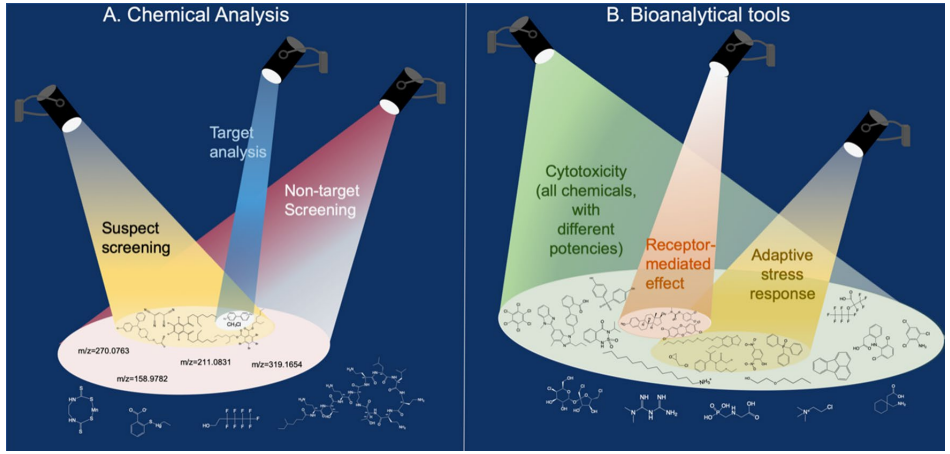
Luxembourg National Research Fund

INSTITUTE FOR ADVANCED STUDIES (IAS)

UNIVERSITÉ DU LUXEMBOURG

Cheminformatics and the Exposome in Health and Disease



Mod. from Escher *et al.* DOI: 10.1126/science.aay6636

Assoc. Prof. Dr. Emma L. Schymanski & Dagny Aurich

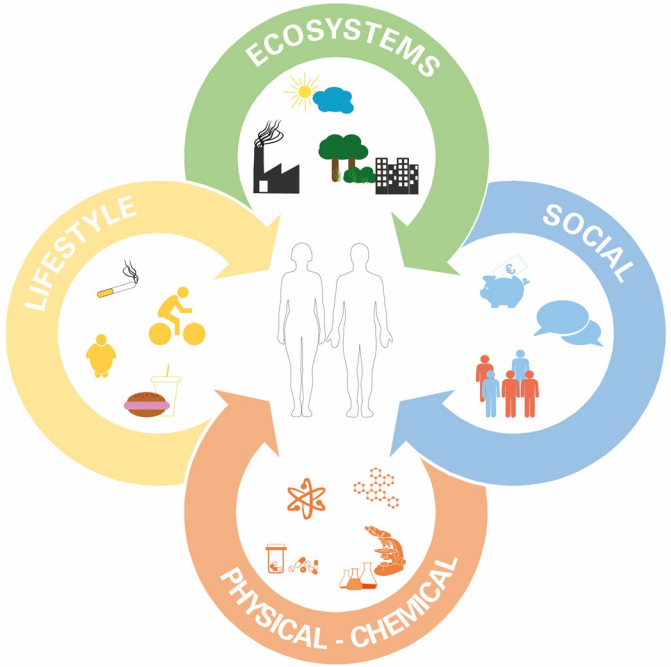
(plus many, many colleagues and collaborators!)

Environmental Cheminformatics Group,
Luxembourg Centre for Systems Biomedicine, University of Luxembourg
Email: emma.schymanski@uni.lu and @ESchymanski
Web: https://wwwen.uni.lu/lcsb/research/environmental_cheminformatics/

Second International Conference (Virtual Edition) on Network Medicine and Big Data, 12-13 April 2021

Slides available at DOI: 10.5281/zenodo.4678461

The Exposome



- Christopher Wild (2005)

  "All exposures from conception onwards, including those from lifestyle, diet and the environment"

- Miller and Jones (2014) functionalized this to:

  "The cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, diet, behaviour and endogenous processes"

- (Under)estimated 16 % (9 million) deaths per year worldwide due to environmental pollution alone (Landrigan *et al.* 2018)

Image c/o Lucie Debroux LCSB, modified from images by Utrecht Uni & Vermeulen *et al.* (2020). Science. DOI: 10.1126/science.aay3164. Wild (2005) DOI: 10.1158/1055-9965.EPI-05-0456; Miller & Jones (2014) DOI: 10.1093/toxsci/kft251; Landrigan *et al* (2018) DOI: 10.1016/S0140-6736(17)32345-0

Bring on Environmental Cheminformatics!

Adelene Lai
@adelene.lai
Given access 1 year ago

Emma Schymanski @emma.schymanski
Given access 1 year ago

Lorenzo Favilli
@lorenzo.favilli
Given access 1 year ago

Anjana Elapavalore
@anjana.elapavalore
Given access 1 year ago

German Andres Preciat Gonzales
@german.preciat
Given access 1 year ago

Mira Narayanan
@mira.narayanan
Given access 2 months ago

Begoña Talavera Andújar
@begona.talavera
Given access 4 months ago

Hiba Mohammed-Taha @hiba.mohammed-taha
Given access 4 months ago

Parviel Chirsir
@parviel.chirsir
Given access 1 month ago

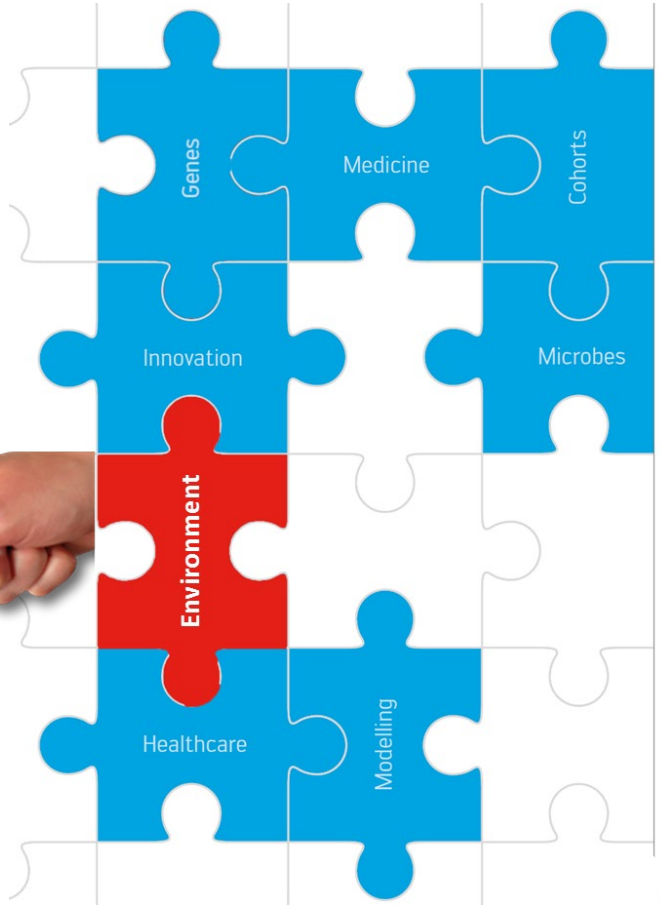
Corey Griffith
@corey.griffith
Given access 1 year ago

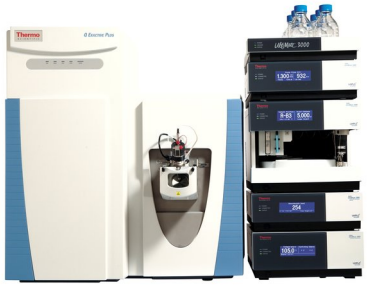
Jessy Krier @jessy.krier
Given access 1 year ago

Todor Kondic
@todor.kondic
Given access 1 year ago

Dagny Aurich
@dagny.aurich
Given access 3 months ago

R3 LCSB





Luxembourg National Research Fund

uni.lu | LCSB

… and International Collaborations!

norman

KEMI Kemikalieinspektionen

INRA SCIENCE & IMPACT

MassBank High Quality Mass Spectral Database

UFZ

NIH U.S. National Library of Medicine National Center for Biotechnology Information

PubChem

776 Data Sources

Explore Data Sources

eawag aquatic research

EPA

Luxembourg National Research Fund

UNIVERSITY of WASHINGTON

UNIVERSITY OF AMSTERDAM

ENVIRONMENTAL INSTITUTE

ontochem

THE GOVERNMENT OF THE GRAND DUCHY OF LUXEMBOURG
Ministry of the Environment, Climate and Sustainable Development

Kanton Zürich
Baudirektion
Amt für Abfall, Wasser, Energie und Luft
Gewässerschutz

UNIVERSITAT JAUME I

UNIVERSITAT ROVIRA i VIRGILI

Research Institute for Pesticides and Water · IUPA

solutions

IPB

CDK

R3

ETH zürich

C²DH

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

COLUMBIA MAILMAN SCHOOL OF PUBLIC HEALTH

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

WIKIDATA

Stockholm University

elixir LUXEMBOURG

National and Kapodistrian UNIVERSITY OF ATHENS

UNIVERSITY OF COPENHAGEN

LIST

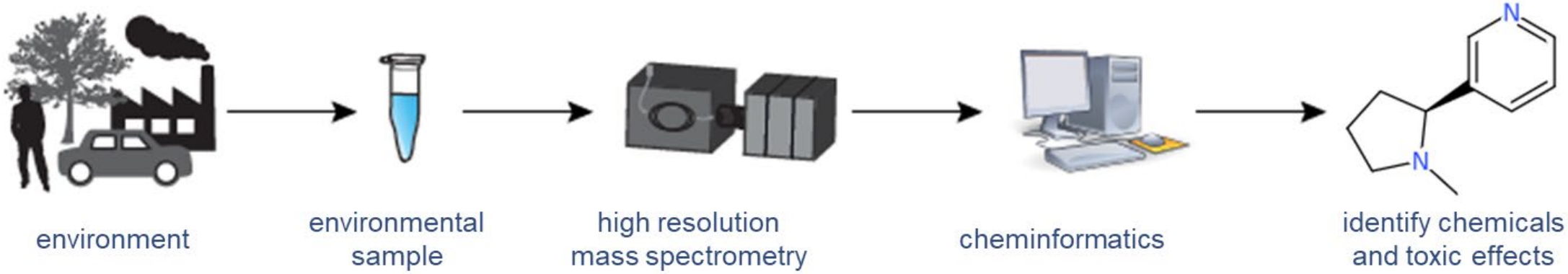
Maastricht University

uni.lu

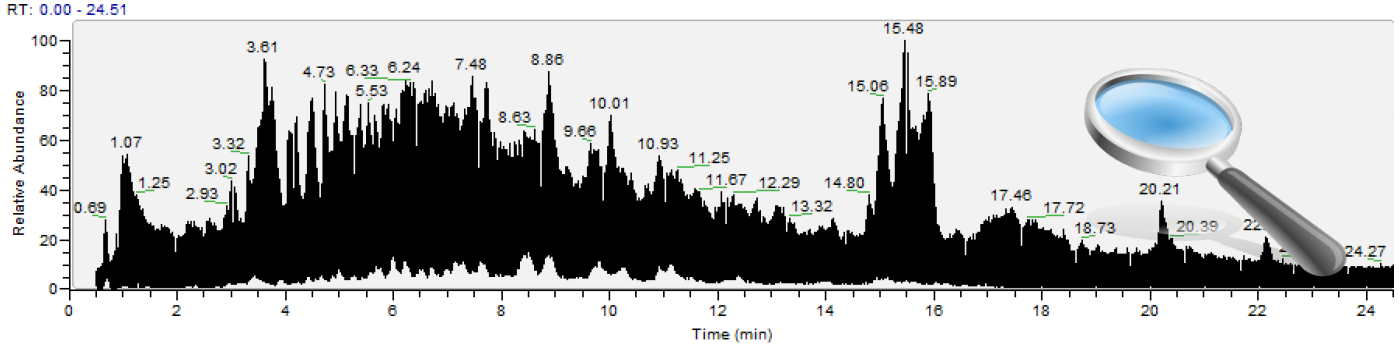
LCSB

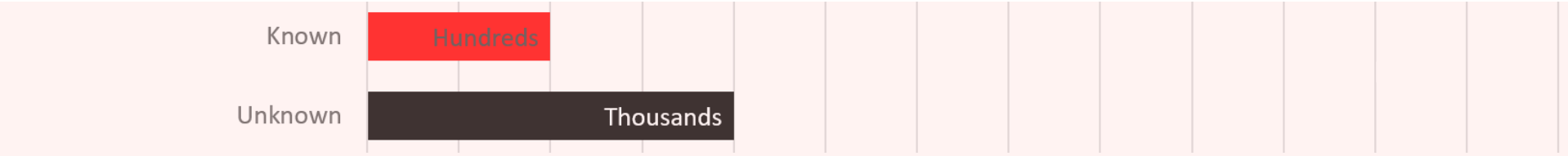
Source: https://en.wikipedia.org/wiki/File:EU-Luxembourg.svg

What is Environmental Cheminformatics?



