analytica
April 10–13, 2018 • Messe München

Finding Small Molecules in Big Data

EPA

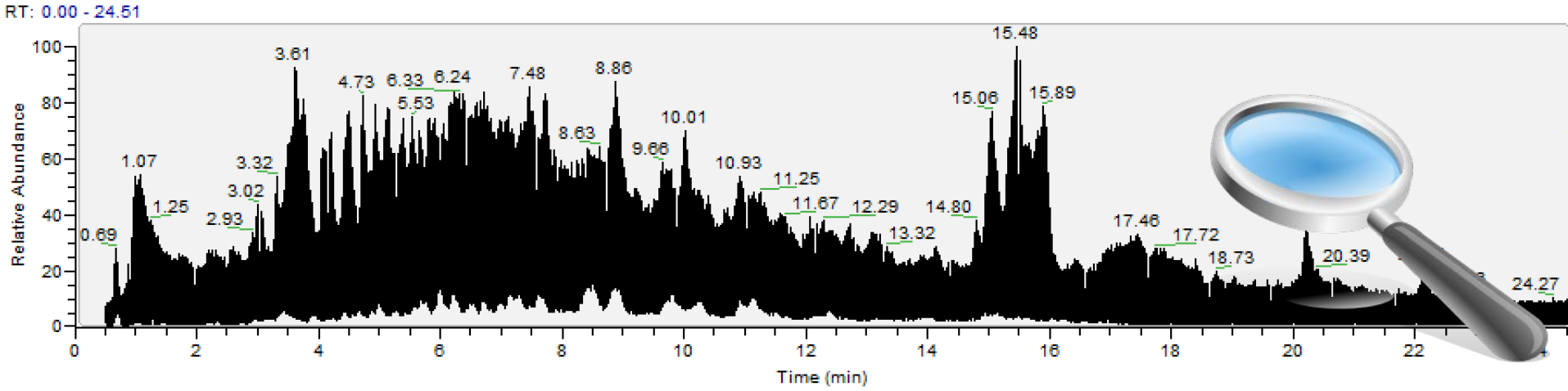


Image © www.seanoakley.com/

Emma Schymanski
Luxembourg Centre for Systems Biomedicine (LCSB),
University of Luxembourg
Email: emma.schymanski@uni.lu

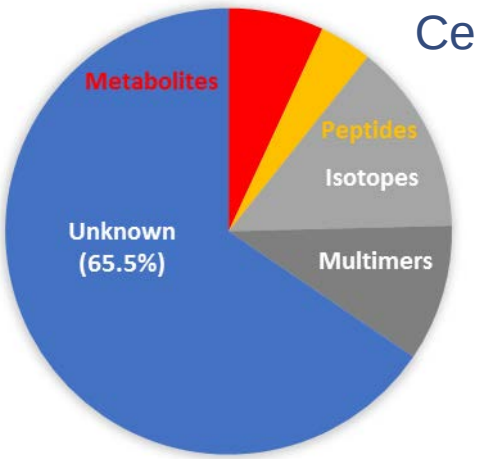
Antony J. Williams
National Centre for Computational Toxicity (NCCT),
US Environmental Protection Agency (US EPA), NC, USA

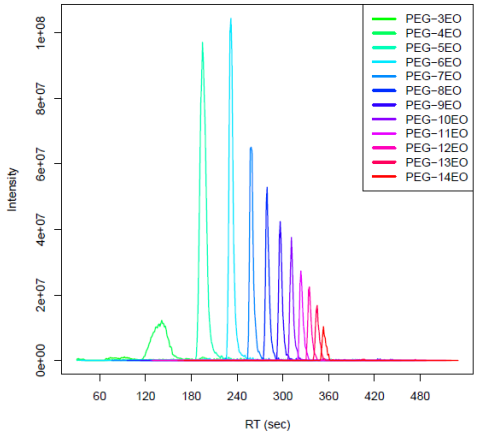
LCSB

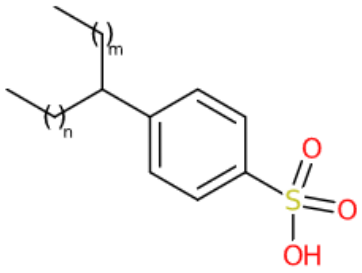
uni.lu UNIVERSITÉ DU LUXEMBOURG

Small molecules … big problems?

- Status quo of small molecules:
  - How many are in compound databases?
  - How many could there be?
  - How many are in spectral libraries?
  - Communicating between libraries
- “Non-target” Identification
  - Metadata
  - Exchanging Information
- Tackling the “unknowns”
  - Exchanging Information on Complex Structures
  - …and how Open Science helps!





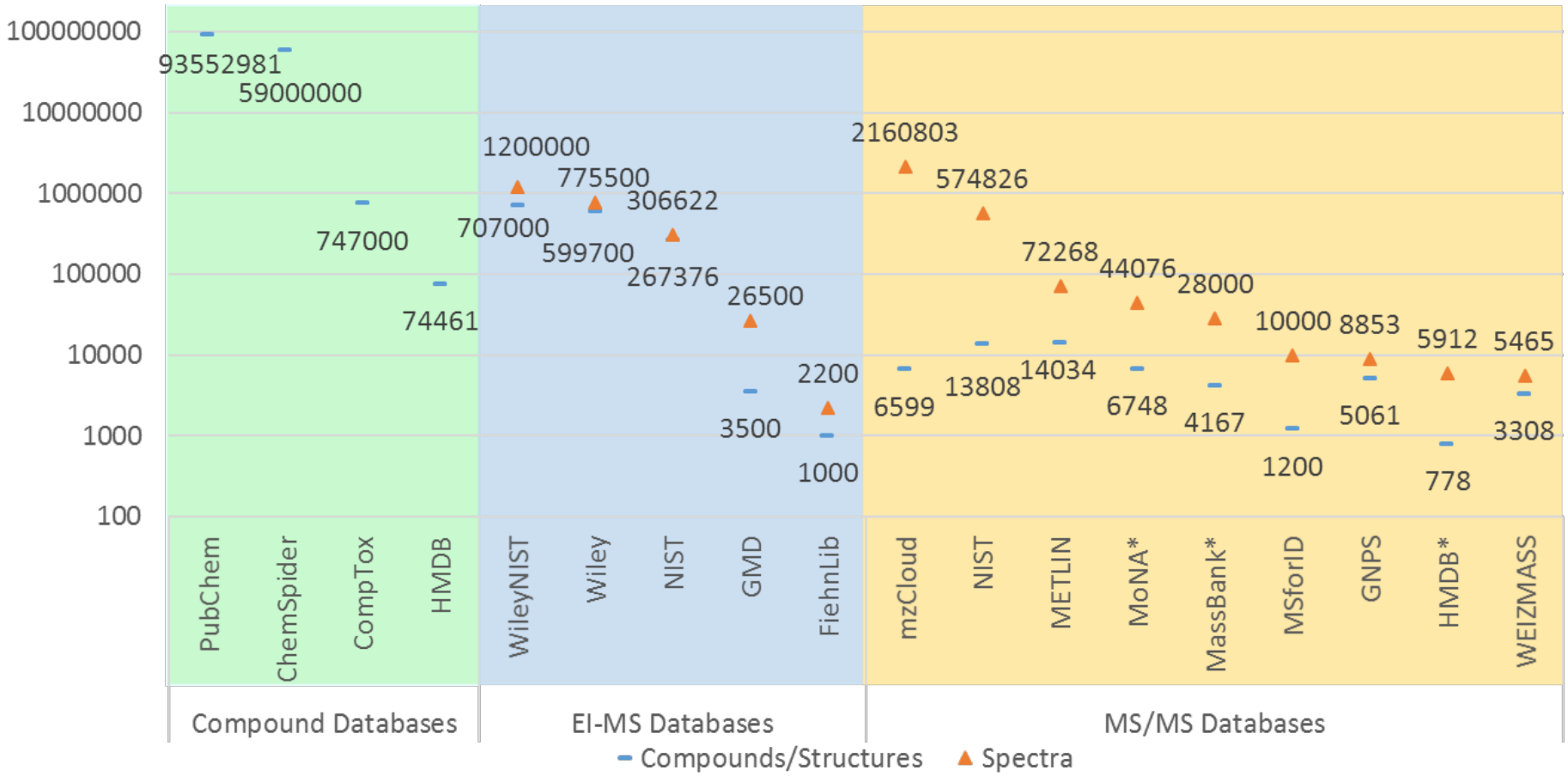


LAS; n+m=7-10

E. coli data provided by N. Zamboni, IMSB, ETH Zürich.

Searching for Small Molecules …

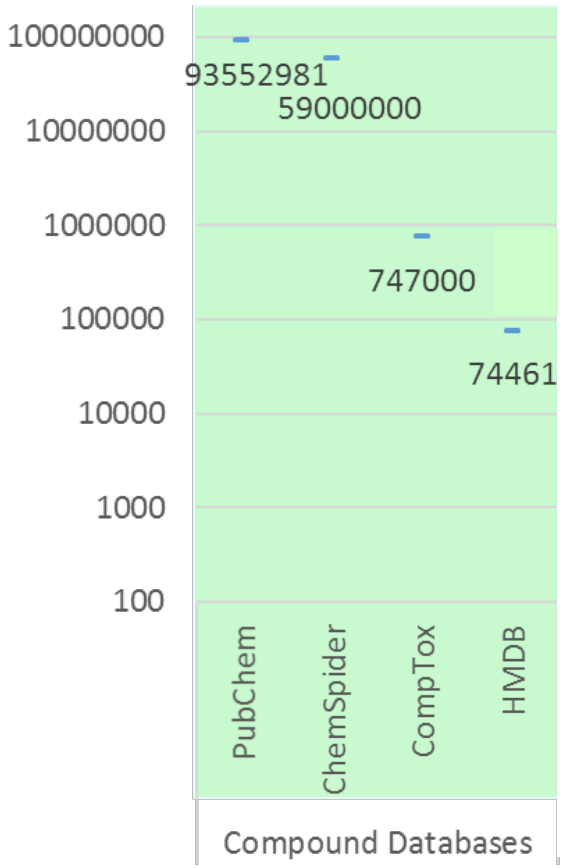
- A subset of current chemical and spectral resources



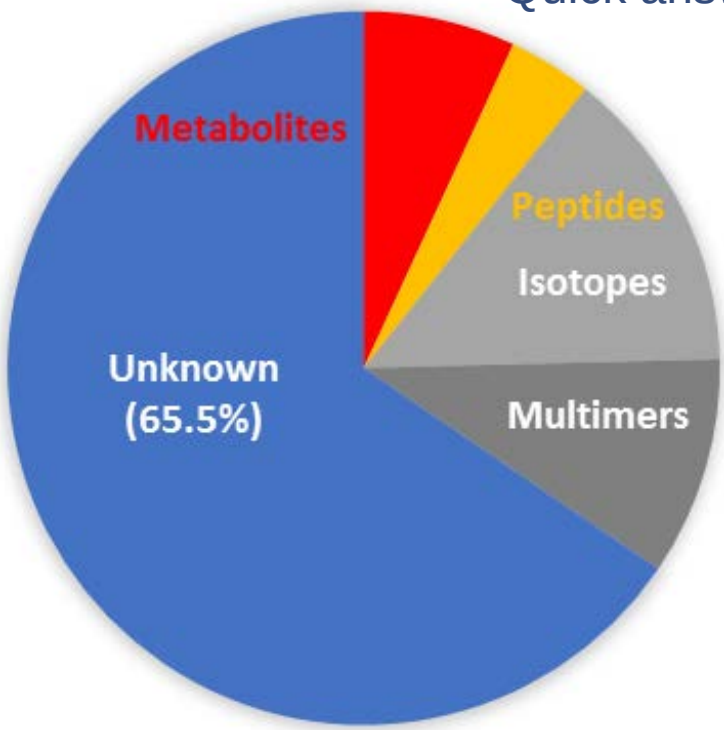
Peisl, Schymanski & Wilmes, 2018 *Anal. Chim. Acta*, DOI: 10.1016/j.aca.2017.12.034

Searching for Small Molecules …

- Compound Databases … isn't 95 million enough?



Quick answer … NO!

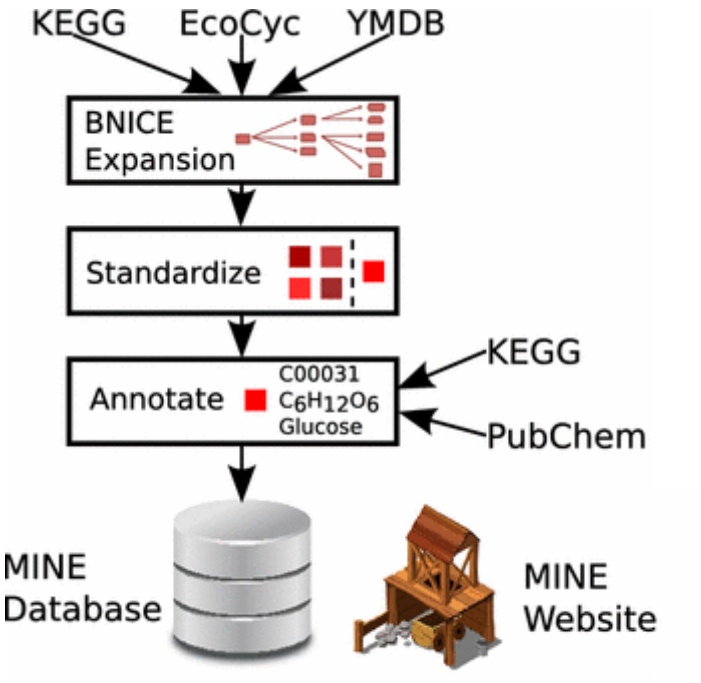


Compounds/Structures ▲ Spectra

Peisl, Schymanski & Wilmes, 2018 *Anal. Chim. Acta*, DOI: 10.1016/j.aca.2017.12.034 *E. coli* data :N. Zamboni, IMSB, ETH Zürich

Searching for More Small Molecules …

- *In silico* metabolite prediction – example of MINE
  - First generation only … combinatorial explosion!



KEGG MINE
13,307 => 571,368

EcoCyc MINE
1,832 => 54,719

YMDB MINE
1,978 => 100,755

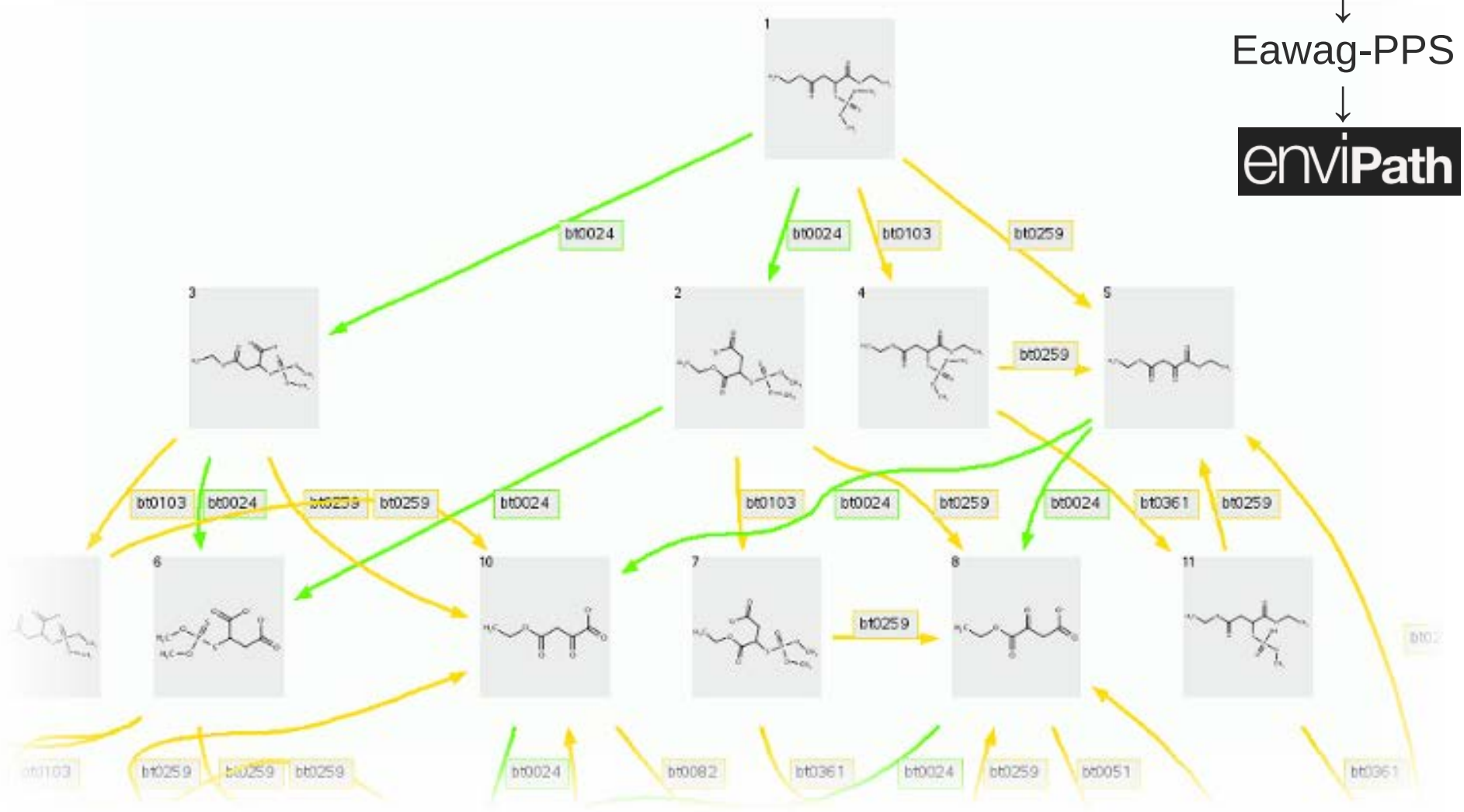
HMDB [15] MINE
23,035 => 400,414

Jeffryes *et al*, 2015, MINEs, *J. Cheminf*, 7:44. DOI: 10.1186/s13321-015-0087-1

