Luxembourg National Research Fund
Robust Nature
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
UNIVERSITÉ DU LUXEMBOURG
Funded by the European Union

Environmental Detective Work: Environmental Cheminformatics and the Exposome

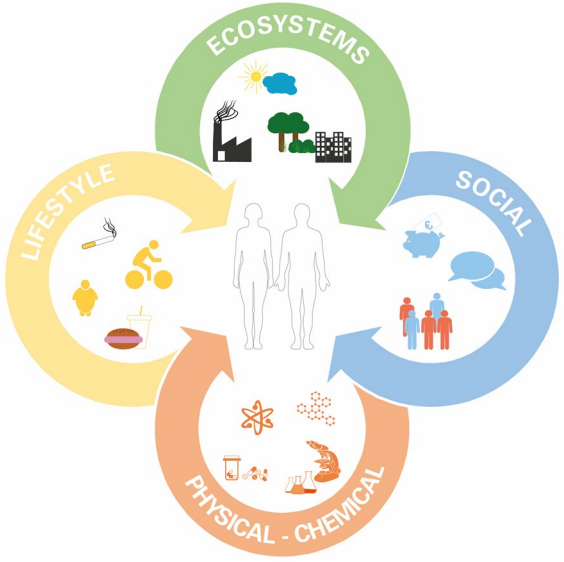


Image mod. from DOI: 10.1126/science.aay3164

Assoc. Prof. Dr. Emma L. Schymanski
(plus many, many colleagues and collaborators!)

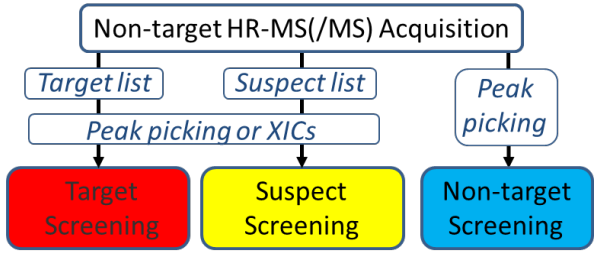
Environmental Cheminformatics Group,
Luxembourg Centre for Systems Biomedicine, University of Luxembourg
Email: emma.schymanski@uni.lu and @ESchymanski
Web: https://wwwen.uni.lu/lcsb/research/environmental_cheminformatics/

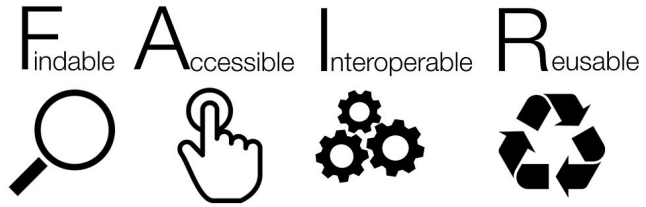
Kolloquium Wintersemester 2021 / 2022
Institut für Ökologie, Evolution und Diversität

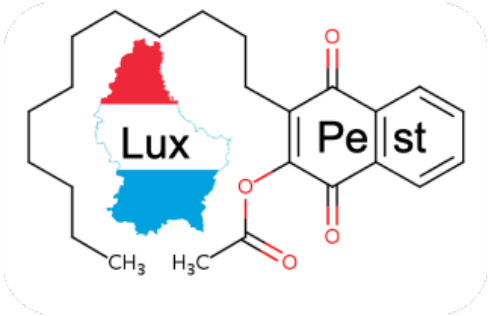
10 January 2022 at 13:15 via Zoom
Slides available at DOI: 10.5281/zenodo.5831546

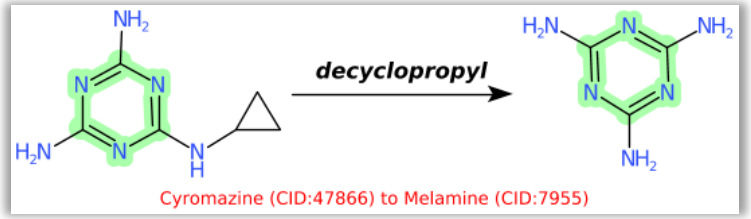
Outline of Today

- Introduction and Background
- Identification & Chemical Space
  - Identification + MetFrag
  - PubChemLite for Exposomics
- Case Study: LuxPest
- Dark Matter and Transformations
- Take-home messages



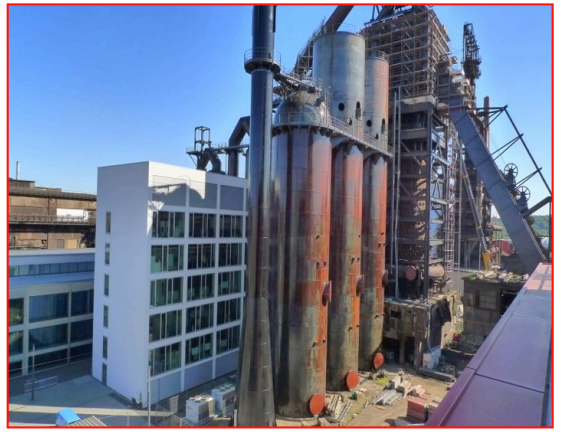






https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

Luxembourg, LCSB and the University



University of Luxembourg
- founded in 2003

Luxembourg Centre for Systems Biomedicine (LCSB)
- founded in 2009

Young & dynamic environment!

https://qbuild.lu/realisations/residence-rouden-eck/

Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885.

Environmental Cheminformatics @ LCSB

Adelene Lai
@adelene.lai
Given access 1 year ago

Anjana Elapavalore
@anjana.elapavalore
Given access 1 year ago

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

Corey Griffith
@corey.griffith
Given access 1 year ago

Dagny Aurich
@dagny.aurich
Given access 3 months ago

Emma Schymanski @emma.schymanski
Given access 1 year ago

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

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

Jessy Krier @jessy.krier
Given access 1 year ago

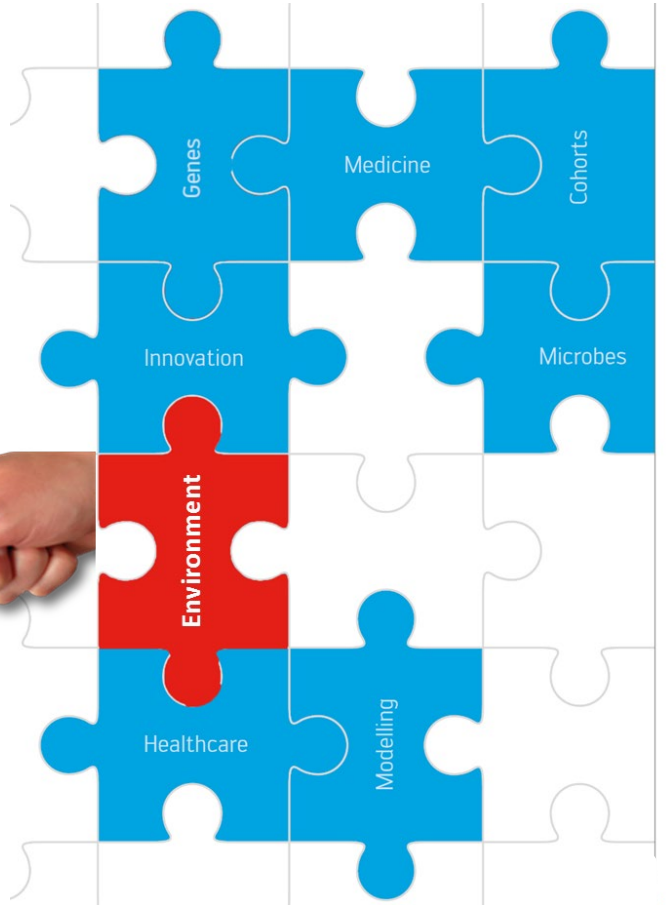
Lorenzo Favilli
@lorenzo.favilli
Given access 1 year ago

Mira Narayanan
@mira.narayanan
Given access 2 months ago

Parviel Chirsir
@parviel.chirsir
Given access 1 month ago

Todor Kondic
@todor.kondic
Given access 1 year ago

R3

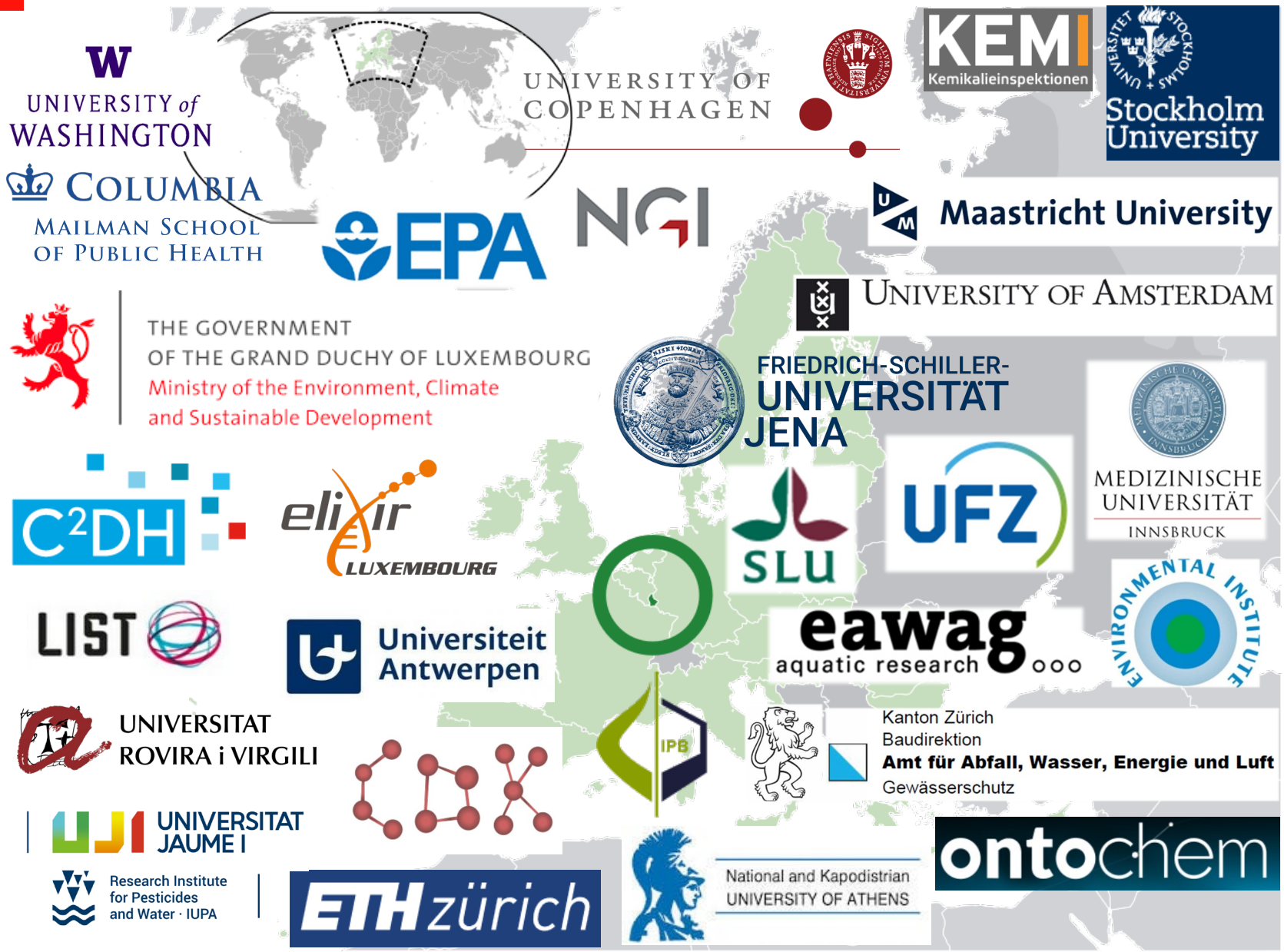


Funded by the European Union

INSTITUTE FOR ADVANCED STUDIES (IAS)

Luxembourg National Research Fund

… plus many National, European & International Collaborations!



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

nprman

PubChem
836 Data Sources

MassBank
High Quality Mass Spectral Database

ZeroPM
Prevent
Prioritize
Remove

Funded by the European Union

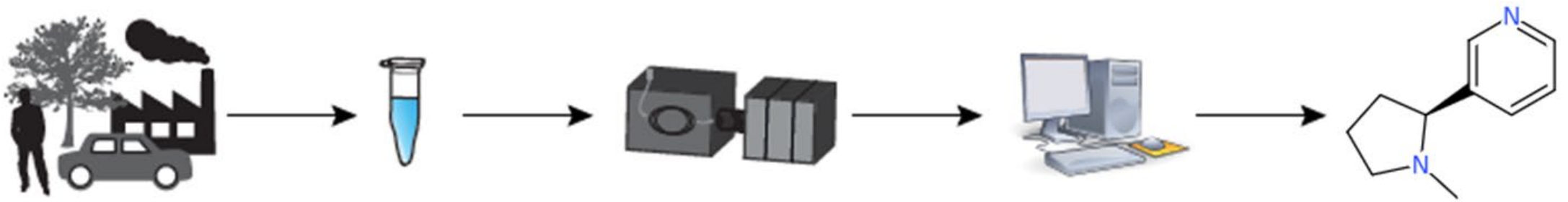
WIKIDATA

R3

Luxembourg National Research Fund

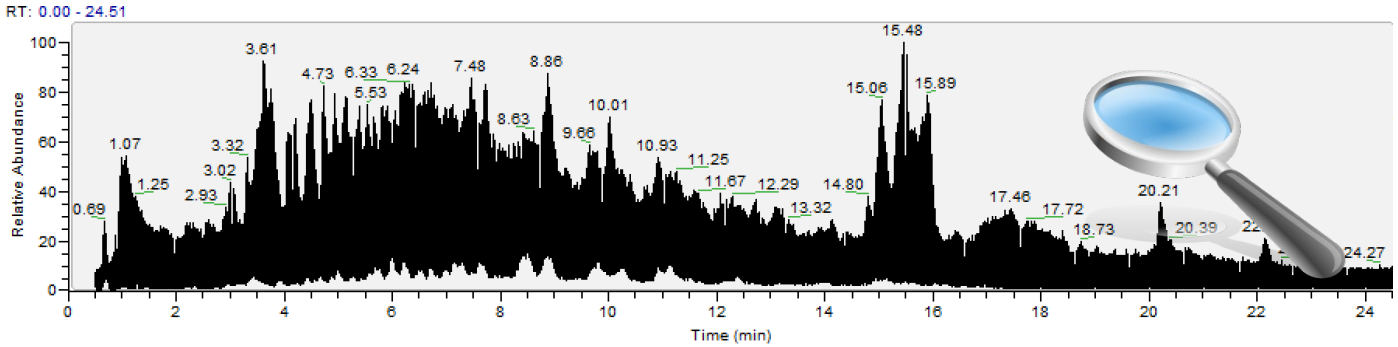
Source:
https://en.wikipedia.org/wiki/File:EU-Luxembourg.svg

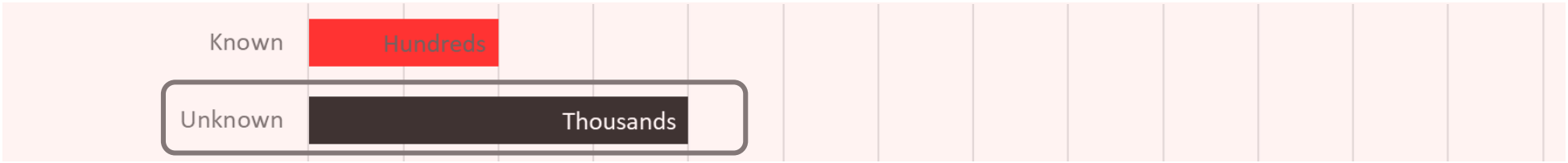
Environmental Cheminformatics & HR-MS



High resolution mass spectrometry

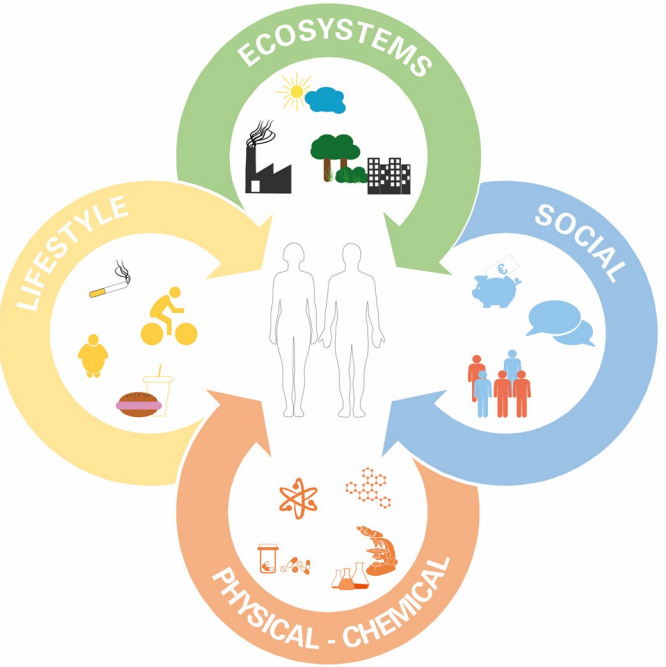
AND connecting chemical knowledge





Samples

Environment … versus … 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)
- Less than approx. 10 % of disease can be explained by genetic factors alone (Vermeulen *et al.* 2020 & refs within)

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

The Exposome and Emerging Contaminants

- Emerging contaminants are part of the exposome …
  - But not the whole picture
- Research by necessity crosses environmental, biological, medical fields and beyond …
  - “exposomics”

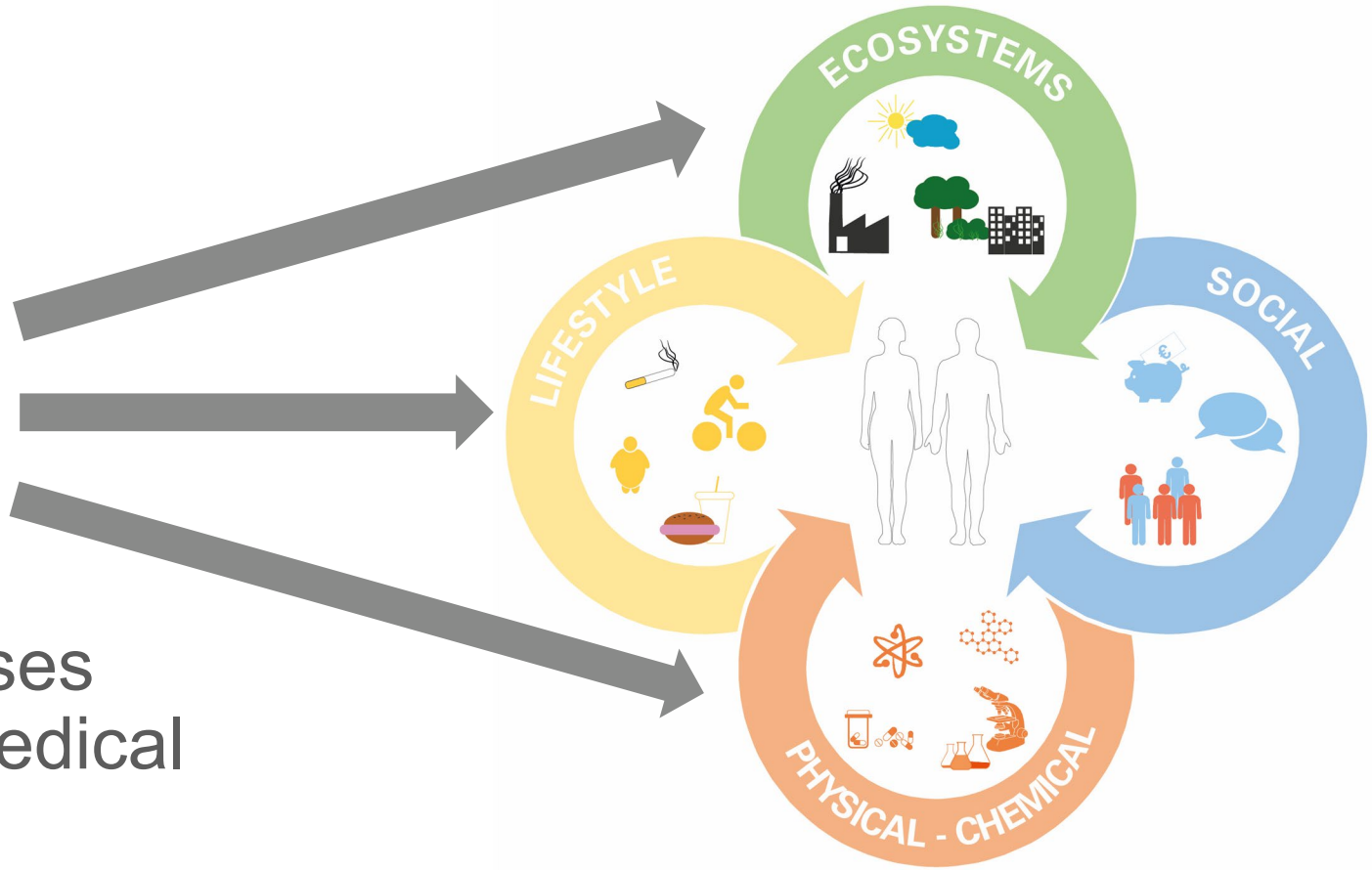
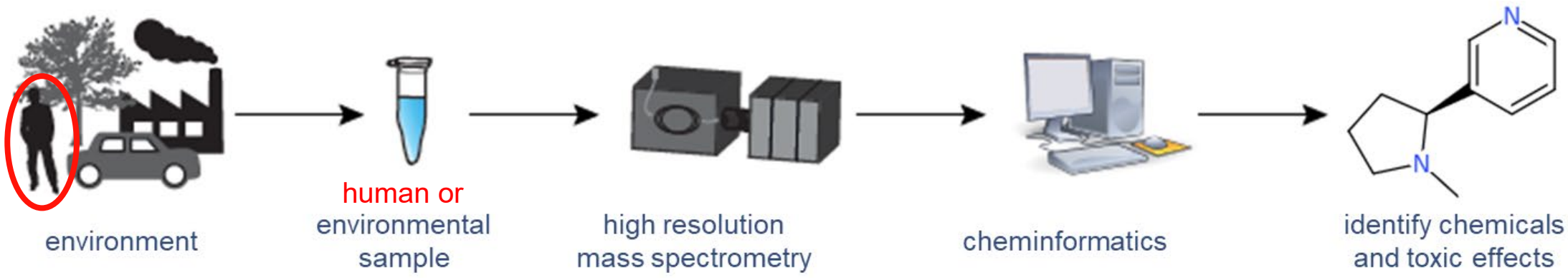


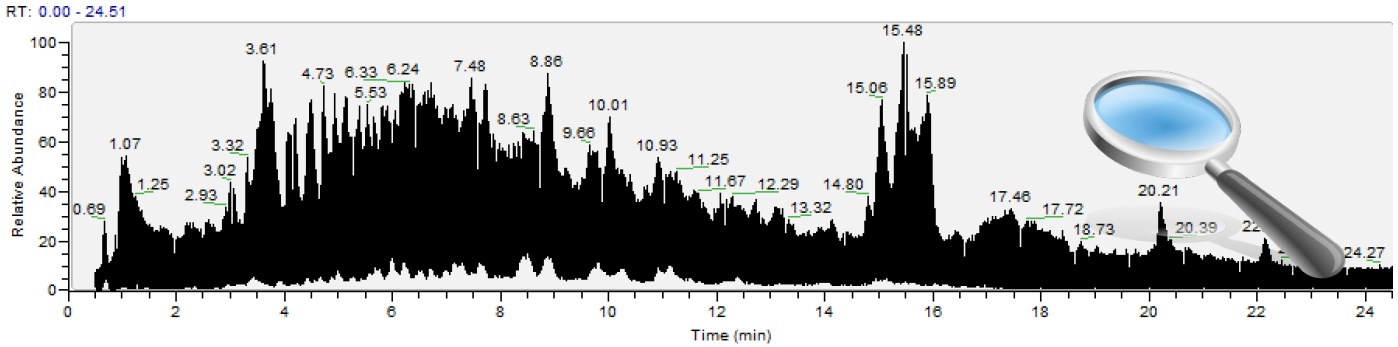
Image c/o Lucie Debroux LCSB, modified from images by Utrecht Uni & Vermeulen *et al.* (2020). Science. DOI: 10.1126/science.aay3164.