High resolution mass spectrometry



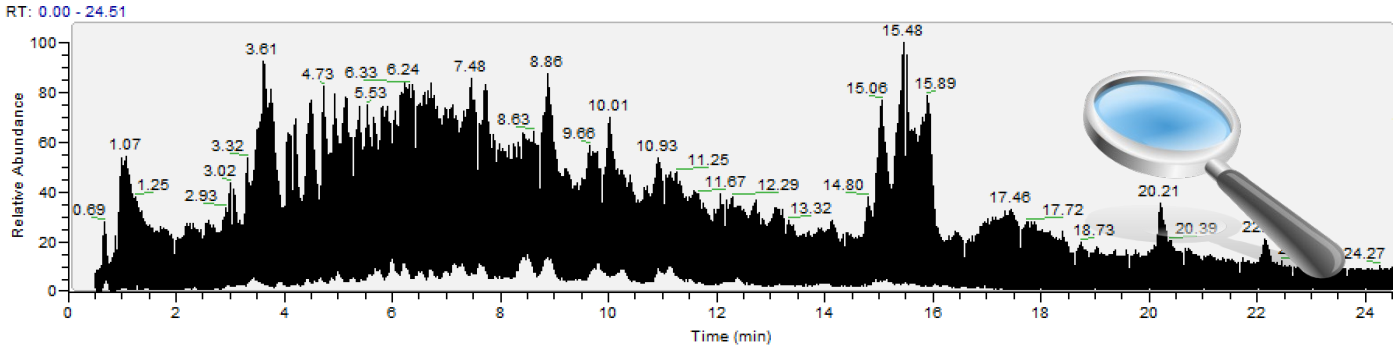


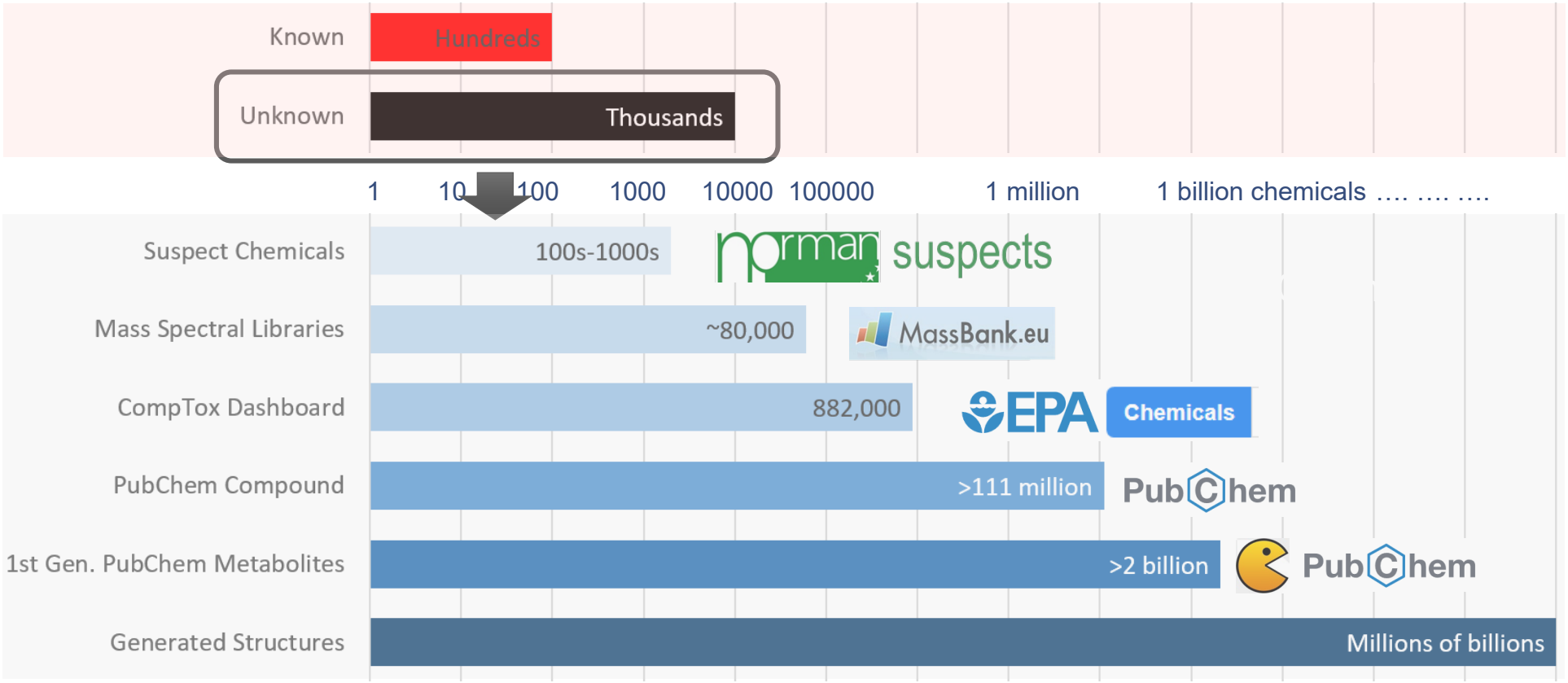
1 10 100 1000 10000 100000 1 million 1 billion chemicals …. …. ….

What is Environmental Cheminformatics?

High resolution mass spectrometry

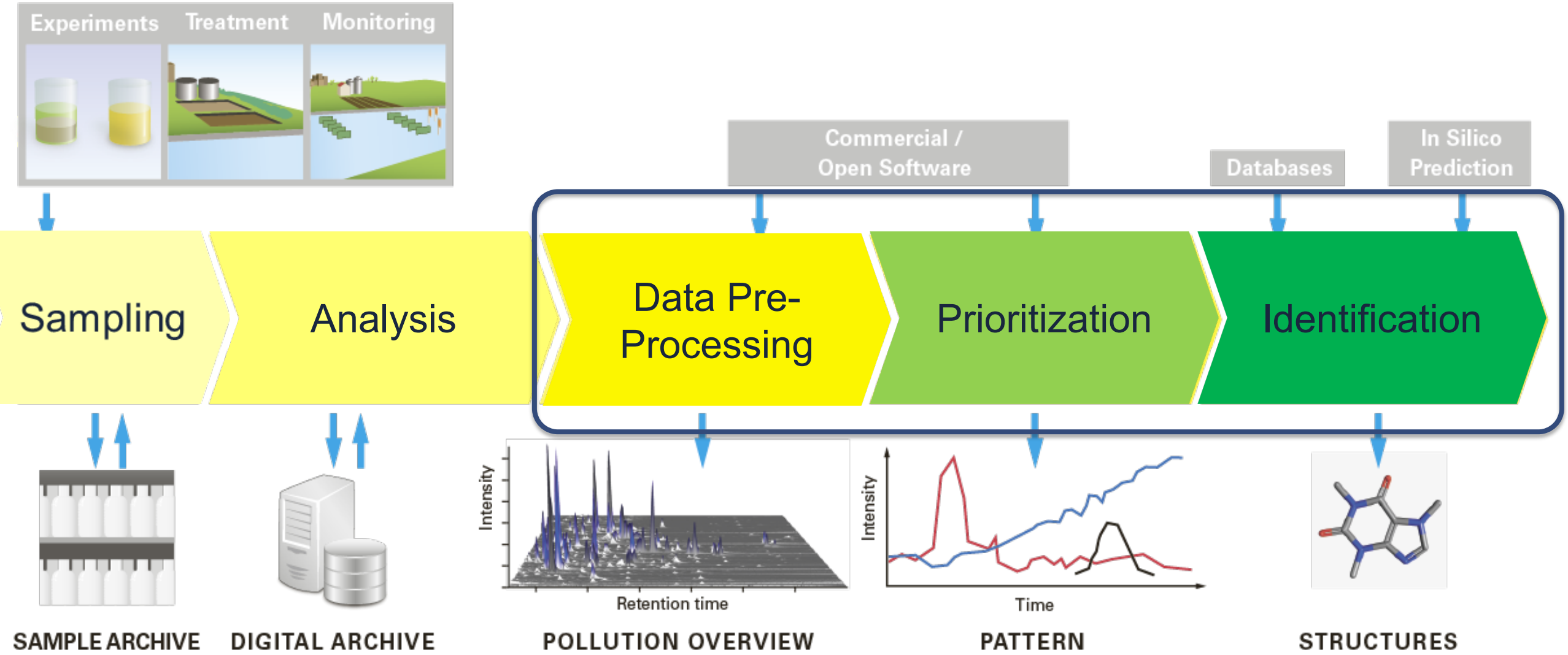
AND connecting chemical knowledge





Schymanski *et al*. (2014) DOI: 10.1021/es4044374; Vermeulen *et al.* (2020) DOI: 10.1126/science.aay3164

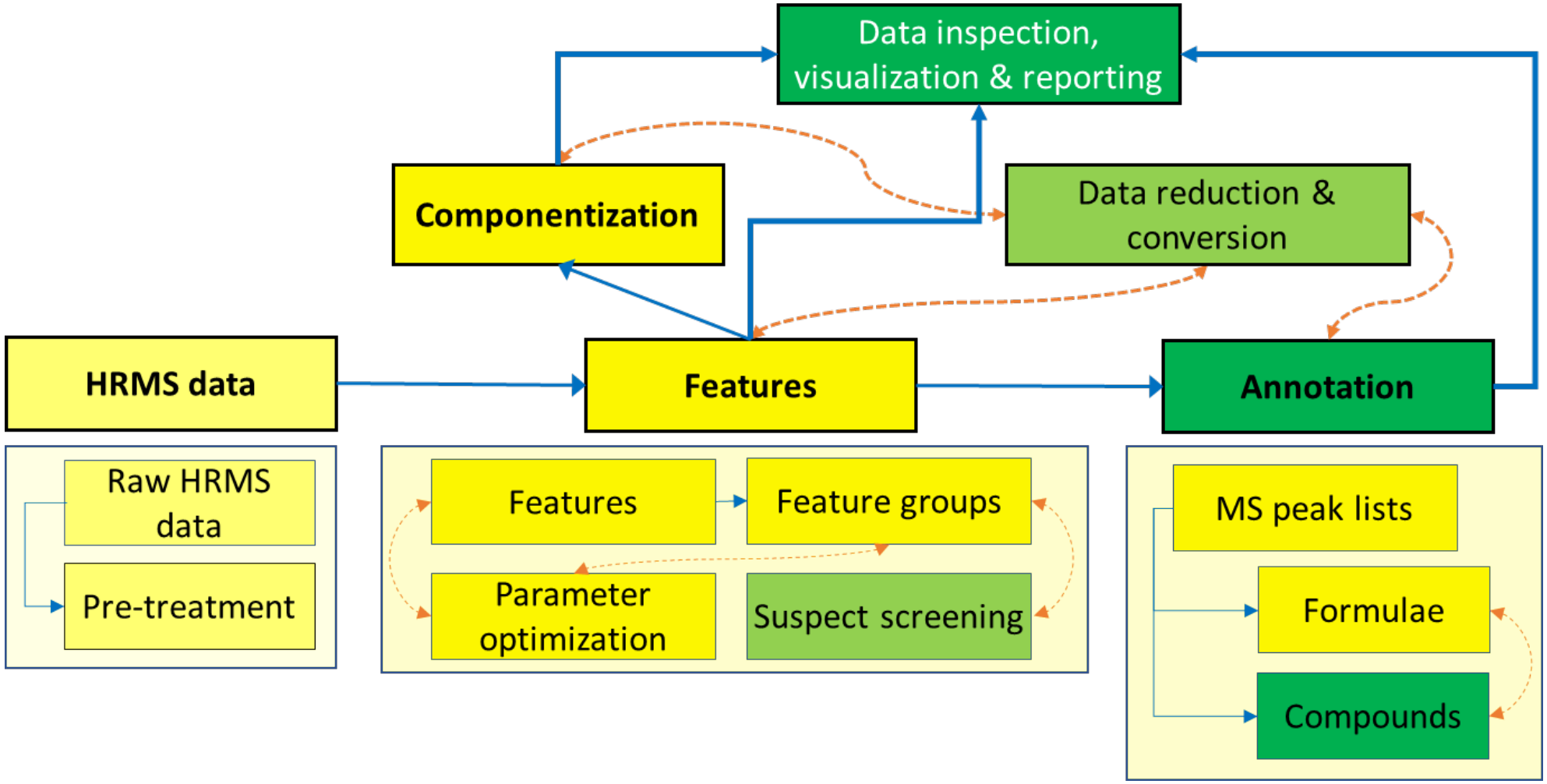
Non-target High Resolution Mass Spectrometry (NT-HRMS)



Modified from: Hollender, Schymanski *et al*. 2018, ES&T Feature, 51:20, 11505-11512. DOI: 10.1021/acs.est.7b02184

Open Source Workflows for NT-HRMS: patRoon

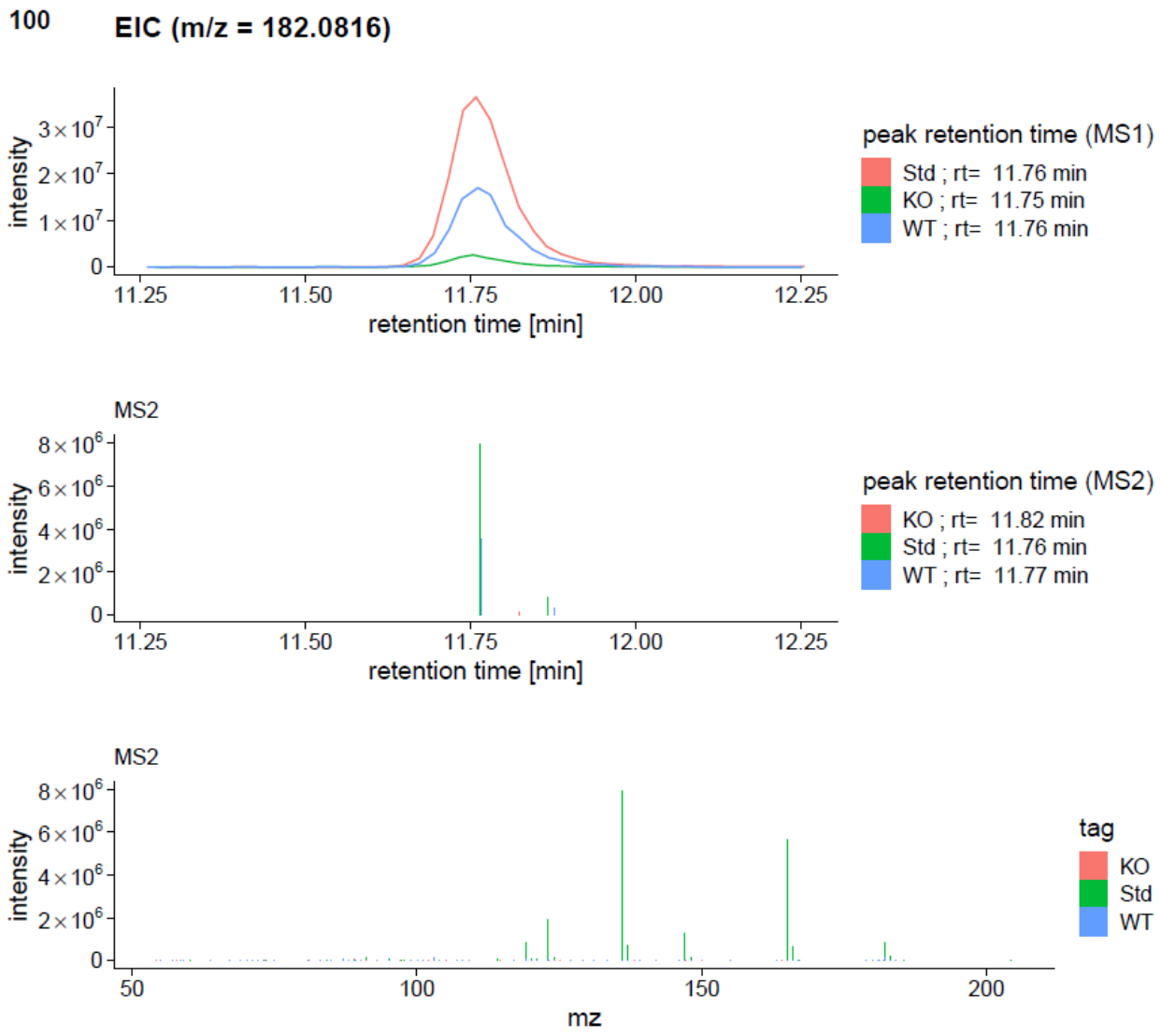
https://rickhelmus.github.io/patRoon/



Modified from: Helmus *et al* (2021), JCheminf. 13:1. DOI: 10.1186/s13321-020-00477-w

