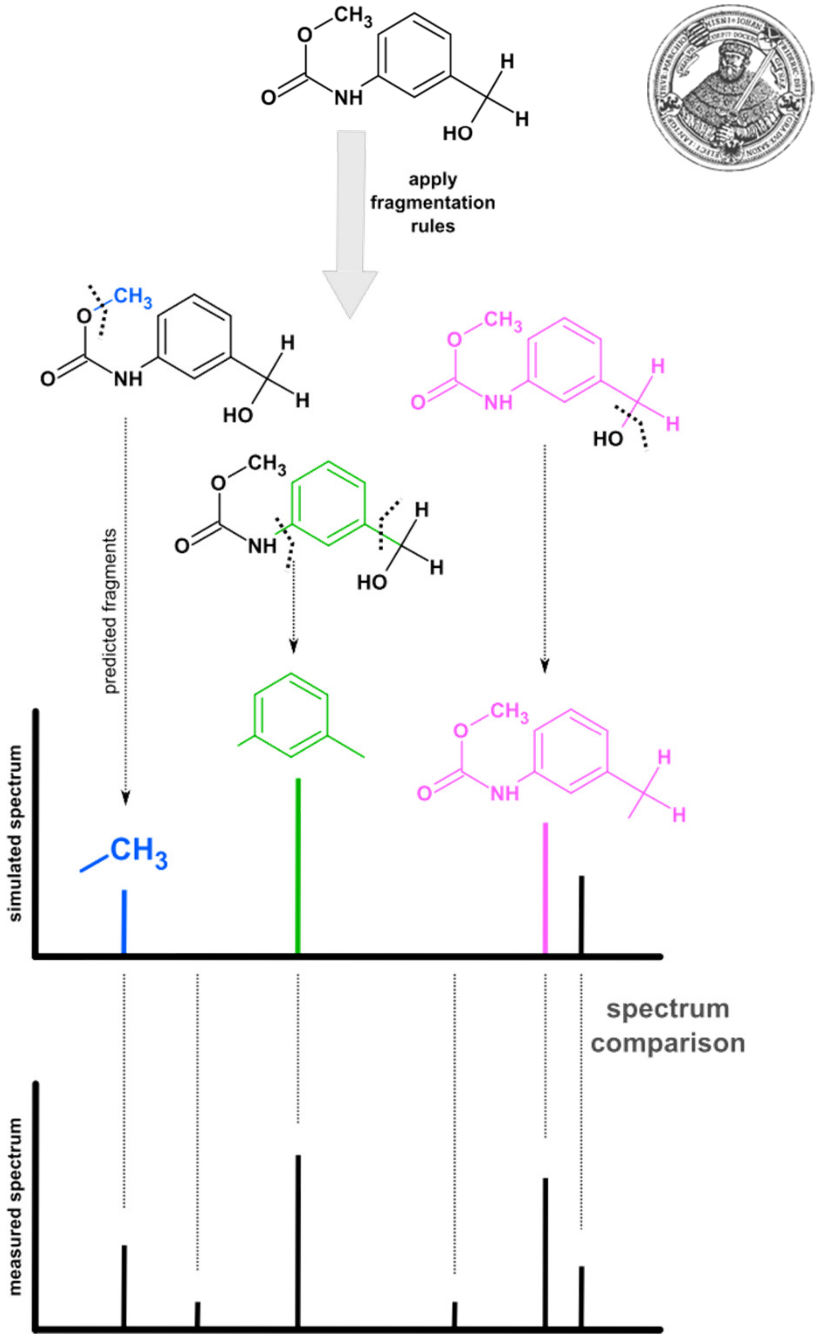
The classic: rule-based prediction

Rule-based prediction



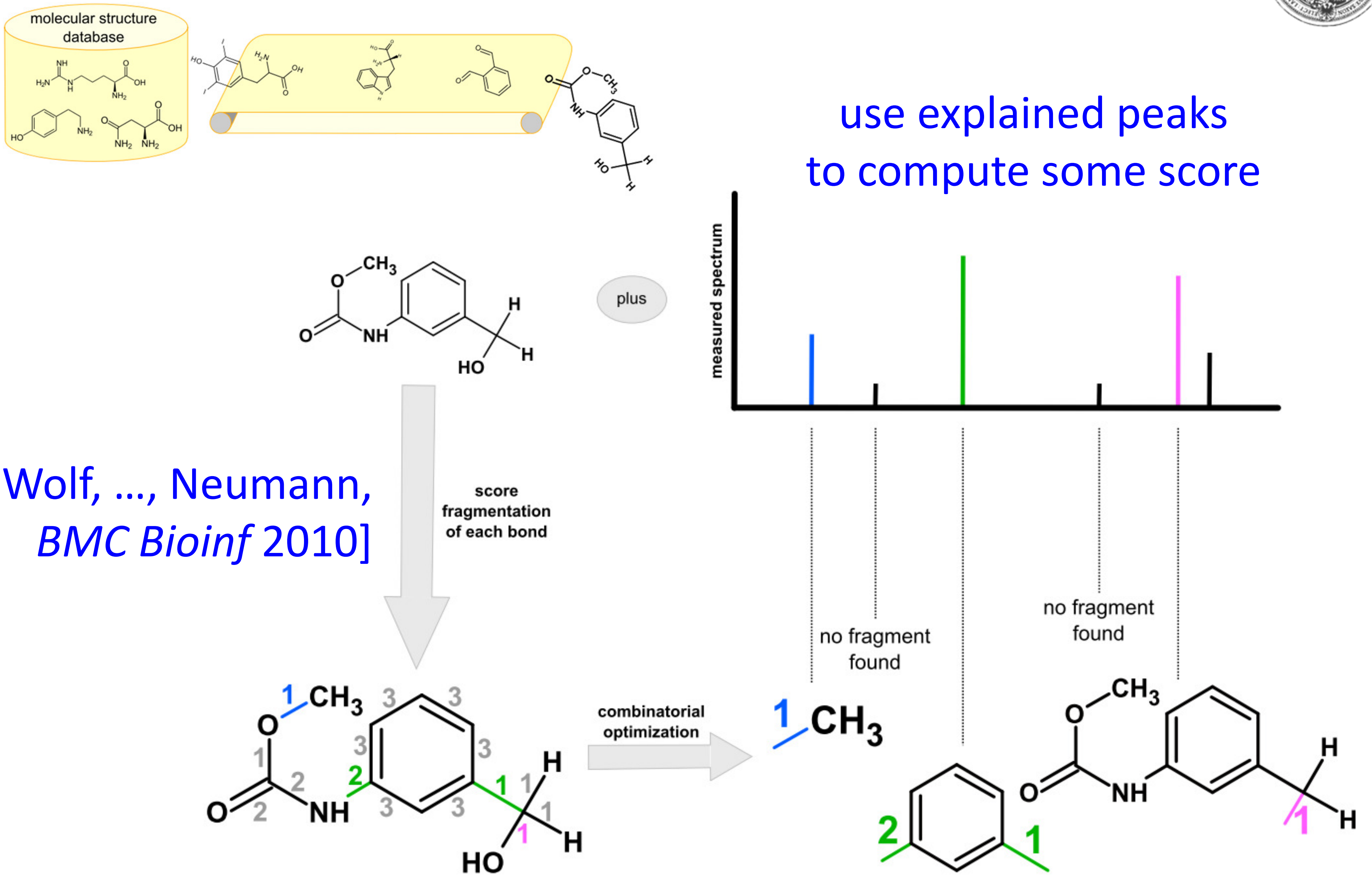


[Hill, ..., Grant, *Anal Chem* 2008]



MetFrag: combinatorial fragmentation

MetFrag (Neumann group)


molecular structure database
plus
measured spectrum
score fragmentation of each bond
combinatorial optimization
no fragment found
no fragment found
CH_3
NH
HO
O
H

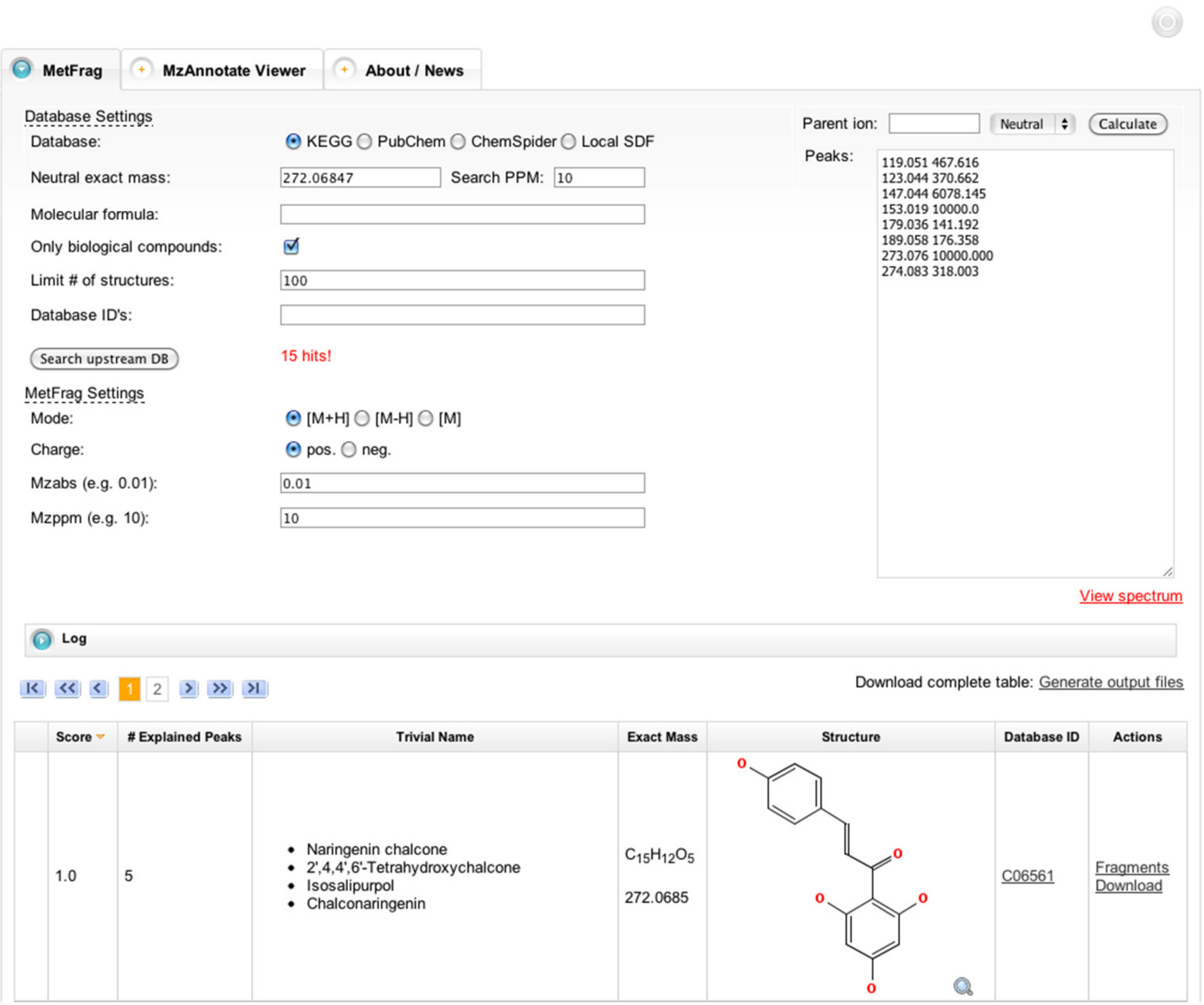
use explained peaks
to compute some score

[Wolf, …, Neumann,
BMC Bioinf 2010]

MetFrag web interface

MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra



Database Settings

Database:	● KEGG ○ PubChem ○ ChemSpider ○ Local SDF
Neutral exact mass:	272.06847 Search PPM: 10
Molecular formula:	
Only biological compounds:	☑
Limit # of structures:	100
Database ID's:	
Search upstream DB	15 hits!

