

Finding Small Molecules in Big Data

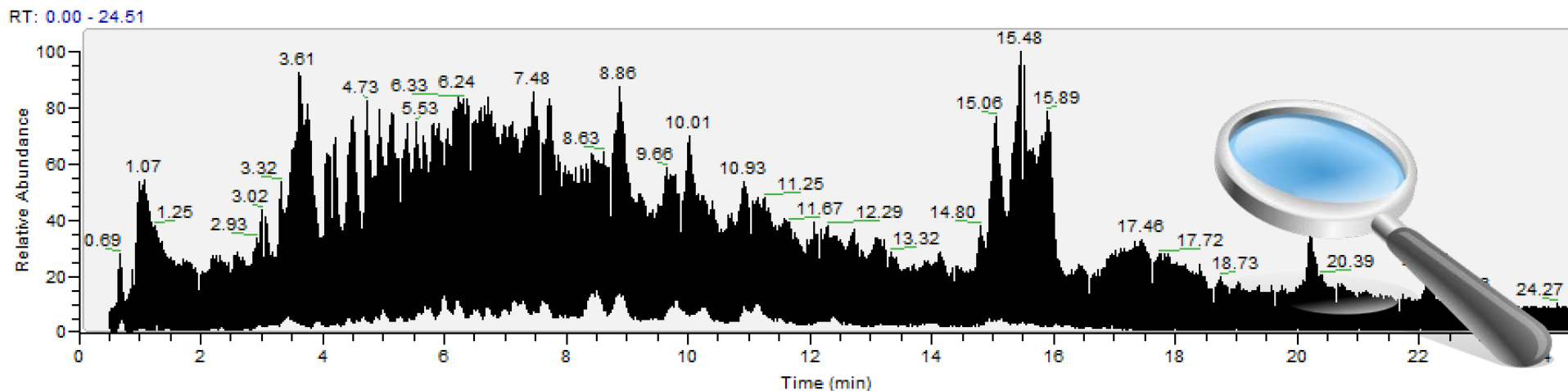


Image © www.seanoakley.com/

Emma Schymanski

Luxembourg Centre for Systems Biomedicine (LCSB),

University of Luxembourg

Email: emma.schymanski@uni.lu

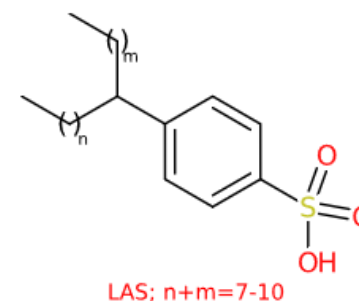
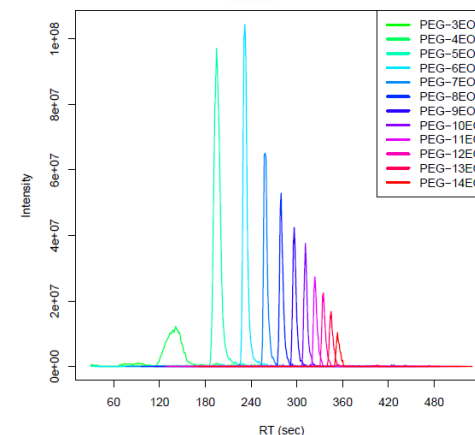
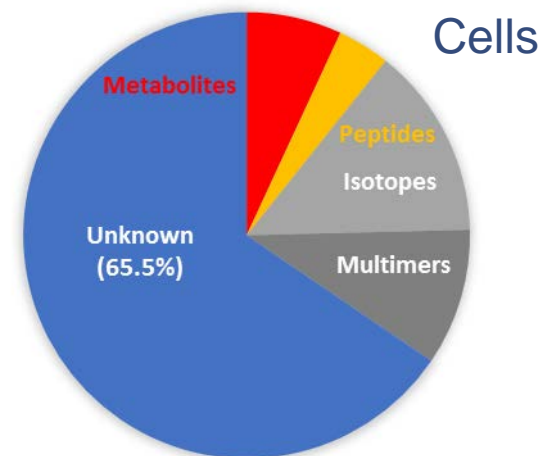
Antony J. Williams

National Centre for Computational Toxicity (NCCT),

US Environmental Protection Agency (US EPA), NC, USA

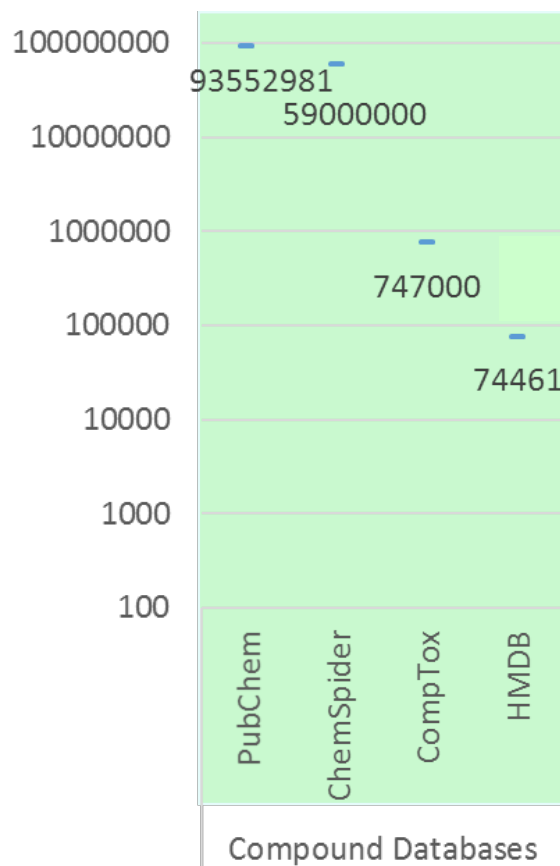
Small molecules ... big problems?

- Status quo of small molecules:
 - How many are in compound databases?
 - How many could there be?
 - How many are in spectral libraries?
 - Mind the Gap!
- Exchanging “expert knowledge”
 - Suspect Lists in Europe
 - Live, retrospective screening & untargeted MS
- Tackling Complex Structures
 - Exchanging Information on Unknowns
 - ...and how Open Science helps!



Searching for Small Molecules ...

○ Compound Databases



PubChem: >95 million

<https://pubchem.ncbi.nlm.nih.gov/>

ChemSpider: >60 million

<http://www.chemspider.com/>

CompTox Chemicals Dashboard: >761 000

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

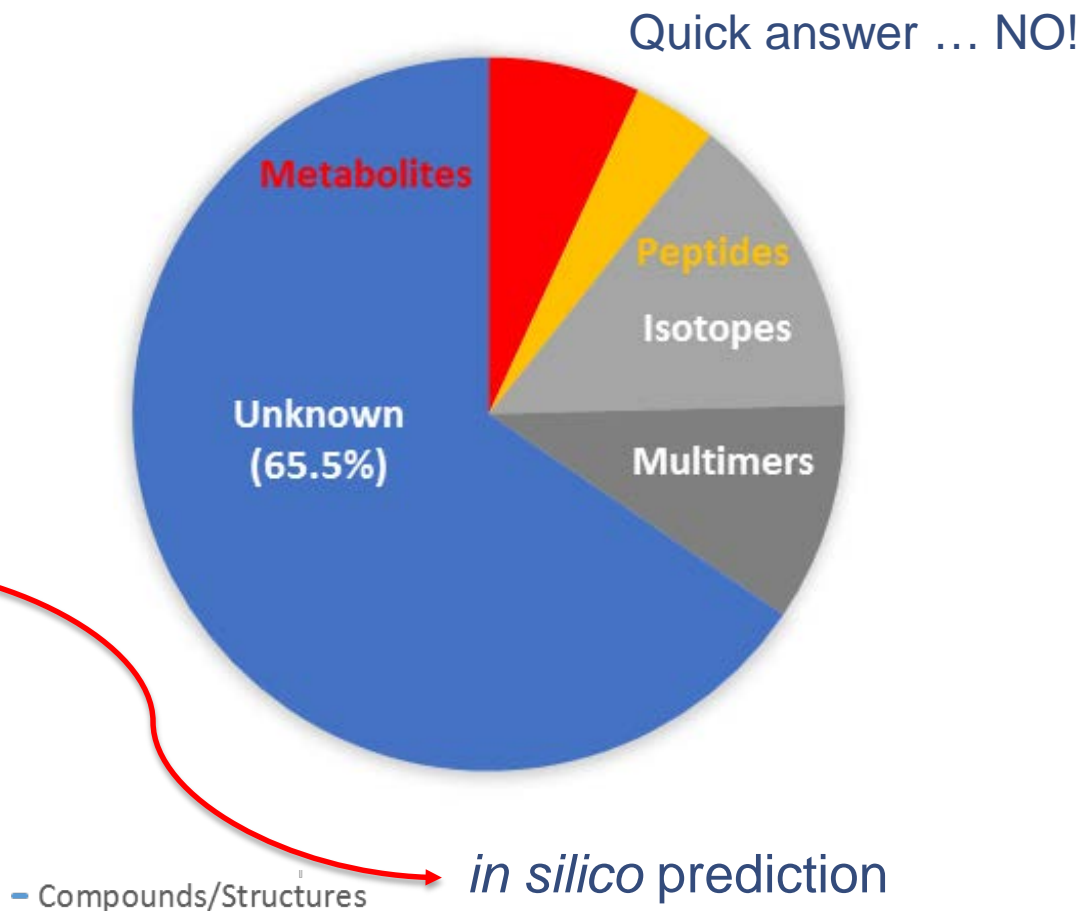
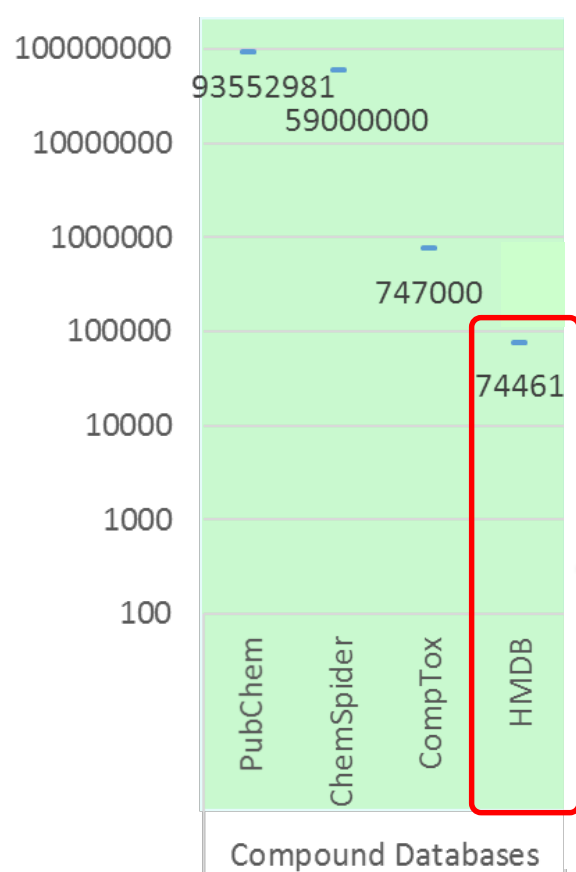
Human Metabolome DB (HMDB): >115 000