Searching for More Small Molecules …

- *In silico* prediction and combinatorial explosion

[UM-PPS]
↓
Eawag-PPS
↓
enviPath



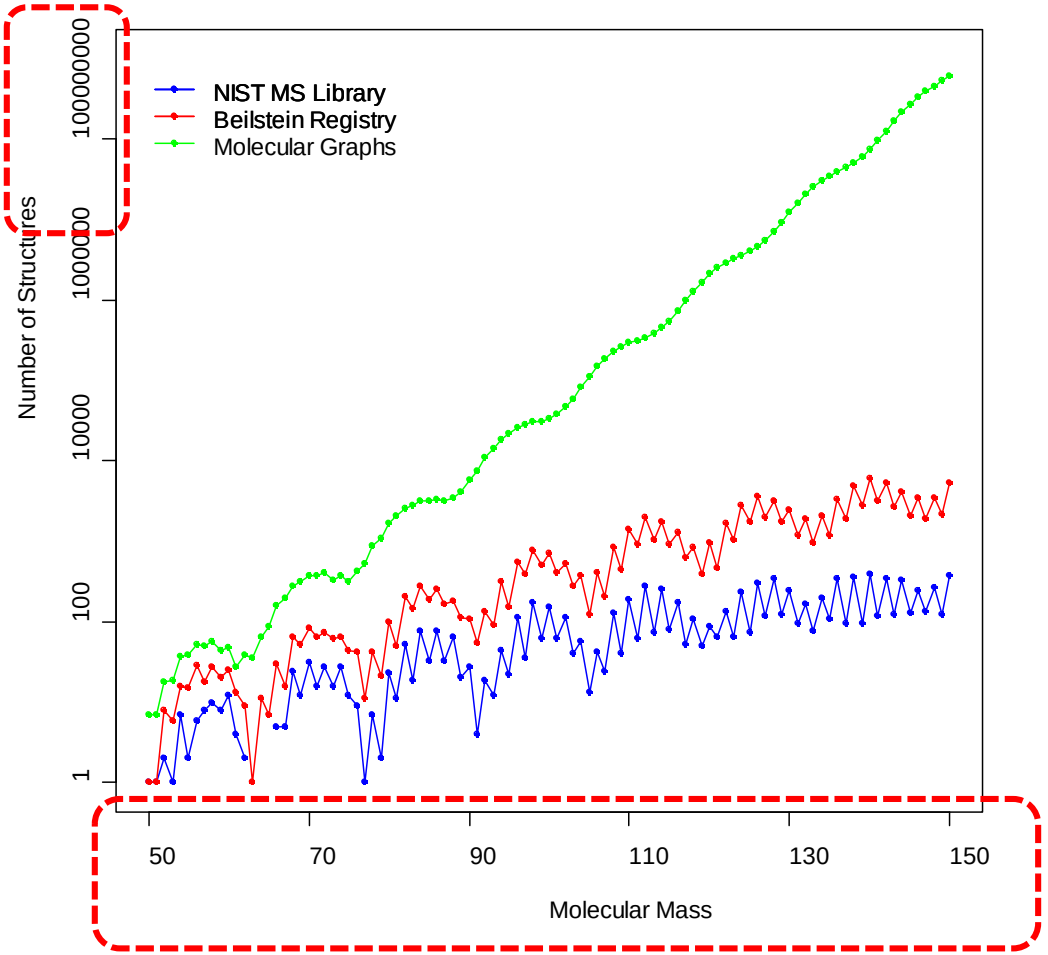
http://eawag-bbd.ethz.ch/predict/aboutPPS.html

Searching for EVEN MORE Small Molecules …

- Structure Generation
  - But of course most of these do not exist

Structures:

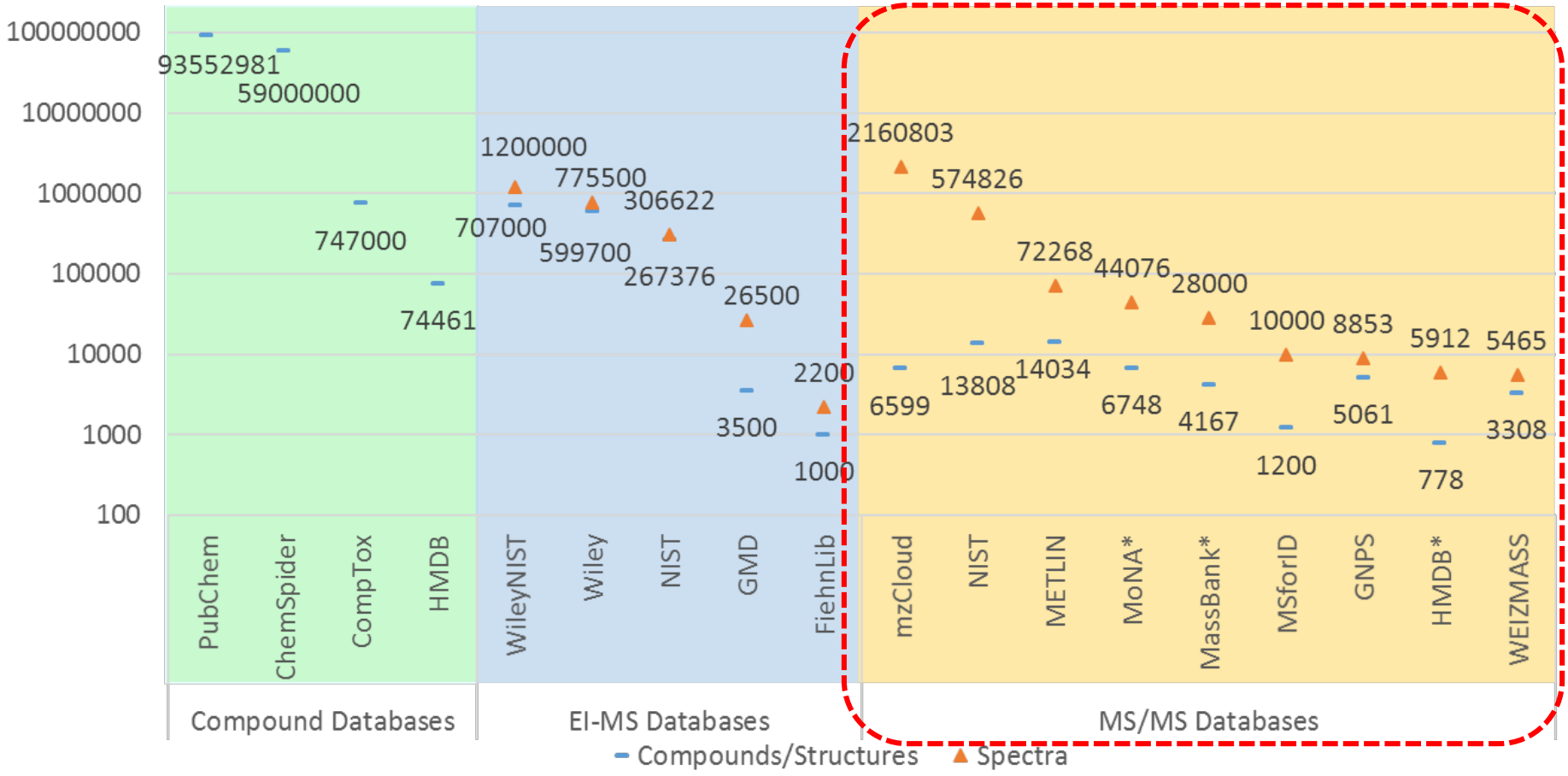
- elements C, H, N, O
- at least 1 C atom
- standard valencies
- no charges
- no radicals
- no stereoisomers
- only connected structures



Source: A. Kerber, R. Laue, M. Meringer, C. Rücker (2005) MATCH 54 (2), 301-312.

Searching for Small Molecules …

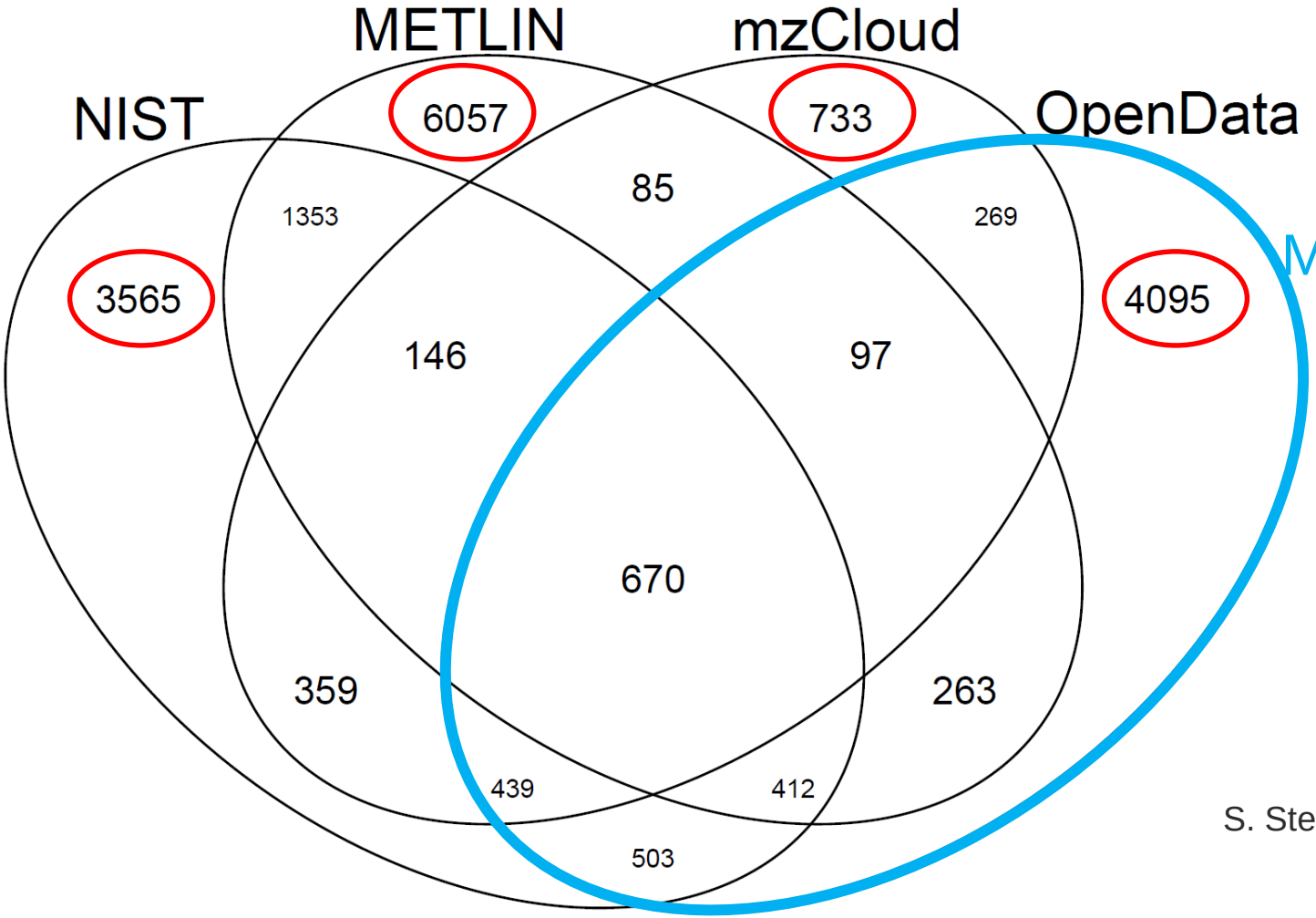
- And finding what we know …
  - Too many different MS/MS libraries (and they are still too small)



Peisl, Schymanski & Wilmes, 2018 *Anal. Chim. Acta*, DOI: 10.1016/j.aca.2017.12.034

Do we need all these libraries?

- Yes … most libraries still have many unique entries



Compound lists provided by:
S. Stein, R. Mistrik, Agilent

Vinaixa, Schymanski, Navarro, Neumann, Salek, Yanes, 2016, TrAC, DOI: 10.1016/j.trac.2015.09.005

SPLASH – Communicate between libraries

http://splash.fiehnlab.ucdavis.edu/

SPectraL hASH – an identifier for mass spectra

splash10 - 0002 - 0900000000 - b112e4e059e1ecf98c5f
[version] - [top10] - [histogram] - [hash of full spectrum]

http://mona.fiehnlab.ucdavis.edu/#/spectra/splash/splash10-0002-0900000000-b112e4e059e1ecf98c5f

https://www.google.ch/search?q=splash10-0002-0900000000-b112e4e059e1ecf98c5f

MassBank Record: EA278005

PK$SPLASH: splash10-0uxr-0973000000-87d07ddd2ed24b9598d7
PK$ANNOTATION: m/z tentative_formula formula_count mass error(ppm)
58.0651 C3H8N+ 1 58.0651 0.25
69.0335 C4H5O+ 1 69.0335 -0.45

Wohlgemuth *et al.*, 2016, Nature Biotechnology 34, 1099-1101, DOI: 10.1038/nbt.3689

SPLASH – Communicate between libraries

http://splash.fiehnlab.ucdavis.edu/

MassBank Record: EA278005

PK$SPLASH: splash10-0uxr-0973000000-87d07ddd2ed24b9598d7
PK$ANNOTATION: m/z tentative_formula formula_count mass error(ppm)
58.0651 C3H8N+ 1 58.0651 0.25
69.0335 C4H5O+ 1 69.0335 -0.45

GNPS Library Spectrum CCMSLIB00000210977

MASSBANK

splash10-0uxr-0973000000-87d07ddd2ed24b9598d7

DRUGBANK

splash10-0uxr-0973000000-87d07ddd2ed24b9598d7 View in MoNA

Identification	
Name	**Codeine**
Accession Number	**DB00318** (APRD00120, DB09471)

m/z CLOUD

splash10-0udi-0009000000-173ad69564777472205e
splash10-0udi-0009000000-a52c1f8d7e070d9019ba

Source: www.massbank.eu; http://gnps.ucsd.edu/, www.drugbank.ca; www.mzcloud.org

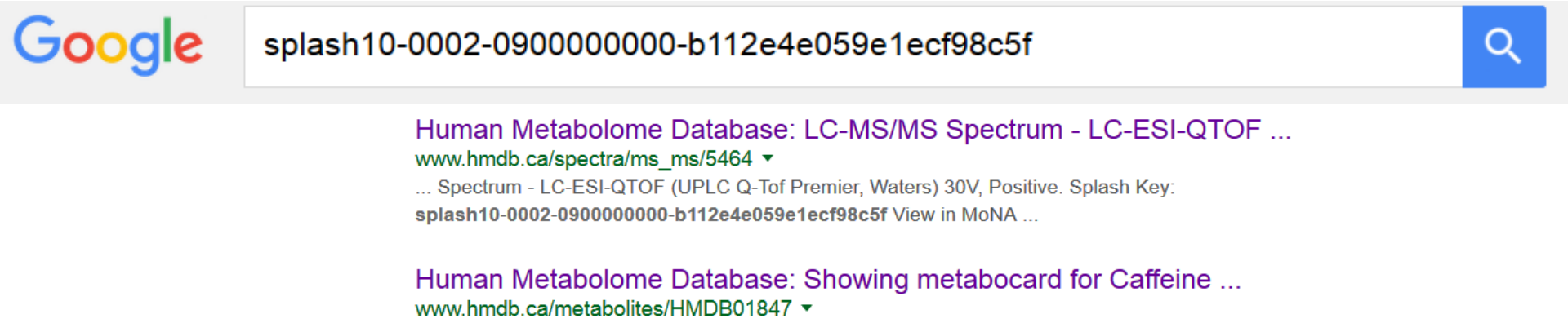
Wohlgemuth *et al.*, 2016, Nature Biotechnology 34, 1099-1101, DOI: 10.1038/nbt.3689