MetFrag Settings

Mode:	● [M+H] ○ [M-H] ○ [M]
Charge:	● pos. ○ neg.
Mzabs (e.g. 0.01):	0.01
Mzppm (e.g. 10):	10

Parent ion: Neutral Calculate

Peaks:

```
119.051 467.616
123.044 370.662
147.044 6078.145
153.019 10000.0
179.036 141.192
189.058 176.358
273.076 10000.000
274.083 318.003
```

View spectrum

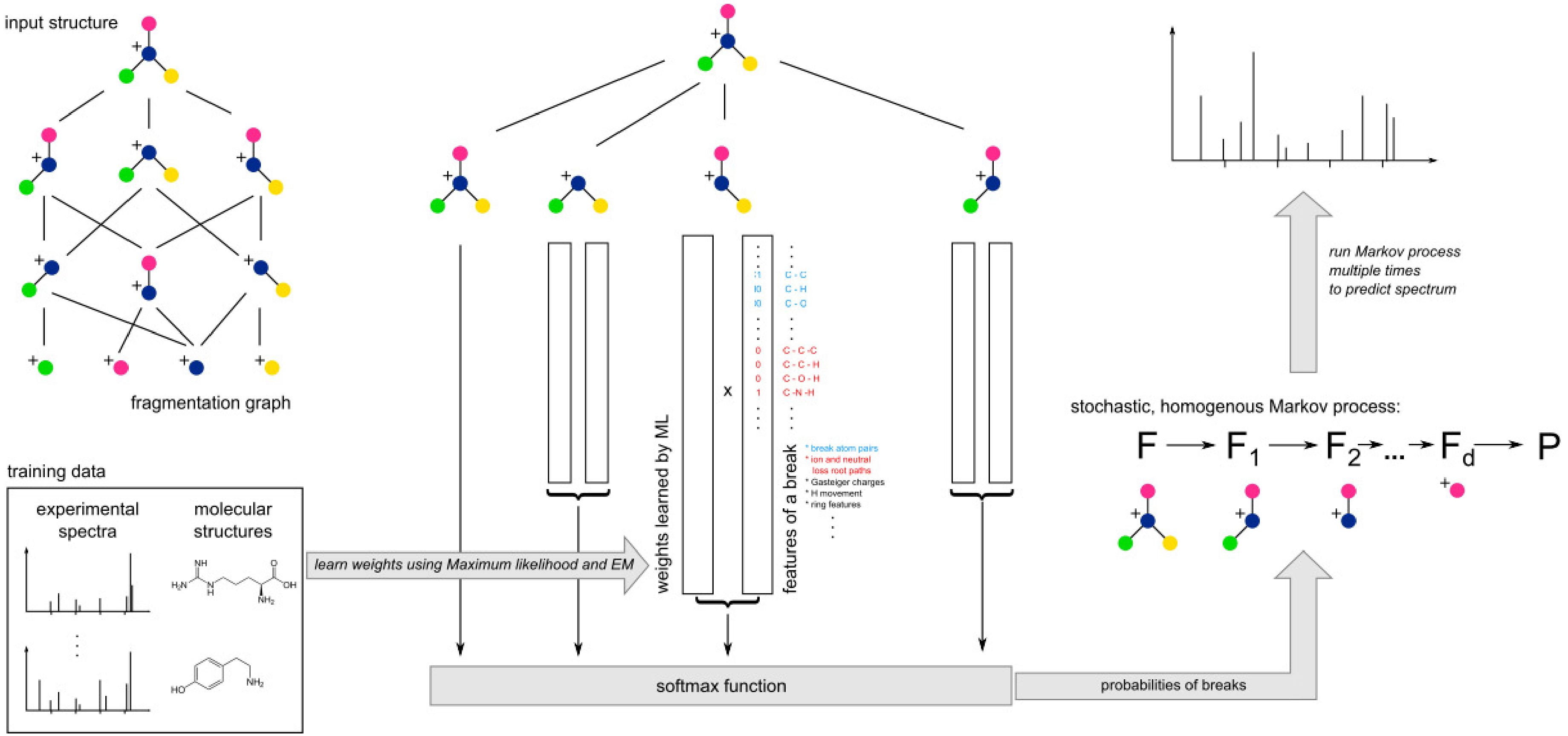
Log

Download complete table: Generate output files

	Score	# Explained Peaks	Trivial Name	Exact Mass	Structure	Database ID	Actions
	1.0	5	• Naringenin chalcone • 2',4,4',6'-Tetrahydroxychalcone • Isosalipurpol • Chalconaringenin	$C_{15}H_{12}O_5$ 272.0685	O O O O O	C06561	Fragments Download

Competitive Fragmentation Modelling

Competitive fragmentation modelling

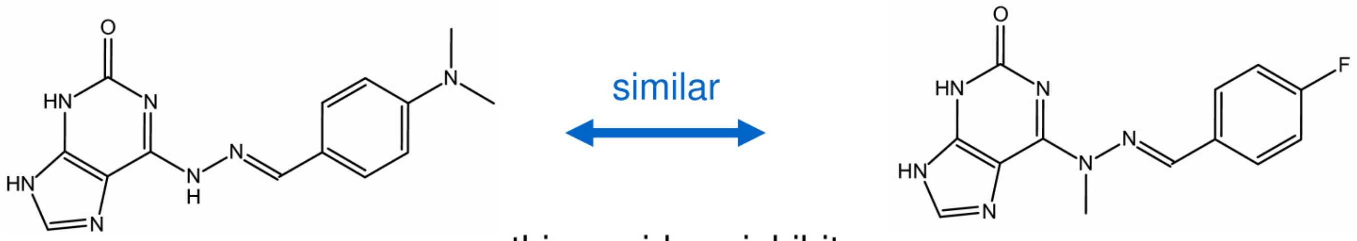


[Allen, Greiner, Wishart, *Metabolomics* 2014]

FingerID: predicting fingerprints

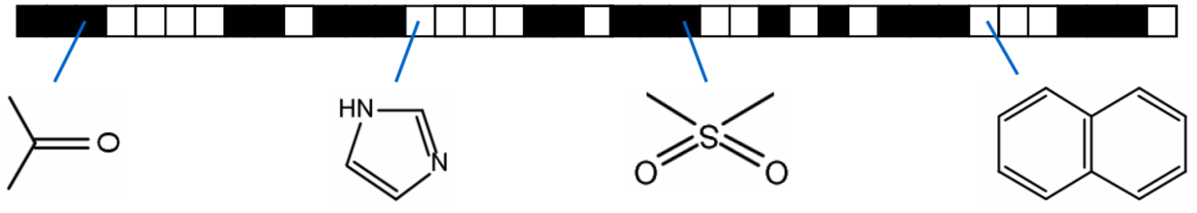
Molecular fingerprints

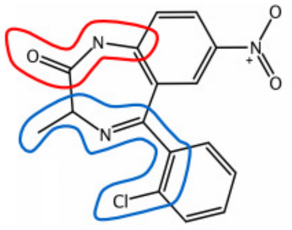
- when are two molecules “similar”?



xanthine oxidase inhibitors

- encode presence/absence of substructures in binary vector

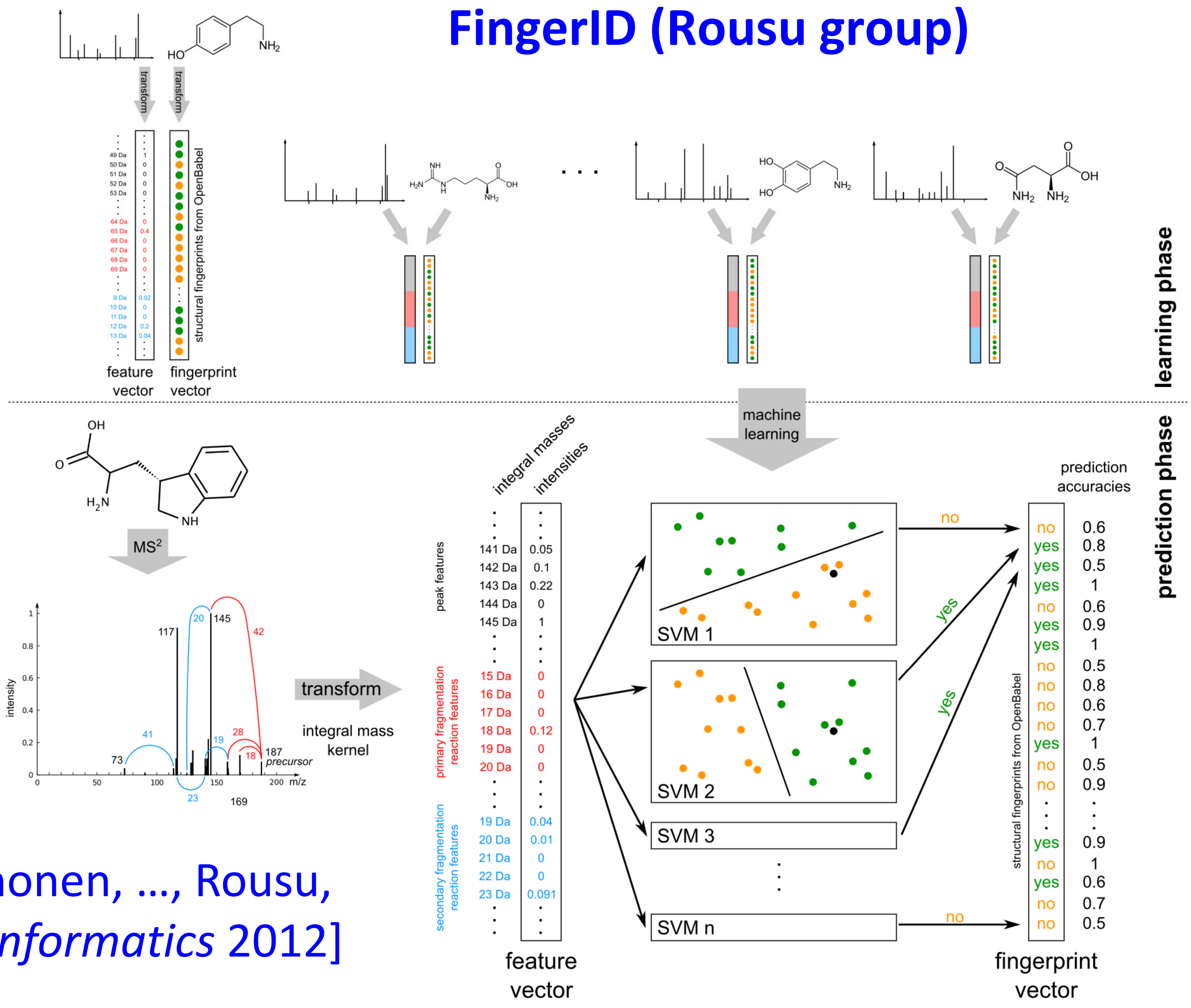




- different types: MACCS, FP1 – FP4, PubChem, …
- used for: virtual screening, estimating chemical similarity, …

Can you predict the molecular fingerprint of an unknown compound directly from the tandem MS data?