<http://www.hmdb.ca/>

— Compounds/Structures

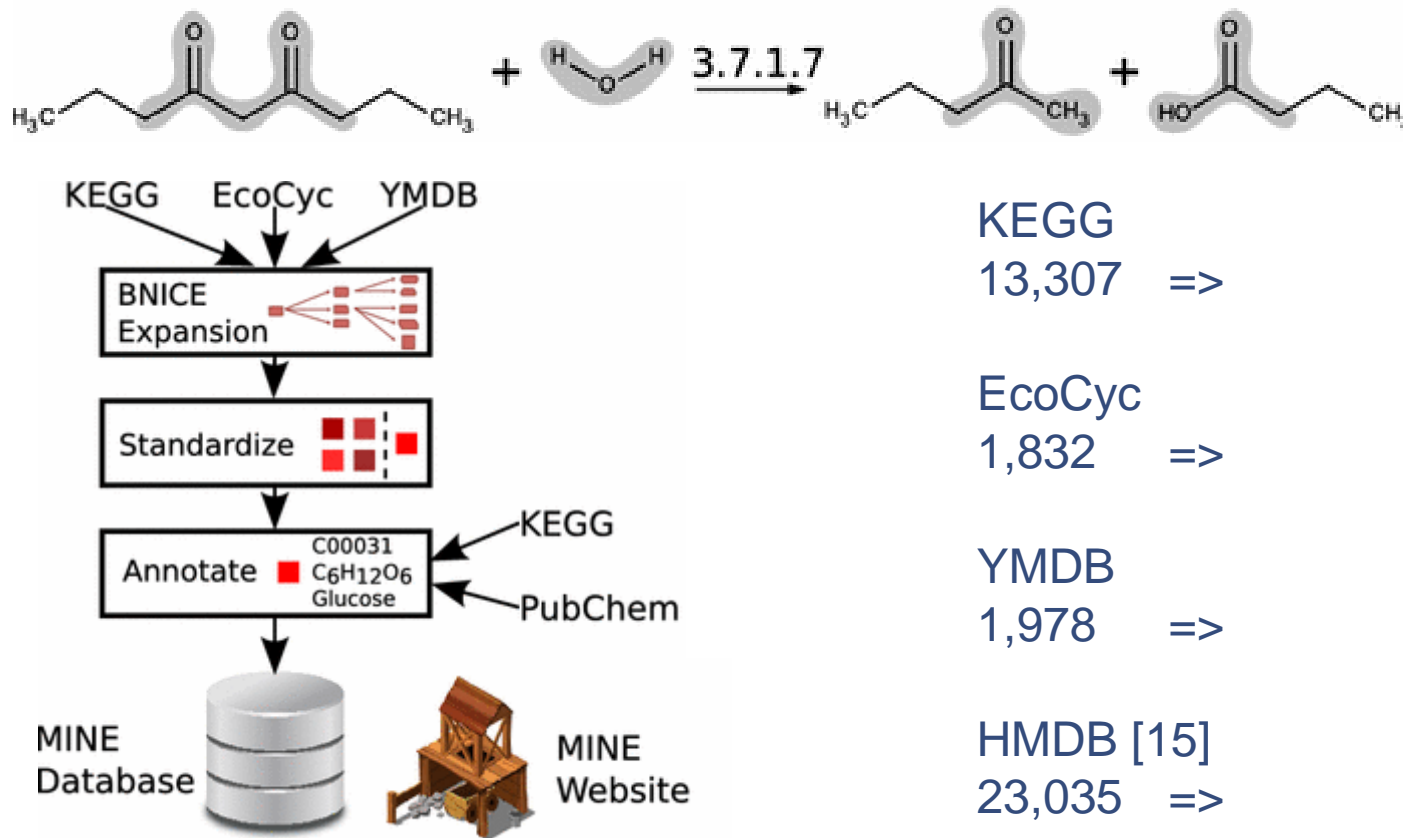
Searching for Small Molecules ...

- Compound Databases ... isn't 95 million enough?



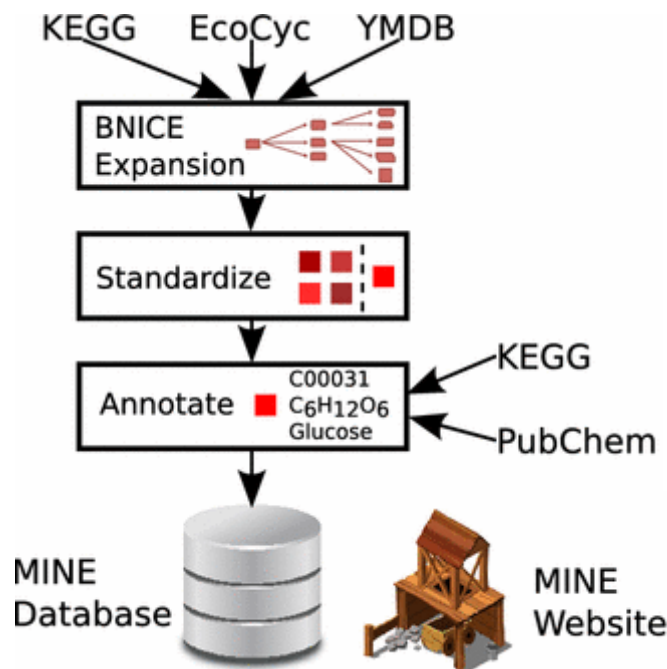
Searching for More Small Molecules ...

- *In silico* metabolite prediction – example of MINE (2015)



Searching for More Small Molecules ...

- *In silico* metabolite prediction – example of MINE (2015)
 - First generation only ... combinatorial explosion!



KEGG
13,307 => **MINE**
571,368

EcoCyc
1,832 => **MINE**
54,719

YMDB
1,978 => **MINE**
100,755

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

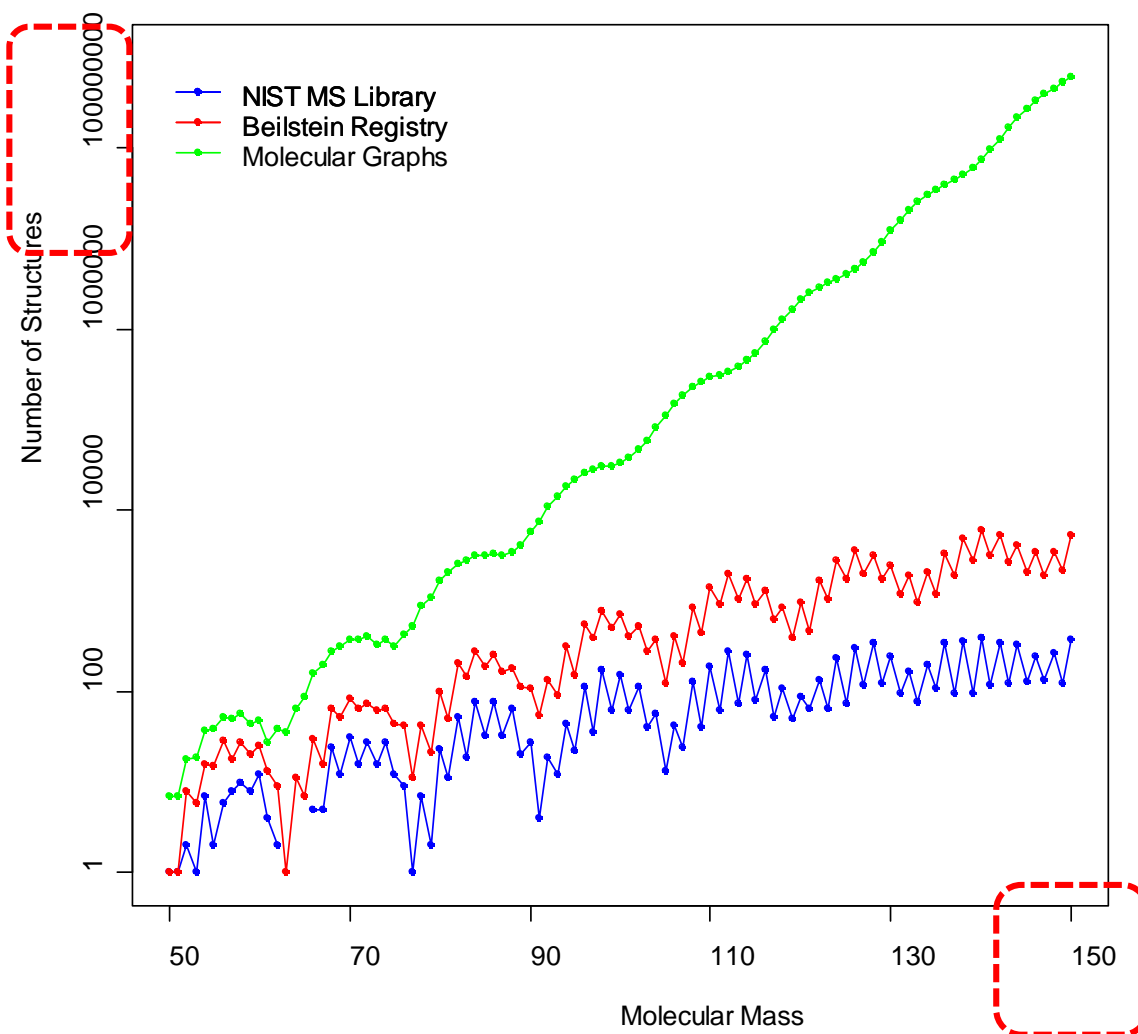
Speculation ...

PubChem
95 million => **MINE**
1.6 billion ... first generation only?!?!

Searching for EVEN MORE Small Molecules ...