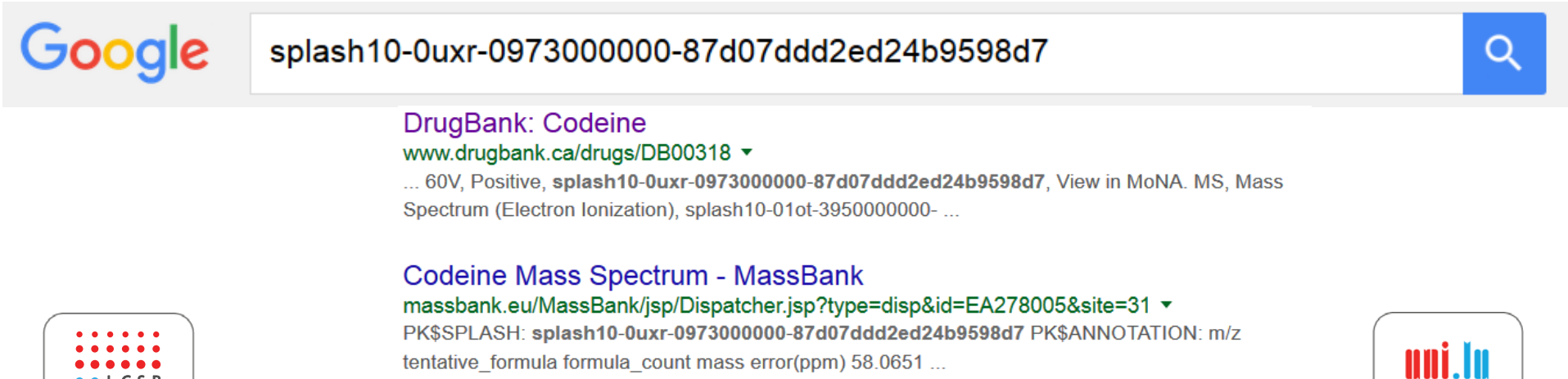
SPLASH – Communicate between libraries

splash10 - 0002 - 0900000000 - b112e4e059e1ecf98c5f
[version] - [top10] - [histogram] - [hash of full spectrum]

http://mona.fiehnlab.ucdavis.edu/#/spectra/splash/splash10-0002-0900000000-b112e4e059e1ecf98c5f

https://www.google.ch/search?q=splash10-0002-0900000000-b112e4e059e1ecf98c5f





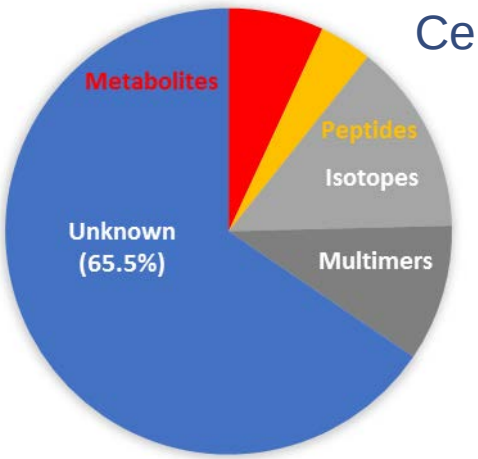
Wohlgemuth *et al.*, 2016, Nature Biotechnology 34, 1099-1101, DOI: 10.1038/nbt.3689

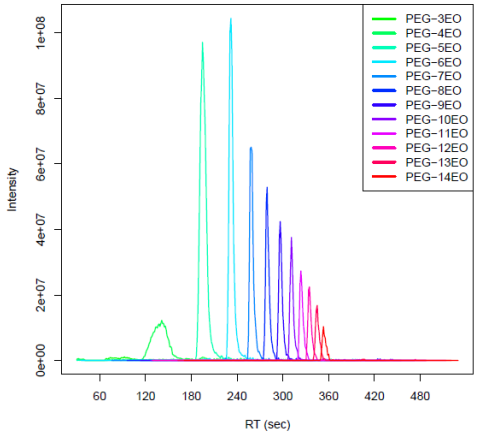
Small molecules … big problems?

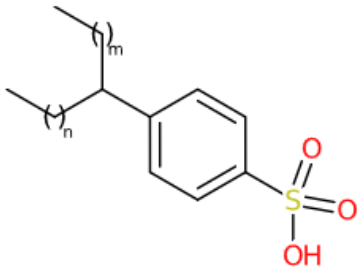
- Status quo of small molecules:
  - How many are in compound databases?
  - How many could there be?
  - How many are in spectral libraries?
  - Communicating between libraries

- "Non-target" Identification
  - Metadata
  - Exchanging Information

- Tackling the "unknowns"
  - Exchanging Information on Complex Structures
  - …and how Open Science helps!







LAS; n+m=7-10

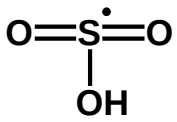
E. coli data provided by N. Zamboni, IMSB, ETH Zürich.

MetFrag2.3: Non-target Identification

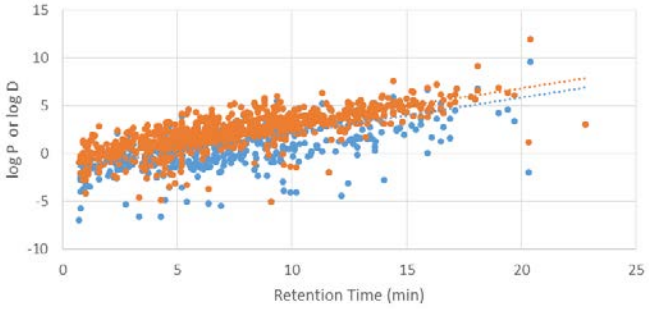
Status: 2010 => 2016

mz [M-H]⁻
213.9637
± 5 ppm

Elements: C,N,S
5 ppm
0.001 Da



RT: 4.54 min
355 InChI/RTs



ChemSpider
Search and share chemistry

or

PubChem OPEN CHEMISTRY DATABASE

References
External Refs
Data Sources
RSC Count
PubMed Count

Literature Patents

MetFrag2.3

Suspect Lists

TOFF IDENT

Chemistry Dashboard

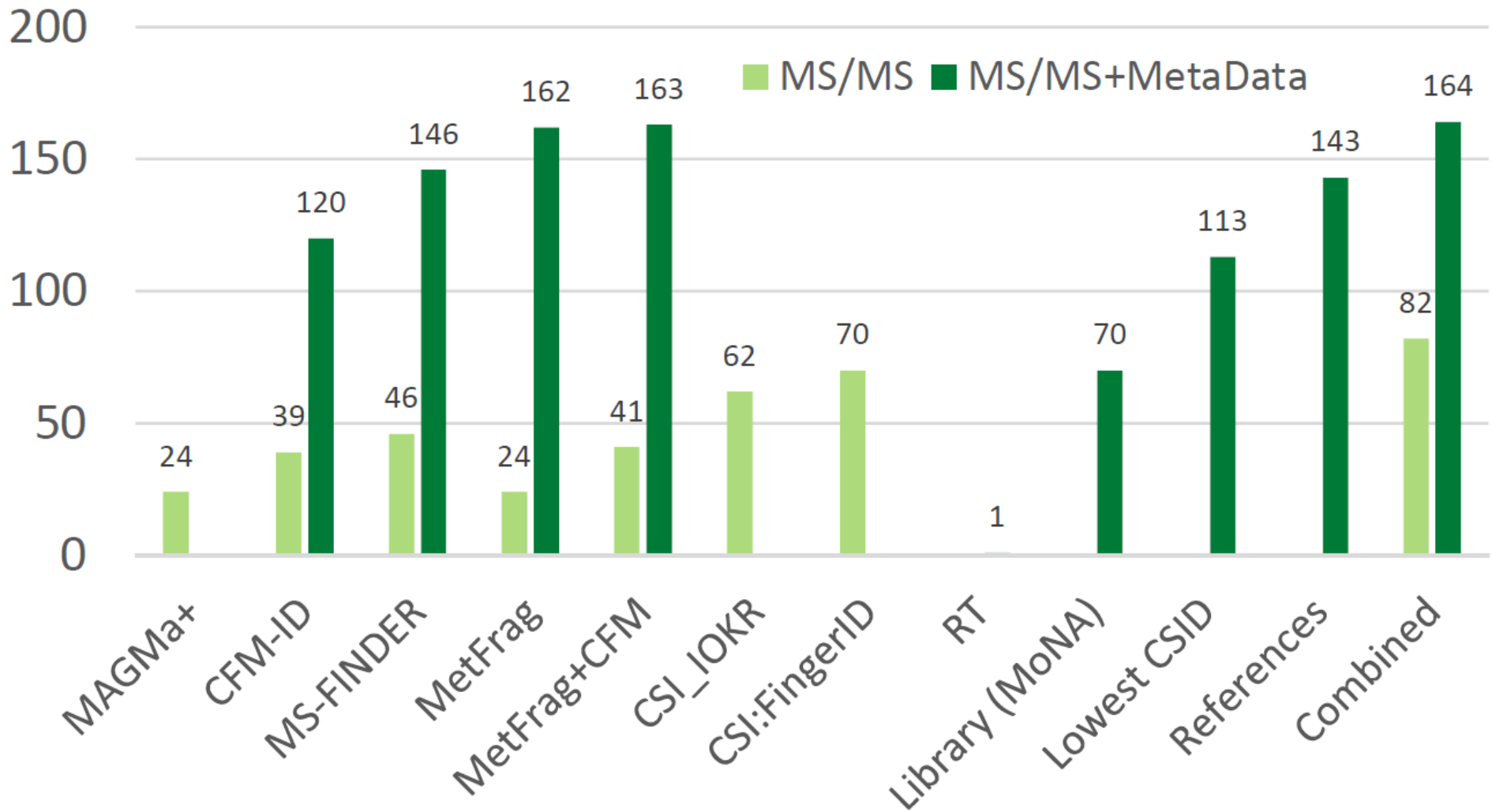
MoNA
MassBank of North America
MassBank.eu

MS/MS

134.0054	339689
150.0001	77271
213.9607	632466

Ruttkies, Schymanski, Wolf, Hollender, Neumann (2016) *J. Cheminf.*, 2016, DOI: 10.1186/s13321-016-0115-9

The Power of the Metadata (Top 1 ranks)



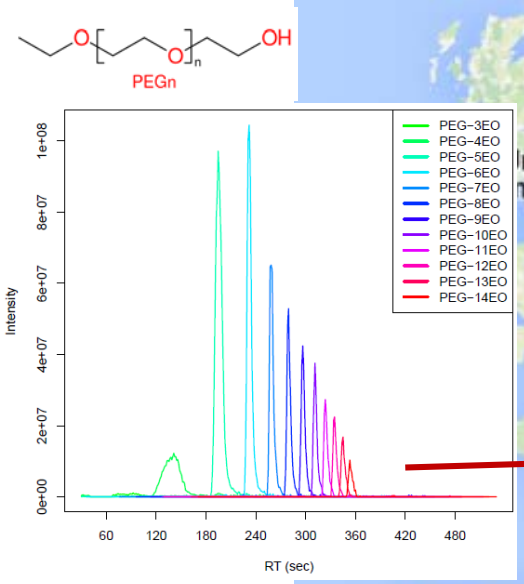
Schymanski *et al*, 2017, J Cheminf., DOI: 10.1186/s13321-017-0207-1

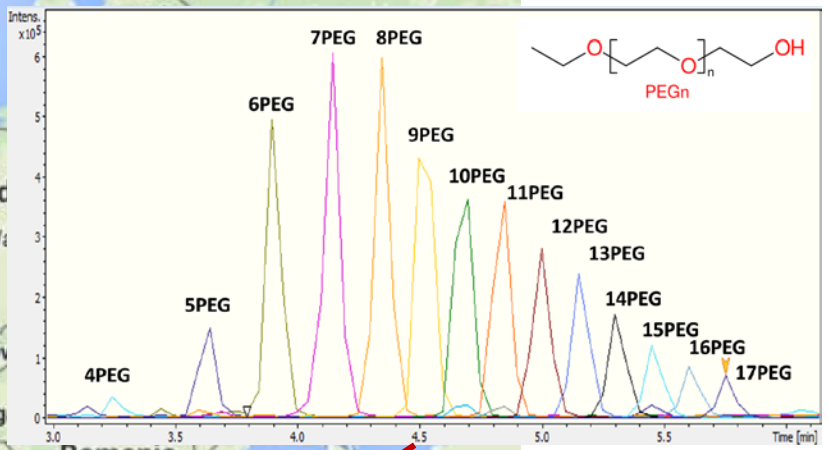
2015: European Non-target Screening Trial

Schymanski et al. 2015, ABC, DOI: 10.1007/s00216-015-8681-7

European (World-)Wide Exchange of Suspects

NORMAN Suspect List Exchange:
http://www.norman-network.com/?q=node/236





Schymanski et al. 2015, *ABC*, DOI: 10.1007/s00216-015-8681-7; Alygizakis *et al*. 2018 *ES&T*, DOI: 10.1021/acs.est.8b00365

NORMAN Lists => CompTox Dashboard

NORMAN Suspect List Exchange

http://www.norman-network.com/?q=node/236

NormaNEWS for retrospective screening of new emerging contaminants	NormaNEWS **CSV**, **XLSX** (3/10/2017) CompTox **NORMANEWS List**	**NormaNEWS InChIKeys** (8/05/2017)	**NormaNEWS** list provided by Nikiforos Alygizakis, Saer Samanipour and Kevin Thomas

NormaNEWS: Norman Early Warning System

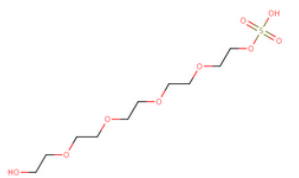
Search NORMANEWS Chemicals

List Details

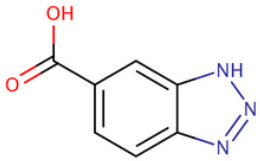
Description: The Norman Early Warning System (NormaNEWS) is a pilot network designed to investigate the spatial and temporal distribution of newly identified contaminants of emerging concern in environmental samples through performing retrospective suspect screening on HRMS data acquired using different instrumental platforms and data processing software. The NormaNEWS pilot study was performed through recruiting eight reference laboratories with available archived HRMS data with the goal of exploring the potential of an early warning network to rapidly establish the occurrence of newly-identified contaminants of emerging concern across Europe and beyond, through the use of retrospective suspect screening employing HRMS. The pilot study was referred to as the Norman Early Warning System, abbreviated to NormaNEWS.