UNIVERSITY OF AMSTERDAM

Open Source Workflows for NT-HRMS: Shinyscreen



https://git-r3lab.uni.lu/eci/shinyscreen

Anjana Elapavalore, Mira Narayanan, Todor Kondic, Jessy Krier, Hiba Mohammed Taha.

Mass Spectral Libraries: MassBank (Open Source & Data!)

https://massbank.eu/MassBank/ https://github.com/MassBank/MassBank-data/

MassBank Search Contents Download Accession Go More▼

MassBank Europe

MassBank
High Quality Mass Spectral Database

>> Search Spectra

>> Learn More

Search for:

Mass calculator

Basic Search | Peak List | Peaks | Peak Differences | InChIKey | SPLASH

Compound Information

Compound name

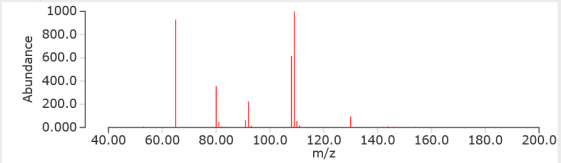
AND Exact Mass Tolerance 0.3

AND Formula (e.g. C6H7N5, C5H*N5, C5*) Search

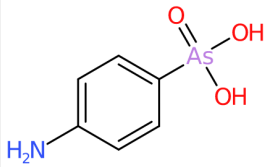
MassBank Record: LU040605

(4-Aminophenyl)arsonic acid; LC-ESI-QFT; MS2; CE: 75; R=17500; [M+H]+

Mass Spectrum



Chemical Structure



Expert Knowledge: NORMAN Database System

https://www.norman-network.com/nds/

NORMAN Database System

NORMAN

NORMAN WEBSITE | NORMAN DATABASE SYSTEM | HOME | LOGIN

NORMAN SUBSTANCE DATABASE

SEAR

Searchi

Note: C

NORMAN Suspect List Exchange – NORMAN SLE

The NORMAN Suspect List Exchange (NORMAN-SLE) was established in 2015 as a central access point for NORMAN members (and others) to find suspect lists relevant for their environmental monitoring question. This Exchange documents all individual collections that form a part of NORMAN SusDat, the merged NORMAN Substance Database (DOI: 10.5281/zenodo.2664077).

Substance Database

A merged list of NORMAN substances; Central Database to access various lists of substances for suspect screening and prioritisation

Chemical Occurrence Data

A database of geo-referenced monitoring data on emerging substances

Ecotoxicology

A platform for systematic collection and evaluation of ecotoxicity studies for harmonised derivation of environmental quality standards

Suspect List Exchange

Central Database to access various lists of substances for suspect screening and prioritisation

Antibiotic Resistance Bacteria/Genes

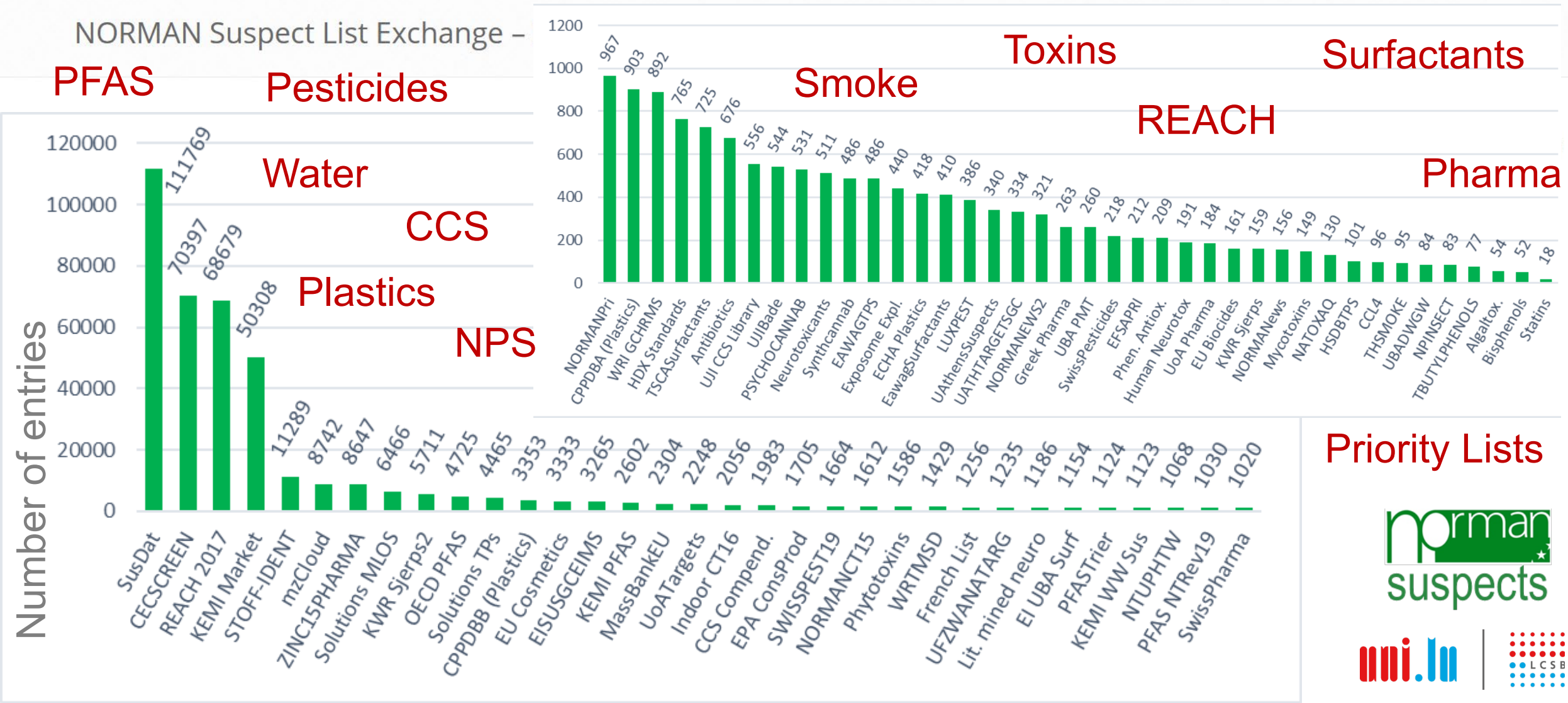
A database of ARBs/ARGs in environmental matrices

MassBank Europe

A database of mass spectra of emerging substances to support identification of unknown substances

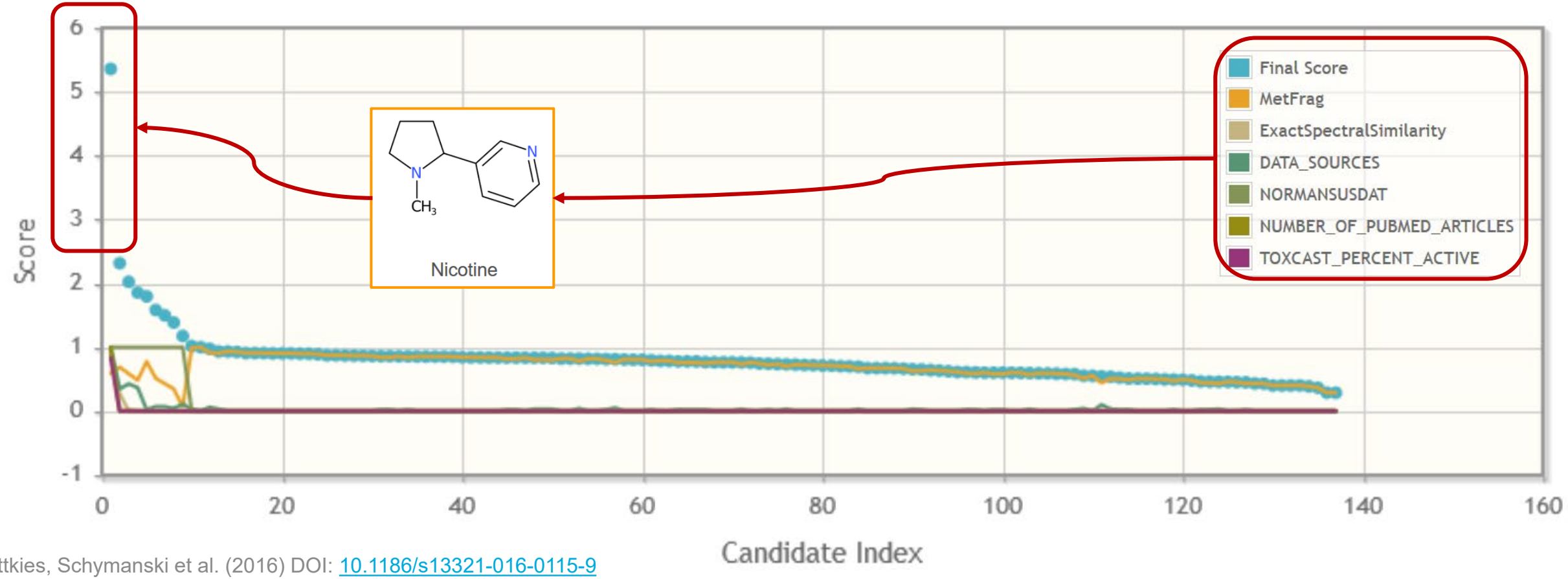
Expert Knowledge: NORMAN Suspect List Exchange [CC-BY-4.0]

https://www.norman-network.com/nds/SLE/ https://zenodo.org/communities/norman-sle



Connecting Knowledge for Chemical Identification: MetFrag

https://msbi.ipb-halle.de/MetFrag/ http://ipb-halle.github.io/MetFrag/



Ruttkies, Schymanski et al. (2016) DOI: 10.1186/s13321-016-0115-9

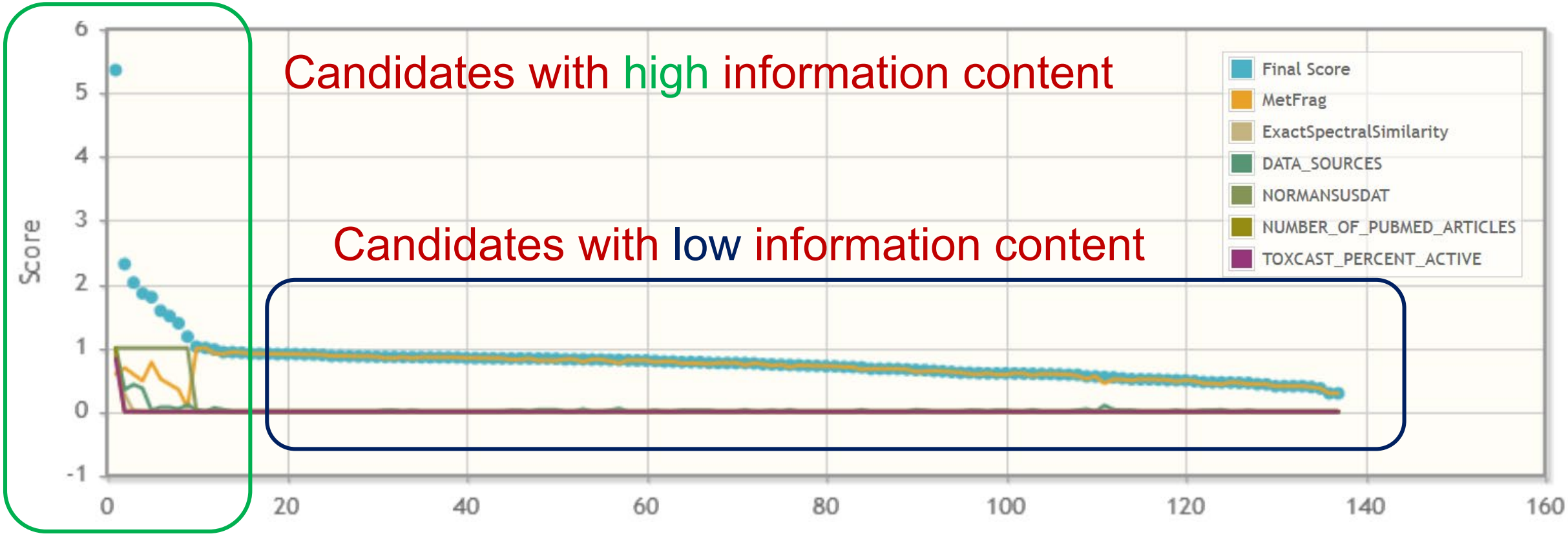
Connecting Knowledge for Chemical Identification: MetFrag

Problem: Exposomics “Chemical Space” is too big!

