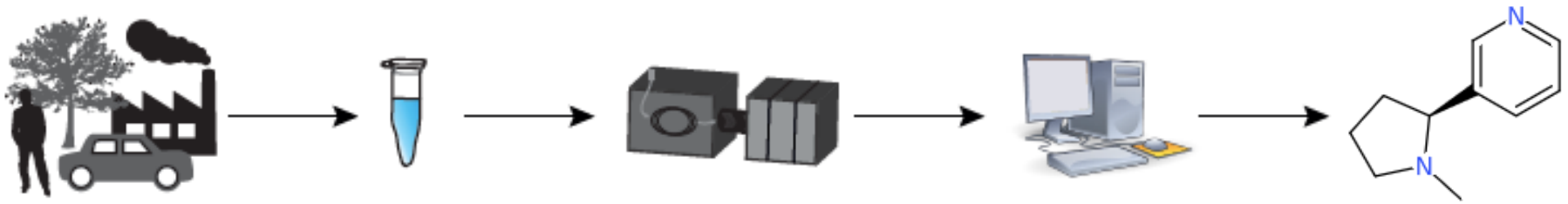
# Finding Small Molecules (and PFAS) with High Resolution Mass Spectrometry




environment → environmental sample → high resolution mass spectrometry → cheminformatics → identify chemicals and toxic effects

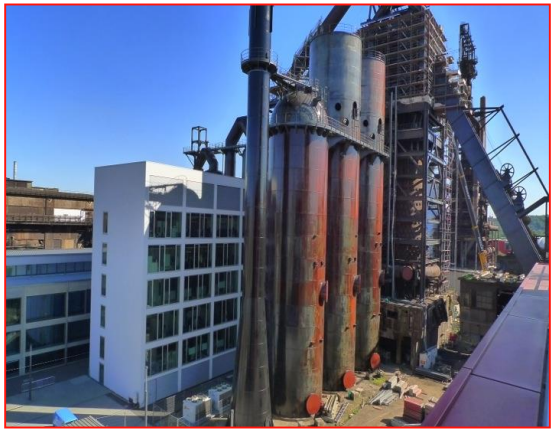
Assoc. Prof. Dr. Emma Schymanski

plus the *Environmental Cheminformatics* team and many, many collaborators!

FNR ATTRACT Fellow; Head of Environmental Cheminformatics Group,
Luxembourg Centre for Systems Biomedicine (LCSB), Luxembourg.
Email: emma.schymanski@uni.lu Twitter: @ESchymanski

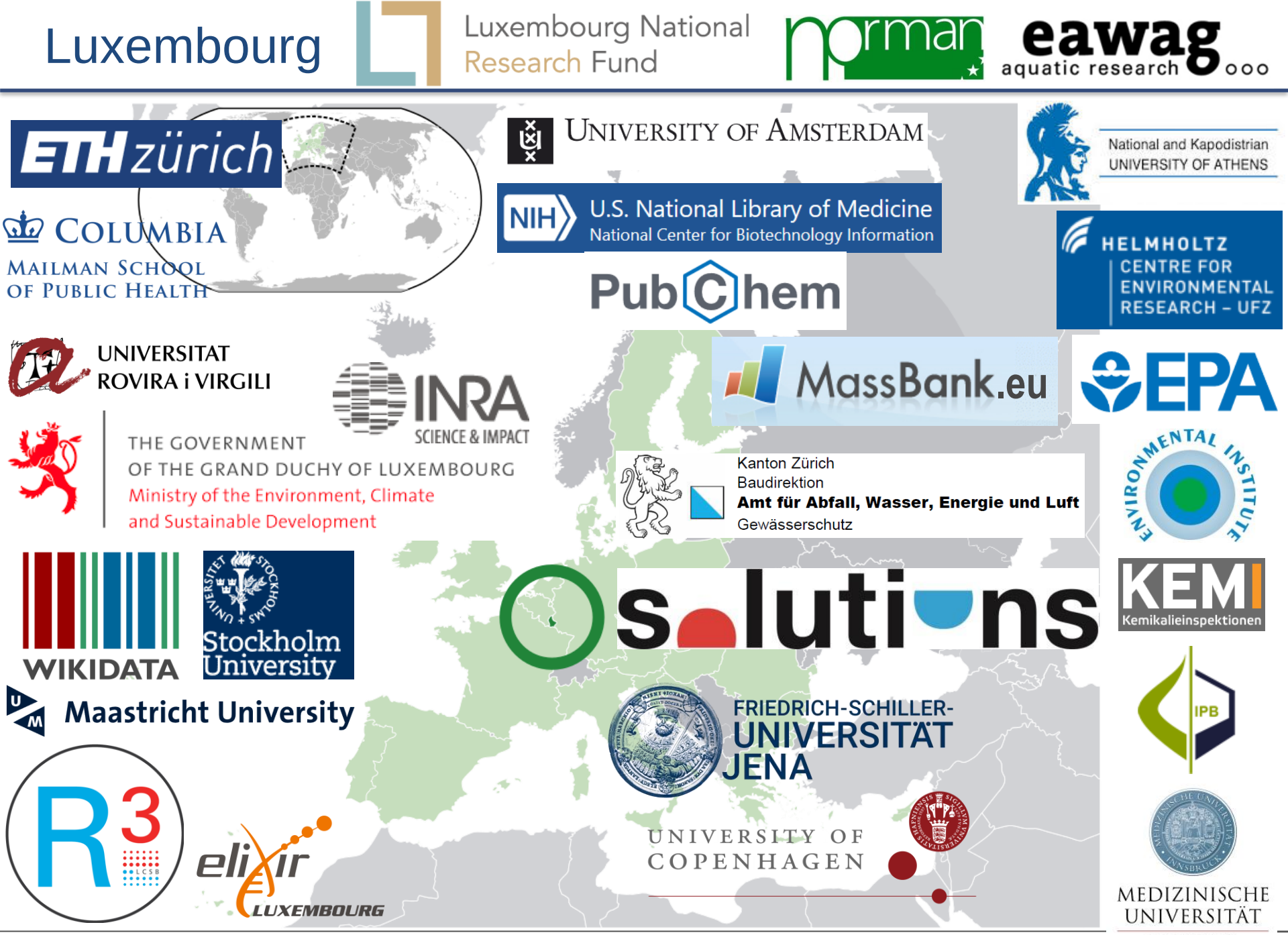
Talk DOI: 10.5281/zenodo.3786338

# University of Luxembourg & LCSB



- Uni Lu was founded in 2003
  - Teenage years!
- LCSB was founded in 2009
  - …and is still pre-teenager
    - Young and very dynamic working environment!

# Luxembourg




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

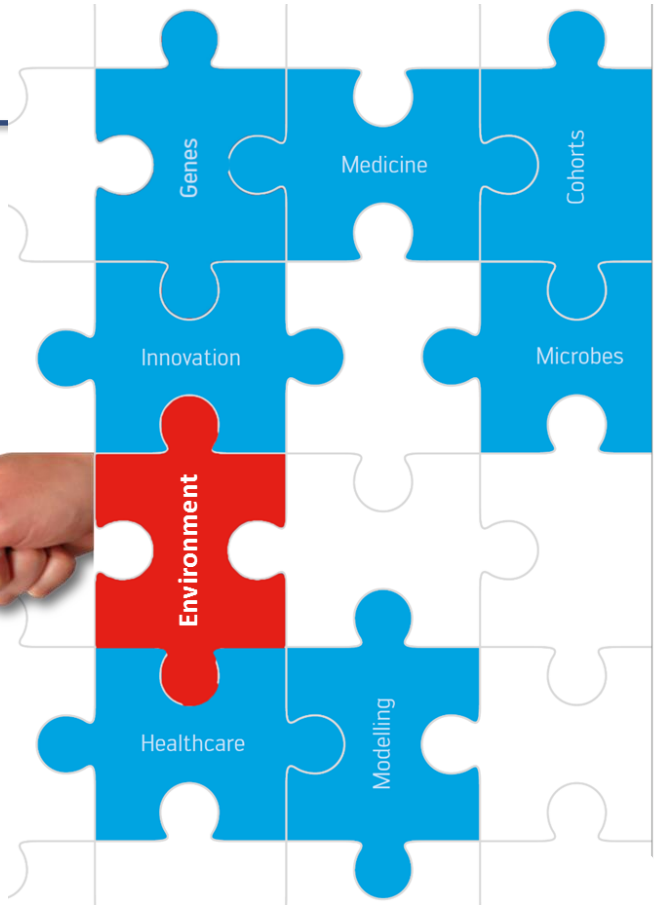
LUXEMBOURG

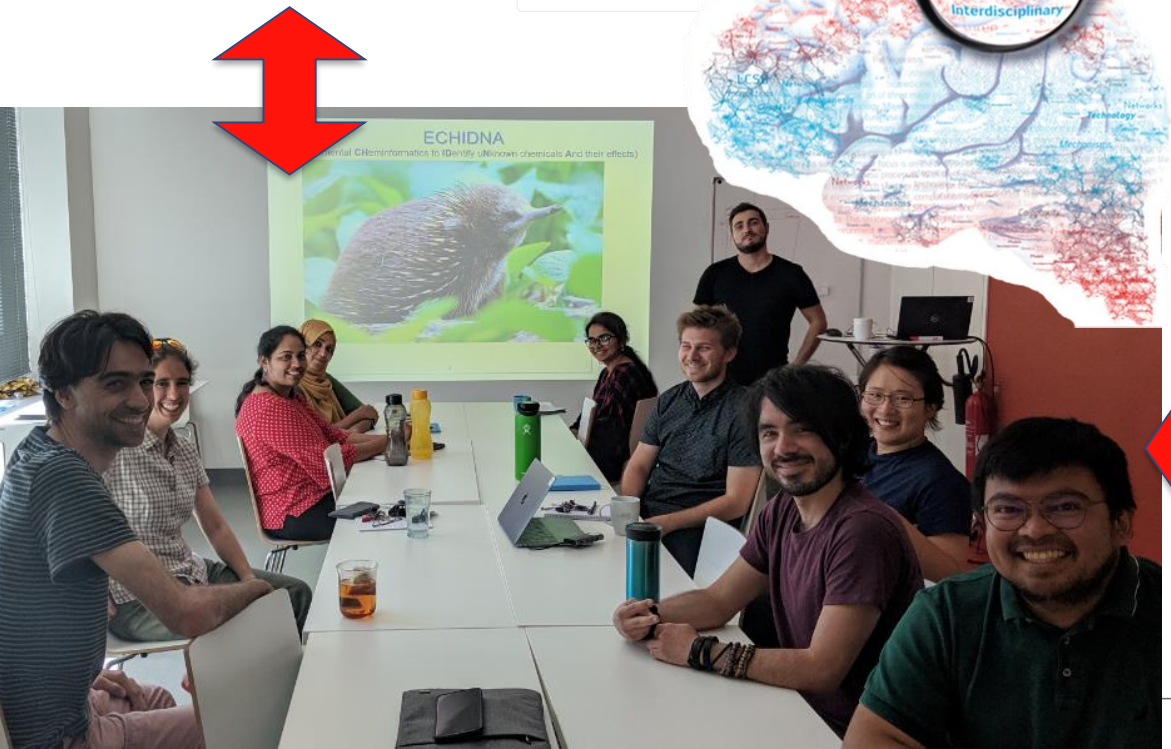
## LET'S MAKE IT HAPPEN

Members with access to **Environmental Cheminformatics**

- **Adelene Lai** @adelene.lai — Given access 2 months ago
- **Anjana Elapavalore** @anjana.elapavalore — Given access 4 weeks ago
- **Corey Griffith** @corey.griffith — Given access 2 months ago
- **Emma Schymanski** @emma.schymanski It's you — Given access 2 months ago
- **German Andres Preciat Gonzales** @german.preciat — Given access 2 months ago
- **Hiba Hiba** @hiba.hiba — Given access 4 weeks ago
- **Jessy Krier** @jessy.krier — Given access 1 week ago
- **Lorenzo Favilli** @lorenzo.favilli — Given access 2 months ago
- **Mira Narayanan** @mira.narayanan — Given access 4 weeks ago
- **Randolph Singh** @randolph.singh — Given access 2 months ago
- **Todor Kondić** @todo — Given access 2

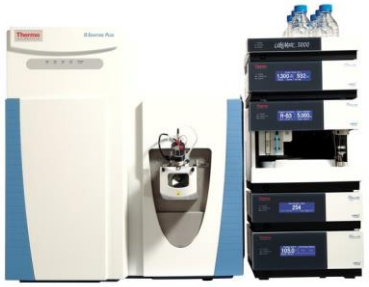
R3 LCSB







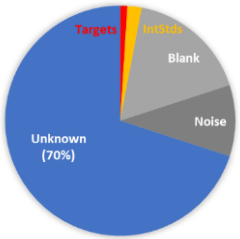


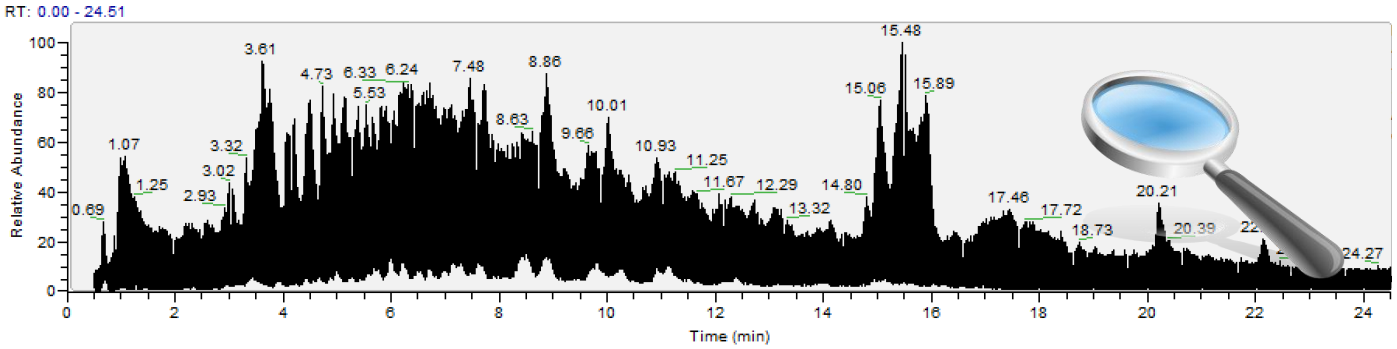
Luxembourg National Research Fund

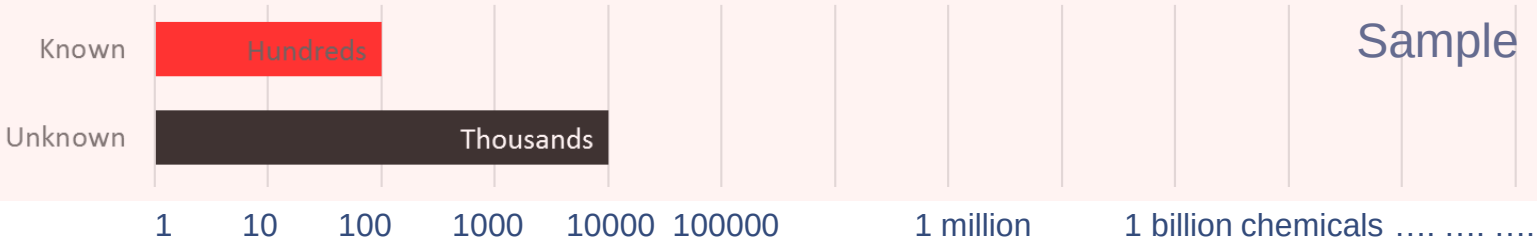
UNIVERSITÉ DU LUXEMBOURG

# Our (Community) Challenge: Identifying Chemicals

## High resolution mass spectrometry



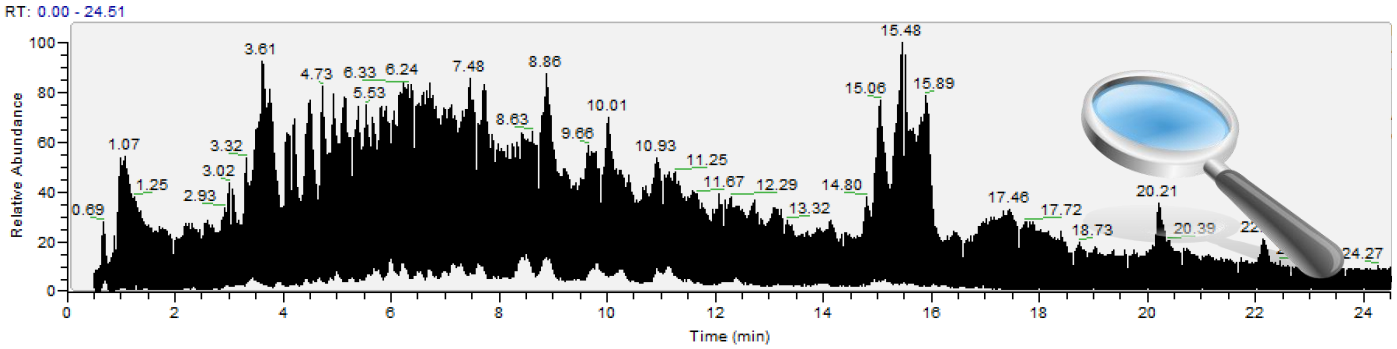


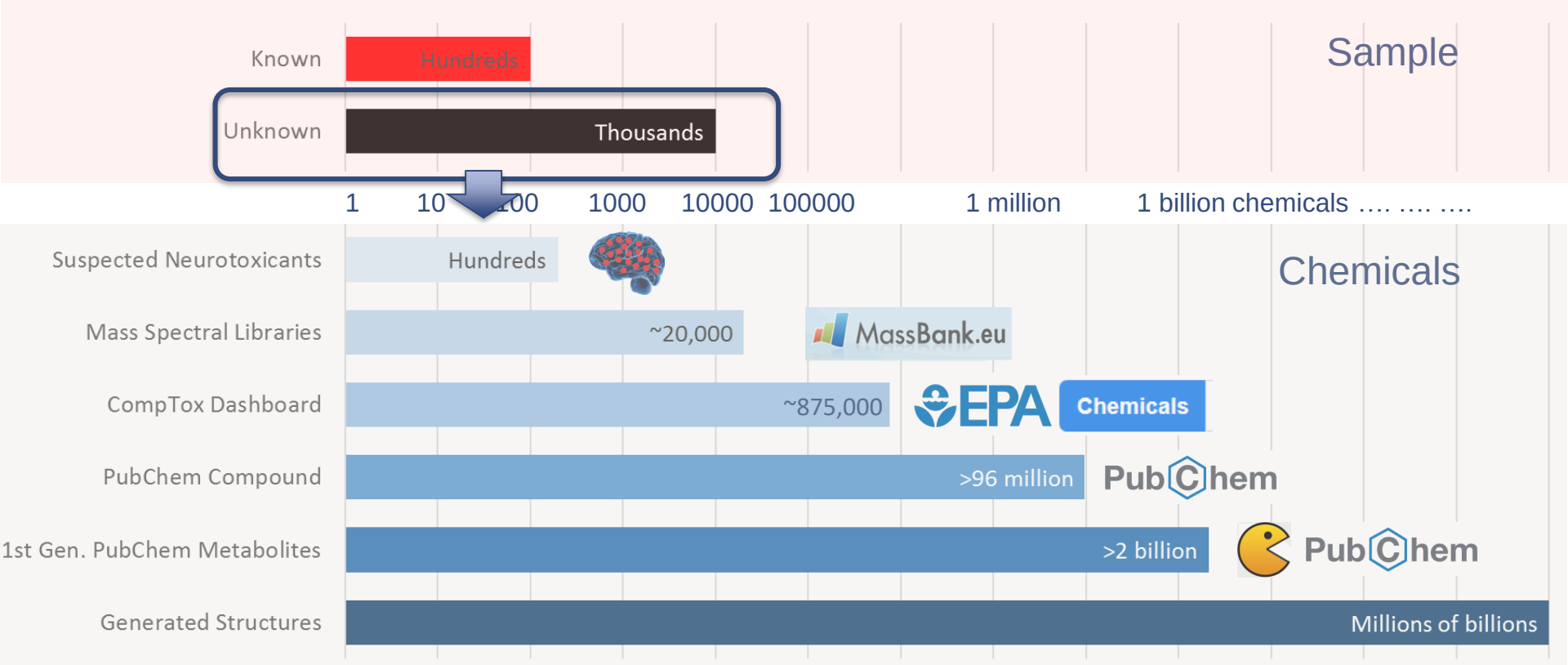

Data: Schymanski *et al* 2014, Environ. Sci. Technol. DOI: 10.1021/es4044374; Hollender *et al* 2017 DOI: 10.1021/acs.est.7b02184