Number of Chemicals: 131

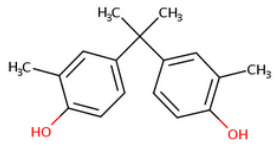
Download / Send | **Sort by:** Name | **4 of 131 chemicals selected** | **Hide:** Unselected × | Deselect all



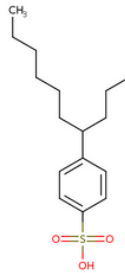
14-hydroxy-3,6,9,12-tetraoxatetradecyl...
NOCAS_881042



1H-Benzotriazole-5-carboxylic acid
23814-12-2



3,3'-Dimethylbisphenol A
79-97-0

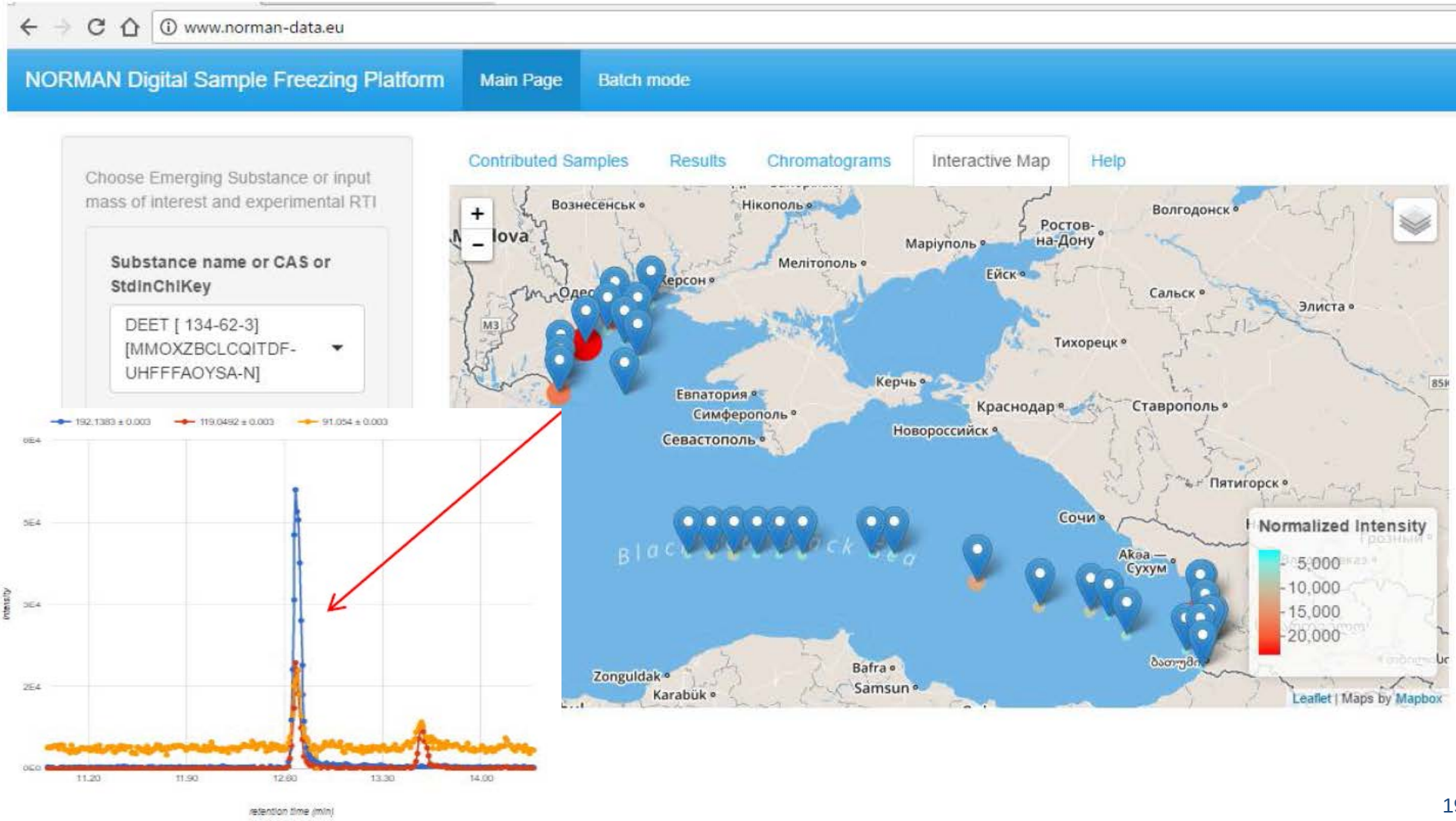


4-(Dodecan-6-yl)benzene-1-sulfonic acid
23003-92-1

https://comptox.epa.gov/dashboard/chemical_lists/normanews

NORMAN Digital Sample Freezing Platform

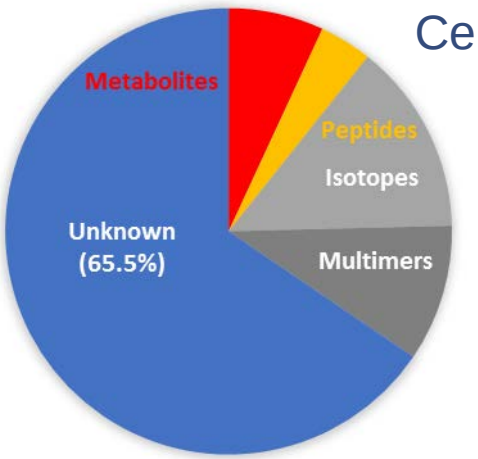
"Live" retrospective screening of known and unknown chemicals in European samples (various matrices)

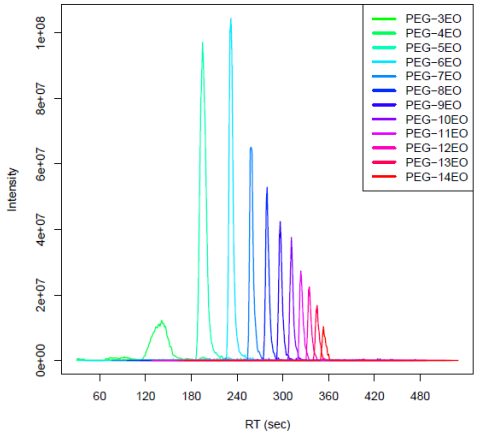


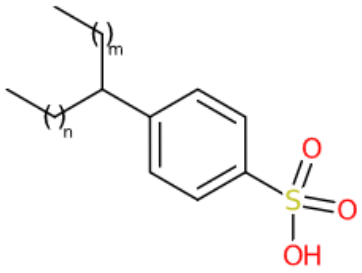
Aligizakis *et al*, in prep.

Small molecules … big problems?

- Status quo of small molecules:
  - How many are in compound databases?
  - How many could there be?
  - How many are in spectral libraries?
  - Communicating between libraries
- “Non-target” Identification
  - Metadata
  - Exchanging Information
- Tackling the “unknowns”
  - Exchanging Information on Complex Structures
  - …and how Open Science helps!





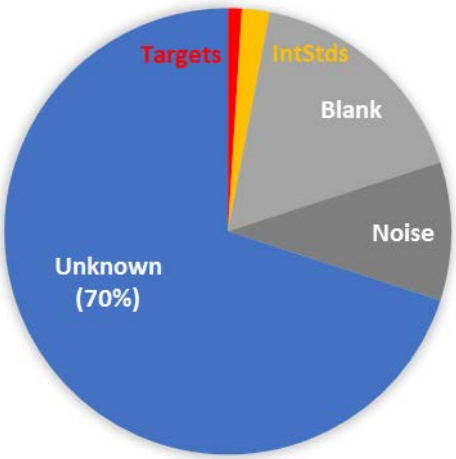


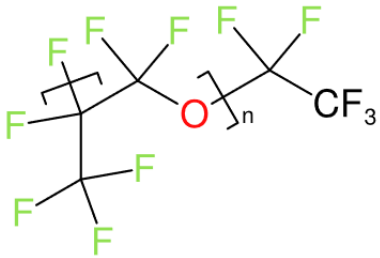
LAS; n+m=7-10

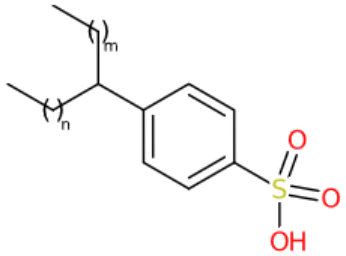
E. coli data provided by N. Zamboni, IMSB, ETH Zürich.

We still have many unknowns …

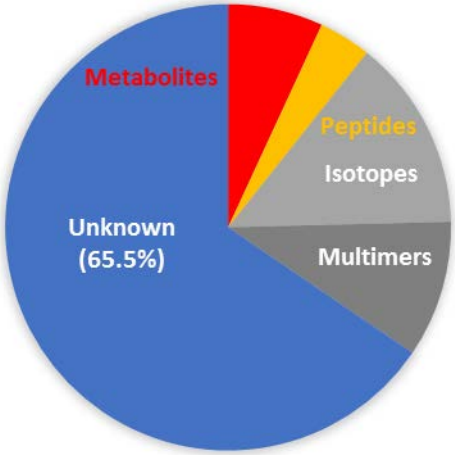
Environment







LAS; n+m=7-10



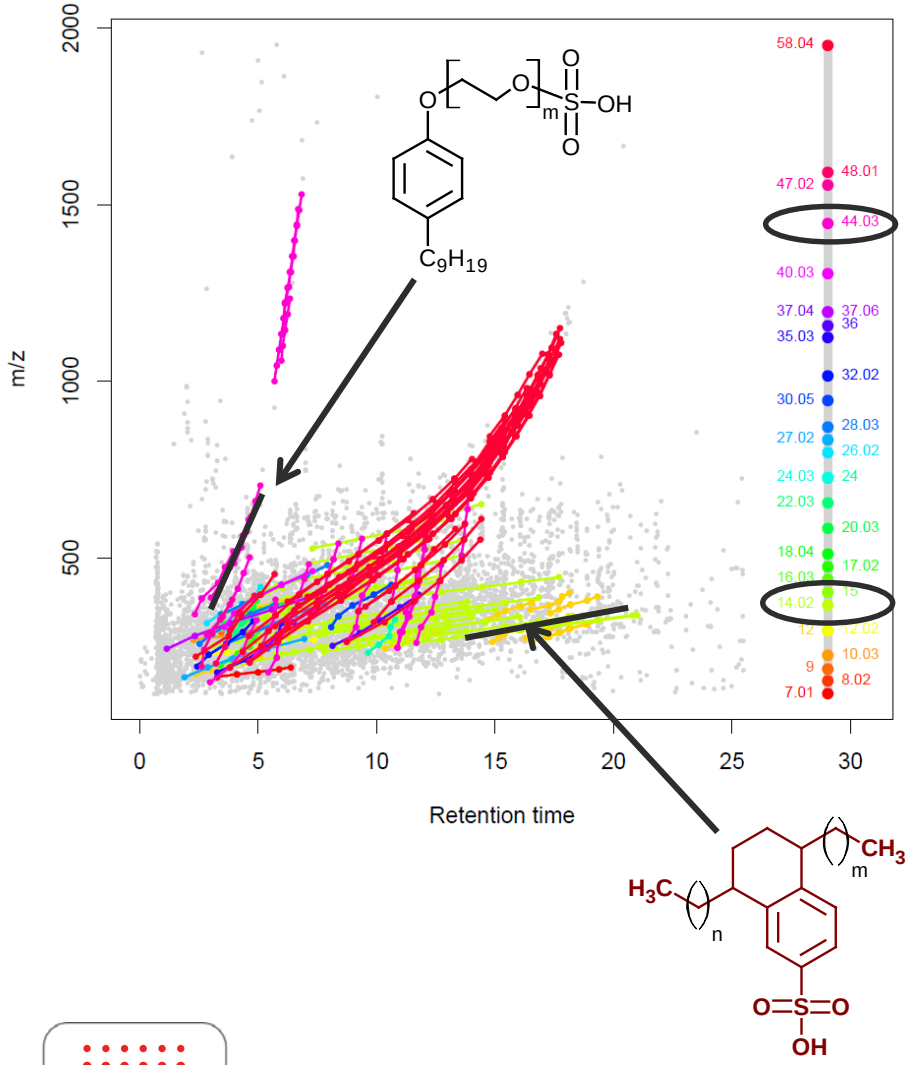
Cells

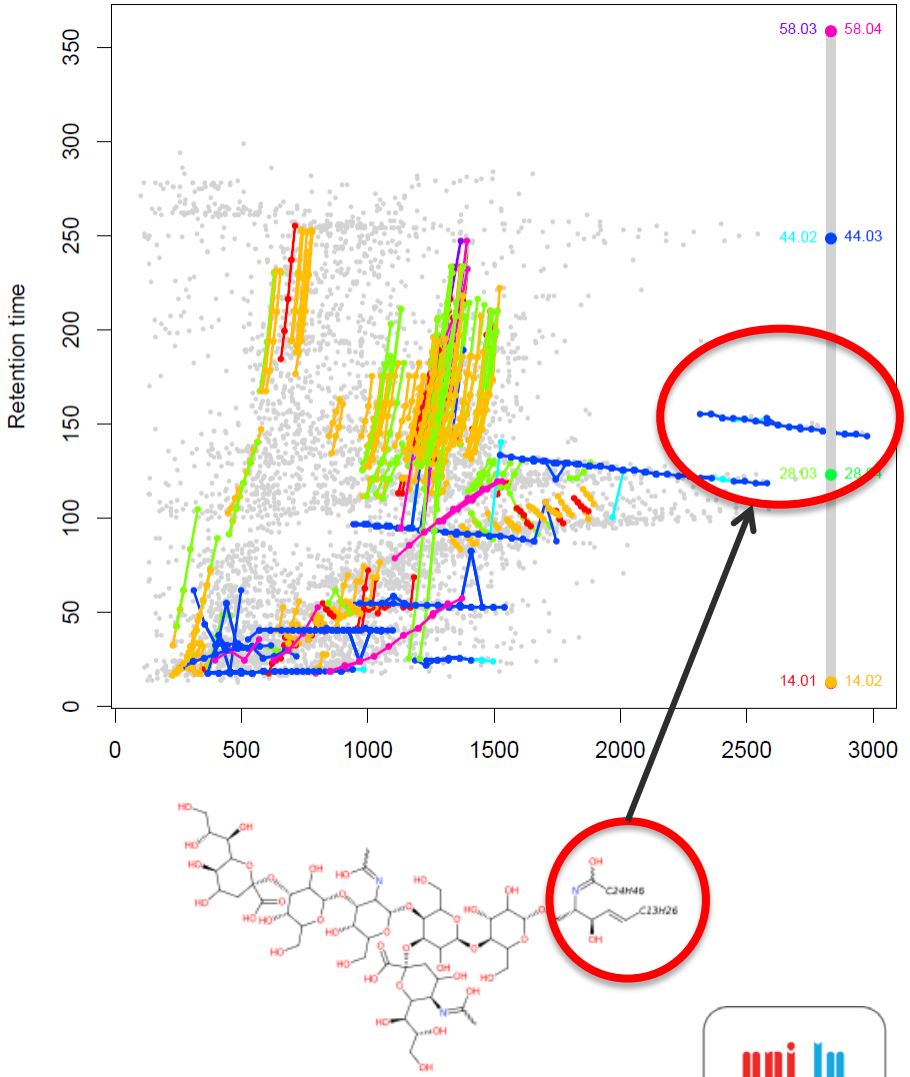
(l) Data from Schymanski *et al* 2014, ES&T DOI: 10.1021/es4044374. (r) *E. coli* data provided by N. Zamboni, IMSB, ETH Zürich.

…and many interconnected by mass

eawag aquatic research
nontarget

Homologous Series in environmental and biological samples ….





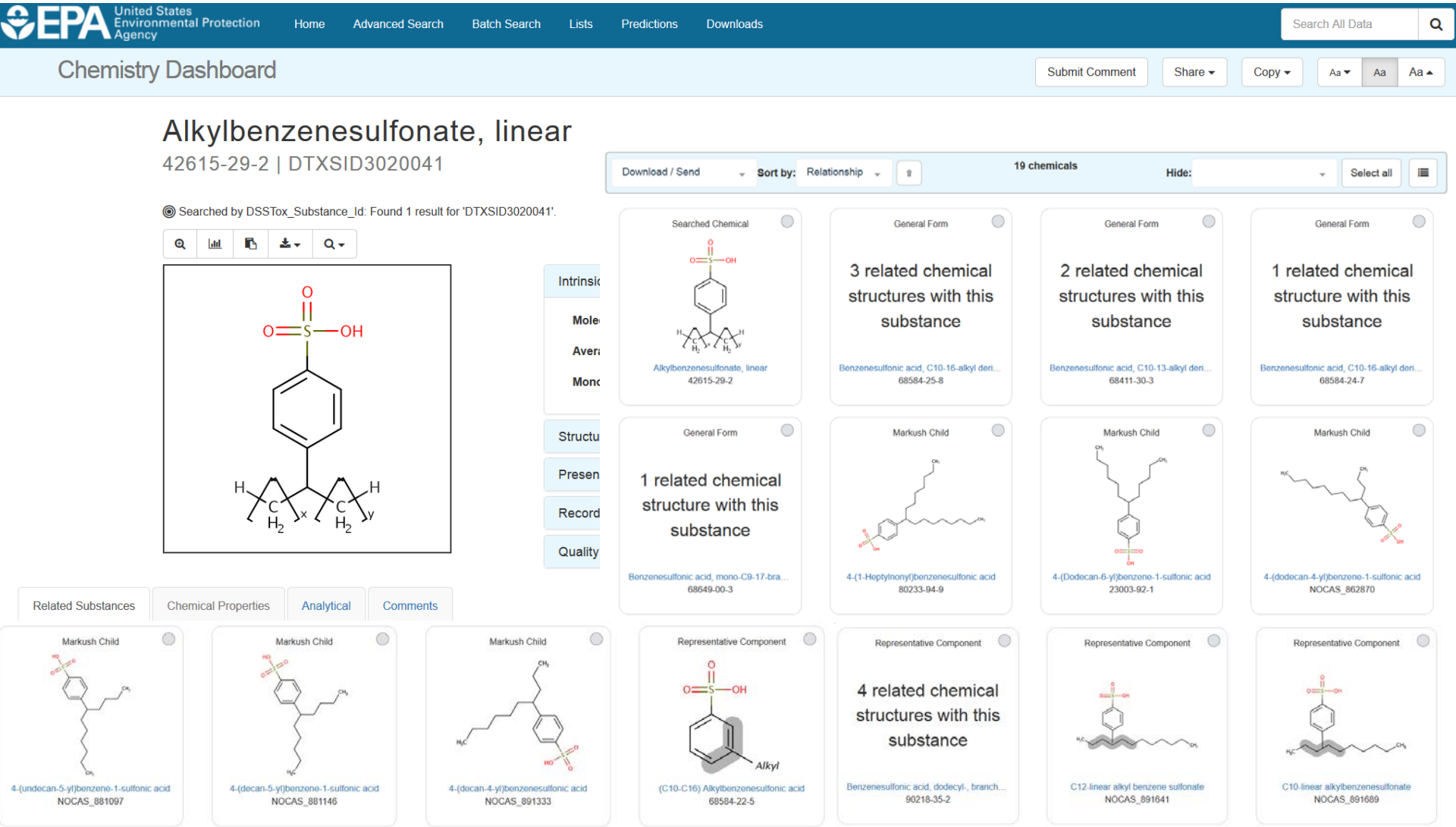
Lipid extract data of *Mycobacterium smegmatis* provided by N. Zamboni, IMSB, ETHZ

LCSB

uni.lu UNIVERSITÉ DU LUXEMBOURG

Schymanski et al. 2014, ES&T, DOI: 10.1021/es4044374; M. Loos & H Singer, 2017. J. Cheminf. DOI: 10.1186/s13321-017-0197-z

How to Store and Access these Structures?

