Luxembourg National Research Fund

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

UNIVERSITÉ DU LUXEMBOURG

PubChemLite Lab: *Cheminformatics* for the *Exposome*

PubChemLite EXPOSOMICS

Assoc. Prof. Dr. Emma L. Schymanski

(plus many, many colleagues and collaborators!)

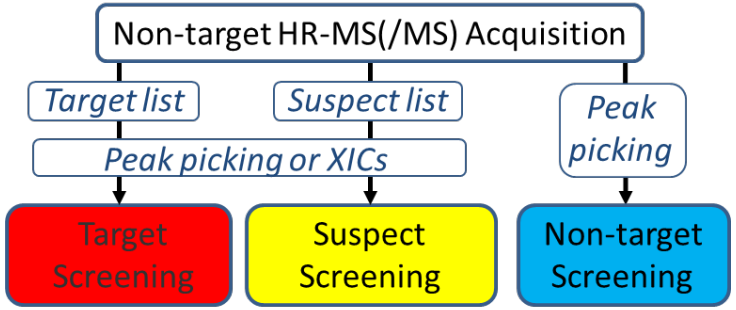
Environmental Cheminformatics Group,
Luxembourg Centre for Systems Biomedicine, University of Luxembourg
Email: emma.schymanski@uni.lu and @ESchymanski
Web: https://wwwen.uni.lu/lcsb/research/environmental_cheminformatics/

The Exposome Boot Camp
Live-Stream, July 22-23, 2021

Slides available at
DOI: 10.5281/zenodo.5115498

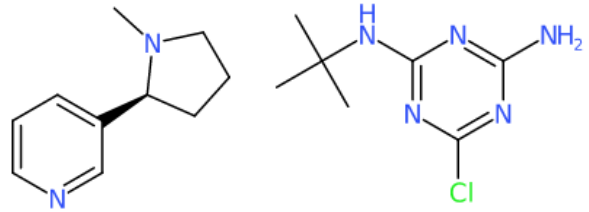
Outline for today

- Background
  - Environmental cheminformatics & HR-MS
  - Non-target screening and identification confidence
  - Compound databases and spectral libraries
- Introduction to PubChemLite and MetFrag
- Examples (MassBank+PubChemLite+MetFrag)
  - Nicotine – relating identifications to disease outcomes
  - Desethylterbutylazine – isobars and the role of metadata
- Closing – Perspectives and Community Contribution