# Our (Community) Challenge: Identifying Chemicals

High resolution mass spectrometry

AND connecting chemical knowledge




Data: Schymanski *et al* 2014, DOI: 10.1021/es4044374; Schymanski & Williams, 2018, DMCM 2018. DOI: 10.5281/zenodo.3242674

# Target, Suspect and Non-Target Screening


**KNOWNS**

**TARGET ANALYSIS**

Targets found

**SUSPECTS**

**SPECTRUM SEARCH**

Spectral match

**SUSPECTS**

**SUSPECT SCREENING**

Suspects found

**No Prior Knowledge**

**NON-TARGET SCREENING**

Masses of interest

(Molecular formula)

**DATABASE SEARCH**

**STRUCTURE GENERATION**

Candidate selection (retention time, MS/MS, calculated properties)

**Time, Effort & Number of Compounds….**

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

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

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

Target List

TARGET ANALYSIS

(enviMass, vendor software)

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

SUSPECT SCREENING

Gather evidence (nontarget, ReSOLUTION, RMassBank)

NON-TARGET SCREENING

Componentization (nontarget)

Prioritization (enviMass)

Masses of interest

Molecular formula determination (enviPat, GenForm)

MS/MS Extraction (RMassBank)

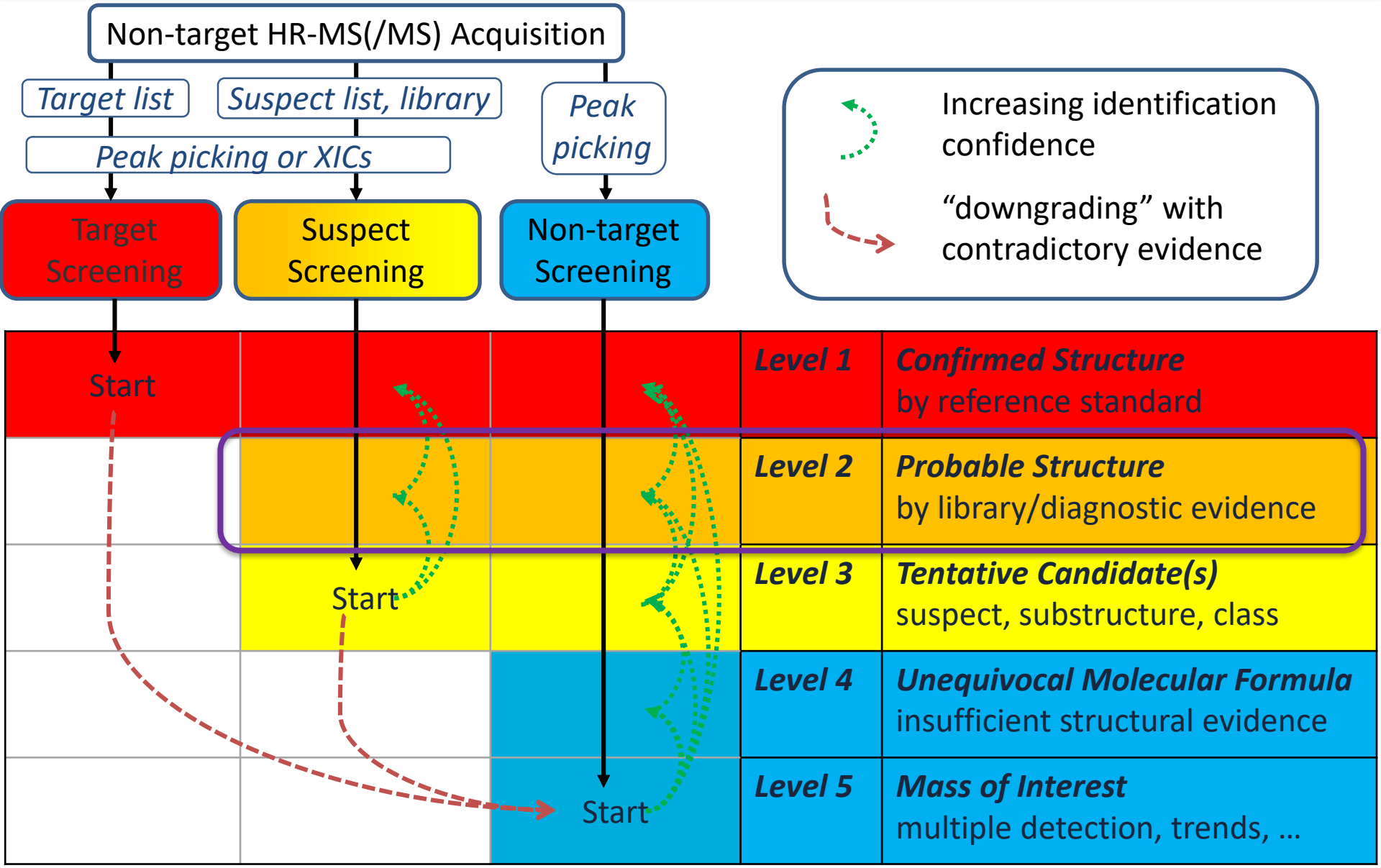
Non-target identification (MetFrag2.3, ReSOLUTION)

Interpretation, confirmation, peak inventory, confidence and reporting

Altenburger *et al*, 2019, Env. Sci. Europe. DOI: 10.1186/s12302-019-0193-1

solutions

# Identification Strategies and Confidence

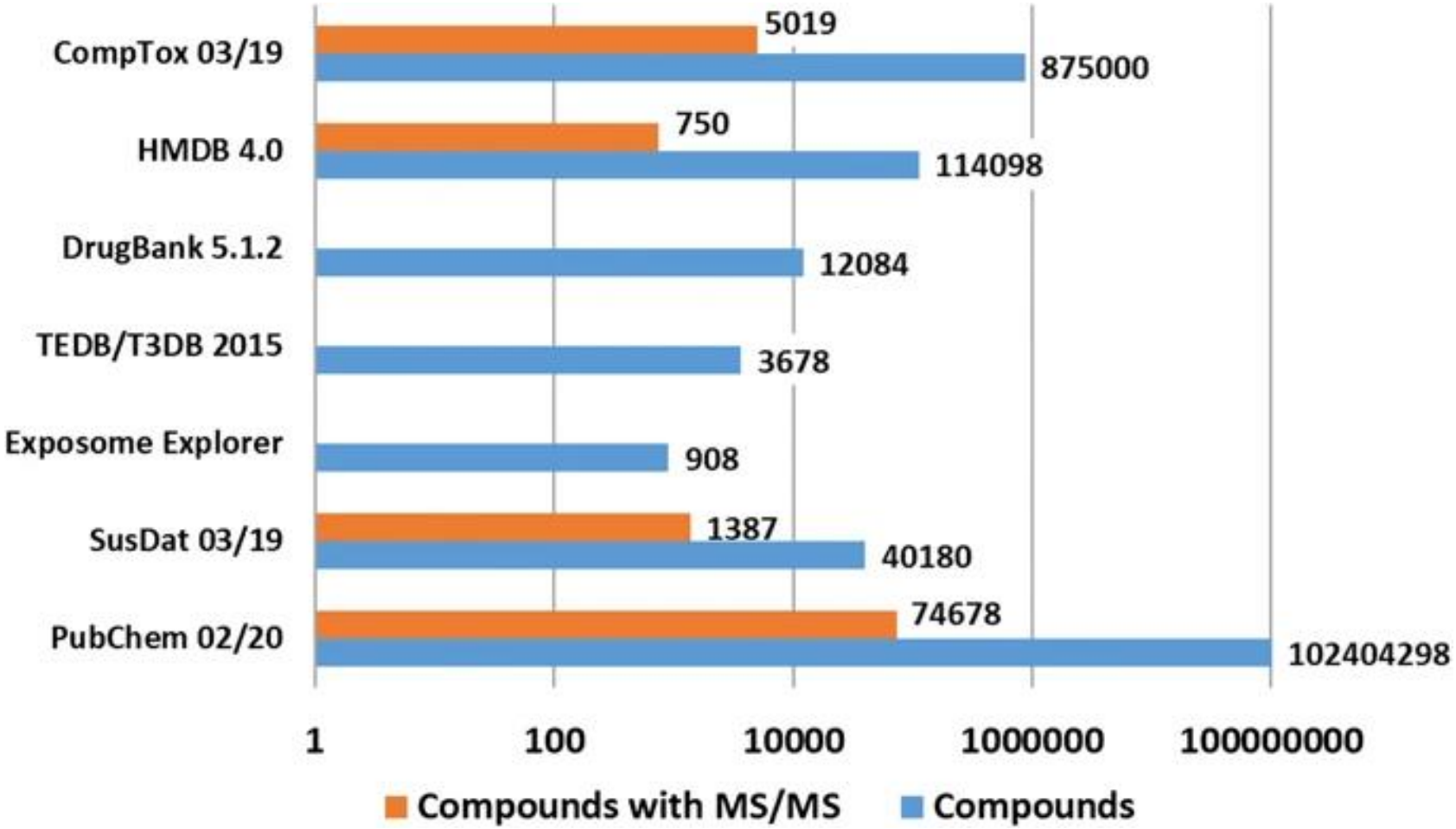





Schymanski *et al*, 2014, ES&T. DOI: 10.1021/es5002105 & Schymanski *et al*. 2015, ABC, DOI: 10.1007/s00216-015-8681-7

# Scarcity of MS/MS Spectra

- Only available for ~0.1-4 % of Exposomics-relevant resources






H. Oberacher *et al.* (2020) Environmental Sciences Europe 32: 43. DOI: 10.1186/s12302-020-00314-9

# Modernising MassBank EU

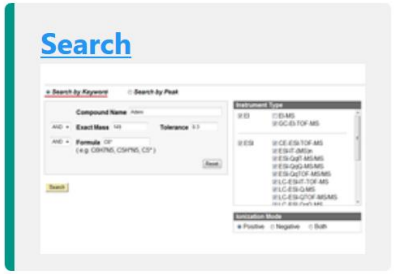
MassBank.eu

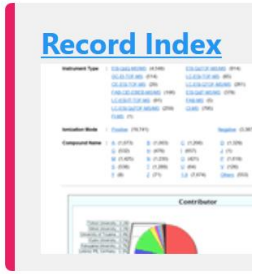
http://massbank.eu/MassBank and https://github.com/MassBank/

MassBank norman High Resolution Mass Spectral Database

European MassBank (NORMAN MassBank)

Home Search Record Index Data Privacy Imprin




>80,000 spectra
~16,000 chemicals
>46 contributors

GitHub, Inc. (US) https://github.com/MassBank

**RMassBank**

Forked from sneumann/RMassBank

Playground for experiments on the official http://bioconductor.org/packages/devel/bioc/html/RMassBank.html

R 11 3 77 1 Updated 2 days ago

**MassBank-web**

The web server application and directly connected components for a MassBank web server