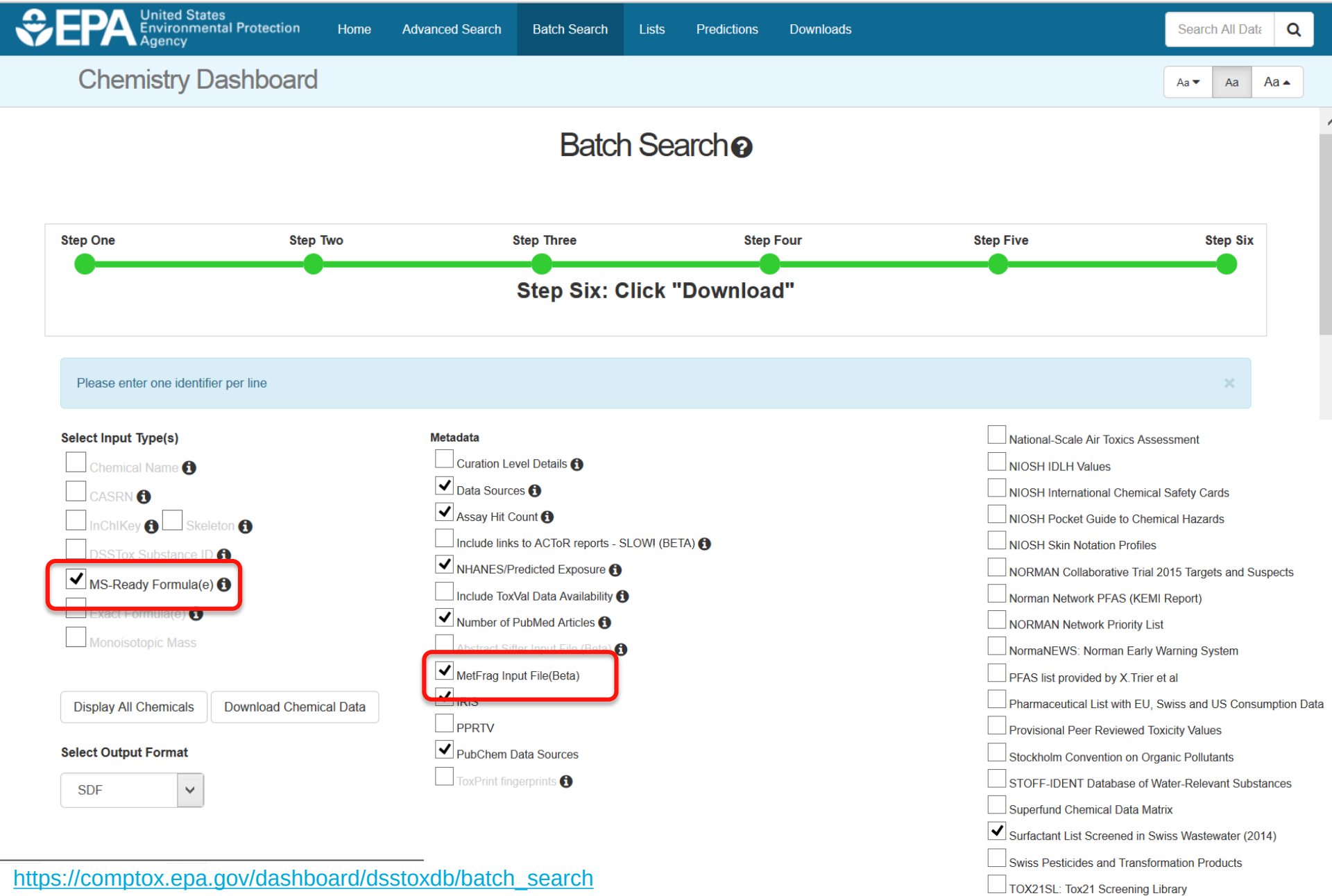


https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID3020041

https://comptox.epa.gov/dashboard/chemical_lists/eawagsurf

Schymanski, Grulke, Williams et al, in prep. & Williams *et al.* 2017 *J. Cheminformatics* 9:61 DOI: 10.1186/s13321-017-0247-6

MS-ready, Dashboard & MetFrag



https://comptox.epa.gov/dashboard/dsstoxdb/batch_search

MS-ready, Dashboard & MetFrag

MetFrag

https://msbi.ipb-halle.de/MetFragBeta/

In silico fragmentation for computer assisted identification of metabolite mass spectra

Database Settings

Database: SDF

Parent Ion: [M+H]+ Calculate

Molecules with aromatic bond order type (4) are discarded as not yet supported by CDK.

Upload File: + Choose

Uploaded ChemistryDashboard-Batch-Search_2018-03-...

Neutral Mass: 298.16038 Search ppm:

Formula: C16H26O3S

Identifiers:

Retrieve Candidates 18 Candidates

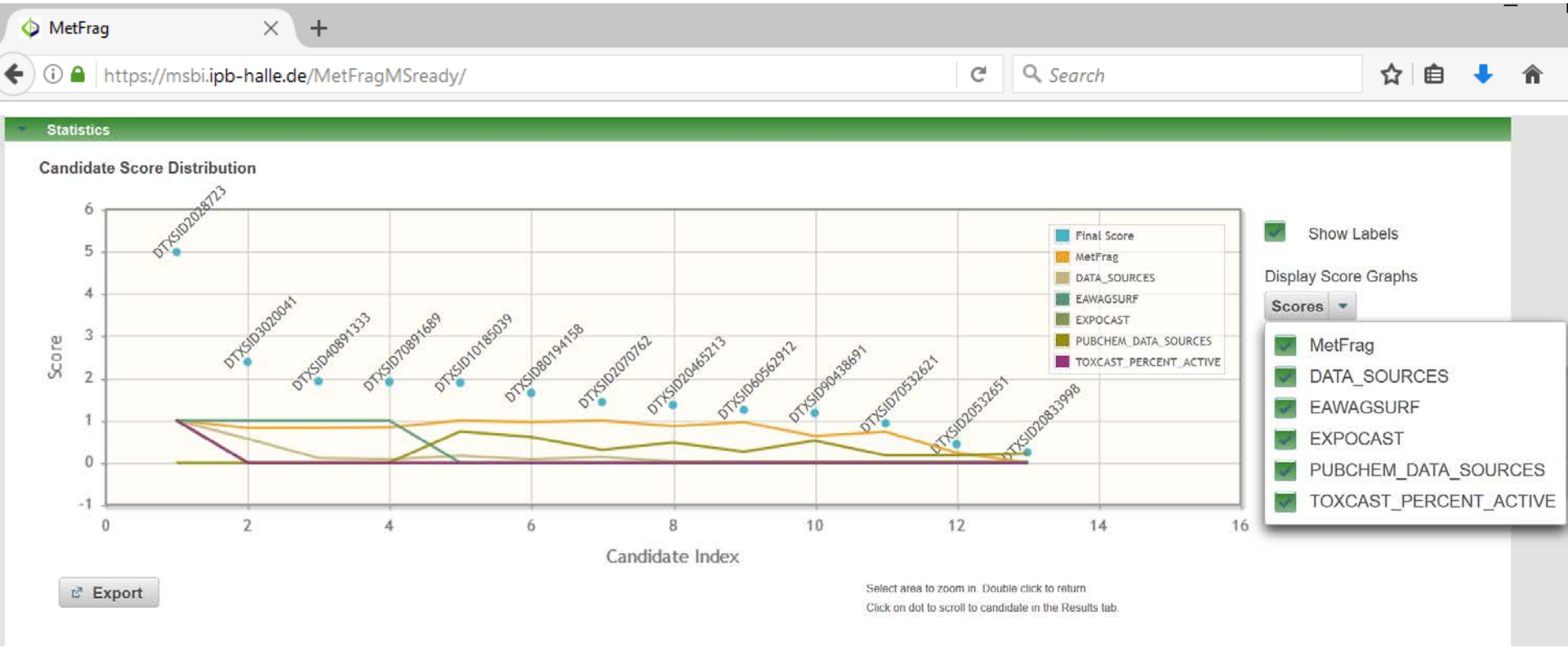
Database Scoring Terms:

- [x] DATA_SOURCES
- [x] EAWAGSURF
- [x] EXPOCAST
- [ ] EXPOCAST_MEDIAN_EXPOSURE_PREDICTION_MG/KG-BW/DAY
- [ ] IRIS
- [ ] NHANES
- [ ] NUMBER_OF_PUBMED_ARTICLES
- [x] PUBCHEM_DATA_SOURCES
- [x] TOXCAST_PERCENT_ACTIVE

https://massbank.eu/MassBank/jsp/RecordDisplay.jsp?id=ETS00012&dsn=Eawag_Addn

MS-ready, Dashboard & MetFrag

- Plot up candidates and contributing metadata



https://msbi.ipb-halle.de/MetFragBeta/

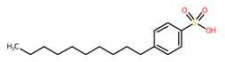
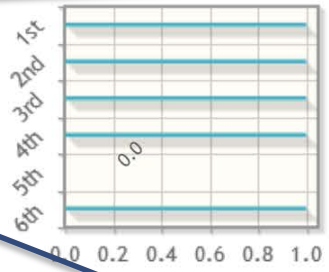
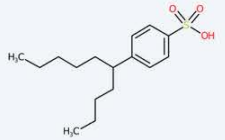
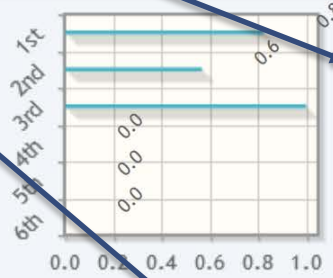
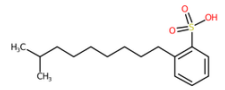
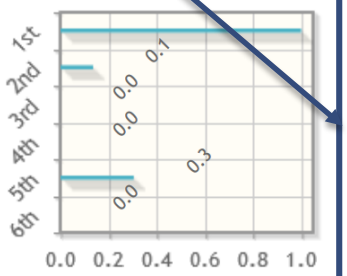
MS-ready, Dashboard & MetFrag

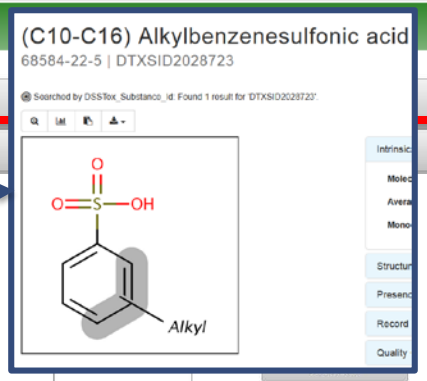
MetFrag

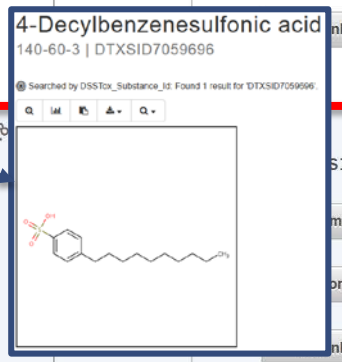
https://msbi.ipb-halle.de/MetFragMSready/

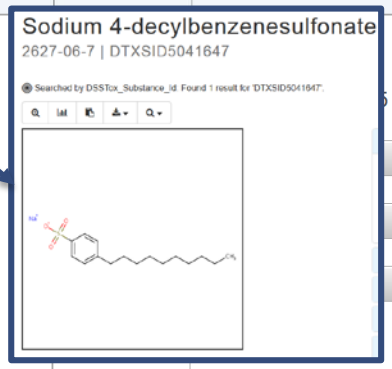
Search

Results

#	Molecule	Identifier	Mass	Formula	Normalized Scores
1	 (C10-C16) Alkylbenzenesulfonic acid	DTXSID2028723 DTXSID7059696 DTXSID5041647 InChIKeyBlock1 = UASQKKHYUPBQJR	298.16027	$C_{16}H_{26}O_3S$	
2	 Alkylbenzenesulfonate, linear	DTXSID3020041 DTXSID70881146 InChIKeyBlock1 = KIIODTIGNUGDLO	298.16027	$C_{16}H_{26}O_3S$	
7	 Benzenesulfonic acid, isodecyl-, compd. with 2,2'-iminobis[ethanol] (1:1)	DTXSID2070762 InChIKeyBlock1 = CKNPXKLNOLAXDF	298.16027	$C_{16}H_{26}O_3S$	

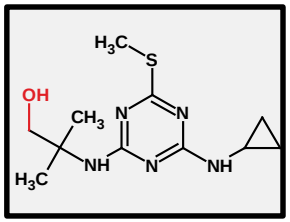
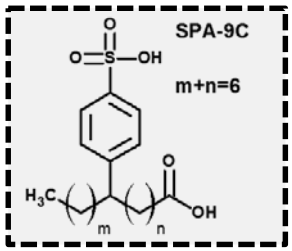
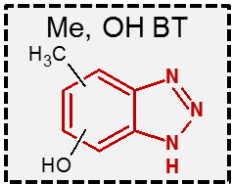






Confidence Levels for Tentative Structures

- Annotation is the key to communicating information

Example	*Identification confidence*	*Minimum data requirements*
	***Level 1: Confirmed structure*** by reference standard	MS, MS^2, RT, Reference Std.
	Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MS^2, Library MS^2 MS, MS^2, Exp. data
	***Level 3: Tentative candidate(s)*** structure, substituent, class	MS, MS^2, Exp. data
		
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct
192.0757	***Level 5: Exact mass*** of interest	MS

Schymanski, Jeon, Gulde, Fenner, Ruff, Singer & Hollender (2014) ES&T, 48 (4), 2097-2098. DOI: 10.1021/es5002105

Automating Confidence Levels

LevelCode	LevelKeyword	COMMENT.CONFIDENCE
1	standard	Reference Standard (Level 1)
1a	standard	Reference Standard (Level 1)
1b	parent	Parent Substance with Reference Standard (Level 1)
1c	confirmed	Identification confirmed with Reference Standard (Level 1)
2	probable	Probable structure, tentative identification (Level 2)
2a	probableLibrary	Probable structure via library match, tentative identification (Level 2a)
2b	probableDiagnostic	Probable structure via diagnostic evidence, tentative identification (Level 2b)
3	tentative	Tentative identification only (Level 3)
3a	tentativeStructure	Tentative identification: most likely structure (Level 3)
3b	tentativeIsomer	Tentative identification: isomers possible (Level 3)
3c	tentativeTPClass	Tentative identification: substance class known (Level 3)
3d	tentativeBestMatch	Tentative identification: best match only (Level 3)
4	formula	Tentative identification: molecular formula only (Level 4)
5	unknown	Tentative identification: structure and formula unknown (Level 5)
5	exactMass	Tentative identification: structure and formula unknown (Level 5)

Schymanski *et al*, 2014, ES&T. DOI: 10.1021/es5002105 & Schymanski *et al*. 2015, ABC, DOI: 10.1007/s00216-015-8681-7