○ Structure Generation

- But of course most of these do not exist



Structure Generation
100 million at MW 150

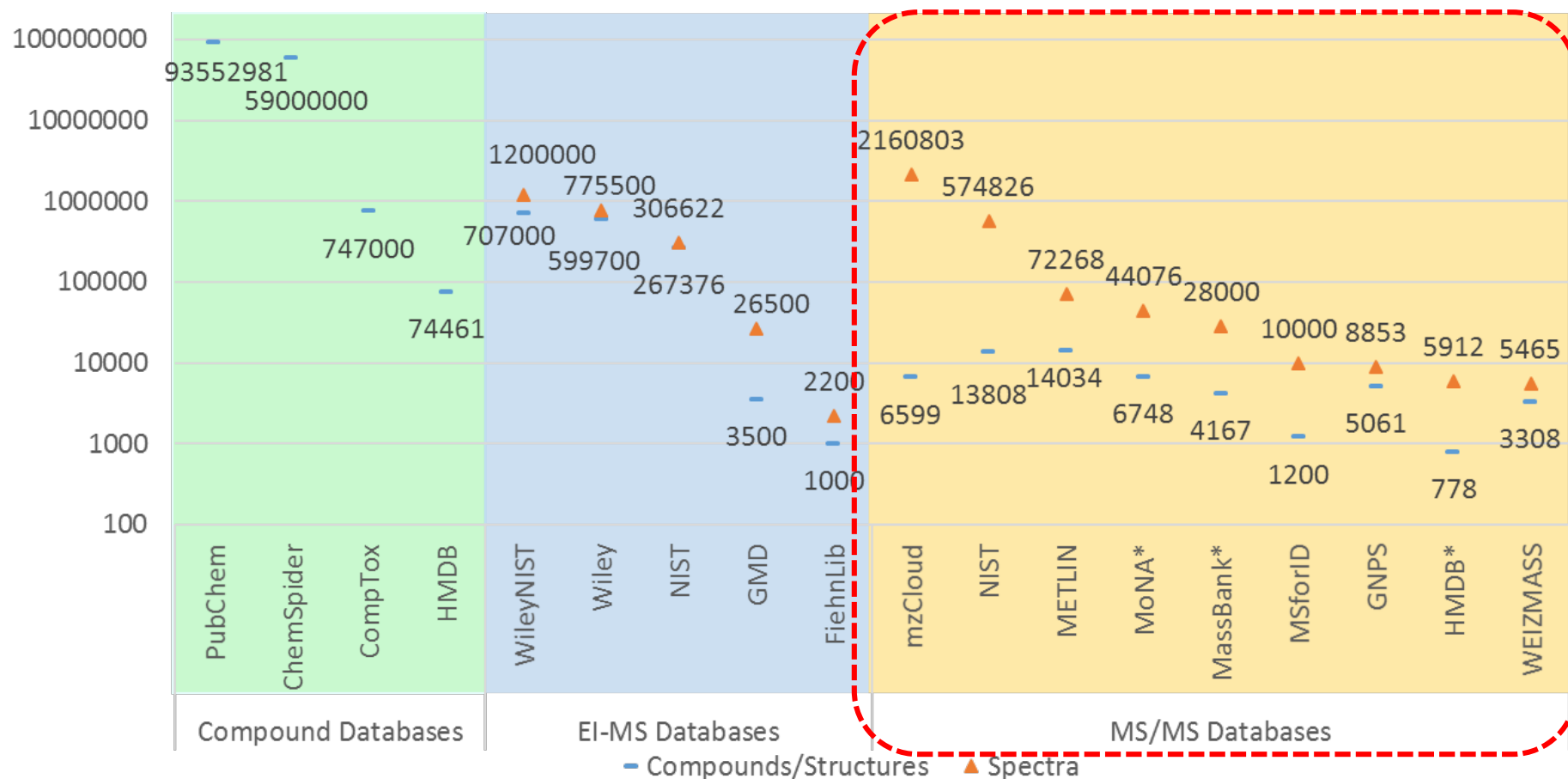
NIST MS Library
~1-200 at MW 150

Spectral Libraries

Searching for Small Molecules in Spectral Libraries

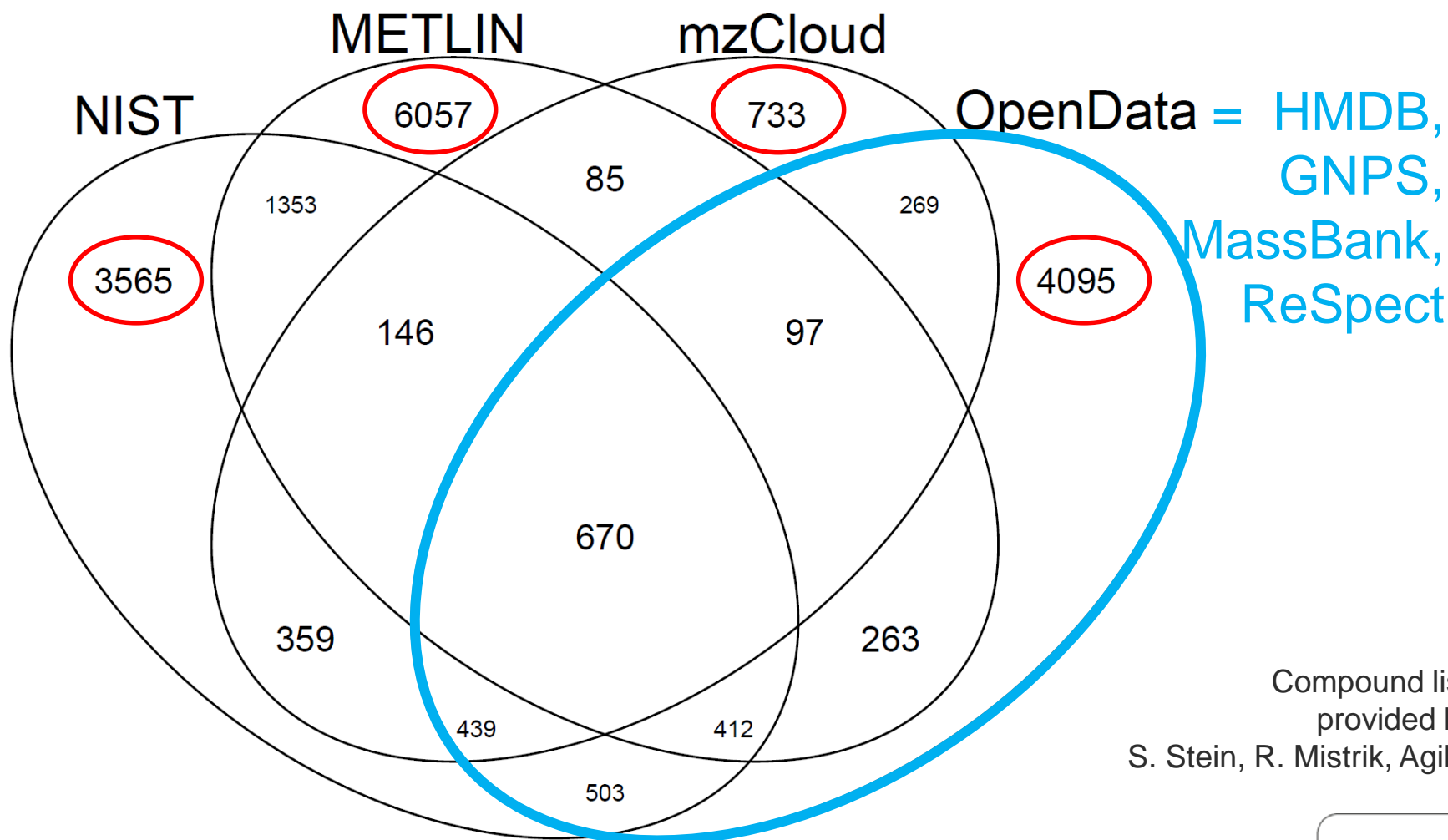
○ ... to find what is “on record”...

- Too many different MS/MS libraries (and they are still too small)



Do we need all these libraries?

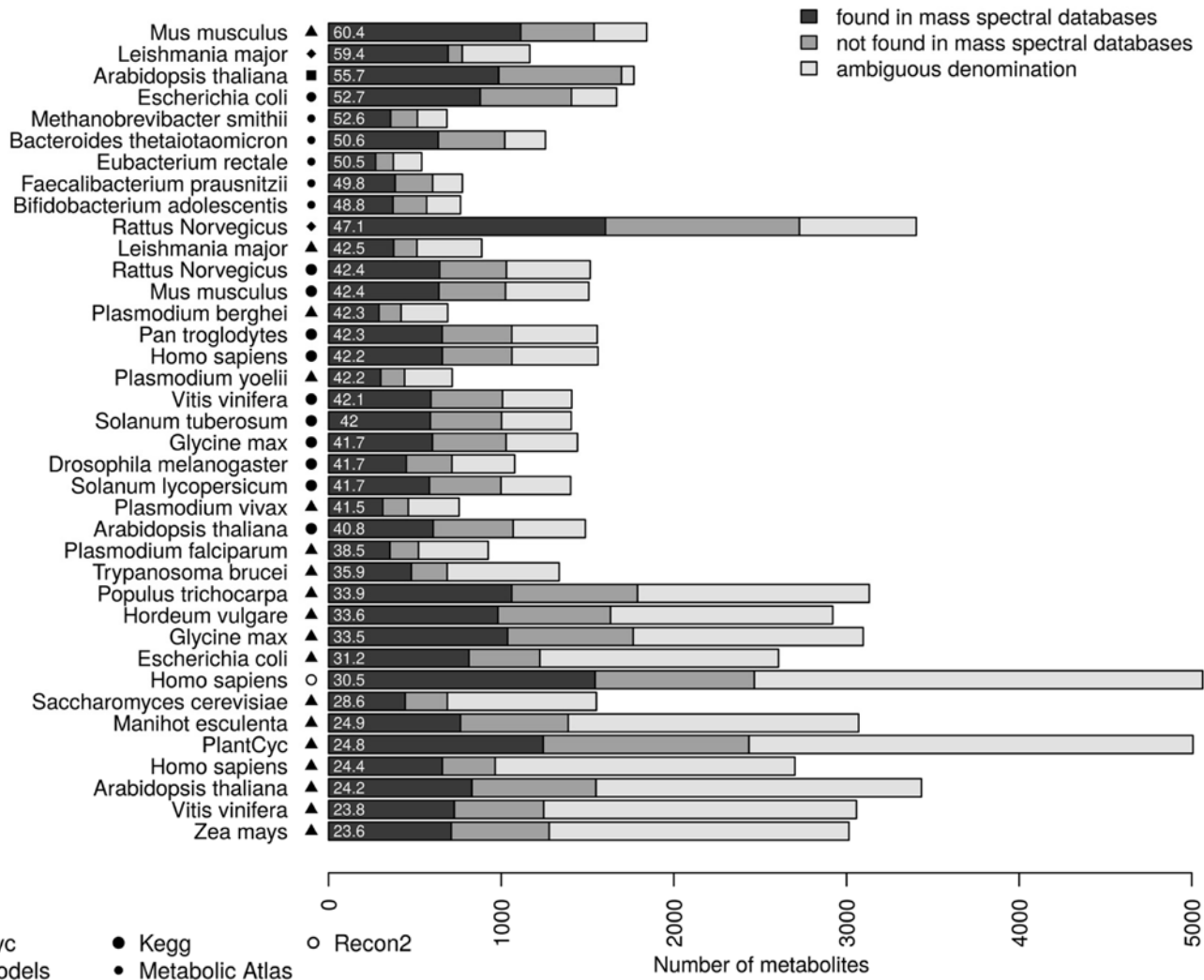
- Yes ... most libraries still have many **unique entries**



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

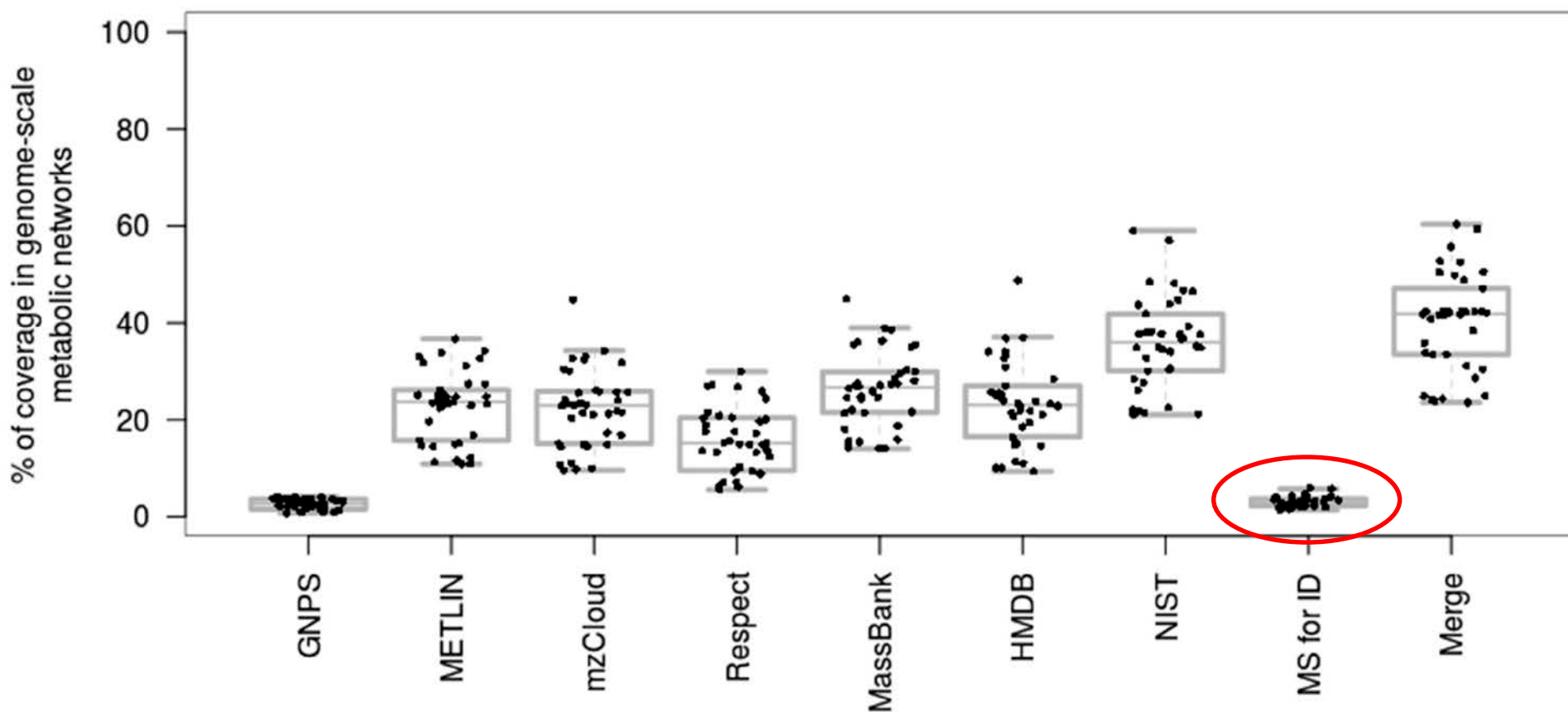
Mind the Gap!