Java 11 5 62 1 Updated 2 days ago

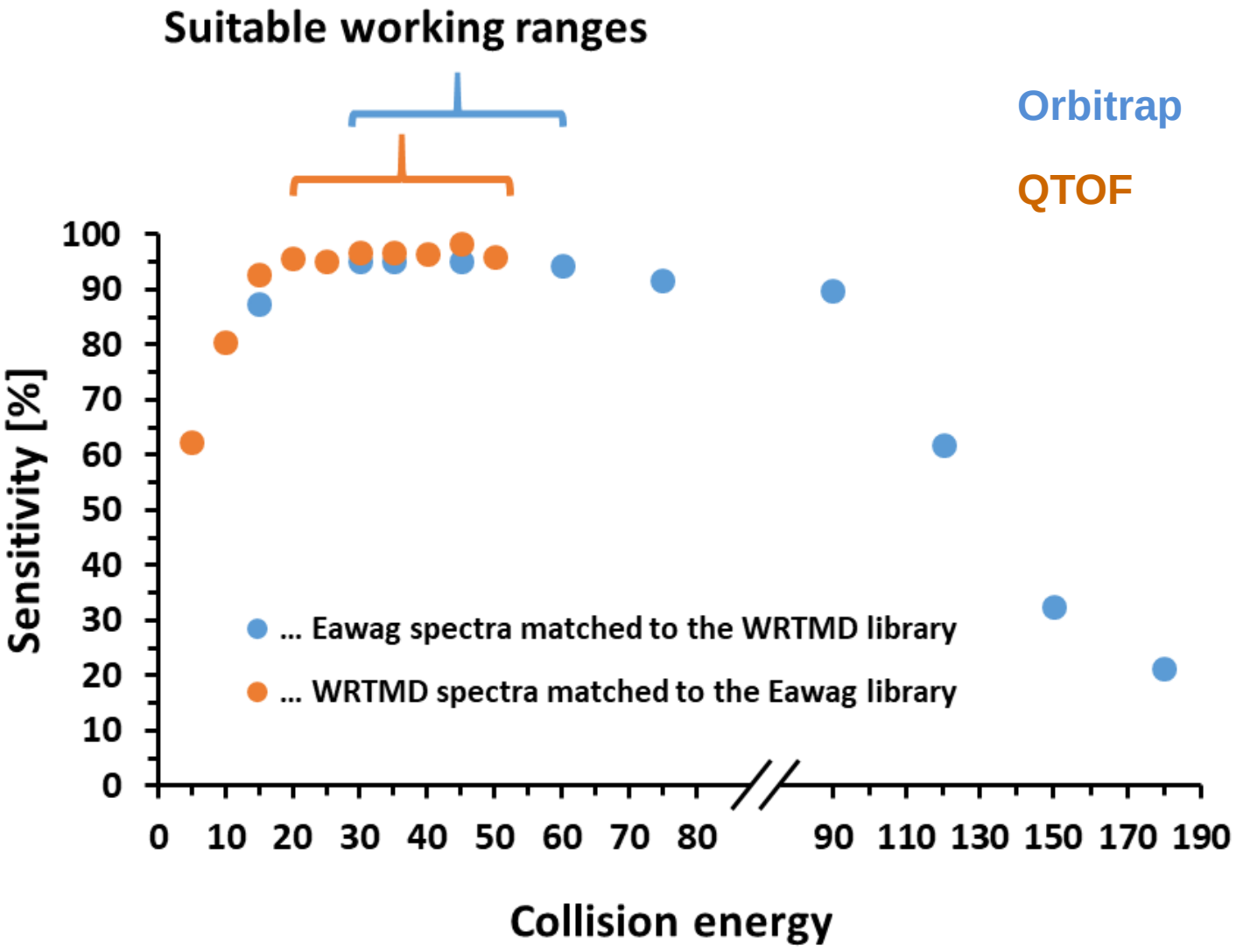
**MassBank-data**

Official repository of open data MassBank records

library repository mass-spectrometry massbank

Shell 23 14 23 0 Updated on Aug 28

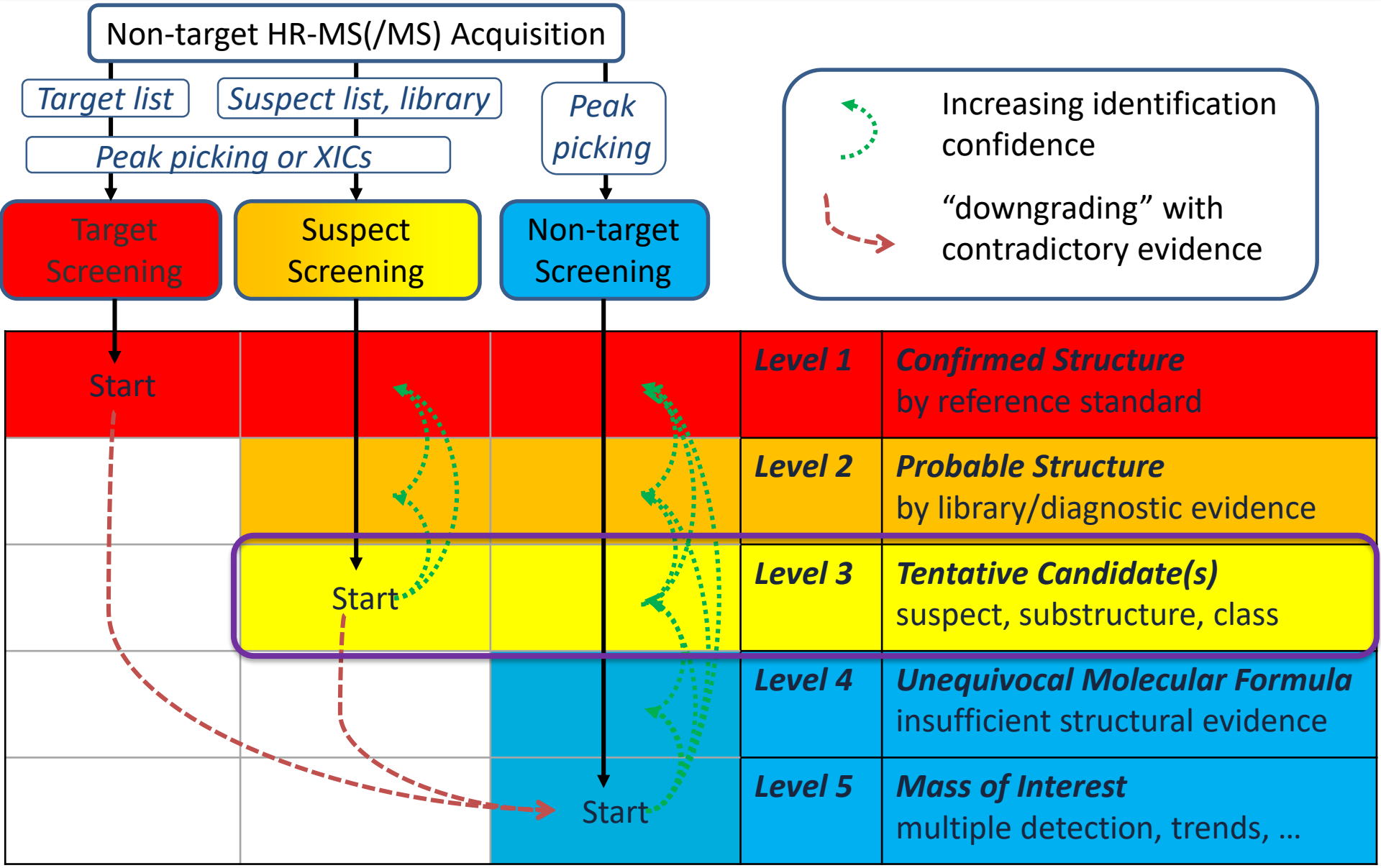
# Comparability of MS/MS Spectra: QTOF v Orbitrap






Oberacher *et al*. 2019, *Metabolites, 9*(1), 3; https://doi.org/10.3390/metabo9010003

# Identification Strategies and Confidence

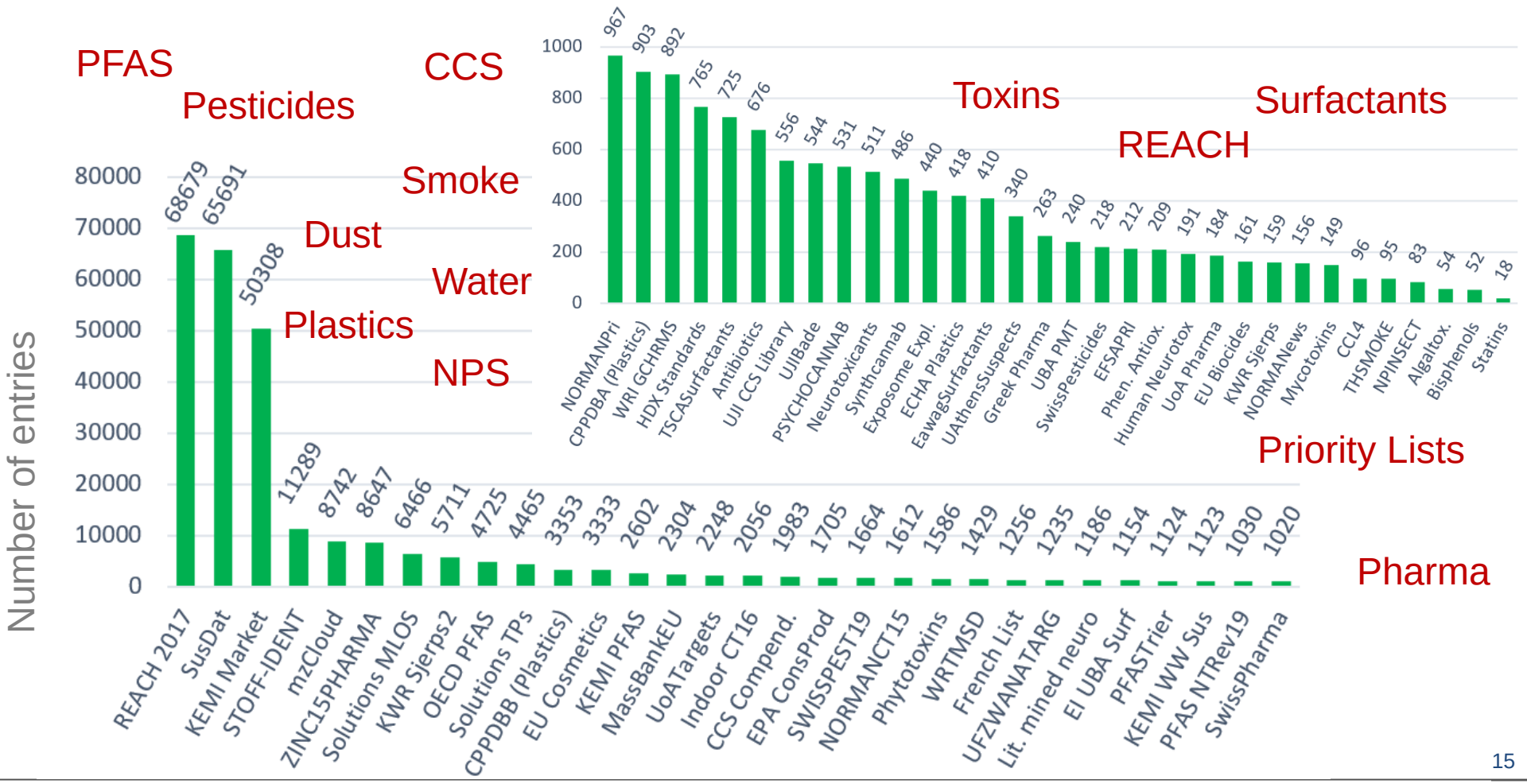





Schymanski *et al*, 2014, ES&T. DOI: 10.1021/es5002105 & Schymanski *et al*. 2015, ABC, DOI: 10.1007/s00216-015-8681-7

# NORMAN Suspect List Exchange (67 lists!)

- https://www.norman-network.com/nds/SLE/
- https://zenodo.org/communities/norman-sle
- https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101






Schymanski *et al*. (in prep.) https://www.norman-network.com/nds/SLE/

# Suspect Screening – NORMAN SLE



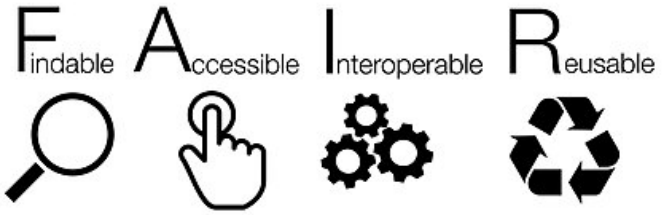
https://www.norman-network.com/nds/SLE/ NORMAN SUBSTANCE DATABASE

## NORMAN Suspect List Exchange – NORMAN SLE

| No. | Abbreviation | Description | Link to full list | Link to InChIKey list | References |
|---|---|---|---|---|---|
| S0 | SUSDAT | **Merged NORMAN Suspect List: SusDat** | Interactive Data table (updating…)<br>CompTox SUSDAT List | MS-ready InChIKeys (1/03/2018) | A merged list of >40,000 structures from suspect lists. See interactive version. Compiled by Reza Aalizadeh, University of Athens, including RTI and toxicity values, support by Nikiforos Alygizakis, EI. *Work in progress ... please report any issues!*<br>DOI: 10.5281/zenodo.2664077 |
| S1 | MASSBANK | **NORMAN Compounds in MassBank** | CSV, XLSX with Fragments (3/10/2017)<br>CompTox MassBank EU Reference List<br>CompTox MassBank EU Special Cases<br>CompTox Fragment Download | MassBankEUInChIKeys (17/06/2019) | www.massbank.eu<br>Stravs *et al.* 2013.<br>DOI: 10.1002/jms.3131<br>DOI: 10.5281/zenodo.2621390 |
| S2 | STOFFIDENT | **HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances** | STOFF-IDENT Contents (6/09/2017)<br>CompTox STOFF-IDENT List | STOFF-IDENT InChIKeys (6/09/2017) | The database enables the search for exact masses from target or unknown lists and the automatic use of a Retention Time Index. See: https://www.lfu.bayern.de/stoffident/#!home (single search for free; batch search after free registration).<br>DOI: 10.5281/zenodo.2621451 |
| S3 | NORMANCT15 | **NORMAN Collaborative Trial Targets and Suspects** | LC-MS: CSV, XLSX (3/10/2017)<br>GC-MS: CSV, XLSX (3/10/2017)<br>CompTox NORMANCT15 List | LC-MS InChIKeys (31/10/2016)<br>GC-MS InChIKeys | Schymanski *et al.* 2015.<br>DOI: 10.1007/s00216-015-8681-7<br>DOI: 10.5281/zenodo.2621478 |

# NORMAN-SLE on Zenodo

https://zenodo.org/communities/norman-sle/




zenodo

Search

emma.schymanski@uni.lu

437 views

528 downloads

See more details...

Tweeted by 3

3

See more details

## NORMAN Suspect List Exch

## Recent uploads

New upload

Search NORMAN Suspect List Exchange

May 6, 2019 (NORMAN-SLE-S52.0.1.0) Dataset Open Access

### S52 | THSMOKE | Thirdhand Smoke (THS) Compounds

Torres, Sonia; Schymanski, Emma; Ramirez, Noelia;

This is the collection associated with list S52 THSMOKE on the NORMAN https://www.norman-network.com/?q=suspect-list-exchange S52 THSM (THS) Compounds THSMOKE XLSX , CSV (06/05/2019) CompTox THSI InChIKeys (06/05/

Uploaded on May 6, 2019

March 1, 2018 (NORMAN-SLE-S0.0.1.0) Dataset Open Access

### S0 | SUSDAT | Merged NORMAN Suspect List: SusDat

**Publication date:**
June 20, 2018

**DOI:**
DOI 10.5281/zenodo.3653165

**Keyword(s):**
Suspect Screening PFAS per- and polyfluoroalkyl substances

**Related identifiers:**
Supplement to ENV/JM/MONO(2018)7

**Communities:**
NORMAN Suspect List Exchange

**License (for files):**
Creative Commons Attribution 4.0 International

MAN Suspect List
ange

s a public repository (under
opment) for suspect lists currently
ble on the NORMAN Suspect List
nge: https://www.norman-
rk.com/?q=suspect-list-exchange

Read more

ed by:
ymane

on policy:
community will collect data that is
relevant and public (or to be public) for

# NORMAN-SLE + more lists on CompTox

- https://comptox.epa.gov/dashboard/chemical_lists
- https://comptox.epa.gov/dashboard/chemical_lists/?search=NORMAN
- https://comptox.epa.gov/dashboard/chemical_lists/?search=PFAS

| List Acronym | List Name | Last Updated | Number of Chemicals | List Description |
|---|---|---|---|---|
| REACH2017 | NORMAN: REACH Chemicals List | 2019-05-19 | 57758 | This REACH list of 57760 REACH chemicals from the NORMAN |
| SUSDAT | | | | |
| KEMIMARKET | | | | |
| SOLUTIONSMLOS | | | | |
| ZINC15PHARMA | | | | |
| KWRSJERPS2 | | | | |
| PFASOECDNA | | | | |

| List Acronym | List Name | Last Updated | Number of Chemicals | List Description |
|---|---|---|---|---|
| PFASMASTER | PFAS Master List of PFAS Substances | 2019-11-11 | 7866 | PFASMASTER is a consolidated list of PFAS substances spanning and bounded by the below lists of current interest to researchers and regulators worldwide. |
| PFASSTRUCT | PFAS\|EPA: PFAS structures in DSSTox | 2019-11-11 | 6648 | List of all structures contained in DSSTox bounded by a set of structure filters used to identify PFAS (per- and polyfluorinated substances) |
| PFASOECD | PFAS: Listed in OECD Global Database | 2018-05-16 | 4729 | OECD released a New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances, (PFASs) listing more than 4700 new PFAS |
| PFASOECDNA | PFAS\|NORMAN: List of PFAS from the OECD Curated by Nikiforos Alygizakis | 2019-05-19 | 3213 | List of PFAS released by the OECD, provided by Zhanyun Wang, curated and mapped to structures by Nikiforos Alygizakis |
| PFASKEMI | PFAS: List from the Swedish Chemicals Agency (KEMI) Report | 2017-02-09 | 2416 | Perfluorinated substances from a Swedish Chemicals Agency (KEMI) Report on the occurrence and use of highly fluorinated substances. |
| PFASDEV1 | PFAS DEVELOPMENT LIST: NO STRUCTURES | 2019-10-13 | 1072 | PFAS List of chemicals - IN DEVELOPMENT NO STRUCTURES |
| PFASTRIER | PFAS\|NORMAN: PFAS Community-Compiled List | 2017-07-16 | 597 | PFASTRIER community-compiled public listing of PFAS (Trier et al, 2015) |

Huge thanks to Antony (Tony) Williams, US EPA for this continued integration of NORMAN content

# NORMAN-SLE on PubChem

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

https://pubchem.ncbi.nlm.nih.gov/source/NORMAN Suspect List Exchange

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

## PubChem NORMAN Suspect List Exchange

| | |
|---|---|
| Organization: | NORMAN Network (c/o Emma Schymanski) |
| Category: | Research and Development |
| URL: | https://www.norman-network.com/nds/SLE/ |
| Contact Name: | Emma Schymanski |
| Address: | 6 avenue du Swing, Belvaux, Luxembourg, 4367 |
| Data Source ID: | 23819 |
| Data in PubChem: | 124,923 Live Substances<br>6,505 Annotations<br>1 Classification |
| Last Updated: | 2020/04/27 |




Huge thanks to Evan Bolton, Jian Zhang (Jeff), Paul Thiessen, Ben Shoemaker and the PubChem team for this!

# NORMAN-SLE on PubChem

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

- https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101

## PubChem Classification Browser

Help

Browse PubChem data using a classification of interest, or search for PubChem records annotated with the desired classification/term (e.g., MeSH: phenylpropionates, or Gene Ontology: DNA repair). More...

**Select classification**

NORMAN Suspect List Exchange

**Search selected classification by**

Keyword | Enter desired search term | Search

**Classification description (from NORMAN Suspect List Exchange)**

The NORMAN Suspect List Exchange (NORMAN-SLE) is a central access point for NORMAN members (and others) to find suspect lists relevant for their environmental monitoring questions. More...

**Data type counts to display**: None | Compound

**Display zero count nodes?**: Yes | No

### Browse NORMAN Suspect List Exchange Tree

- NORMAN Suspect List Exchange Classification ? 133,878
  - S25 | OECDPFAS | List of PFAS from the OECD ? 3,680
  - S60 | SWISSPEST19 | Swiss Pesticides and Metabolites ? 1,358
- S00 | SUSDAT | Merged NORMAN Suspect List: SusDat ? 64,540
- S01 | MASSBANK | NORMAN Compounds in MassBank EU ? 15,917
- S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances ? 11,452
- S03 | NORMANCT15 | NORMAN Collaborative Trial Targets and Suspects ? 640

# NORMAN-SLE on PubChem

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

- https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101

**Select classification** NORMAN Suspect List Exchange

**Search selected classification by** Keyword PFAS Search Clear

**Classification description (from NORMAN Suspect List Exchange)**

1. **S25 | OECDPFAS | List of PFAS from the OECD** 3680

   List of PFAS (per- and polyfluoroalkyl substances) from the OECD. PFAS are a group of man-made chemicals that includes PFOA, PFOS, GenX, and many other chemicals.