CAS A DIVISION OF THE AMERICAN CHEMICAL SOCIETY 180 million

PubChem 110 million

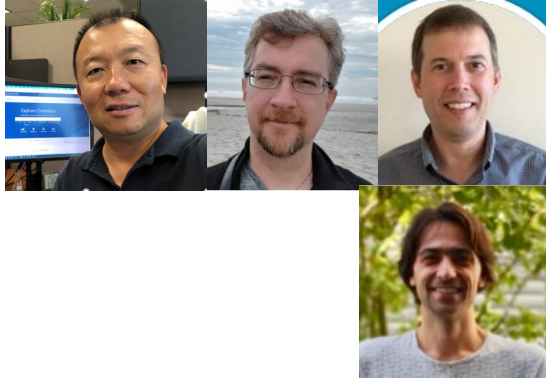
ChemSpider Search and share chemistry 103 million



Ruttkies, Schymanski et al. (2016) DOI: 10.1186/s13321-016-0115-9

Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9

DIY Databases: PubChemLite for Exposomics



- PubChem Compound TOC ? 49,493,641
  - Agrochemical Information ? 3,045
  - Associated Disorders and Diseases ? 20,847
  - Biologic Description ? 2,056,521
  - Biological Test Results ? 3,622,920
  - Biomolecular Interactions and Pathways ? 125,253
  - Chemical and Physical Properties ? 263,015
  - Classification ? 1,454,824
  - Drug and Medication Information ? 17,922
  - Food Additives and Ingredients ? 8,414
  - Identification ? 4,968
  - Information Sources ? 20,271,277
  - Literature ? 1,833,941
  - Names and Identifiers ? 1,275,170
  - Patents ? 36,351,418
  - Pharmacology and Biochemistry ? 110,628
  - Related Records ? 9,224,590
  - Safety and Hazards ? 149,319
  - Spectral Information ? 480,730
  - Structures ? 9,117,635
  - Toxicity ? 114,012
  - Use and Manufacturing ? 115,321
- Chemical Safety ? 147,023

110 million … OR …
the most relevant / annotated?

PubChemLite for Exposomics
371,663 entries

To find out more:
https://www.nordicmetsoc.org/MetExposome2021.html

zenodo | Search | Upload | Communities | emma.schymanski@uni.lu

October 31, 2020 — Dataset — Open Access

PubChemLite for Exposomics

Bolton, Evan; Schymanski, Emma; Kondic, Todor; Thiessen, Paul; Zhang, Jeff

PubChemLite is a subset of PubChem (https://pubchem.ncbi.nlm.nih.gov/) selected from major
ontents page at the PubChem Classification Browser
ih.gov/classification/#hid=72). With this release, there is now just one
is the former tier1 plus two new categories (Associated Disorders &
:

is 371,663 compounds (31 Oct 2020) compiled from 10 categories:
DrugMedicInfo, FoodRelated, PharmacoInfo, SafetyInfo, ToxicityInfo,
, Identification.

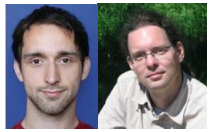
ollapsed by InChIKey first block, reporting the structure from the most
CIDs. Entries that will be ignored by MetFrag (salts, disconnected
(e.g. transition metals) have been removed. The Patent and PubMed ID
les on the PubChem FTP site. The "AnnoTypeCount" term counts how
epresented, the subsequent column (named per category) counts the
ories available in the next sub-category of the TOC entry.

Edit
New version

Communities
LCSB Environmental Cheminformatics Group — Remove

1,084 views — 1,170 downloads
See more details...

MetFrag



Database Settings

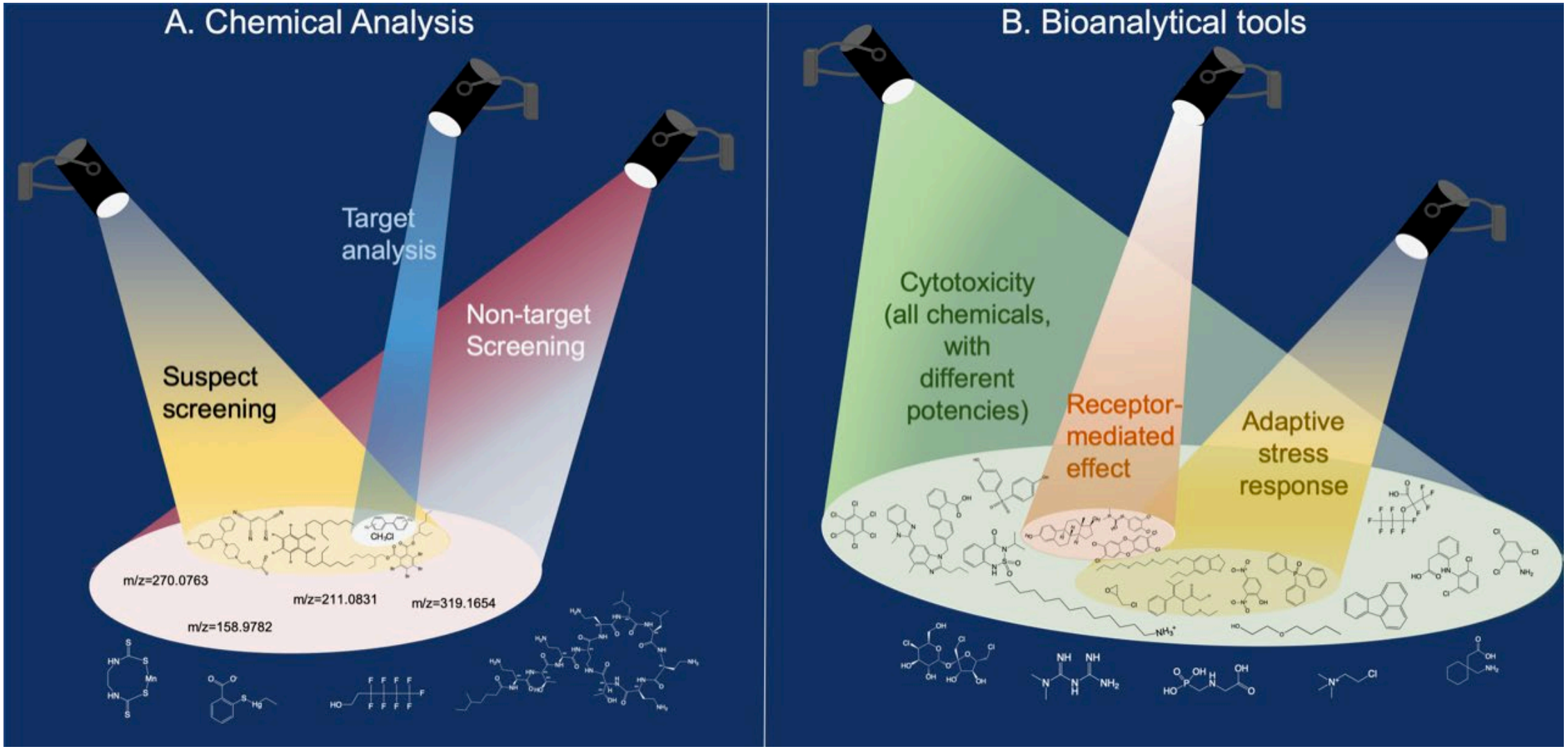
Database: PubChemLite_31Oct2020_
Neutral Mass: 229.10948 Search ppm: 5
Formula: C9H16ClN5
Identifiers:

Retrieve Candidates — 4 Candidates

Schymanski *et al.* (2021) DOI: 10.1186/s13321-016-0115-9

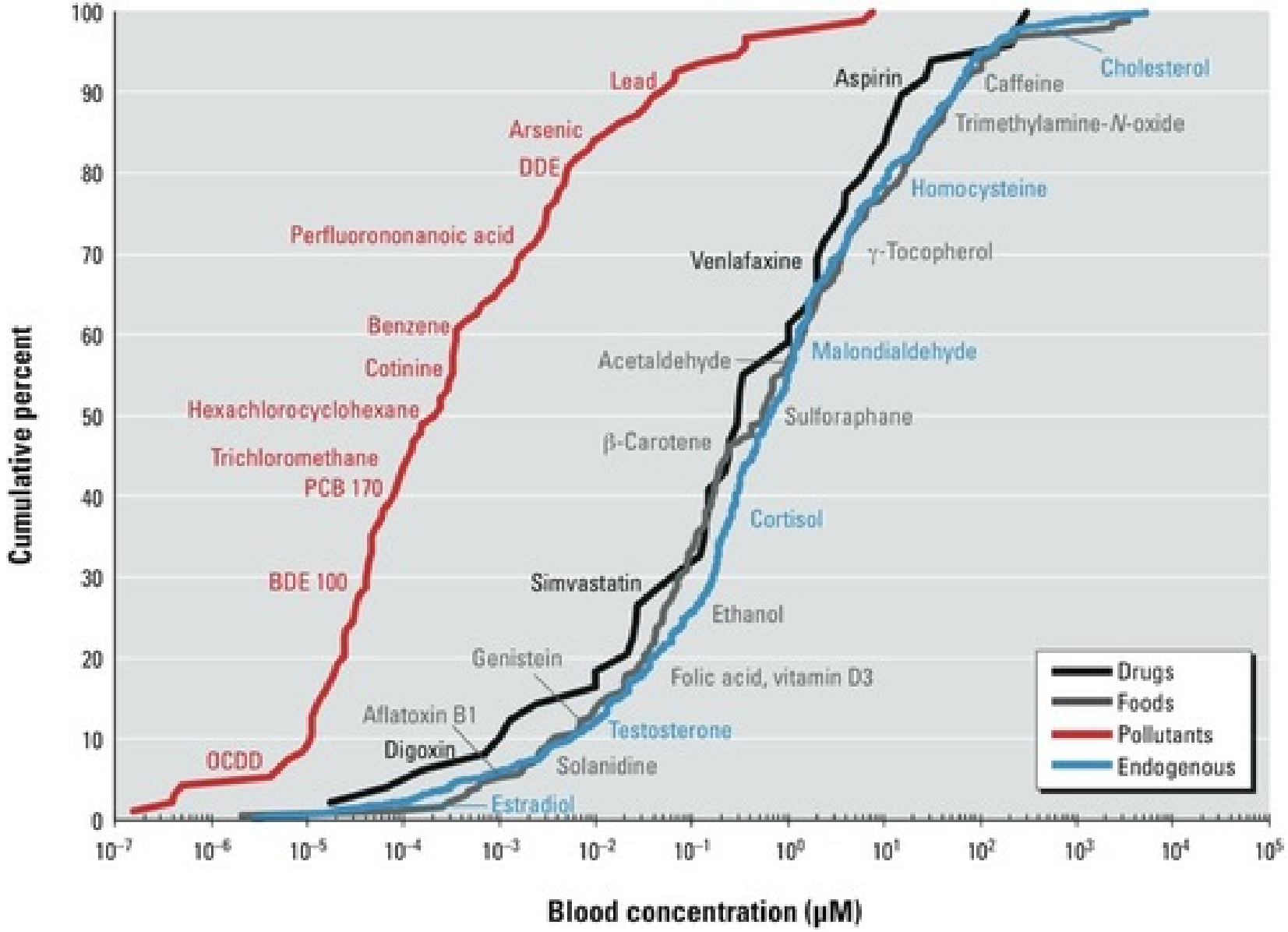
NIH U.S. National Library of Medicine National Center for Biotechnology Information

Complex Mixtures – Connecting Chemicals and Effects



Modified from Escher, Stapleton, Schymanski (2020). Science. DOI: 10.1126/science.aay6636