- Only 23-60 % of (defined) metabolites in Genome-Scale Metabolic Networks are covered by (combined!) Mass Spectral Libraries



Mind the Gap!

- Best library to choose depends highly on your dataset
 - Example: MSforID (<https://msforid.com/>) is poor for metabolic networks – but great for forensic toxicology!



SPectraL hASH (SPLASH) – Search between libraries

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

splash10 - 0002 - 09000000000 - b112e4e059e1ecf98c5f
[version] - [top10] - [histogram] - [hash of full spectrum]



Human Metabolome Database: LC-MS/MS Spectrum - LC-ESI-QTOF ...

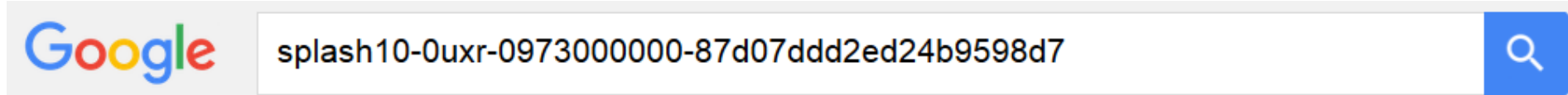
www.hmdb.ca/spectra/ms_ms/5464 ▼

... Spectrum - LC-ESI-QTOF (UPLC Q-ToF Premier, Waters) 30V, Positive. Splash Key:

splash10-0002-09000000000-b112e4e059e1ecf98c5f View in MoNA ...

Human Metabolome Database: Showing metabocard for Caffeine ...

www.hmdb.ca/metabolites/HMDB01847 ▼



DrugBank: Codeine

www.drugbank.ca/drugs/DB00318 ▼

... 60V, Positive, **splash10-0uxr-0973000000-87d07ddd2ed24b9598d7**, View in MoNA. MS, Mass

Spectrum (Electron Ionization), splash10-01ot-3950000000- ...

Codeine Mass Spectrum - MassBank

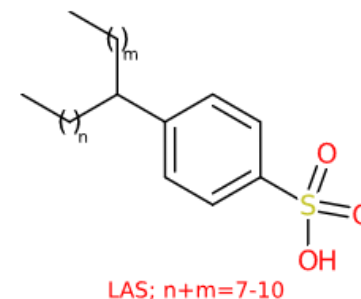
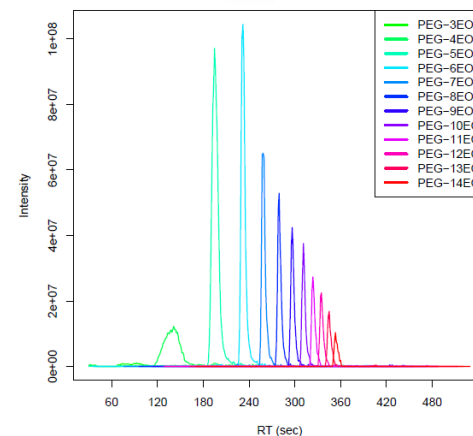
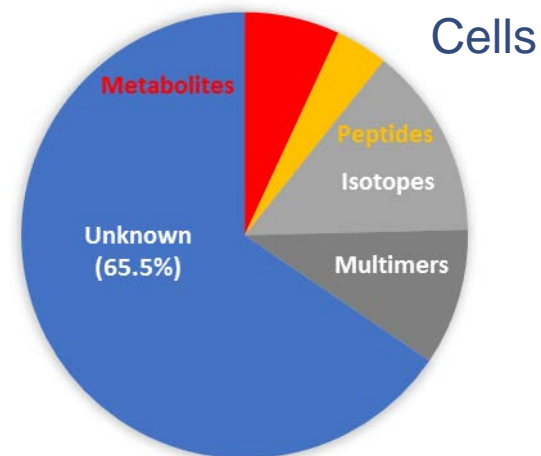
massbank.eu/MassBank/jsp/Dispatcher.jsp?type=disp&id=EA278005&site=31 ▼

PK\$SPLASH: **splash10-0uxr-0973000000-87d07ddd2ed24b9598d7** PK\$ANNOTATION: m/z

tentative_formula formula_count mass error(ppm) 58.0651 ...

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?
 - Mind the Gap!
- Exchanging “expert knowledge”
 - Suspect Lists in Europe
 - Live, retrospective screening & untargeted MS
- Tackling Complex Structures
 - Exchanging Information on Unknowns
 - ...and how Open Science helps!



2015: European Non-target Screening Trial

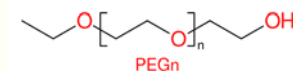
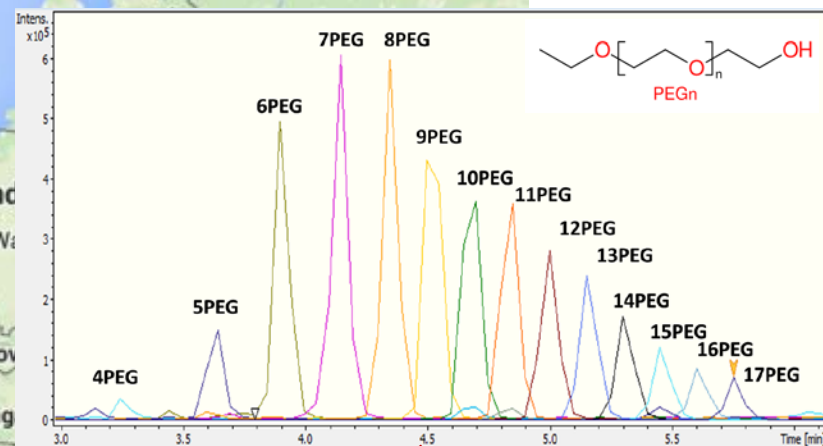
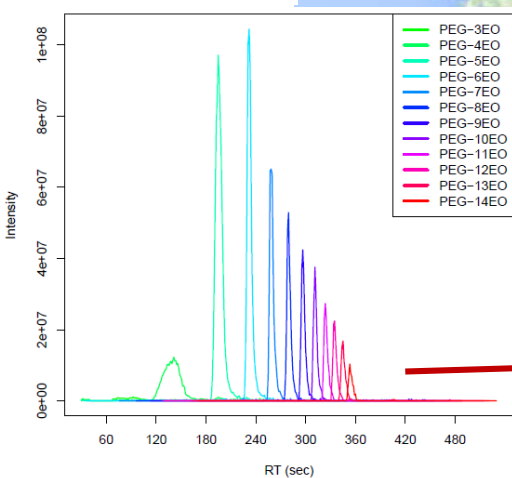
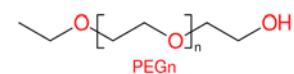


European (World-)Wide Exchange of Suspects



NORMAN Suspect List Exchange:

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



NORMAN Suspect Exchange Lists

- <http://www.norman-network.com/?q=node/236>
- 21 lists available ... specialist collections to market lists
 - Integrated in NORMAN Databases & CompTox Chemistry Dashboard



Emma Louise Schymanski
added an **update**

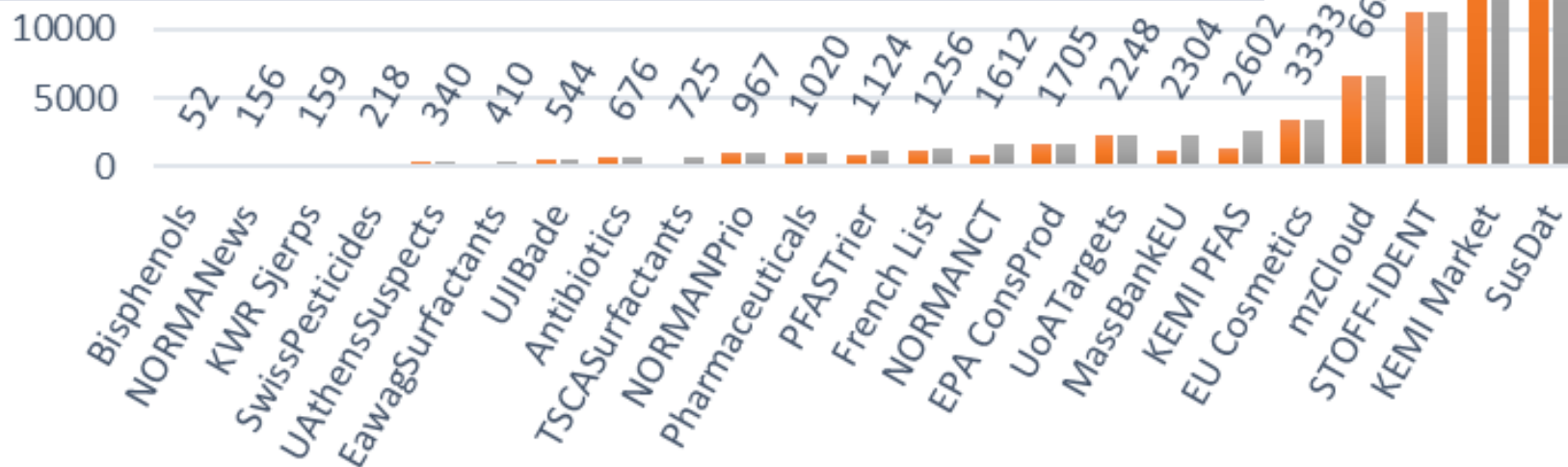
1d ago ▾

Coordinated with publications!

NormaNEWS: retrospective screening of emerging contaminants

More news: one of our favourite examples, the NORMAN Network's pilot trial for global retrospective screening of emerging contaminants has just been accepted in ES&T - full list on the NORMAN Suspect Exchange and the CompTox Dashboard.