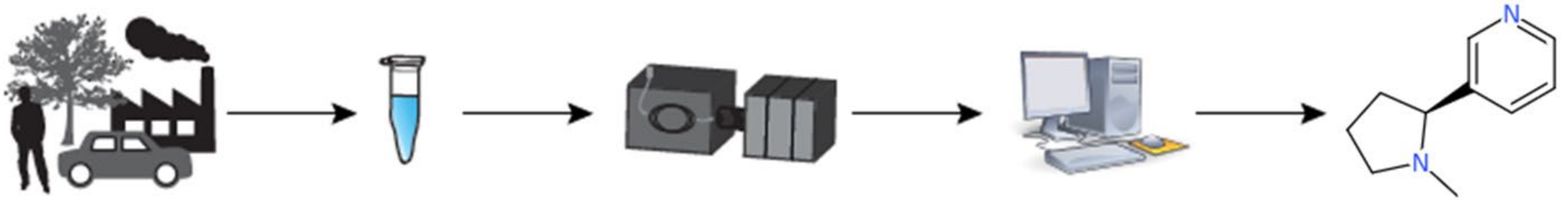


PubChemLite EXPOSOMICS

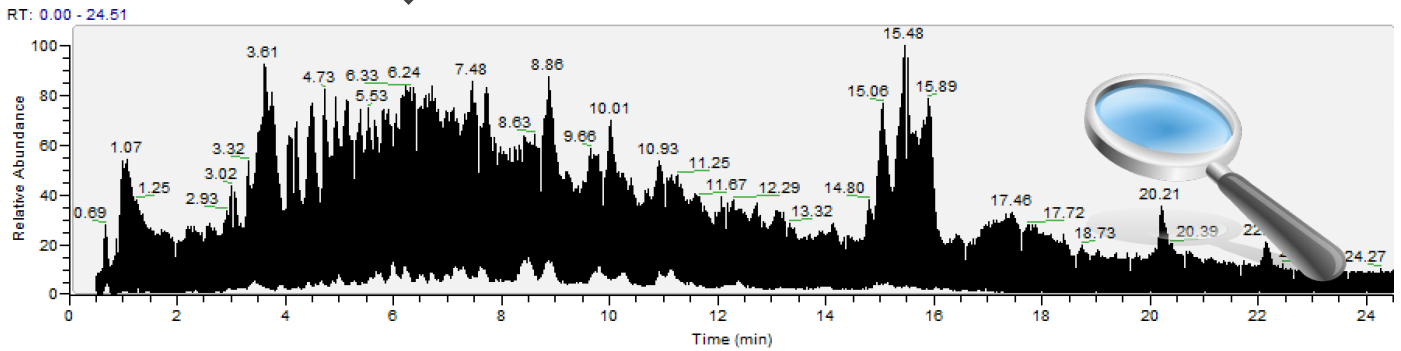
MetFrag

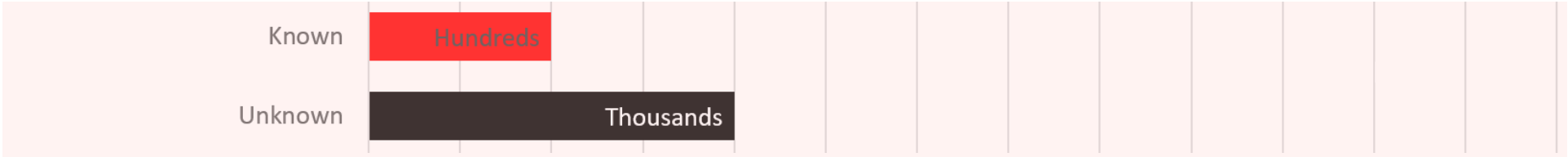


Background: Environmental Cheminformatics & HR-MS



High resolution mass spectrometry

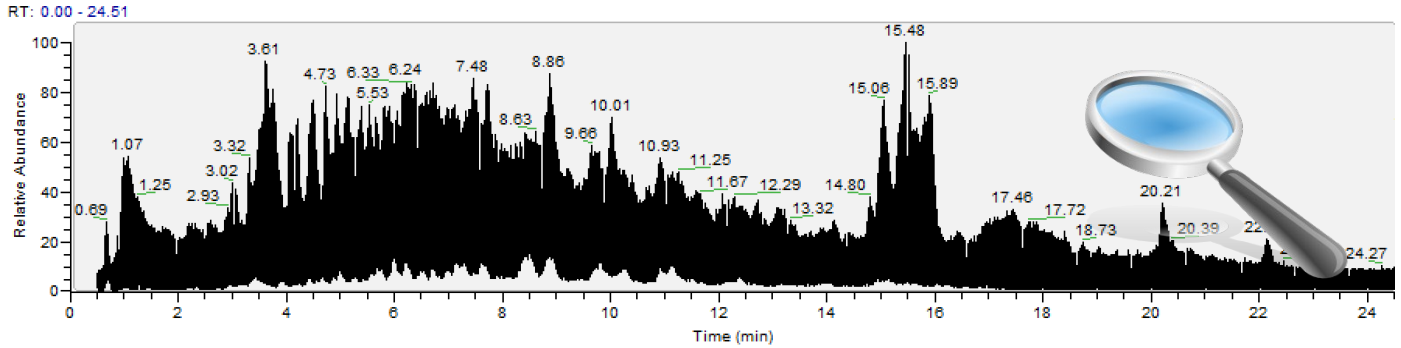


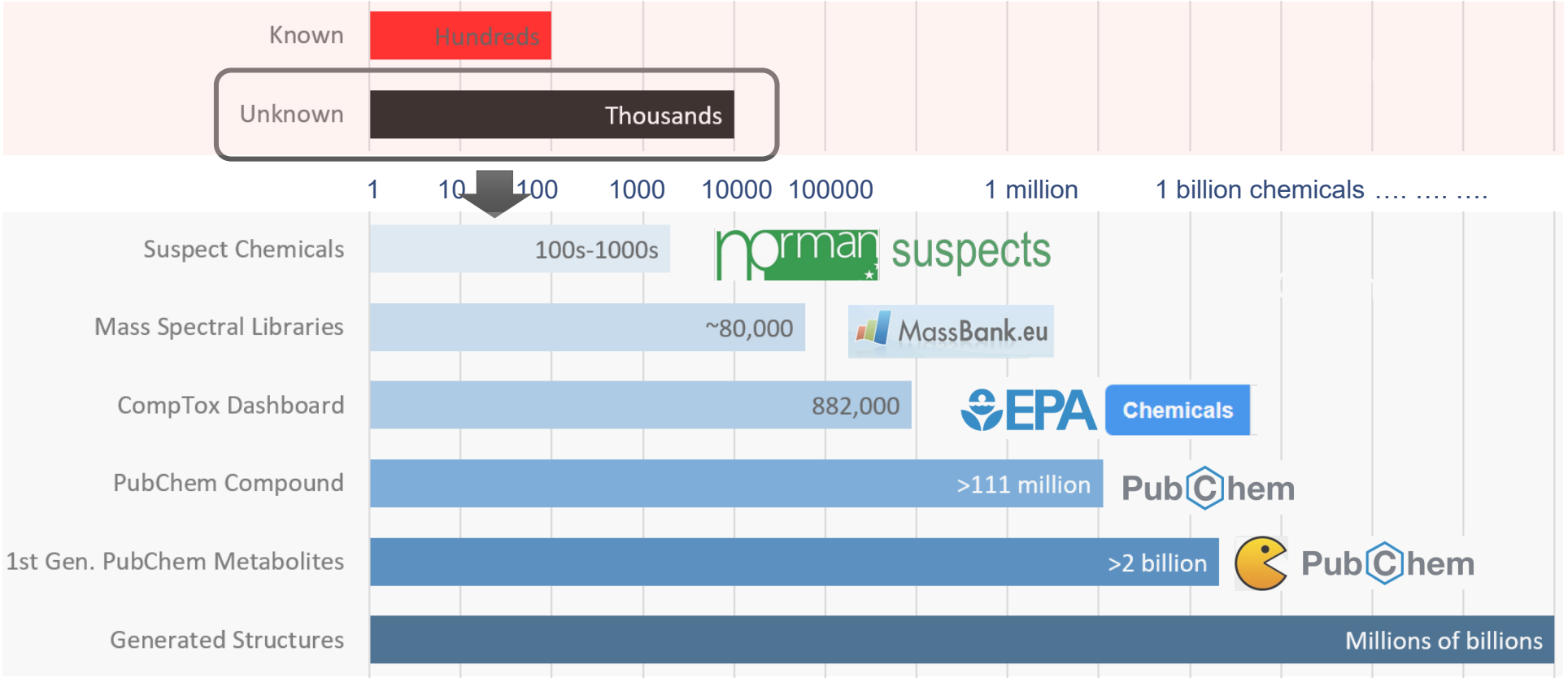


Background: Environmental Cheminformatics & HR-MS

High resolution mass spectrometry

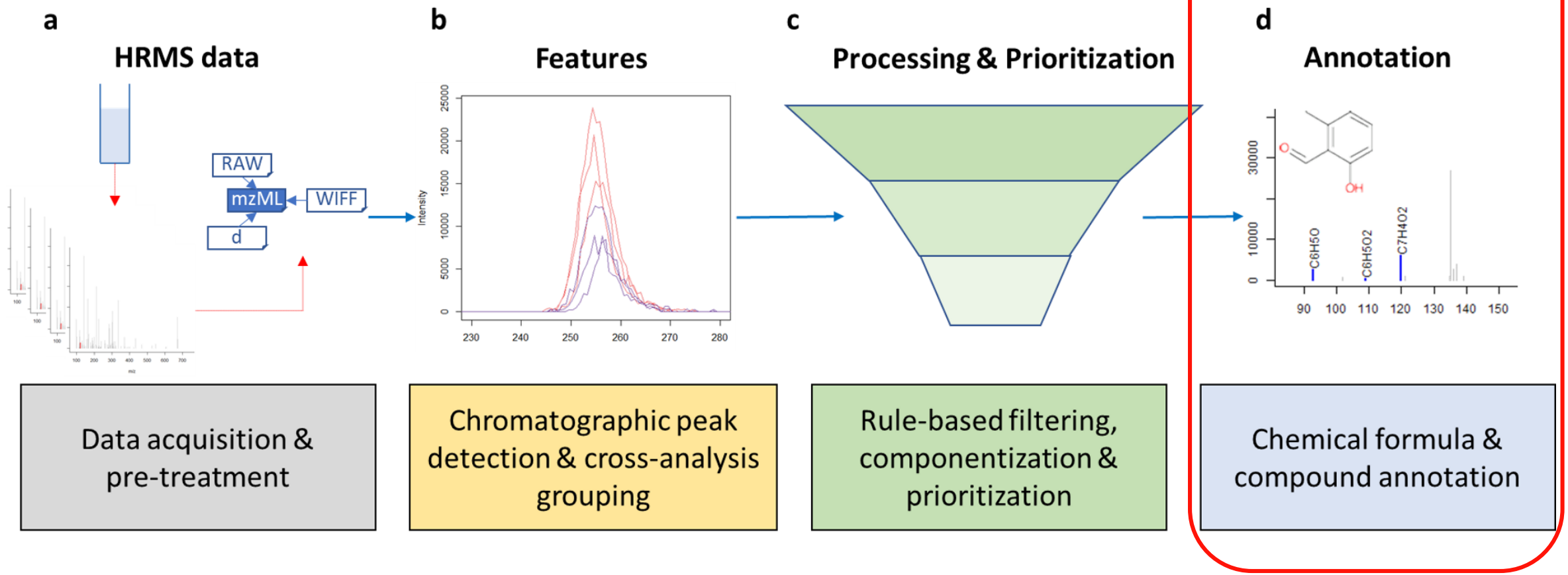
AND connecting chemical knowledge





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

Background: Identification with HR-MS



Helmus *et al* (2021). patRoon. *J. Cheminformatics*. DOI: 10.1186/s13321-020-00477-w