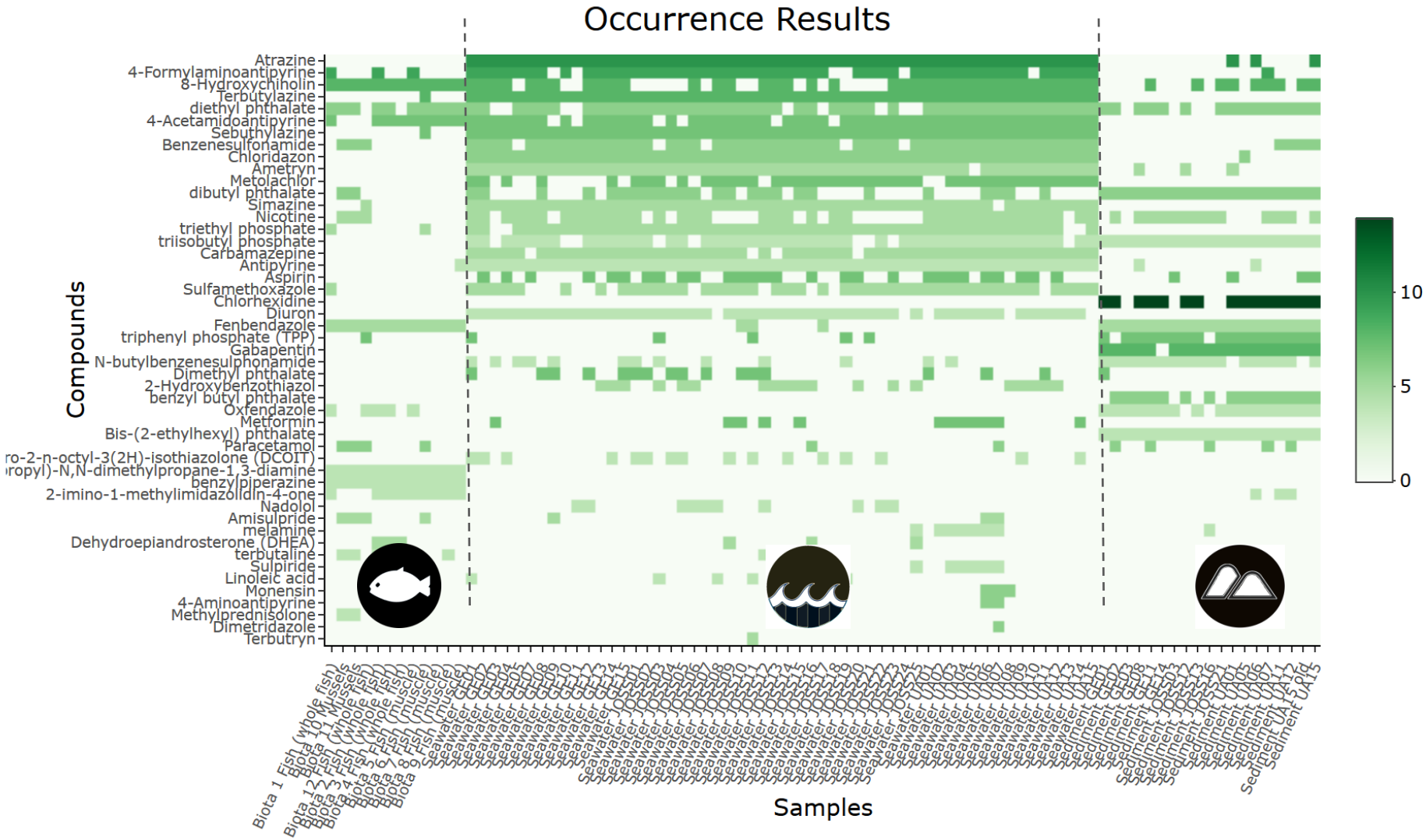
“The concentration challenge”



Rappaport *et al* (2014) *Environmental Health Perspectives*. DOI: 10.1289/ehp.1308015

Examples: Distribution of Chemicals in Various Matrices

Retrospective screening of REACH chemicals in Black Sea samples



Alygizakis NA *et al* (2019) DOI: 10.1016/j.trac.2019.04.008

Interactive heat map at http://norman-data.eu/NORMAN-REACH

Examples: NORMAN Digital Sample Freezing Platform

"Live" retrospective screening of known and unknown chemicals in European samples

www.norman-data.eu

NORMAN Digital Sample Freezing Platform | Main Page | Batch mode

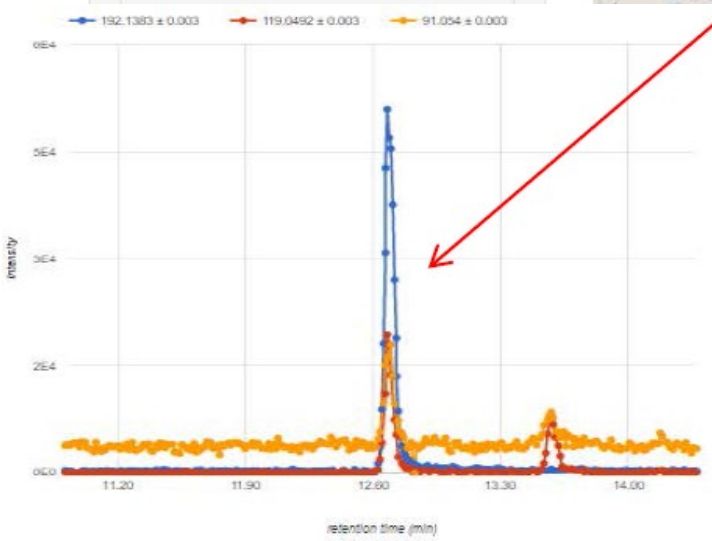
norman suspects

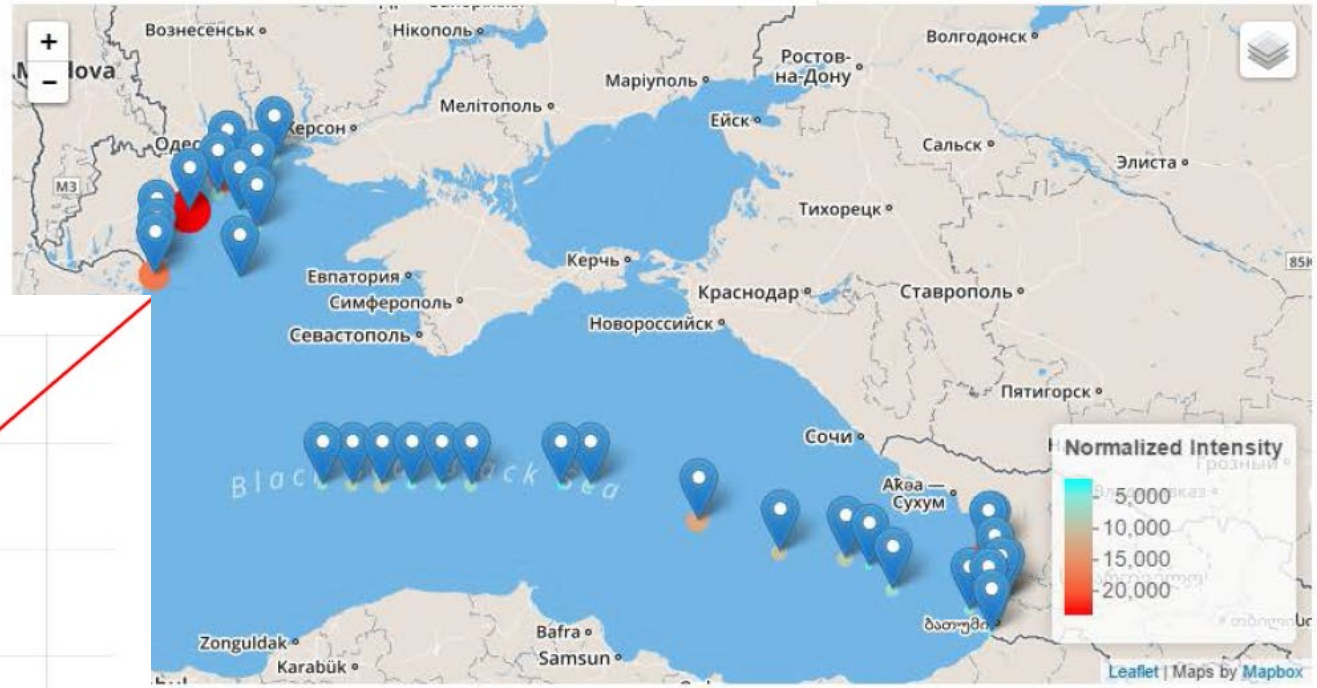
Choose Emerging Substance or input mass of interest and experimental RTI

Substance name or CAS or StdInChIKey

DEET [134-62-3] [MMOXZBCLCQITDF-UHFFFAOYSA-N]

Contributed Samples | Results | Chromatograms | Interactive Map | Help



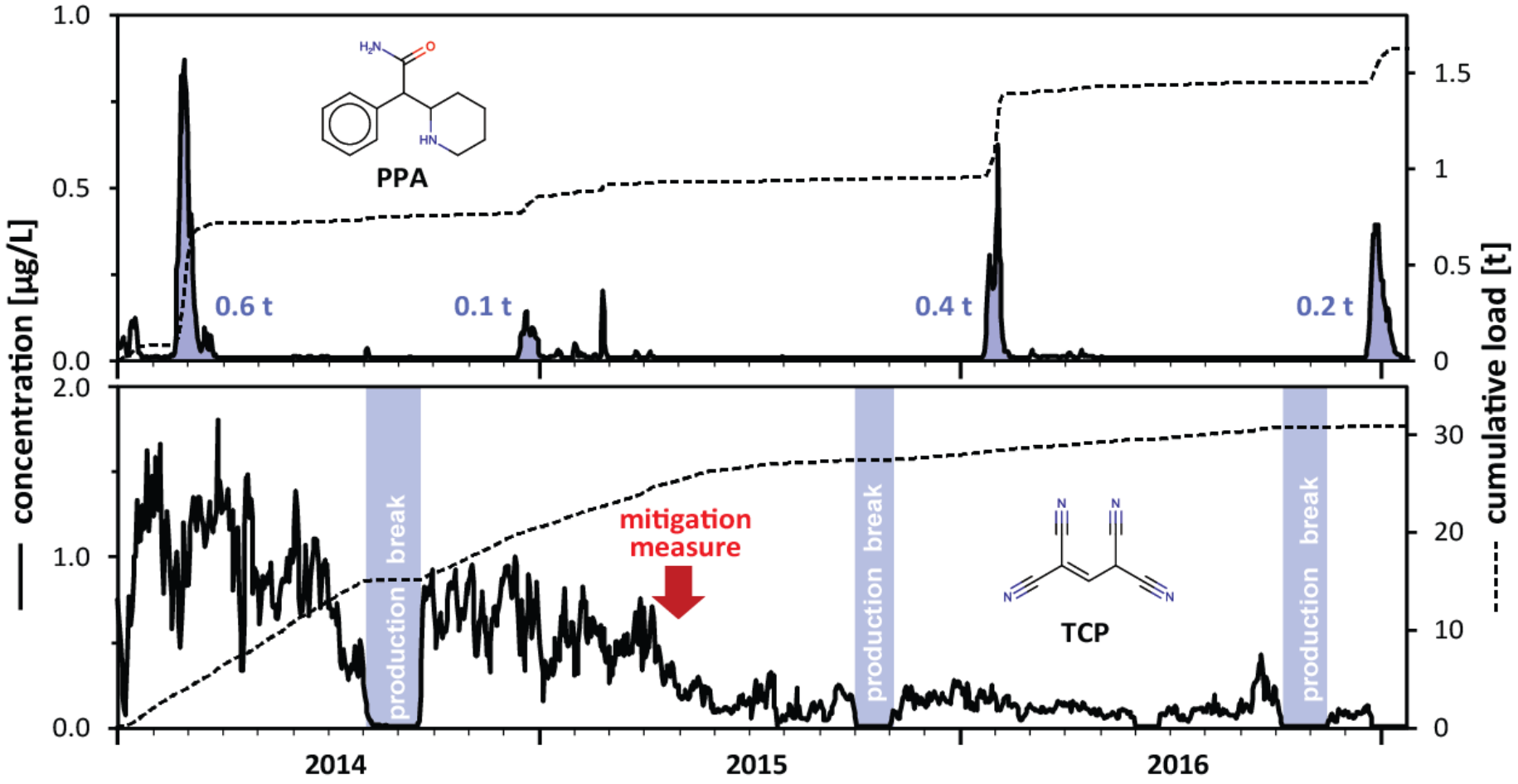


Alygizakis NA *et al* (2019) DOI: 10.1016/j.trac.2019.04.008

NORMAN-DSFP available at http://norman-data.eu/

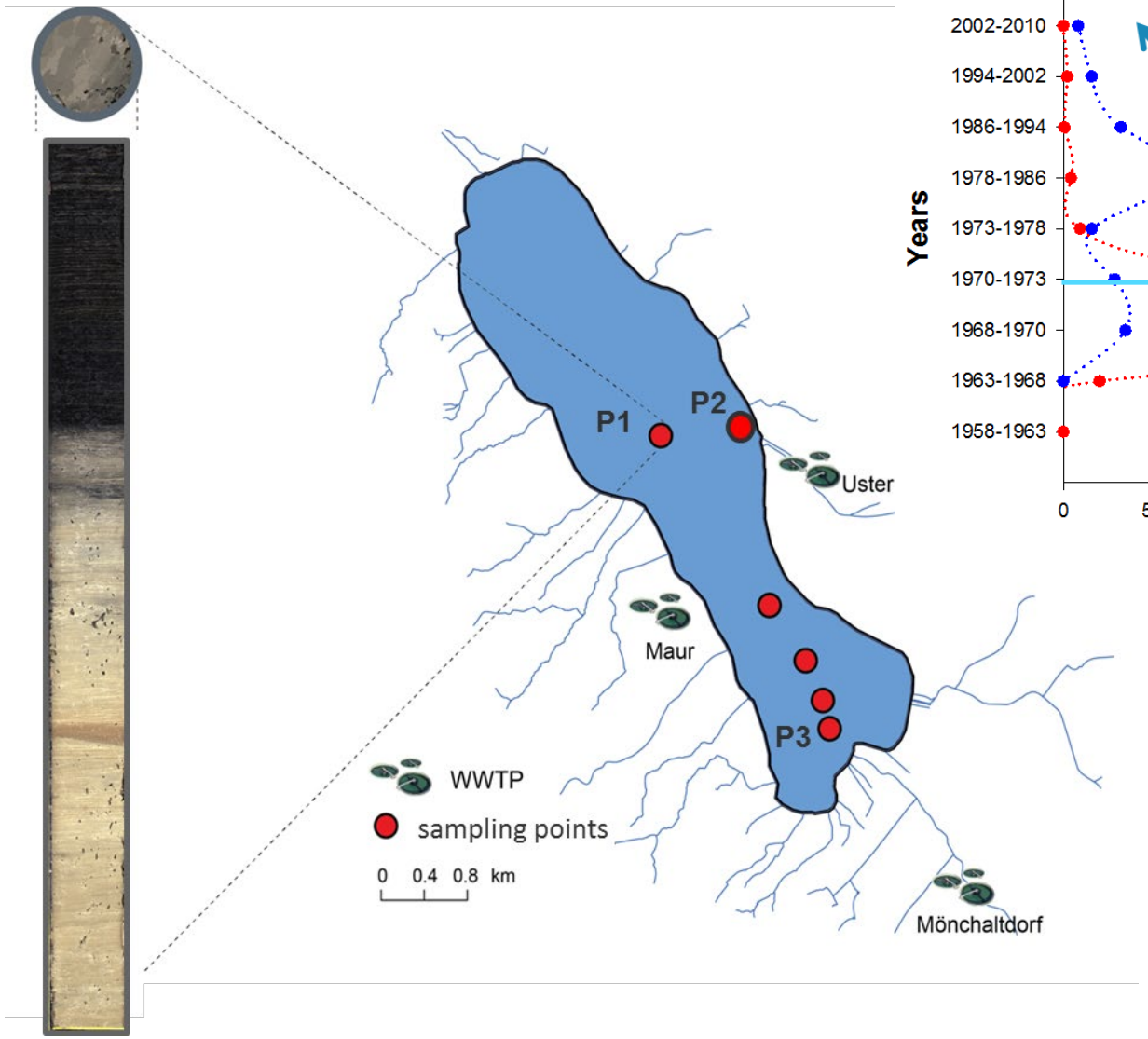
Examples: Real Time Monitoring of the Rhine River