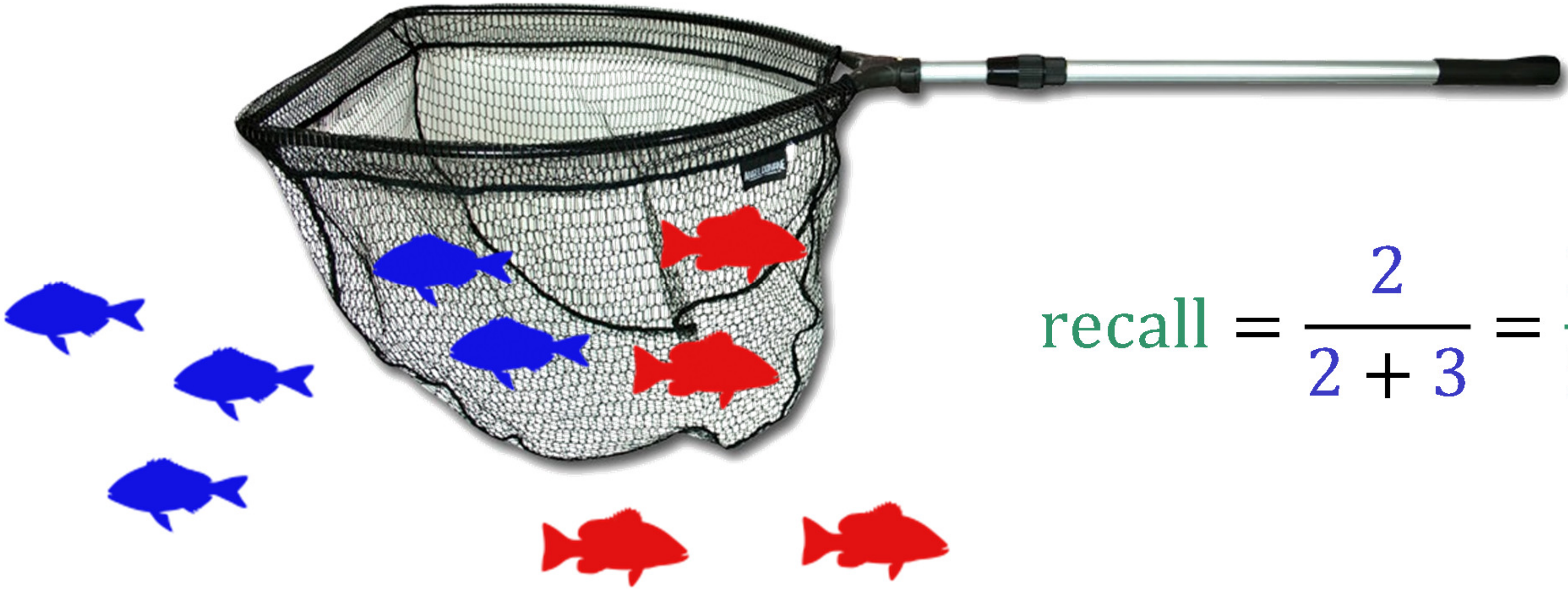
FingerID (Rousu group)



[Heinonen, …, Rousu, *Bioinformatics* 2012]

Precision, Recall, F-score

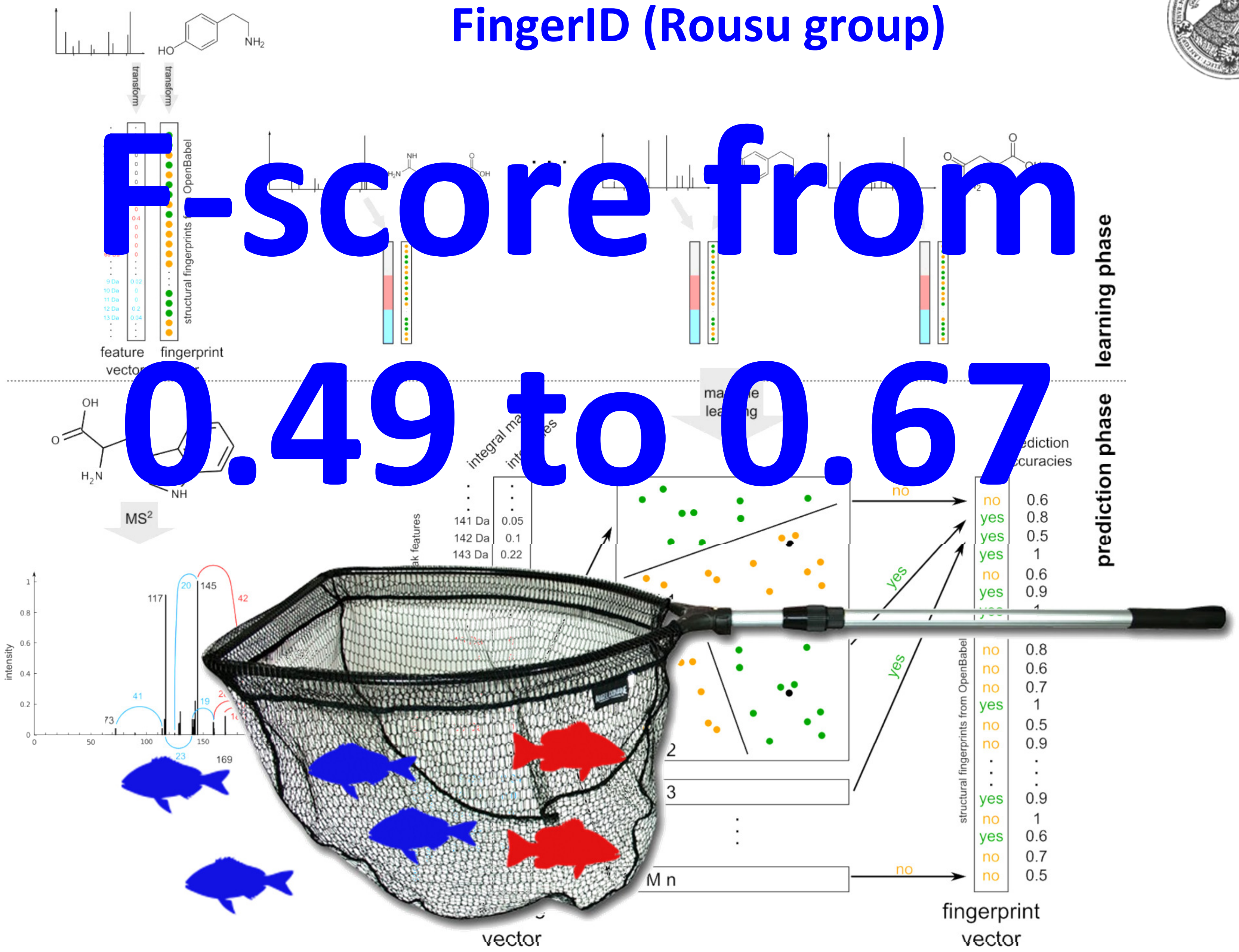
$$\text{precision} = \frac{2}{2+2} = \frac{1}{2}$$



$$\text{recall} = \frac{2}{2+3} = \frac{2}{5}$$

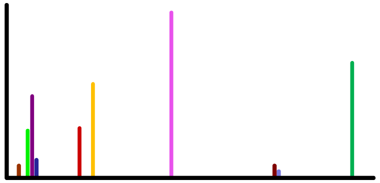
$$\text{F-score} = \frac{2}{2/1 + 5/2} = \frac{4}{9}$$

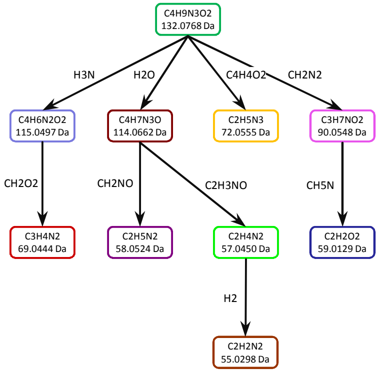
FingerID (Rousu group)