   ***Classification:***

   NORMAN Suspect List Exchange Classification > S25 | OECDPFAS | List of PFAS from the OECD

2. **S09 | PFASTRIER | PFAS Suspect List of fluorinated substances from X. Trier and colleagues** 517

   A list of PFAS (per- and polyfluoroalkyl substances) kindly supplied by Xenia Trier, David Lunderberg and colleagues. Detailed reference information in source files. Dataset DOI: 10.5281/zenodo.2621988

   ***Classification:***

   NORMAN Suspect List Exchange Classification > S09 | PFASTRIER | PFAS Suspect List of fluorinated substances from X. Trier and colleagues

3. **S14 | KEMIPFAS | PFAS Highly Fluorinated Substances List from KEMI** 1400

   PFAS Highly Fluorinated Substances List, taken from Appendix 2 of the Swedish Chemicals Agency (KEMI) Report 7/15. Provided by Stellan Fischer, KEMI. Registration and mapping to CompTox Dashboard by Antony Williams, US EPA. Dataset DOI: 10.5281/zenodo.2621524

   ***Classification:***

   NORMAN Suspect List Exchange Classification > S14 | KEMIPFAS | PFAS Highly Fluorinated Substances List from KEMI

4. **S25 | OECDPFAS | List of PFAS from the OECD** 3770

   List of PFAS (per- and polyfluoroalkyl substances) from the OECD, provided by Zhanyun Wang. PFAS are a group of man-made chemicals that includes PFOA, PFOS, GenX, and many other chemicals. Details in the Zenodo dataset. Extensive registration and curation performed by CompTox Dashboard team (see CompTox files). Dataset DOI: 10.5281/zenodo.2648775

   ***Classification:***

# NORMAN-SLE on PubChem

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

- NORMAN Suspect List Exchange Classification ? ↗ 133,878
  - S25 | OECDPFAS | List of PFAS from the OECD ? 3,680
    - Linear vs Branched Isomer 3,680
    - Non-polymer vs Polymer 3,680
    - Number of Functional Groups 3,680
    - Perfluorocarbon Chain Length 3,656
    - Potential Precursor to PFAAs in Environment or Biota 3,680
    - Single Compound vs Mixture 3,680
    - Sources 3,679
    - Structure Category 3,680
      - 100 Perfluoroalkyl carbonyl compounds (CnF2n+1-C(=O)-R) 492
      - 200 Perfluoroalkane sulfonyl compounds (CnF2n+1-S(=O)(=O)-R) 458
      - 300 Perfluoroalkyl phosphate compounds (CnF2n+1-P(O)-R) 16
      - 400 Fluorotelomer-related compounds 1,393
      - 500 Per- and polyfluoroalkyl ether-based compounds (CnF2n+1-O-CmF2m+1-R) 322
      - 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 282
      - 700 Other PFAA precursors or related compounds - semifluorinated 716
      - 800 Fluoropolymers 1

# NORMAN-SLE on PubChem

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

- Information in individual records (Classification, Use)
- https://pubchem.ncbi.nlm.nih.gov/compound/139598272#section=Classification

PubChem 1,1,2,2-Tetrafluoro-2-(1,1,2,2,3,3,4,4,5,6,6-undecafluorohexoxy)... (Compound)

## 5 Classification

### 5.1 Ontologies

#### 5.1.1 NORMAN Suspect List Exchange Classification

Showing 1 of 1 — View in Classification Browser — Download

**S46 | PFASNTREV19 | List of PFAS reported in Non-Target HRMS Studies from Liu et al 2019**

List of PFAS (per- and polyfluoroalkyl substances) compiled in the non-target high resolution mass spectrometry (HRMS) PFAS review by Liu et al 2019, DOI: 10.1016/j.trac.2019.02.021. MS-ready list prepared by Yanna Liu, Lisa D'Agostino, Emma Schymanski and Jon Martin. Note not all entries have structures. Dataset DOI: 10.5281/zenodo.2656744

LINKED RECORDS

Compounds: 682

CLASSIFICATION (PARENT NODES)

NORMAN Suspect List Exchange Classification

▸ NORMAN Suspect List Exchange

Share Tweet Email

Cite Download

CONTENTS

# NORMAN-SLE on PubChem

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

- Information in individual records (Classification, Use)
- https://pubchem.ncbi.nlm.nih.gov/compound/74483#section=Use-Classification
- https://pubchem.ncbi.nlm.nih.gov/compound/74483#section=NORMAN-Suspect-List-Exchange-Classification&fullscreen=true

## 8.2 Use Classification

OECD Category -> PFAS (per- and polyfluoroalkyl substances)

▼ NORMAN Suspect List Exchange

**Source:** NORMAN Suspect List Exchange
**Record Name:** S25 | OECDPFAS | List of PFAS from the OECD | CID74483
**Description:** The NORMAN network enhances the exchange of information on emerging environmental substances, and encourages the validation and harmonisation of common measurement methods and monitoring tools so that the requirements of risk assessors and risk managers can be better met. It specifically seeks both to promote and to benefit from the synergies between research teams from different countries in the field of emerging substances.

PUBCHEM > PERFLUOROOCTANESULFO... > NORMAN SUSPECT LIST EXCHANGE CLASSIFICATION

**202 perfluoroalkane sulfonic acids (PFSAs), their salts and esters (R = OH, ONa, OCH3, etc.)**

CID

LINKED RECORDS

Compounds: 117

CLASSIFICATION (PARENT NODES)

NORMAN Suspect List Exchange Classification > S25 | OECDPFAS | List Of PFAS Fr... > Structure Category > 200 Perfluoroalkane Sulfonyl Co...

# But what do we do with all this information?

collaborative

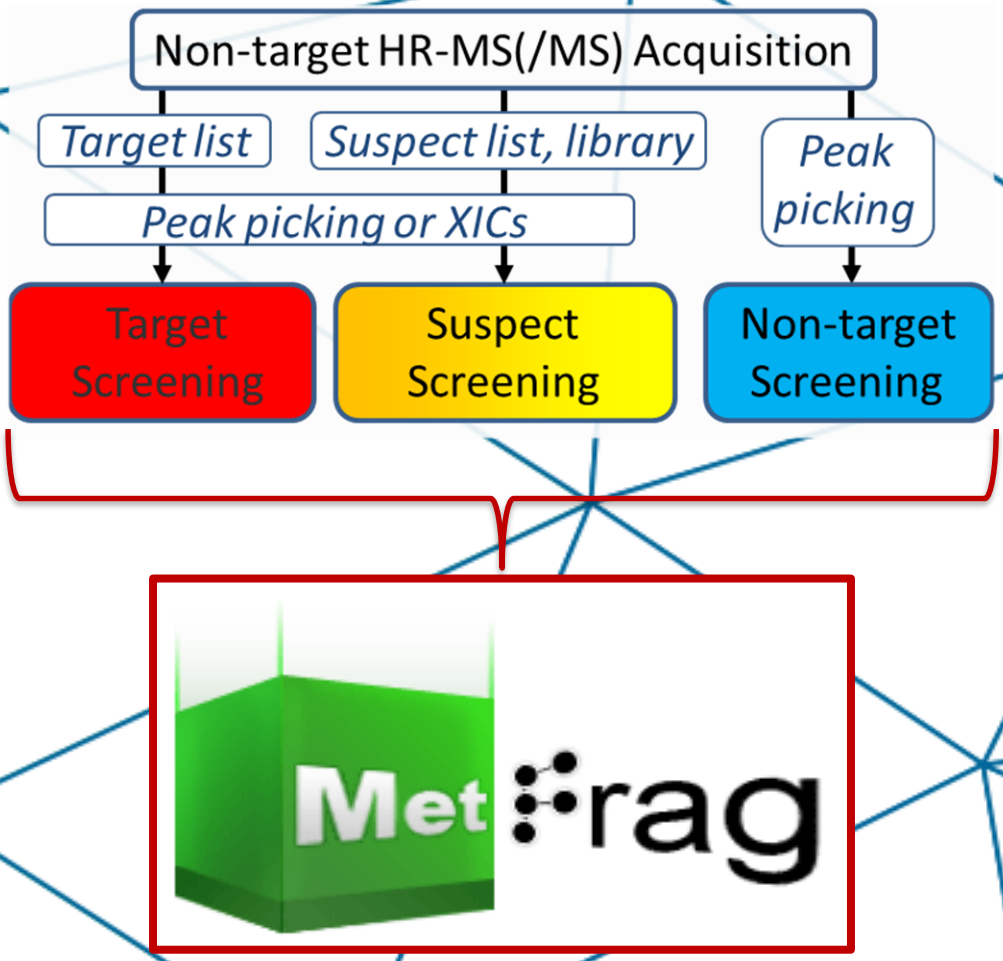



Image: https://www.wikidata.org/wiki/Wikidata:Main_Page

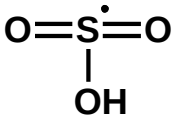
# MetFrag – Adding RT + MetaData



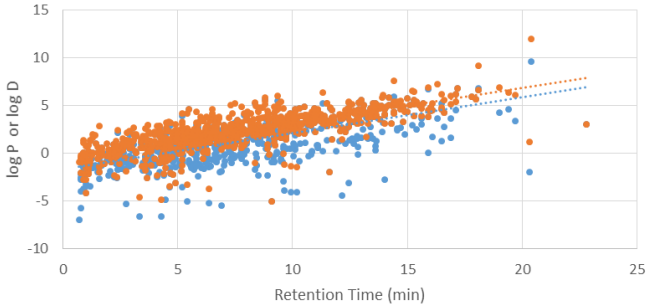
*2010 => 2016*

*mz* [M-H]⁻
213.9637
± 5 ppm

***Elements***: C,N,S
5 ppm
0.001 Da




***RT***: 4.54 min
355 InChI/RTs




ChemSpider
Search and share chemistry

or

PubChem

References
External Refs
Data Sources
RSC Count
PubMed Count

Literature Patents

MetFrag

MoNA
MassBank of North America
MassBank.eu

Suspect Lists

TOFF IDENT

norman suspects

EPA Chemicals

**MS/MS**

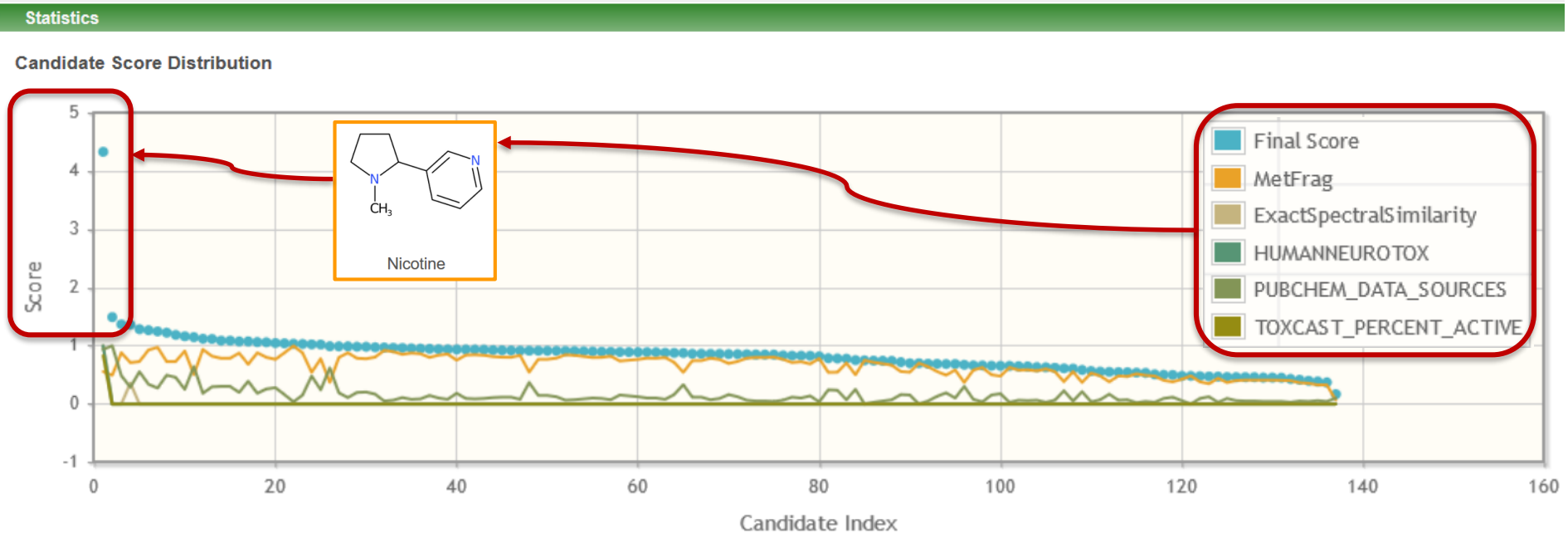
| | |
|---|---|
| 134.0054 | 339689 |
| 150.0001 | 77271 |
| 213.9607 | 632466 |



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

# Connecting multiple lines of evidence for identification

MetFrag · MassBank.eu · norman suspects · PubChem · EPA Chemicals

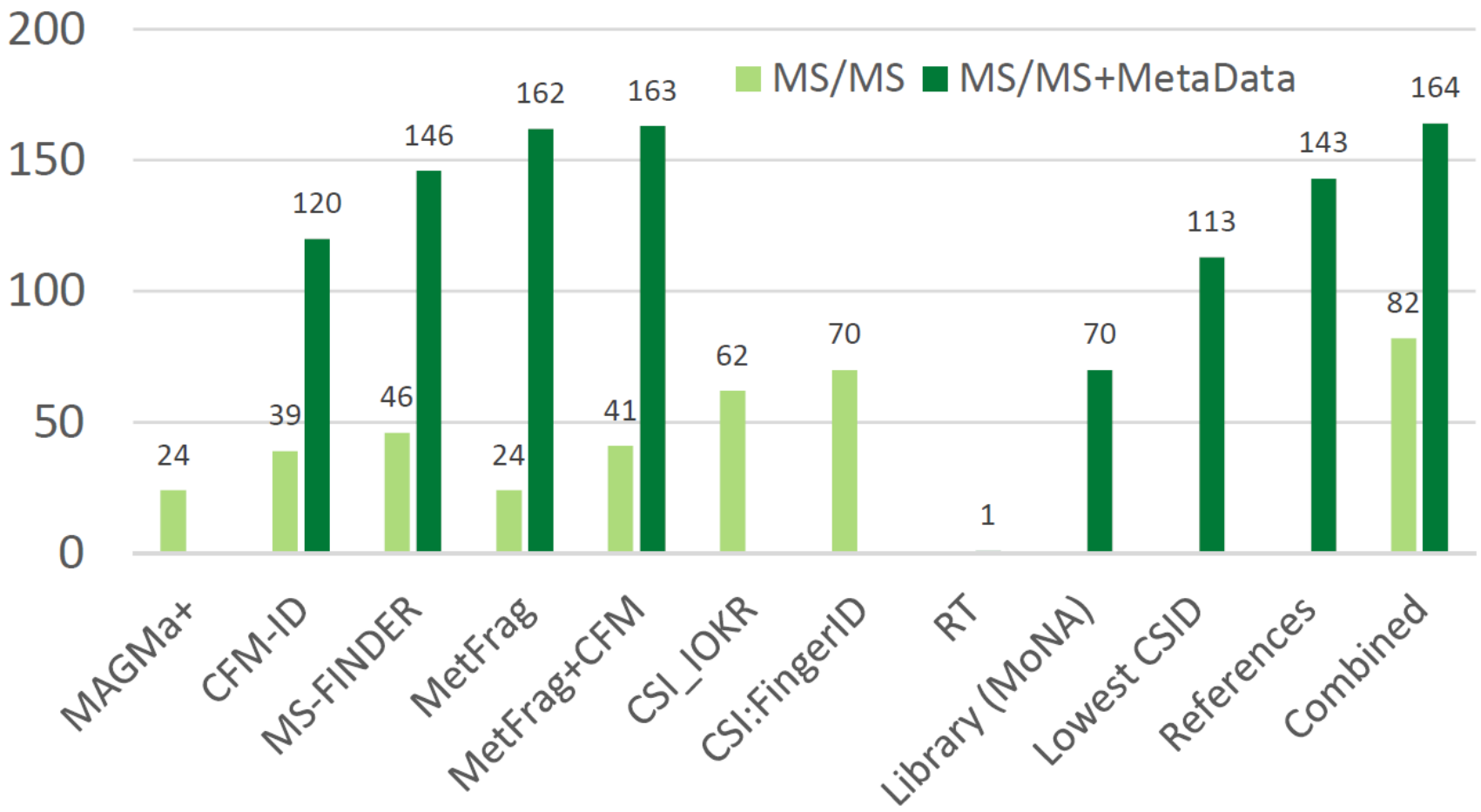





Data: Schymanski *et al* 2014, DOI: 10.1021/es4044374; https://www.slideshare.net/EmmaSchymanski/small-molecules-in-big-data-analytica-munich