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

CAS
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

180 million

PubChem

110 million

ChemSpider
Search and share chemistry

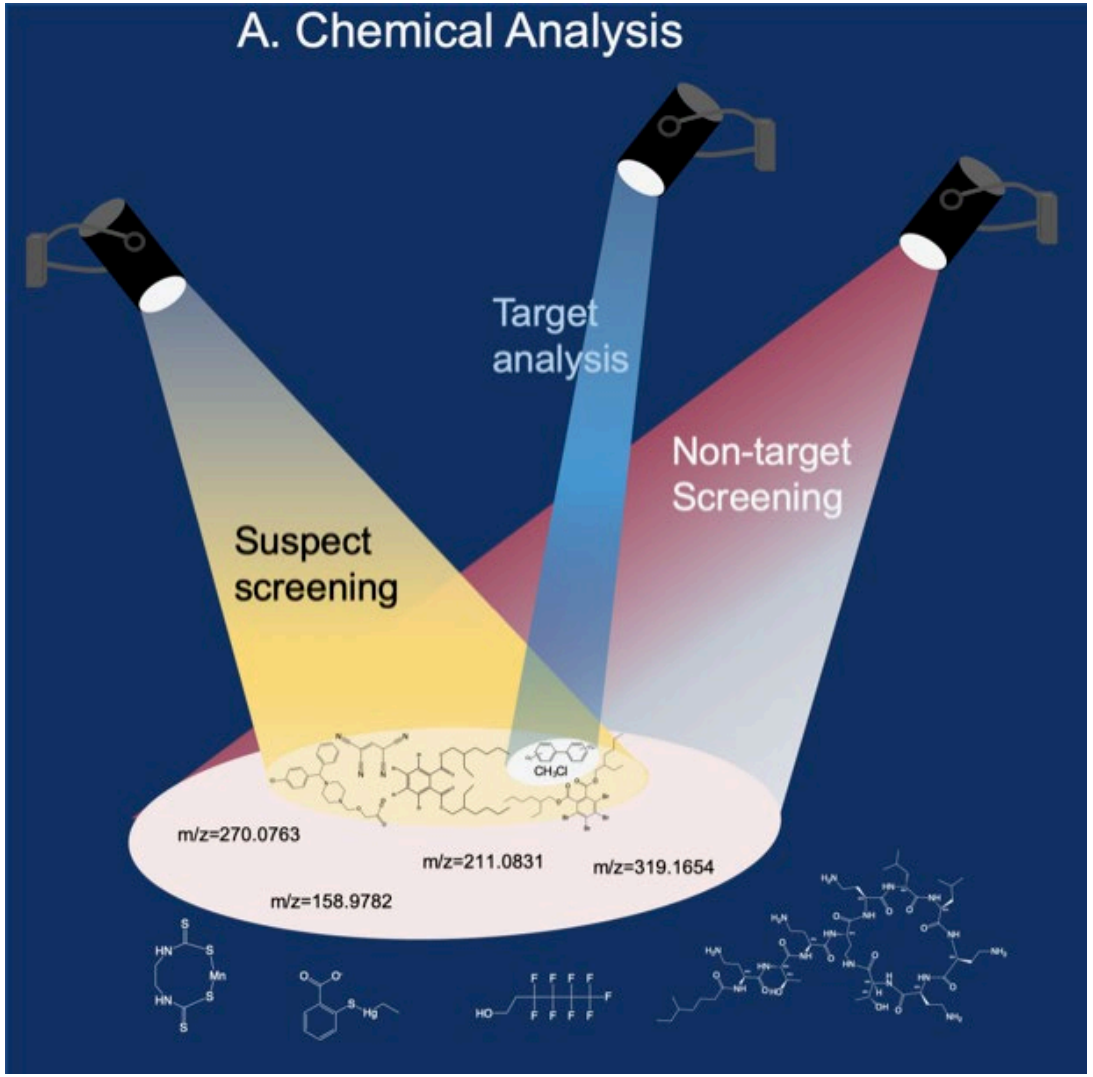
103 million

EPA Chemicals

883,000

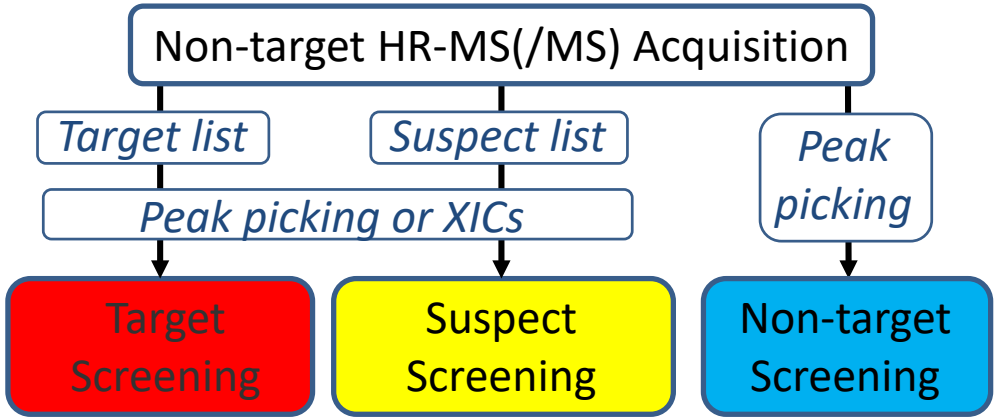
norman suspects

113,000



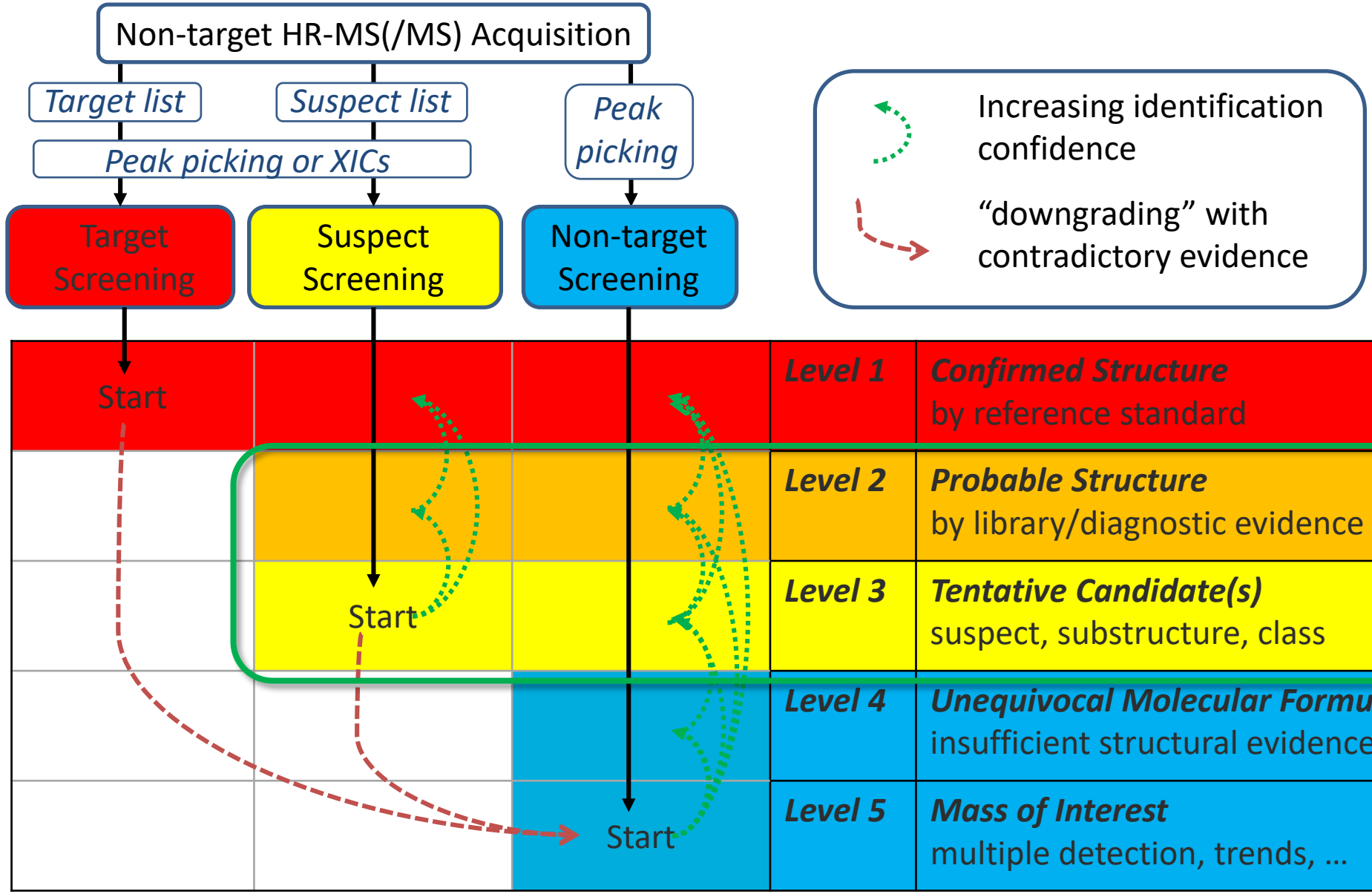
Source: ESO/IDA/Danish 1.5 m http://www.eso.org/public/images/potw1015a/ Mod. from Escher *et al* (2020). Science. DOI: 10.1126/science.aay6636

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



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)



MassBank
High Quality Mass Spectral Database

MetFrag

eawag
aquatic research

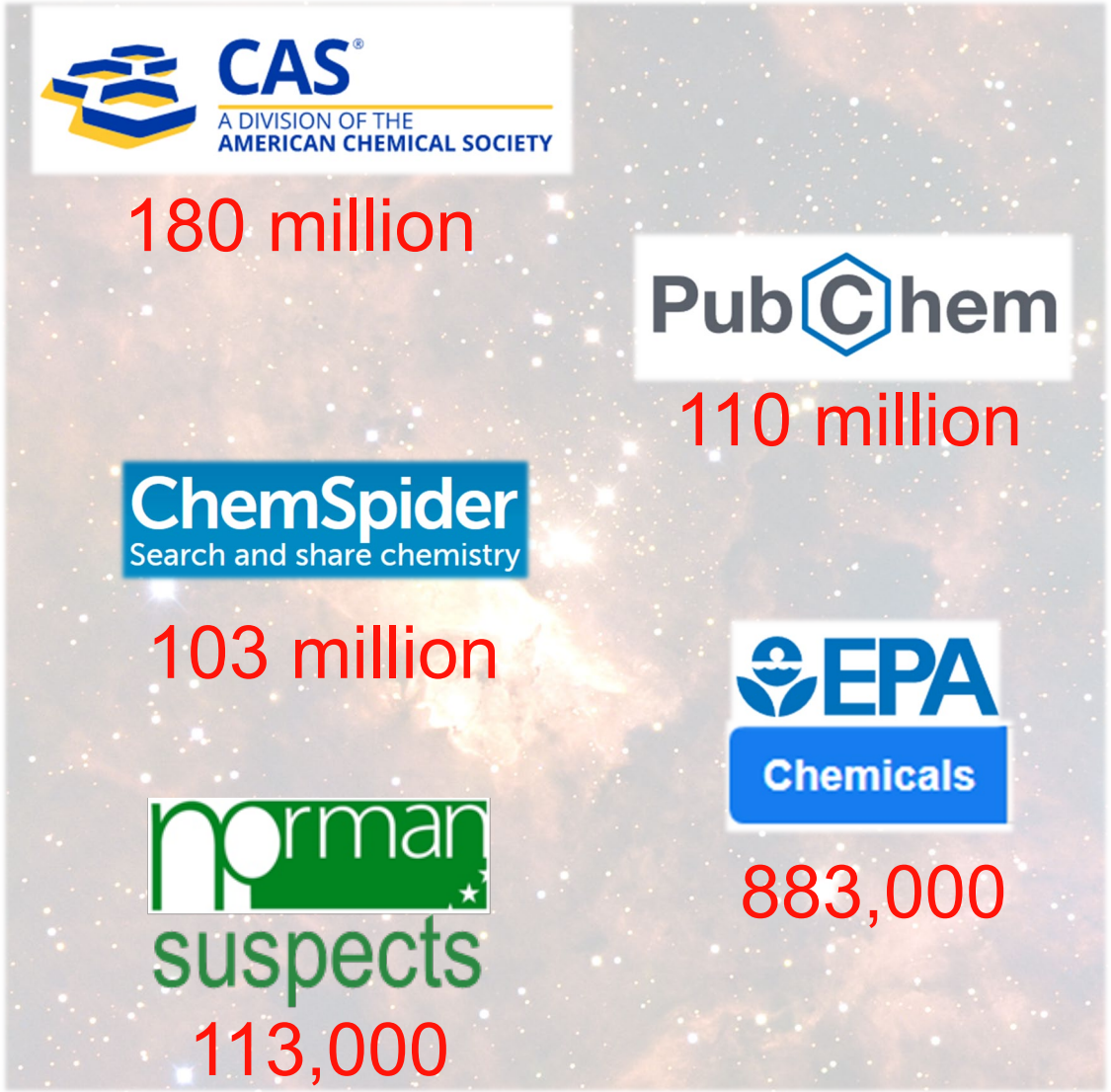
Schymanski *et al*, 2014, ES&T. DOI: 10.1021/es5002105 & Schymanski *et al*. 2015, ABC, DOI: 10.1007/s00216-015-8681-7

Background: Key Resources for Identification / Annotation

- Compound Databases
  - A collection of structures, their properties and associated information
  - Generally little or no spectral data, rather structures & links
  - Largest have >100 million structures in them
    - But don't assume that everything is in there – it isn't!!!!

PubChemLite EXPOSOMICS

370,000 entries "small"



Background: Key Resources for Identification / Annotation

- Mass Spectral Databases or Libraries
  - A collection of structures, mass spectra and associated information
  - NIST and Wiley are widely accepted for GC-EI/MS
    - Together >1.2 million spectra of 707,000 compounds
- MS/MS databases are growing, none are yet “established”
  - Together > 2 million spectra, but only approx. 40-80,000 compounds
  - MS/MS available for only ~0.1-4 % of relevant exposomics resources (see next slide)