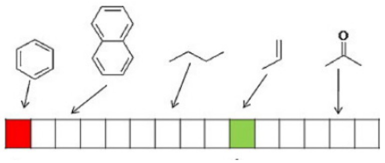


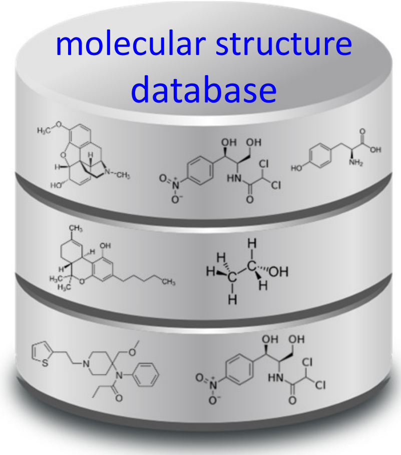
F-score from 0.49 to 0.67

Our method

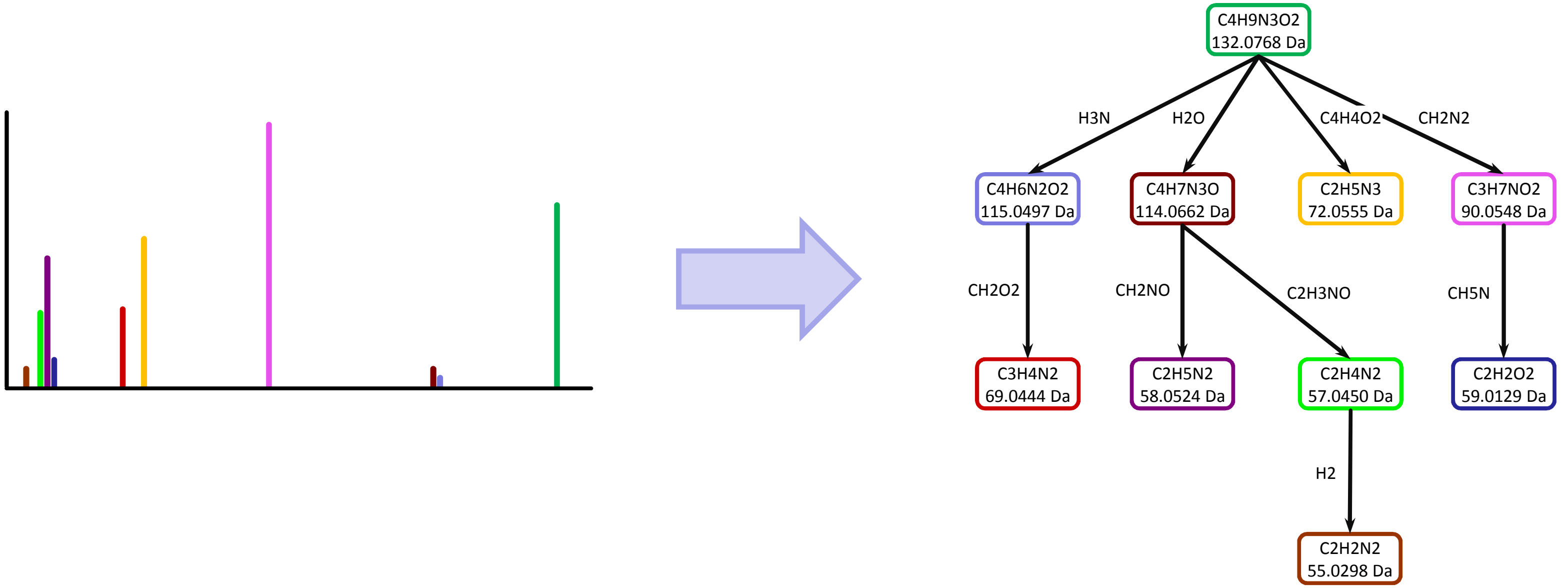



C4H9N3O2
132.0768 Da
H3N
H2O
C4H4O2
CH2N2
C4H6N2O2
115.0497 Da
C4H7N3O
114.0662 Da
C2H5N3
72.0555 Da
C3H7NO2
90.0548 Da
CH2O2
CH2NO
C2H3NO
CH5N
C3H4N2
69.0444 Da
C2H5N2
58.0524 Da
C2H4N2
57.0450 Da
C2H2N2O2
59.0129 Da
H2
C2H2N2
55.0298 Da




molecular structure database

1st step Fragmentation trees



C4H9N3O2 132.0768 Da

H3N

H2O

C4H4O2

CH2N2

C4H6N2O2 115.0497 Da

C4H7N3O 114.0662 Da

C2H5N3 72.0555 Da

C3H7NO2 90.0548 Da

CH2O2

CH2NO

C2H3NO

CH5N

C3H4N2 69.0444 Da

C2H5N2 58.0524 Da

C2H4N2 57.0450 Da

C2H2O2 59.0129 Da

H2

C2H2N2 55.0298 Da

Our fragmentation trees

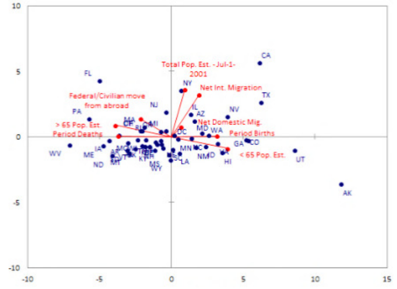
- tandem MS, multiple MS not required

- fully automated method

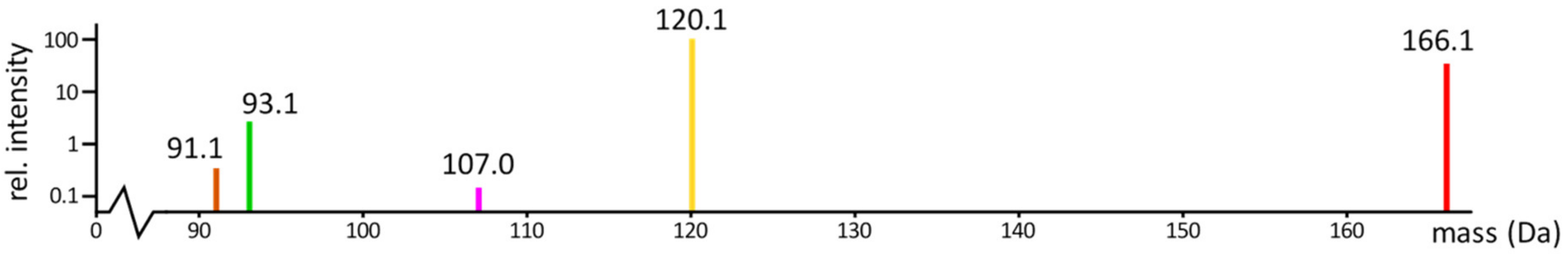
- best explains experimental data

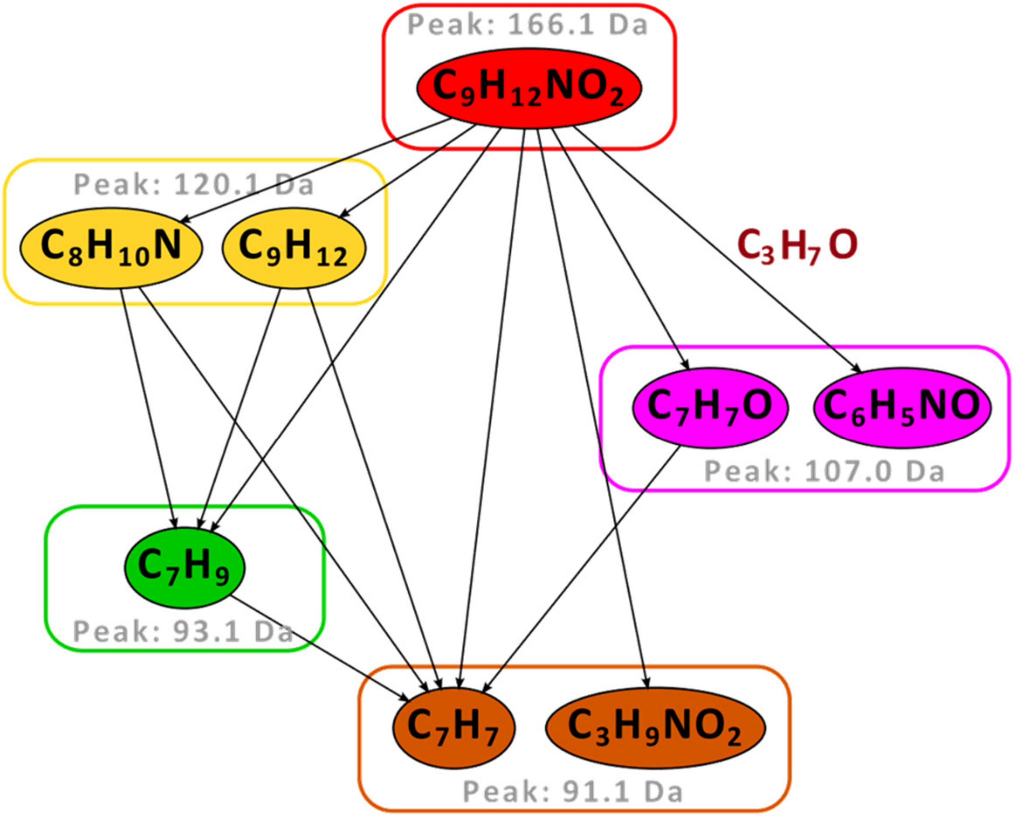
- combinatorial optimization

MS^n not required



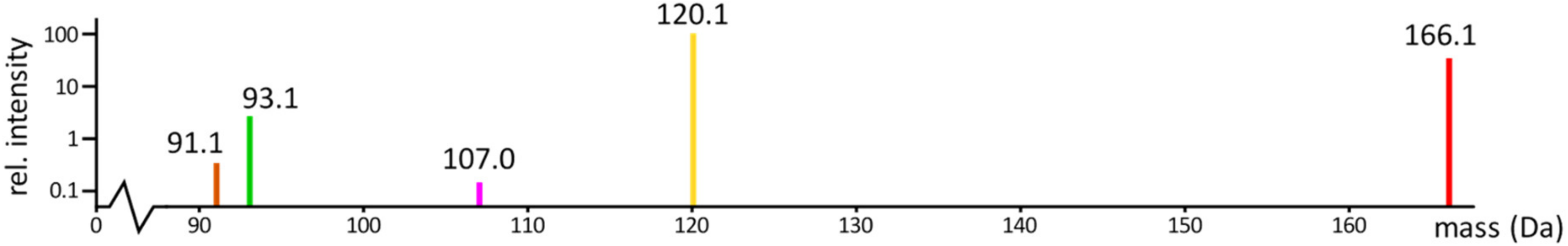
Our fragmentation trees

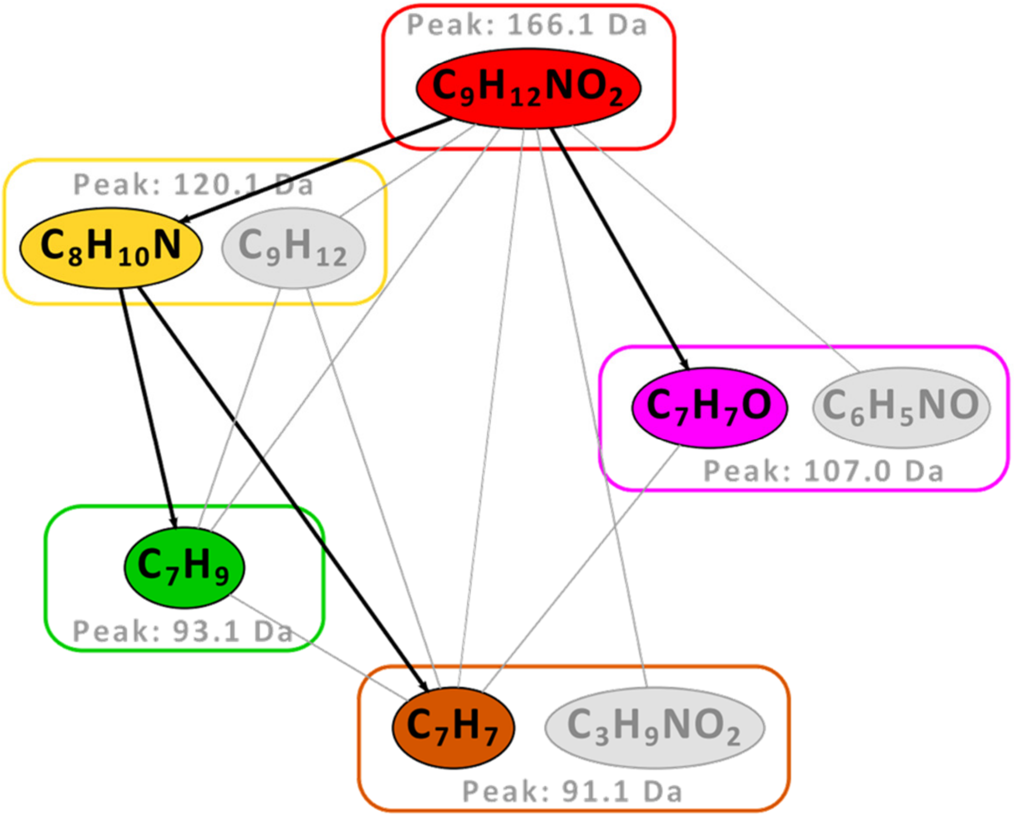




- Böcker and Rasche, *Bioinf* 2008
- Rasche *et al.*, *Anal Chem* 2011
- Dührkop and Böcker, unpublished
- White *et al.*, unpublished