<https://pubs.acs.org/doi/pdf/10.1021/acs.est.8b00365>



NormaNEWS: Norman Early Warning System

Q

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

Sort by:

Name

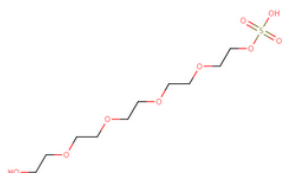


4 of 131 chemicals selected

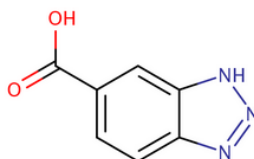
Hide:

Unselected x

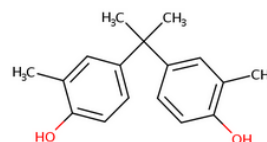
Deselect all



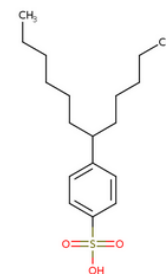
14-hydroxy-3,6,9,12-tetraoxatetradecyl...
NOCAS 881042



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

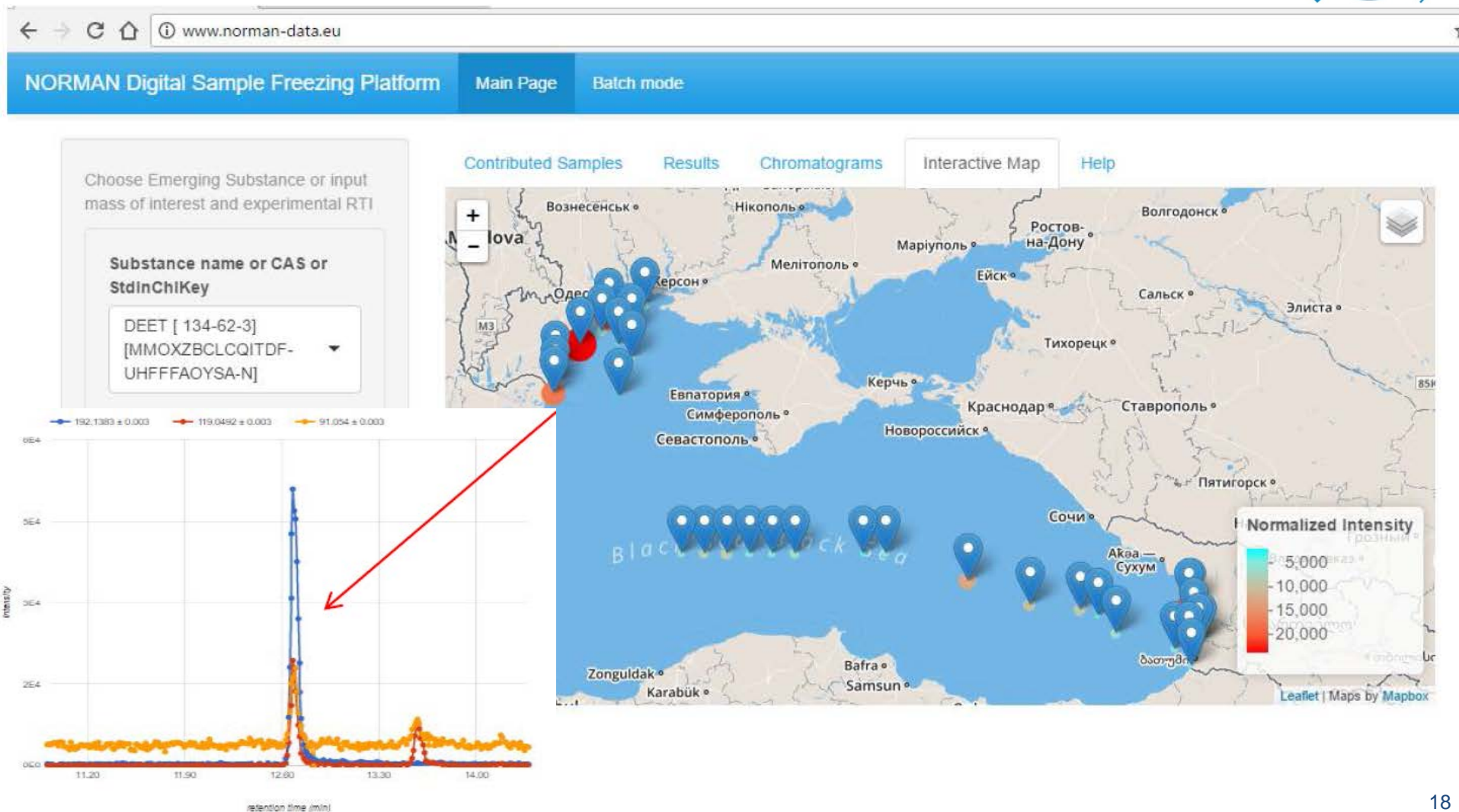


3,3'-Dimethylbisphenol A
79-97-0



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

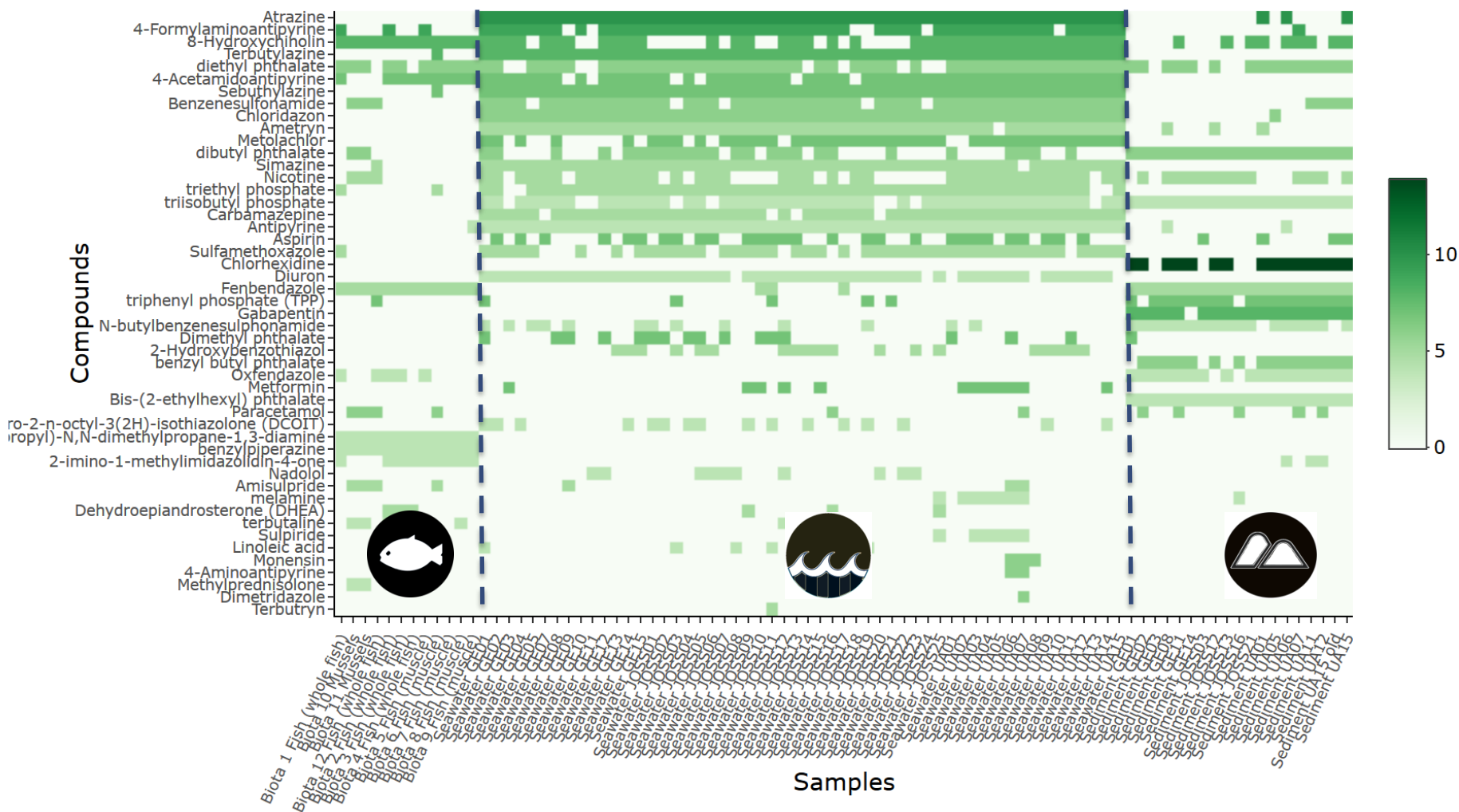
“Live” retrospective screening of **known** and unknown chemicals in European samples (various matrices)



Retrospective screening of REACH chemicals in Black Sea samples (various matrices)



Occurrence Results



norman

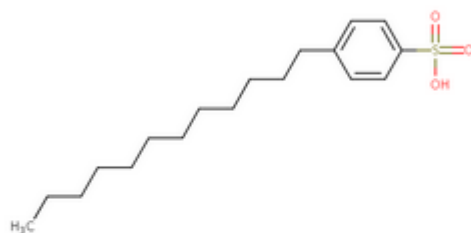


Fatty acids, coco, reaction products with...
CASRN:68604-76-2
CPDAT:0
Sources:6

Alcohols, C12-13, ethers with polyethyle...
CASRN:68908-51-0
CPDAT:0
Sources:6

Challenge 2: Mass Spec “sees” **ONE** part at a time

- What do all these chemicals have in common?

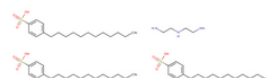


4-Dodecylbenzenesulfonic acid

CASRN:121-65-3

Sources:19

TOXCAST:0

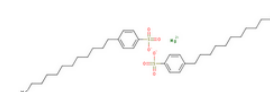


Benzenesulfonic acid, 4-dodecyl-, com...

CASRN:67924-18-9

Sources:5

TOXCAST:0

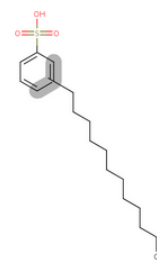


Magnesium bis(4-dodecylbenzene-1-s...

CASRN:77860-72-1

Sources:2

TOXCAST:0

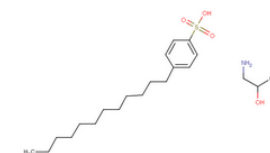


Dodecylbenzenesulfonic acid

CASRN:27176-67-6

Sources:62

TOXCAST:176/617



Benzenesulfonic acid, 4-dodecyl-, com...

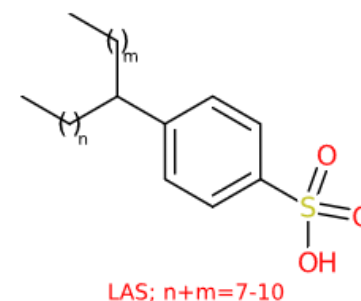
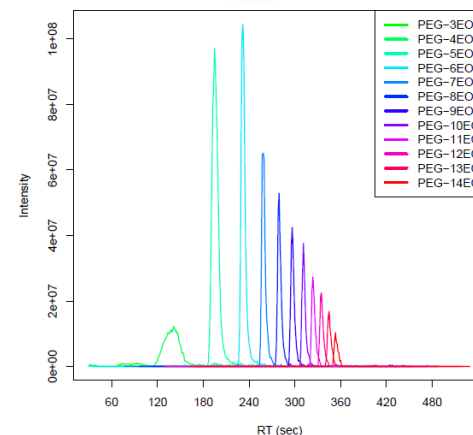
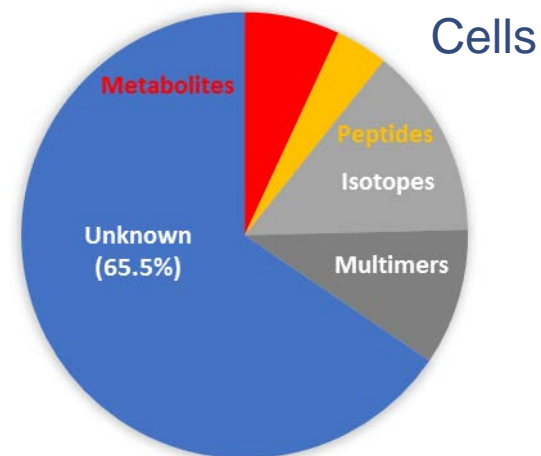
CASRN:54590-52-2