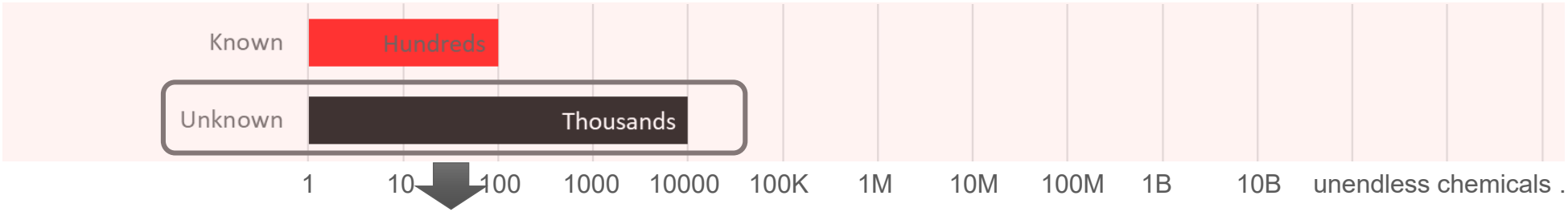
Environmental Cheminformatics, *Exposomics* & HR-MS



High resolution mass spectrometry

AND connecting chemical knowledge



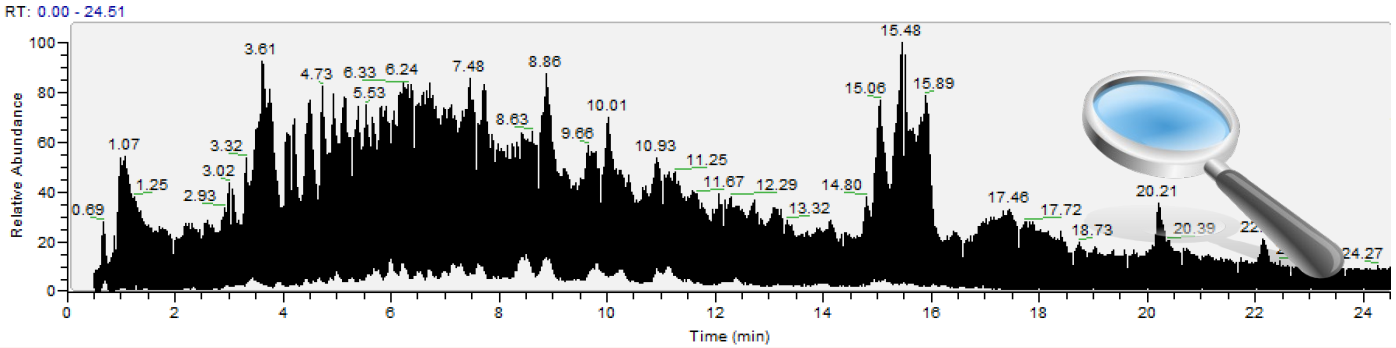


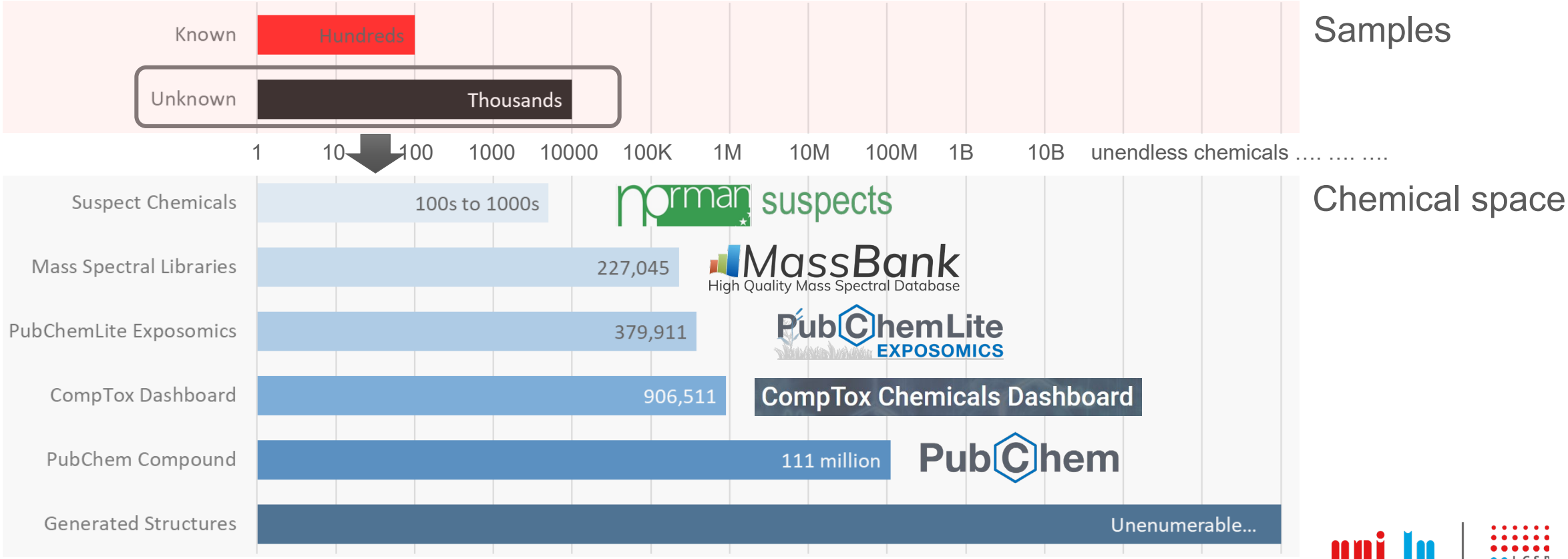
Samples

Environmental Cheminformatics, *Exposomics* & HR-MS

High resolution mass spectrometry

AND connecting chemical knowledge



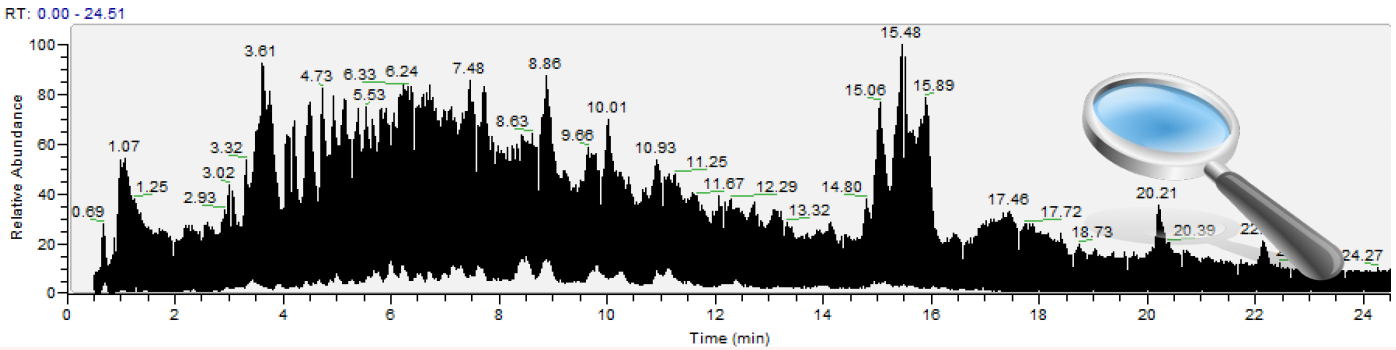


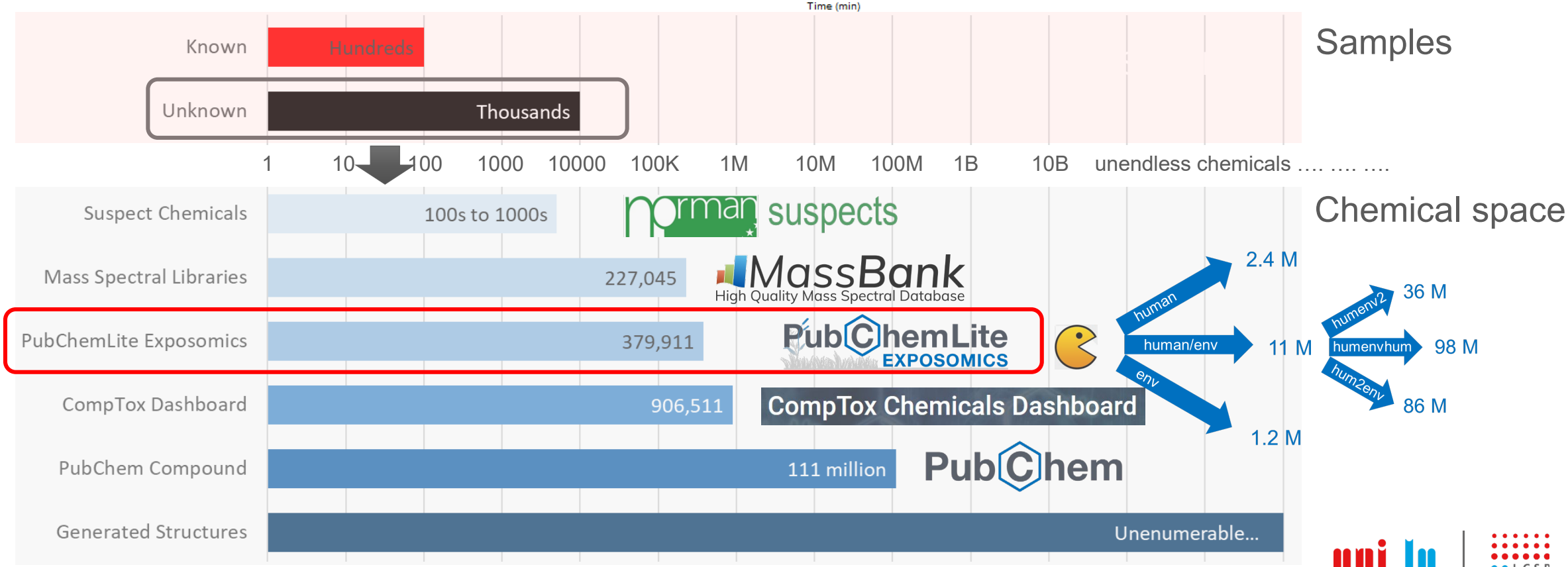
Chromatogram: Schymanski *et al*. (2014) DOI: 10.1021/es4044374; Resource numbers updated Dec 2021.

Environmental Cheminformatics, *Exposomics*, MS & *Metabolites*

High resolution mass spectrometry

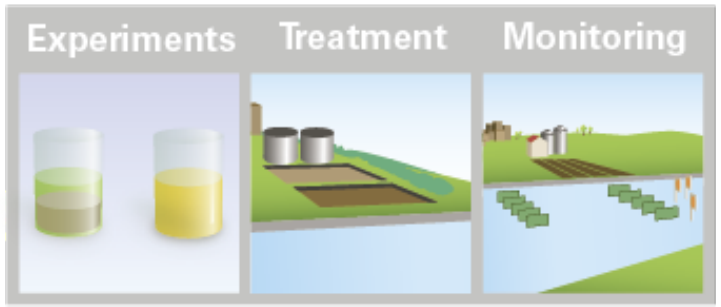
AND connecting chemical knowledge





Chromatogram: Schymanski *et al*. (2014) DOI: 10.1021/es4044374; Resource numbers updated Dec 2021. BT numbers: Bolton *et al* 2021 *in prep.*

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



Commercial / Open Software

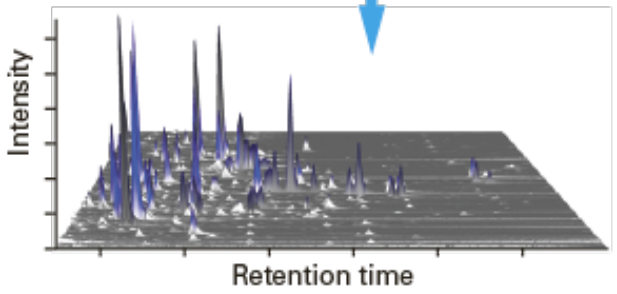
Databases

In Silico Prediction

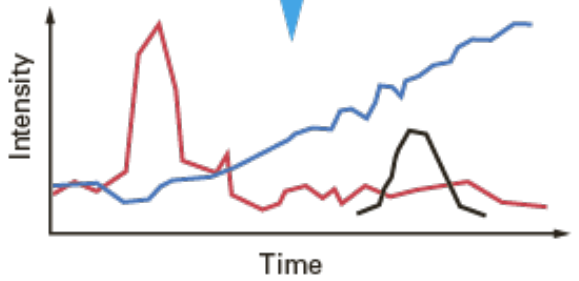
Sampling → Analysis → Data Pre-Processing → Prioritization → Identification

SAMPLE ARCHIVE

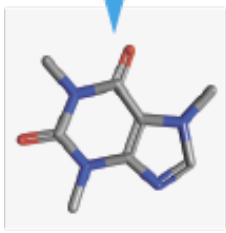
DIGITAL ARCHIVE



POLLUTION OVERVIEW



PATTERN



STRUCTURES

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/

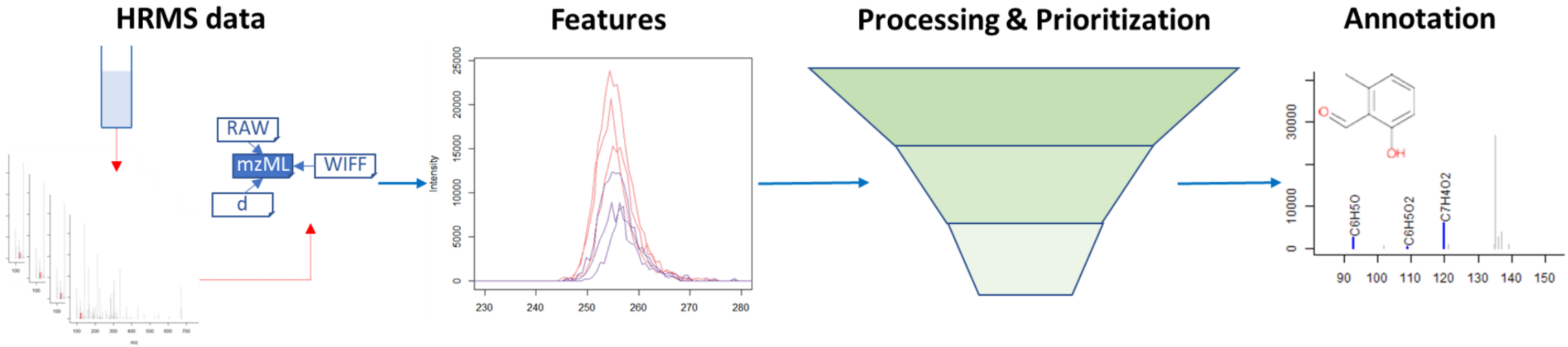
Rick Helmus

Software | Open Access | Published: 06 January 2021

patRoon: open source software platform for environmental mass spectrometry based non-target screening

Journal of Cheminformatics **13**, Article number: 1 (2021) | Cite this article

Rick Helmus, Thomas L. ter Laak, Annemarie P. van Wezel, Pim de Voogt & Emma L. Schymanski



Data acquisition & pre-treatment

Chromatographic peak detection & cross-analysis grouping

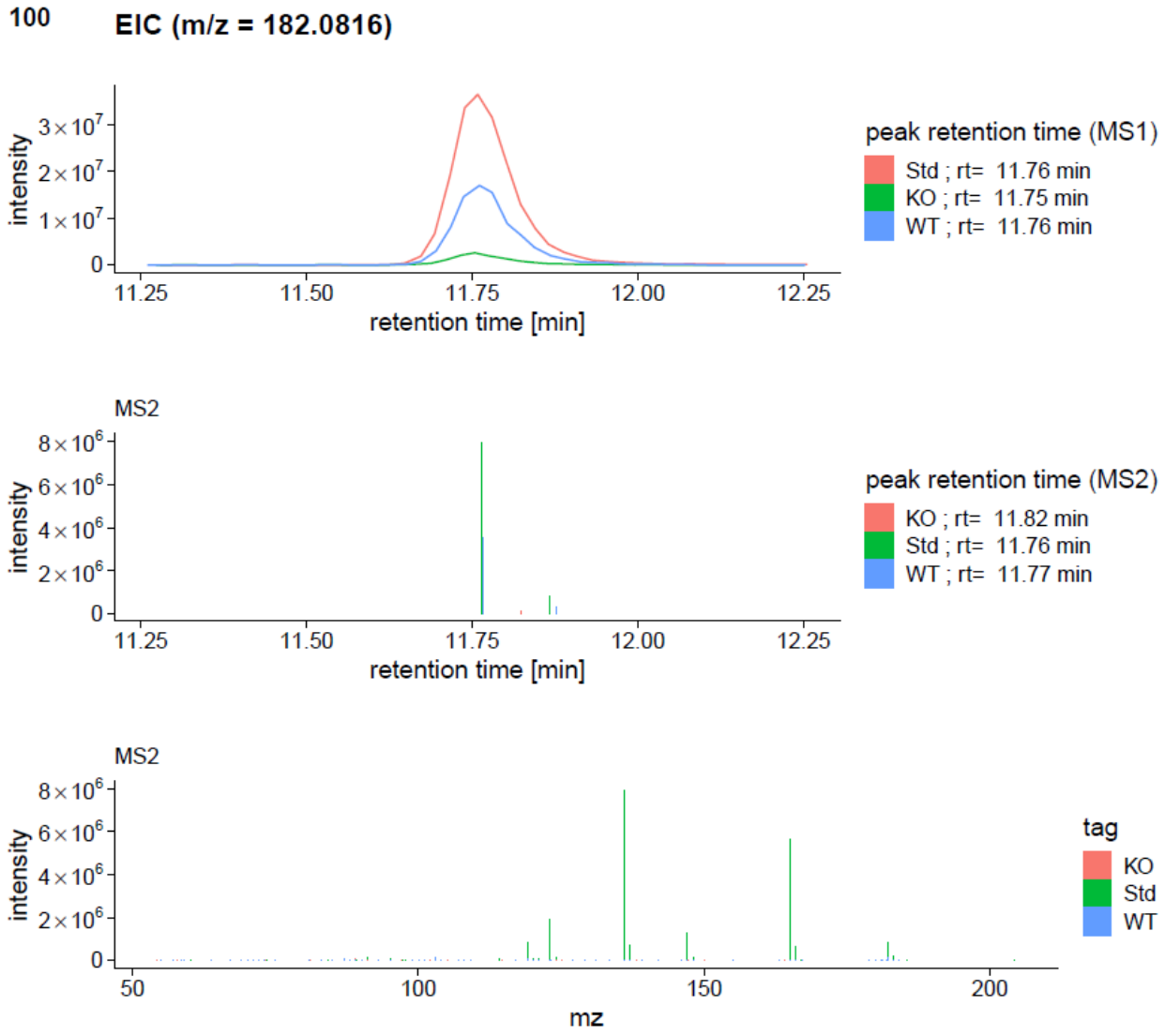
Rule-based filtering, componentization & prioritization

Chemical formula & compound annotation

Helmus *et al.* (2021) patRoon. *J. Cheminformatics*. 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.

Kondić *et al.* (in prep) https://gitlab.lcsb.uni.lu/eci/shinyscreen

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

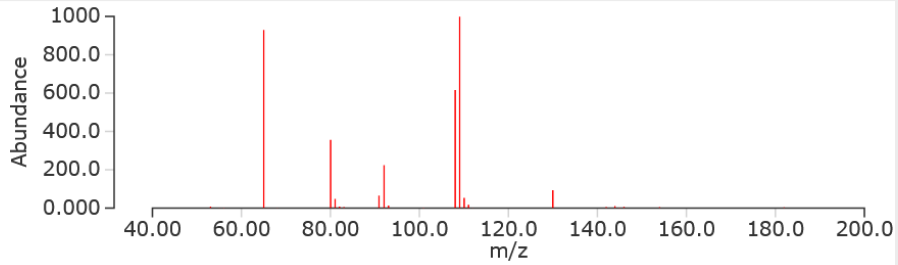
MassBank
High Quality Mass Spectral Database

>> Search Spectra

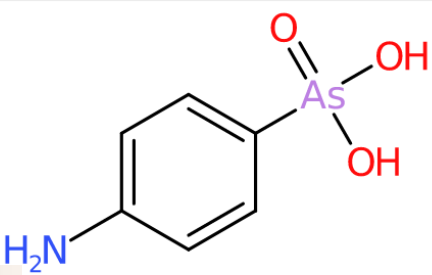
MassBank Record: LU040605

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

Mass Spectrum



Chemical Structure



Search for:

Basic Search | Peak List | Peaks | Peak Differences

Compound Information

Compound name

Exact Mass

AND

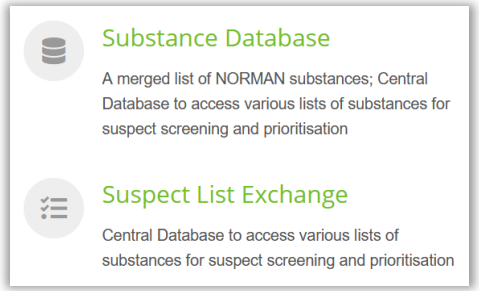
Formula (e.g. C6H7N5, C5H*N5, C5*)

AND

Search

Expert Knowledge: NORMAN Database System

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



NORMAN Database System

NORMAN organises the development and maintenance of various web-based databases for the collection & evaluation of data / information on emerging substances in the environment

SEARCH All Databases

Searching for individual substance or group(s) of substances in all databases

Note: Click on a link below to go to an individual database home page

SARS-CoV-2 in sewage

A database with the latest information on SARS-CoV-2 in sewage across Europe and internationally; including a common protocol for sample collection, storage, extraction, analysis and data sharing to support the development of an international comparable data set.

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

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

Substance Database

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

Suspect List Exchange

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

Contact us:
normansle@uni.lu

NORMAN WEBSITE | NORMAN DATABASE SYSTEM | HOME | LOGIN

NORMAN SUBSTANCE DATABASE

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 questions. The NORMAN-SLE documents all individual collections that form a part of the merged collection NORMAN SusDat. The original SLE lists should be consulted to verify SusDat information if necessary (see Source column in SusDat). NORMAN-SLE versions are tracked on Zenodo.

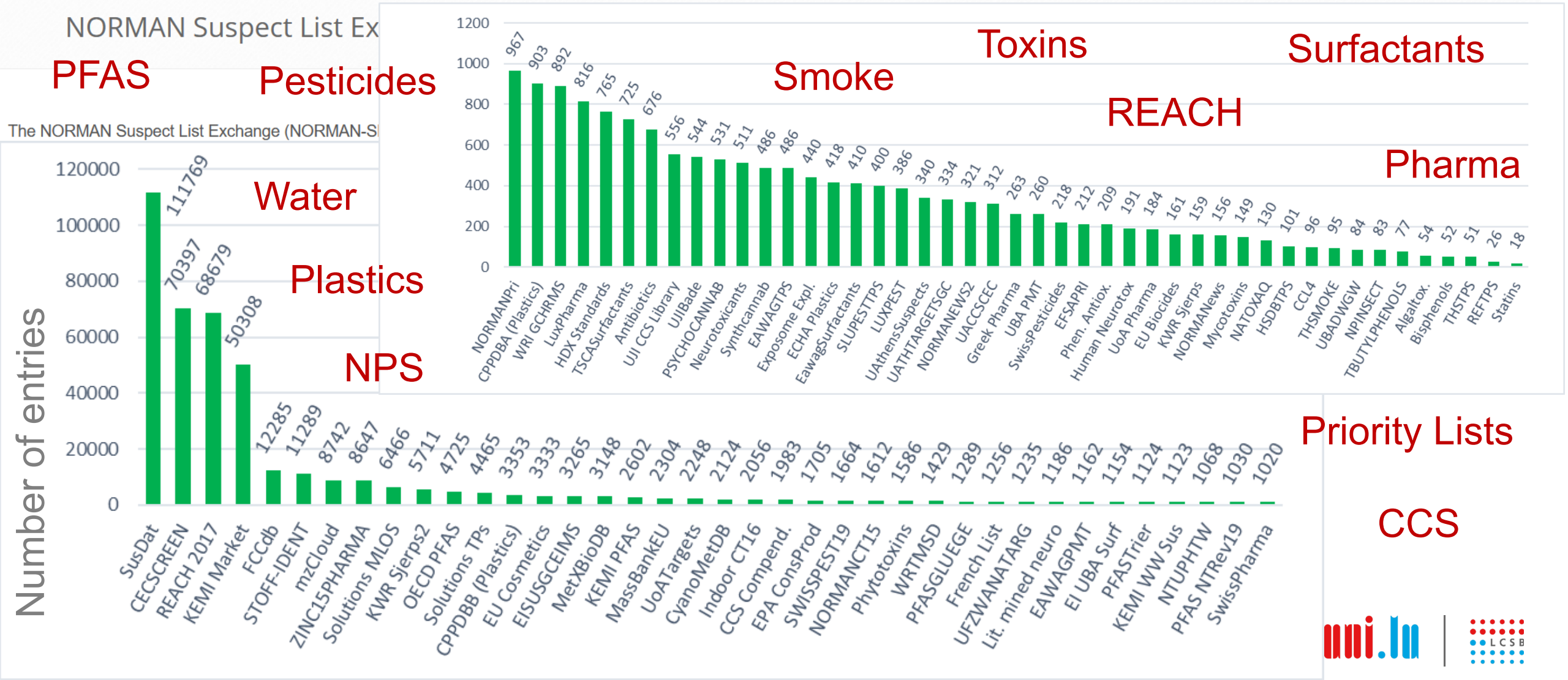
Comments and contributions are welcome - please email us at suspects@normandata.eu.

Please refer to our documentation pages for: citation instructions, credits, updates, license details, SDFs and other useful tips!