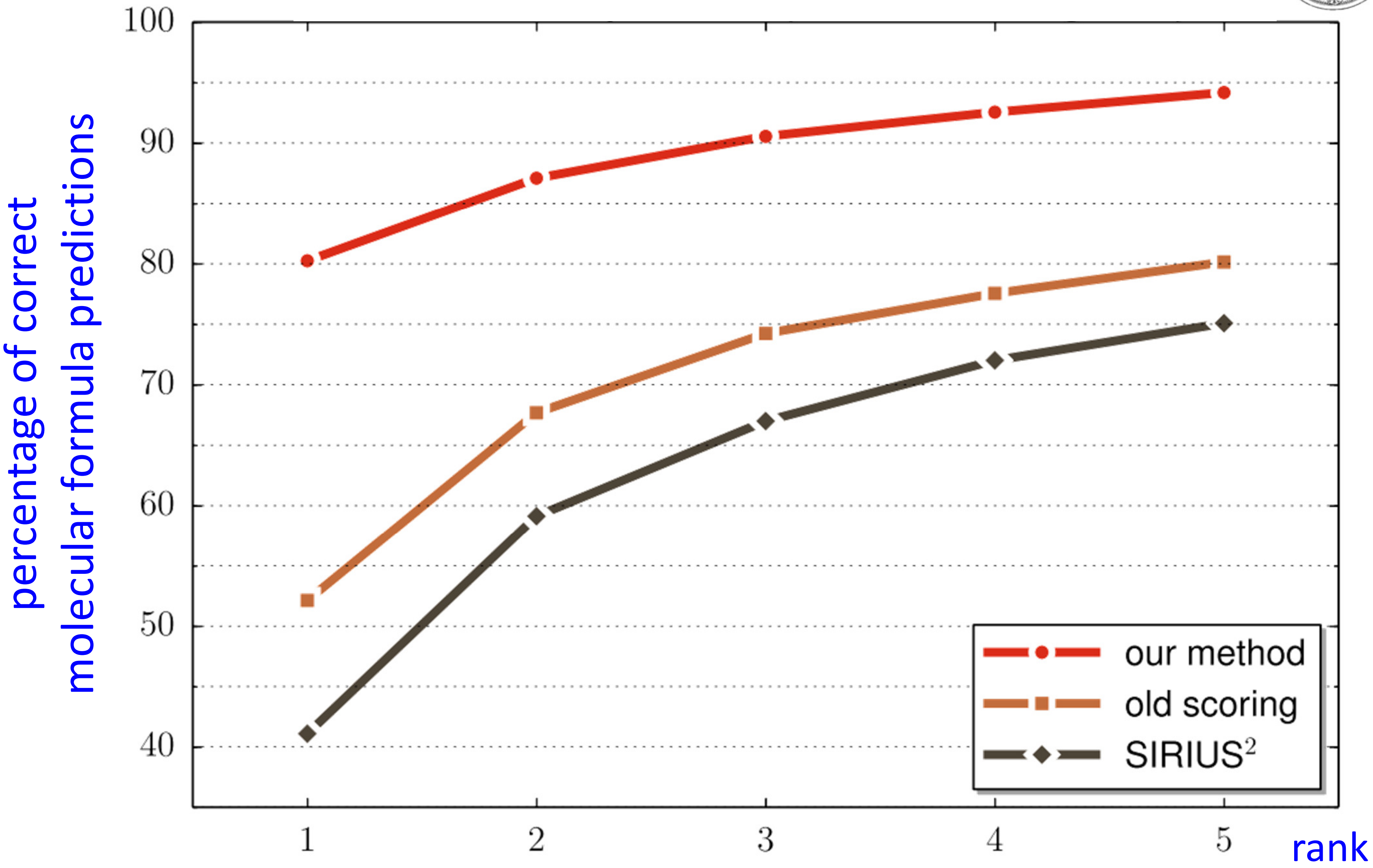
Our fragmentation trees



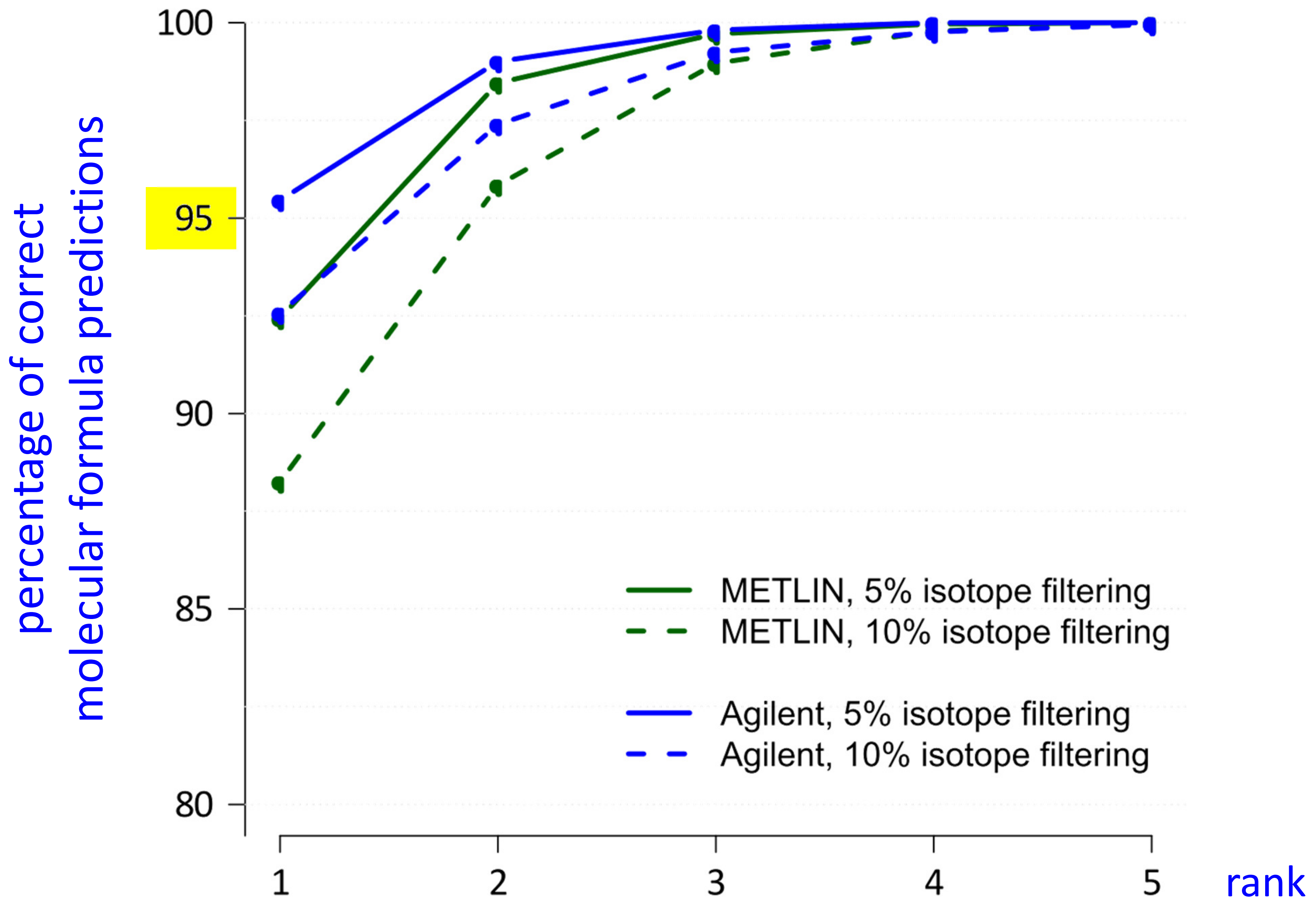


- Böcker and Rasche, *Bioinf* 2008
- Rasche *et al.*, *Anal Chem* 2011
- Dührkop and Böcker, unpublished
- White *et al.*, unpublished

Molecular formula prediction



Molecular formulas with isotope patterns

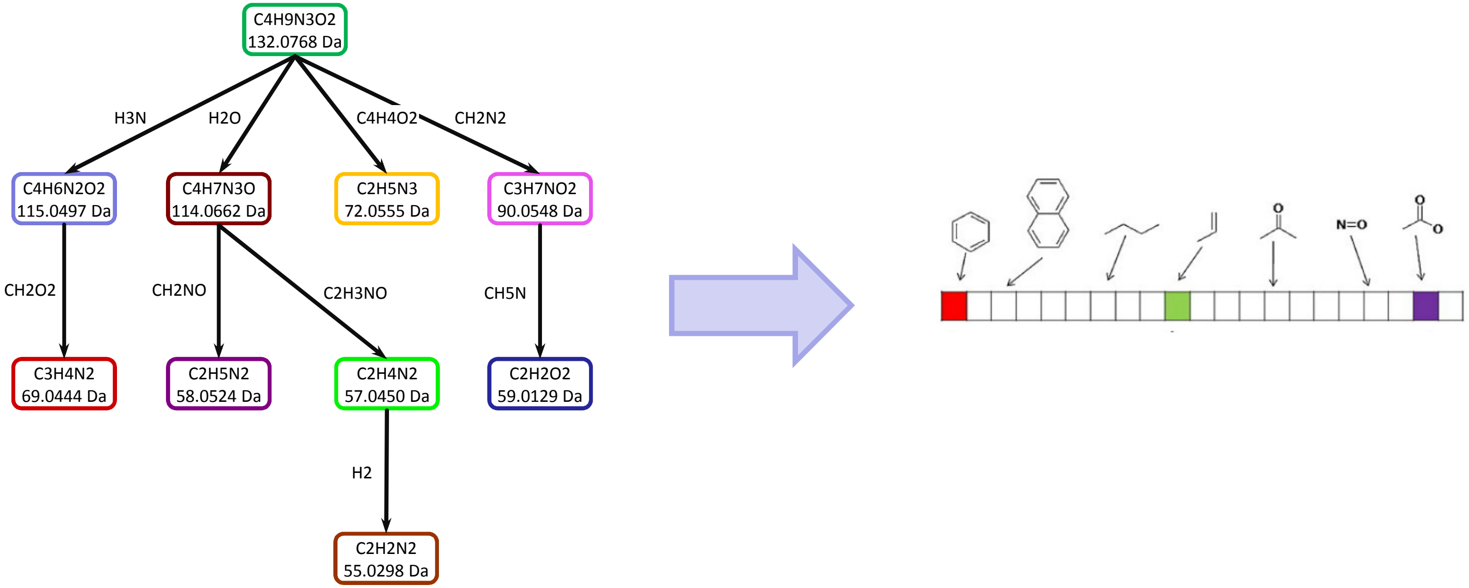


CASMI challenge 2013

- Critical Assessment of Small Molecule Identification, http://www.casmi-contest.org
- we got 12 out of 14 molecular formulas correct
- 2nd place, winner manually analyzed the challenges
- we were the only contestants that **did not search PubChem**, but instead considered all possible molecular formulas