Sources:5

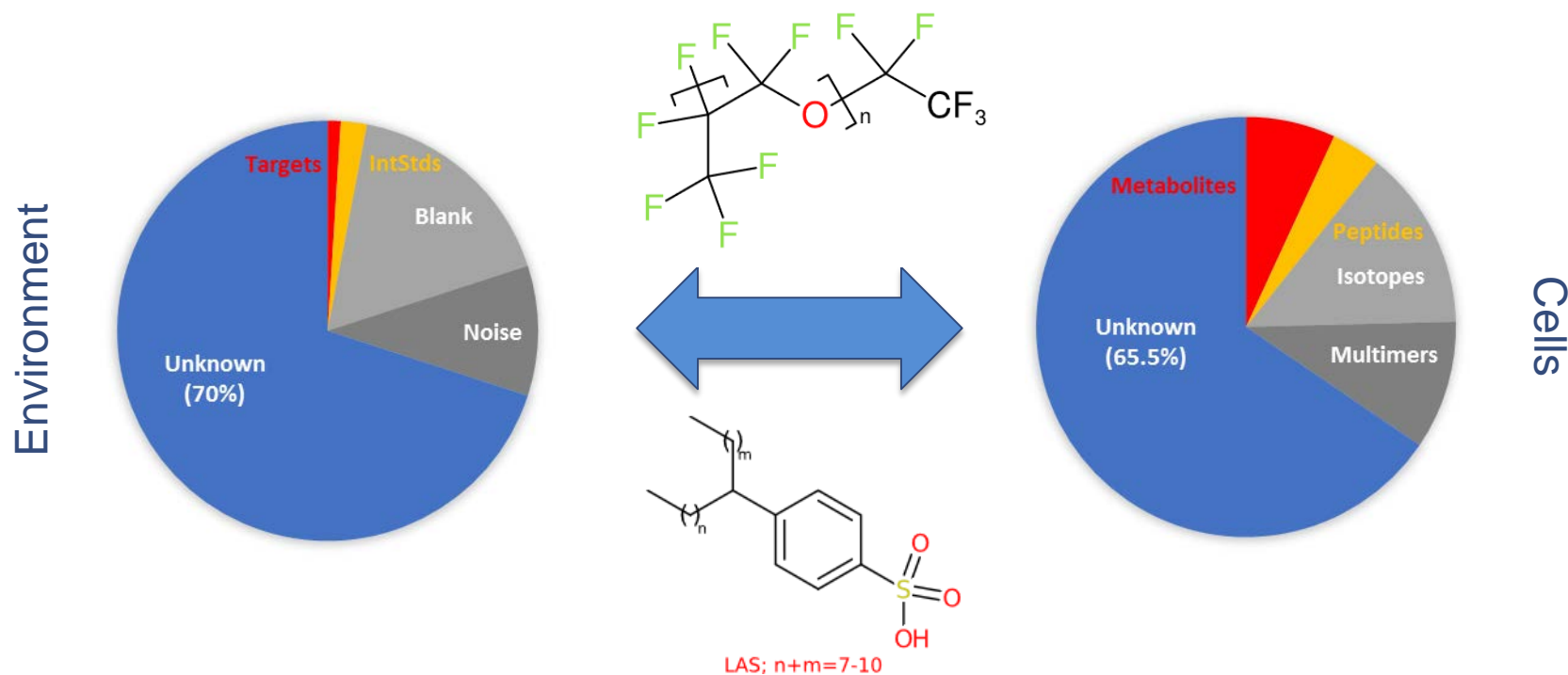
TOXCAST:0

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?
 - Mind the Gap!
- Exchanging “expert knowledge”
 - Suspect Lists in Europe
 - Live, retrospective screening & untargeted MS
- Tackling Complex Structures
 - Exchanging Information on Unknowns
 - ...and how Open Science helps!

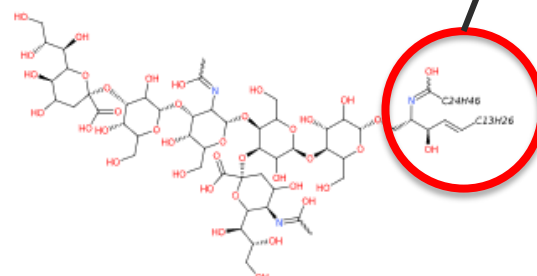
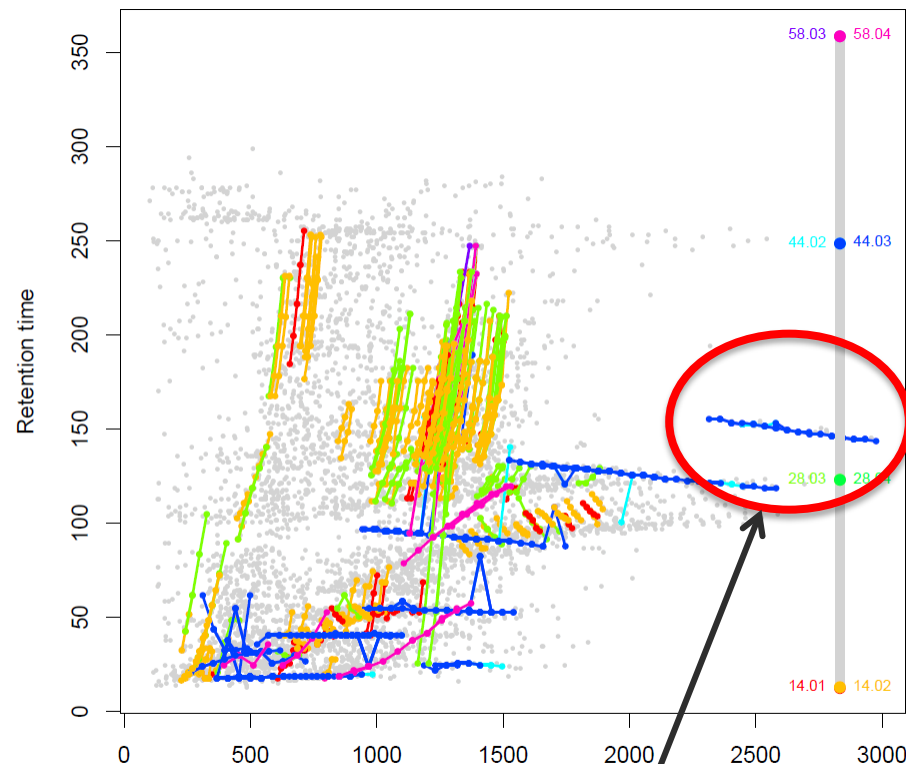
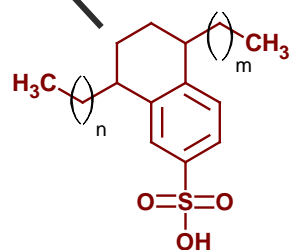
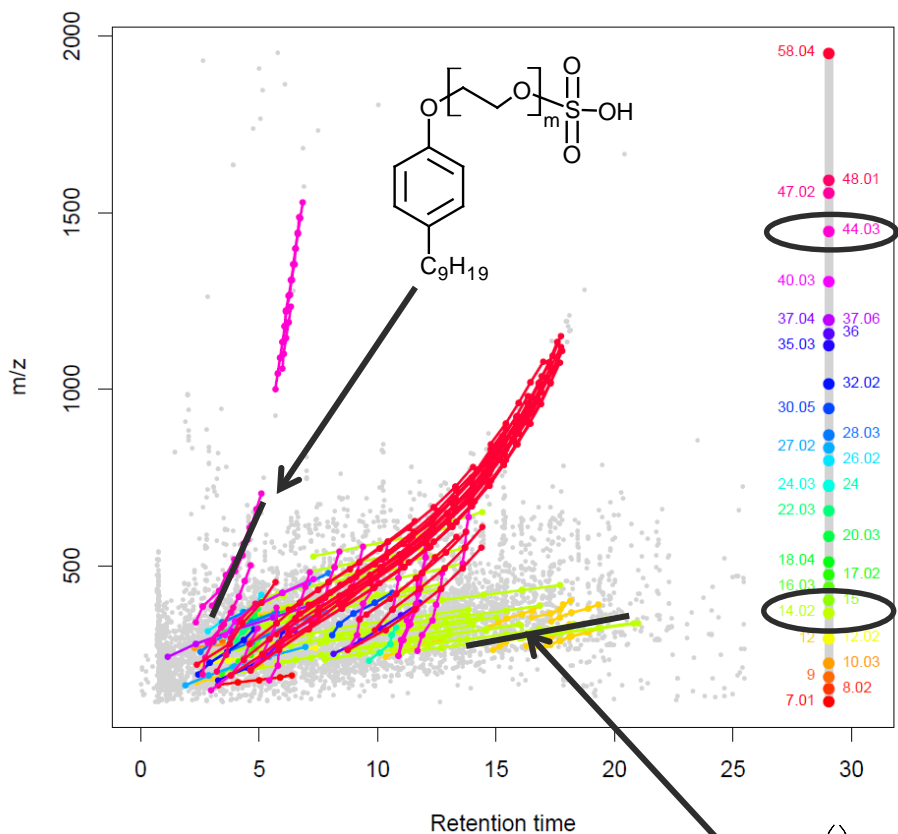


We still have many unknowns ...



...and many are interconnected by mass

Homologous Series in environmental and biological samples



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

New Ways to Store and Access Homologues



Chemistry Dashboard

Submit Comment

Share

Copy

Aa

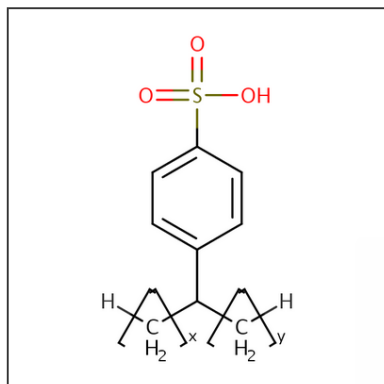
Aa

Aa

Alkylbenzenesulfonate, linear

42615-29-2 | DTXSID3020041

© Searched by DSSTox_Substance_Id: Found 1 result for 'DTXSID3020041'.



Intrinsic

Molecular

Availability

Monomer

Structure

Record

Quality

Download / Send

Sort by:

Relationship

19 chemicals

Hide:

Select all

Searched Chemical

Alkylbenzenesulfonate, linear



42615-29-2

General Form

3 related chemical structures with this substance

Benzenesulfonic acid, C10-16-alkyl der...

General Form

2 related chemical structures with this substance

Benzenesulfonic acid, C10-13-alkyl der...

General Form

1 related chemical structure with this substance

Benzenesulfonic acid, C10-16-alkyl der...

General Form

Markush Child

Markush Child

Markush Child

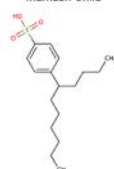
Related Substances

Chemical Properties

Analytical

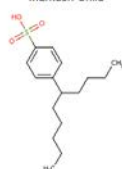
Comments

Markush Child



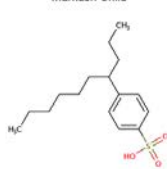
4-(undecan-5-yl)benzene-1-sulfonic acid
NOCAS_881097

Markush Child



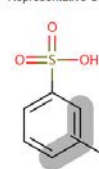
4-(decan-5-yl)benzene-1-sulfonic acid
NOCAS_881146

Markush Child



4-(docan-4-yl)benzenesulfonic acid
NOCAS_891333

Representative Component



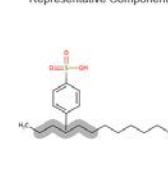
(C10-C16) Alkylbenzenesulfonic acid
68584-22-5

Representative Component

4 related chemical structures with this substance

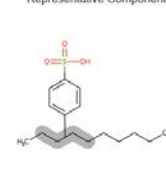
Benzenesulfonic acid, dodecyl-, branch...

Representative Component



C12-linear alkyl benzene sulfonate
NOCAS_891641

Representative Component



C10-linear alkylbenzenesulfonate
NOCAS_891689



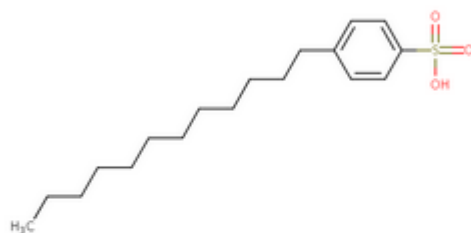
<https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID3020041>

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



RECAP: Mass Spec “sees” ONE part at a time

- What do all these chemicals have in common?

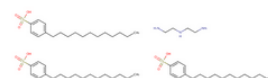


4-Dodecylbenzenesulfonic acid

CASRN:121-65-3

Sources:19

TOXCAST:0

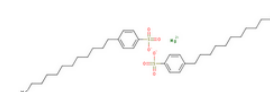


Benzenesulfonic acid, 4-dodecyl-, com...