No.	Abbreviation	Description	Link to full list	Link to InChIKey list	References
S0	SUSDAT	**Merged NORMAN Suspect List: SusDat**	Interactive Data table SusDat with Haz and Expo scores as XLSX, CSV (06/11/2020) MetFrag CSV (03/03/2020) CompTox SUSDAT List	SusDat InChIKeys: All, MS-ready (18/06/2020)	A merged list of >111,000 structures from SLE suspect lists. See interactive version. Compiled by Reza Aalizadeh, Nikiforos Alygizakis and Lubos Cirka, University of Athens/EI, including RTI and toxicity values, with Hazard and Exposure values provided by Stellan Fischer, KEMI, documented here. *Work in progress ... please report any issues!* DOI: 10.5281/zenodo.2664077
S89	PRORISKPFAS	**List of PFAS Compiled from NORMAN SusDat**	PRORISKPFAS in XLSX, CSV (09/12/2021) CompTox PRORISKPFAS List (coming soon…)	PRORISKPFAS InChIKeys (09/12/2021)	A compiled list of 4777 PFAS occurring in NORMAN SusDat, created by merging SLE lists S9, S14, S25, S46, and S80 and searching for additional fluorinated content in SusDat. Provided by Kelsey Ng, EI (manuscript in prep.). Dataset DOI: 10.5281/zenodo.5769582

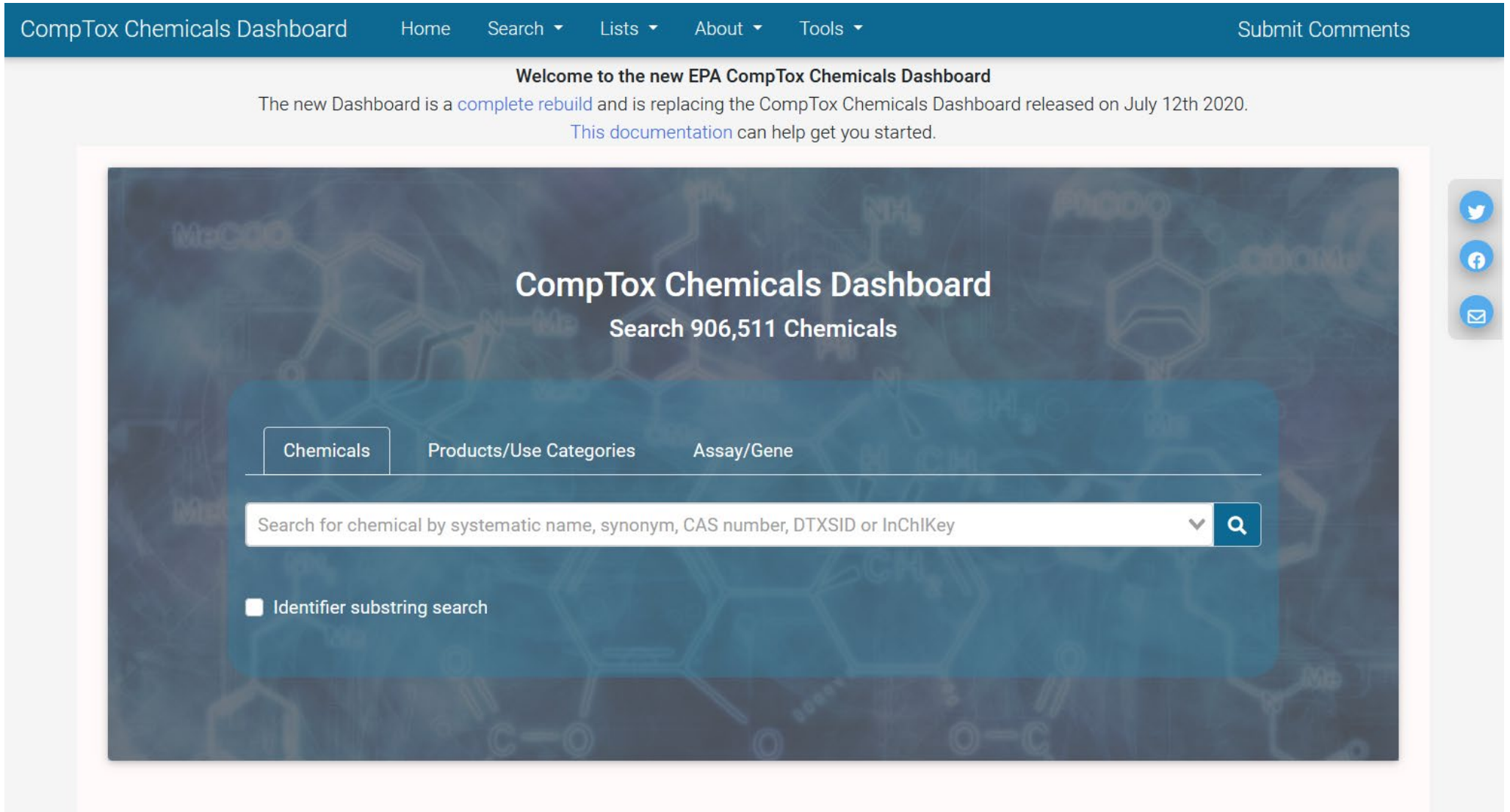
NORMAN Suspect List Exchange (now 90 lists+)

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



Toxicology: CompTox Chemicals Dashboard

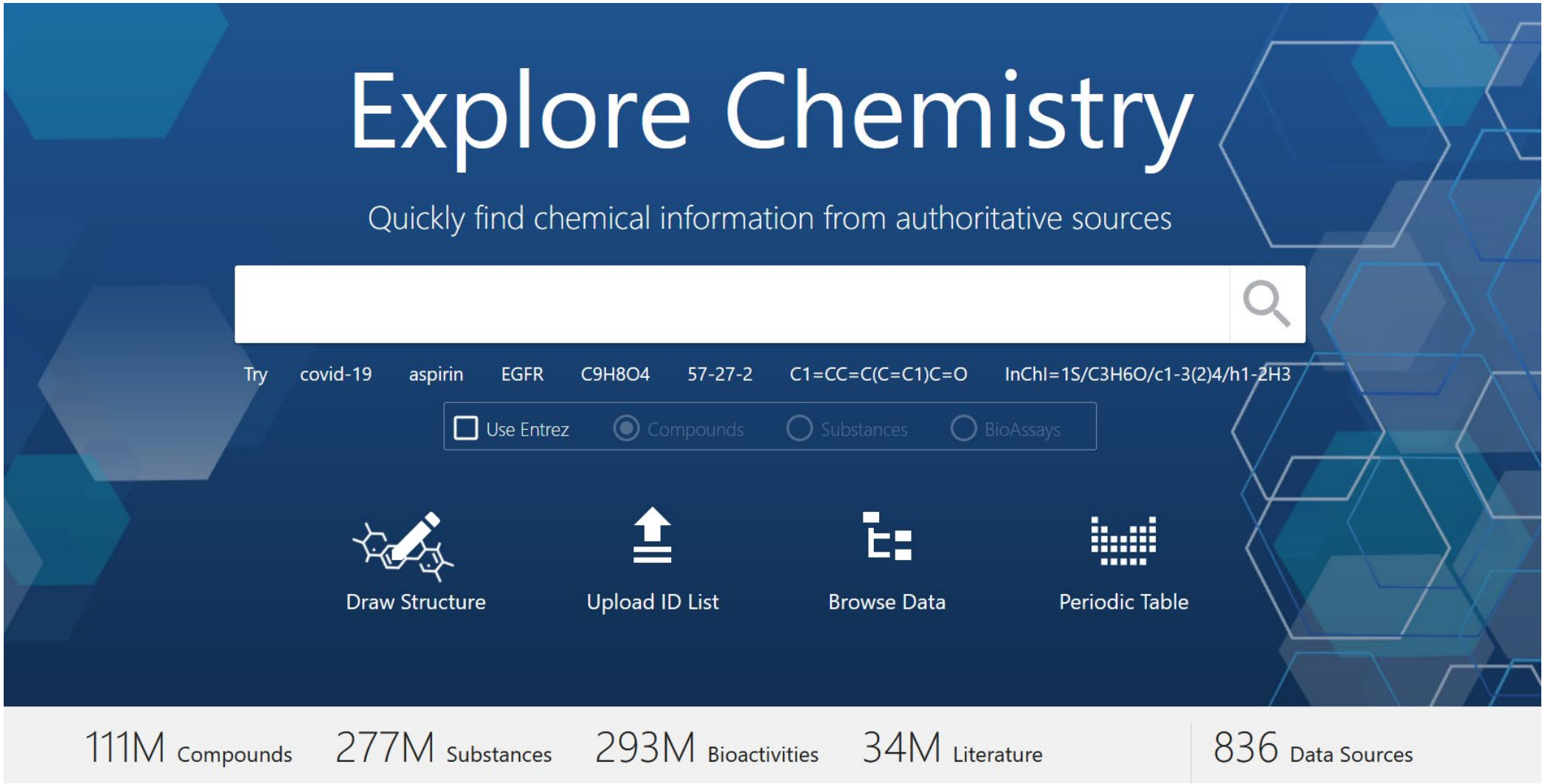
https://comptox.epa.gov/dashboard/



Open Chemistry Knowledge Bases: PubChem

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

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



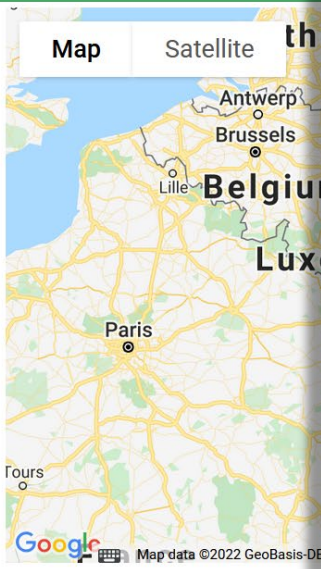
NORMAN-SLE meets PubChem

DATA SOURCES

NORMAN Suspect List Exchange

The NORMAN network enhances the exchange of information on emerging environmental substances, and
harmonisation of common measurement methods and monitoring tools so that the requirements of risk ass
met. It specifically seeks both to promote and to benefit from the synergies between research teams from d
emerging substances.

Organization	NORMAN Network (c/o UniLu)
Category	Research and Development
URL	https://www.norman-network.com/nds/SLE/
License Note	Data: CC-BY 4.0; Code (hosted by ECI, LCSB): Artistic-2.0
License URL	https://creativecommons.org/licenses/by/4.0/
Contact Name	Emma Schymanski
Address	6 avenue du Swing, Belvaux, Luxembourg, 4367
Data Source ID	23819
Data in PubChem	116,423 Live Substances 21,090 Annotations 1 Classification
Last Updated	2021/12/28



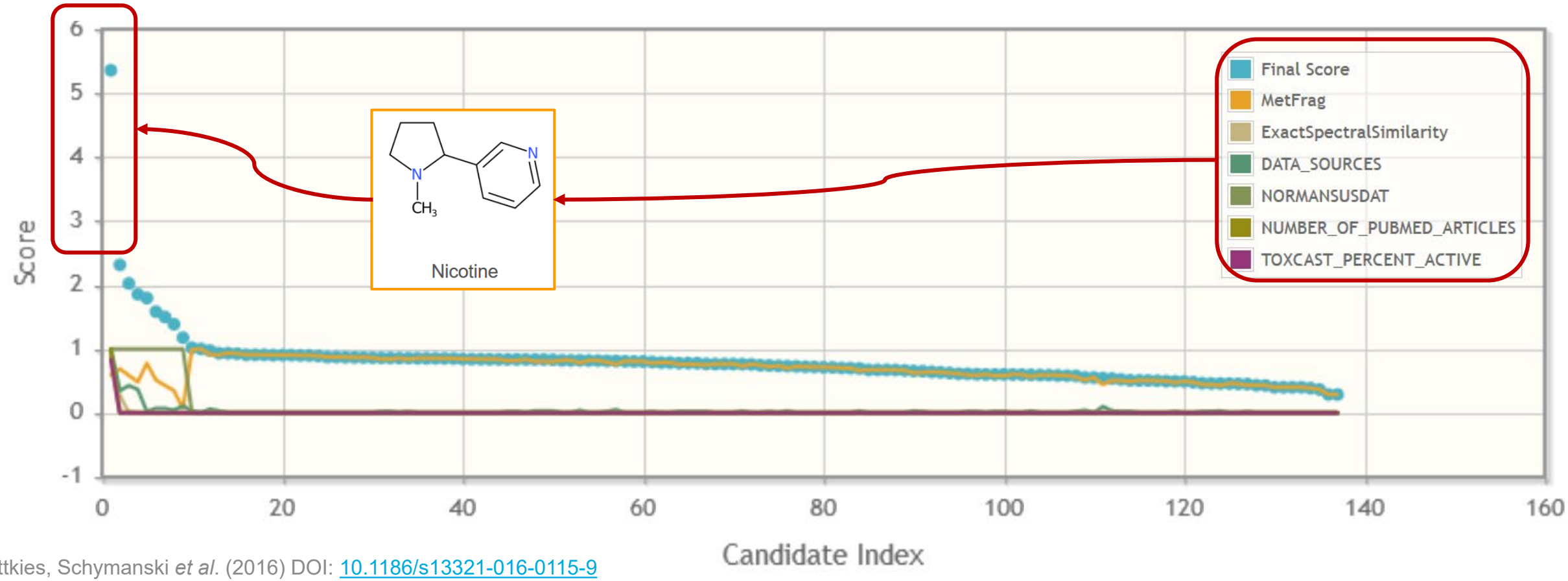
- NORMAN Suspect List Exchange Classification ? ↗ 114,929
  - S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) ? 3,906
  - S25 | OECDPFAS | List of PFAS from the OECD ? 3,677
  - S36 | UBAPMT | Potential Persistent, Mobile and Toxic (PMT) substances ? 254
  - S47 | ECHAPLASTICS | A list from the Plastic Additives Initiative Mapping Exercise by ECHA ? 241
  - S50 | CCSCOMPEND | The Unified Collision Cross Section (CCS) Compendium ? 885
  - S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019 ? 1,347
  - S61 | UJICCSLIB | Collision Cross Section (CCS) Library from UJI ? 574
  - S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag ? 258
  - S68 | HSDBTPS | Transformation Products Extracted from HSDB Content in PubChem ? 102
  - S69 | LUXPEST | Pesticide Screening List for Luxembourg ? 386
  - S72 | NTUPHTW | Pharmaceutically Active Substances from National Taiwan University ? 1,068
  - S75 | CyanoMetDB | Comprehensive database of secondary metabolites from cyanobacteria ? 2,088
  - S79 | UACCSCEC | Collision Cross Section (CCS) Library from UAntwerp ? 148
  - S00 | SUSDAT | Merged NORMAN Suspect List: SusDat ? 99,129
  - S01 | MASSBANK | NORMAN Compounds in MassBank EU ? 7,168
  - S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances ? 11,261

Data source: https://pubchem.ncbi.nlm.nih.gov/source/23819
Classification tree: https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101

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

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

How? Community Efforts … through *Open* and *FAIR* Science

Enabling easier (and FAIRer) data exchange with simple templates…

…can make great things happen!

FAIR chemical structures in the Journal of Cheminformatics

Emma L. Schymanski and Evan E. Bolton

Letter to the Editor | 7 July 2021

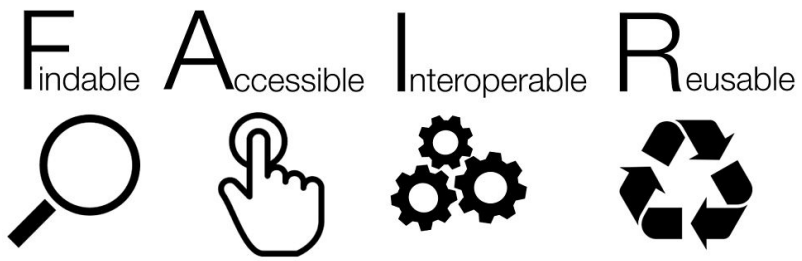
The Letter Response to this article has been published in *Journal of Cheminformatics* 2021 **13**:49

Reply to "FAIR chemical structure in the Journal of Cheminformatics"

Rajarshi Guha, Nina Jeliazkova, Egon Willighagen and Barbara Zdrazil

Letter Response | 7 July 2021

The Letter to the Editor to this article has been published in *Journal of Cheminformatics* 2021 **13**:50



PubChem_CID	Name	SMILES	InChIKey
2256	Atrazine	CCNC1=NC(=NC(=N1)Cl)NC(C)C	MXWJVTOOROXGIU-UHFFFAOYSA-N
2328	Bentazone	CC(C)N1C(=O)C2=CC=CC=C2NS1(=O)=O	ZOMSMJKLGFBRBS-UHFFFAOYSA-N
3030	Dicamba	COC1=C(C=CC(=C1C(=O)O)Cl)Cl	IWEDIXLBFLAXBO-UHFFFAOYSA-N
3120	Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl	XMTQQYYKAHVGBJ-UHFFFAOYSA-N

Predecessor_CID	Predecessor_Name	Transformation	Successor_CID	Successor_Name	Biosystem
13101	6PPD	Ozone	154926030	6PPD-quinone	Environment
2256	Atrazine	Environmental	13878	Deisopropyl-atrazine	Soil
2256	Atrazine	Mammalian metabolism	135408770	Ammeline	Mammal
2256	Atrazine	Fungal metabolism	22563	Desethyl-atrazine	Fungus

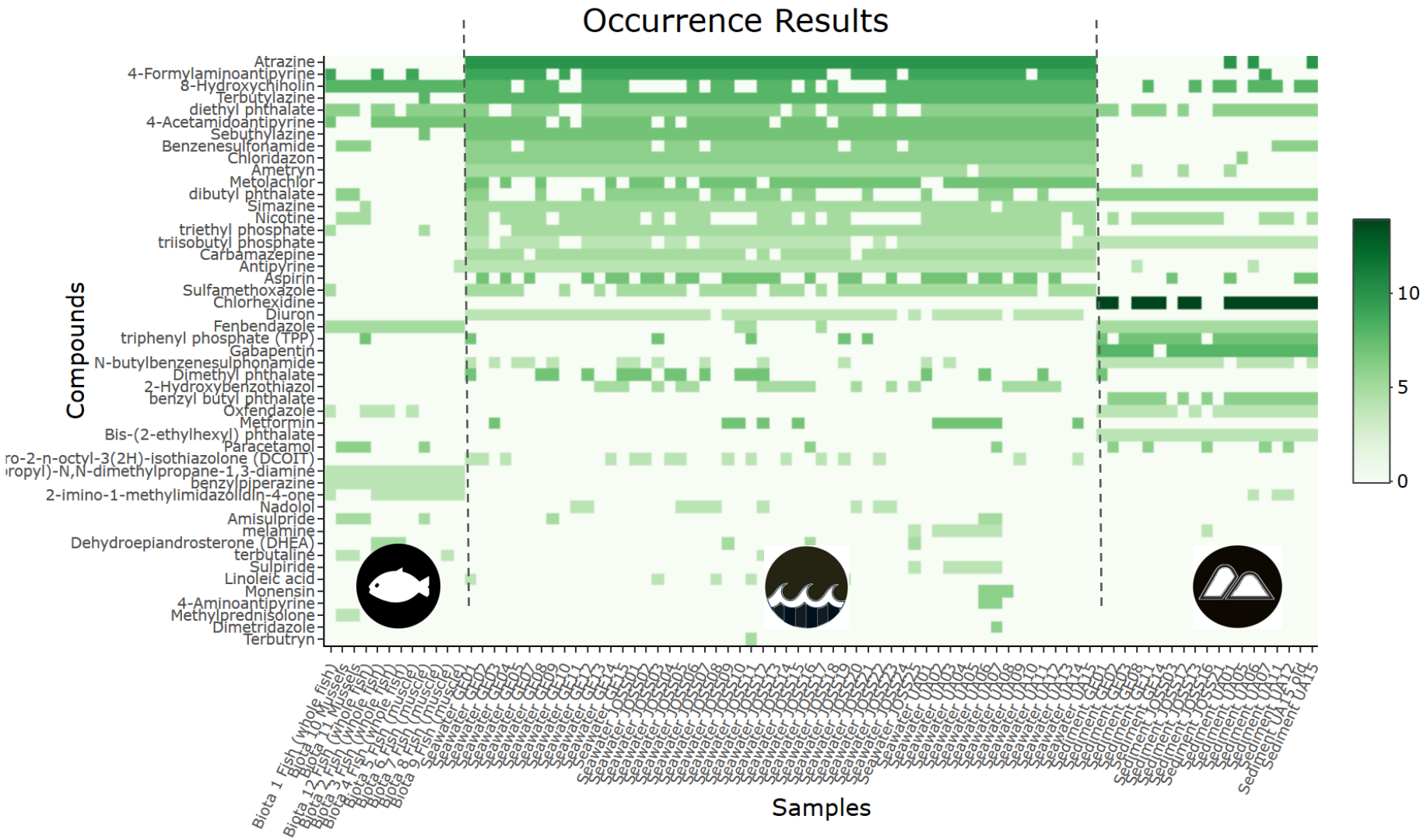
https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

Schymanski & Bolton (2021) FAIR Chemical Structures. *J. Cheminform*. DOI: 10.1186/s13321-021-00520-4

Schymanski & Bolton (2021) FAIR-ifying the Exposome: Preprint DOI: 10.5281/zenodo.5495109; Preproof: DOI: 10.1093/exposome/osab006

Examples: Distribution of Chemicals in Various Matrices

Retrospective screening of REACH chemicals in Black Sea samples

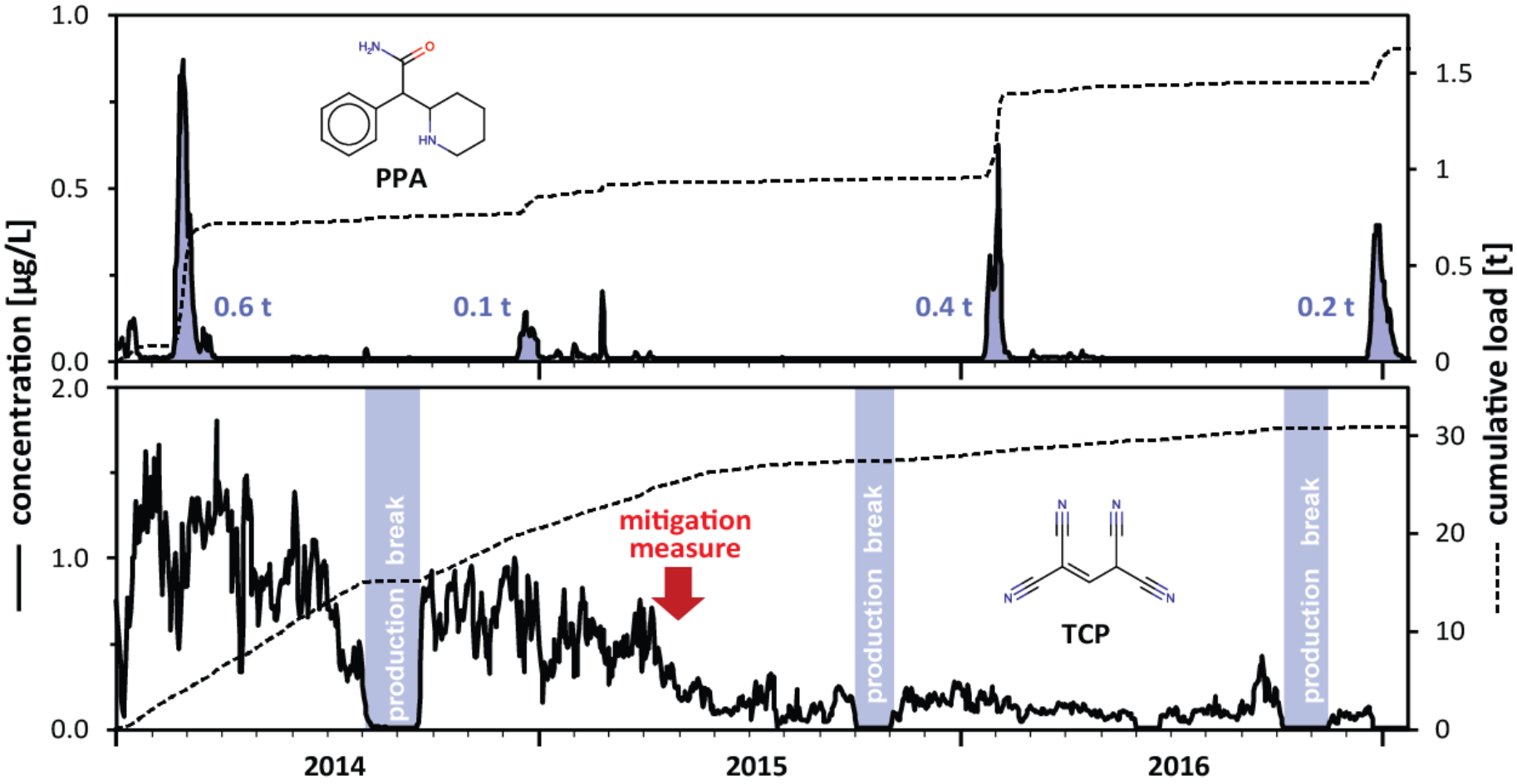


Alygizakis NA *et al* (2019) DOI: [10.1016/j.trac.2019.04.008](https://doi.org/10.1016/j.trac.2019.04.008)

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

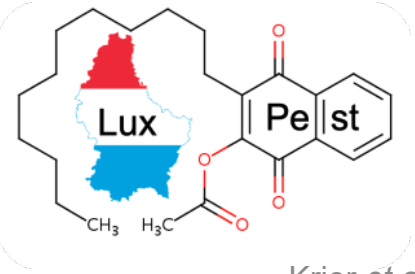
Examples: Real Time Monitoring of the Rhine River