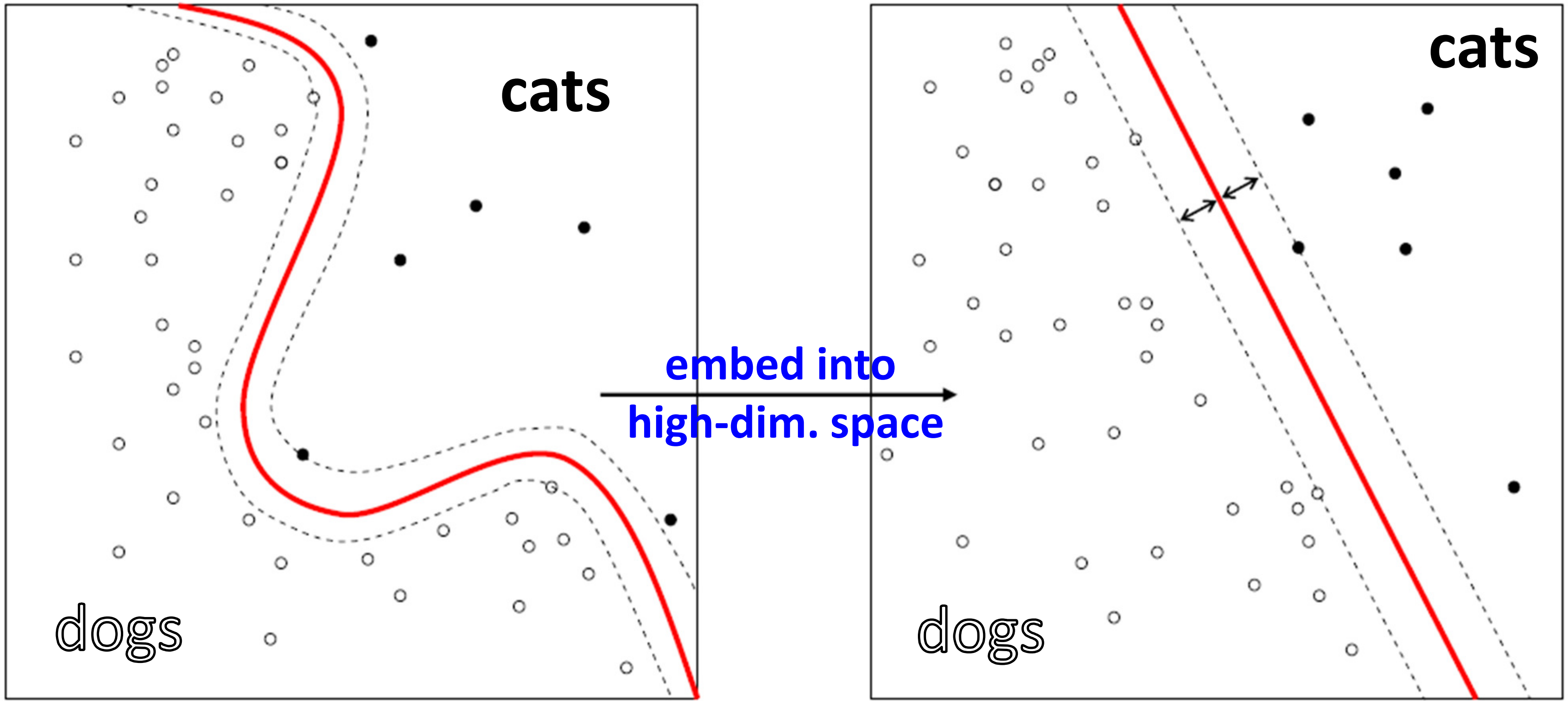
2nd step

Fingerprints



Machine Learning

Support Vector Machines



cats

dogs

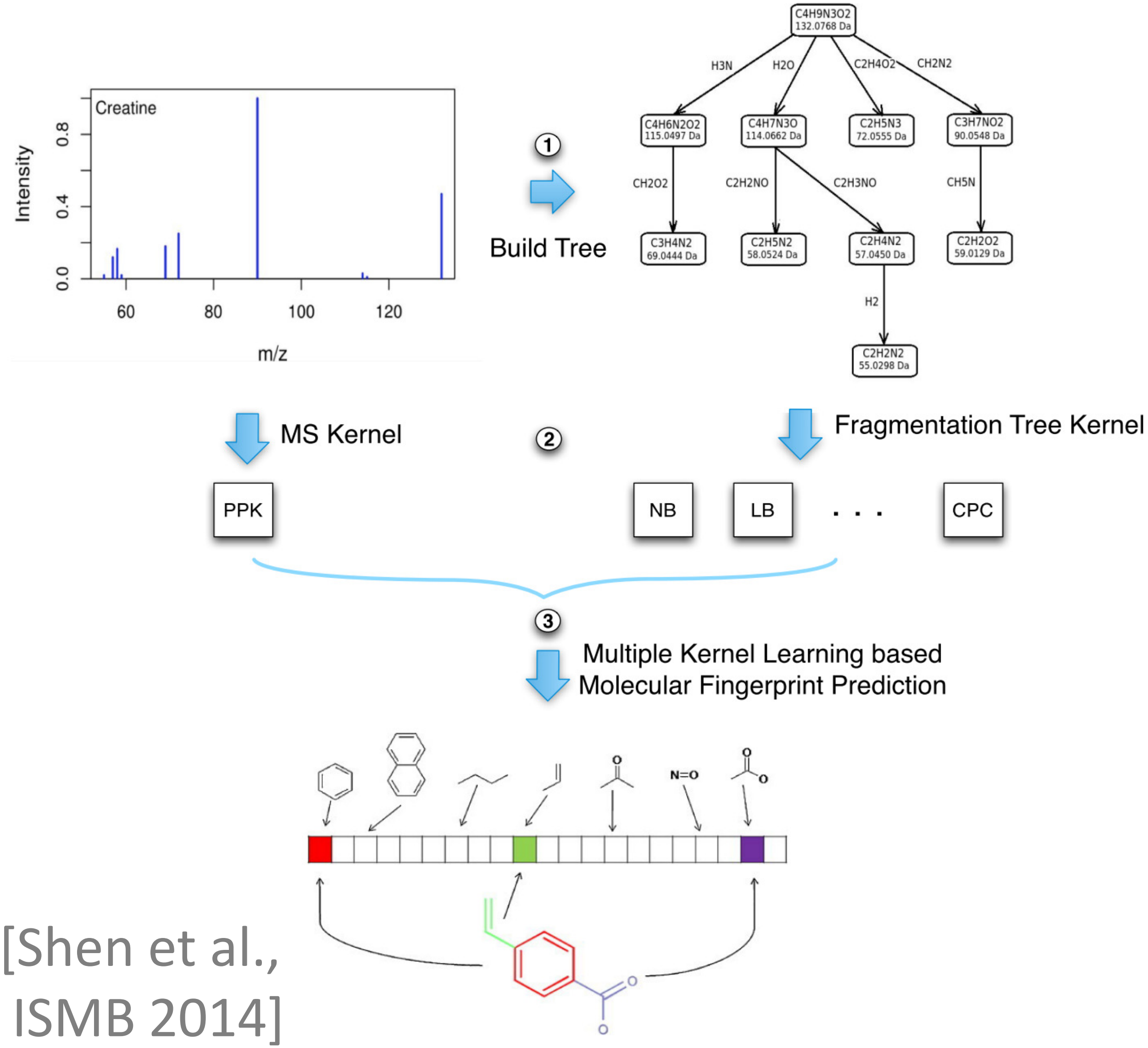
embed into high-dim. space

cats

dogs

- separate cats and dogs via features (weight, height, …)
- map features so that linear separation is possible

Use fragmentation trees as input



Build Tree

MS Kernel

Fragmentation Tree Kernel

PPK

NB

LB

. . .

CPC

Multiple Kernel Learning based Molecular Fingerprint Prediction

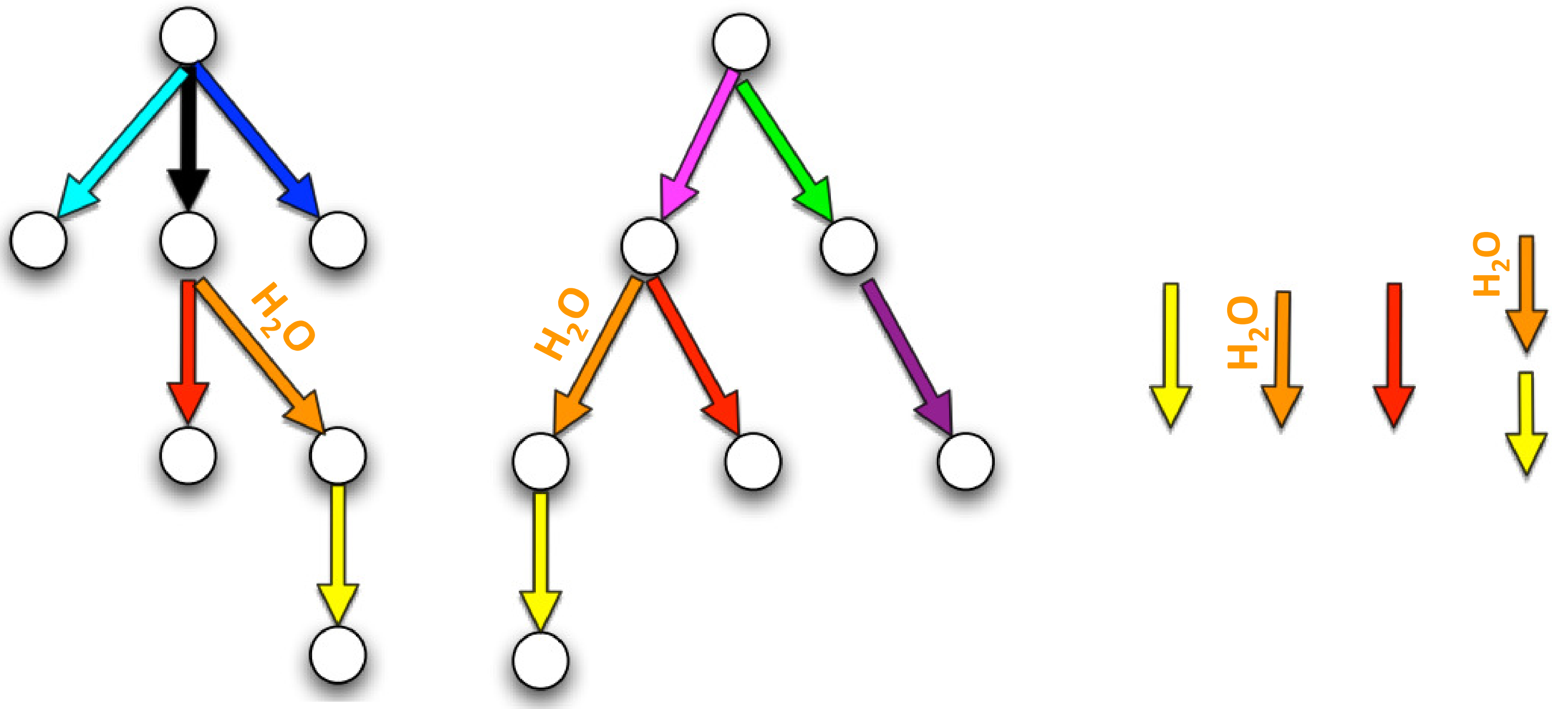
[Shen et al., ISMB 2014]