CASRN:67924-18-9

Sources:5

TOXCAST:0

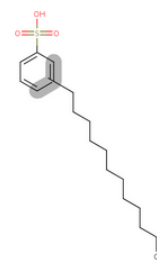


Magnesium bis(4-dodecylbenzene-1-s...

CASRN:77860-72-1

Sources:2

TOXCAST:0

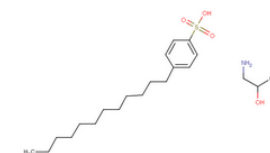


Dodecylbenzenesulfonic acid

CASRN:27176-67-6

Sources:62

TOXCAST:176/617



Benzenesulfonic acid, 4-dodecyl-, com...

CASRN:54590-52-2

Sources:5

TOXCAST:0

MS-Ready: Accessing Data from Salts and Mixtures



MetFrag

In silico f

#	Molecule	Identifier
---	----------	------------

Database Settings

Database:

Neutral Mass: Search p

Formula:

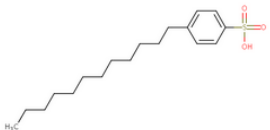
Identifiers:

Retrieve Candidates

72 Candidates

Weights

MetFrag (1st)	<input type="range"/>
ExactSpectralSimilarity (2nd)	<input type="range"/>
DATA_SOURCES (3rd)	<input type="range"/>
PUBCHEM_DATA_SOURCES (4th)	<input type="range"/>
TOXCAST_PERCENT_ACTIVE (5th)	<input type="range"/>



4-Dodecylbenzenesulfonic acid
Sources:19
TOXCAST:0
PubChem:56

DTXSID8050443

DTXSID9042413
DTXSID8069226
DTXSID9065786
DTXSID5074264
DTXSID2065759
DTXSID5070864

InChIKeyBlock1 =
KWXCIGTUELQSQ

DTXSID0041642

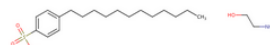
DTXSID2049391
DTXSID9068031
DTXSID7071078
DTXSID8065244
DTXSID9067433
DTXSID0067923
DTXSID0068000

InChIKeyBlock1 =
WBIQQQGBSDOWNP


DTXSID0041644

DTXSID40860241
DTXSID2071106
DTXSID20892388


InChIKeyBlock1 =
HRPQWSOMACYCRG



Benzenesulfonic acid, 4-dodecyl-, com...
Sources:11
TOXCAST:0
PubChem:16



N,N-Dimethyl-1,3-propanediamine 4-do...
Sources:7
TOXCAST:0
PubChem:9

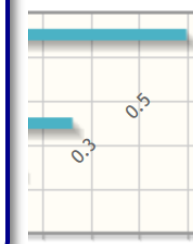


Benzenesulfonic acid, 4-dodecyl-, com...
Sources:5
TOXCAST:0
PubChem:13

Normalized Scores



0.4 0.6 0.8 1.0



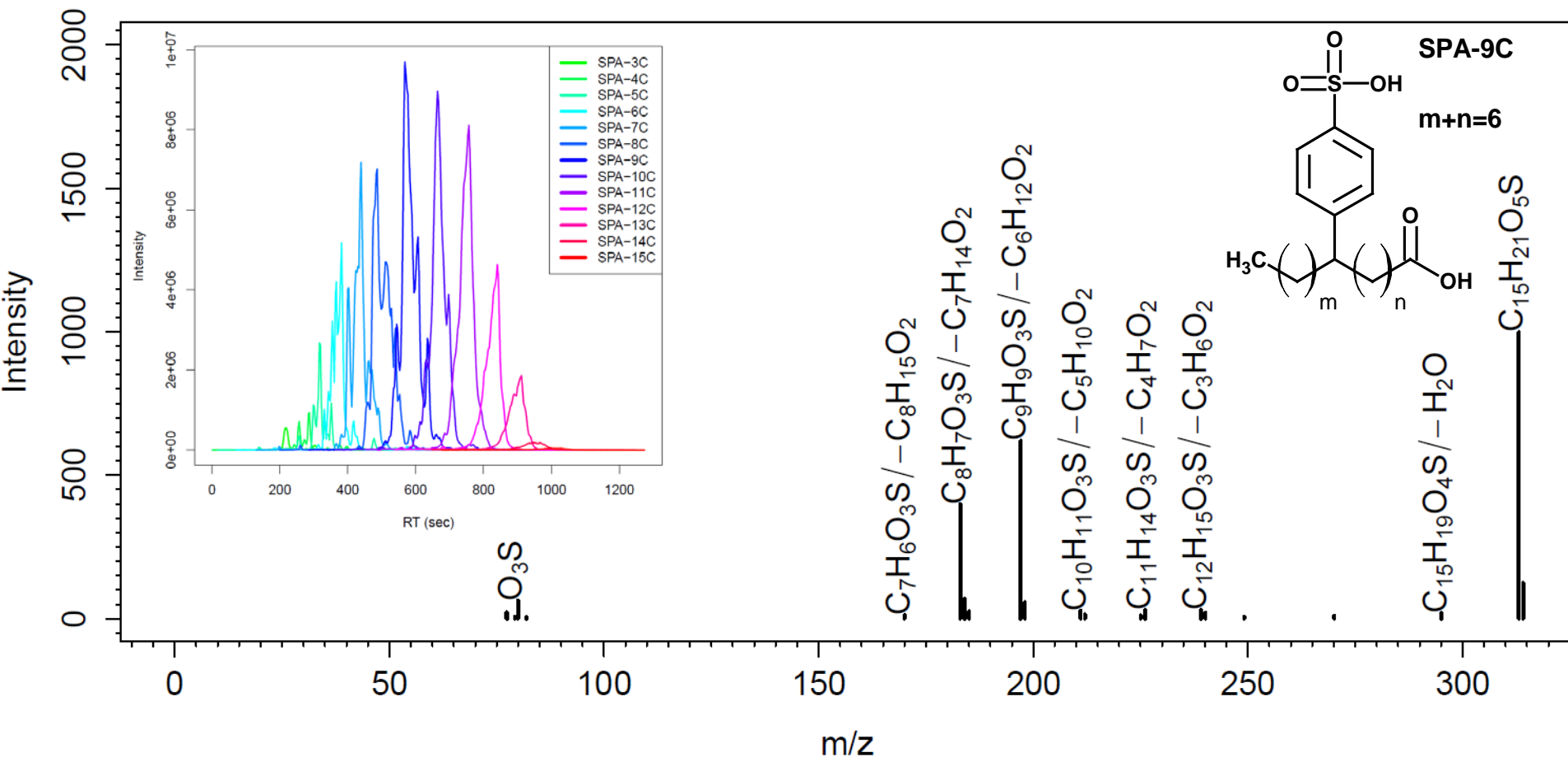
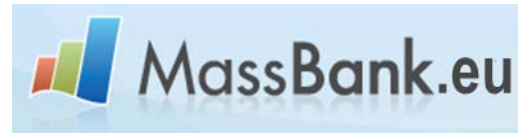
0.4 0.6 0.8 1.0

Annotated Spectra of Homologues in MassBank.EU

Tentatively Identified Spectra:

<http://goo.gl/0t7jGp>

RMassBank



www.massbank.eu

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

ACCESSIONS (LAS, SPACs):

Literature MS/MS

LIT00034, LIT00037

Std Mix., Sample

ETS00012, ETS00018

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>



[Back to main page](#)

[Back to status page](#)

[Collapse all](#)

[Download](#)

Continuous ID Search: MSV000078934 - GNPS_CAICE_CARB_C18_Aerosol_Headspace_Samples_NEGATIVE_POLARITY_Maxis_Impact_LCMS_

Hits 1 ~ 30 out of 3072

Go to

Go

Select columns

Filter	ClusterIndex	NumSpectra	PrecursorMZ	PrecursorInt	RTMean	DefaultGroups	LibraryID
Show Analogs 1	9752	8	311.77600	648800.00000	436	G1,	MassbankEU:ETS00014 C11-LAS (STANDARD MIX) C11-linear alkylbenzyl sulfonate 4-(undecan-5-yl)benzenesulfonic acid
Show Analogs 2	9776	76	311.16800	18631200.00000	505	G1,	MassbankEU:ETS00014 C11-LAS (STANDARD MIX) C11-linear alkylbenzyl sulfonate 4-(undecan-5-yl)benzenesulfonic acid
Show Analogs 3	1	26	159.12100	11030200.00000	483	G1,	
Show Analogs 4	4	7	173.11400	5976280.00000	140	G1,	

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

“Live” retrospective screening of known and **unknown** chemicals in European samples (various matrices)



*Future work: use results of **unknowns** to drive prioritization efforts*

EMBLAS-II: EU/UNDP project Improving Environmental Monitoring in the Black Sea
Non-target Screening Data

NTS Data

Joint Black Sea Surveys (JBSS)
year

2017

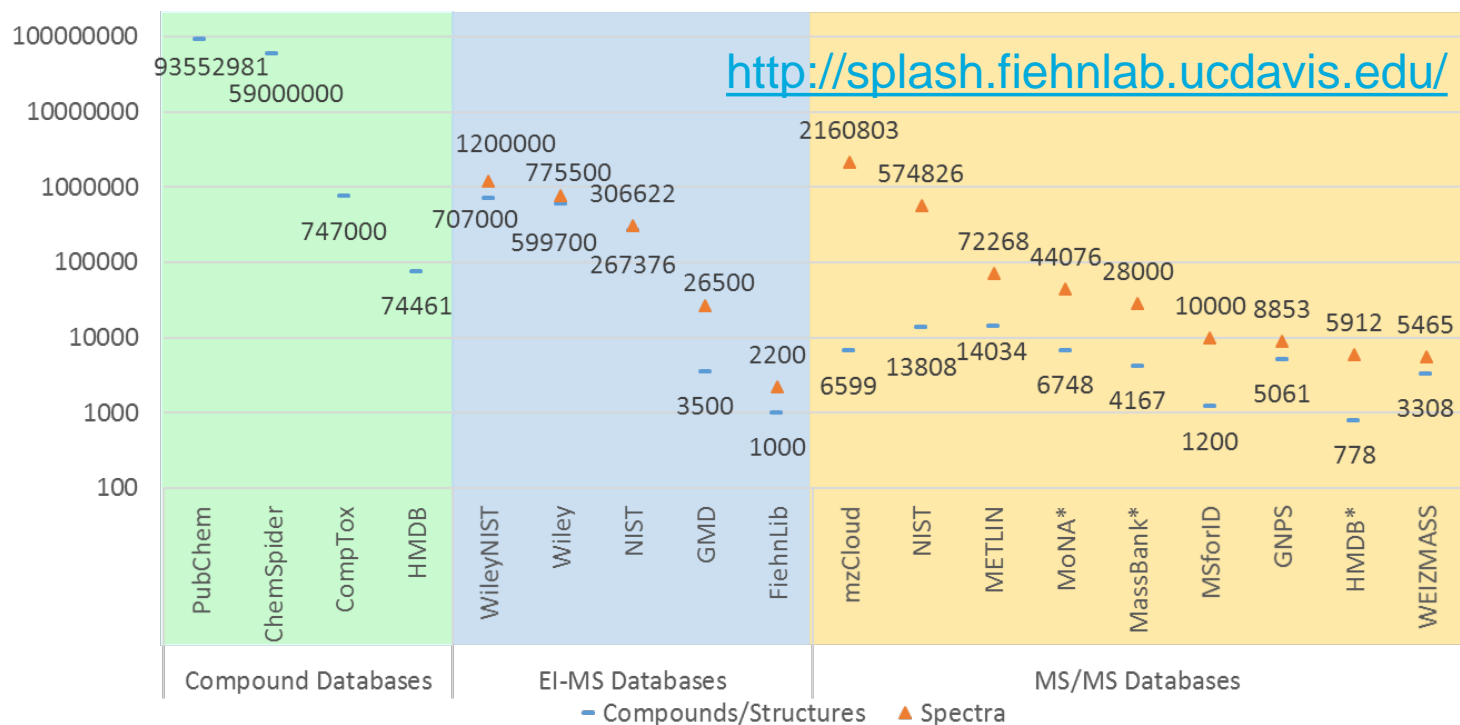
Choose compound

Positive ESI m/z=354.2876
rt=10.87



Small molecules ... big ~~problems~~ 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



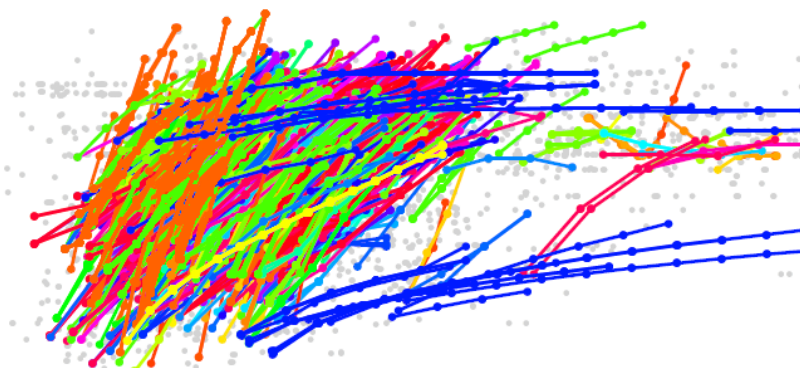
Small molecules ... ~~big problems~~ OPPORTUNITIES!