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

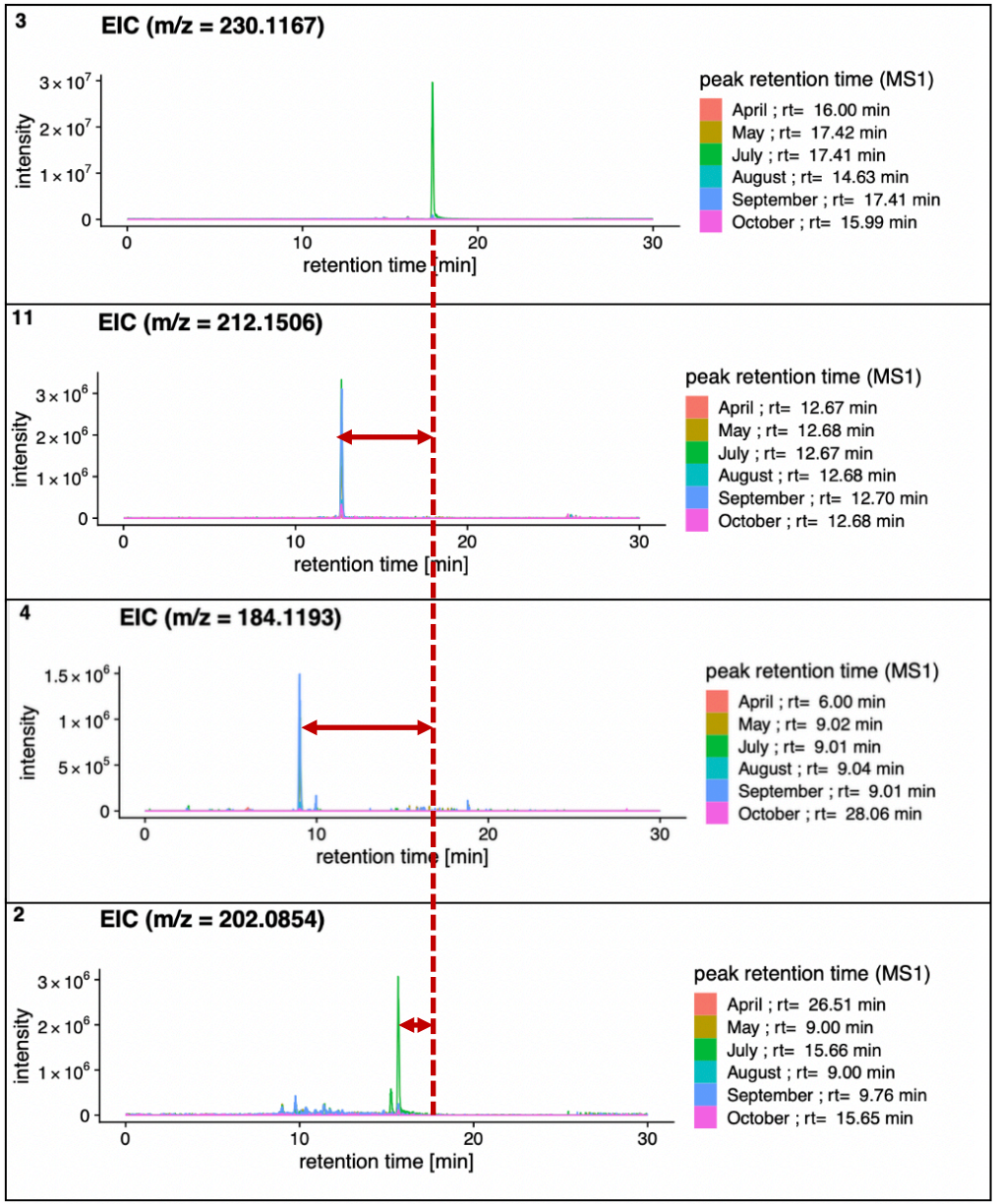


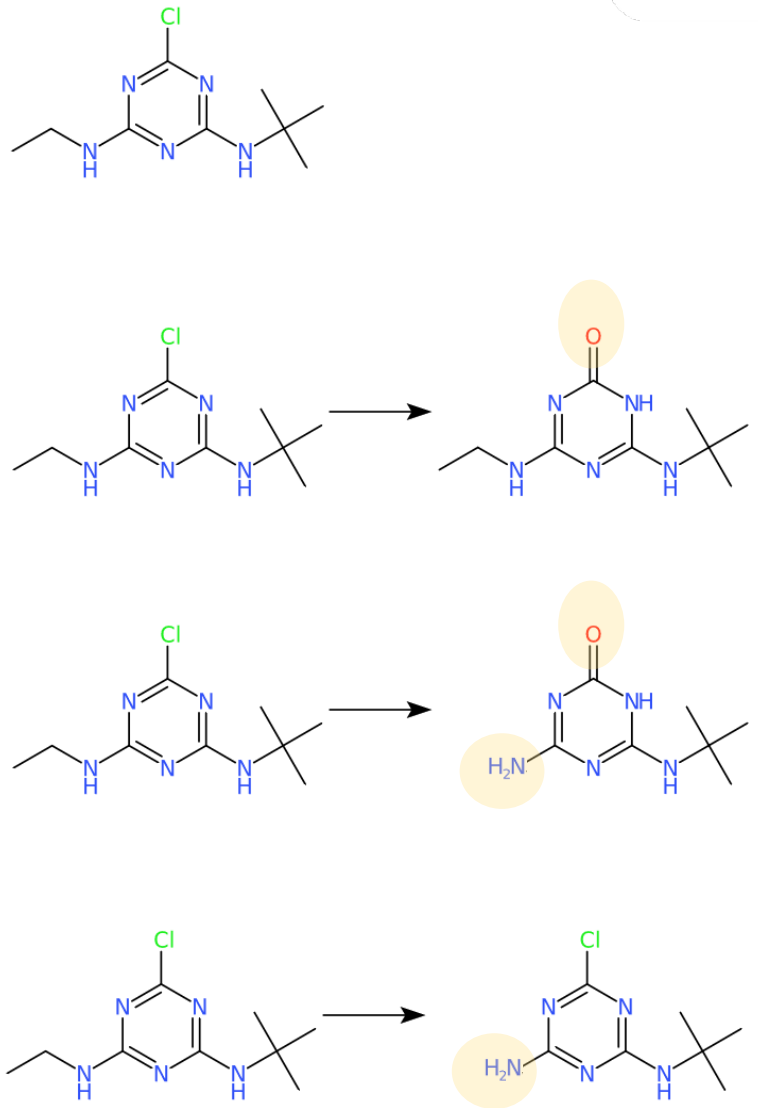
Hollender, Schymanski, Singer & Ferguson (2018) ES&T Feature, 51:20, 11505-11512. DOI: 10.1021/acs.est.7b02184

Examples: LuxPest and Transformation Products

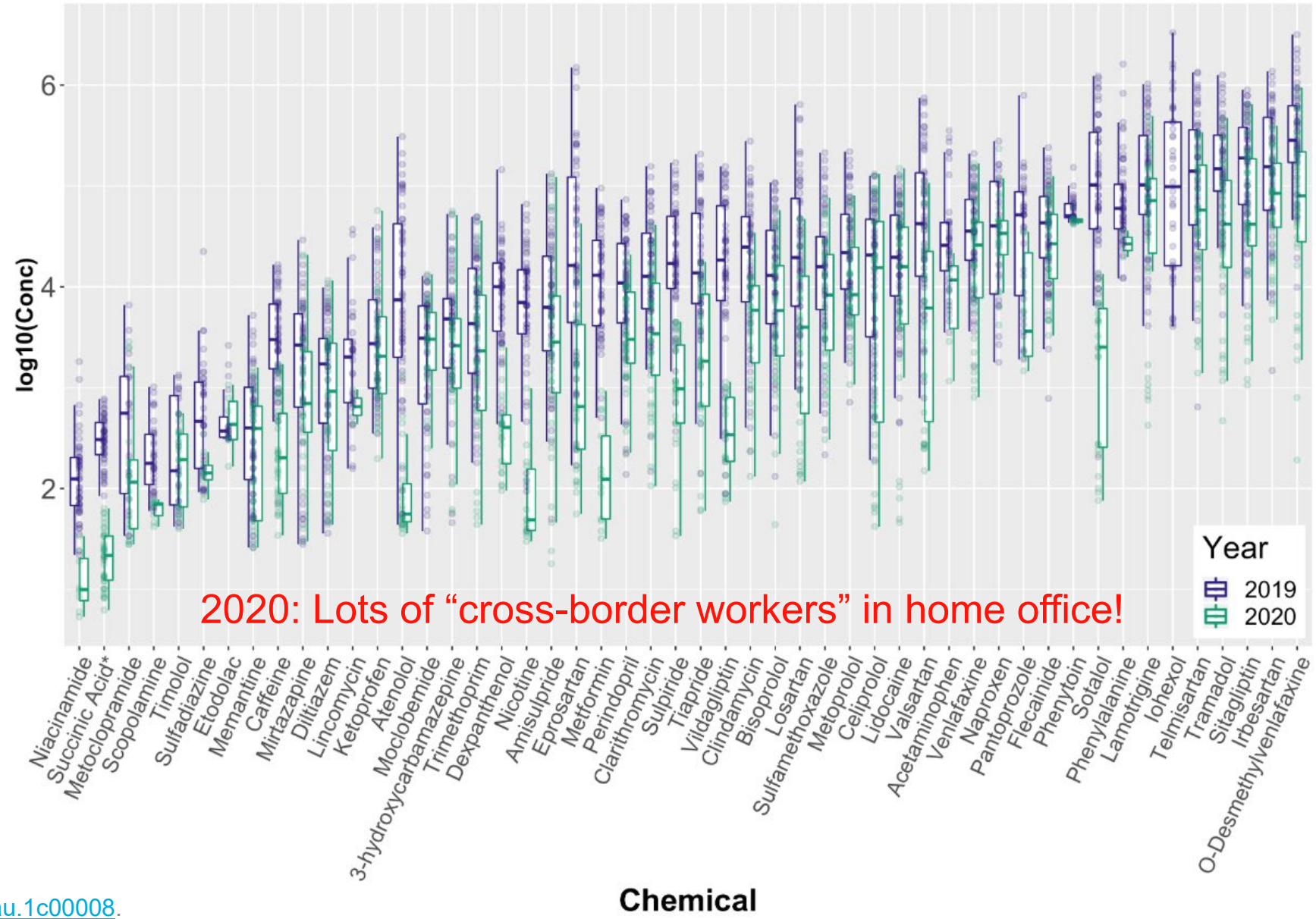


Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885





Examples: LuxPharma – 2019 versus 2020 & COVID?



Singh *et al* (2021).
DOI: [10.1021/acsenvironau.1c00008](https://doi.org/10.1021/acsenvironau.1c00008).

The Problem: Which chemicals are relevant? How to find them?

CAS
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

180 million

PubChem

111 million

CompTox Chemicals Dashboard

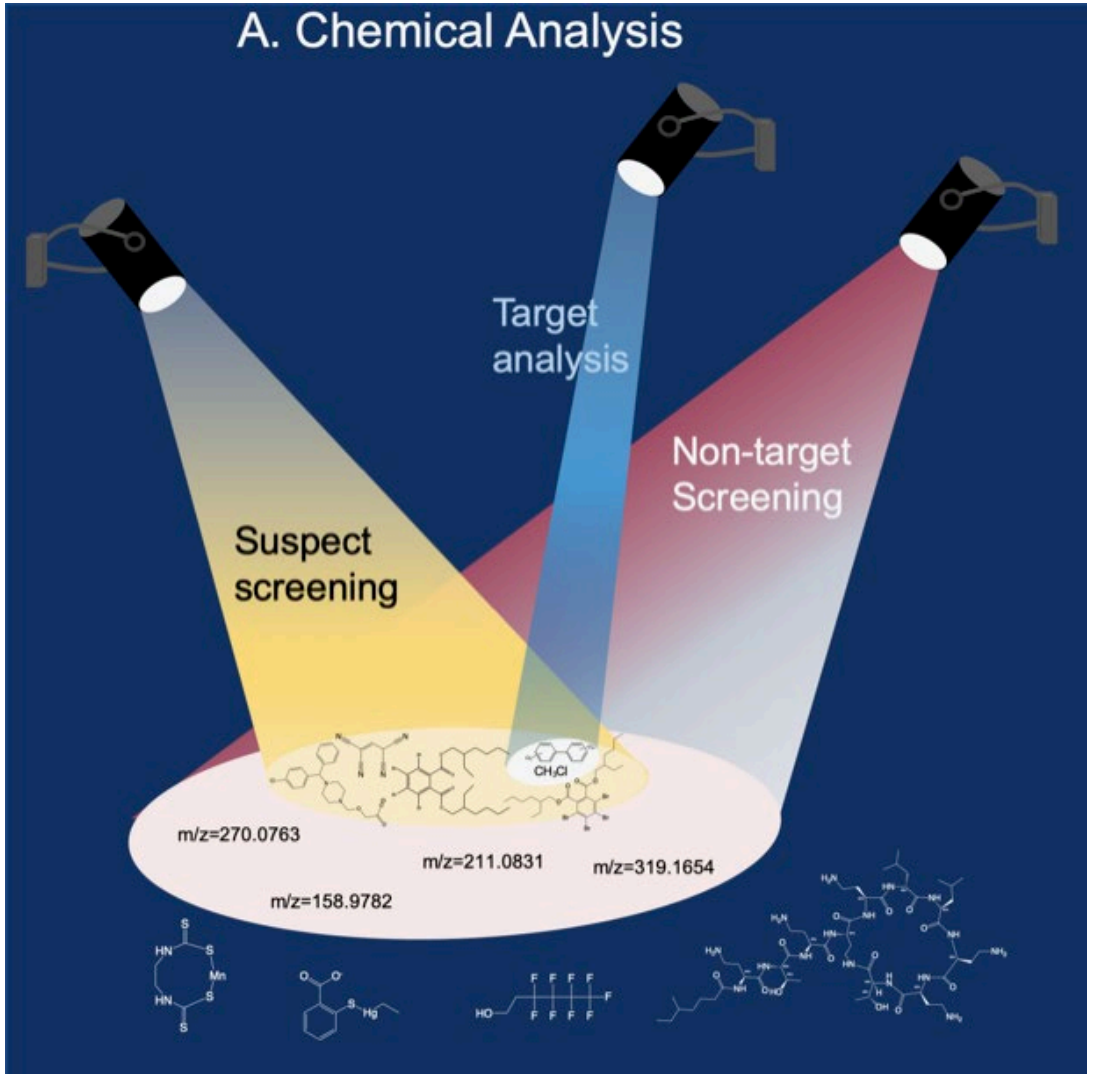
906,511

PubChemLite EXPOSOMICS

~370,000

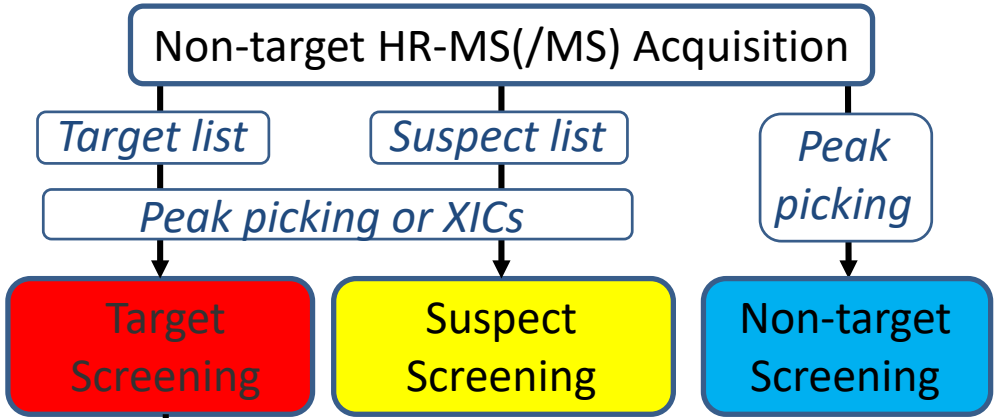
norman suspects

~113,000



Source: ESO/IDA/Danish 1.5 m http://www.eso.org/public/images/potw1015a/ Mod. from Escher *et al* (2020). Science. DOI: 10.1126/science.aay6636

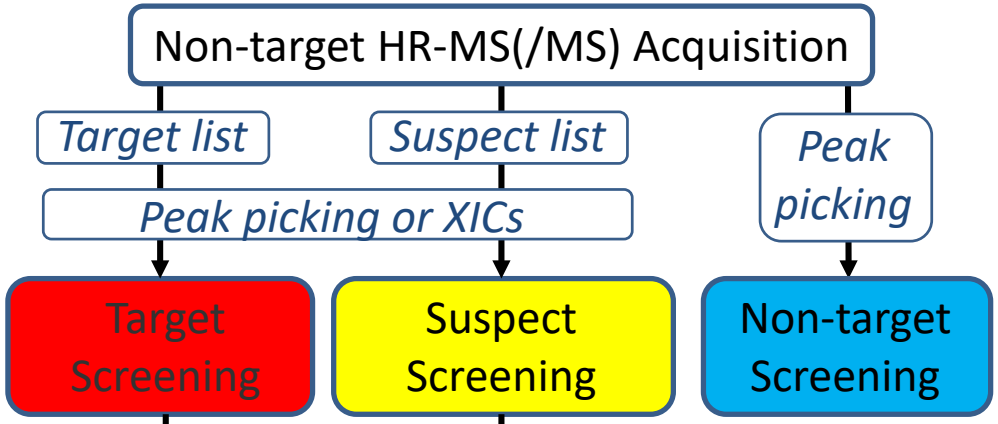
Identification Strategies and Confidence in NT-HRMS(/MS)



Limited by ref. std. availability …

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

Identification Strategies and Confidence in NT-HRMS(/MS)



Start			**Level 1**	**Confirmed Structure** by reference standard
			Level 2	**Probable Structure** by library/diagnostic evidence
	Start		**Level 3**	**Tentative Candidate(s)** suspect, substructure, class

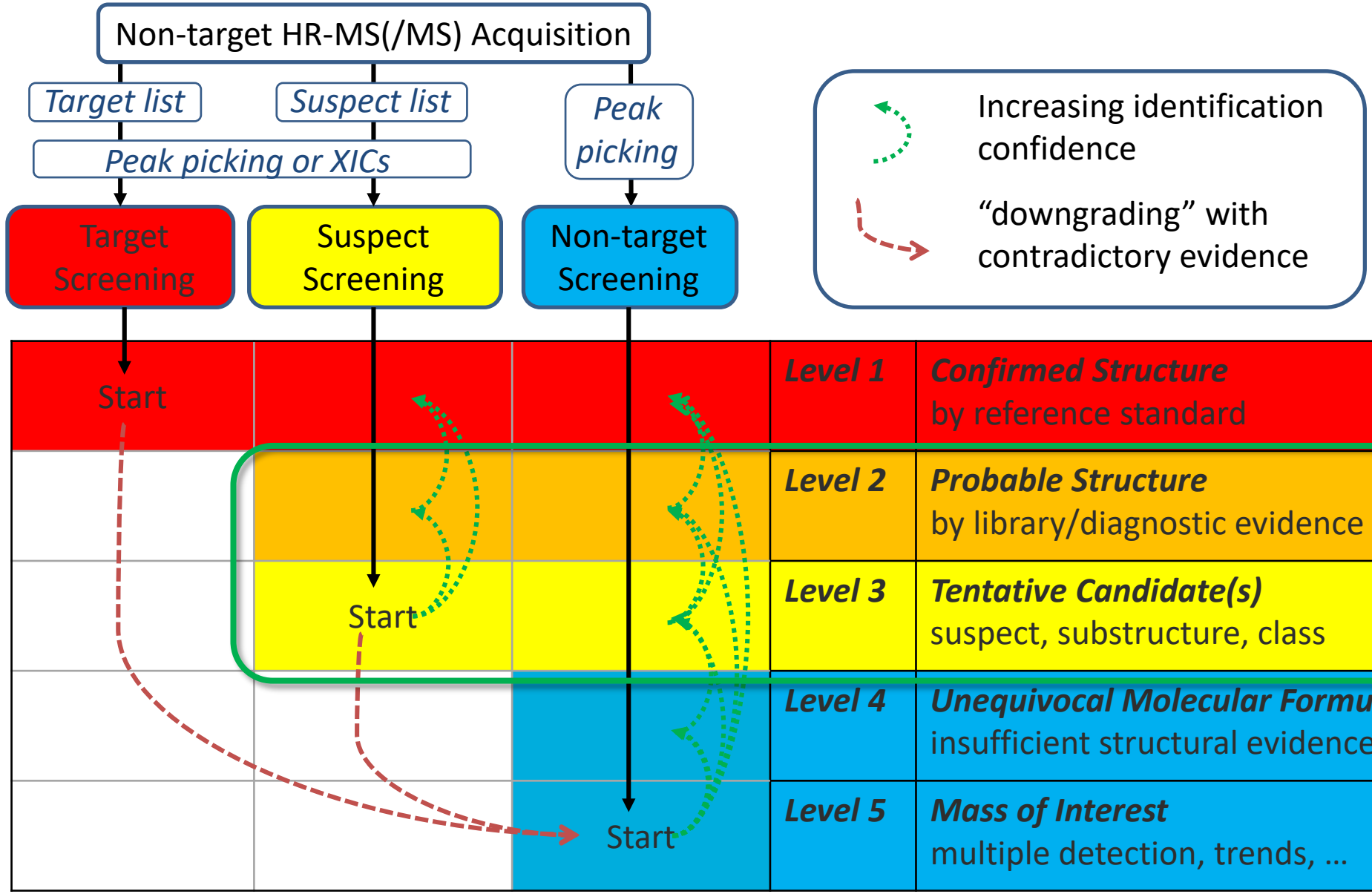
Limited by ref. std. availability …

MassBank
High Quality Mass Spectral Database

MetFrag

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

Identification Strategies and Confidence in NT-HRMS(/MS)



Limited by ref. std. availability …

MassBank
High Quality Mass Spectral Database

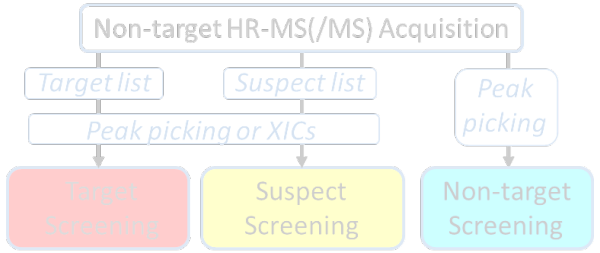
MetFrag

The "dark matter"

eawag aquatic research

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

Outline of Today



- Introduction and Background
- Identification & Chemical Space
  - Identification + MetFrag
  - PubChemLite for Exposomics
- Case Study: LuxPest
- Dark Matter and Transformations
- Take-home messages!

https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

Introduction to MetFrag

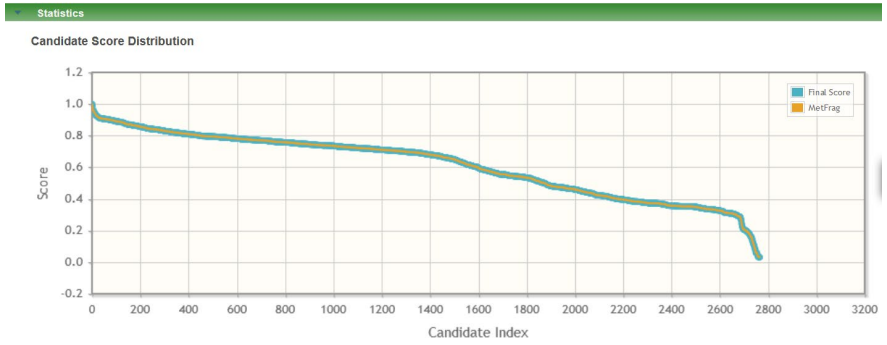
https://msbi.ipb-halle.de/MetFrag/

mz [M−H]⁻
213.9637
± 5 ppm

5 ppm
0.001 Da

PubChem

Ranked Candidates



MS/MS

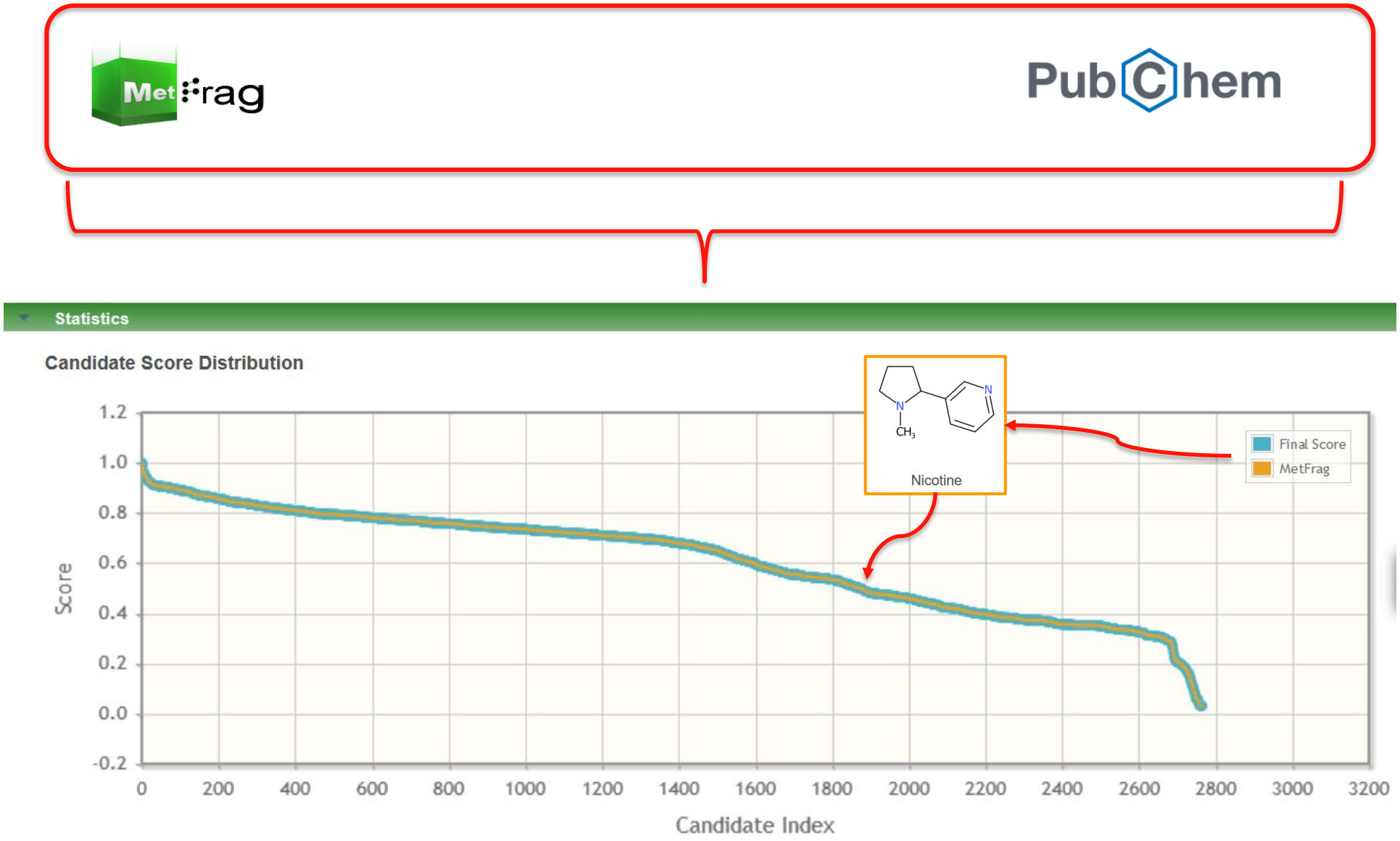
134.0054	339689
150.0001	77271
213.9607	632466

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

Key Challenge: MS and MS/MS alone is not enough!

Ruttkies, Schymanski *et al*. (2016) DOI: 10.1186/s13321-016-0115-9 Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

Key Challenge: MS and MS/MS alone is not enough!