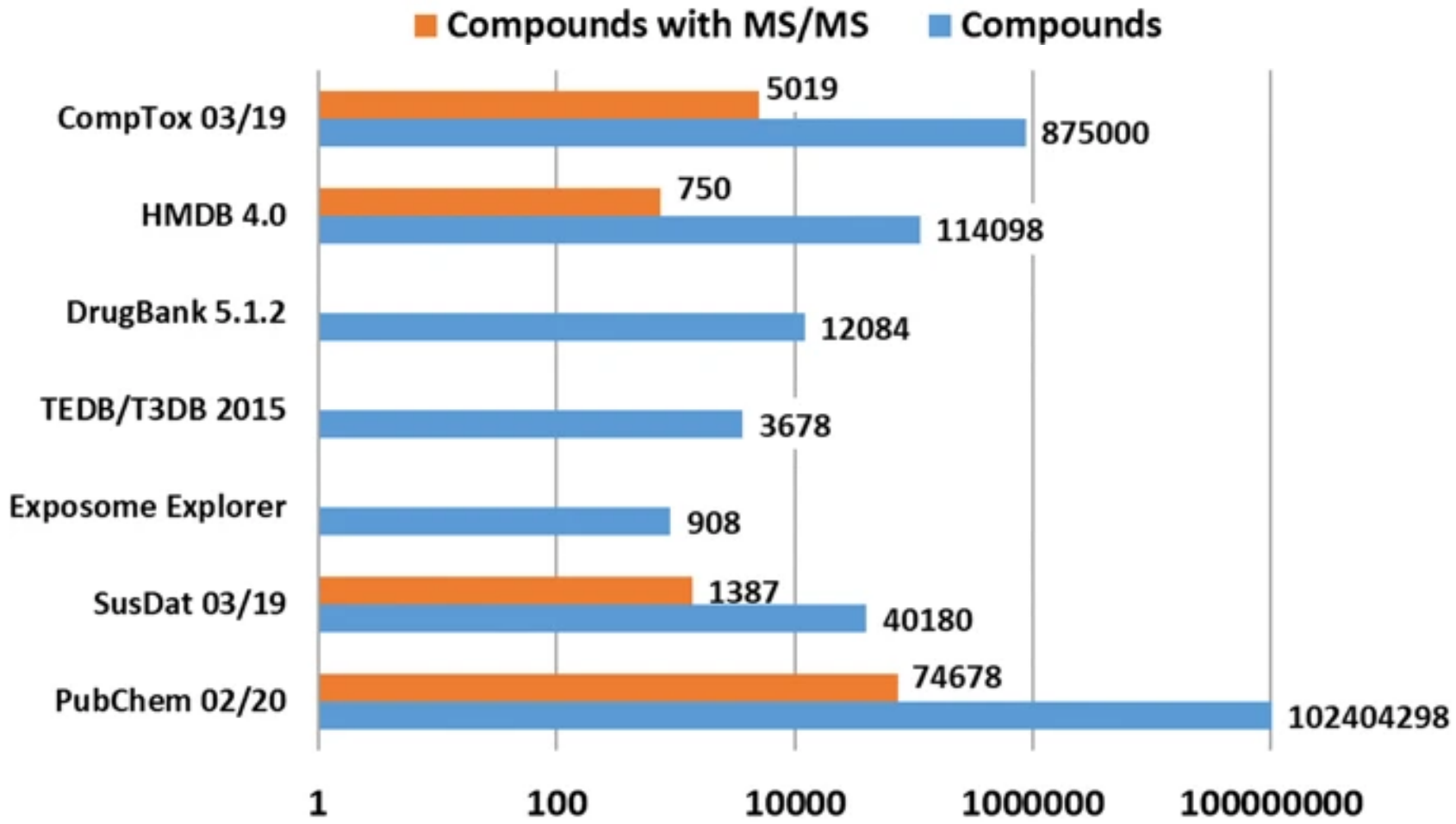
- Key resource(s) mentioned today:
  - MassBank EU: http://massbank.eu/MassBank
  - MoNA: https://mona.fiehnlab.ucdavis.edu/

MassBank
High Quality Mass Spectral Database

MoNA
MassBank of North America

Background: Scarcity of MS/MS Spectra in Exposomics

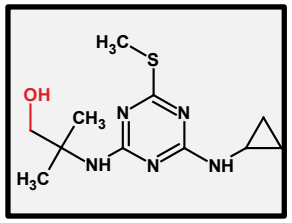
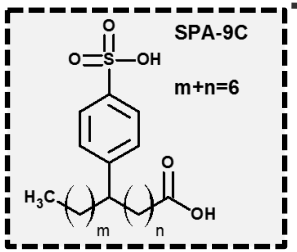
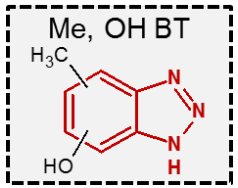
- MS/MS avail. for ~0.1-4 % of relevant exposomics resources



MEDIZINISCHE UNIVERSITÄT INNSBRUCK

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

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

Example	*Identification confidence*	*Minimum data requirements*
	***Level 1: Confirmed structure*** by reference standard	MS, MS^2, RT, **Reference Std.**
	Level 2: Probable structure a) by library spectrum match	MS, MS^2, **Library MS^2 >0.9 sim**
	Level 2: Probable structure b) by diagnostic evidence	MS, MS^2, Exp. data
	***Level 3: Tentative candidate(s)*** structure, substituent, class	MS, MS^2, Exp. data
		
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct
192.0757	***Level 5: Exact mass*** of interest	MS

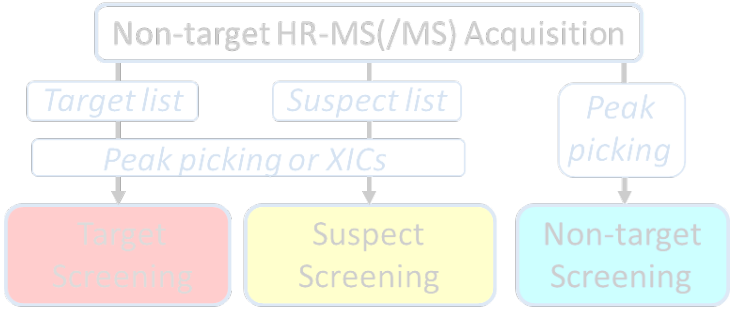
MassBank
High Quality Mass Spectral Database

MetFrag

Schymanski *et al*, 2014, ES&T. DOI: 10.1021/es5002105

Outline for today

- Background
  - Environmental cheminformatics & HR-MS
  - Non-target screening and identification confidence
  - Compound databases and spectral libraries
- Introduction to PubChemLite and MetFrag
- Examples (MassBank+PubChemLite+MetFrag)
  - Nicotine – relating identifications to disease outcomes
  - Desethylterbutylazine – isobars and the role of metadata
- Closing – Perspectives and Community Contribution



PubChemLite EXPOSOMICS

MetFrag

Introduction to MetFrag

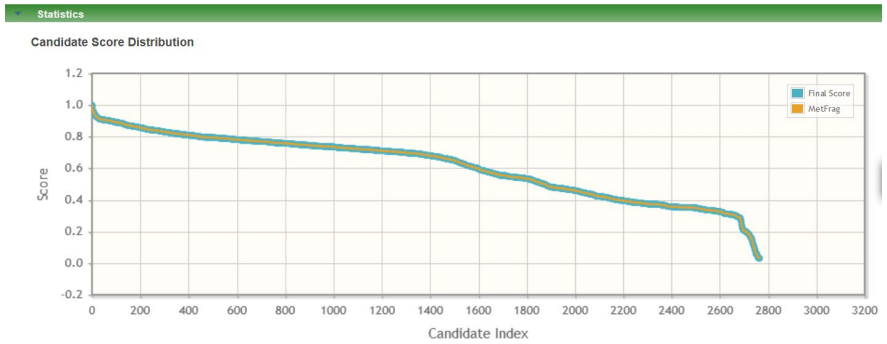
https://msbi.ipb-halle.de/MetFrag/

mz [M−H]⁻
213.9637
± 5 ppm

5 ppm
0.001 Da

PubChem

Ranked Candidates



MS/MS

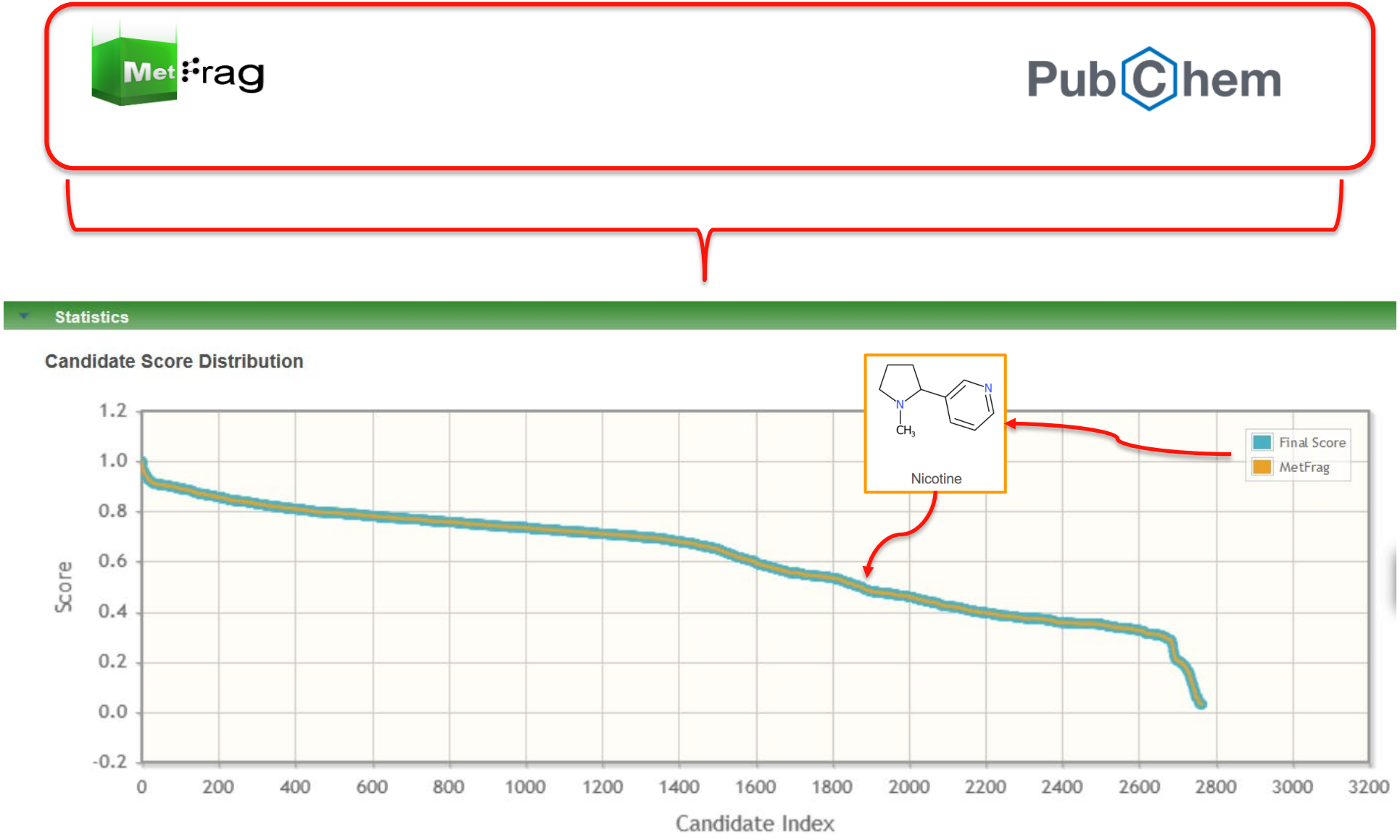
134.0054	339689
150.0001	77271
213.9607	632466