- 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
- Exchanging expert knowledge worldwide
 - Community efforts contribute greatly to improved cross-annotation



Small molecules ... big ~~problems~~ OPPORTUNITIES!

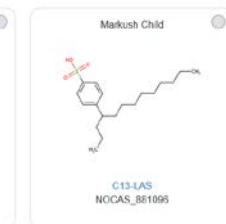
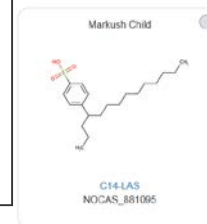
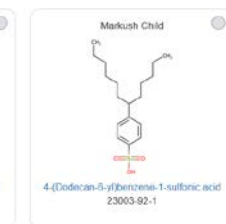
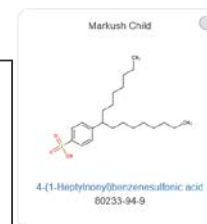
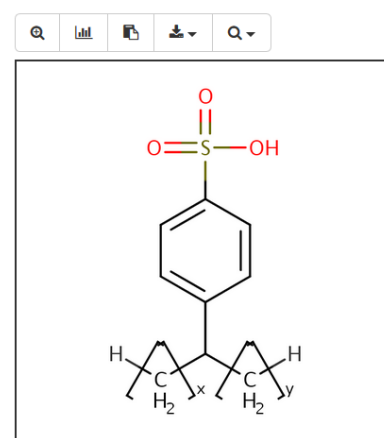
- Identifying small molecules requires information from many sources
 - Extensive compound databases available
 - Many mass spectral libraries available
 - Many excellent workflows available to collate this information
 - Community initiatives to improve communication between resources
- Exchanging expert knowledge worldwide
 - Community efforts contribute greatly to improved cross-annotation
- Tackling complex structures
 - Huge progress in cheminformatics approaches in very short time ...



Alkylbenzenesulfonate, linear

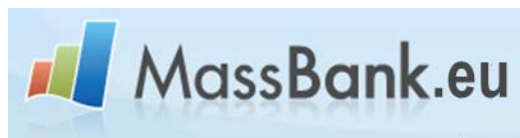
42615-29-2 | DTXSID3020041

© Searched by DSSTox_Substance_Id: Found 1 result for 'DTXSID3020041'.



Small molecules ... big ~~problems~~ OPPORTUNITIES!

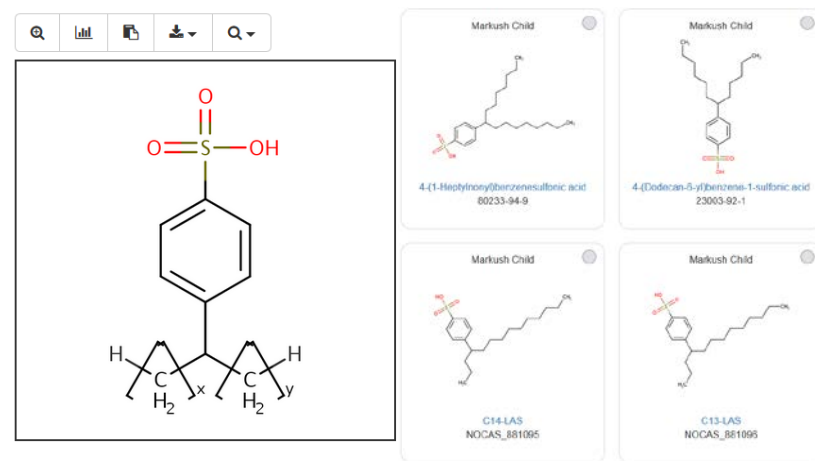
- Identifying small molecules requires information from many sources
 - Extensive compound databases available
 - Many mass spectral libraries available
 - Many excellent workflows available to collate this information
 - Community initiatives to improve communication between resources
- Exchanging expert knowledge worldwide
 - Community efforts contribute greatly to improved cross-annotation
- Tackling complex structures
 - Huge progress in cheminformatics approaches in very short time ...
 - Information in the public domain helps everyone!
(you never know when it will help you!)



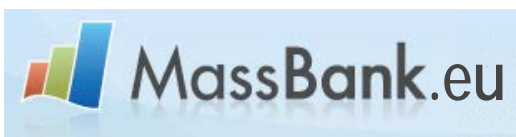
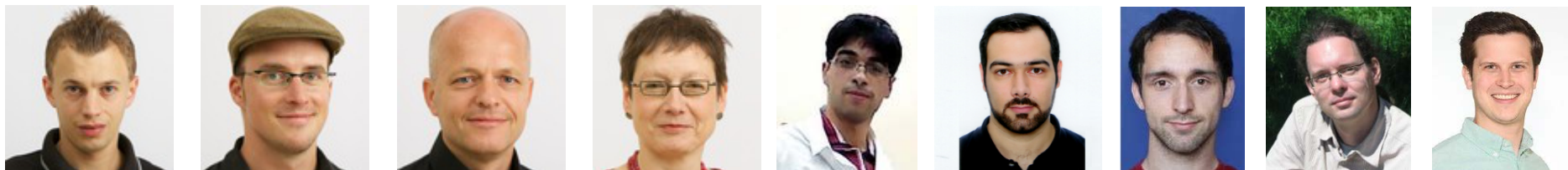
Alkylbenzenesulfonate, linear

42615-29-2 | DTXSID3020041

© Searched by DSSTox_Substance_Id: Found 1 result for 'DTXSID3020041'.



Acknowledgements I



emma.schymanski@uni.lu

Further Information:

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

<https://massbank.eu/MassBank/>

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

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

<https://github.com/MassBank/>



Community Efforts!

More opportunities: <https://tinyurl.com/lcsb-eci-jobs>



MassBank consortium



NORMAN Suspect List Exchange



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

NORMAN
Network of reference laboratories, research centres and related organisations


Home | Menu

Emerging
DATABASE
Topics ar
Worksho
QA/QC
Glossary

User login

Username
Password
Request n

Log in




Emma Louise Schymanski
added an **update**

NormaNEWS: retrospective screening of emerging contaminants

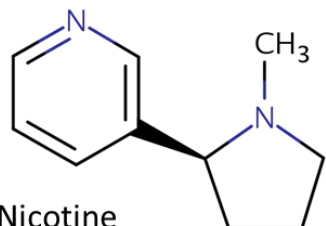
More news: one of our favourite examples, the NORMAN Network's pilot trial for global retrospective screening of emerging contaminants has just been accepted in ES&T - full list on the NORMAN Suspect List Exchange and the CompTox Dashboard.

<https://pubs.acs.org/doi/pdf/10.1021/acs.est.8b00365>

	Interactive Data table (updating...)		See interactive version . Compiled by Reza Aalizadeh, University of Athens, including RTI and toxicity values, support by Nikiforos Alygizakis, EI. <i>Work in progress ... please report any issues!</i>
MassBank	CSV, XLSX with Fragments (3/10/2017) CompTox MassBank EU Reference List CompTox MassBank EU Special Cases CompTox Fragment Download	MassBankEUIInChIKeys (11/04/2017)	www.massbank.eu Stravs <i>et al.</i> 2013. DOI: 10.1002/jms.3131
ST database of water-relevant substances	STOFF-IDENT Contents (6/09/2017) CompTox STOFF-IDENT List	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
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

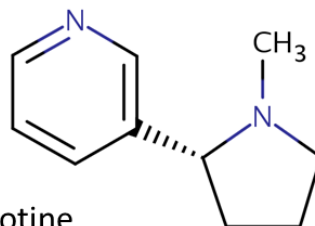
art of the
pect lists
contains
ctures as
pective
ank and
iewpoint

lowacka,



Nicotine

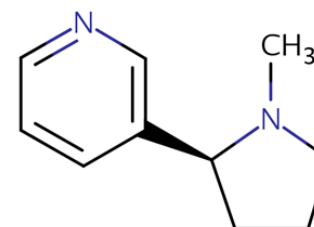
CN1CCC[C@H]1C1=CN=CC=C1
 DTXSID1020930 | SNICXCGAKADSCV
 54-11-5 | **162.1157** | 0.929 | **72**
 Tox: **yes** | Expo: **yes** | Bioassay: **yes**



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**

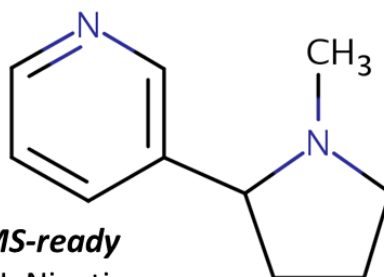
LEGEND: Name, SMILES
 DTXSID | InChIKey 1st Block
 CAS | **Monoiso.** Mass | logP | **Sources**
 Data on: **Toxicity** | **Exposure** | **Bioassays**



HCl

Nicotine hydrochloride

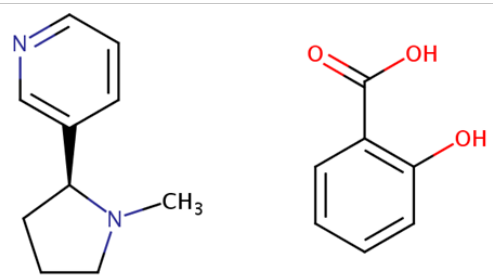
Cl.CN1CCC[C@H]1C1=CN=CC=C1
 DTXSID6020931 | HDJBTCAJIMNXEW
 2820-51-1 | **198.0924** | 0.929 | **9**
 Tox: **no** | Expo: **yes** | Bioassay: **yes**



MS-ready

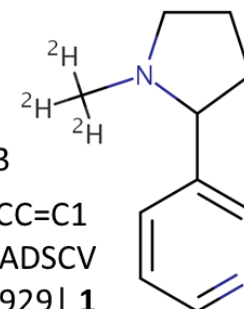
DL-Nicotine

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=CC(=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**

MetFrag2.3: Non-target Identification

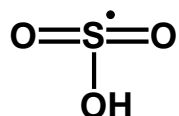
Status: 2010 => 2016

m/z $[M-H]^-$
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



Suspect Lists

?TOFF IDENT

Chemistry Dashboard

MetFrag2.3

MoNA

MassBank of North America

MassBank.eu

MS/MS

134.0054	339689
150.0001	77271
213.9607	632466