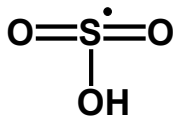
MetFrag + PubChem + Formula Search + https://massbank.eu/MassBank/RecordDisplay?id=EQ300804

Status Quo in 2016: MetFrag Relaunched …

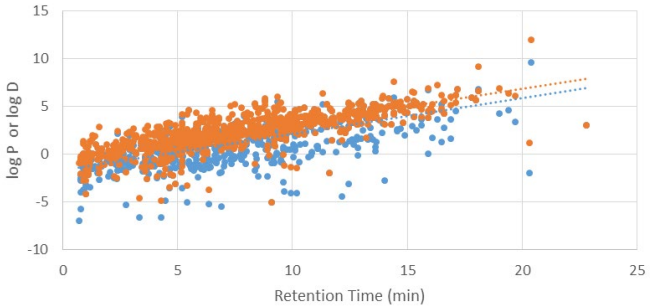
mz [M-H]⁻
213.9637
± 5 ppm

Elements: C,N,S

5 ppm
0.001 Da



RT: 4.54 min
355 InChI/RTs



CompTox Chemicals Dashboard
or
PubChem

References
Tox. Data
Data Sources
Exposure Info
MS-ready links

Literature Patents

Suspect Lists

?TOFF !DENT

norman suspects

CompTox Chemicals Dashboard

MoNA
MassBank of North America
MassBank
High Quality Mass Spectral Database

MS/MS

134.0054	339689
150.0001	77271
213.9607	632466



IPB

eawag aquatic research

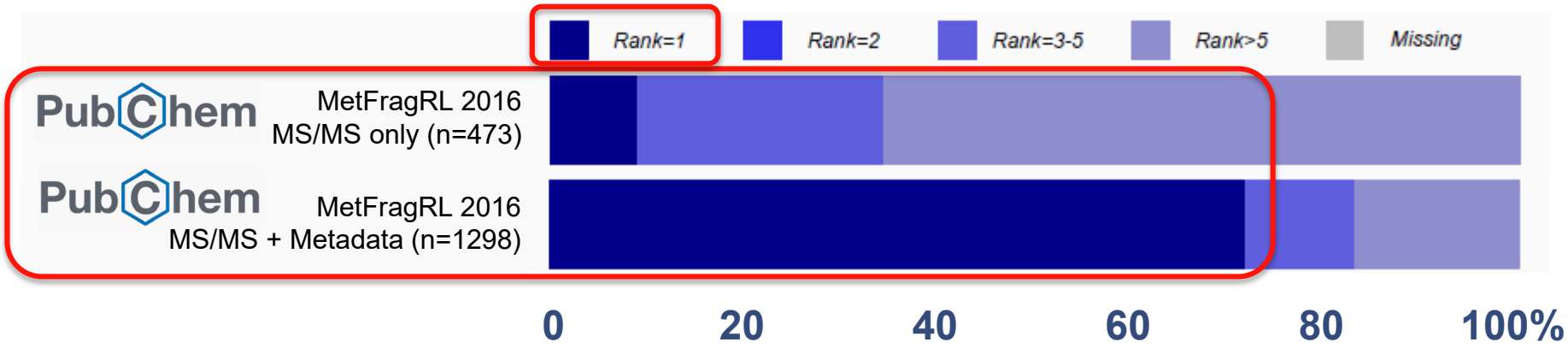
solutions

uni.lu | LCSB

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

MetFragRL + PubChem + MS/MS + Metadata

- Adding literature, references & RT boosts to ~71 % rank 1!

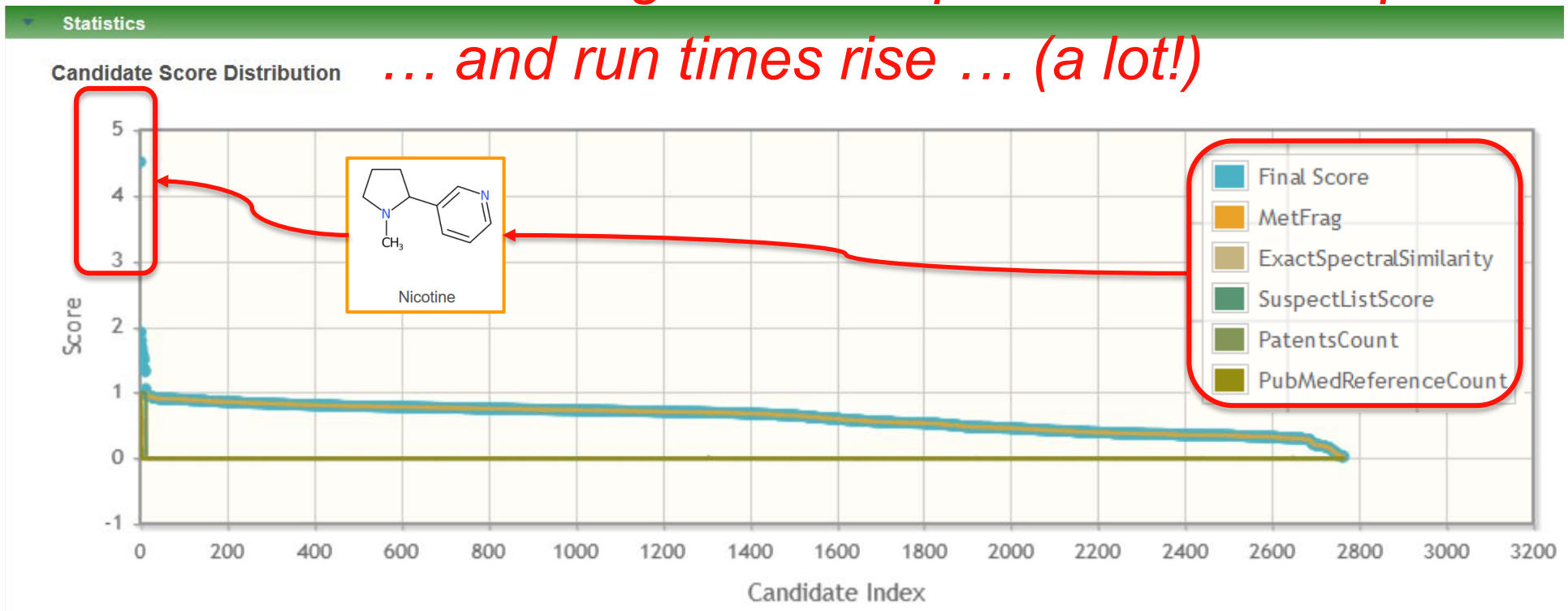


Ruttkies, Schymanski *et al*. (2016) DOI: 10.1186/s13321-016-0115-9 Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

MetFragRL + PubChem + MS/MS + Metadata

BUT …databases grow … ID performance drops

… and run times rise … (a lot!)



MetFrag + PubChem + Formula + MoNA + SusDat + Pat + Refs + https://massbank.eu/MassBank/RecordDisplay?id=EQ300804

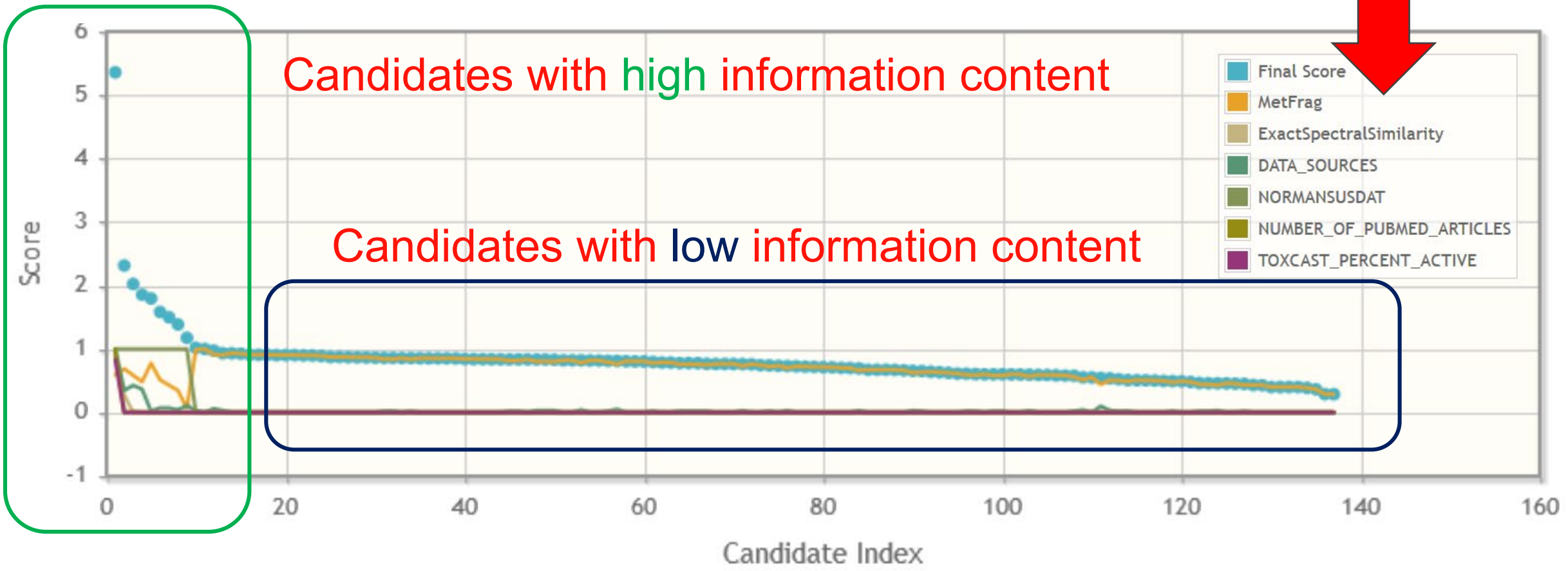
Problem: Exposomics “Chemical Space” is too big!

CAS A DIVISION OF THE AMERICAN CHEMICAL SOCIETY — 180 million

PubChem — 111 million

ChemSpider Search and share chemistry — 114 million

CompTox Chemicals Dashboard — 906,511



Can we break down PubChem into useful bits?

- 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

- Agrochemical Information 3,045
  - Agrochemical Category 1,903
  - Agrochemical Transformations 1,383
  - EU Pesticides Data 1,213
  - USDA Pesticide Data Program 579
- Use and Manufacturing 115,321
  - Uses 27,314
  - Consumption Patterns 1,168
  - Formulations/Preparations 4,742
  - General Manufacturing Information 52,353
  - Impurities 870
  - Methods of Manufacturing 6,192
  - Overview 1,179
  - Sampling Procedures 579
  - U.S. Exports 1,038
  - U.S. Imports 1,547
  - U.S. Production 5,500
  - Use Classification 63,558

PubChem Furathiocarb (Compound) CONTENTS

7 Agrochemical Information

7.1 Agrochemical Category

Insecticides

- EU Pesticides Database

7.2 Agrochemical Transformations

Furathiocarb has known environmental transformation products that include carbofuran.

S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019 | DOI:10.5281/zenodo.3544759

- NORMAN Suspect List Exchange

7.3 EU Pesticides Data

Active Substance	furathiocarb
Status	Not Approved [Reg. (EC) No 1107/2009]
Categories	Insecticides

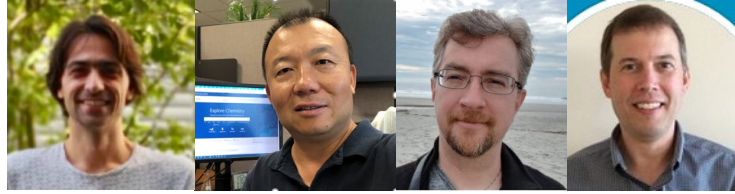
https://pubchem.ncbi.nlm.nih.gov/compound/Furathiocarb#section=Agrochemical-Information

Schymanski *et al.* (2021) DOI: 10.1186/s13321-021-00489-0

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

LCSB

Introducing …

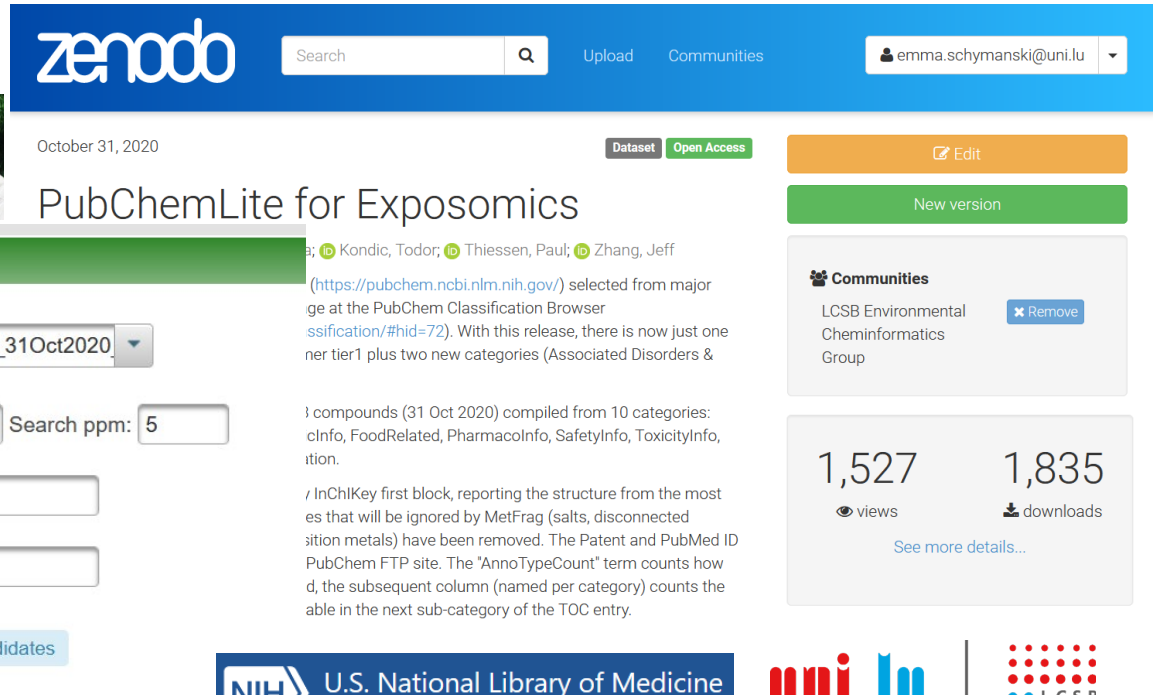


- PubChem Compound TOC ? 49,493,641
  - Agrochemical Information ? 3,045
  - Associated Disorders and Diseases ? 20,847
  - Biologic Description ? 2,056,521
  - Biological Test Results ? 3,622,920
  - Biomolecular Interactions and Pathways ? 125,253
  - Chemical and Physical Properties ? 263,015
  - 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

PubChemLite EXPOSOMICS

~370,000 entries “small”

R3



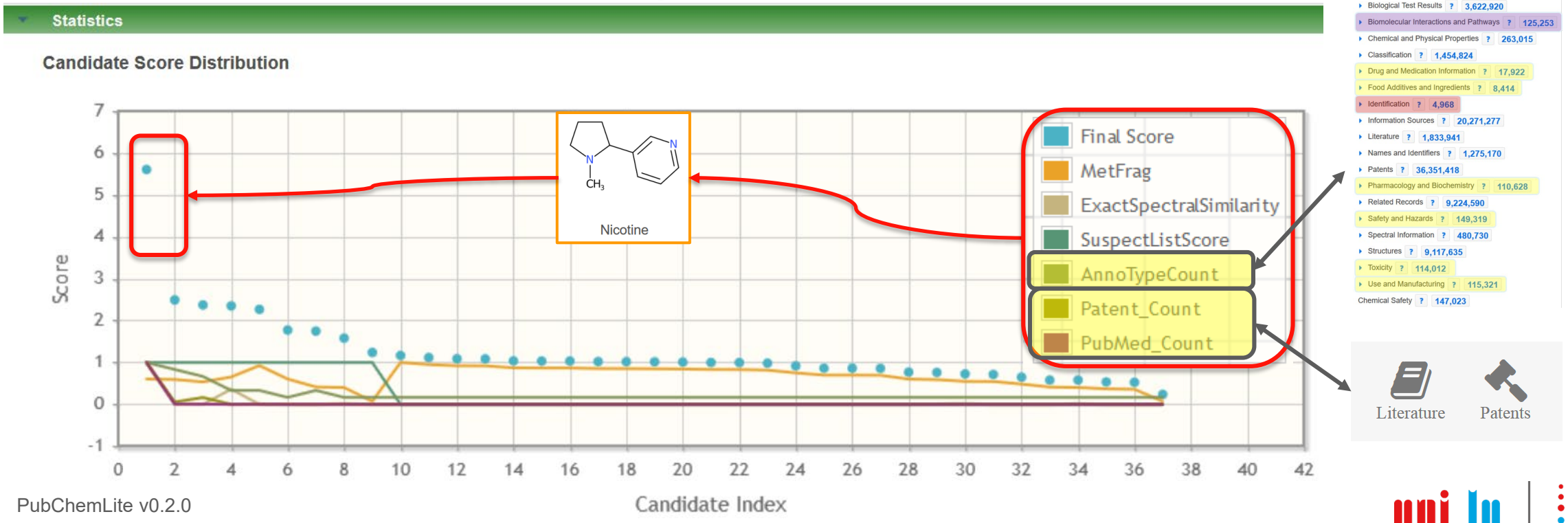
Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

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

uni.lu | LCSB

PubChemLite: Fewer and more relevant candidates – with context

MetFrag · MassBank High Quality Mass Spectral Database · norman suspects · PubChemLite EXPOSOMICS

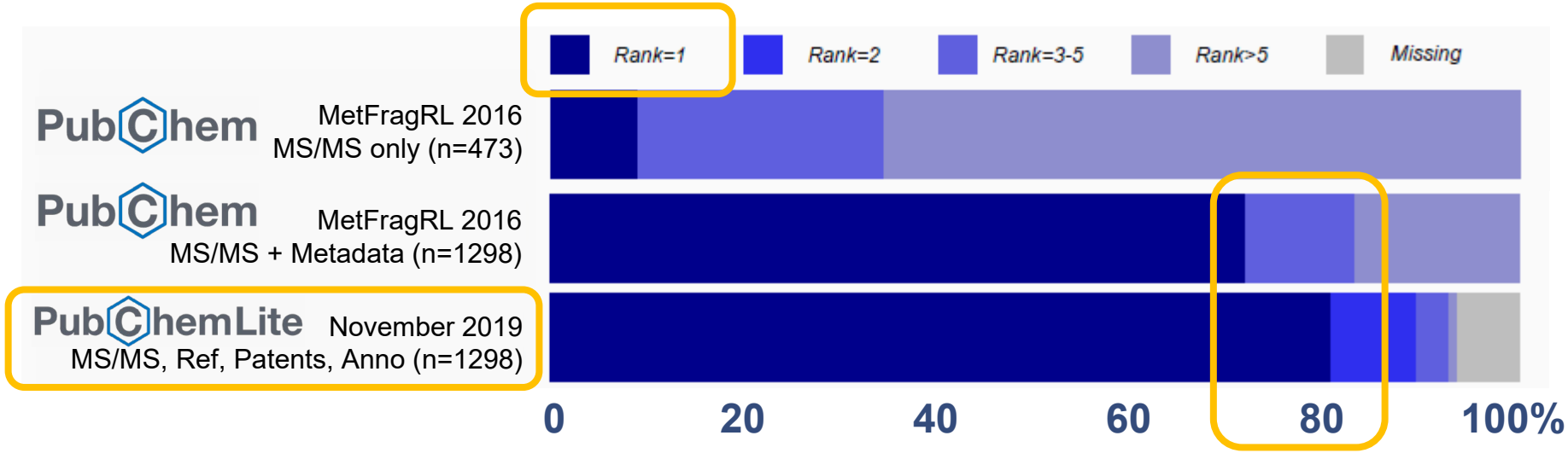


PubChemLite v0.2.0

MetFrag+PubChemLite+Formula+MoNA+SusDat+Pat+Refs+Anno + https://massbank.eu/MassBank/RecordDisplay?id=EQ300804

How does PubChemLite perform?

- ~110 M => ~370 K … how does this influence performance?



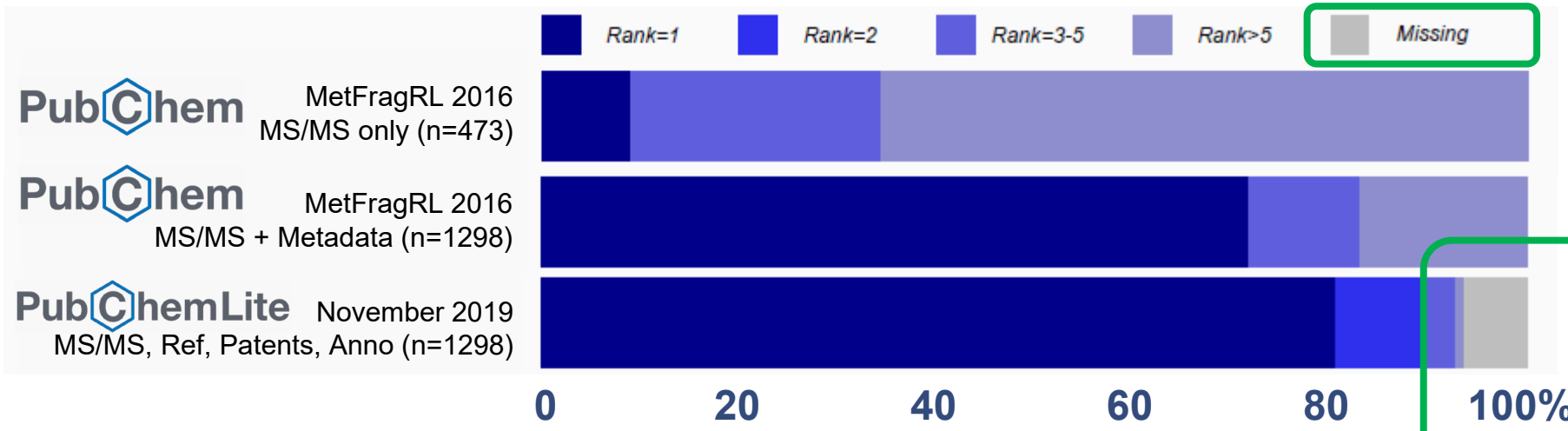
>80 % ranked in first place
~90 % ranked first or second!

} high throughput identification

Ruttkies, Schymanski *et al*. (2016) DOI: 10.1186/s13321-016-0115-9 Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

How does PubChemLite perform?

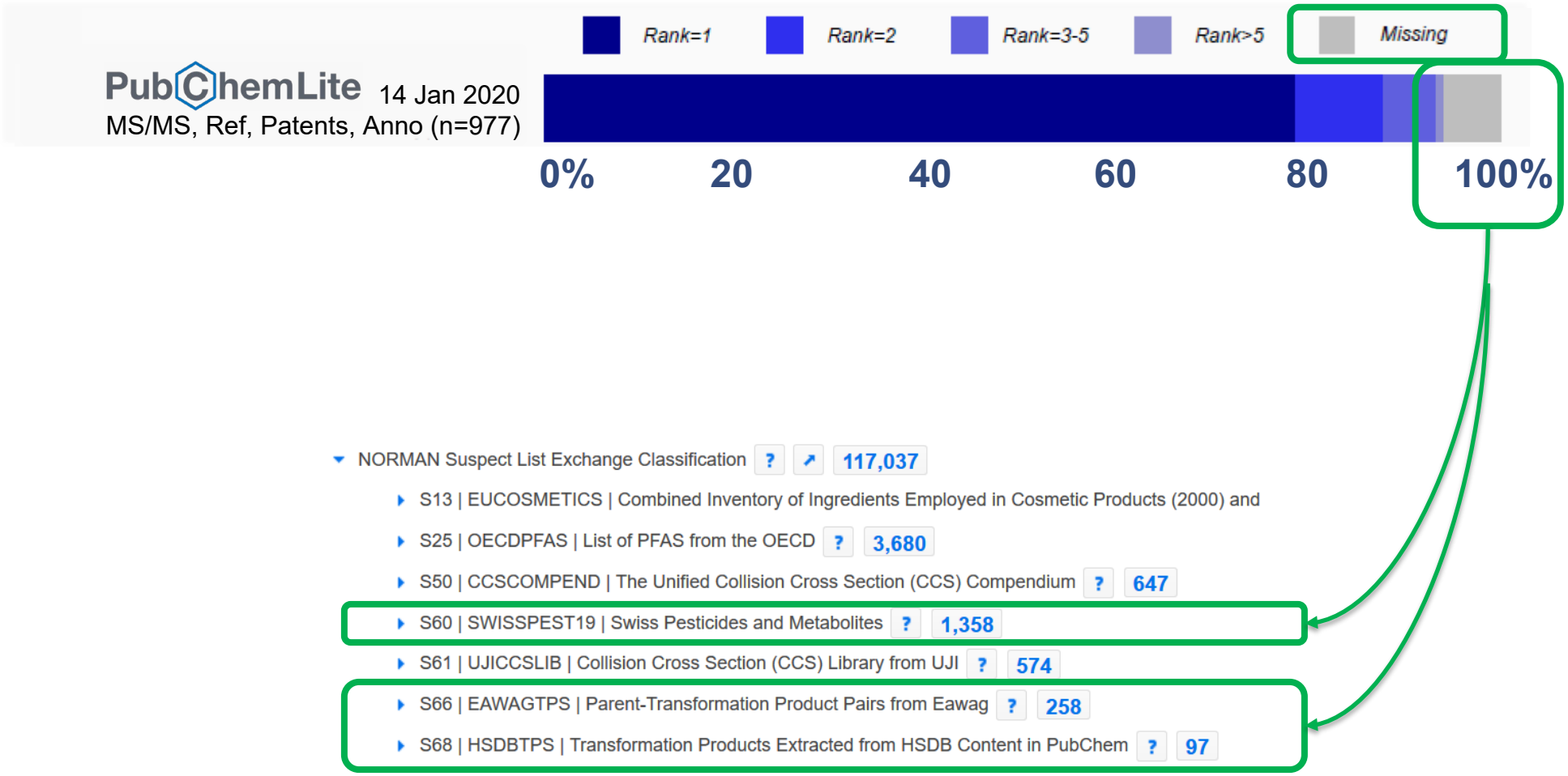
- ~110 M => ~370 K … how does this influence performance?