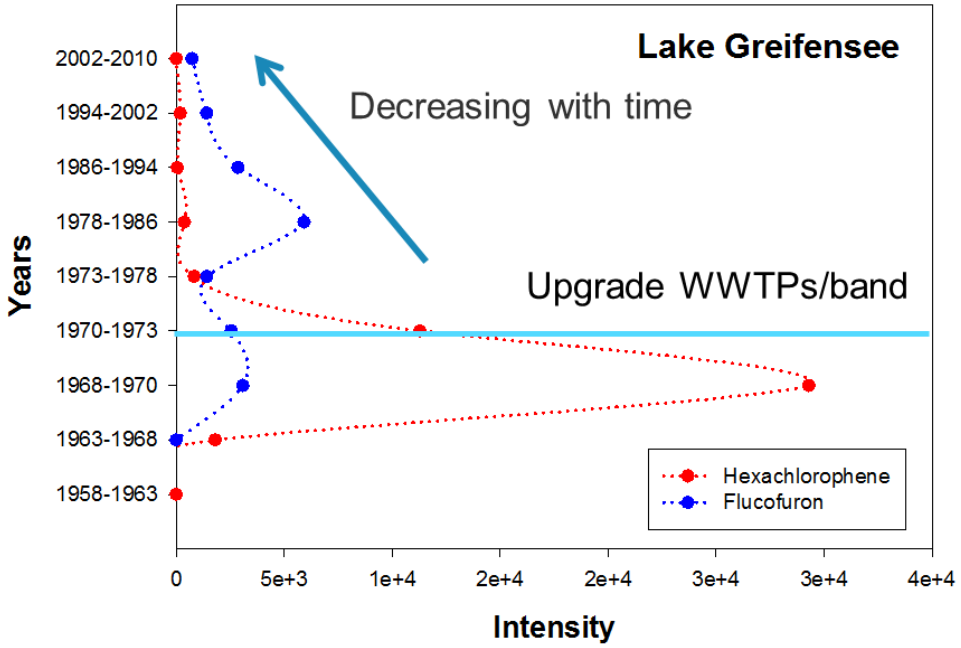
Previously unknown chemicals detected due to “stand-out” patterns



Hollender, Schymanski, Singer & Ferguson (2018) ES&T Feature, 51:20, 11505-11512. DOI: 10.1021/acs.est.7b02184

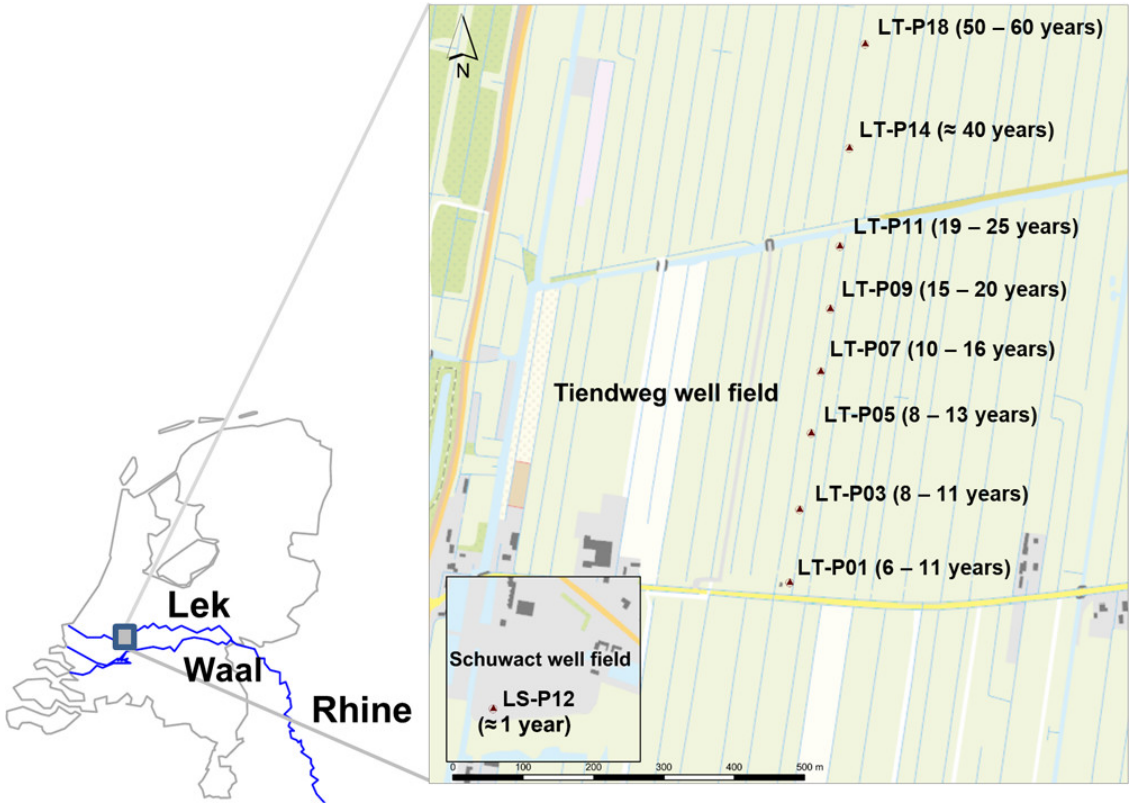
Examples: Historical Contamination in Lake Sediments

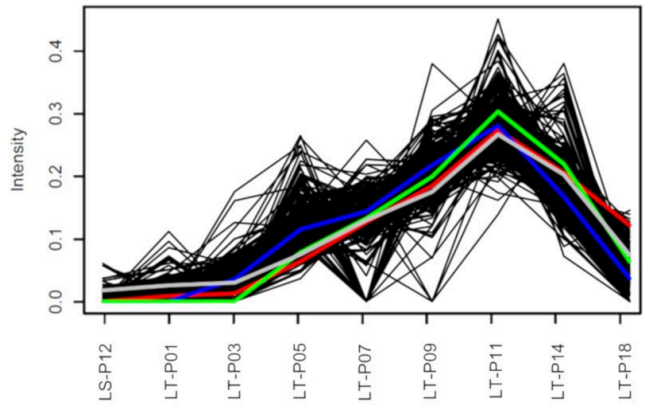


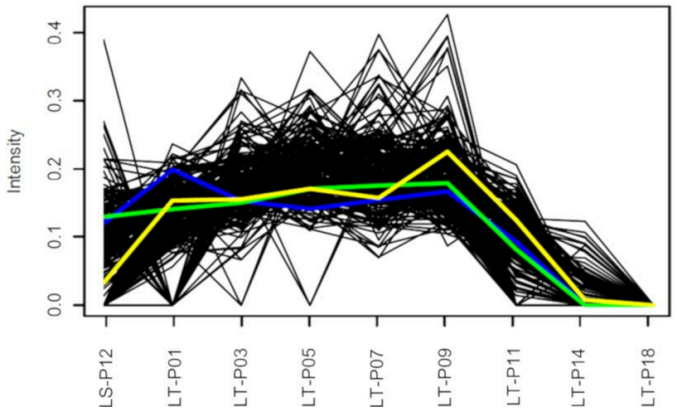


Chiaia-Hernandez *et al.* (2014) Anal. Bioanal. Chem. 406 (28), pp 7323-7335. DOI: 10.1007/s00216-014-8166-0

Examples: Time Trends in Riverbank Filtration Systems



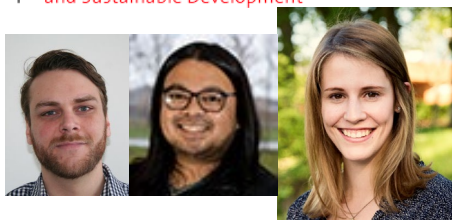


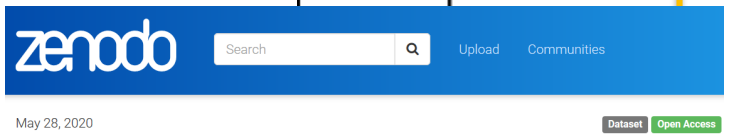


Albergamo *et al*. (2019) *Environ. Sci. Technol.* 53 (13), 7584-7594, DOI: 10.1021/acs.est.9b01750

Examples: Pesticides & Metabolites in Luxembourg

THE GOVERNMENT
OF THE GRAND DUCHY OF LUXEMBOURG
Ministry of the Environment, Climate
and Sustainable Development

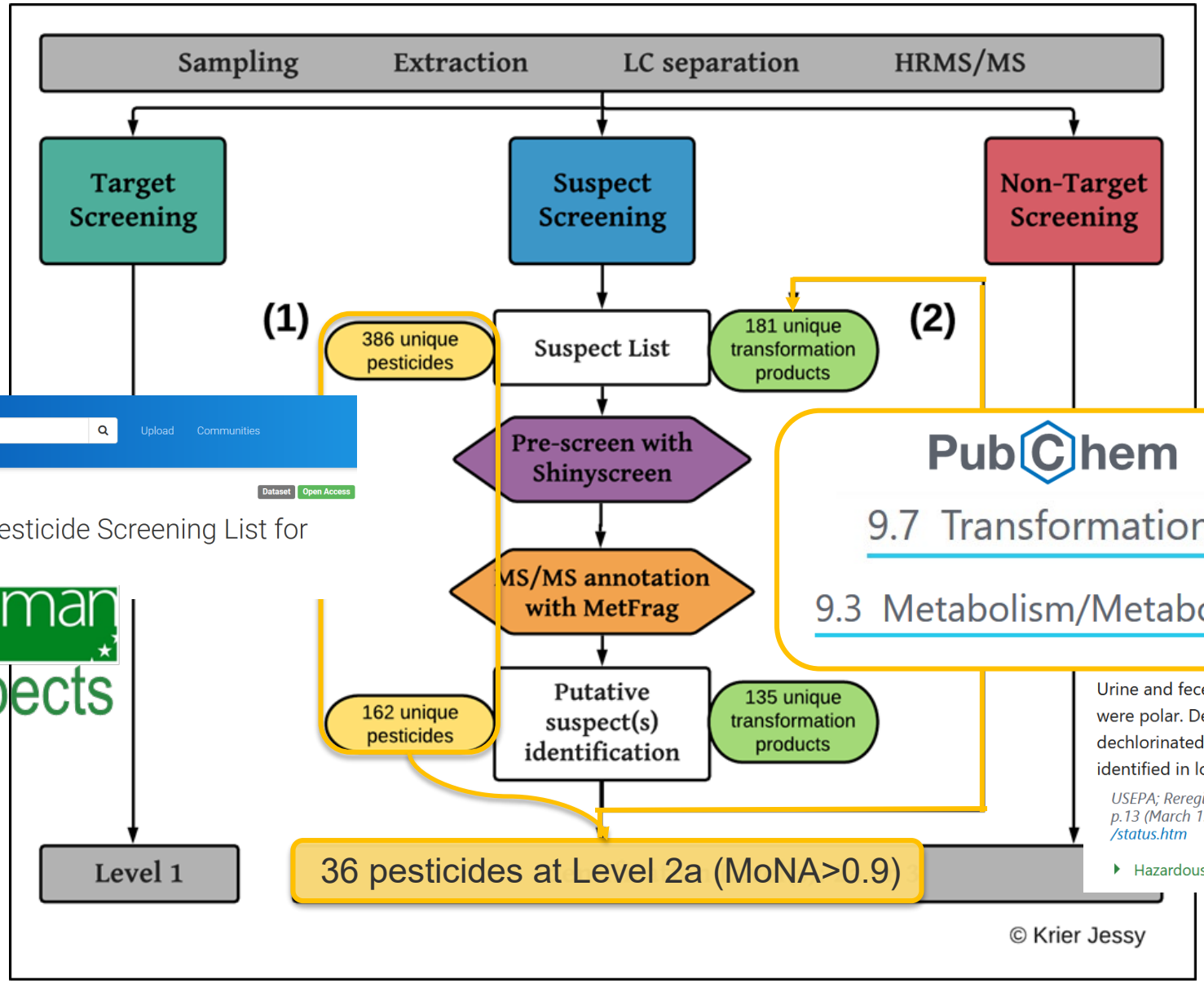




May 28, 2020 Dataset Open Access

S69 | LUXPEST | Pesticide Screening List for Luxembourg

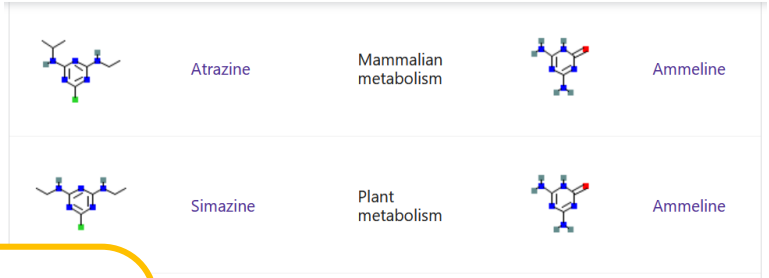
Krier, Jessy



PubChem 9.7 Transformations

PubChem 9.3 Metabolism/Metabolites

PubChem Ammeline (Compound)

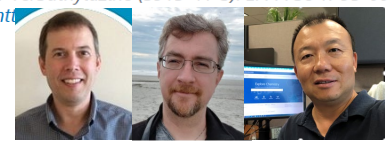


Schymanski *et al.* (2021)
DOI: 10.1186/s13321-016-0115-9

Urine and feces contained up to 25 and 15 identified metabolites, respectively, most of which were polar. Degradation of the triazine ring did not occur. Ammeline and ammelide, 2 dechlorinated and dealkylated/hydroxylated metabolites common to all triazines, were identified in low amounts in the feces.