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

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

Schymanski, Kondic, Neumann, Thiessen, Zhang & Bolton (2021) *J. Cheminf.*, DOI: 10.1186/s13321-021-00489-0
Ruttkies, Schymanski, Wolf, Hollender, Neumann (2016) *J. Cheminf.*, DOI: 10.1186/s13321-016-0115-9

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



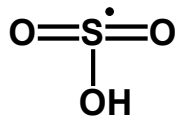
MetFrag + PubChem + Formula Search + https://massbank.eu/MassBank/RecordDisplay?id=EQ300804

Status Quo in 2016: MetFrag Relaunched …

mz [M-H]⁻
213.9637
± 5 ppm

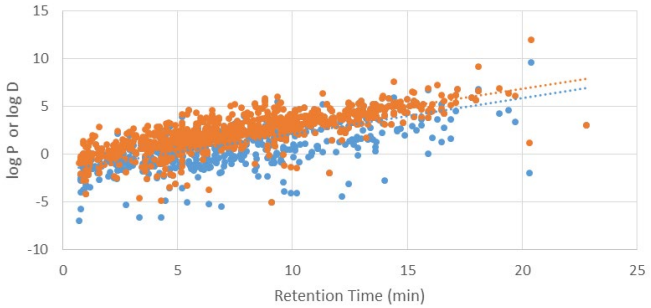
Elements: C,N,S

5 ppm
0.001 Da



RT: 4.54 min

355 InChI/RTs



EPA Chemicals
or
PubChem

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

Literature Patents

Suspect Lists

TOFF IDENT

norman suspects

EPA Chemicals

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

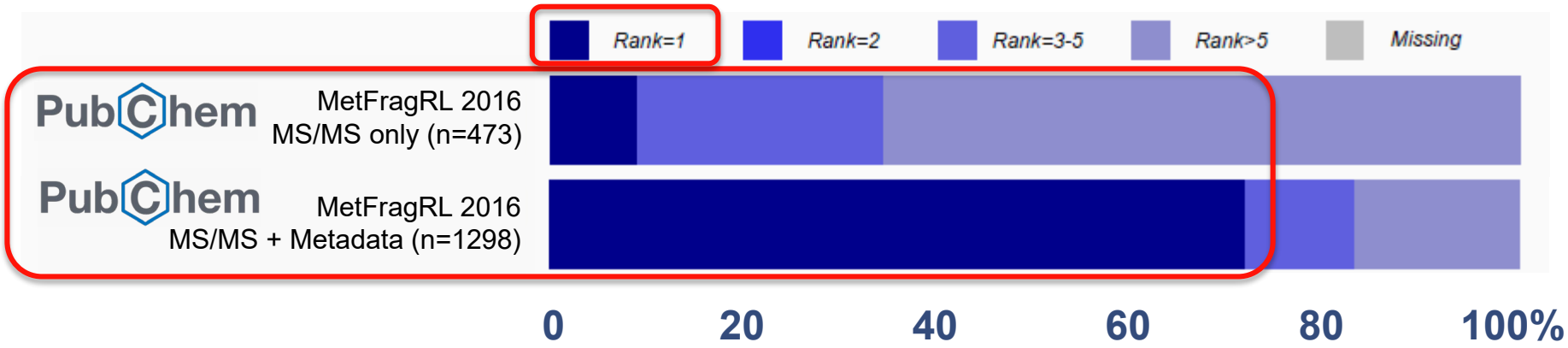
solutions

uni.lu | LCSB

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

MetFragRL + PubChem + MS/MS + Metadata

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



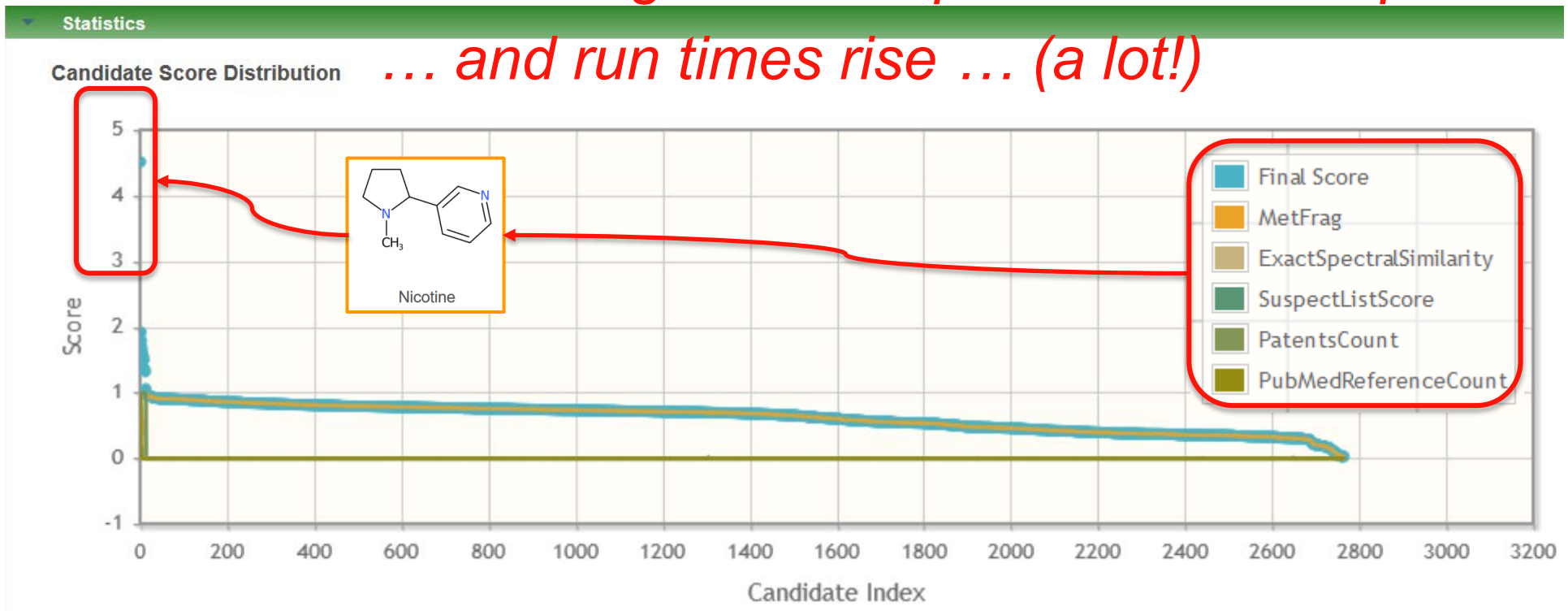
Schymanski, Kondic, Neumann, Thiessen, Zhang & Bolton (2021) *J. Cheminf.*, DOI: 10.1186/s13321-021-00489-0

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

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

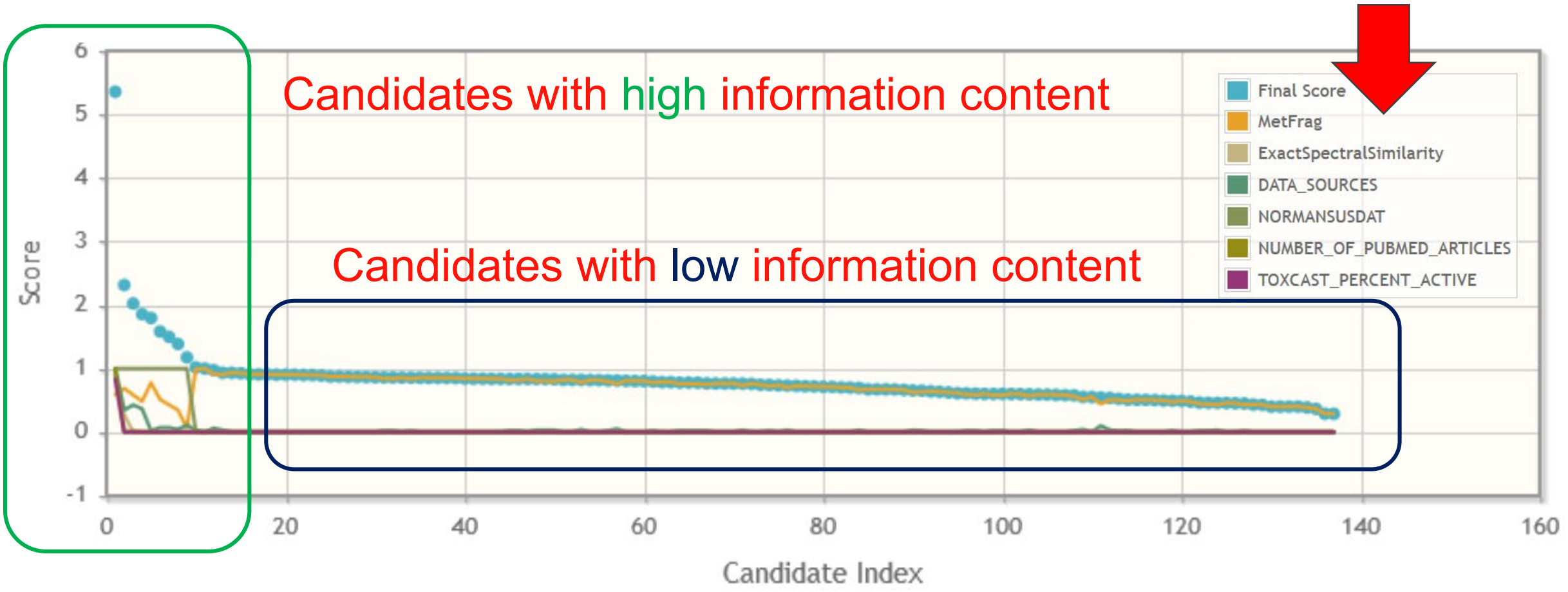
Problem: Exposomics “Chemical Space” is too big!