FT **structure** kernels

- nodes binary
- nodes intensity
- loss binary
- loss count
- loss intensity
- root loss binary
- root loss intensity
- common path counting
- common paths of length 2
- common paths with peak scores
- common subtree counting

One example: Common Path Counting kernel

- tree kernels measure the structural similarity of two fragmentation trees



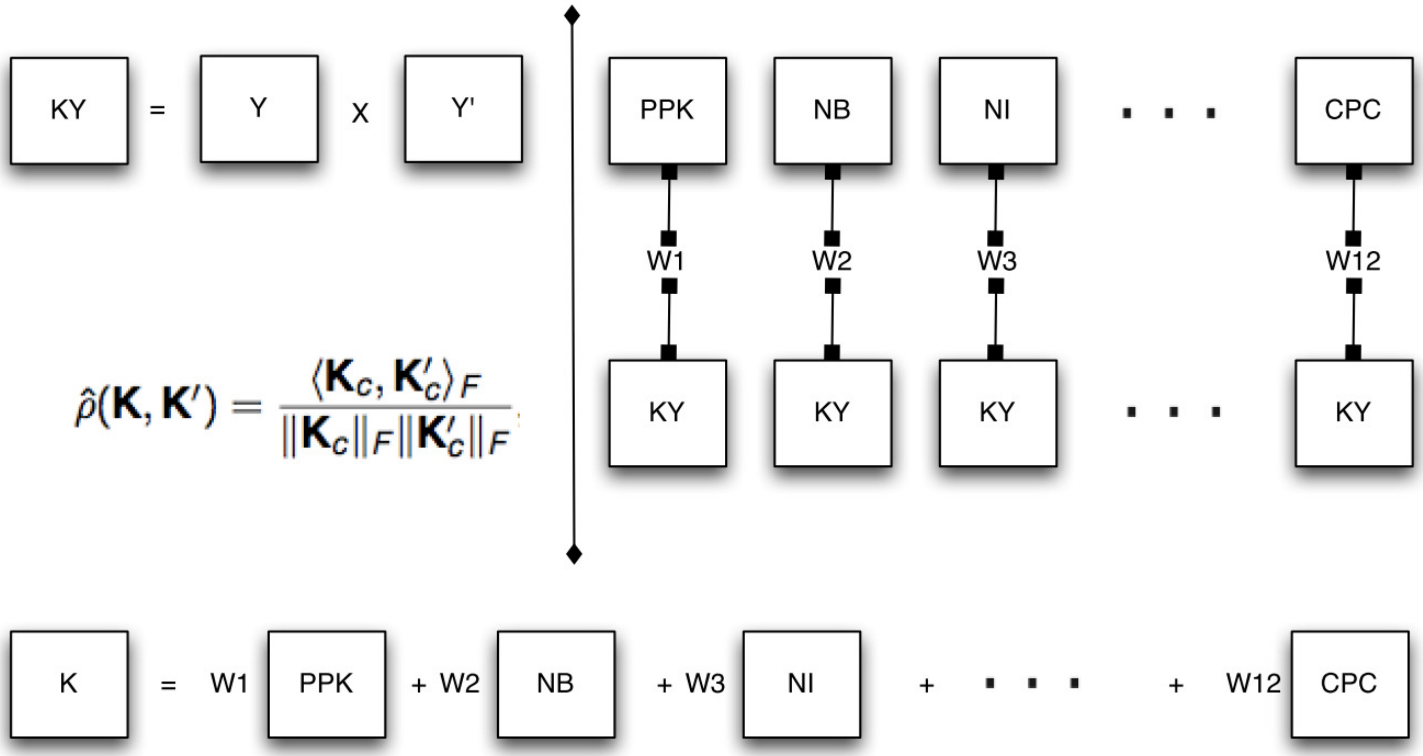
1st fragmentation tree

2nd fragmentation tree

4 common paths

Multiple kernel learning

- combine predictions of all 12 kernels into one new kernel
- ALIGNF: Learn weights by comparing kernels to target kernel

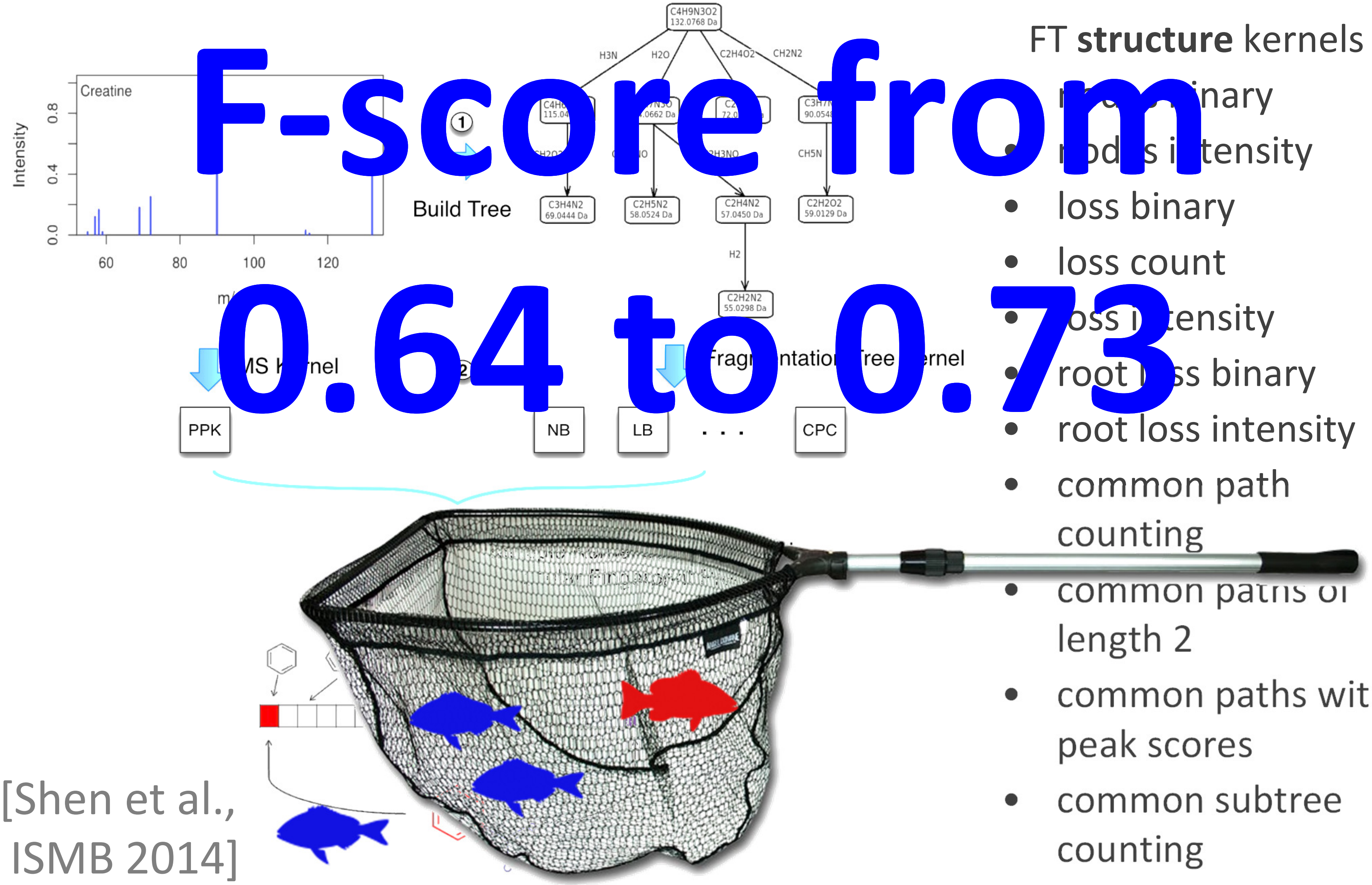


$$\hat{\rho}(\mathbf{K}, \mathbf{K}') = \frac{\langle \mathbf{K}_c, \mathbf{K}'_c \rangle_F}{\|\mathbf{K}_c\|_F \|\mathbf{K}'_c\|_F}$$

Was it all worth it?

Use fragmentation trees as input

F-score from 0.64 to 0.73



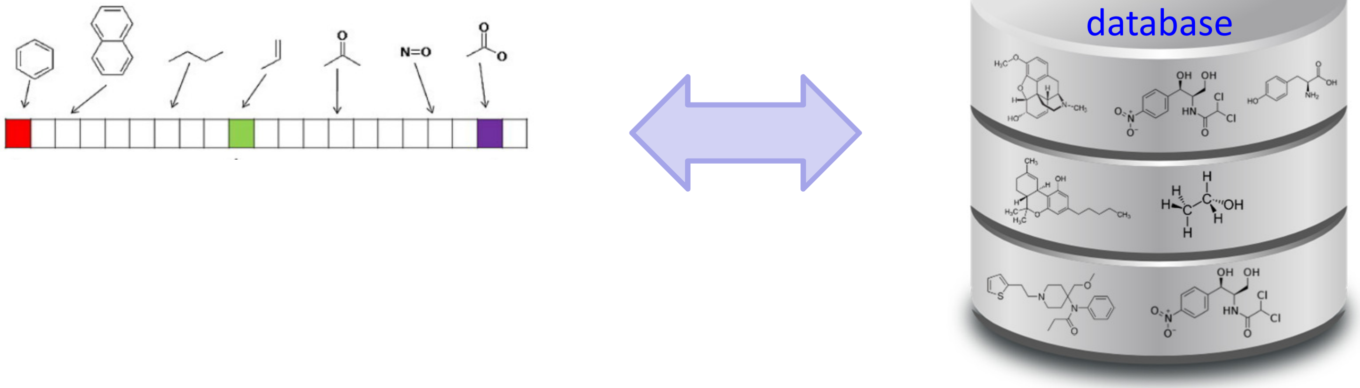
FT **structure** kernels

- node binary
- nodes intensity
- loss binary
- loss count
- loss intensity
- root loss binary
- root loss intensity
- common path counting
- common paths of length 2
- common paths with peak scores
- common subtree counting

[Shen et al., ISMB 2014]