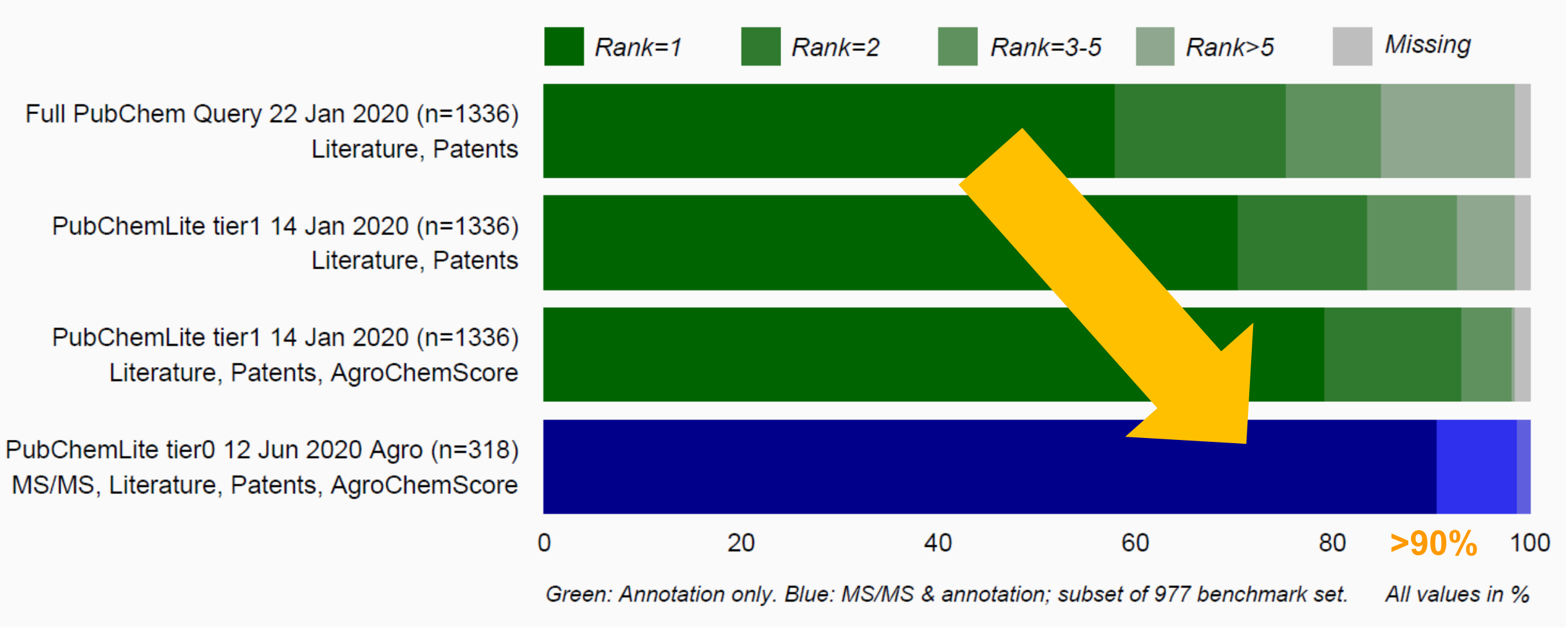
USEPA; Reregistration Eligibility Decision (RED) Database for Terbuthylazine (5915-41-3). EPA 738-R-95-005 p.13 (March 1995). Available from, as of October 11, 2012: htt /status.htm

Hazardous Substances Data Bank (HSDB)



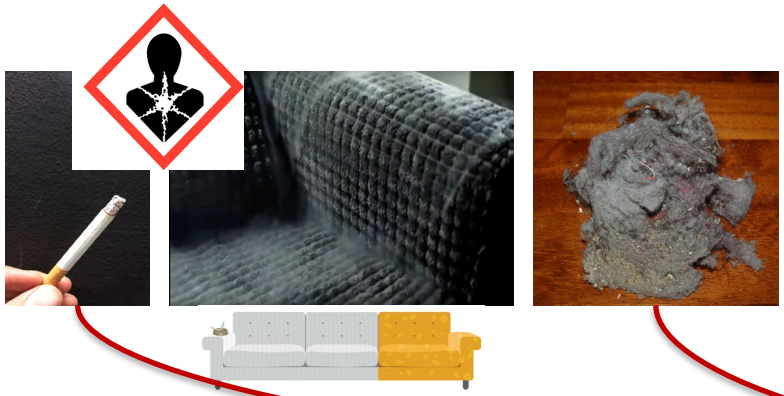
Jessy Krier (2020) Master's Thesis, defended July 2020. Figure 8 & Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

“PubChemLite for Exposomics” … so what now?

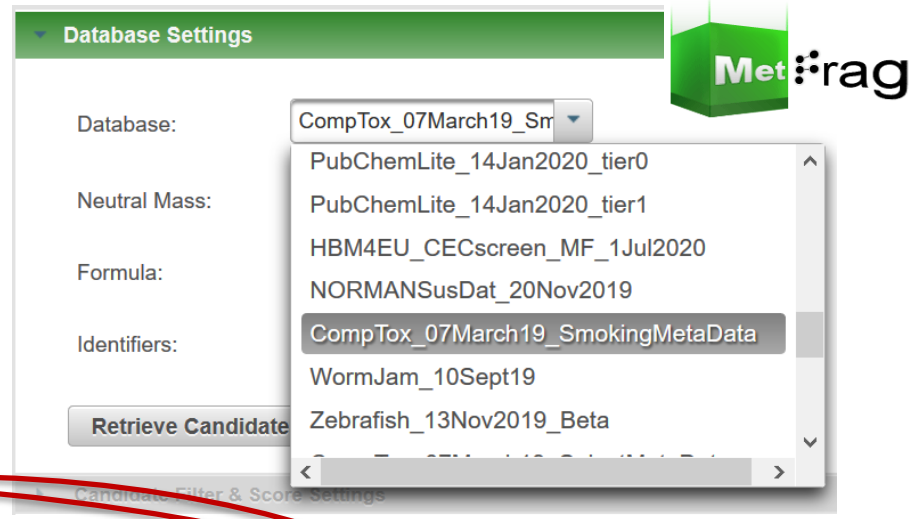


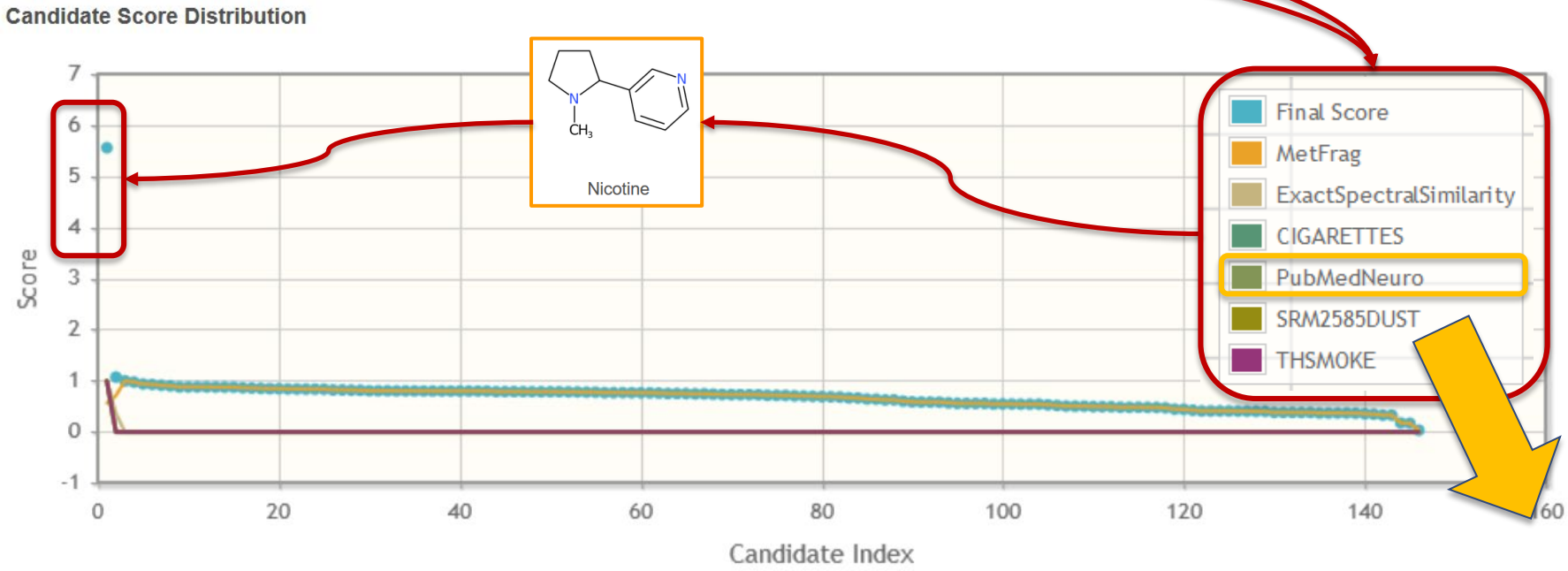
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9

Application to Health and Disease: Thirdhand Smoke



+



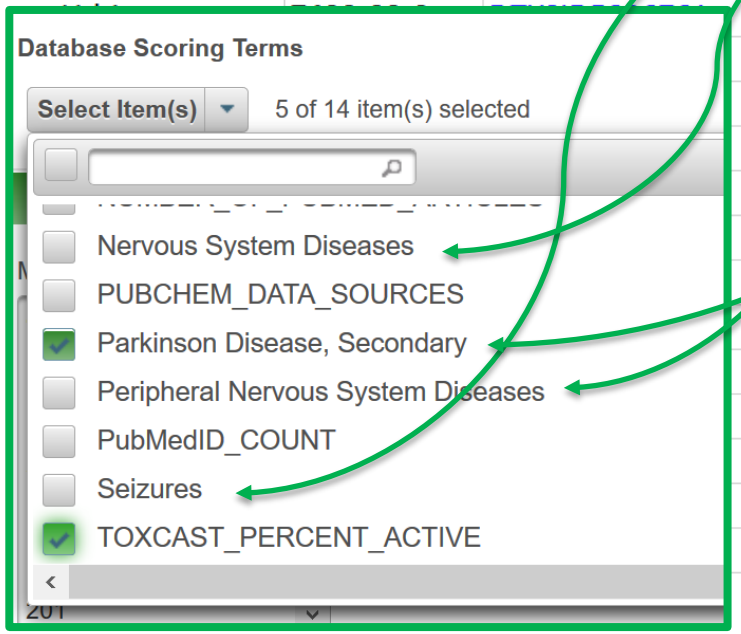


UNIVERSITAT ROVIRA i VIRGILI

Schymanski, EL, Torres, S, & Ramirez, N. (2020). Thirdhand Smoke in House Dust. Zenodo. http://doi.org/10.5281/zenodo.3613472

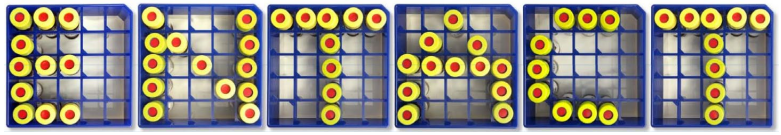
Application: Disease-specific Reference Counts