CAS A DIVISION OF THE AMERICAN CHEMICAL SOCIETY — 180 million

PubChem — 110 million

ChemSpider Search and share chemistry — 103 million

EPA Chemicals — 883,000



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

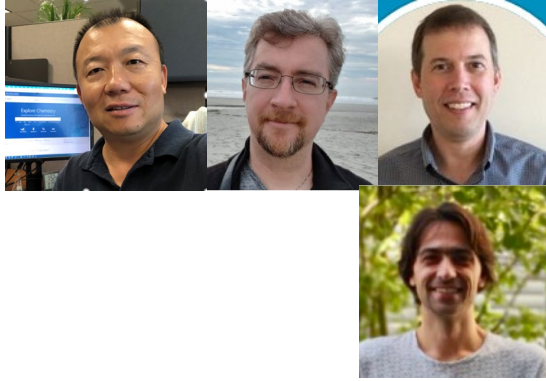
Active Substance	furathiocarb
Status	Not Approved [Reg. (EC) No 1107/2009]
Categories	Insecticides

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

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

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

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”

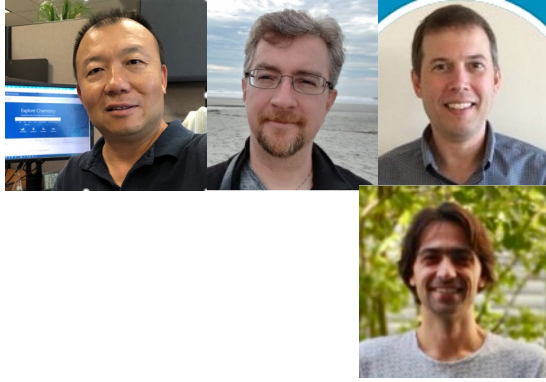
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9

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

uni.lu | LCSB

Introducing …

PubChemLite EXPOSOMICS



Collapsed by InChIKey First Block (skeleton) *and* by presence of annotation content

~370,000 entries “small”

PubChem INCSWYKICIYAHB

Compounds (6) | Substances (2)

Searching chemical names and synonyms including IUPAC names and InChIKeys across the compound collection. Note that annotations text searched. Read More...

6 results | Filters | SORT BY Compound C

2,3-dihydroxy-2,3-dihydrobenzoic Acid; 5,6-dihydroxycyclohe
Acid; 100459-00-5; 2,3-dihydro-2,3-dihydroxybenzoic Acid; ACM

Compound CID (sort by): 3
MF: $C_7H_8O_4$ MW: 156.14g/mol
InChIKey: INCSWYKICIYAHB-UHFFFAOYSA-N
IUPAC Name: 5,6-dihydroxycyclohexa-1,3-diene-1-carboxylic acid
Create Date: 2004-09-16

Summary | Similar Structures Search | Related Records

(2S,3S)-2,3-dihydroxy-2,3-dihydrobenzoic Acid; 176487-06-2; (2S
2,3-dihydroxybenzoate; (5S,6S)-5,6-dihydroxycyclohexa-1,3-dien
(2S,3S)-2,3-Dihydro-2,3-dihydroxybenzoate; ...

Compound CID (sort by): 9964159
MF: $C_7H_8O_4$ MW: 156.14g/mol
InChIKey: INCSWYKICIYAHB-WDSKDSINSA-N
IUPAC Name: (5S,6S)-5,6-dihydroxycyclohexa-1,3-diene-1-carboxylic acid
Create Date: 2006-10-25

MetFrag

Database Settings

Database: PubChemLite_31Oct2020_

Neutral Mass: 229.10948 Search ppm: 5

Formula: C9H16ClN5

Identifiers:

Retrieve Candidates | 4 Candidates

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

October 31, 2020 | Dataset | Open Access

PubChemLite for Exposomics

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

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

is 371,663 compounds (31 Oct 2020) compiled from 10 categories:
DrugMedicInfo, FoodRelated, PharmacoInf Info,
, Identification.

llapsed by InchIKey first block, reporting th most
CIDs. Entries that will be ignored by MetFra
(e.g. transition metals) have been removed Med ID
les on the PubChem FTP site. The "AnnoTy how
epresented, the subsequent column (named per category) counts the
ories available in the next sub-category of the TOC entry.

Edit | New version

Communities
LCSB Environmental Cheminformatics Group | Remove

1,084 views | 1,170 downloads

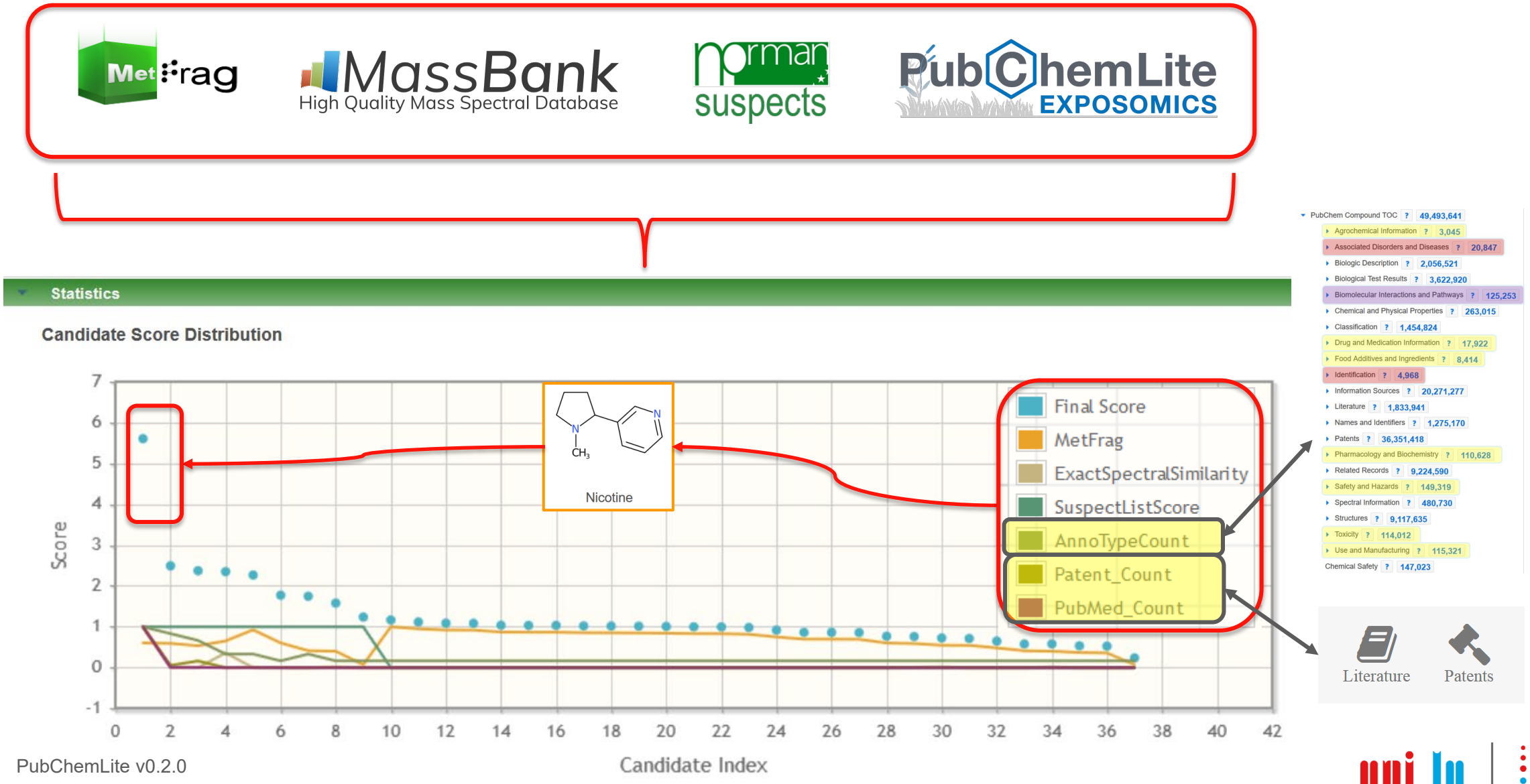
See more details...

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

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

uni.lu | LCSB

MetFragRL + PubChemLite: tailor-made database + metadata

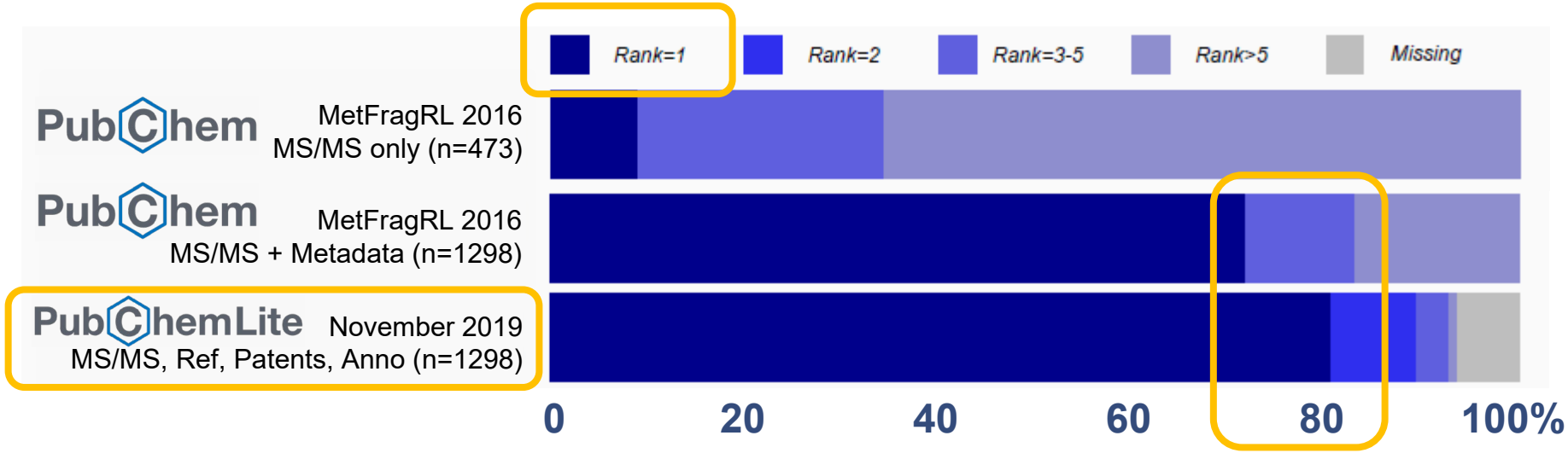


PubChemLite v0.2.0

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?