3rd step

Searching PubChem



Searching a molecular structure database

- retrieve all compounds with correct molecular formula
- for each compound in the database, we know its structure and, hence, we know its correct molecular fingerprint
- compare predicted fingerprint to those of all candidates
- simplest score is unit costs

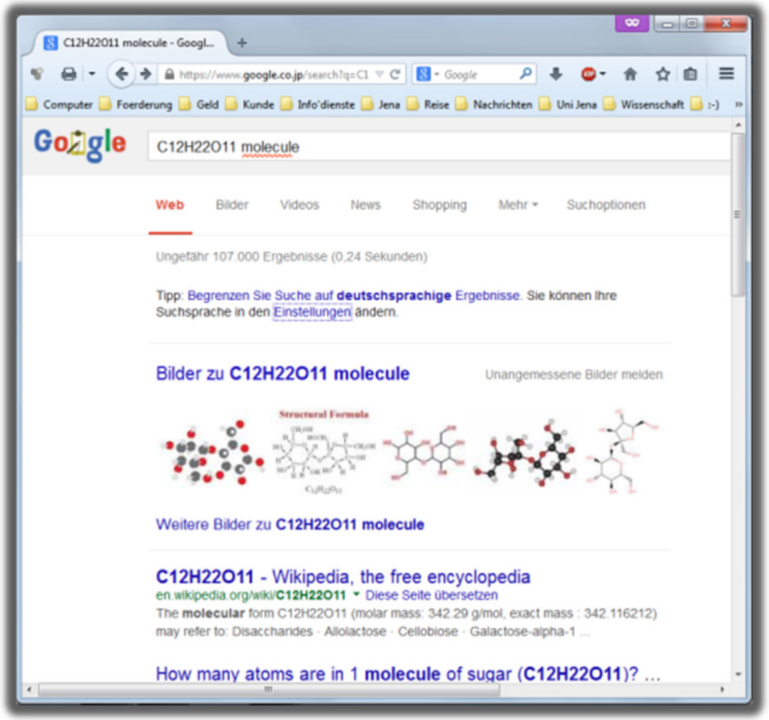
predicted fingerprint	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0
candidate 1853	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	1	0	0
differences: 4	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓

- rank candidates according to score

rank	1st	2nd	3rd	4th	…
candidate	765	2271	1853	61	…
differences	1	3	4	7	…

Evaluation setup

- “Evaluation is the process of judging something or someone based on a set of standards.”
- retrieve all compounds with correct molecular formula
- for each compound in the evaluation dataset, we know its correct molecular structure
- at what position do we find the correct answer? (TOP-*k*)
- we only evaluate plain structures (**no stereochemistry** etc)



C12H22O11 molecule

Ungefähr 107.000 Ergebnisse (0,24 Sekunden)

Tipp: Begrenzen Sie Suche auf **deutschsprachige** Ergebnisse. Sie können Ihre Suchsprache in den Einstellungen ändern.

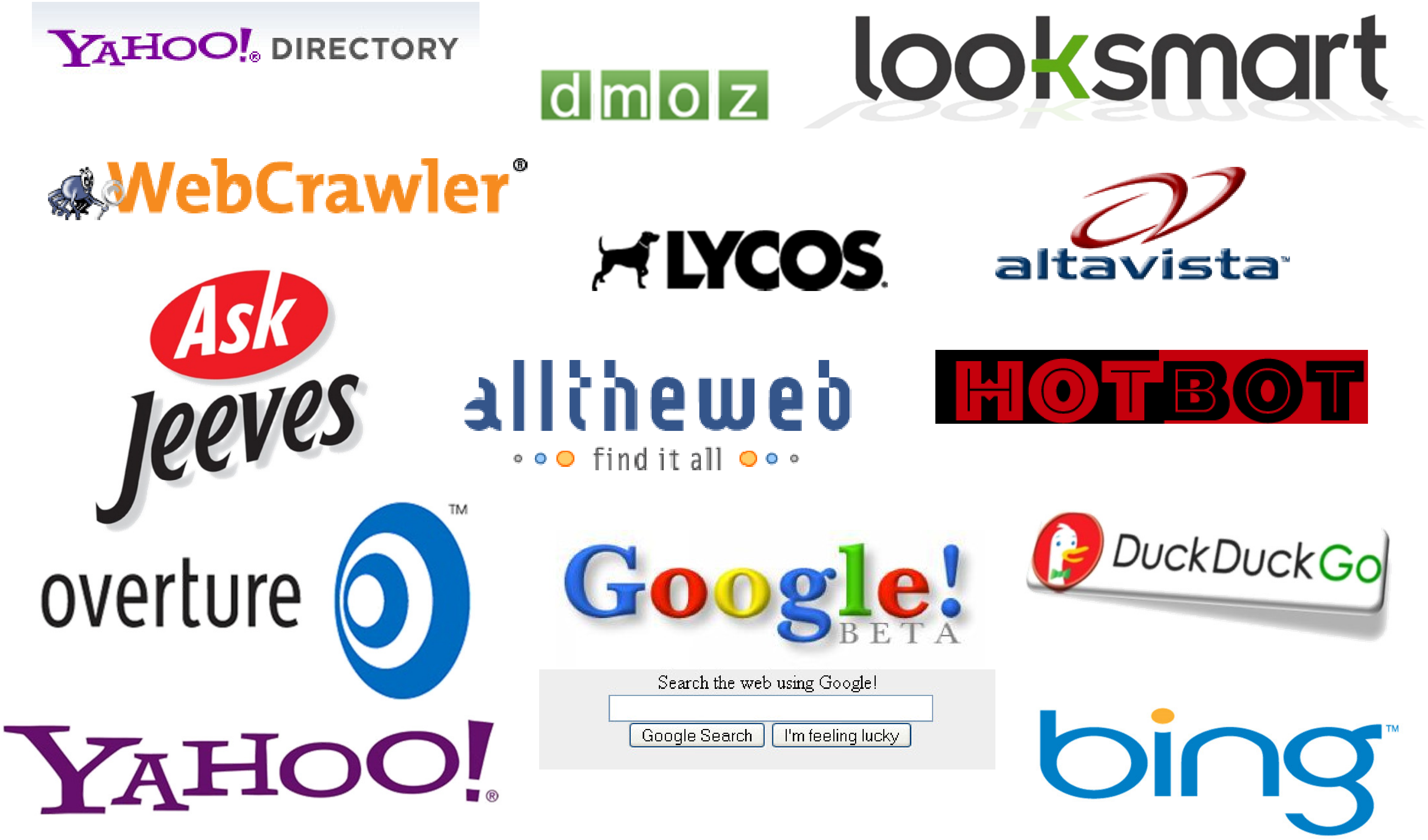
Bilder zu **C12H22O11 molecule** Unangemessene Bilder melden

Weitere Bilder zu **C12H22O11 molecule**

C12H22O11 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**C12H22O11** Diese Seite übersetzen
The **molecular** form C12H22O11 (molar mass: 342.29 g/mol, exact mass : 342.116212) may refer to: Disaccharides · Allolactose · Cellobiose · Galactose-alpha-1 ...

How many atoms are in 1 **molecule** of sugar (**C12H22O11**)? ...

Intermission: WWW search engines

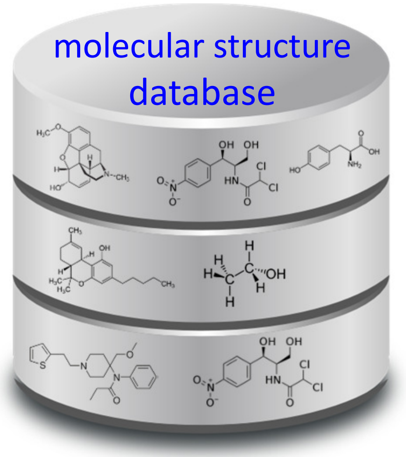


Search the web using Google!

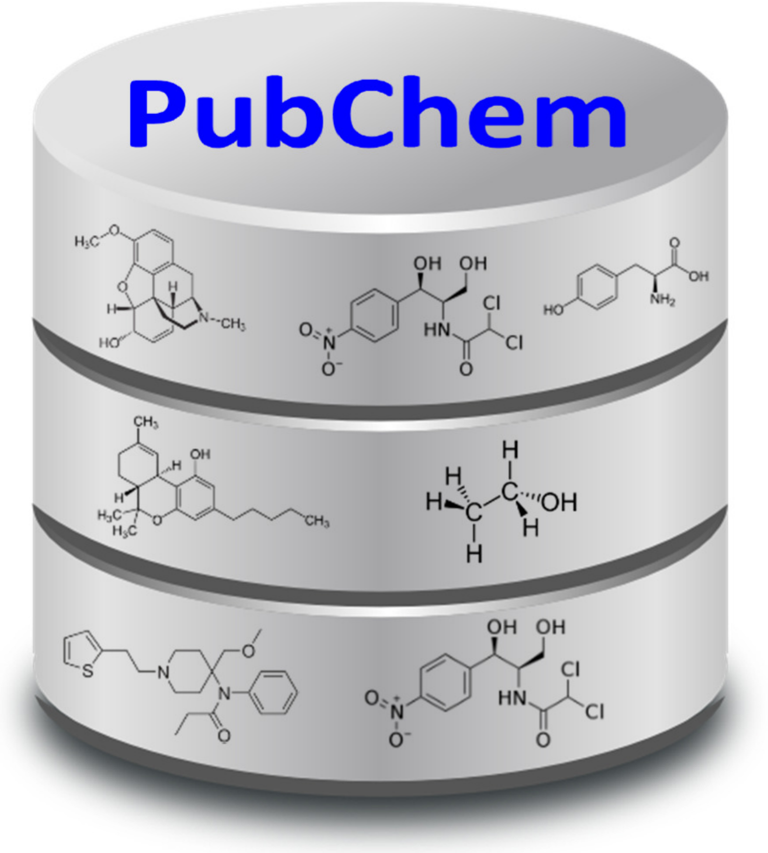
Training and cross validation data

- GnPS database (UCSD, San Diego)
- Forensic database (Agilent Technologies)
- QTOF MS instruments
- ≈ 2800 + 2200 = 5000 compounds
- tandem mass spectra (CID) at different frag. energies
- mass accuracy usually 10 ppm or better
- used to train and evaluate: 10x cross validation

Where to search: molecular structure databases



- PubChem compounds that have a citation in PubMed
- plus HMDB, Knapsack, ChEBI, METLIN, contaminants
- total 400 000 compounds



- full PubChem: more than 50 million compounds

Conclusion

- searching in molecular structure dbs using tandem MS data has become an option
- for the complete PubChem dataset (40 million structures) our method currently reaches 35% hits (correct IDs)

Outlook

- better kernels, better scores, better search results
- significances: False Discovery Rates, q-values, p-values
- and much more to come…

Credits

Kai Dührkop

Huibin Shen

Marvin Meusel

Juho Rousu

Aalto University, Helsinki, Finland

Thank you

- GNPS db, Pieter Dorrestein
- Agilent Technologies

Funding

- Deutsche Forschungsgemeinschaft

Thank you for your attention!