RMassBank: More than “just” MassBank

RMassBank

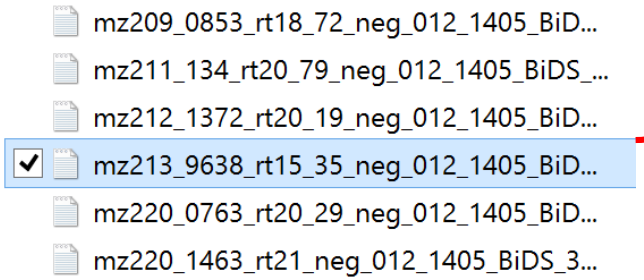
LevelCode	LevelKeyword	COMMENT.CONFIDENCE
1	standard	Reference Standard (Level 1)
1a	standard	Reference Standard (Level 1)
1b	parent	Parent Substance with Reference Standard (Level 1)
1c	confirmed	Identification confirmed with Reference Standard (Level 1)
2	probable	Probable structure, tentative identification (Level 2)
2a	probableLibrary	Probable structure via library match, tentative identification (Level 2a)
2b	probableDiagnostic	Probable structure via diagnostic evidence, tentative identification (Level 2b)
3	tentative	Tentative identification only (Level 3)
3a	tentativeStructure	Tentative identification: most likely structure (Level 3)
3b	tentativeIsomer	Tentative identification: isomers possible (Level 3)
3c	tentativeTPClass	Tentative identification: substance class known (Level 3)
3d	tentativeBestMatch	Tentative identification: best match only (Level 3)
4	formula	Tentative identification: molecular formula only (Level 4)
5	unknown	Tentative identification: structure and formula unknown (Level 5)
5	exactMass	Tentative identification: structure and formula unknown (Level 5)

LC-MS/MS raw data

online resources: CTS, CACTUS

Automatic MS and MS/MS Recalibration and Clean-up Remove interfering peaks

Spectral Annotation with - Experimental Details - Compound Information



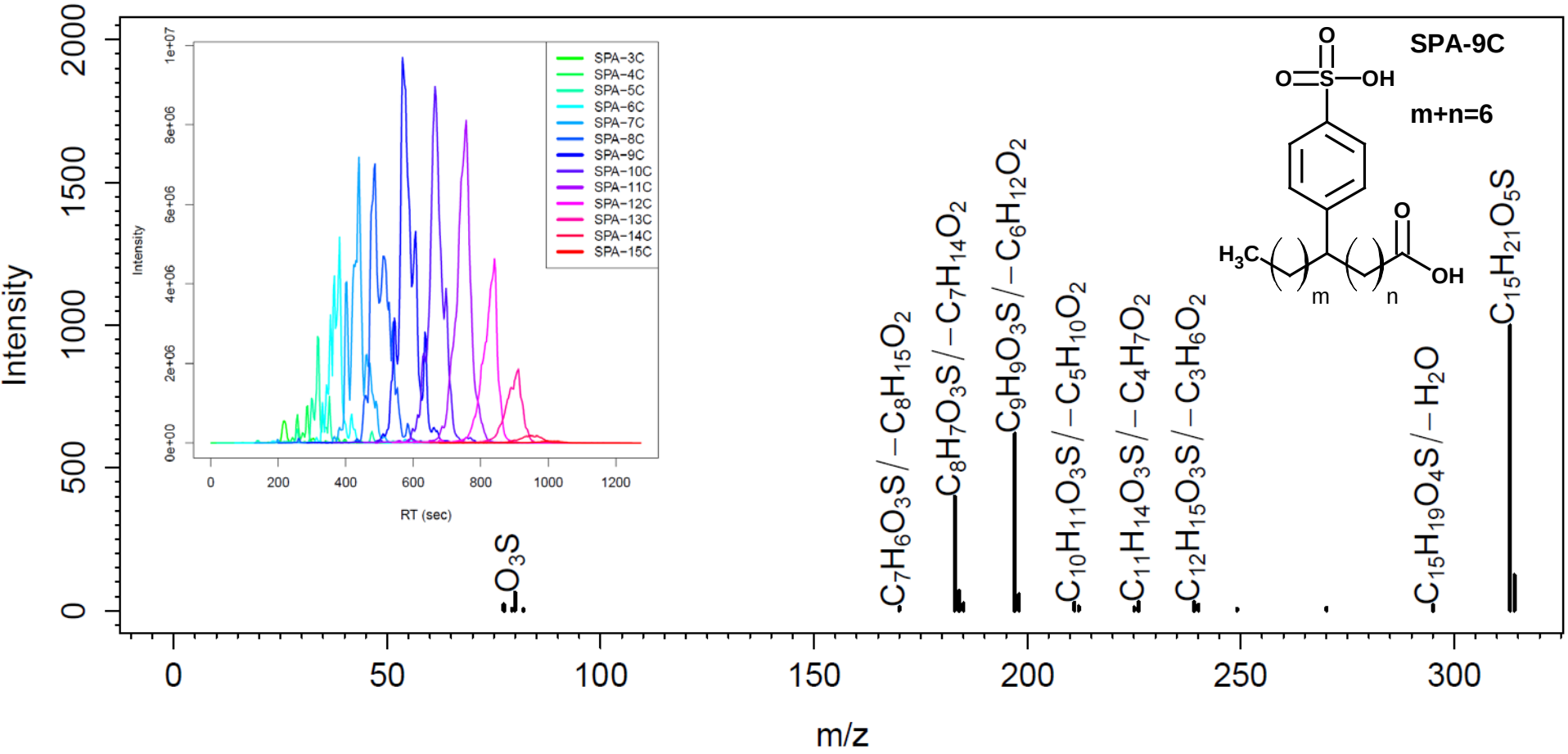
MS/MS for further processing

https://github.com/MassBank/RMassBank/
http://bioconductor.org/packages/RMassBank/

Supporting Evidence for Homologues

Chromatography and MS/MS Annotation

https://github.com/MassBank/RMassBank/



Formulas: http://sourceforge.net/projects/genform/
Meringer *et al,* 2011, *MATCH* 65, 259-290
Data: Schymanski *et al*. 2014, ES&T, 48: 1811-1818. DOI: 10.1021/es4044374

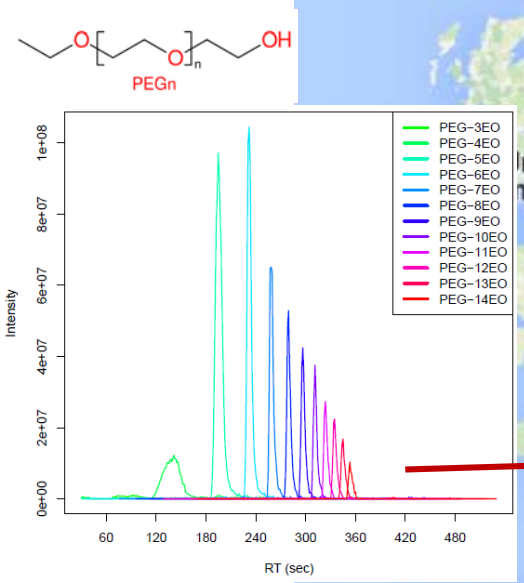
Literature: LIT00034,35
Sample: ETS00002
Standard: ETS00016,17,19,20

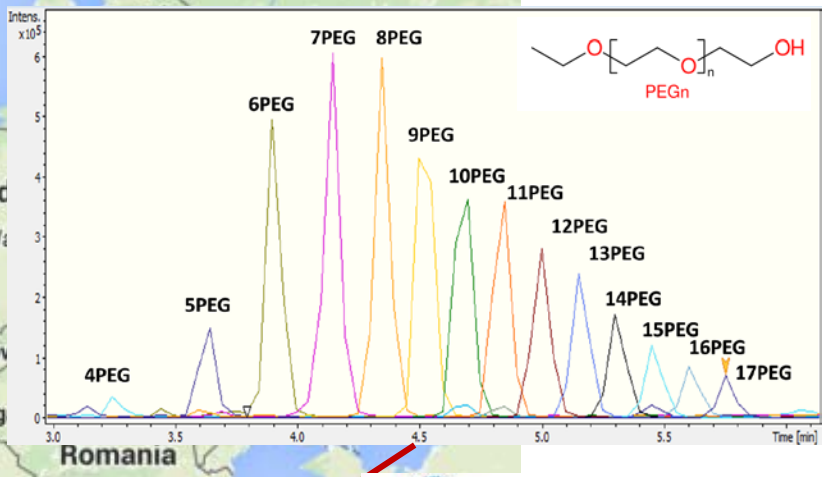
Stravs *et al*. *(*2013), *J. Mass Spectrom,* 48(1):89-99. DOI: 10.1002/jms.3131

European (World-)Wide Exchange of Suspects

Tentatively Identified Spectra:
http://goo.gl/0t7jGp
Hits in GNPS MassIVE datasets:
TPs in skin: http://goo.gl/NmO4tx
Surfactants: http://goo.gl/7sY9Pf

NORMAN Suspect List Exchange:
http://www.norman-network.com/?q=node/236

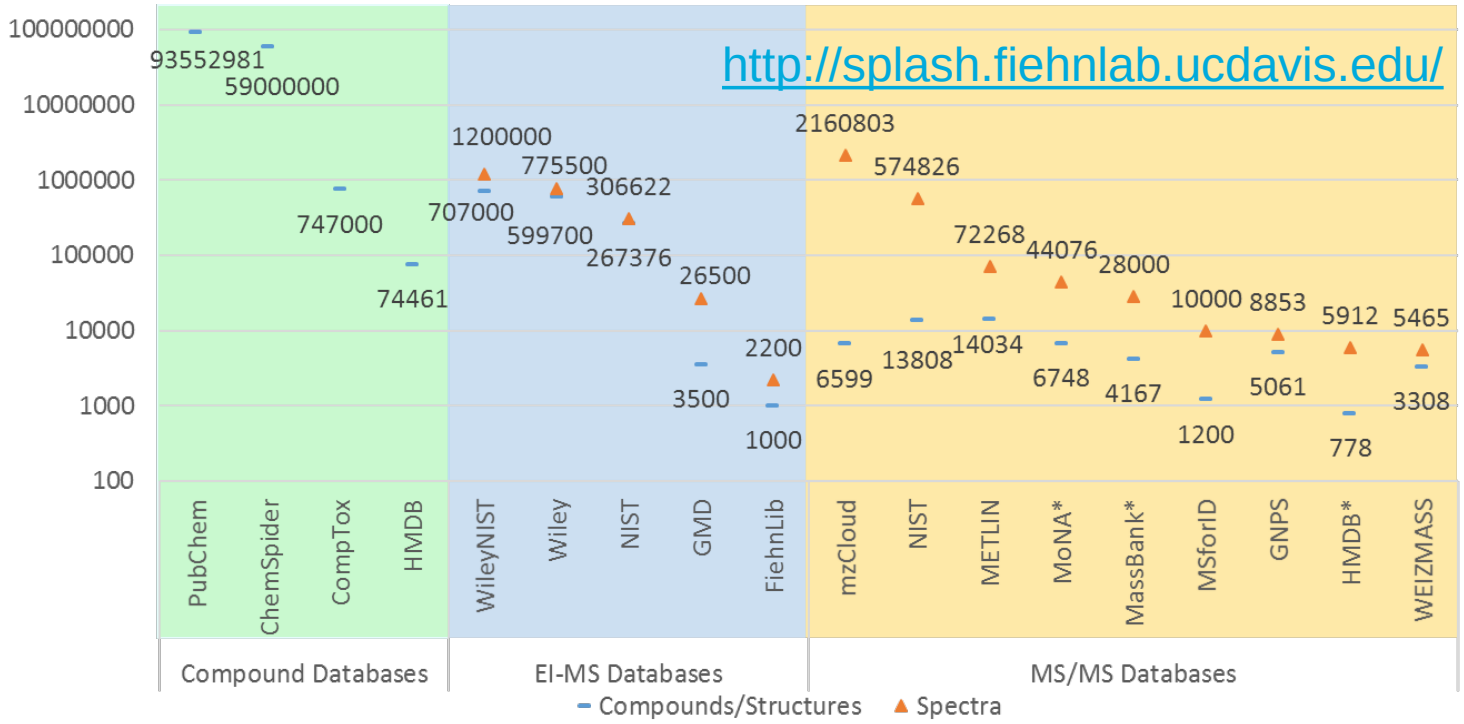




Schymanski et al. 2015, *ABC*, DOI: 10.1007/s00216-015-8681-7; Alygizakis *et al*. 2018 *ES&T*, DOI: 10.1021/acs.est.8b00365

Small molecules … big OPPORTUNITIES!

- Identifying small molecules requires information from many sources
  - Extensive compound databases available
  - Many mass spectral libraries available
  - Many excellent workflows available to collate this information
  - Community initiatives to improve communication between resources



Peisl, Schymanski & Wilmes, 2018 *Anal. Chim. Acta*, DOI: 10.1016/j.aca.2017.12.034

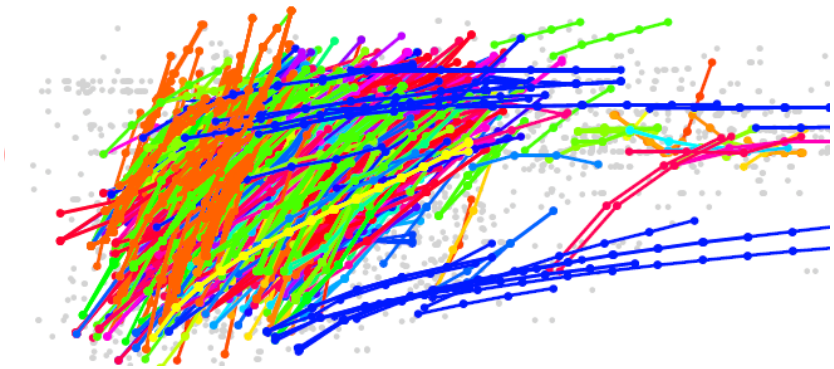
Small molecules … big OPPORTUNITIES!

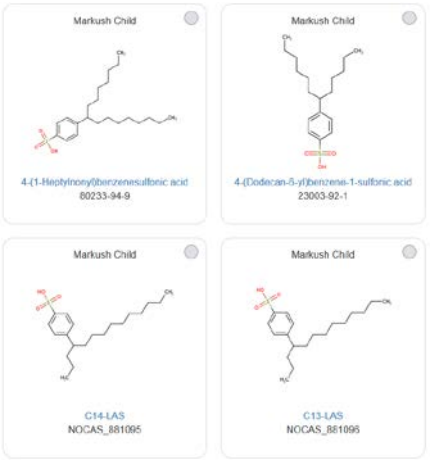
- Identifying small molecules requires information from many sources
  - Extensive compound databases available
  - Many mass spectral libraries available
  - Many excellent workflows available to collate this information
  - Community initiatives to improve communication between resources

- "Non-target" Identification
  - Metadata & Exchanging Information is critical to success

Schymanski et al. 2015, *ABC*, DOI: 10.1007/s00216-015-8681-7; Alygizakis *et al*. 2018 *ES&T*, DOI: 10.1021/acs.est.8b00365

Small molecules … big OPPORTUNITIES!

- Identifying small molecules requires information from many sources
  - Extensive compound databases available
  - Many mass spectral libraries available
  - Many excellent workflows available to collate this information
  - Community initiatives to improve communication between resources
- “Non-target” Identification
  - Metadata & Exchanging Information is critical to success
- Tackling the “unknowns”
  - Huge progress in cheminformatics approaches in very short time …
  - 

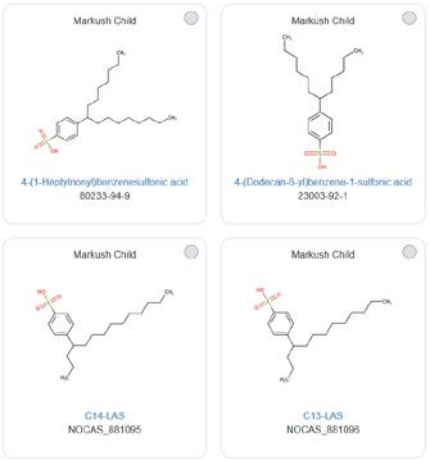


https://www.researchgate.net/project/Supporting-Mass-Spectrometry-Through-Cheminformatics

Small molecules … big OPPORTUNITIES!

- Identifying small molecules requires information from many sources
  - Extensive compound databases available
  - Many mass spectral libraries available
  - Many excellent workflows available to collate this information
  - Community initiatives to improve communication between resources
- “Non-target” Identification
  - Metadata & Exchanging Information is critical to success
- Tackling the “unknowns”
  - Huge progress in cheminformatics approaches in very short time …
  - Information in the public domain helps everyone!

    (you never know when it will help you!)