"Transformation products are missing from databases"
[General complaint; environmental community]

Ruttkies, Schymanski *et al*. (2016) DOI: 10.1186/s13321-016-0115-9 Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

Assessing the Missing Entries in PubChemLite



Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

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

Transformation Products: Filling the Data Gaps!

PubChem NORMAN Suspect List Exchange

- NORMAN Suspect List Exchange Classification ? 113,080
  - S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) ? 3,85
  - S25 | OECDPFAS | List of PFAS from the OECD ? 3,677
  - S36 | UBAPMT | Potential Persistent, Mobile and Toxic (PMT) substances ? 254
  - S50 | CCSCOMPEND | The Unified Collision Cross Section (CCS) Compendium ? 885
  - S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019 ? 1,343
  - S61 | UJICCSLIB | Collision Cross Section (CCS) Library from UJI ? 574
  - S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag ? 258
  - S68 | HSDBTPS | Transformation Products Extracted from HSDB Content in PubChem ? 102
  - S69 | LUXPEST | Pesticide Screening List for Luxembourg ? 386
  - S72 | NTUPHTW | Pharmaceutically Active Substances from National Taiwan University ? 1,068
  - S75 | CyanoMetDB | Comprehensive database of secondary metabolites from cyanobacteria ? 2,088
  - S00 | SUSDAT | Merged NORMAN Suspect List: SusDat ? 99,130
  - S01 | MASSBANK | NORMAN Compounds in MassBank EU ? 7,164
  - S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances ? 11,261
  - S03 | NORMANCT15 | NORMAN Collaborative Trial Targets and Suspects ? 624
  - S04 | UJIBADE | Target List from UJI used in Bade et al 2015 ? 542

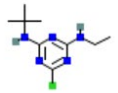
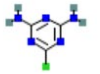
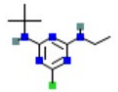
- Pharmacology and Biochemistry ? 112,039
  - Human Metabolite Information ? 64,199
  - Metabolism/Metabolites ? 8,204
  - Transformations ? 5,857

PubChem Terbuthylazine (Compound)

8.5 Transformations

Page 3 of 25 items View More Rows & Details Download

SORT BY Please Choose One

Predecessor Image	Predecessor Name	Transformation	Successor Image	Successor Name	Evidence DOI
	Terbuthylazine	Mammalian metabolism		6-Chloro-1,3,5-triazine-2,4-diamine	10.5281/zenodo.382
	Terbutylazine	Deethylation		Terbutylazine-desethyl	10.1007/s13361-017-

Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

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

Transformation Products: Filling the Data Gaps!

PubChem Terbuthylazine (Compound)

7 Agrochemical Information

7.1 Agrochemical Category

Pesticides -> Herbicides -> Triazine herbicides -> Chlorotriazine herbicides

S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag | DOI:10.5281/zenodo.3754448

▸ NORMAN Suspect List Exchange

7.2 Agrochemical Transformations

Terbutylazine has known environmental transformation products that include Terbutylazine-2-hydroxy, Terbutylazine-desethyl, and Terbutylazine-desethyl-2-hydroxy.

S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag | DOI:10.5281/zenodo.3754448

▸ NORMAN Suspect List Exchange

Terbutylazine has known environmental transformation products that include CSAA036479, CSAA04949, CSCD648241, CSCD692760, GS31398, MT1, GS 26379, MT13, GS 23158, Terbutylazine metabolite MT14, Terbutylazine metabolite MT23, and Terbutylazine metabolite MT24.

S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019 | DOI:10.5281/zenodo.3544759

▸ NORMAN Suspect List Exchange

▾ Agrochemical Information ? 3,045

Agrochemical Category ? 1,903

Agrochemical Transformations ? 1,383

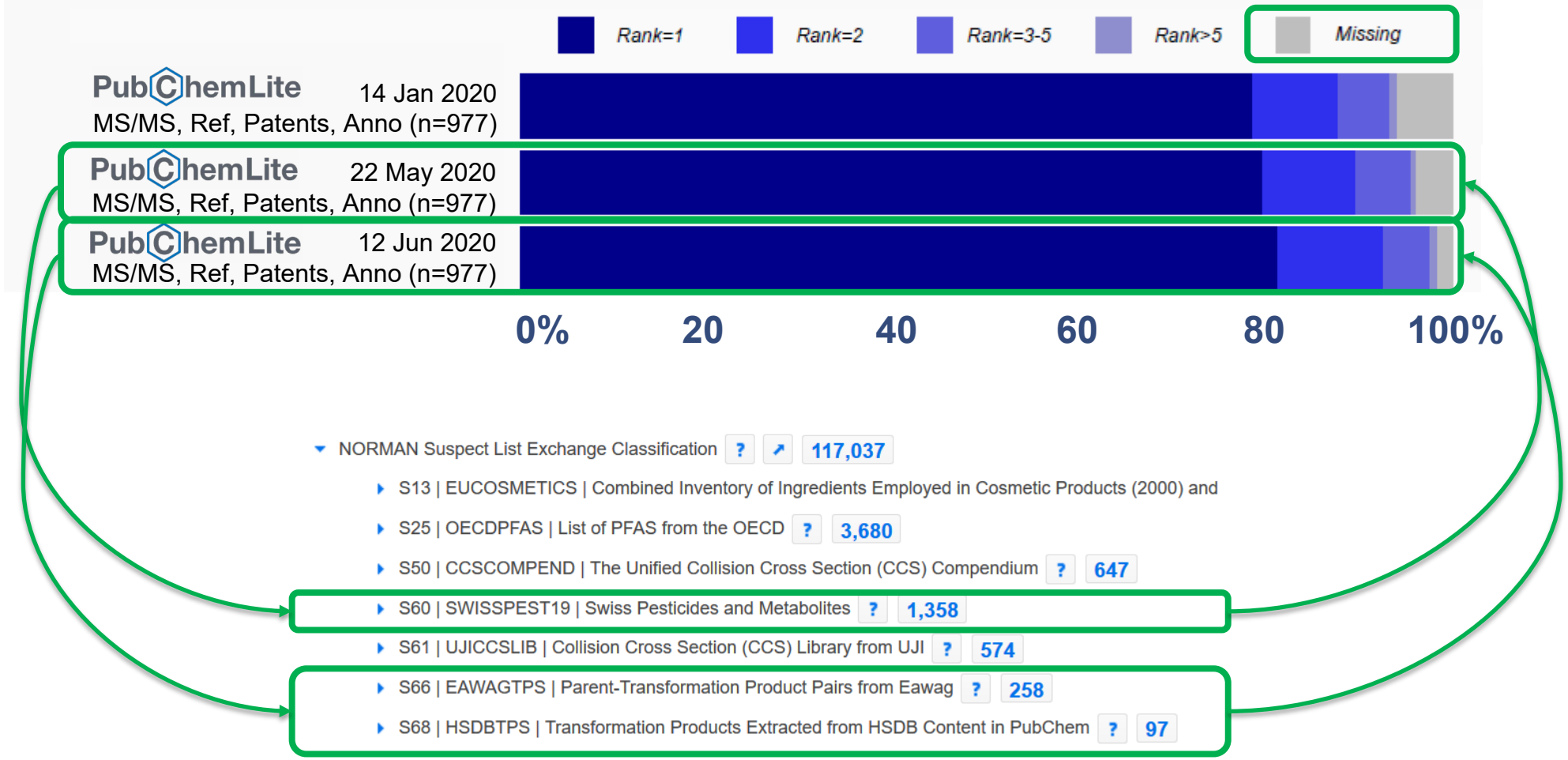
EU Pesticides Data ? 1,213

USDA Pesticide Data Program ? 579

PubChemLite EXPOSOMICS

Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

Assessing the Missing Entries in PubChemLite

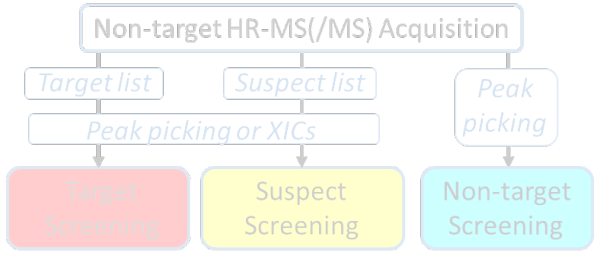


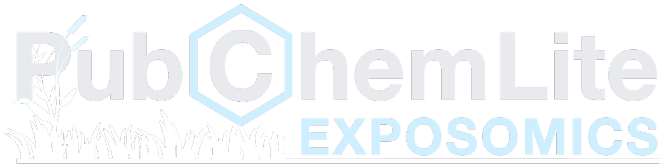
Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

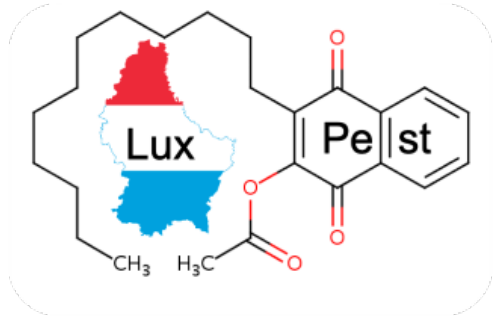
U.S. National Library of Medicine
National Center for Biotechnology Information

Outline of Today

- Introduction and Background
- Identification & Chemical Space
  - Identification + MetFrag
  - PubChemLite for Exposomics
- Case Study: LuxPest
- Dark Matter and Transformations
- Take-home messages!

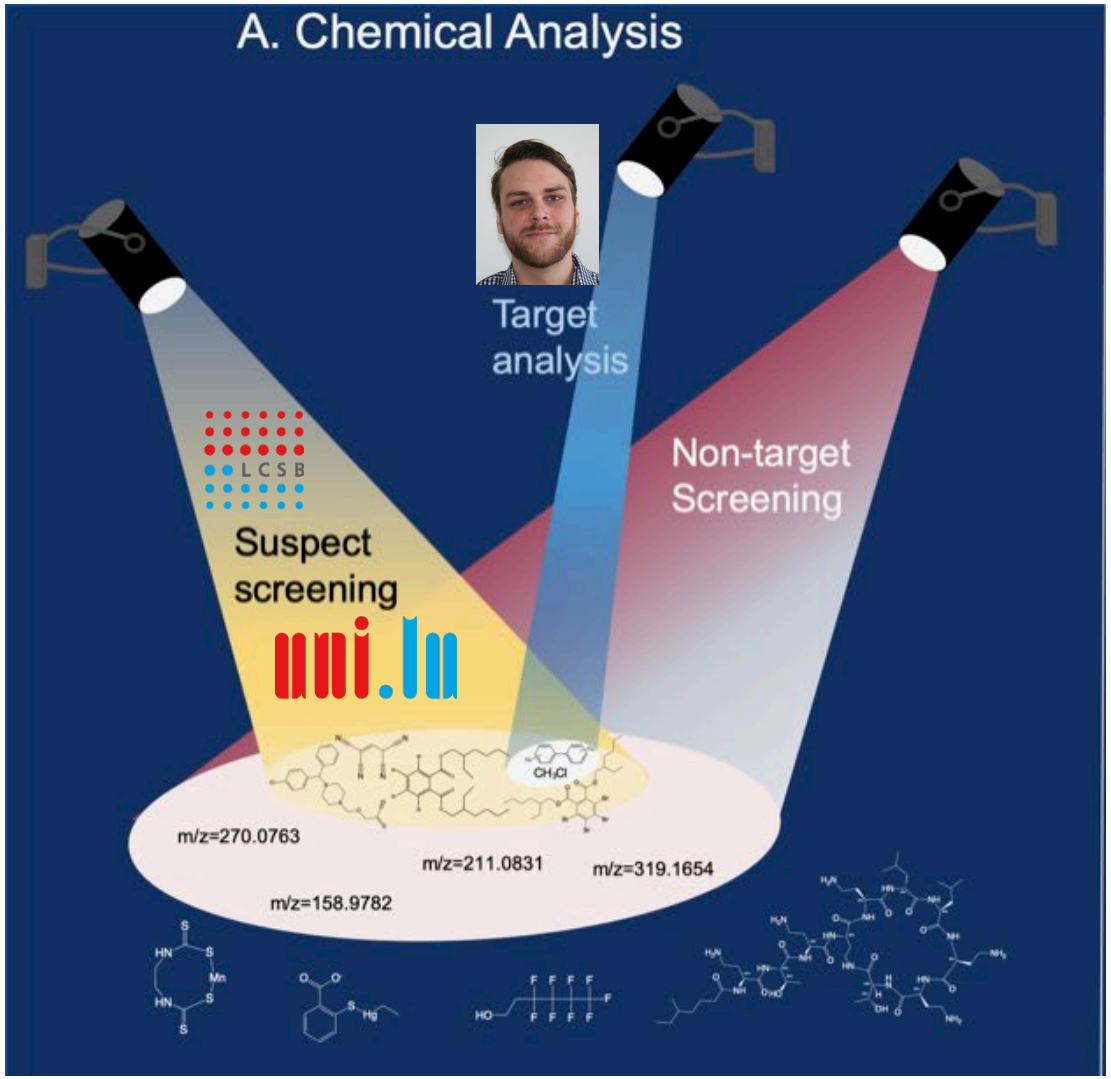






https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

The Problem: Which chemicals are relevant? How to find them?

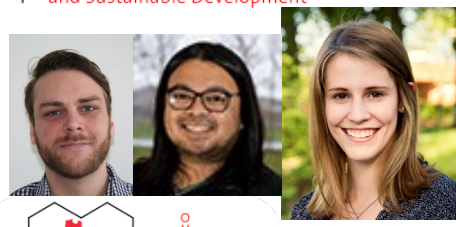


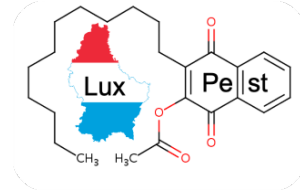
Monthly sampling
8-9 locations / year
4 fixed, 4-5 rotate
3 year cycle

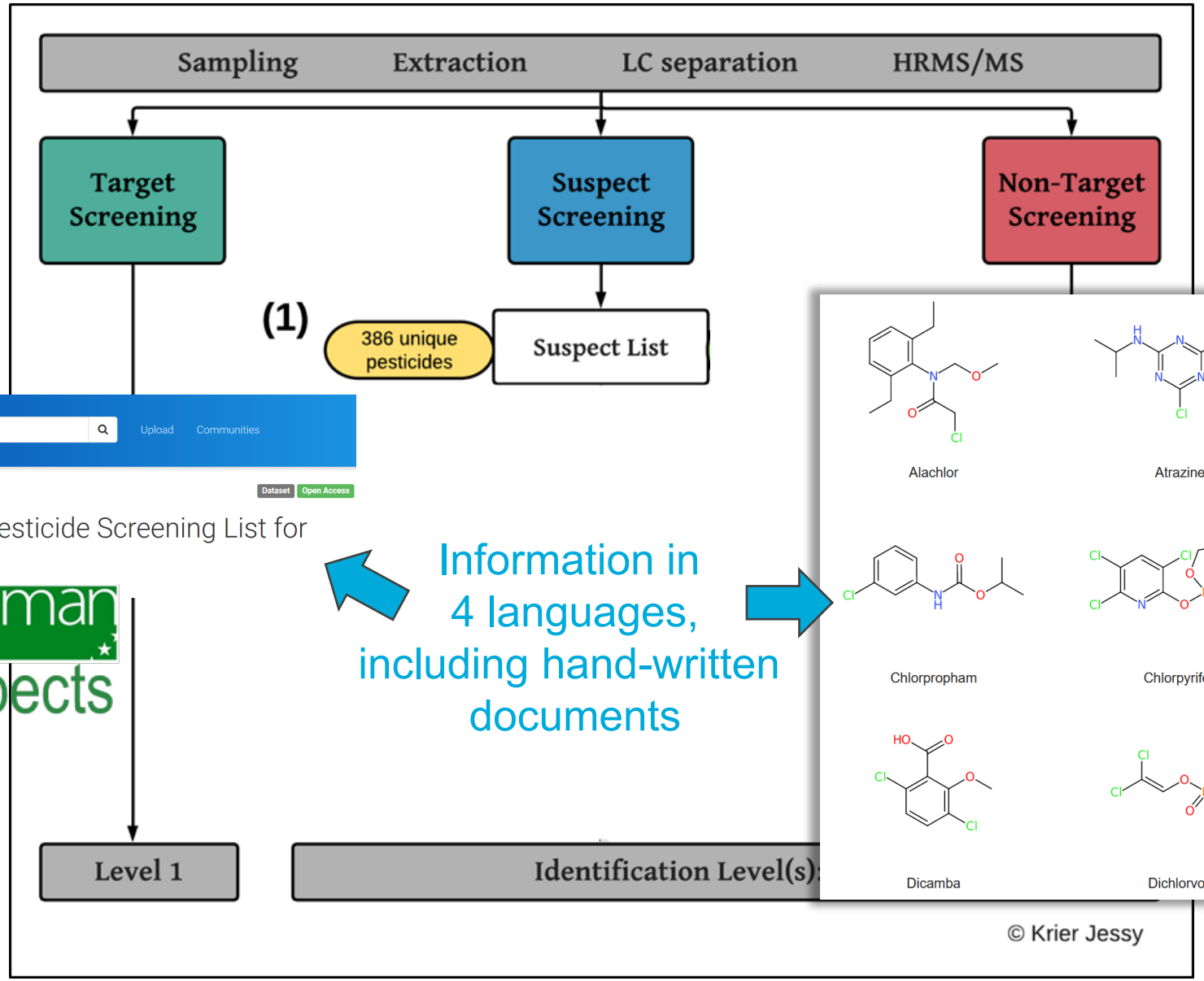
Modified from Escher, Stapleton, Schymanski (2020). Science. DOI: 10.1126/science.aay6636. Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885.

LuxPest – Suspect List Generation

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







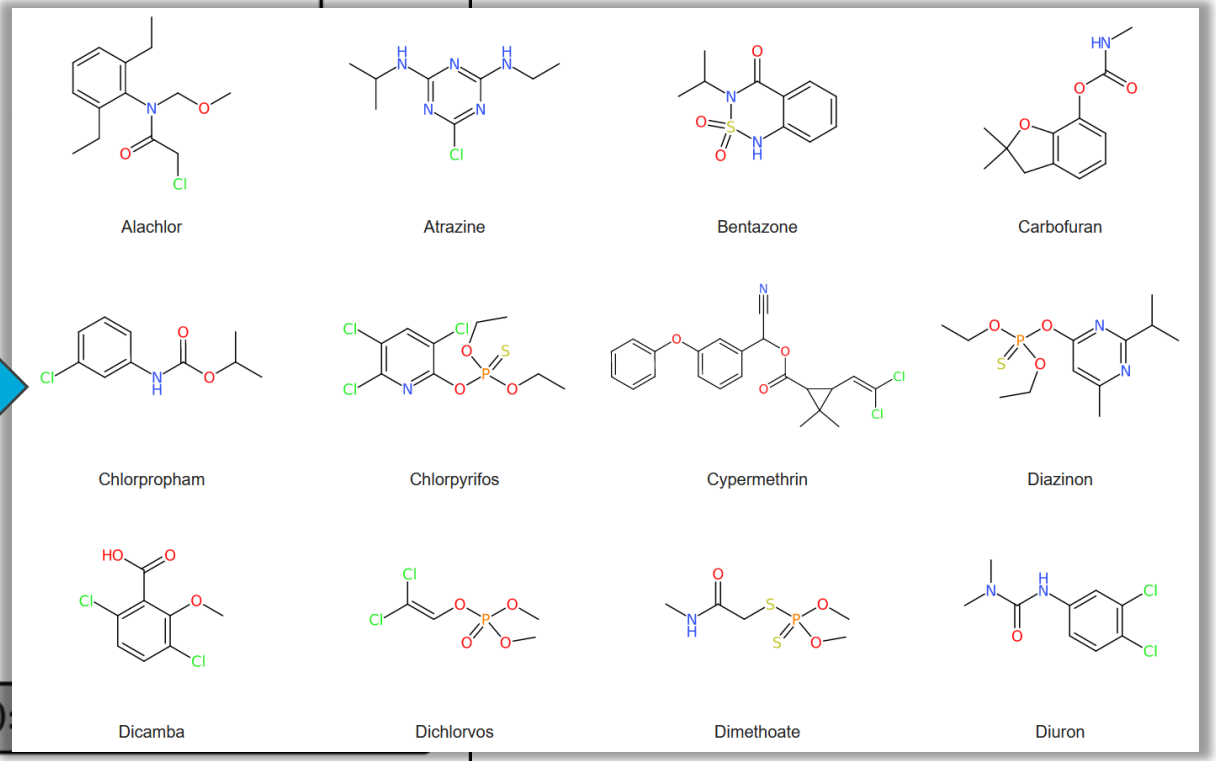
May 28, 2020 Dataset Open Access

S69 | LUXPEST | Pesticide Screening List for Luxembourg

Krier, Jessy

norman suspects

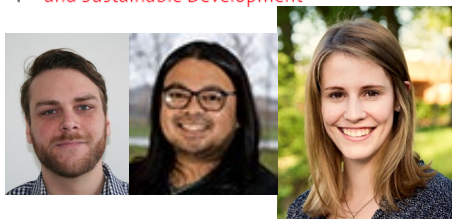
Information in 4 languages, including hand-written documents

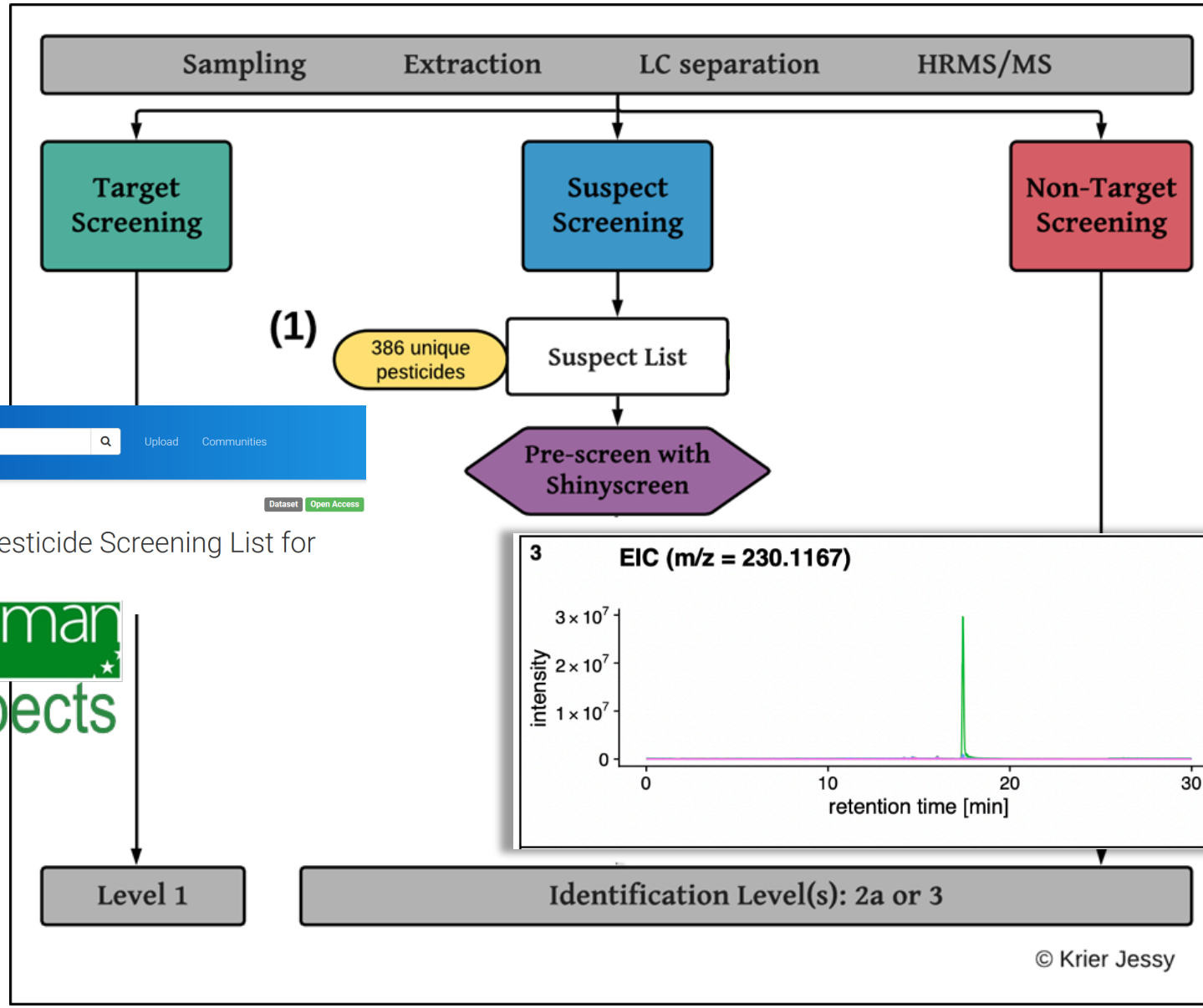


Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885. Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

LuxPest – Pre-screening (+QC) with Shinyscreen

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





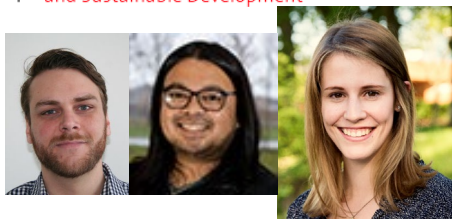
S69 | LUXPEST | Pesticide Screening List for Luxembourg