# State of the Art in Small Molecule Identification

**Metadata** is critical to improving annotation of known unknowns!

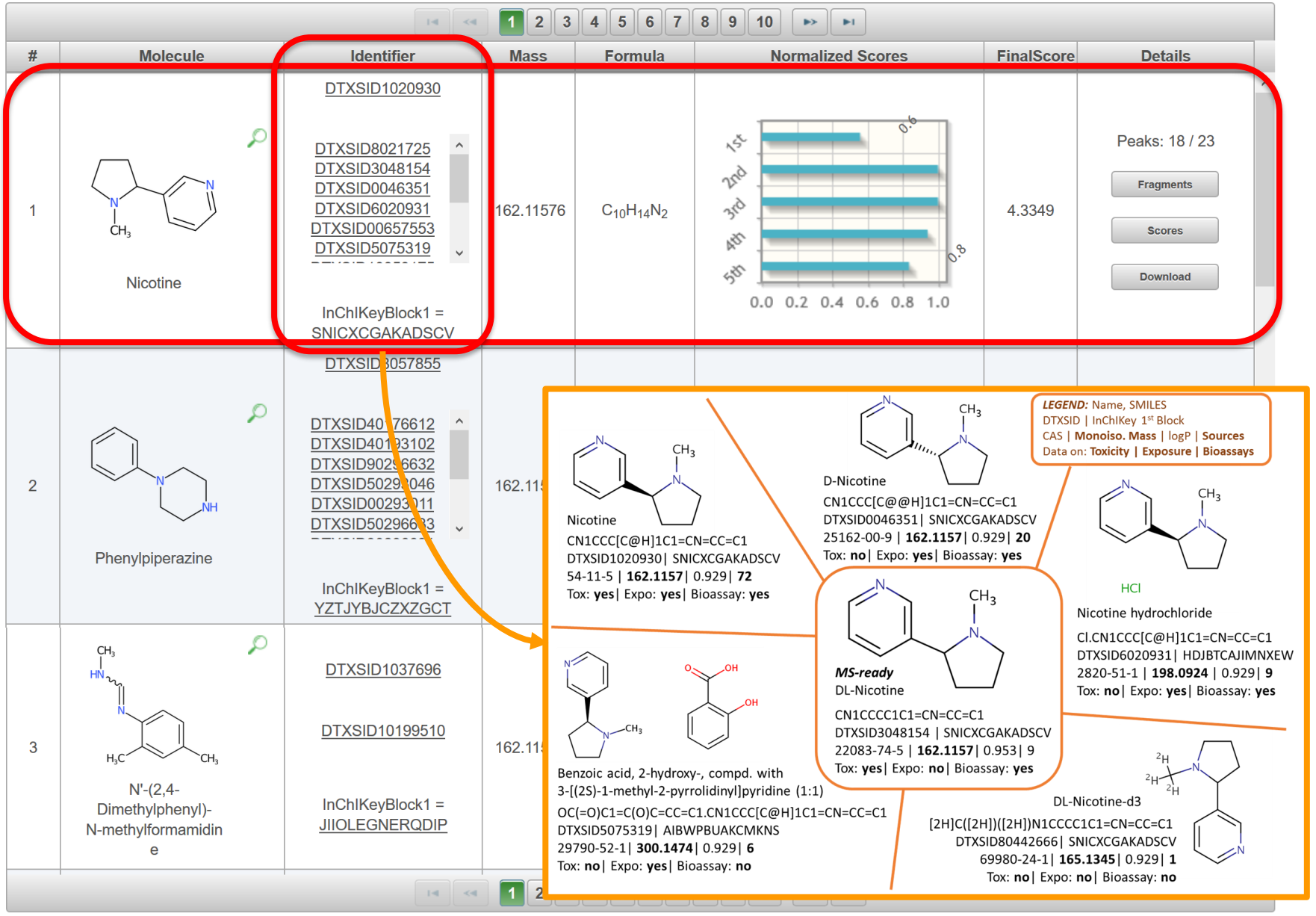





Schymanski *et al*, 2017, J Cheminf., DOI: 10.1186/s13321-017-0207-1

www.casmi-contest.org

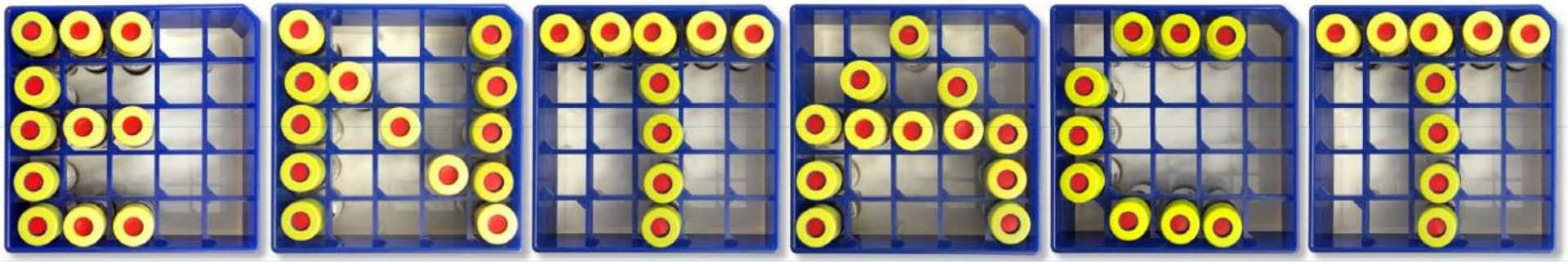
# “MS-ready” Form for MetaData in MetFrag



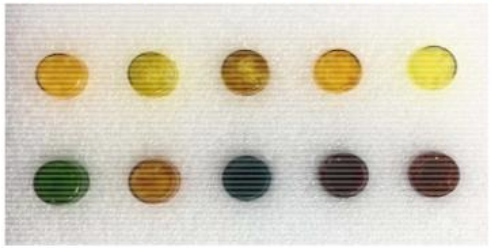



McEachran *et al*. 2018, DOI: 10.1186/s13321-018-0299-2; Schymanski & Williams, 2017 *ES&T* DOI: 10.1021/acs.est.7b01908

# US EPA ENTACT Trial (2016-18)



95, 185 or 365 substances/mixture



with RMassBank and MetFrag

**85-96 %** coverage over both modes

- High potential for successful automated identification
- Mixes 499-506: **83-91 % Level 1 IP4.5**

| Mix | 499 | 500 | 501 | 502 | 503 | 504 | 505 | 506 | 507 | 508 |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Present | 95 | 95 | 95 | 95 | 185 | 185 | 365 | 365 | 95 | 365* |
| *% Level 1 IP4.5* | *86%* | *86%* | *91%* | *83%* | *86%* | *86%* | *88%* | *87%* | *42%* | *38%* |
| *% Level 1 IP4.5 and 2* | *87%* | *87%* | *91%* | *85%* | *86%* | *86%* | *89%* | *89%* | *43%* | *38%* |
| *% All Levels* | *89%* | *89%* | *91%* | *85%* | *89%* | *86%* | *94%* | *89%* | *96%* | *90%* |
| | | | | | | | | | | |
| *%NDs* | *11%* | *11%* | *9%* | *15%* | *11%* | *14%* | *6%* | *10%* | *4%* | *7%* |
| *%Level1 IP4.5 Both* | *35%* | *23%* | *33%* | *21%* | *26%* | *23%* | *31%* | *25%* | *16%* | *16%* |

Elin Ulrich *et al.* 2018, https://www.epa.gov/sites/production/files/2018-06/documents/comptox_cop_6-28-18.pdf

# Automated Level Assignment for THS NTs

UNIVERSITAT ROVIRA i VIRGILI

Schymanski, Torres & Ramirez. (2019, May). Zenodo DOI: 10.5281/zenodo.3046373

| ID | mz | Name | RT | Int | MS/MS | RT_RMB | #Cand | MaxScore | SMILES | Name_maxScore | ExplPeaks | max #Peaks | #Peaks | max MoNA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3027 | 163.1221 | FT0532 | 1.073 | 9E+07 | TRUE | 1.052 | 137 | 7.270015 | CN1CCC | Nicotine | 58.0658_9 | 9 | 28 | 0.99701 |
| 3008 | | | | +07 | TRUE | 1.01 | 41 | 7.22864 | CN1CC(= | Creatinine | 57.0454_1 | 3 | 14 | 0.54699 |
| 3131 | | | | +06 | TRUE | 19.102 | 71 | 7.028138 | CC(C=CC | all-trans-Retinoic acid | 57.0706_5 | 8 | 39 | 0.10337 |
| 3321 | | | | +07 | TRUE | 26.205 | 25 | 6.093 | CCCCCC | Didecyl phthalate | 69.0704_2 | 22 | 71 | 0 |
| 3484 | | | | +06 | TRUE | 24.088 | 1 | 6 | CC(C)(C) | 2,2'-Oxamidodiethyl bis[3- | 57.0706_3 | 4 | 42 | 0 |
| 3206 | 346.1096 | FT3193 | 21.08 | 3E+06 | TRUE | 21.144 | 68 | 5.973483 | CCCN(CC | Nitralin | NA | 9 | 51 | 0 |
| 3044 | | | | +07 | TRUE | 1.073 | 253 | 5.960103 | CCN(CC) | N,N-Diethylnicotinamide | 78.0342_4 | 9 | 44 | 0.01075 |
| 3006 | | | | +06 | TRUE | 1.287 | 20 | 5.94264 | OCCOCC | Diethylene glycol | NA | 0 | 8 | 0 |
| 3043 | | | | +07 | TRUE | 3.163 | 253 | 5.721703 | CCN(CC) | N,N-Diethylnicotinamide | 84.0811_1 | 7 | 24 | 0.00209 |
| 3055 | | | | +07 | TRUE | 11.202 | 114 | 5.712626 | COC1=C( | Scopoletin | 53.0393_1 | 9 | 28 | 0.94271 |
| 3046 | 183.0796 | FT0741 | 24.77 | 1E+07 | TRUE | 24.744 | 72 | 5.689798 | O=C(C1= | Benzophenone | 50.0159_2 | 5 | 45 | 0 |
| 3039 | | | | +07 | TRUE | 7.635 | 219 | 5.467659 | CN1C(CC | Cotinine | 53.0393_1 | 9 | 23 | 0.9987 |
| 3038 | | | | +07 | TRUE | 8.017 | 219 | 5.354236 | CC1CN(N | 4-Methyl-1-phenylpyrazoli | 53.0393_2 | 9 | 40 | 0.66408 |
| 3183 | | | | +07 | TRUE | 19.655 | 74 | 4.928842 | CCCCCC | Dihexyl phthalate | 54.0452_2 | 9 | 51 | 0 |
| 3095 | | | | +08 | TRUE | 21.548 | 74 | 4.860782 | CC(C)CO | Triisobutyl phosphate | 57.0707_8 | 5 | 15 | 0.99613 |
| 3020 | 151.1111 | FT0427 | 13.16 | 2E+06 | FALSE | | 307 | 3.943886 | CC(C)(C) | 4-tert-Butylphenol | NA | 0 | 0 | 0 |
| 3029 | | | | +05 | FALSE | | 242 | 3.021459 | COC1=CC | 1,2,4-Trimethoxybenzene | NA | 0 | 0 | 0 |
| 3123 | | | | +06 | FALSE | | 18 | 3 | CCCCCCC | Octadecyl isocyanate | NA | 0 | 0 | 0 |
| 3128 | | | | +06 | FALSE | | 13 | 3 | CCCCCCC | Octadecanoic acid, hydraz | NA | 0 | 0 | 0 |
| 3172 | | | | +06 | FALSE | | 15 | 3 | CCCCCCC | Stearylbenzene | NA | 0 | 0 | 0 |

Level 1 Target

Level 2 MSMS Match

Level 3 Tentative ID

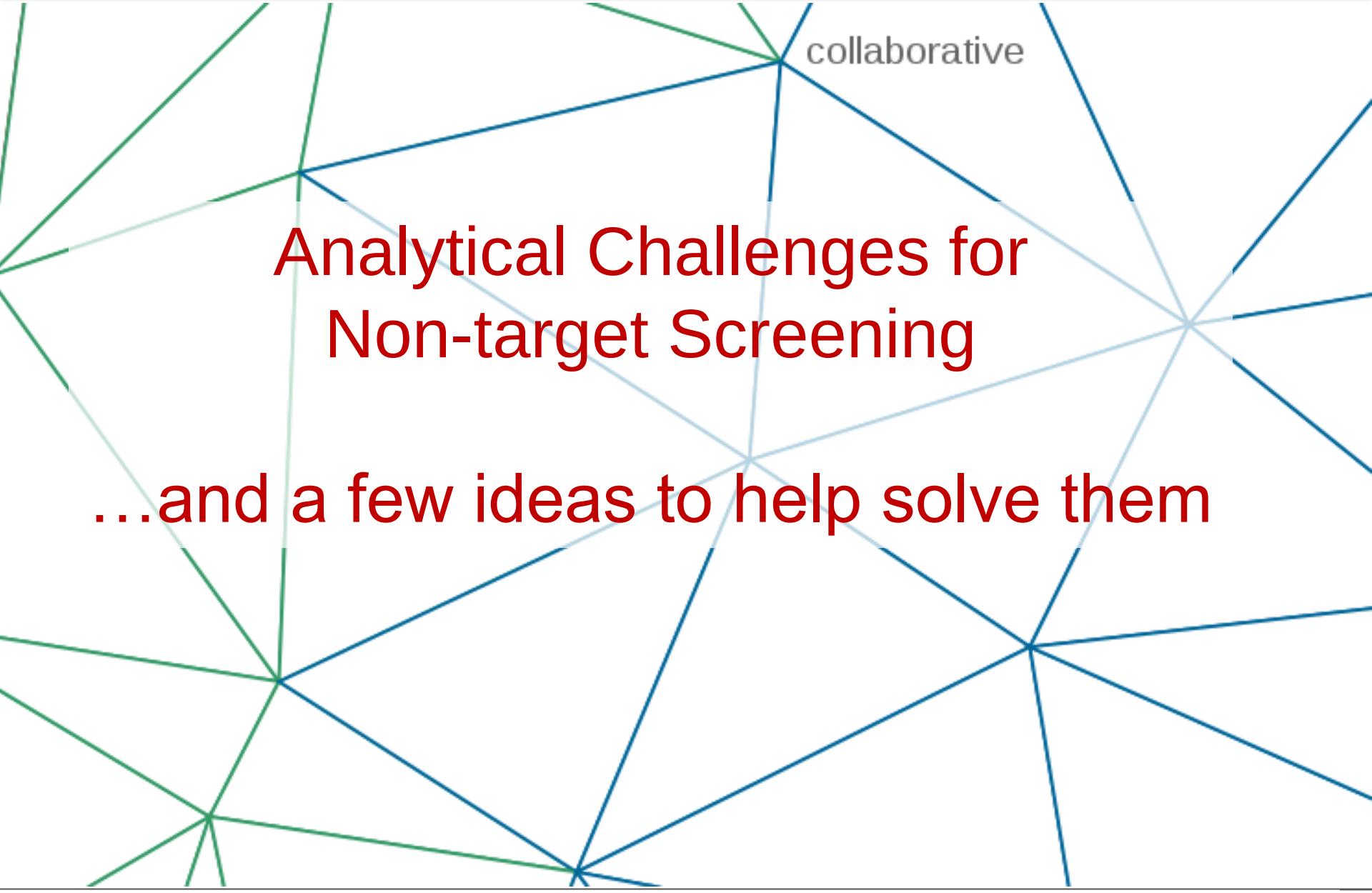
Level 5 No MS/MS

⇒ Approaching automatic assignment of confidence levels

⇒ Quick, high throughput prioritization for data reacquisition



TIBP well documented in dust, e.g. Rostkowski *et al* 2019 DOI: 10.1007/s00216-019-01615-6 and DOI: 10.5281/zenodo.2653206