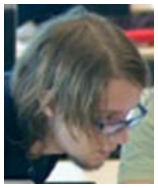


Acknowledgements I







RMassBank

nontarget

enviPat Web

enviPick

enviMass

MetFrag 2.3


MassBank.eu

Stellan Fischer KEMI


ETH zürich
Department of Biology
Institute of Molecular Systems Biology

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

norman

solutions

emma.schymanski@uni.lu

Further Information:

http://www.norman-network.com/?q=node/236

http://c-ruttkies.github.io/MetFrag/

https://comptox.epa.gov/dashboard/

https://www.researchgate.net/project/Supporting-Mass-Spectrometry-Through-Cheminformatics

https://github.com/MassBank/

EU Grant 603437

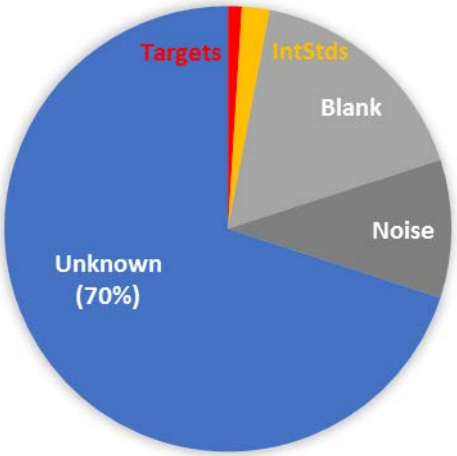
LCSB

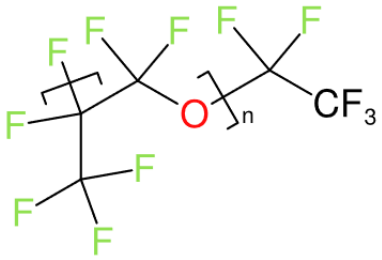
SEVENTH FRAMEWORK PROGRAMME

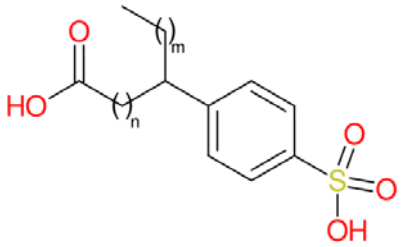
uni.lu UNIVERSITÉ DU LUXEMBOURG

We still have many unknowns …

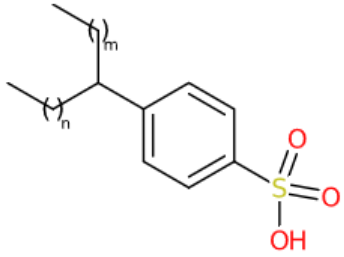
Environment



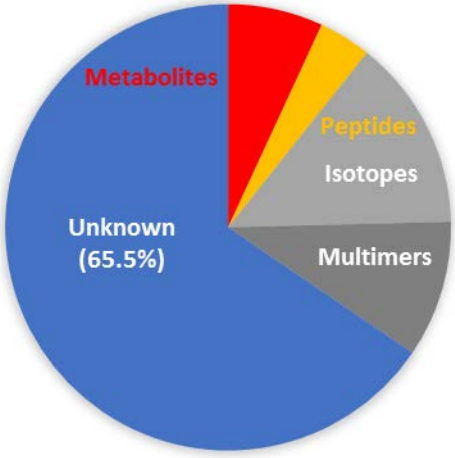




SPACs; n+m>=0



LAS; n+m=7-10



Cells

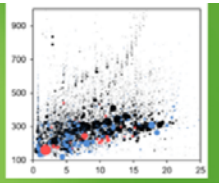
(l) Data from Schymanski *et al* 2014, ES&T DOI: 10.1021/es4044374. (r) *E. coli* data provided by N. Zamboni, IMSB, ETH Zürich.

Enter: NORMAN Suspect Exchange

- http://www.norman-network.com/?q=node/236

NORMAN

Network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances



Home | NORMAN Network | Working Groups | Membership | NORMAN Bulletin | Success Stories | Publications | Job opportunities | Contact | Gallery | NORMAN GA meetings

Menu

- Emerging Substances
- DATABASES
- Topics and Activities
- Workshops and Events
- QA/QC Issues
- Glossary

User login

Username *

Password *

Request new password

Log in

NORMAN Suspect List Exchange

In September 2014, NORMAN members expressed the need to exchange various lists of substances to improve their suspect screening efforts. This website was established aspart of the 2015 Joint Programme of Activities as a central access point for NORMAN members (and others) to find suspect lists relevant for their environmental monitoring question. All suspect lists currently available are compiled in the table below and on the US EPA CompTox Chemistry Dashboard (**website**, **downloads**, **chemical lists**). The "Link to full list" column below contains an excel or comma-separated file (csv) with all available information, e.g. as provided as supporting information for the publication, while the third column provides a list of the structures as InChIKeys only, which allows suspect searching using MetFrag or other workflows. The fourth column contains references for the data: please cite these references if you use the respective datasets. Recent Suspect Exchange and Dashboard presentations/publications include: **ICCE Oslo 2017**: **NORMAN Suspects meet the Dashboard** and **NORMAN MassBank and Suspect Exchange**; SETAC **Mixtures** Denver: **Identifying Complex Mixtures with Cheminformatics and HR-MS**; ACS Fall 2017: **Markush Enumeration for UVCBs** and a **viewpoint article**.

Coordination: Emma Schymanski, LCSB; Curation/RTI/toxicity: Reza Aalizadeh & Nikos Thomaidis, Uni. Athens; CompTox: Antony Williams, US EPA; Webmaster: Natalia Glowacka, Environmental Institute; IT: Lubos Cirka, Environmental Institute; Contributors: see below.

If you have any feedback or a list that you would like included, please contact **suspects@normandata.eu**.

Interactive merged list of all suspect substances

NEW (March 2018): SusDat is now at >40,000 chemicals; new US EPA Consumer Product Suspect List, Uni Athens Target List.

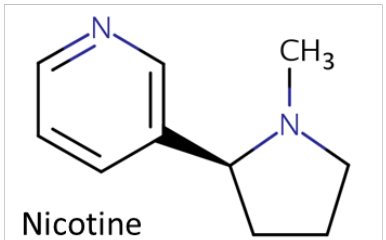
Full Lists | **InChIKeys** | **References**

Name and Description	Link to full list	Link to InChIKey list	References
Merged NORMAN Suspect List "SusDat"	**Updated Excel File** (1/03/2018) **Interactive Data table** (updating…)	**MS-ready InChIKeys** (1/03/2018)	A merged list of >40,000structures from suspect lists. See **interactive version**. Compiled by Reza Aalizadeh, University of Athens, including RTI and toxicity values, support by Nikiforos Alygizakis, EI. *Work in progress ... please report any issues!*
NORMAN Compounds in MassBank	**CSV**, **XLSX** with Fragments (3/10/2017) CompTox **MassBank EU Reference List** CompTox **MassBank EU Special Cases** CompTox **Fragment Download**	**MassBankEUInChIKeys** (11/04/2017)	**www.massbank.eu** Stravs *et al.* 2013. DOI: **10.1002/jms.3131**
HSWT/LfU STOFF-IDENT database of water-relevant substances	**STOFF-IDENT Contents** (6/09/2017) CompTox **STOFF-IDENT List** Further curation in progress…	**STOFF-IDENT InChIKeys** (6/09/2017)	The database enables the search for exact masses from target or unknown lists and the automatic use of a Retention Time Index. See: **https://www.lfu.bayern.de/stoffident/#!home** (single search for free; batch search after free registration).

Schymanski, Aalizadeh *et al.* in prep.

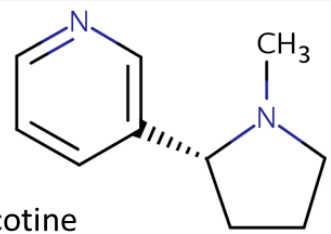
Metadata & The Chemical Identity Crisis

LEGEND: Name, SMILES
DTXSID | InChIKey 1st Block
CAS | **Monoiso. Mass** | logP | **Sources**
Data on: **Toxicity | Exposure | Bioassays**



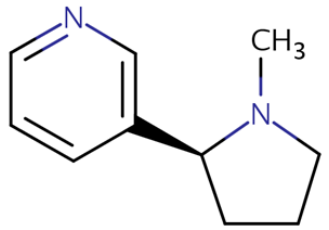
Nicotine

CN1CCC[C@H]1C1=CN=CC=C1
DTXSID1020930| SNICXCGAKADSCV
54-11-5 | **162.1157**| 0.929| **72**
Tox: **yes**| Expo: **yes**| Bioassay: **yes**



D-Nicotine

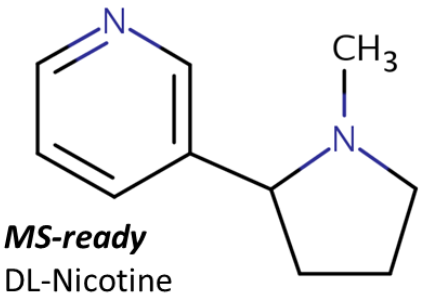
CN1CCC[C@@H]1C1=CN=CC=C1
DTXSID0046351| SNICXCGAKADSCV
25162-00-9 | **162.1157**| 0.929| **20**
Tox: **no**| Expo: **yes**| Bioassay: **yes**



HCl

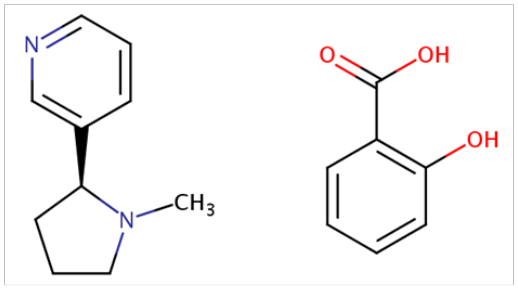
Nicotine hydrochloride

Cl.CN1CCC[C@H]1C1=CN=CC=C1
DTXSID6020931| HDJBTCAJIMNXEW
2820-51-1 | **198.0924** | 0.929| **9**
Tox: **no**| Expo: **yes**| Bioassay: **yes**



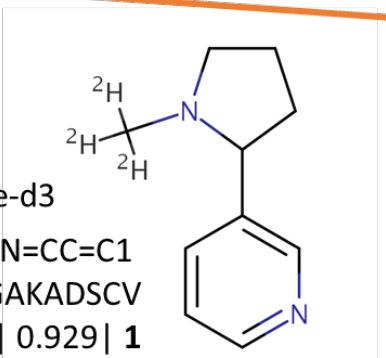
MS-ready
DL-Nicotine

CN1CCCC1C1=CN=CC=C1
DTXSID3048154 | SNICXCGAKADSCV
22083-74-5 | **162.1157**| 0.953| 9
Tox: **yes**| Expo: **no**| Bioassay: **yes**



Benzoic acid, 2-hydroxy-, compd. with
3-[(2S)-1-methyl-2-pyrrolidinyl]pyridine (1:1)

OC(=O)C1=C(O)C=CC=C1.CN1CCC[C@H]1C1=CN=CC=C1
DTXSID5075319| AIBWPBUAKCMKNS
29790-52-1| **300.1474**| 0.929| **6**
Tox: **no**| Expo: **yes**| Bioassay: **no**



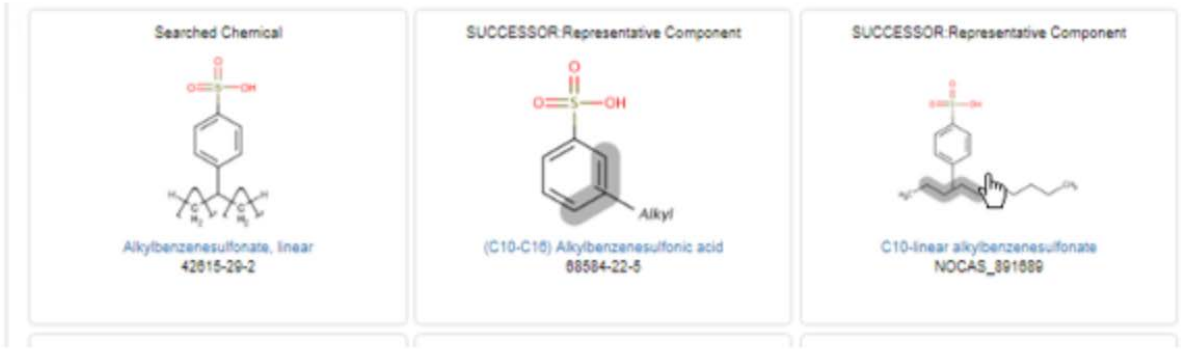
DL-Nicotine-d3

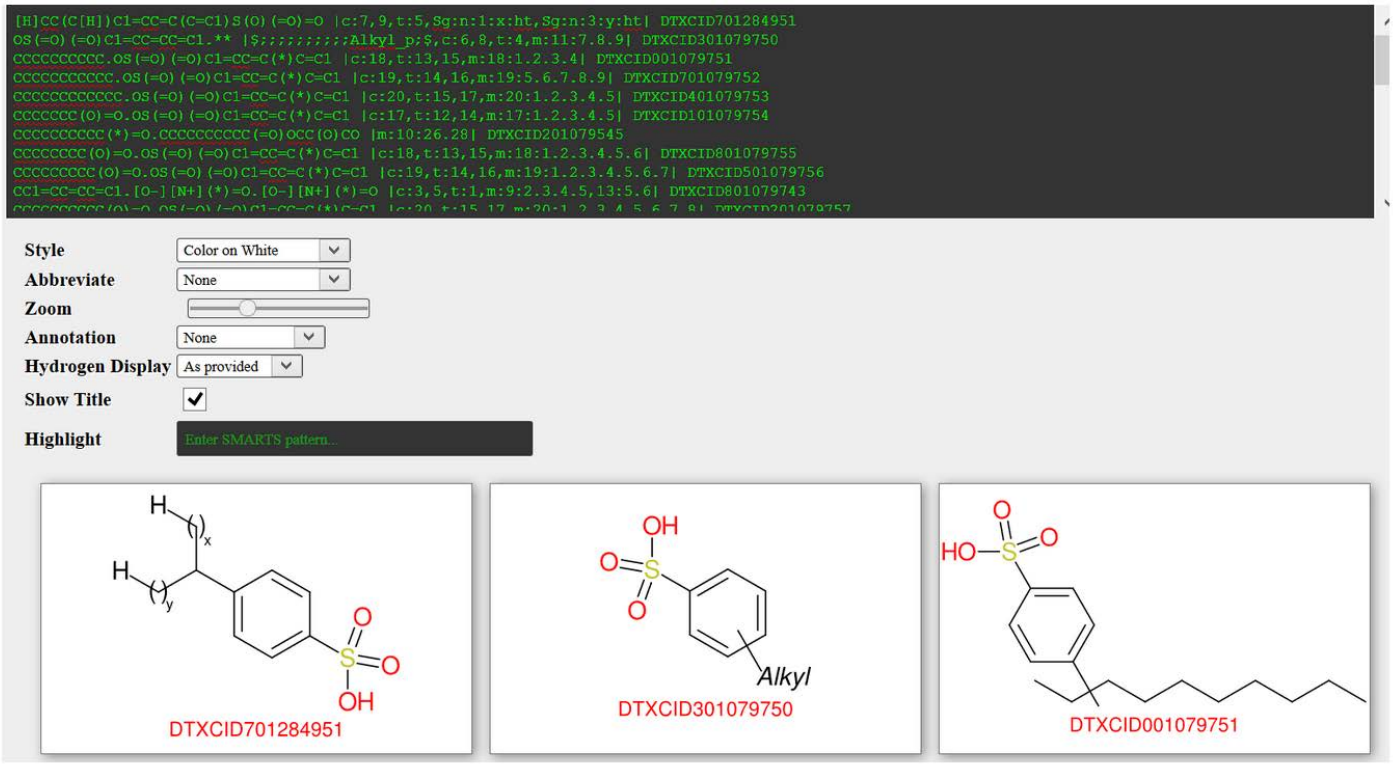
[2H]C([2H])([2H])N1CCCC1C1=CN=CC=C1
DTXSID80442666| SNICXCGAKADSCV
69980-24-1| **165.1345**| 0.929| **1**
Tox: **no**| Expo: **no**| Bioassay: **no**

Schymanski & Williams, 2017, ES&T51 (10), pp 5357–5359. DOI: 10.1021/acs.est.7b01908