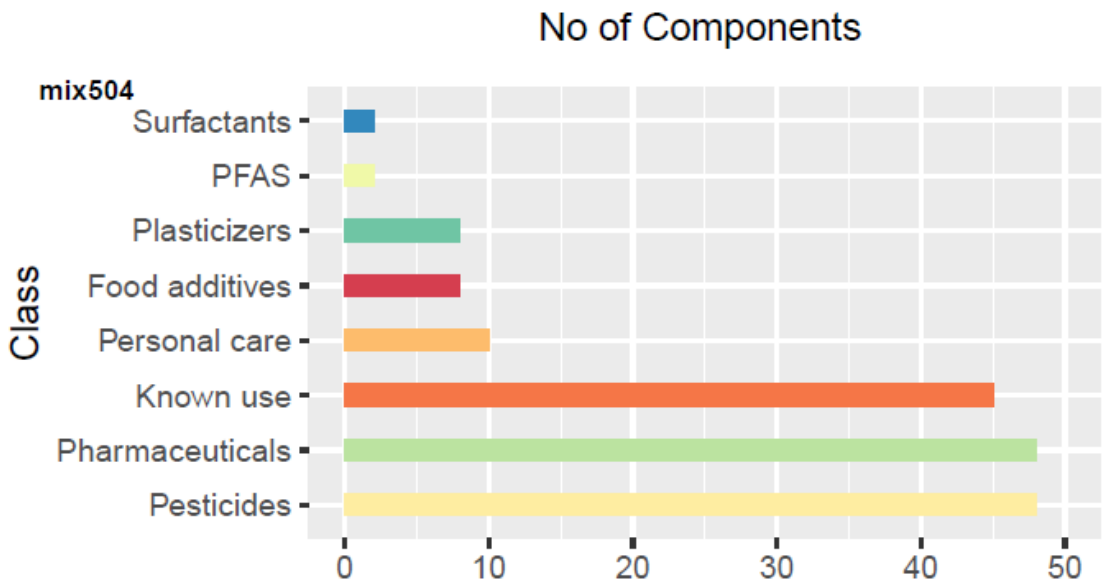
Chemical	CAS RN	DSSToxID	PMID Ct	Seizures	Nervous System Diseases	Peripheral Nervous System Diseases	Brain Diseases	Muscular Diseases	Basal Ganglia Diseases	Parkinson Disease, Secondary	Coma	Hallucinations	Tremor	Memory Disorders	Central Nervous
Cisplatin	15663-27-1	DTXSID4024983	1032	20	47	140	13	0	4	1	1	0	1	2	4
Ethanol	64-17-5	DTXSID9020584	768	100	23	11	18	26	1	3	20	6	17	54	2
Lead	7439-92-1	DTXSID2024161	740	28	107	68	102	4	2	2	1	3	4	19	30
[illegible]	[illegible]	[illegible]	689	30	50	9	22	5	36	13	25	6	93	12	15
[illegible]	[illegible]	[illegible]	666	32	10	3	65	6	10	18	45	5	18	4	2
[illegible]	[illegible]	[illegible]	638	1	24	0	11	0	6	289	0	0	5	0	1
[illegible]	[illegible]	[illegible]	567	17	59	125	15	5	1	1	5	3	2	1	8
[illegible]	[illegible]	[illegible]	560	37	24	25	16	9	3	1	9	3	8	4	6
[illegible]	[illegible]	[illegible]	555	6	6	1	10	6	153	51	4	4	11	1	0
[illegible]	[illegible]	[illegible]	530	151	16	0	8	0	2	3	3	8	6	12	11
[illegible]	[illegible]	[illegible]	489	8	3	0	3	2	2	0	9	4	1	0	5
[illegible]	[illegible]	[illegible]	485	4	43	217	9	14	0	0	0	0	0	1	2
[illegible]	[illegible]	[illegible]	477	13	41	1	105	4	0	0	1	0	1	13	12
[illegible]	[illegible]	[illegible]	464	150	26	15	3	2	0	0	8	4	6	2	10
[illegible]	[illegible]	[illegible]	451	17	25	1	79	4	0	1	5	0	1	9	18
[illegible]	[illegible]	[illegible]	450	6	79	22	23	2	3	5	2	2	38	7	25



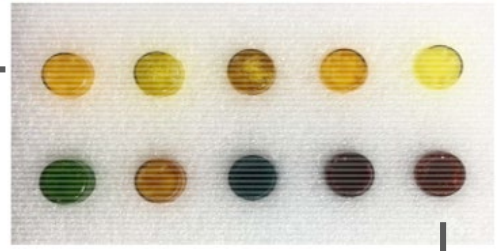
Schymanski *et al.* (2020) DOI: 10.1039/C9EM00068B (Perspective) Environ. Sci.: Processes Impacts (with cover image).

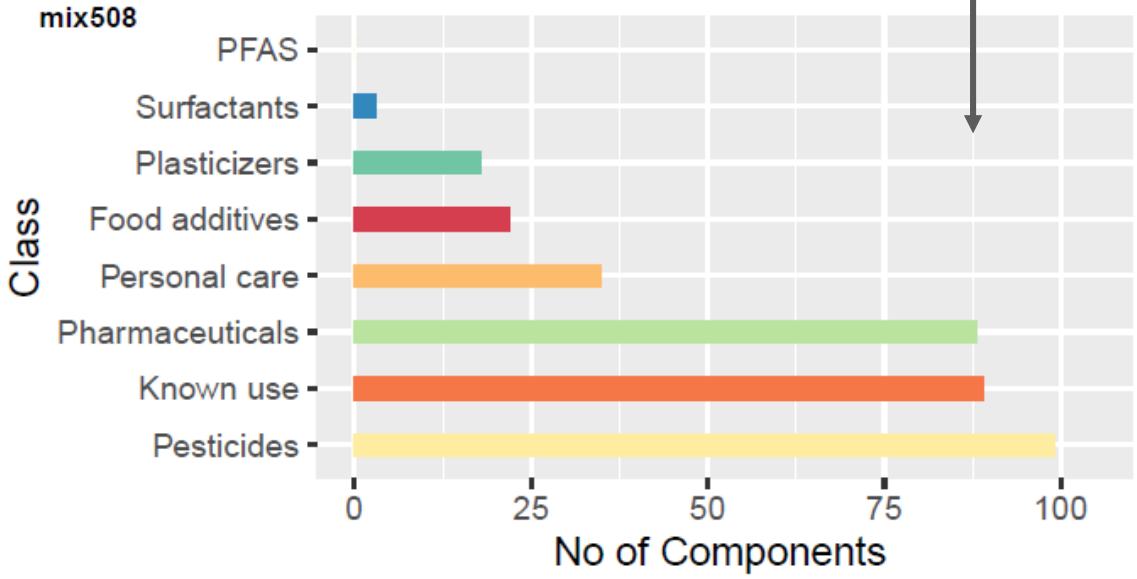
Rapid Classification: All a question of perspective





95, 185 or 365 substances/mixture

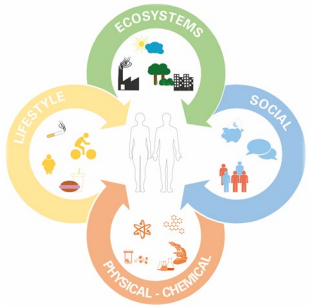




ENTACT: Elin Ulrich *et al.* 2018, https://www.epa.gov/sites/production/files/2018-06/documents/comptox_cop_6-28-18.pdf and Ulrich, EM, et al. (2019) Anal Bioanal Chem. DOI: 10.1007/s00216-018-1435-6
Classification Figures:
Anjana Elapavalore, ECI (Master's thesis)

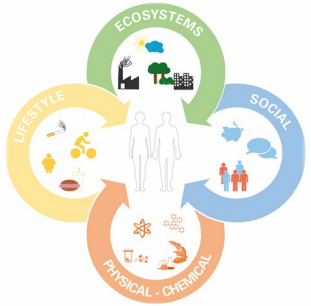
Take Home: Cheminformatics, Exposome, Health & Disease

- “The Exposome” involves many factors …
  - Environmental and biological; over lifespan

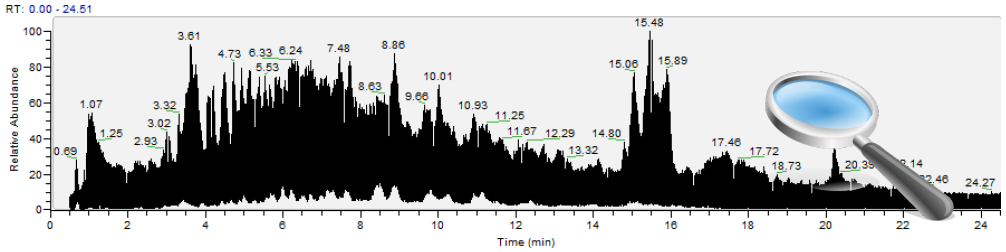


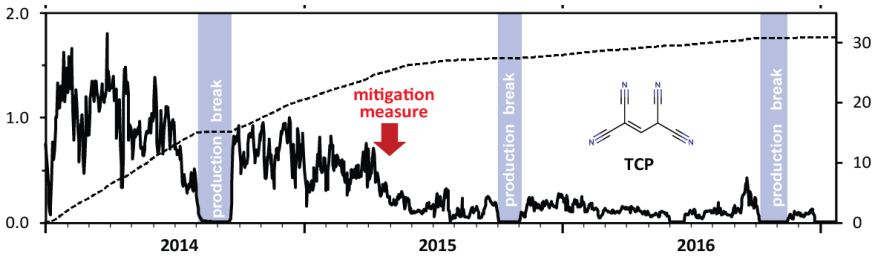
Take Home: Cheminformatics, Exposome, Health & Disease

- “The Exposome” involves many factors …
  - Environmental and biological; over lifespan



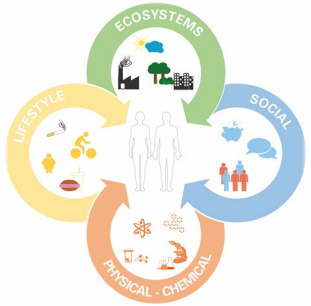
- Mass Spectrometry & Cheminformatics perform the “detective work”
  - Capturing expert knowledge
    - Human and machine readable forms
  - Greatly improved identification and interpretation
  - Methods are applied throughout the world



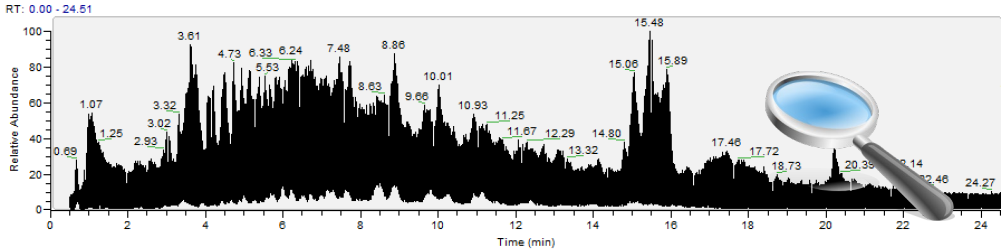


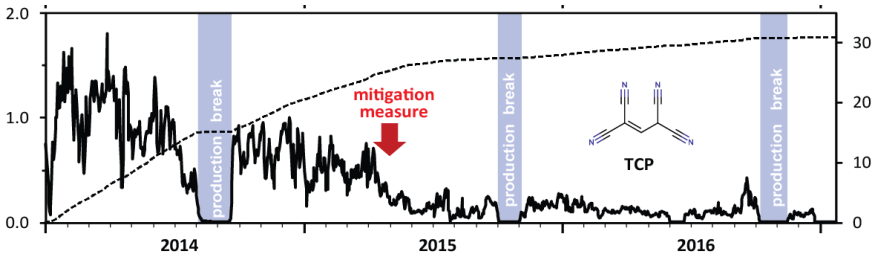
Take Home: Cheminformatics, Exposome, Health & Disease

- "The Exposome" involves many factors …
  - Environmental and biological; over lifespan



- Mass Spectrometry & Cheminformatics perform the "detective work"
  - Capturing expert knowledge
    - Human and machine readable forms
  - Greatly improved identification and interpretation
  - Methods are applied throughout the world





- Tackling the Exposome will require a joint, international effort of many
  - A major motivation for our collaborative #OpenScience work

Thank you!

emma.schymanski@uni.lu and @ESchymanski

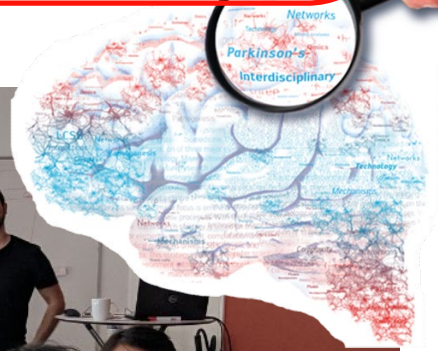
Further Information:

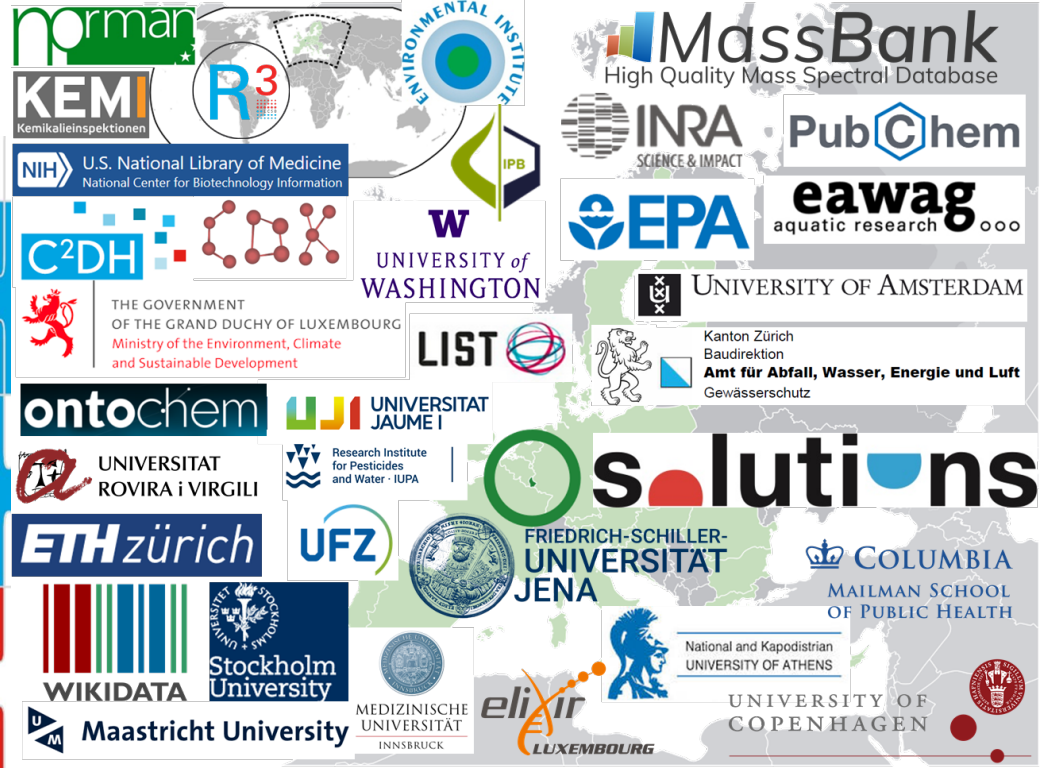
Meet ECI@DOI: 10.5281/zenodo.4596021

https://www.norman-network.com/nds/SLE/

https://zenodo.org/communities/lcsb-eci/

https://wwwen.uni.lu/lcsb/research/environmental_cheminformatics/





uni.lu | INSTITUTE FOR ADVANCED STUDIES (IAS)

Slides available at DOI: 10.5281/zenodo.4678461

Luxembourg National Research Fund

uni.lu | LCSB