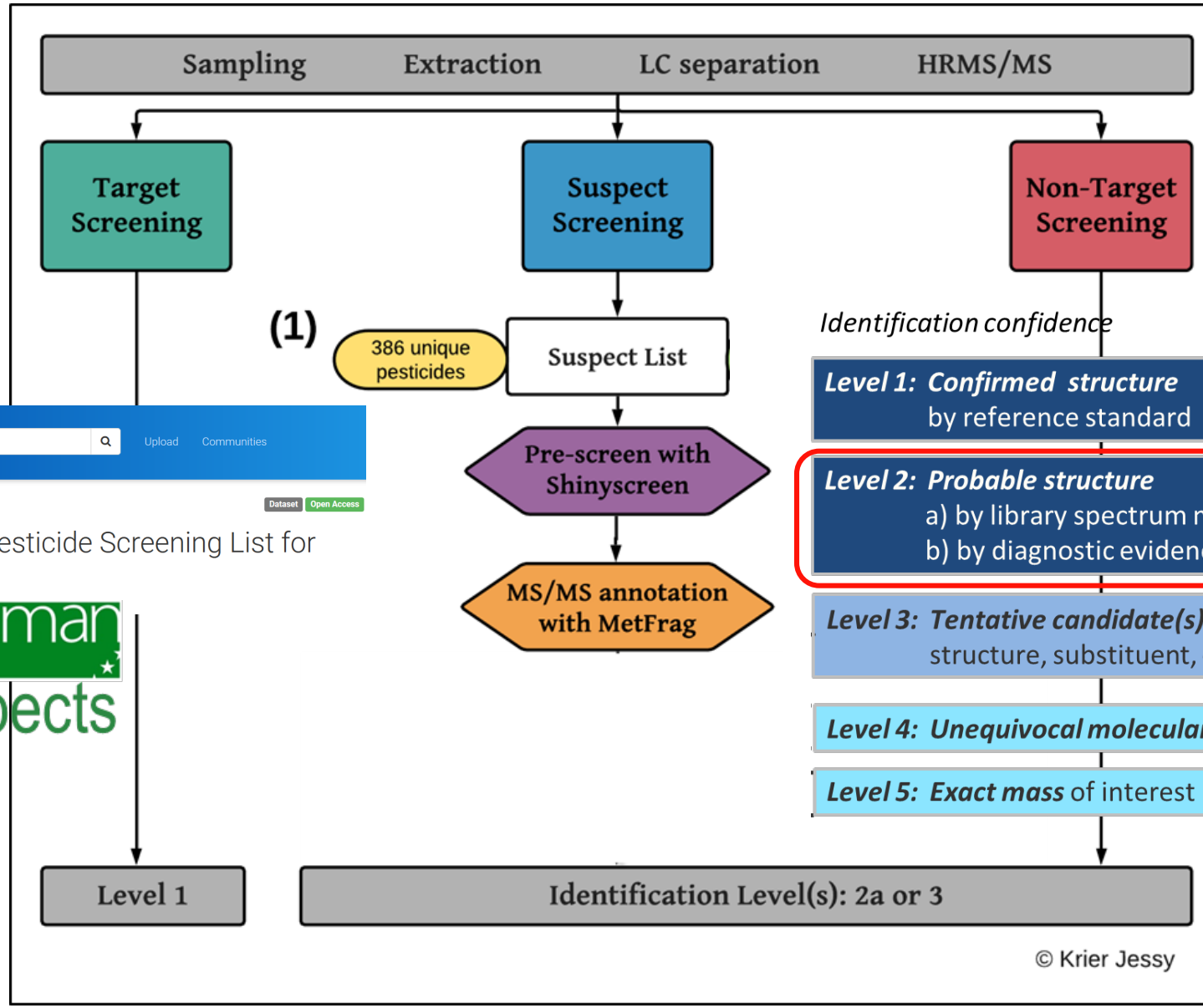
Krier, Jessy

Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885. Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

LuxPest – MS/MS Annotation with MetFrag

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





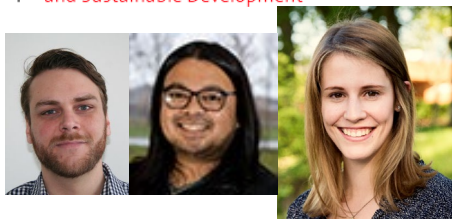
Identification confidence | *Minimum data requirements*

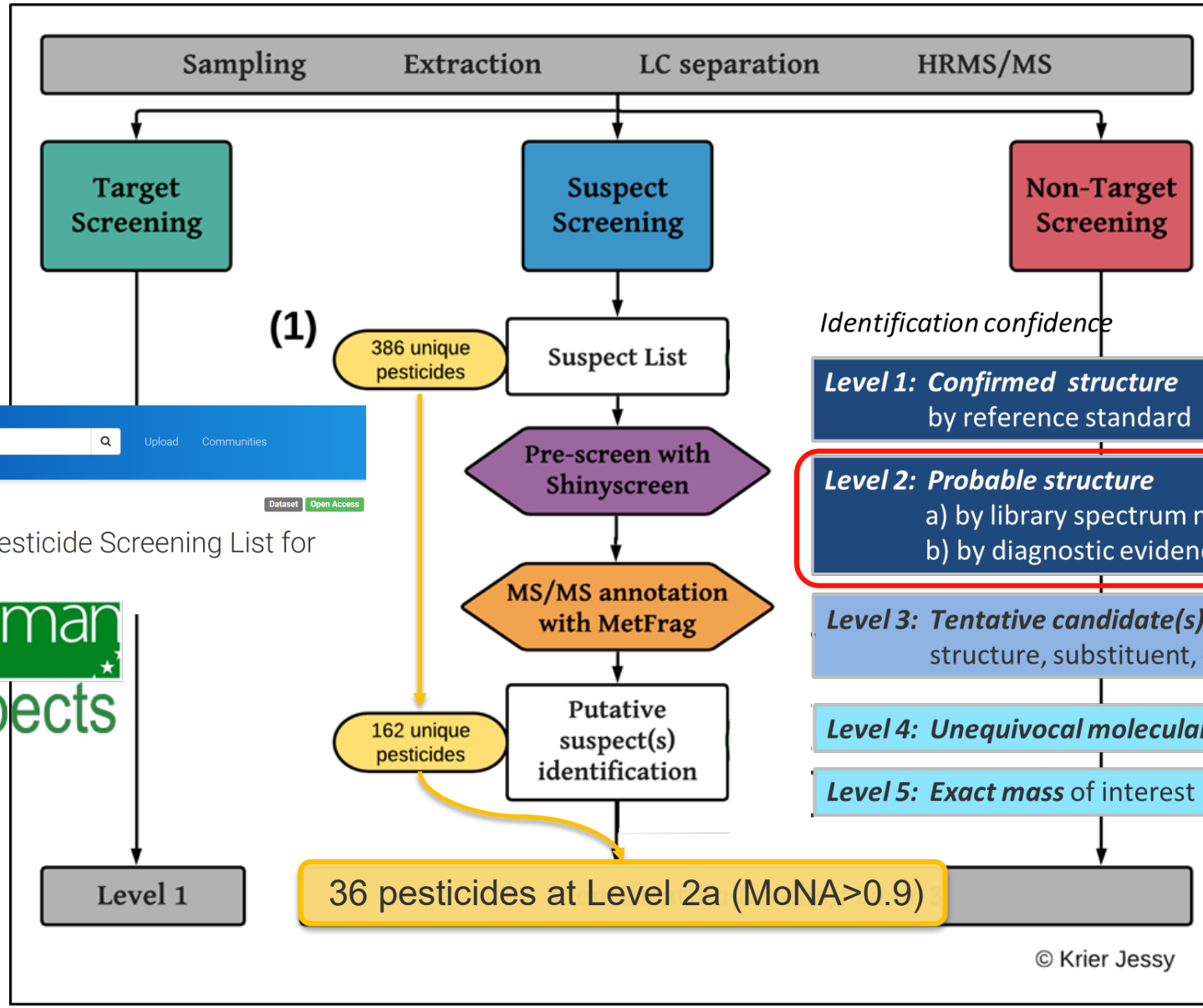
Level 1: Confirmed structure by reference standard	MS, MS^2, RT, Reference Std.
Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MassBank MS, High Quality Mass Spectral Database
Level 3: Tentative candidate(s) structure, substituent, class	MS, MS^2, Exp. data
Level 4: Unequivocal molecular formula	MS isotope/adduct
Level 5: Exact mass of interest	MS

Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885. Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

LuxPest – MS/MS Annotation with MetFrag

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





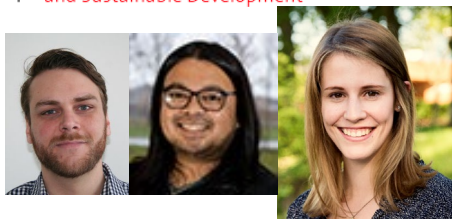
Identification confidence | *Minimum data requirements*

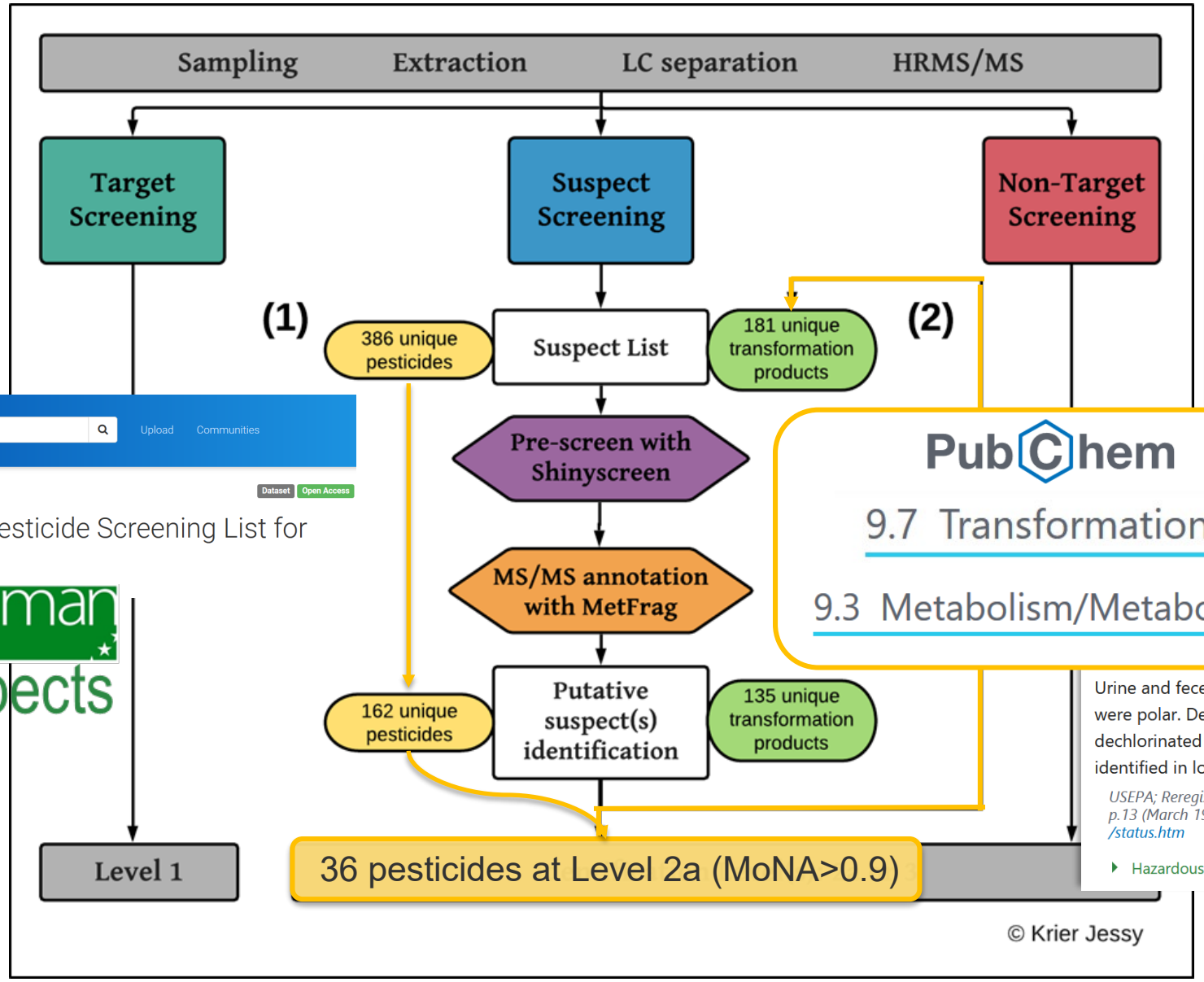
Level 1: Confirmed structure by reference standard	MS, MS^2, RT, Reference Std.
Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MassBank MS, High Quality Mass Spectral Database
Level 3: Tentative candidate(s) structure, substituent, class	MS, MS^2, Exp. data
Level 4: Unequivocal molecular formula	MS isotope/adduct
Level 5: Exact mass of interest	MS

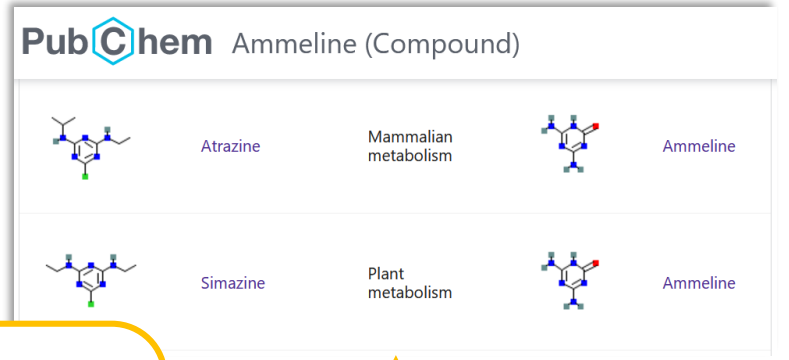
Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885. Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

LuxPest – Transformation Product Workflow

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





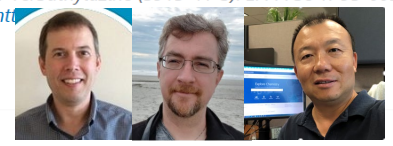


Schymanski *et al*. (2021)
DOI: 10.1186/s13321-016-0115-9

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: ht /status.htm

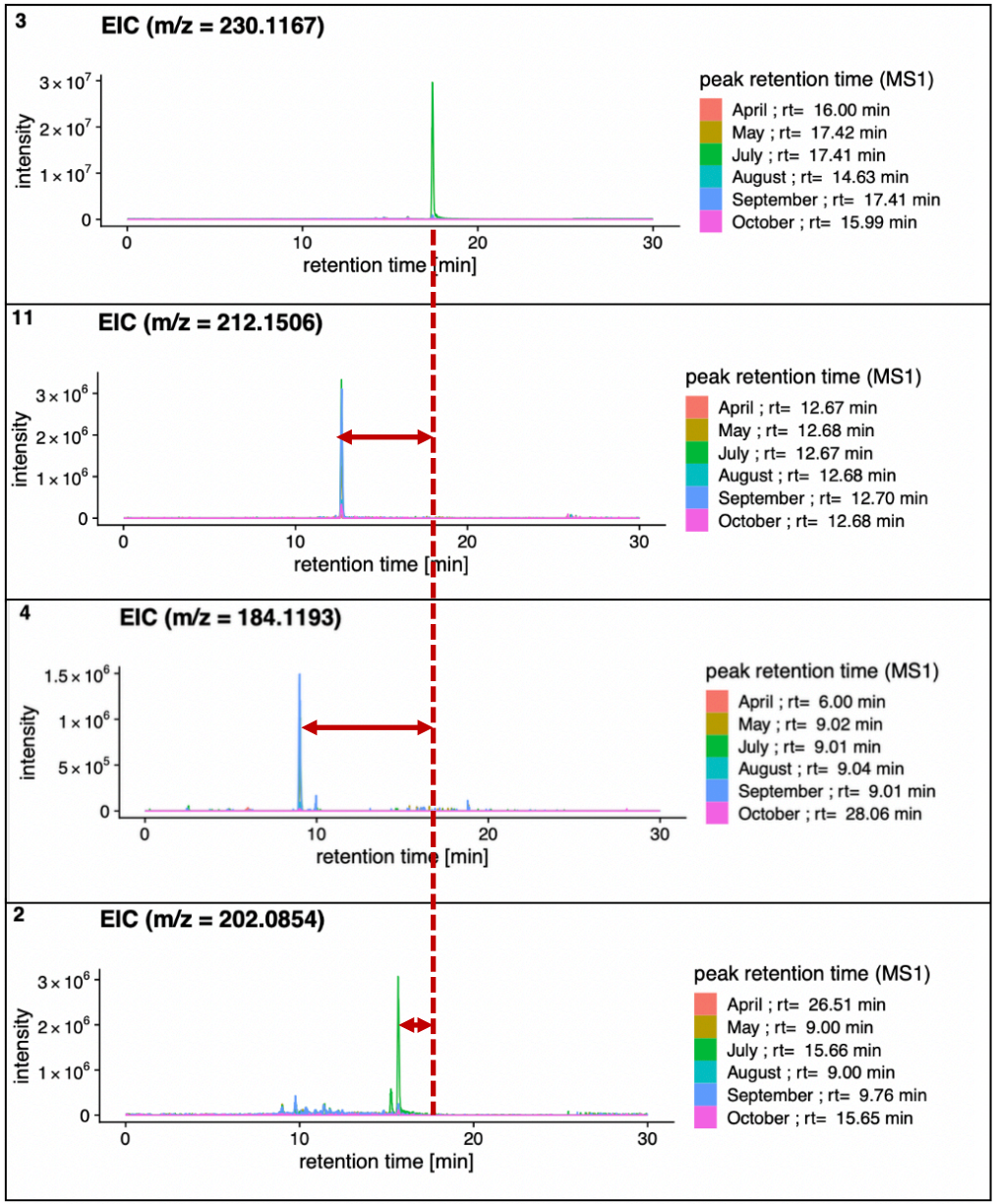
Hazardous Substances Data Bank (HSDB)

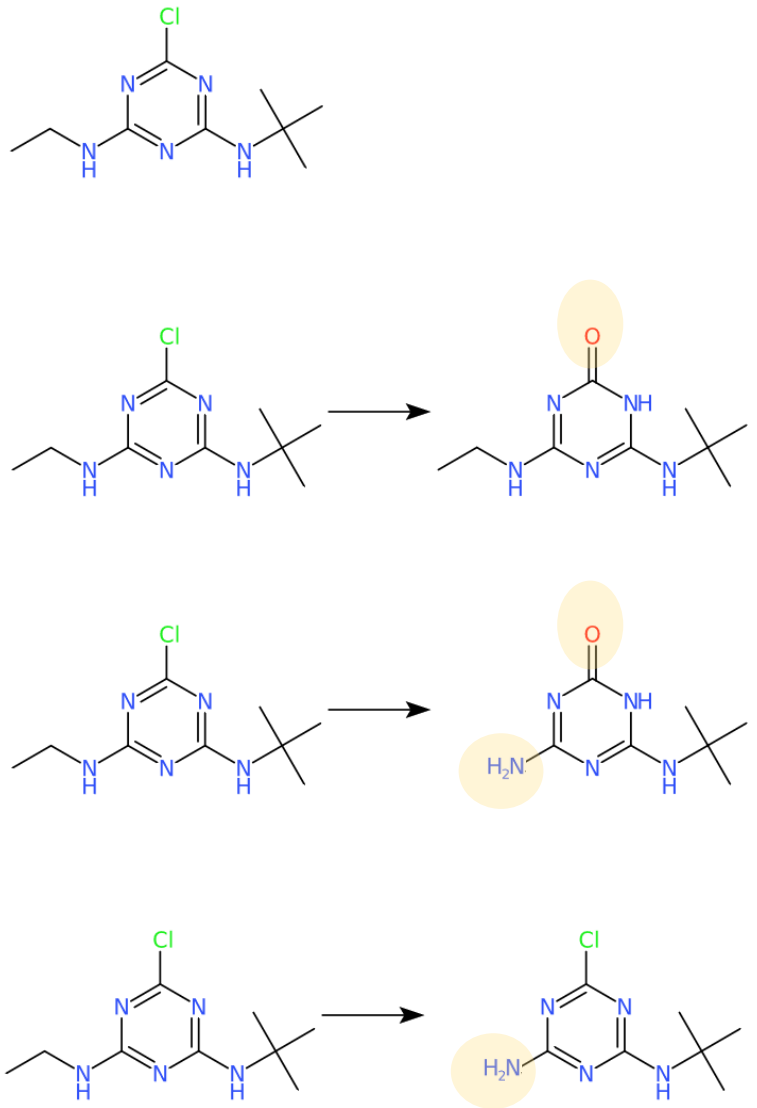


Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885. Jessy Krier (2020) S69 | LUXPEST. DOI: 10.5281/zenodo.3862689

LuxPest – Terbutylazine and (tentative) TPs

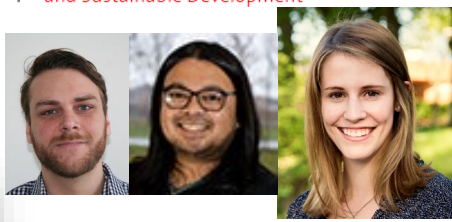
Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885

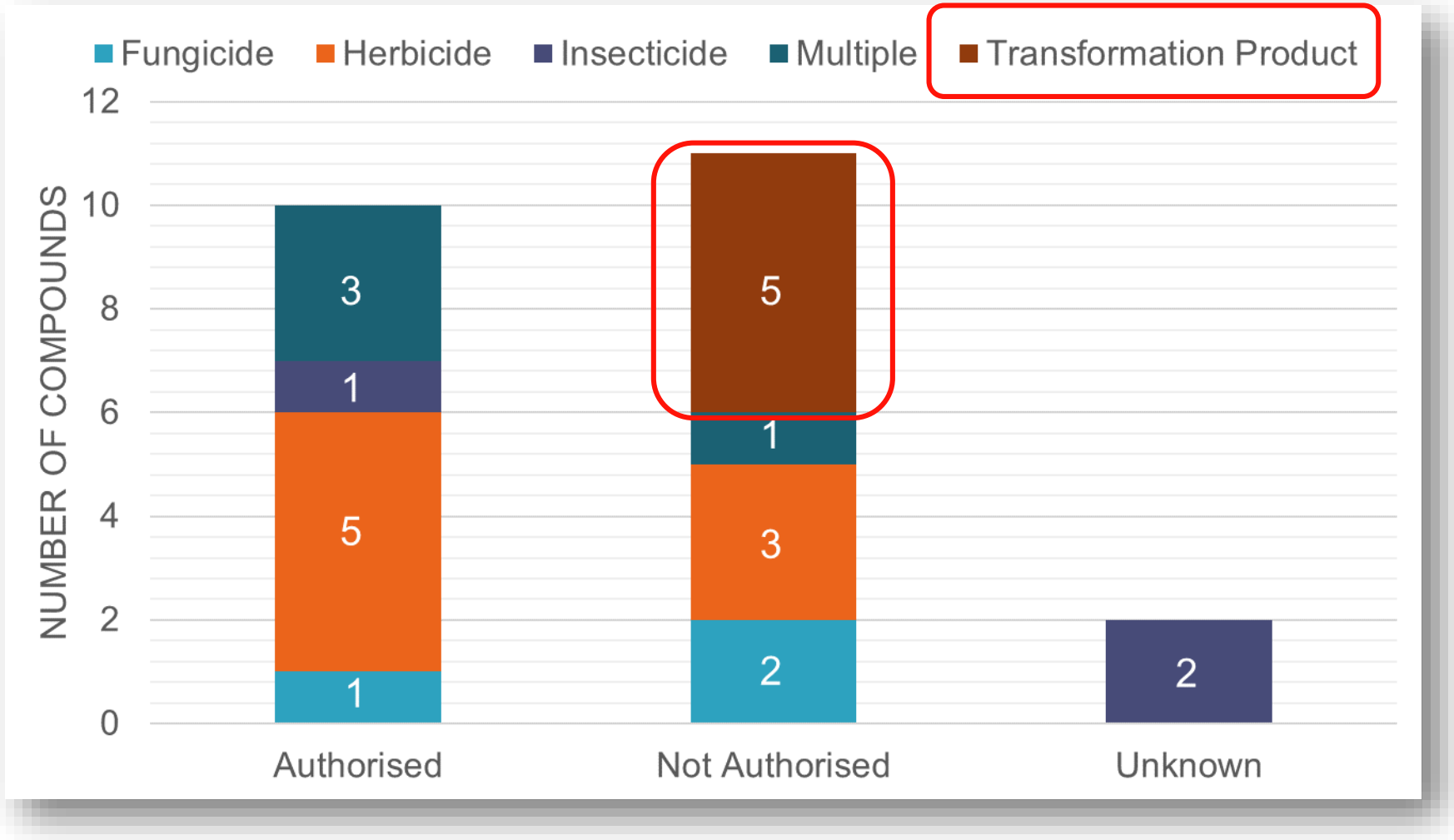




LuxPest – Verification and Quantification

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

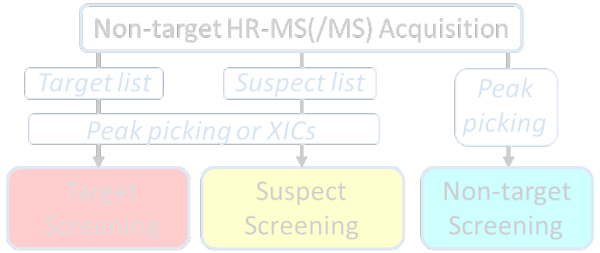


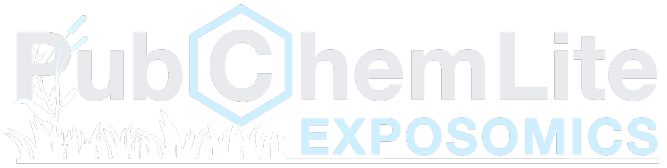


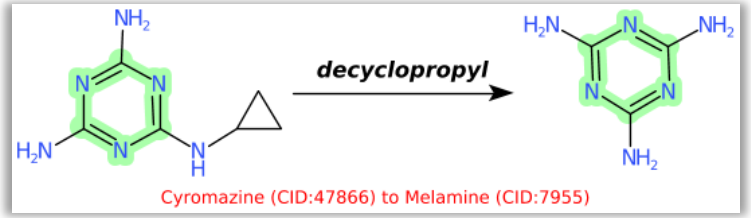
Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885.