# Analytical Challenges for Non-target Screening

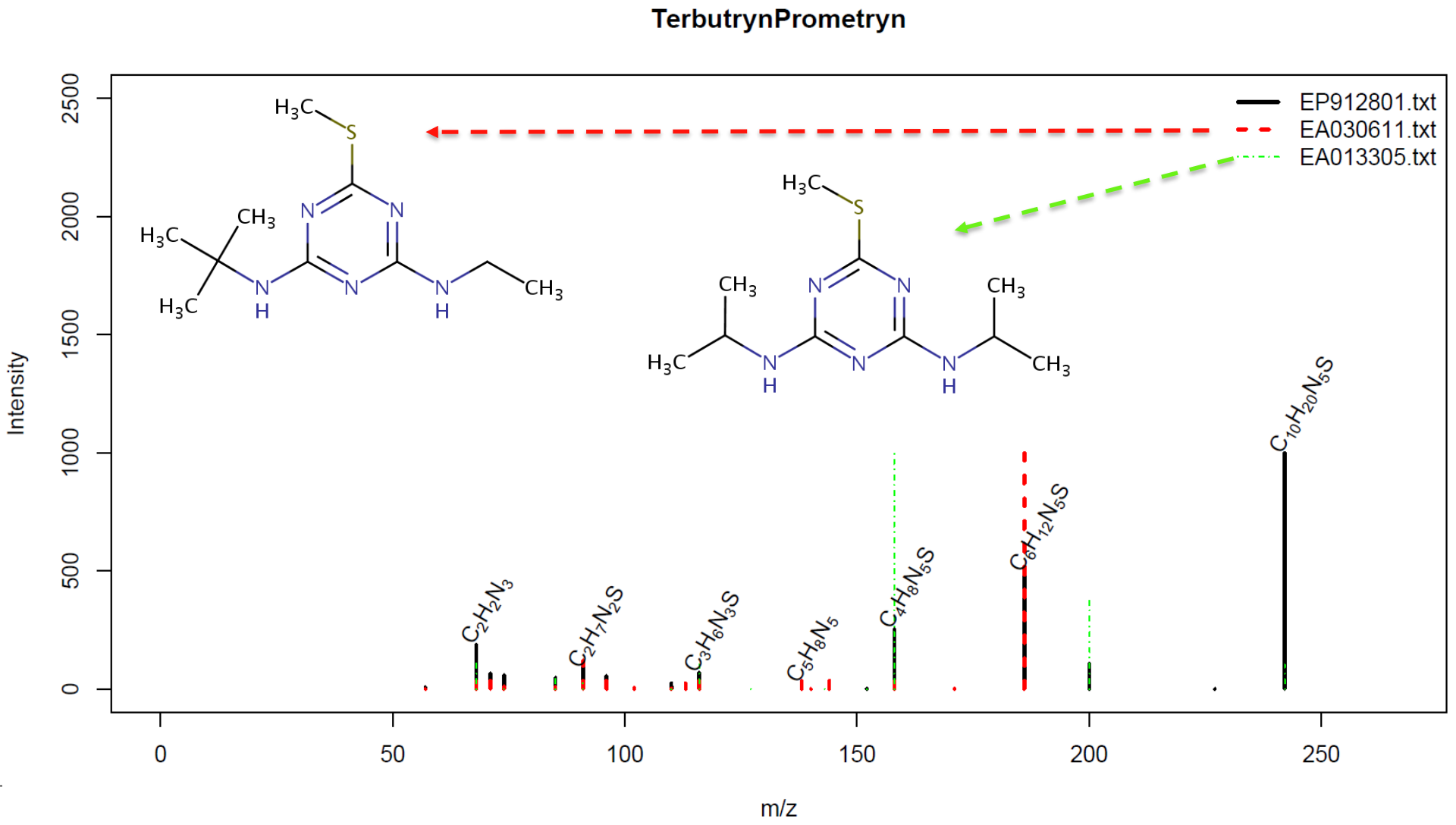
## …and a few ideas to help solve them

Image: https://www.wikidata.org/wiki/Wikidata:Main_Page

# ENTACT Observation: MS/MS for co-eluting peaks

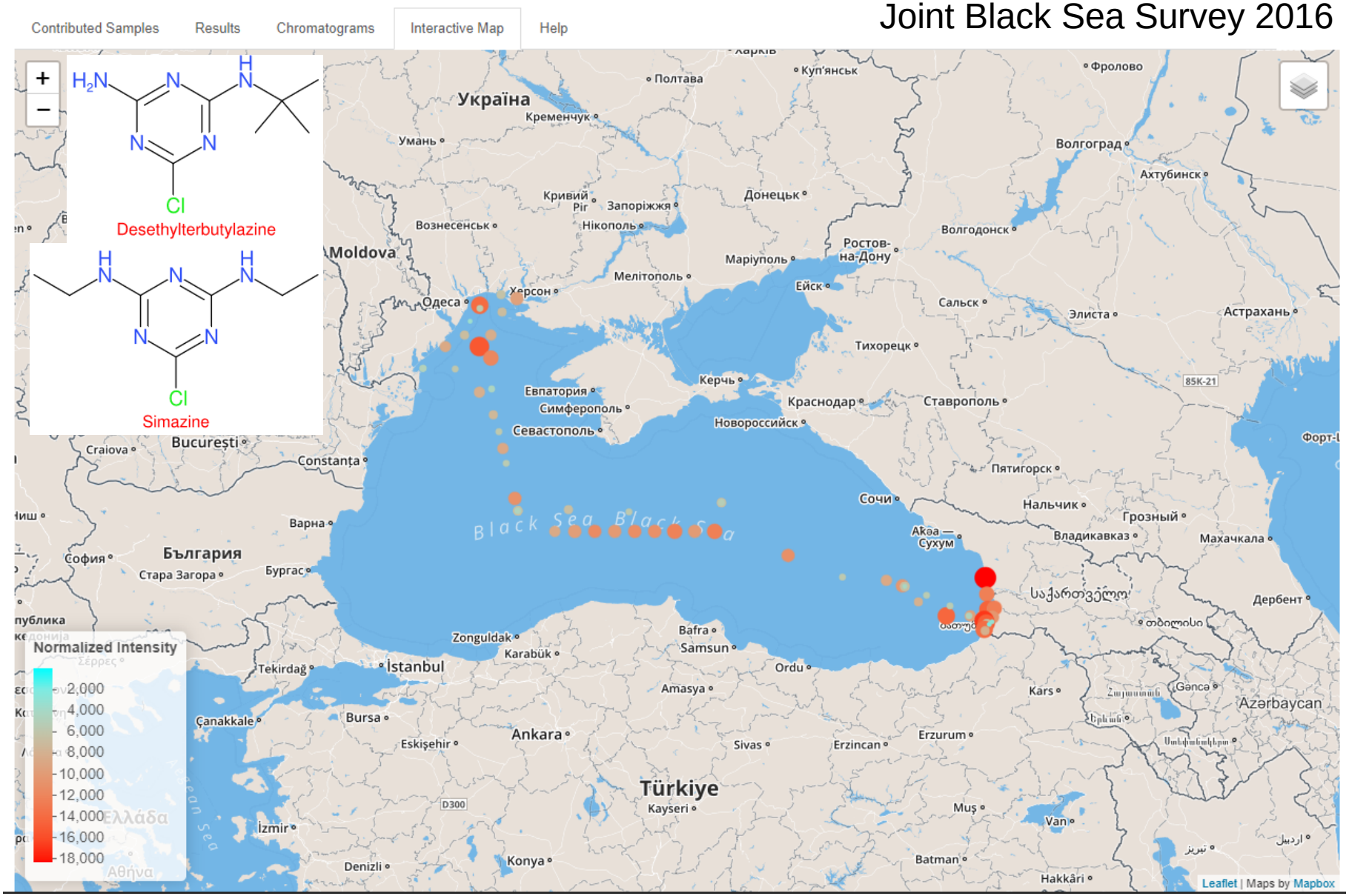
Identical retention time for reference standards … “combined” match value

*Still a challenge for automating NTS*



# Isobars in Automated Screening

Joint Black Sea Survey 2016

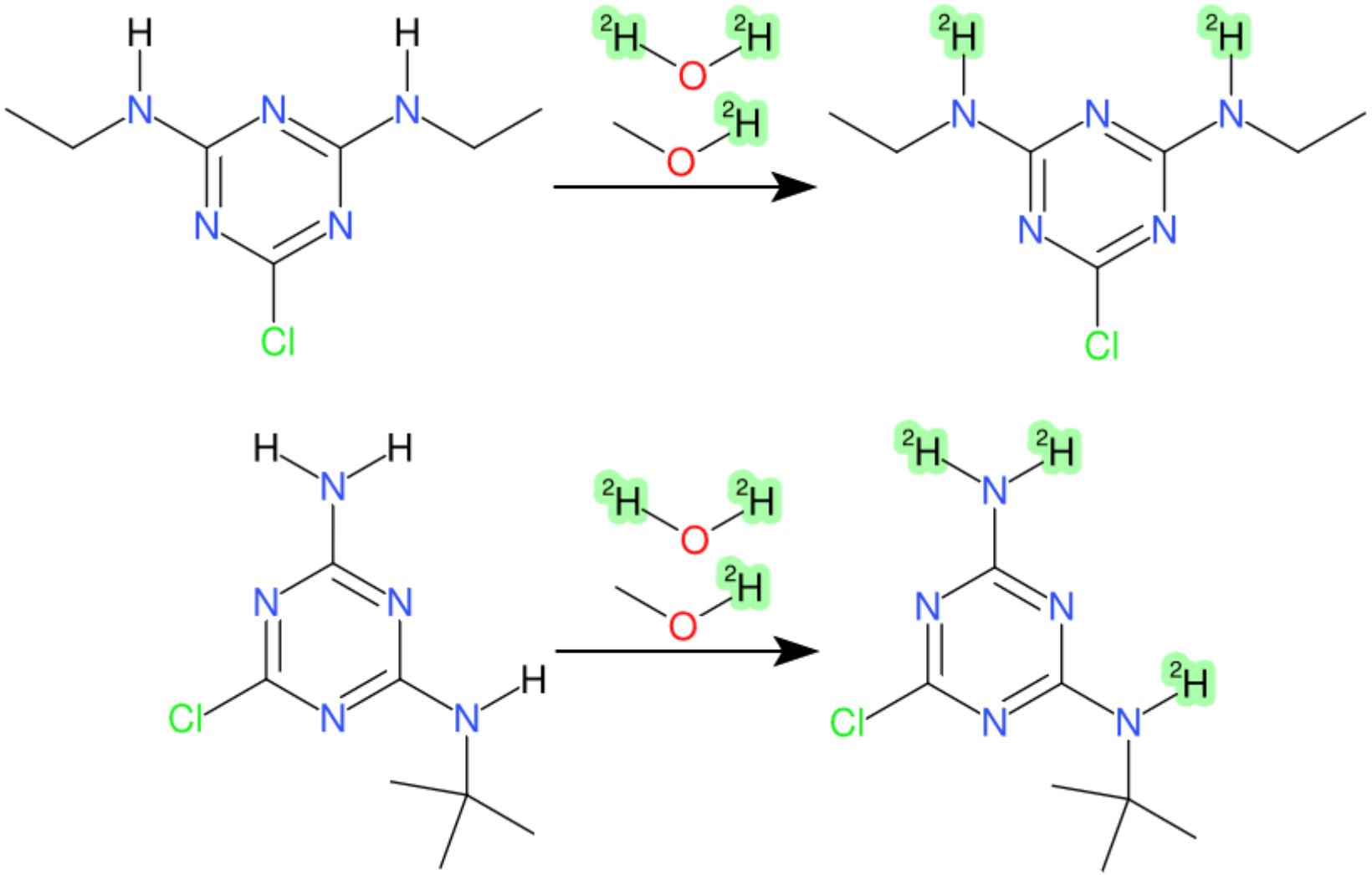





NORMAN Digital Sample Freezing Platform Image provided by Nikiforos Alygizakis

# Separating/Identifying Co-eluting Peaks - HDX

- Hydrogen Deuterium Exchange

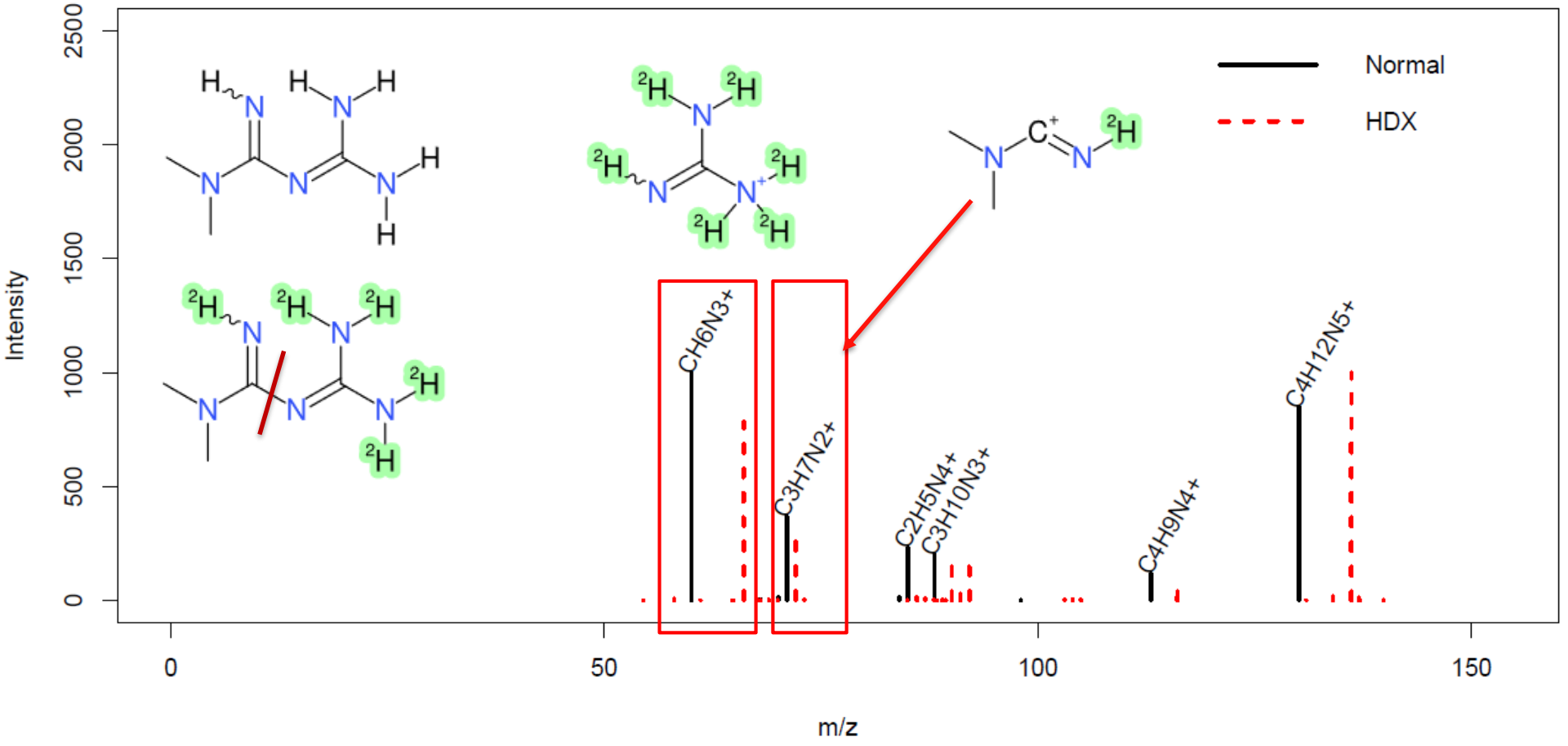





Ruttkies, Schymanski, *et al*. 2019, Anal. Bioanal. Chem. (Paper in Forefront). DOI: 10.1007/s00216-019-01885-0

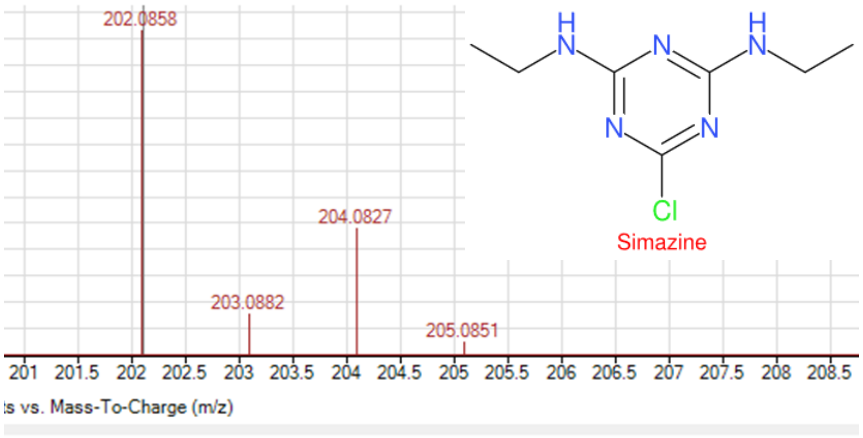
# Hydrogen Deuterium Exchange

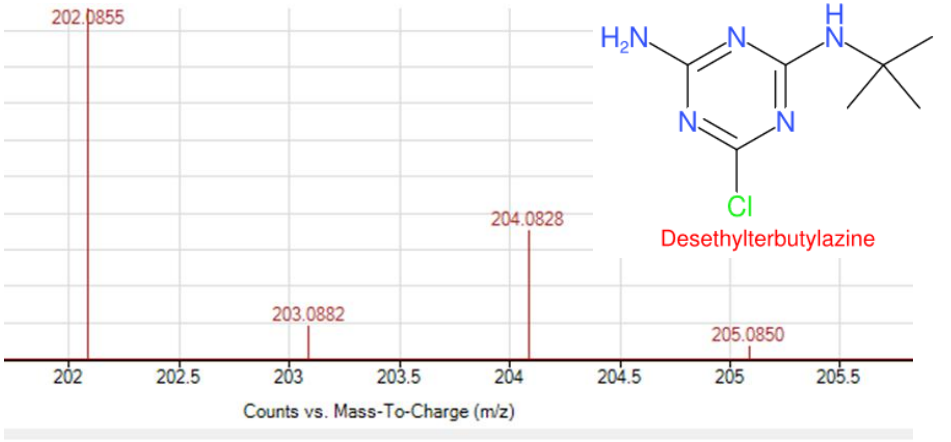
- Evaluated on 762 compounds of environmental interest
  - See reference below for more
- Example with Metformin

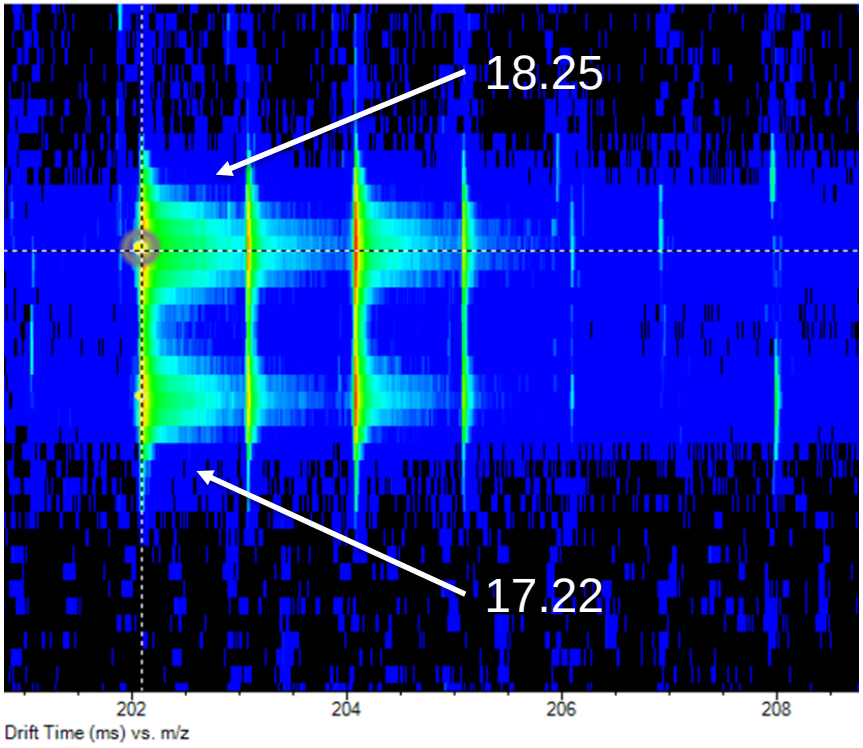


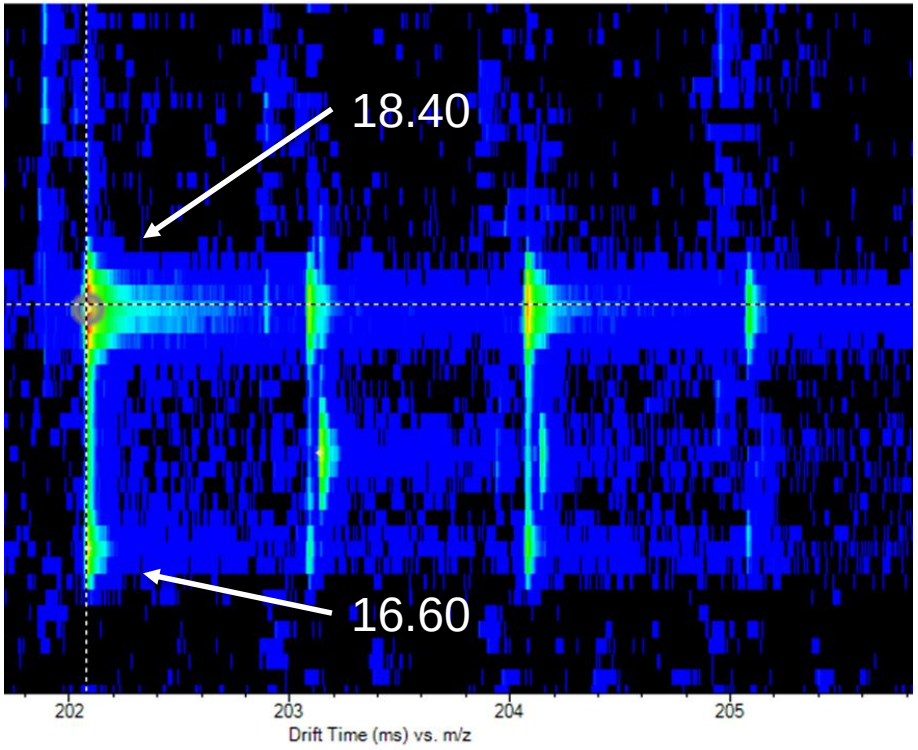

Ruttkies, Schymanski, *et al*. 2019, Anal. Bioanal. Chem. (Paper in Forefront). DOI: 10.1007/s00216-019-01885-0