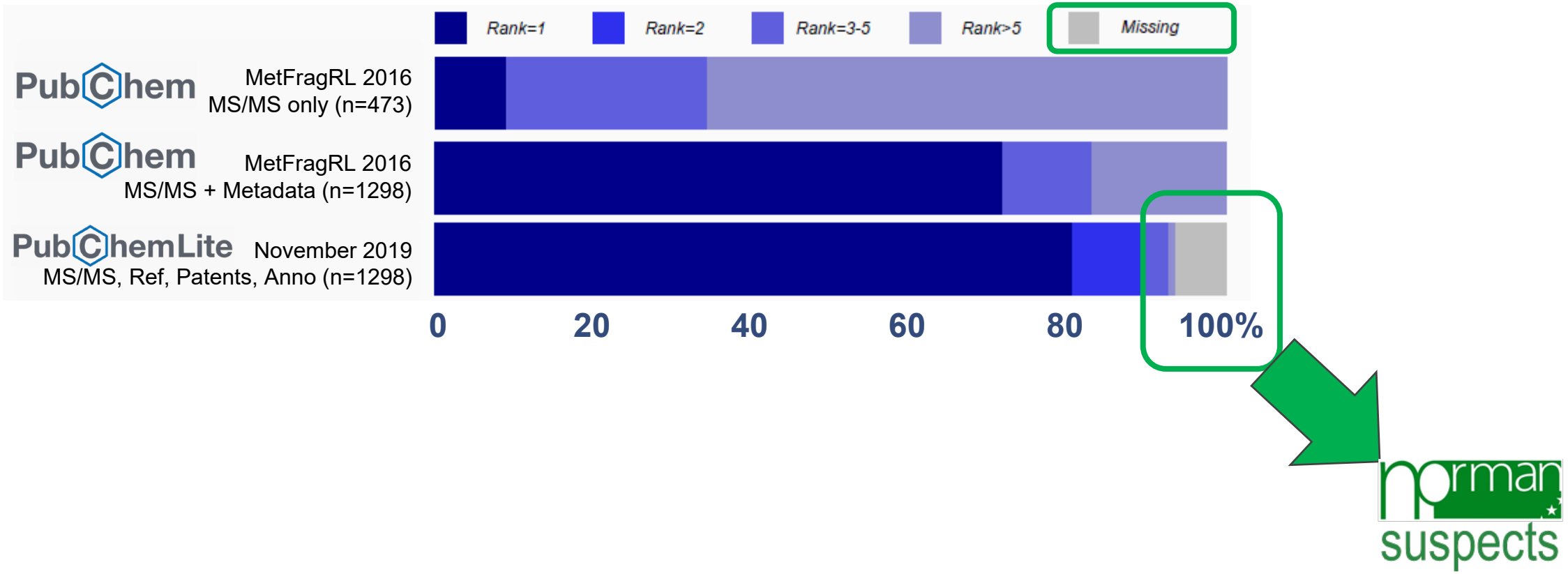


Schymanski, Kondic, Neumann, Thiessen, Zhang & Bolton (2021) *J. Cheminf.*, DOI: 10.1186/s13321-021-00489-0

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

How does PubChemLite perform?

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

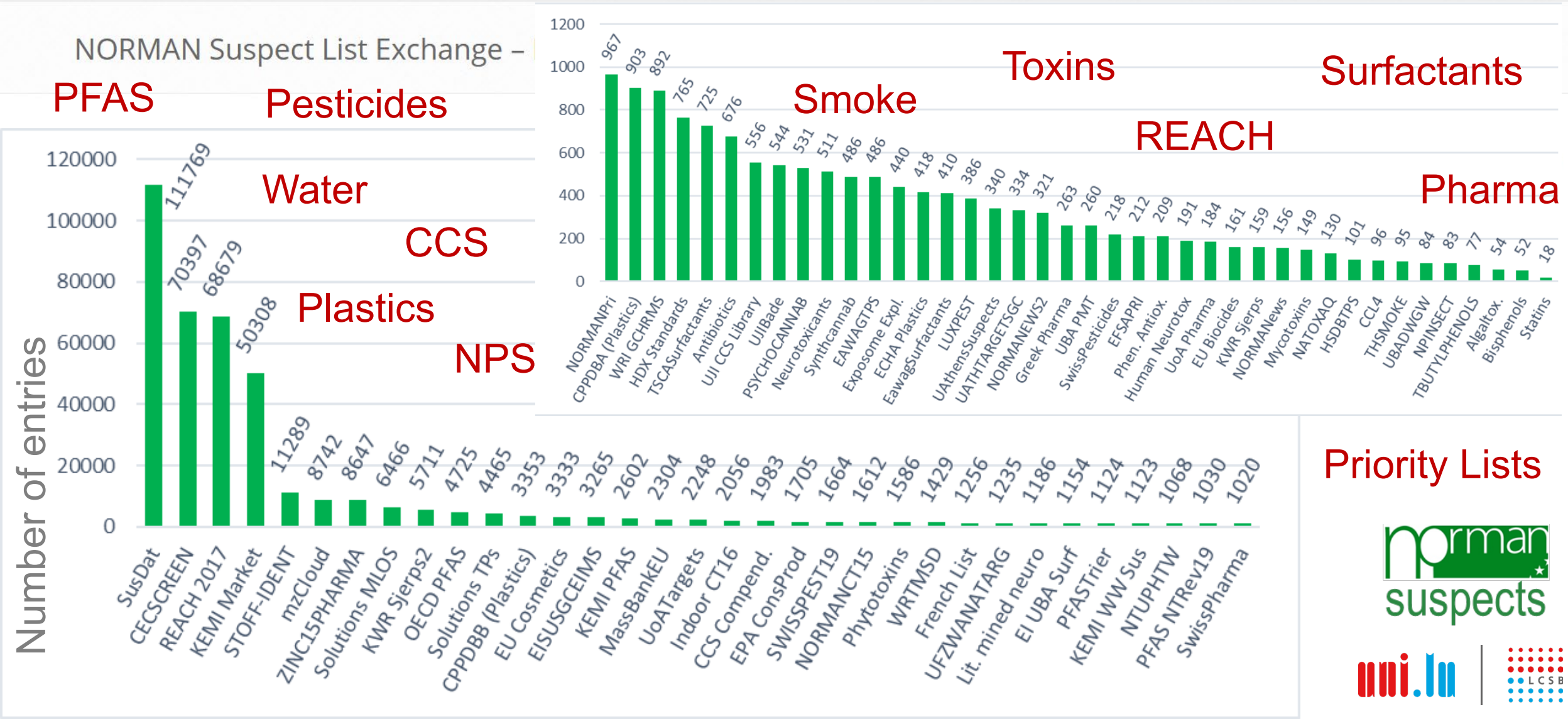


Schymanski, Kondic, Neumann, Thiessen, Zhang & Bolton (2021) *J. Cheminf.*, DOI: 10.1186/s13321-021-00489-0

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

Expert Knowledge: NORMAN Suspect List Exchange (>80 lists!)

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



Priority Lists

norman suspects

uni.lu LCSB

Filling Gaps: Integrating NORMAN-SLE

https://www.norman-network.com/nds/SLE/ => https://pubchem.ncbi.nlm.nih.gov/source/23819

PubChem NORMAN Suspect List Exchange

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	114,573 Live Substances 16,423 Annotations 1 Classification
Last Updated	2021/04/15