Toolkit compatibility will be vital for us …

- https://github.com/cdk/depict/issues/7

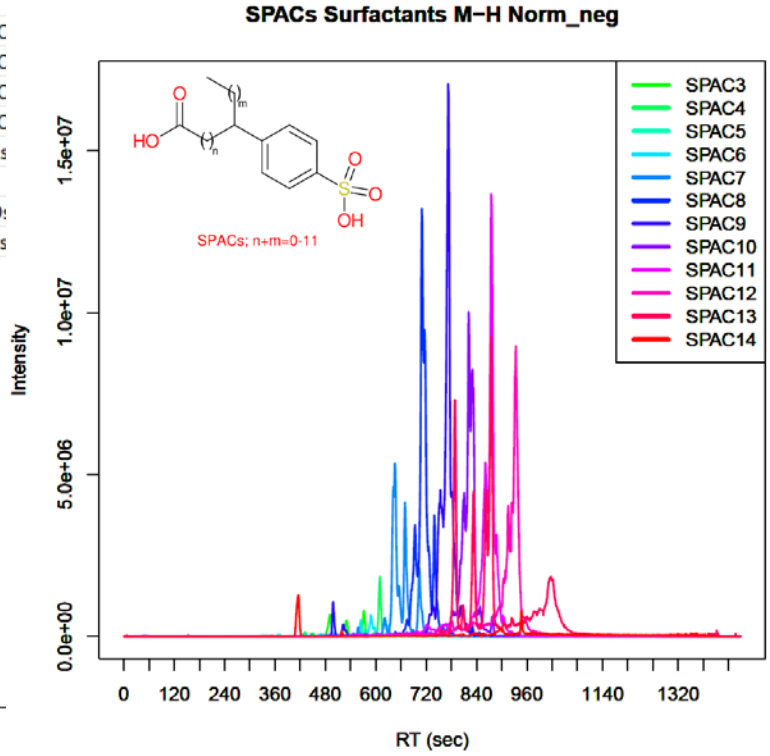


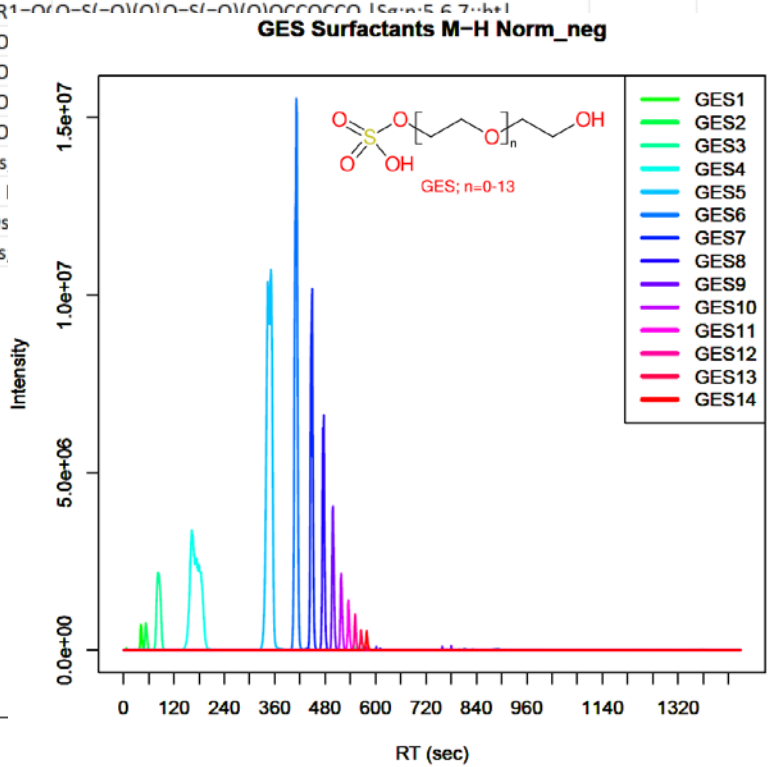


…to get high throughput MS screening of UVCBs

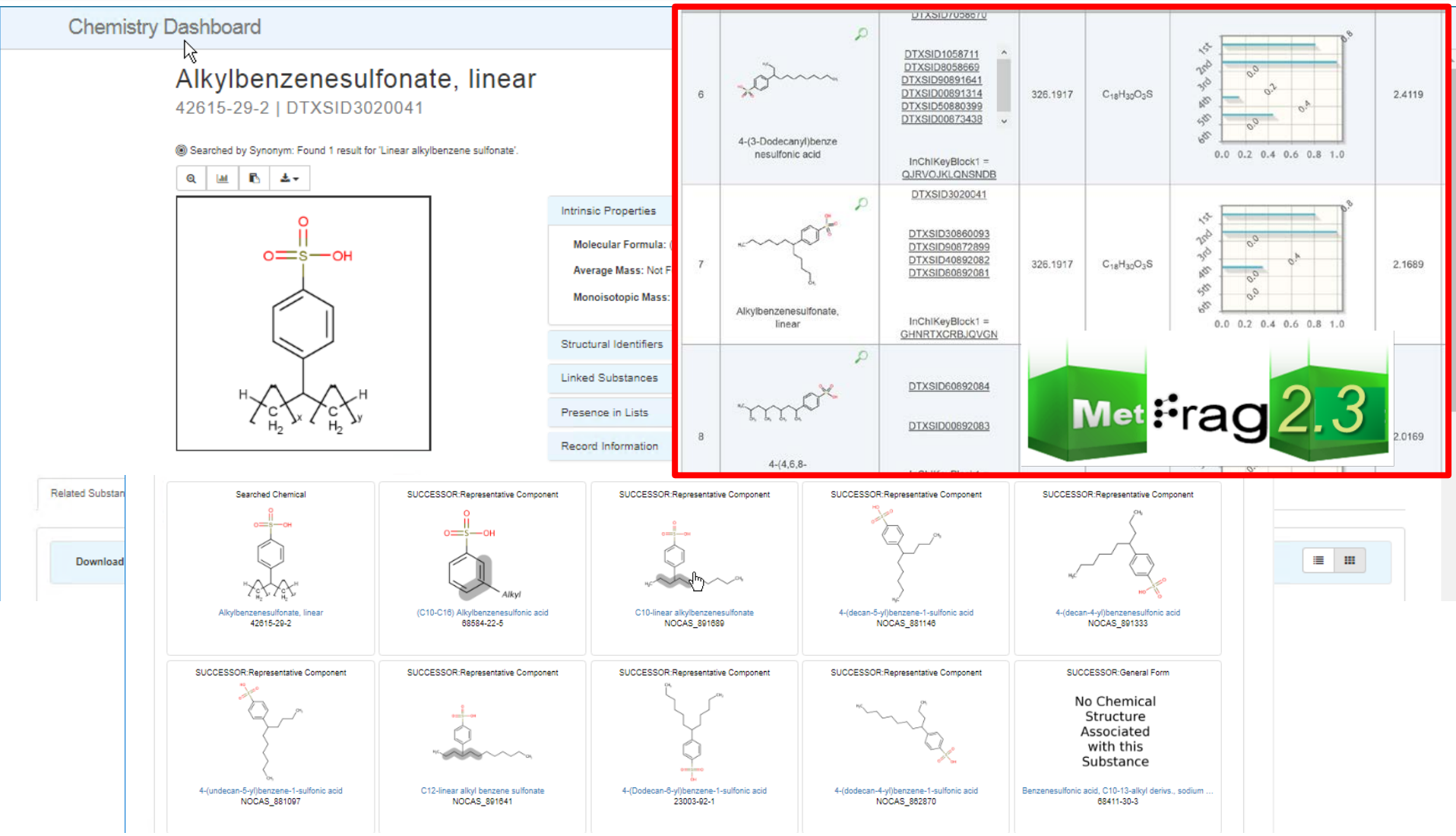
- https://github.com/schymane/RChemMass/

Series_ID	GenSMILES	R1toN	nR1toN	ExtraAtom	RDB_R1toN	SeriesName	Source	BabelSmile	Label1	CxSMILES_	CxSMILES_noTitle
LAS	OS(=O)(=O)	{C,C}	{9-13}			LAS	http://pubs	OS(=O)(=O)	LAS; R1=C,I	OS(=O)(=O)	OS(=O)(=O)c1ccc(cc1)C(CC)CC |Sg:n:13:m:ht,Sg:n:11:n:ht|
SPACs	OS(=O)(=O)	{C,C}	{1-13}			SPACs	http://pubs	OS(=O)(=O)	SPACs; R1=	OS(=O)(=O)	OS(=O)(=O)c1ccc(cc1)C(CC(=O)O)CC |Sg:n:15:m:ht,Sg:n:11:n:ht|
SPADCs	OS(=O)(=O)	{C,C}	{0-13}			SPADCs	http://pubs	OS(=O)(=O)	SPADCs; R1	OS(=O)(=O)	OS(=O)(=O)c1ccc(cc1)C(CC(=O)O)CC(=O)O |Sg:n:15:m:ht,Sg:n:11:n:ht|
DATS	OS(=O)(=O)	{C,C}	{0-15}			DATS	http://pubs	OS(=O)(=O)	DATS; R1=(	OS(=O)(=O)	OS(=O)(=O)c1ccc2c(c1)C(CCC2CC)CC |Sg:n:14:m:ht,Sg:n:16:n:ht|
STACs	OS(=O)(=O)	{C,C}	{1-13}			STACs	http://pubs	OS(=O)(=O)	STACs; R1=	OS(=O)(=O)	OS(=O)(=O)c1ccc2c(c1)C(CCC2CC(=O)O)CC |Sg:n:14:m:ht,Sg:n:18:n:ht|
STADCs	OS(=O)(=O)	{C,C}	{0-13}			STADCs	http://pubs	OS(=O)(=O)	STADCs; R1	OS(=O)(=O)	OS(=O)(=O)c1ccc2c(c1)C(CCC2CC(=O)O)CC(=O)O |Sg:n:14:m:ht,Sg:n:18:n:ht|
AS	O=S(=O)(O)	{C}	{11-15}			AS	http://pubs	O=S(=O)(O)	AS; R1=C(1	O=S(=O)(O)	O=S(=O)(O)OCC |Sg:n:5:n:ht|
C12AES	O=S(O)(=O)	{CCO,C}	{2-11,11-11}			C12AES	http://pubs	O=S(O)(=O)	C12AES; R1	O=S(O)(=O)	O=S(O)(=O)OCCOCCC |Sg:n:5,6,7::ht,Sg:n:9:m:ht|
C13AES	O=S(O)(=O)	{CCO,C}	{2-11,12-12}			C13AES	http://pubs	O=S(O)(=O)	C13AES; R1	O=S(O)(=O)	O=S(O)(=O)OCCOCCC |Sg:n:5,6,7::ht,Sg:n:9:m:ht|
SAS	O=S(O)(=O)	{C,C}	{9-20}			SAS	http://pubs	O=S(O)(=O)	SAS; R1=C,I	O=S(O)(=O)	O=S(O)(=O)C(CC)CC |Sg:n:5:m:ht,Sg:n:7:n:ht|
NPEOSO4s	O=S(=O)(O)	{OCC,C}	{0-2,9-9}			NPEOSO4s	http://pubs	O=S(=O)(O)	NPEOSO4s;	O=S(=O)(O)	O=S(=O)(O)OCCOc1ccc(cc1)CC |Sg:n:5,6,7::ht,Sg:n:14:m:ht|
GES											
CxAEC											
CxAEC											
CxAEC											
CxAEC											
CDEAs											
PEGs											
NPEOs											
NPECs											





…and a HUGE thanks to you all!

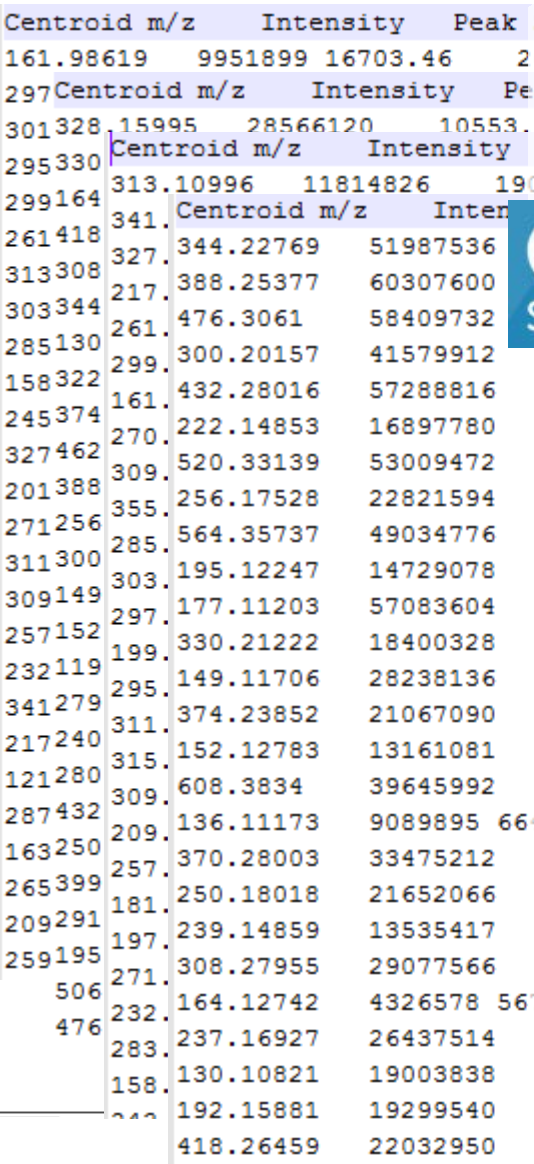


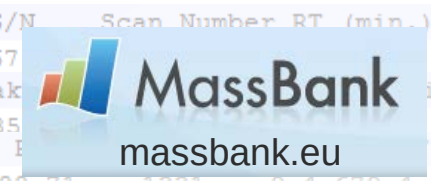
https://comptox.epa.gov/dashboard/chemical_lists/eawagsurf

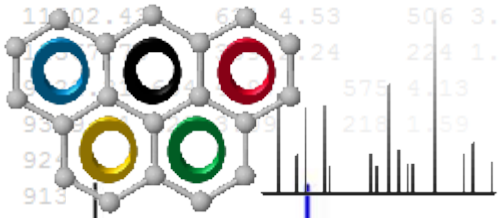
Schymanski, Grulke, Williams et al, in prep. & Williams *et al.* 2017 *J. Cheminformatics* 9:61 DOI: 10.1186/s13321-017-0247-6

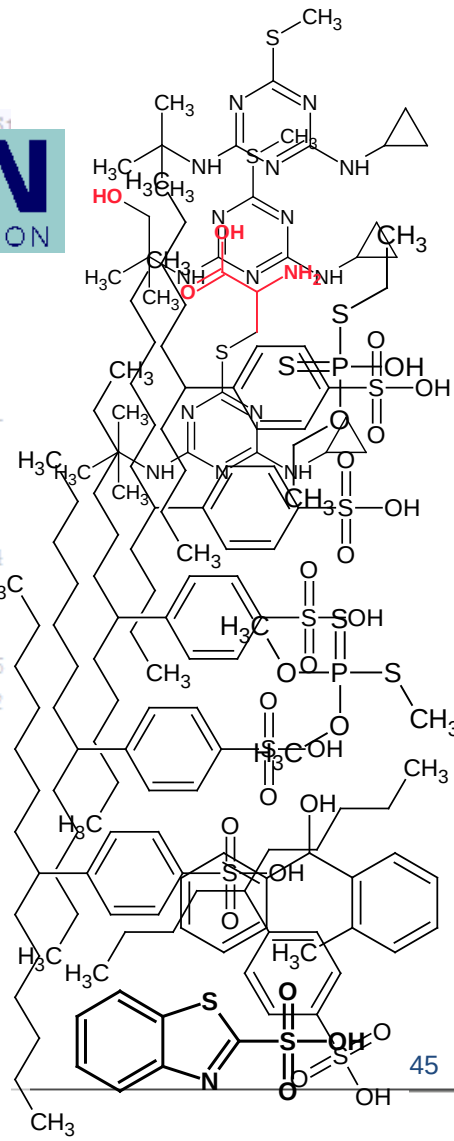
Background: From Molecule to Mass …and back

- Identification = turning numbers into structures









Sampling  extraction (SPE)  HPLC separation  HR-MS/MS 

Conversion (Proteowizard) and Peak Picking (enviPick, xcms, MZmine, …)

Detection of blank/blind/noise/internal standards; time trend analysis (enviMass)

Target List

TARGET ANALYSIS

(enviMass, vendor software)

Suspect List (e.g. NORMAN, LMC, Eawag-PPS, ReSOLUTION)

SUSPECT SCREENING

Gather evidence (nontarget, ReSOLUTION, RMassBank)

NON-TARGET SCREENING

Componentization (nontarget)

Prioritization (enviMass)

Masses of interest

Molecular formula determination (enviPat, GenForm)

MS/MS Extraction (RMassBank)

Non-target identification (MetFrag2.3, ReSOLUTION)

Interpretation, confirmation, peak inventory, confidence and reporting