# Ion Mobility Separation






Images provided by Randolph Singh, LCSB-ECI

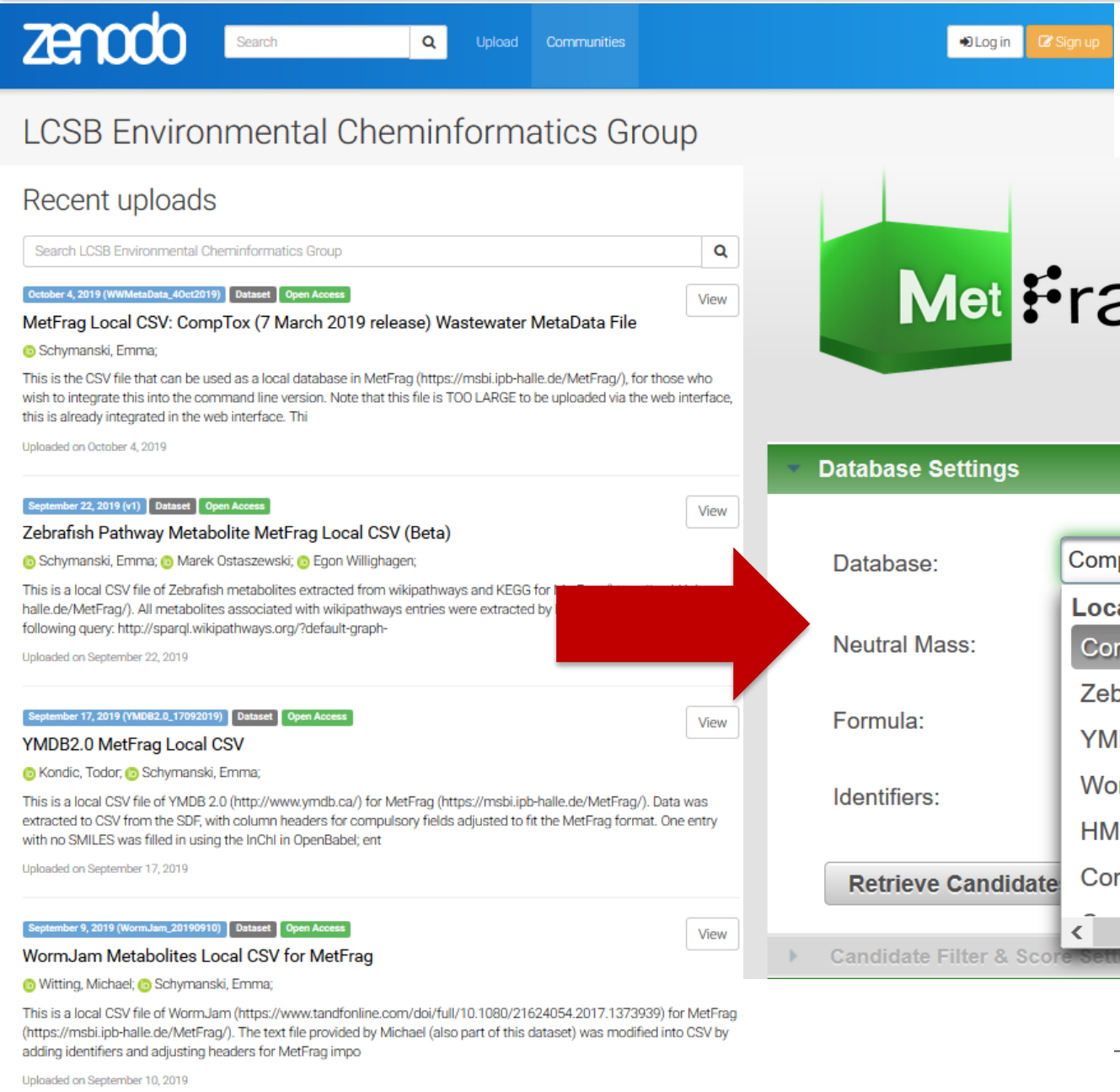
collaborative

# Data Challenges for Non-target Screening

## …and a few ideas to help solve them

Image: https://www.wikidata.org/wiki/Wikidata:Main_Page

# Empowering MetFrag for a Broader Audience



# New MetaData: Disease-Specific Reference Counts

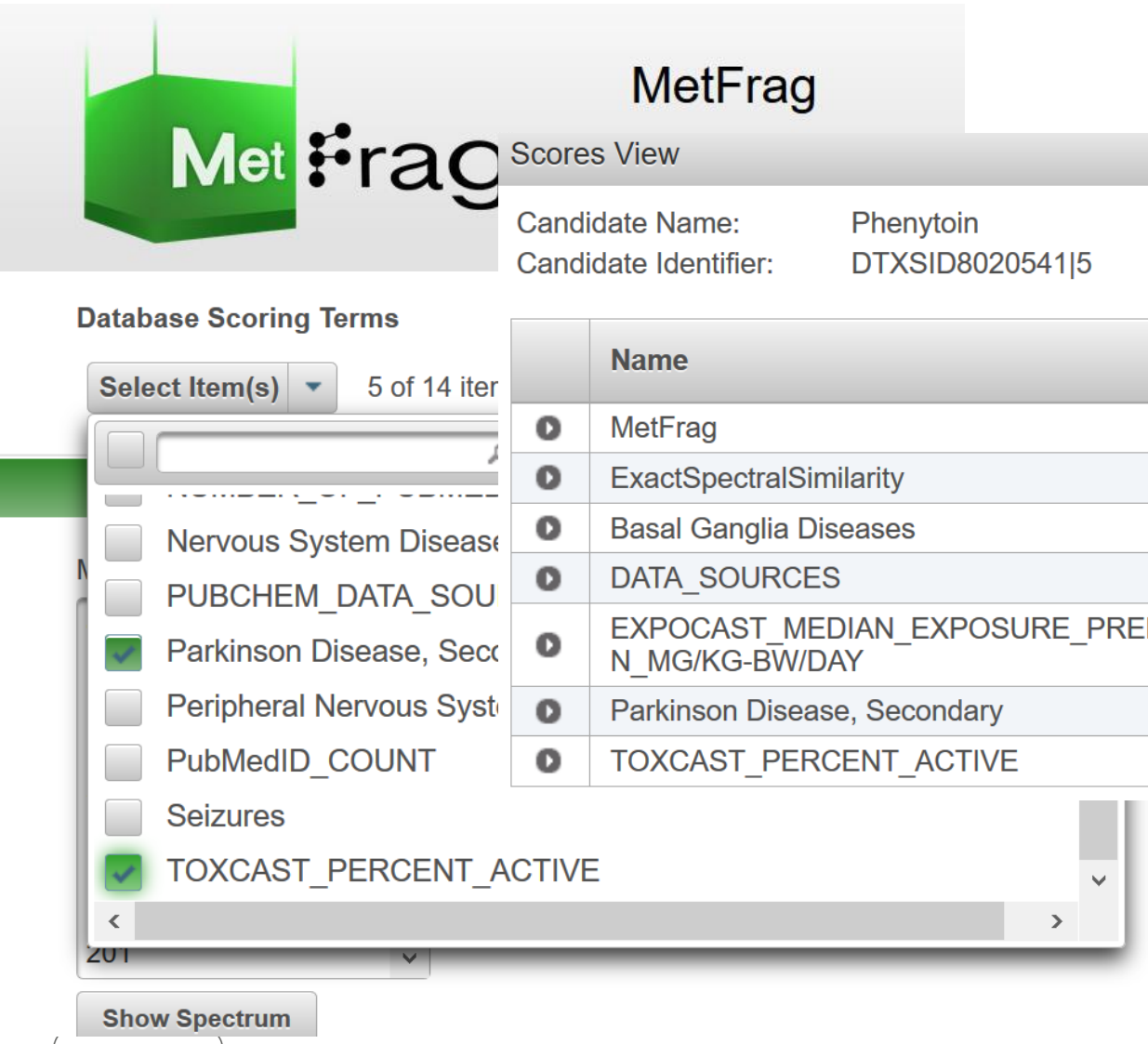
https://comptox.epa.gov/dashboard/chemical_lists/litminedneuro

| 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 |
| Lithium | 7439-93-2 | DTXSID5036761 | 689 | 30 | 50 | 9 | 22 | 5 | 36 | 13 | 25 | 6 | 93 | 12 | 15 |
| Valproic Acid | 76584-70-8 | DTXSID70227388 | 666 | 32 | 10 | 3 | 65 | 6 | 10 | 18 | 45 | 5 | 18 | 4 | 2 |
| 1-Methyl-4-phe | 28289-54-5 | DTXSID8040933 | 638 | 1 | 24 | 0 | 11 | 0 | 6 | 289 | 0 | 0 | 5 | 0 | 1 |
| Vincristine | 2068-78-2 | DTXSID8044331 | 567 | 17 | 59 | 125 | 15 | 5 | 1 | 1 | 5 | 3 | 2 | 1 | 8 |
| Phenytoin | 57-41-0 | DTXSID8020541 | 560 | 37 | 24 | 25 | 16 | 9 | 3 | 1 | 9 | 3 | 8 | 4 | 6 |
| Haloperidol | 52-86-8 | DTXSID4034150 | 555 | 6 | 6 | 1 | 10 | 6 | 153 | 51 | 4 | 4 | 11 | 1 | 0 |
| Cocaine | 50-36-2 | DTXSID2038443 | 530 | 151 | 16 | 0 | 8 | 0 | 2 | 3 | 3 | 8 | 6 | 12 | 11 |
| Aspirin | 50-78-2 | DTXSID5020108 | 489 | 8 | 3 | 0 | 3 | 2 | 2 | 0 | 9 | 4 | 1 | 0 | 5 |
| Paclitaxel | 33069-62-4 | DTXSID9023413 | 485 | 4 | 43 | 217 | 9 | 14 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| Aluminum | 7429-90-5 | DTXSID3040273 | 477 | 13 | 41 | 1 | 105 | 4 | 0 | 0 | 1 | 0 | 1 | 13 | 12 |
| Lidocaine | 6108-05-0 | DTXSID80209953 | 464 | 150 | 26 | 15 | 3 | 2 | 0 | 0 | 8 | 4 | 6 | 2 | 10 |
| Methotrexate | 59-05-2 | DTXSID4020822 | 451 | 17 | 25 | 1 | 79 | 4 | 0 | 1 | 5 | 0 | 1 | 9 | 18 |
| Mercury | 7439-97-6 | DTXSID1024172 | 450 | 6 | 79 | 22 | 23 | 2 | 3 | 5 | 2 | 2 | 38 | 7 | 25 |



Schymanski *et al.* (Perspective) Environ. Sci.: Processes Impacts, 2019, DOI: 10.1039/C9EM00068B [cover image]

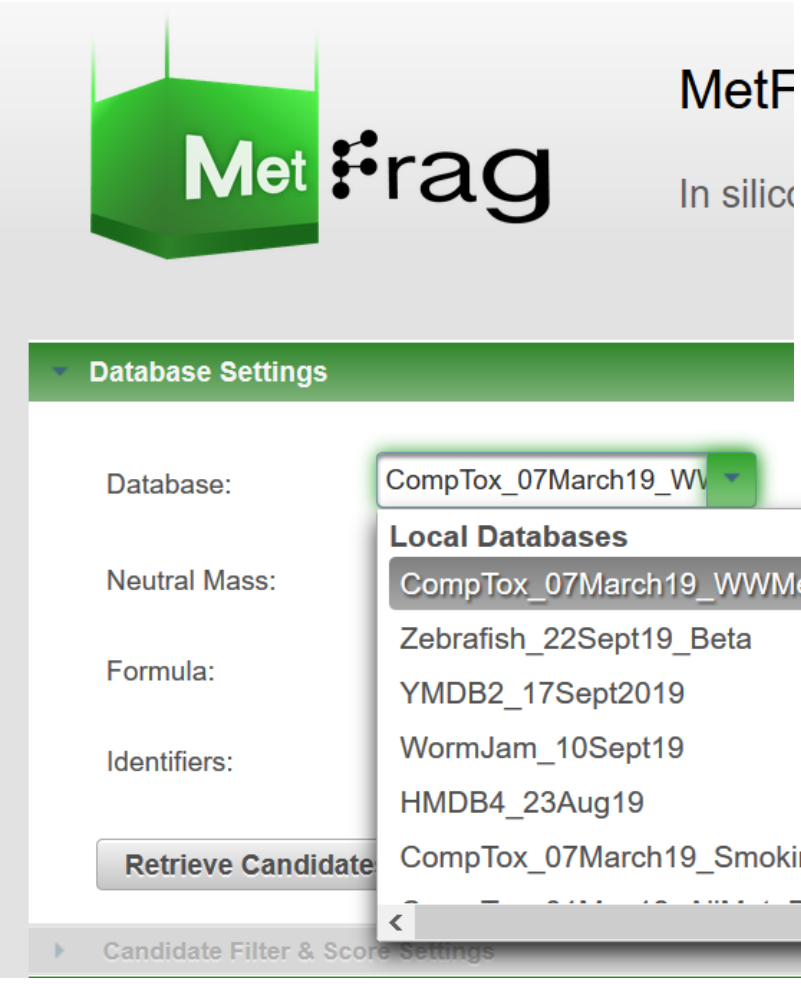
# New MetaData: Disease-Specific Reference Counts

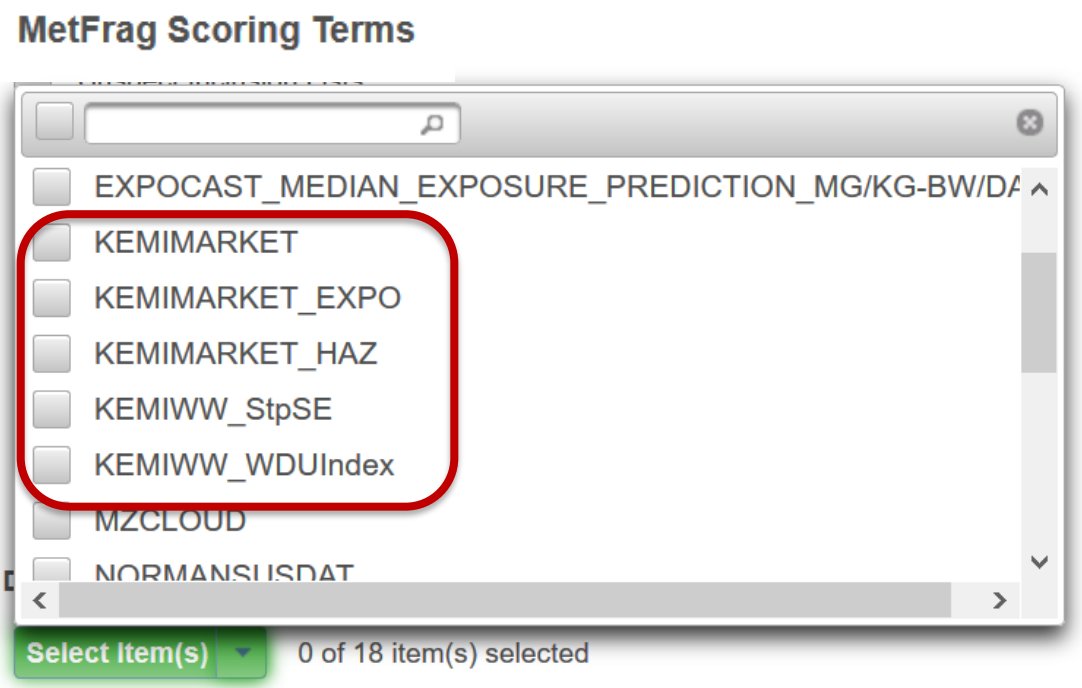



| | Name | Normalized Value | Raw Value |
|---|---|---|---|
| | MetFrag | 1.0 | 92.9258 |
| | ExactSpectralSimilarity | 1.0 | 1.0 |
| | Basal Ganglia Diseases | 1.0 | 3.0 |
| | DATA_SOURCES | 1.0 | 105.0 |
| | EXPOCAST_MEDIAN_EXPOSURE_PREDICTION_MG/KG-BW/DAY | 0.2006 | 0.0 |
| | Parkinson Disease, Secondary | 1.0 | 1.0 |
| | TOXCAST_PERCENT_ACTIVE | 1.0 | 2.5605 |

# Beyond “just” Suspect Lists: KEMI Metadata Scores

KEMI
Kemikalieinspektionen




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

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

UNIVERSITY OF
COPENHAGEN

MEDIZINISCHE
UNIVERSITÄT
INNSBRUCK

# Beyond “just” Suspect Lists: KEMI Metadata Scores

KEMI
Kemikalieinspektionen

## Scores View

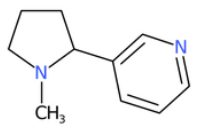
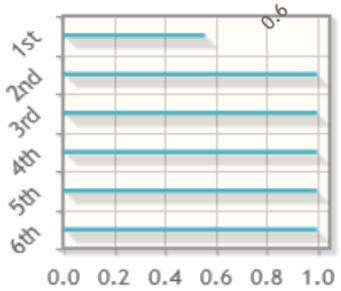
Candidate Name: Nicotine
Candidate Identifier: DTXSID1020930|6575

| | Name | Normalized Value | Raw Value |
|---|---|---|---|
| ⓘ | MetFrag | 0.5585 | 294.7878 |
| ⓘ | ExactSpectralSimilarity | 1.0 | 1.0 |
| ⓘ | KEMIMARKET_EXPO | 1.0 | 19.0 |
| ⓘ | KEMIMARKET_HAZ | 1.0 | 6.0 |
| ⓘ | KEMIWW_StpSE | 1.0 | 5.0 |
| ⓘ | KEMIWW_WDUIndex | 1.0 | 2.0 |

Results

MetFrag (1st)
ExactSpectralSimilarity (2nd)
KEMIMARKET_EXPO (3rd)
KEMIMARKET_HAZ (4th)
KEMIWW_StpSE (5th)
KEMIWW_WDUIndex (6th) 100 %

1 2 3 4 5 6 7 8 9 10

| # | Molecule | Identifier | Mass | Formula | Normalized Scores | FinalScore | Details |
|---|---|---|---|---|---|---|---|
| 1 |  Nicotine | DTXSID1020930<br>DTXSID8021725<br>DTXSID20237235<br>DTXSID60993843<br>DTXSID80983589<br>DTXSID20982596<br>DTXSID60982595<br>InChIKeyBlock1 = SNICXCGAKADSCV | 162.11576 | $C_{10}H_{14}N_2$ |  | 5.5585 | Peaks: 18 / 23<br>Fragments<br>Scores<br>Download |

# The 103 million PubChem Challenge …

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

## 500 masses ≈ 2-3 hrs with CompTox or 2-3 DAYS with PubChem

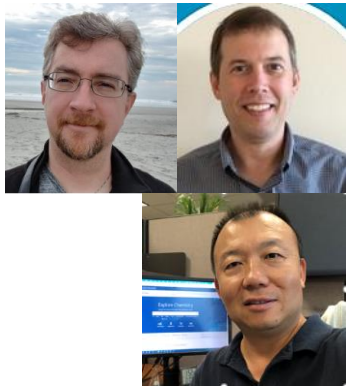
- PubChem Compound TOC ? 33,765,953
  - Agrochemical Information ? 2,002
  - Biologic Description ? 1,539,532
  - Biological Test Results ? 3,467,416
  - Biomolecular Interactions and Pathways ? 109,610
  - Chemical and Physical Properties ? 237,729
  - Classification ? 18,569,627
  - Diseases ?
  - Drug and Medication Information ? 15,955
  - Food Additives and Ingredients ? 7,447
  - Identification ? 5,746
  - Information Sources ? 20,654,780
  - Literature ? 1,668,437
  - Names and Identifiers ? 1,310,169
  - Patents ? 22,144,888
  - Pharmacology and Biochemistry ? 130,367
  - Related Records ? 5,297,096
  - Safety and Hazards ? 125,607
  - Spectral Information ? 761,478
  - Structures ? 5,926,225
  - Toxicity ? 114,554
  - Use and Manufacturing ? 108,745

Chemical Safety ? 122,739

103 million … OR …
the most relevant / annotated?