Outline of Today

- Introduction and Background
- Identification & Chemical Space
  - Identification + MetFrag
  - PubChemLite for Exposomics
- Case Study: LuxPest
- Dark Matter and Transformations
- Take-home messages!







https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

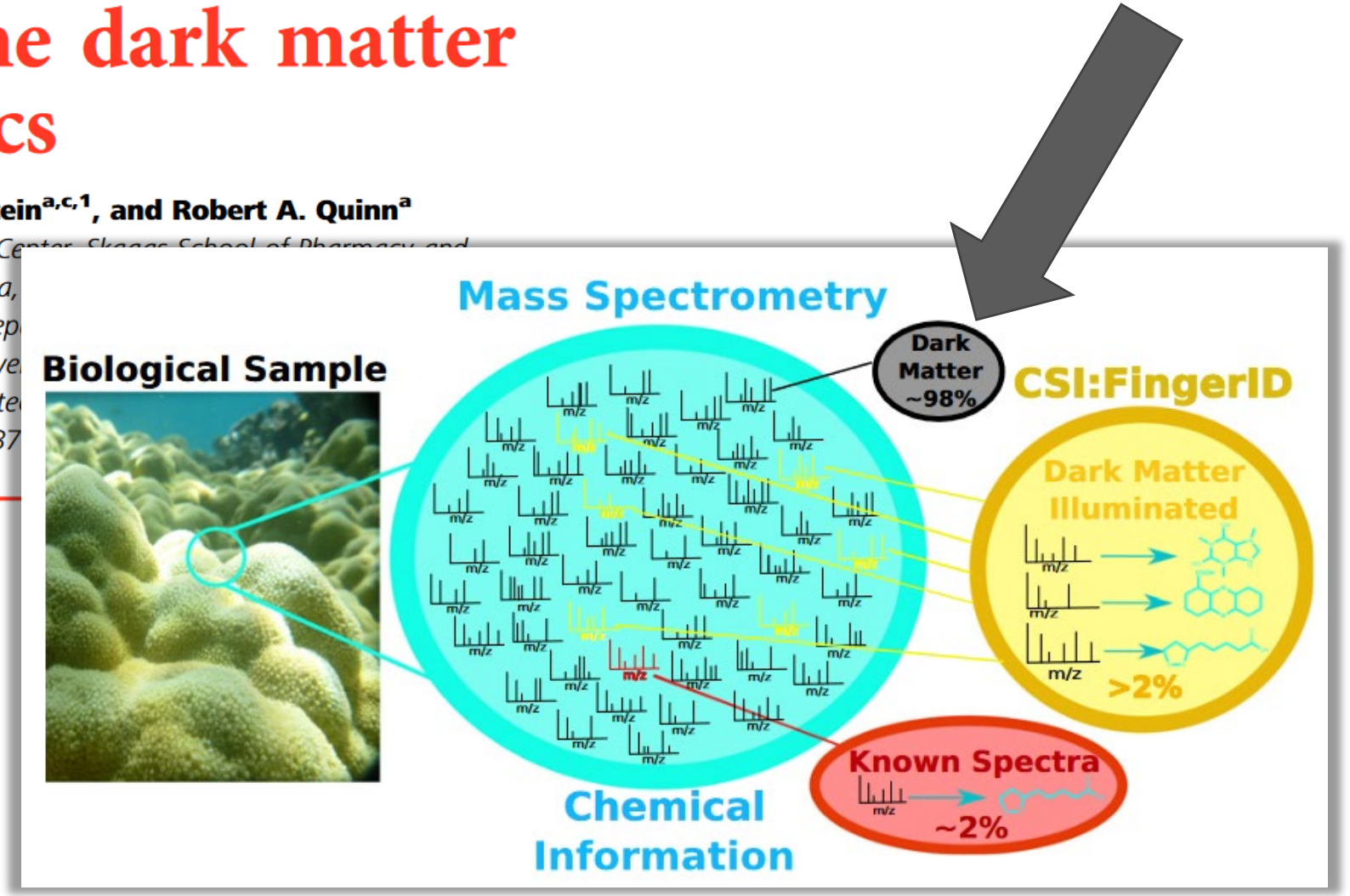
Dark Matter, Exposomics and Transformations

PNAS

Illuminating the dark matter in metabolomics

Ricardo R. da Silva[a,b], Pieter C. Dorrestein[a,c,1], and Robert A. Quinn[a]

[a]*Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, Pesquisa em Produtos Naturais e Sintéticos, Dep Ciências Farmacêuticas de Ribeirão Preto, Unive 14040-903, Brazil; and [c]Center for Marine Biote Institution of Oceanography, La Jolla, CA 92037*



da Silva et al. (2015), PNAS, DOI: 10.1073/pnas.1516878112

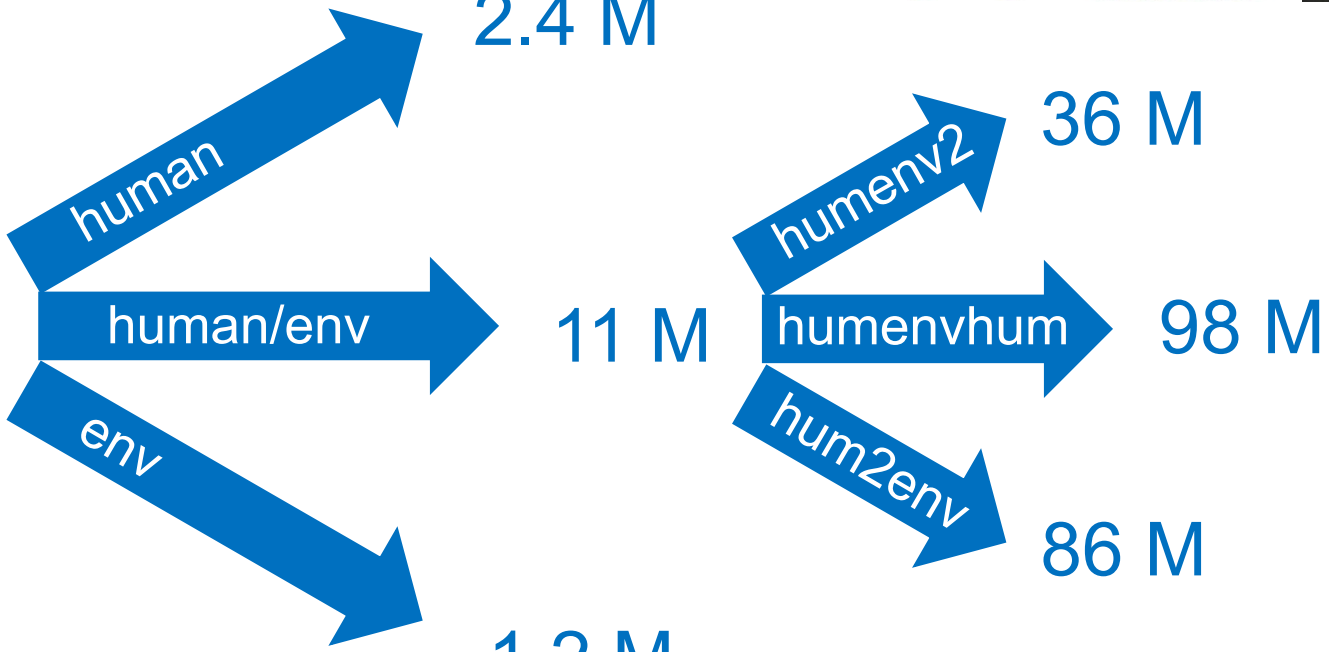
Transforming PubChemLite with BioTransformer 3.0


http://biotransformer.ca/

Wishart Lab
UNIVERSITY OF ALBERTA

PubChemLite
EXPOSOMICS

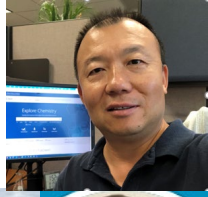
379,199 entries

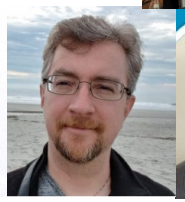


Combinatorial explosion – more data needed to predict reaction likelihoods

Djoumbou Feunang *et al* (2019). BioTransformer, JCheminf. DOI: 10.1186/s13321-018-0324-5.
Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0; Bolton *et al* (2021) *preliminary calculations;*
Schymanski, Bolton, Cheng, Thiessen, Zhang, Helmus (2021) Transformations in PubChem, DOI: 10.5281/zenodo.5644560

FAIR Transformations in PubChem and NORMAN-SLE





norman suspects

- EAWAGTPS
- HSDB
- MetXBioDB
- REFTPS
- SLUPESTTPS
- SWISSPEST19
- THSTPS
- UACCSCEC

ChEMBL

PubChem Transformations

6,091 CIDs

>double the BioTransformer library

zenodo

Search | Upload | Communities

November 4, 2021

Dataset | Open Access

Transformations in PubChem - Full Dataset

Schymanski, Emma; Bolton, Evan; Cheng, Tiejun; Thiessen, Paul; Zhang, Jian (Jeff); Helmus, Rick

This is an archive of the data contained in the "Transformations" section in PubChem for integration into patRoon and other workflows.

For further details see the ECI GitLab site: README and main "tps" folder.

Credits:

Concepts: E Schymanski, E Bolton, J Zhang, T Cheng;

Code (in R): E Schymanski, R Helmus, P Thiessen

Transformations: E Schymanski, J Zhang, T Cheng and many contributors to various lists!

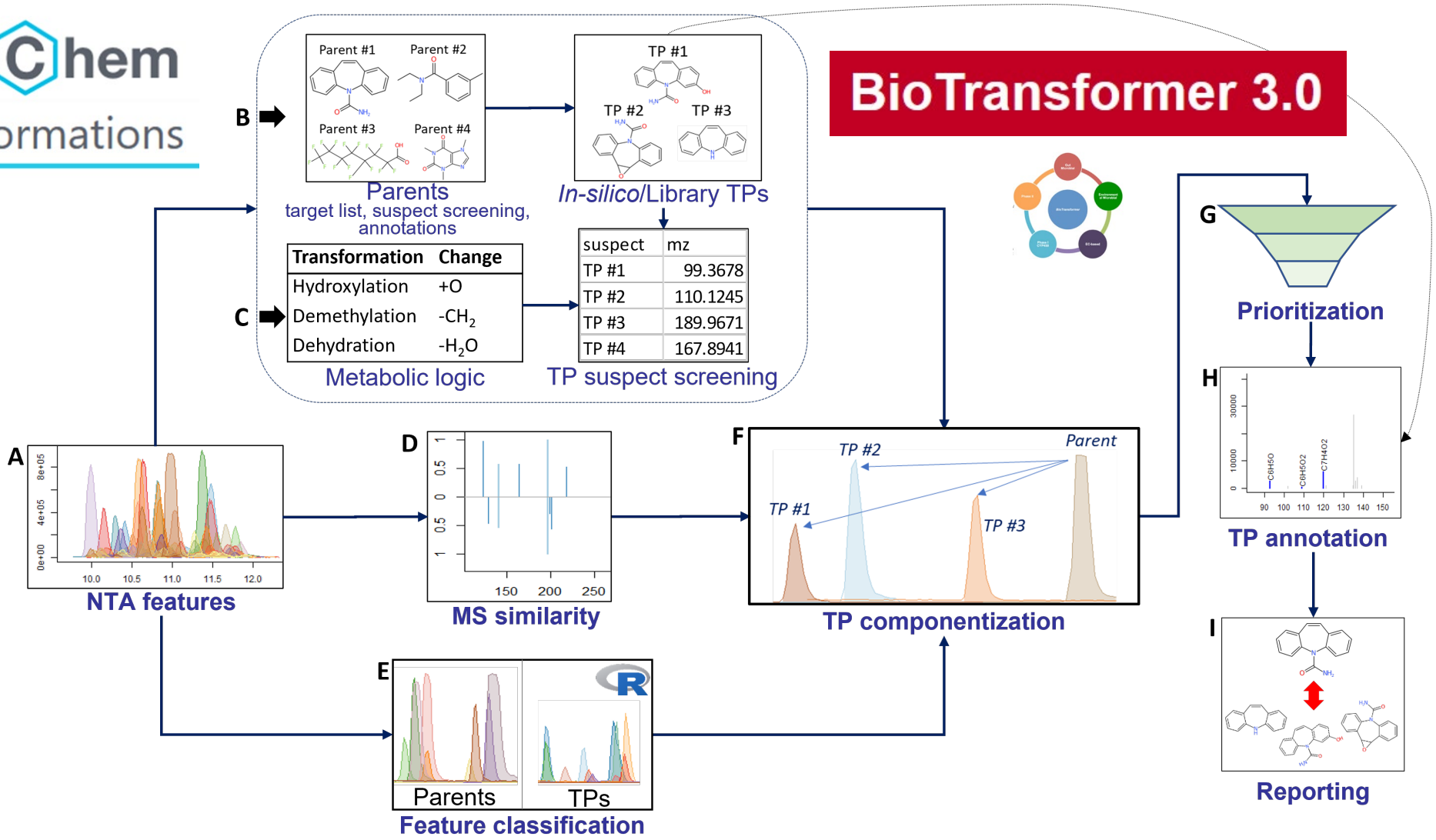
PubChem infrastructure: PubChem team.

GitLab

Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0; Helmus *et al* (under review), DOI: 10.21105.joss.03978
Schymanski, Bolton, Cheng, Thiessen, Zhang, Helmus (2021) Transformations in PubChem, DOI: 10.5281/zenodo.5644560

Open Transformation Products Workflows in patRoon 2.0

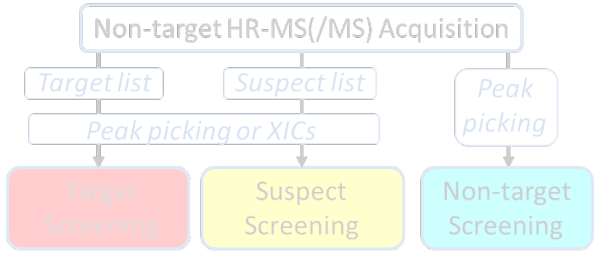
PubChem Transformations

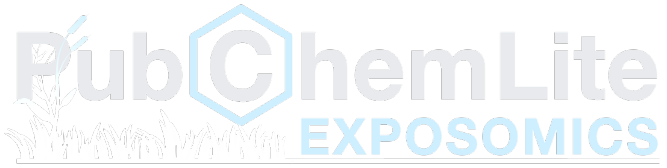


Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0; Helmus *et al* (under review), DOI: 10.21105.joss.03978
Schymanski, Bolton, Cheng, Thiessen, Zhang, Helmus (2021) Transformations in PubChem, DOI: 10.5281/zenodo.5644560

Outline of Today

- Introduction and Background
- Identification & Chemical Space
  - Identification + MetFrag
  - PubChemLite for Exposomics
- Case Study: LuxPest
- Dark Matter and Transformations
- Take-home messages!





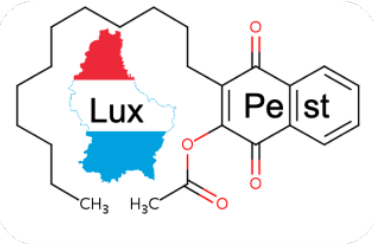
https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

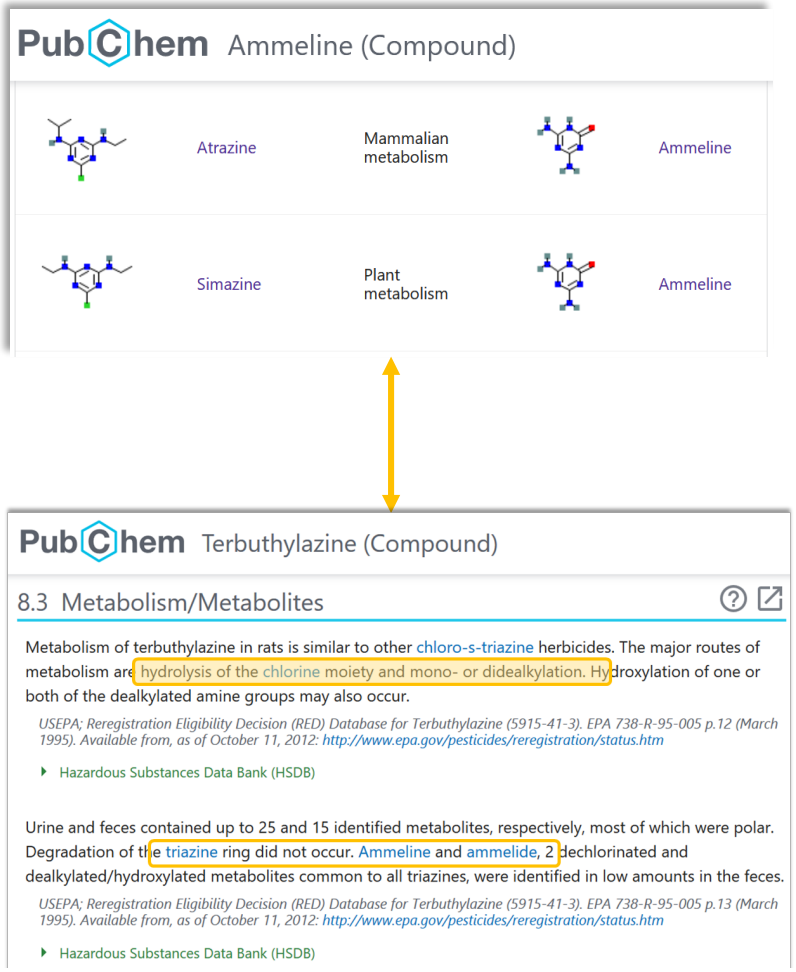
Environmental Detective Work Involves …

- Open and FAIR Expert Knowledge Exchange
- Open Source workflows
- with auto-QC & manual review
- Comprehensive & open annotation combining MetFrag & MassBank

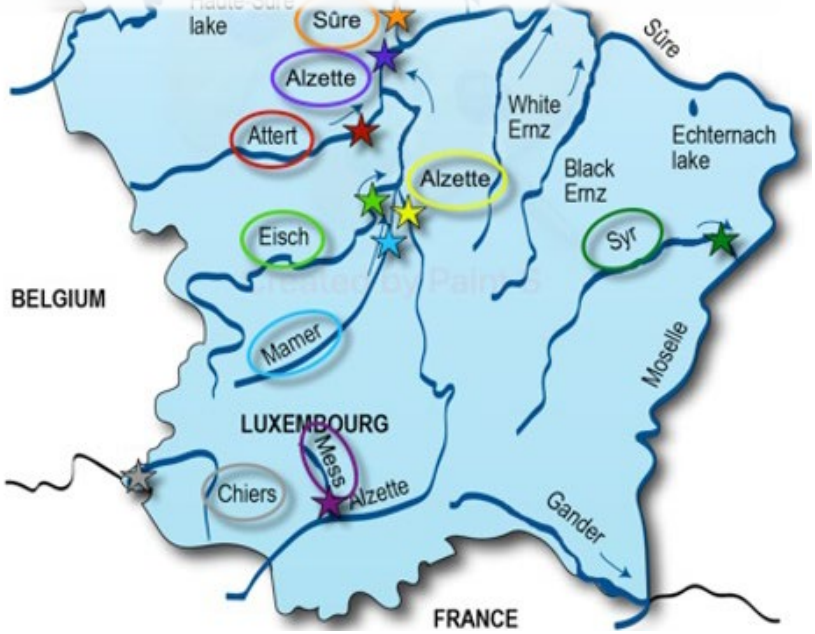
Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885.

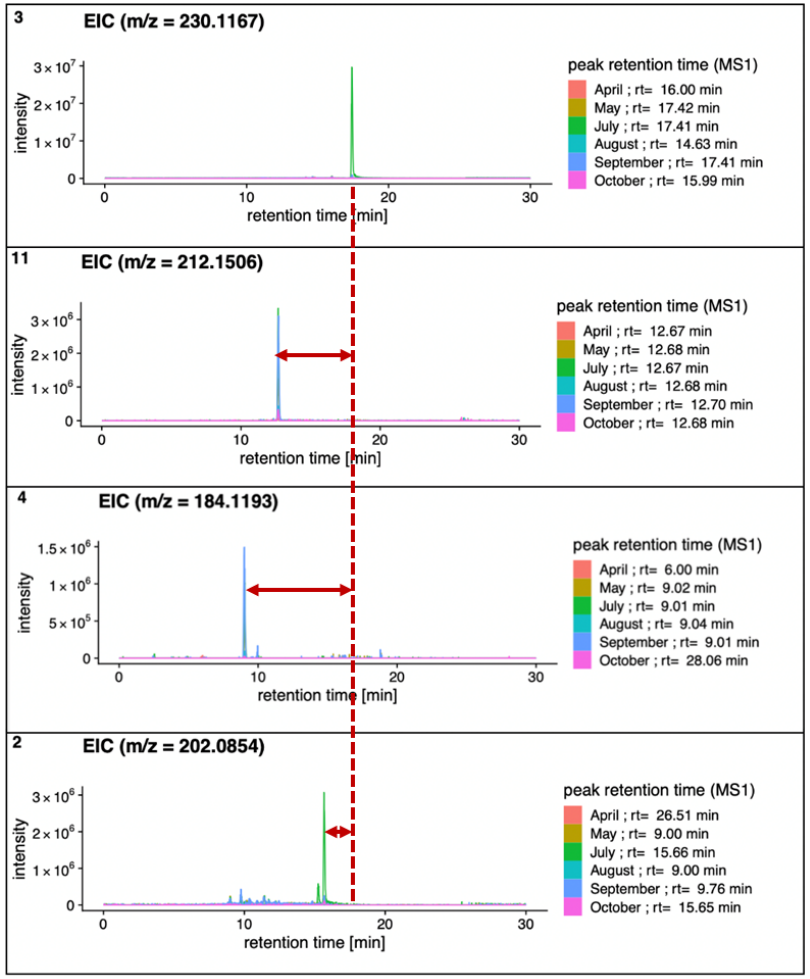
Environmental Detective Work Involves …





- Extensive transformation product efforts
- Enhanced visualisation and interpretation





Krier *et al* (2022). DOI: 10.1016/j.envint.2021.106885.

Outcomes (1)

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

Continued efforts for improved monitoring of chemicals (and actions!) in Luxembourg ... ... and the world!

- 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

PubChemLite
EXPOSOMICS

~370,000 entries “small”

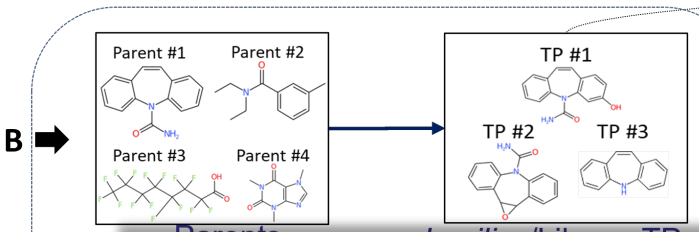
Schymanski *et al*. (2021) DOI: 10.1186/s13321-021-00489-0

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

uni.lu | LCSB

Outcomes (2) – New Open Source Transformations Workflows

PubChem Transformations



BioTransformer 3.0

patRoon 2.0.0 | Reference | Tutorial | Handbook | Changelog

Installation

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

`patRoon` itself can be installed as any other `R` package, however, some additional installation steps are needed to install its dependencies. Alternatively, R Studio based Docker images are available to easily deploy a complete `patRoon` environment. Please see the installation section in the handbook for more information.

Getting started

For a very quick start:

```
library(patRoon)
newProject()
```

The `newProject()` function will pop-up a dialog screen (requires R Studio), which will allow you to quickly select the analyses and common workflow options to subsequently generate a template `R` processing script.

However, for a better guide to get started it is recommended to read the tutorial. Afterwards the handbook is a recommended read if you want to know more about advanced usage of `patRoon`. Finally, the reference outlines all the details of the `patRoon` package.

Citing

When you use `patRoon` please cite its publications:

Links

Browse source code at https://github.com/rickhelmus/patRoon/

Report a bug at https://github.com/rickhelmus/patRoon/issues

License

GPL-3

Citation

Citing patRoon

Developers

Rick Helmus
Author, maintainer

All authors...

Dev status

PASSED

University of Amsterdam

Helmus *et al* (under review), DOI: 10.21105.joss.03978
Schymanski, Bolton, Cheng, Thiessen, Zhang, Helmus (2021) Transformations in PubChem, DOI: 10.5281/zenodo.5644560

Acknowledgements!

Email: emma.schymanski@uni.lu
Twitter: @ESchymanski

Slides @ DOI:10.5281/zenodo.5831546

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

R3 PubChemLite EXPOSOMICS

Wishart Lab
UNIVERSITY OF ALBERTA

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

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

PubChem

norman

MassBank
High Quality Mass Spectral Database

UNIVERSITY OF COPENHAGEN

Stockholm University

KEMI
Kemikalieinspektionen

Luxembourg National Research Fund

R3

ZeroPM

NGI

Maastricht University

UNIVERSITY OF AMSTERDAM

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

Funded by the European Union

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

C²DH

elixir LUXEMBOURG

SLU

UFZ

Universiteit Antwerpen

LIST

eawag aquatic research

ENVIRONMENTAL INSTITUTE

COLUMBIA MAILMAN SCHOOL OF PUBLIC HEALTH

UNIVERSITY of WASHINGTON

IPB

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

UNIVERSITAT JAUME I

Research Institute for Pesticides and Water · IUPA

EPA

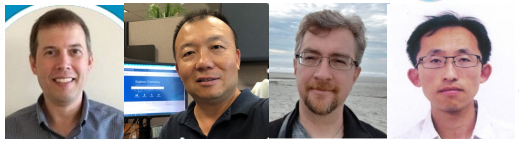
UNIVERSITAT ROVIRA i VIRGILI

ETH zürich

National and Kapodistrian UNIVERSITY OF ATHENS

ontochem

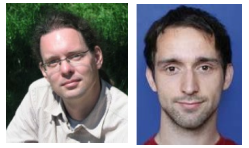
PubChem



… and team

Community Efforts!

MetFrag



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

Luxembourg National Research Fund

uni.lu

LCSB