PubChemLite tier0: 316 K

PubChemLite tier1: 360 K



January 14, 2020 Dataset Open Access

### PubChemLite tier0 and tier1

Bolton, Evan; Schymanski, Emma

PubChemLite is a subset of PubChem (https://pubchem.ncbi.nlm.nih.gov/) selected from major categories of the Table of Contents page at the PubChem Classification Browser (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72). So far we are providing two "flavours":

tier0 is 316,810 compounds (14 Jan 2020) compiled from 7 categories: AgroChemInfo, DrugMedicInfo, FoodRelated, PharmacoInfo, SafetyInfo, ToxicityInfo, KnownUse

tier1 is 363,911 compounds (14 Jan 2020) compiled from 8 categories (tier0 + BioPathway): AgroChemInfo, BioPathway, DrugMedicInfo, FoodRelated, PharmacoInfo, SafetyInfo, ToxicityInfo, KnownUse

PubChemCIDs have been collapsed by InChIKey first block, reporting the structure from the most annotated CID, plus related CIDs. Entries that will be ignored by MetFrag (salts, disconnected substances) or cause errors (e.g. transition metals) have been removed. The Patent and PubMed ID counts are extracted from files on the PubChem FTP site. The "AnnoTypeCount" term counts how many of the categories are represented, the subsequent column (named per category) counts the number of annotation categories available in the next sub-category of the TOC entry.

These files can be used "as is" as localCSV for MetFrag Command Line (https://ipb-halle.github.io/MetFrag/) - please do NOT upload these files directly to the web interface, they are too large and will be available in a drop-down menu.

Edit

New version

**Communities**

LCSB Environmental Cheminformatics Group — Remove

375 views 333 downloads

See more details...

1 Tweeted by 2

See more details

Bolton & Schymanski (2020). PubChemLite tier0 and tier1 (Version PubChemLite.0.2.0) DOI: 10.5281/zenodo.3611238

# The 103 million PubChem Challenge ..

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

MetFrag

In silico fra

## Scores View

Candidate Name: 3-[(2S)-1-methylpyrrolidin-2-yl]pyridine
Candidate Identifier: 89594|47009

| | Name | Normalized Value | Raw Value |
|---|---|---|---|
| ⊙ | MetFrag | 0.6058 | 294.7878 |
| ⊙ | ExactSpectralSimilarity | 1.0 | 1.0 |
| ⊙ | AnnoTypeCount | 1.0 | 6.0 |
| ⊙ | Patent_Count | 1.0 | 78331.0 |
| ⊙ | PubMed_Count | 1.0 | 35157.0 |

**Database Settings**

Database:

Neutral Mass:

Formula:

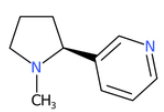
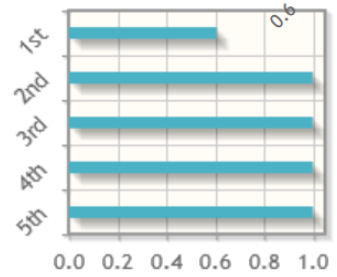
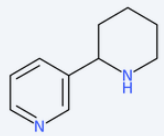
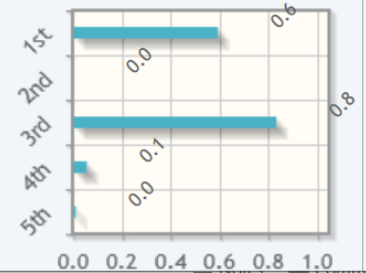
Identifiers:

**Retrieve Candidates**

MetFrag (1st)
ExactSpectralSimilarity (2nd)
AnnoTypeCount (3rd)
Patent_Count (4th)
PubMed_Count (5th)

100 %

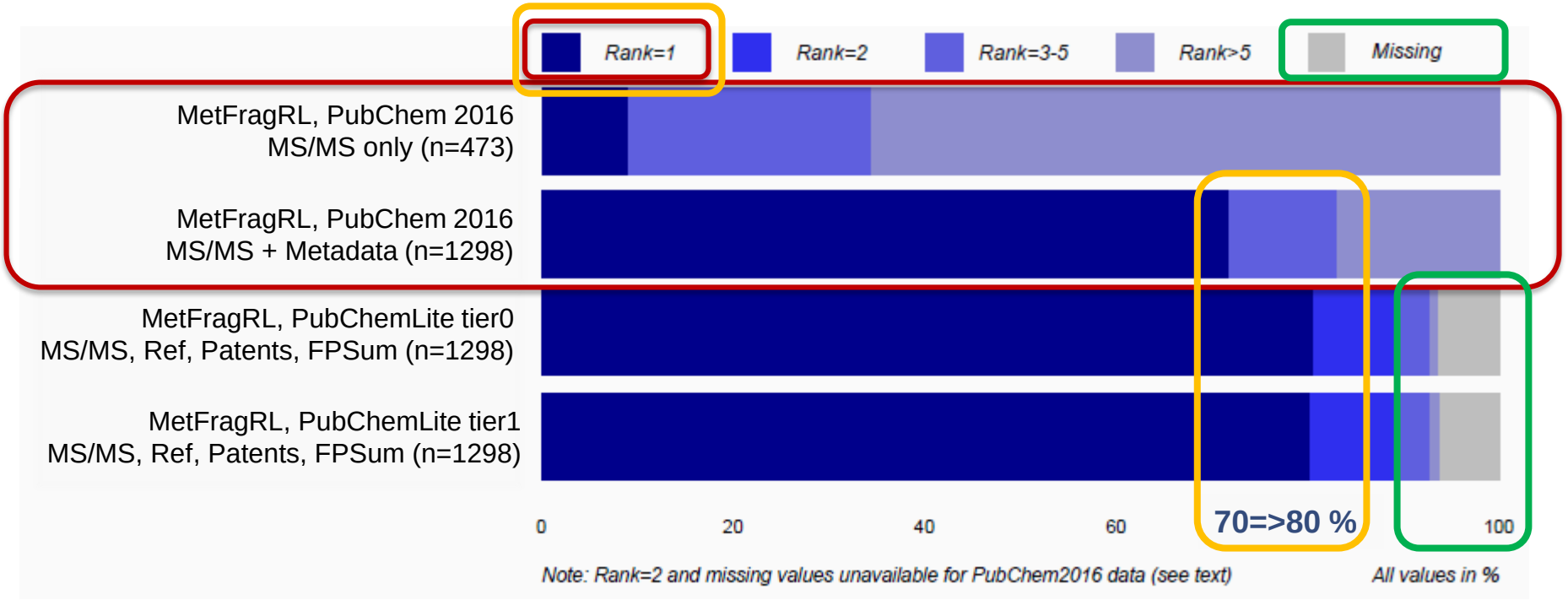
1 2 3 4

| # | Molecule | Identifier | Mass | Formula | Normalized Scores | FinalScor | Details |
|---|---|---|---|---|---|---|---|
| 1 |  3-[(2S)-1-methylpyrrolidin-2-yl]pyridine | 89594 InChIKeyBlock1 = SNICXCGAKADSCV | 162.1157 | $C_{10}H_{14}N_2$ |  | 4.6058 | Peaks: 18 / 23 Fragments Scores Download |
| 2 |  3-piperidin-2-ylpyridine | 2181 InChIKeyBlock1 = MTXSIJUGVMTTMU | 162.1157 | $C_{10}H_{14}N_2$ |  | 1.4939 | Peaks: 18 / 23 Fragments Scores Download |

# Evaluating PubChemLite



- 103 M => 300 K … how does this influence performance?

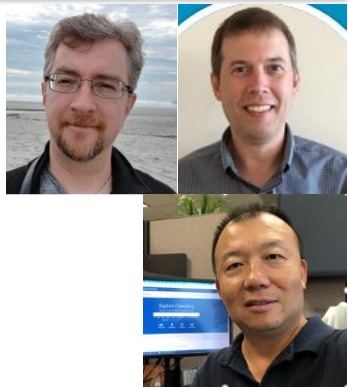



MetFragRL: Ruttkies et al. (2016) DOI: 10.1186/s13321-016-0115-9
Bolton, Schymanski et al., in prep. +
Bolton & Schymanski (2020). PubChemLite tier0 and tier1, DOI: 10.5281/zenodo.3611238

# Transformation Products: Filling the Data Gaps!

PubChem NORMAN Suspect List Exchange

- NORMAN Suspect List Exchange Classification ? 115,510
  - S25 | OECDPFAS | List of PFAS from the OECD ? 3,680
  - S60 | SWISSPEST19 | Swiss Pesticides and Metabolites ? 1,358
    - Pesticide transformation products (TPs, metabolites, successors) ? 1,107
    - Pesticides (parents, predecessors) ? 267
  - S66 | EAWAGTPS | Parent-Transformation Product Pa
  - S00 | SUSDAT | Merged NORMAN Suspect List: SusDat
  - S01 | MASSBANK | NORMAN Compounds in MassBank
  - S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Databas
  - S03 | NORMANCT15 | NORMAN Collaborative Trial Targ
  - S04 | UJIBADE | Target List from UJI used in Bade et al 2
  - S05 | KWRSJERPS | KWR Drinking Water Suspect List
  - S06 | ITNANTIBIOTIC | Antibiotic List from the ITN MSCA
  - S07 | EAWAGSURF | Eawag Surfactants Suspect List ?
  - S08 | ATHENSSUS | University of Athens Surfactants and
  - S09 | PFASTRIER | PFAS Suspect List of fluorinated subs
  - S10 | SWISSPHARMA | Pharmaceutical List with Consun
  - S11 | SWISSPEST | Swiss Insecticides, Fungicides and T
  - S12 | NORMANEWS | NormaNEWS for Retrospective Sc

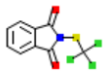
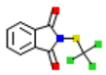
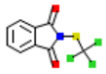


PubChem Folpet (Compound)

## 8.7 Transformations

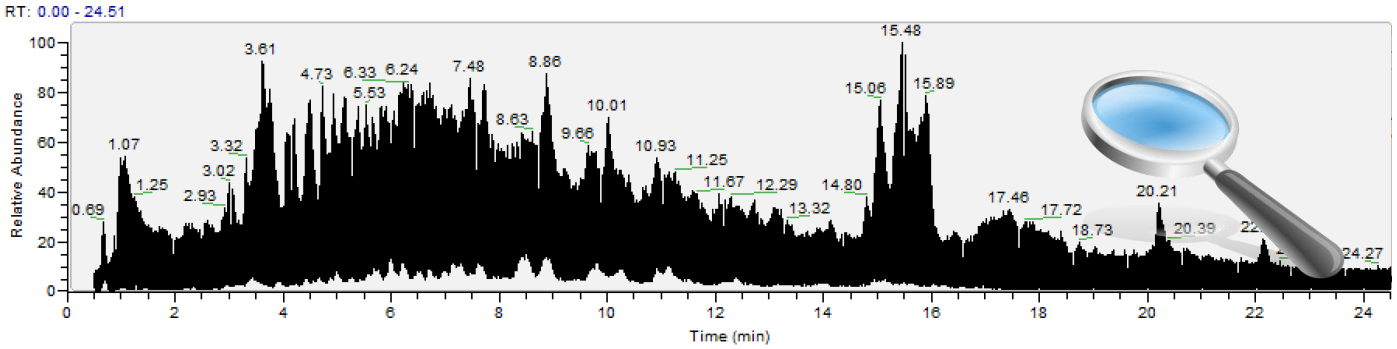
3 items View More Details Download

SORT BY Please Choose One

| Predecessor Image | Predecessor Name | Transformation | Successor Image | Successor Name | Evidence DOI |
|---|---|---|---|---|---|
|  | Folpet | Environmental Transformation |  | Phtalimide | 10.1016/j.watres.2019.114972 |
|  | Folpet | Environmental Transformation |  | Phthalamic acid | 10.1016/j.watres.2019.114972 |
|  | Folpet | Environmental Transformation |  | Phthalic acid | 10.1016/j.watres.2019.114972 |

# Take Home Messages

- Over 60 % of HR-MS peaks are potentially ***relevant*** but ***unknown***

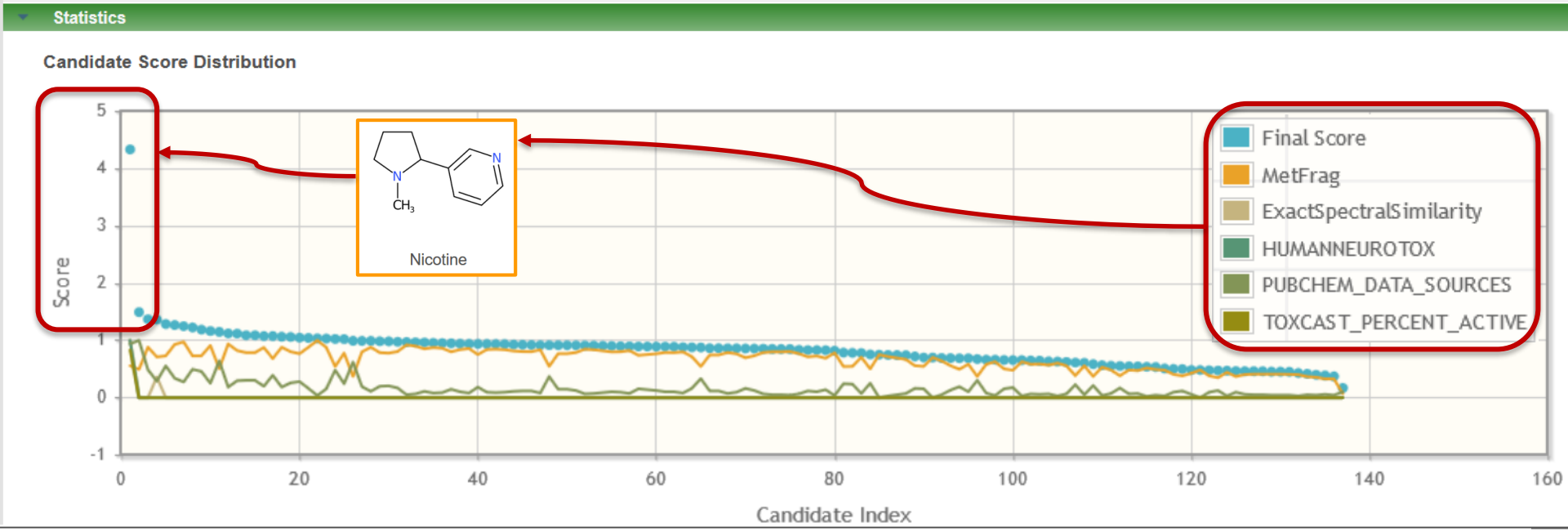


# Take Home Messages

- Over 60 % of HR-MS peaks are potentially ***relevant*** but ***unknown***
- ***Annotating unknowns*** requires data and evidence from ***many different sources***
  - Many excellent workflows available to collate this information
  - Incorporation of ***all available metadata*** is critical to success!

# Take Home Messages

- Over 60 % of HR-MS peaks are potentially ***relevant*** but ***unknown***
- ***Annotating unknowns*** requires data and evidence from ***many different sources***
- Non-target Screening is “Ready to go”!
  - Community efforts contribute greatly to improved cross-annotation
  - Information in the public domain helps everyone!
    - You never know when it will help you ☺



# Acknowledgements

Luxembourg National Research Fund

ART2 CURE

solutions

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

norman

eawag aquatic research

IPB

HELMHOLTZ CENTRE FOR ENVIRONMENTAL RESEARCH – UFZ

UNIVERSITY OF AMSTERDAM

Maastricht University

UNIVERSITAT ROVIRA i VIRGILI

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

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

PubChem

National and Kapodistrian UNIVERSITY OF ATHENS

KEMI Kemikalieinspektionen

WIKIDATA

UNIVERSITY OF COPENHAGEN

ETH zürich

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

ENVIRONMENTAL INSTITUTE

R3 LCSB

INRA SCIENCE & IMPACT

elixir LUXEMBOURG

EPA

emma.schymanski@uni.lu and @ESchymanski

Further Information:

https://massbank.eu/MassBank/

https://ipb-halle.github.io/MetFrag/

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

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


SEVENTH FRAMEWORK PROGRAMME
eawag
aquatic research
Community Efforts!
modelkey
solutions
MassBank consortium
ECHIDNA
norman
ART2 CURE