Filling Gaps: Integrating NORMAN-SLE

NORMAN-SLE Classification: 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

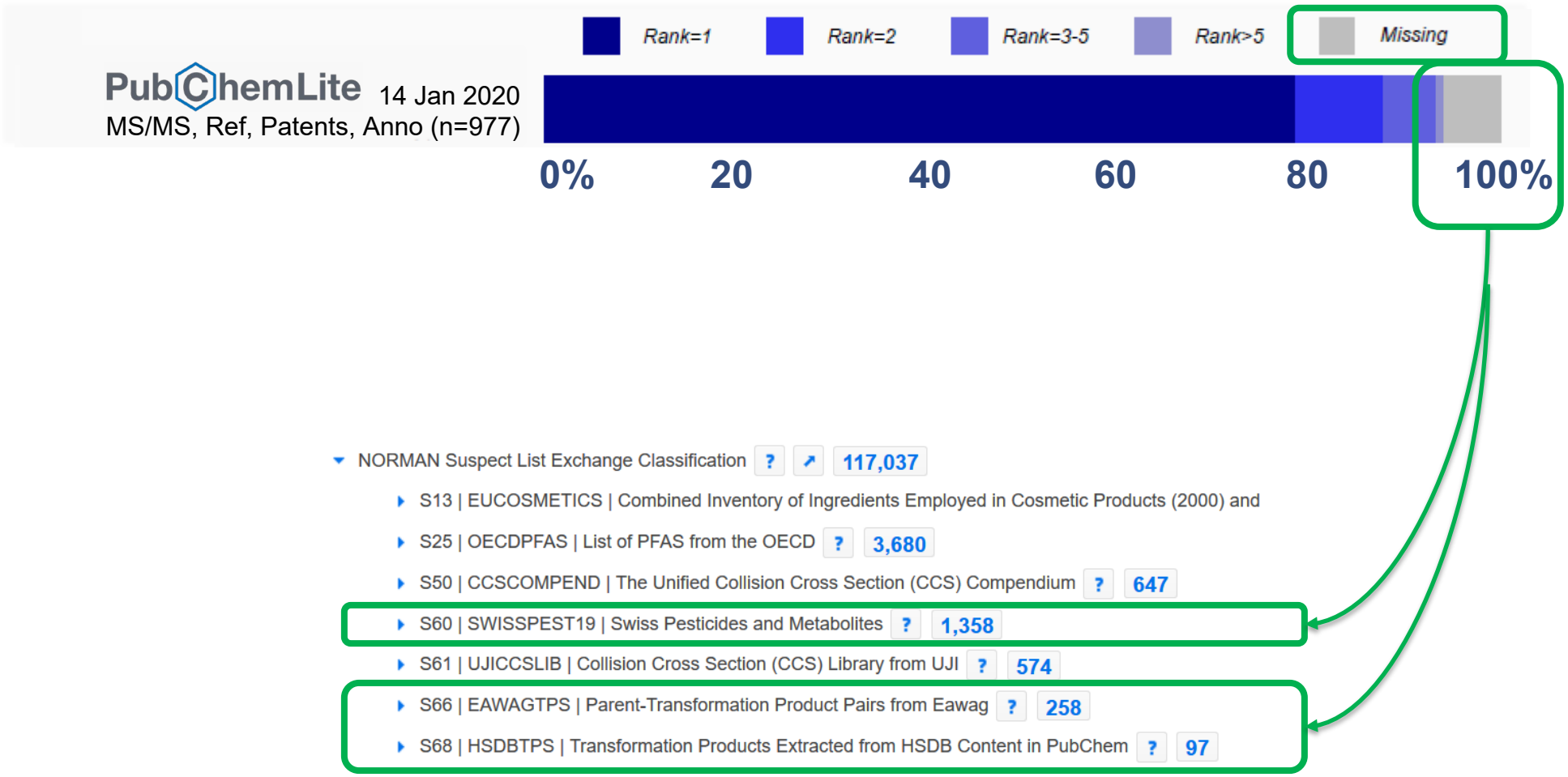
Search selected classification by Keyword | Enter desired search term | Search

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...

- NORMAN Suspect List Exchange Classification ? 113,080
  - S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) ? 3,859
  - 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

Assessing the Missing Entries in PubChemLite



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

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

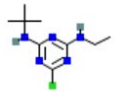
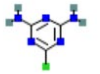
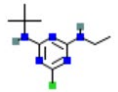
- Pharmacology and Biochemistry ? 112,039
  - Human Metabolite Information ? 64,199
  - Metabolism/Metabolites ? 8,204
  - Transformations ? 5,857

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/s13321-016-0115-9

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

uni.lu LCSB

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

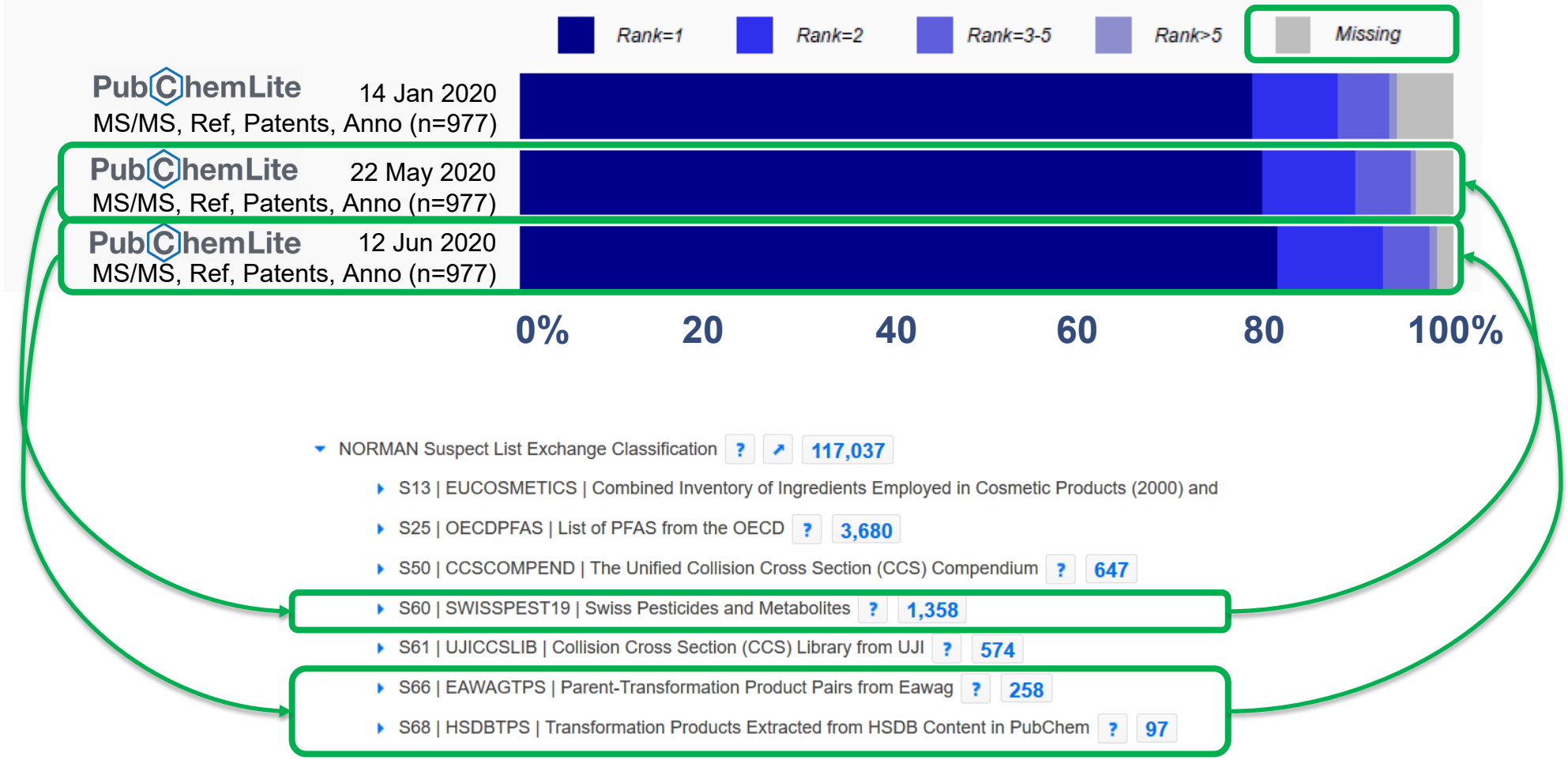
- Agrochemical Information ? 3,045
  - Agrochemical Category ? 1,903
  - Agrochemical Transformations ? 1,383
  - EU Pesticides Data ? 1,213
  - USDA Pesticide Data Program ? 579

PubChemLite EXPOSOMICS

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

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

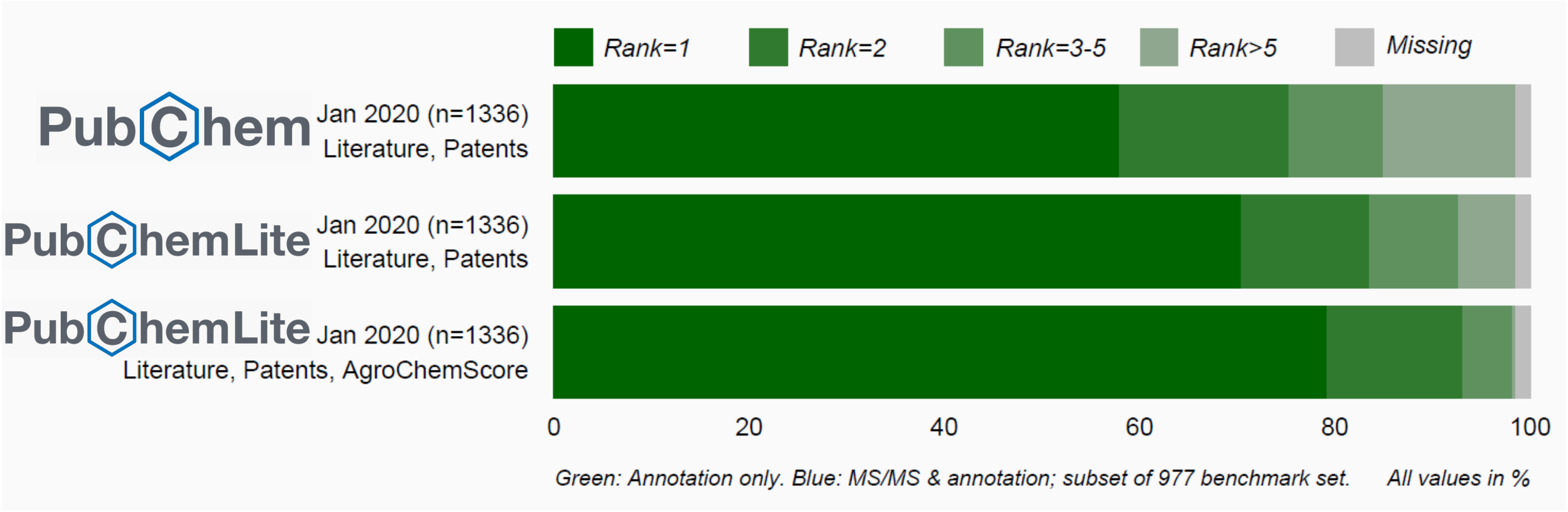
Assessing the Missing Entries in PubChemLite



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

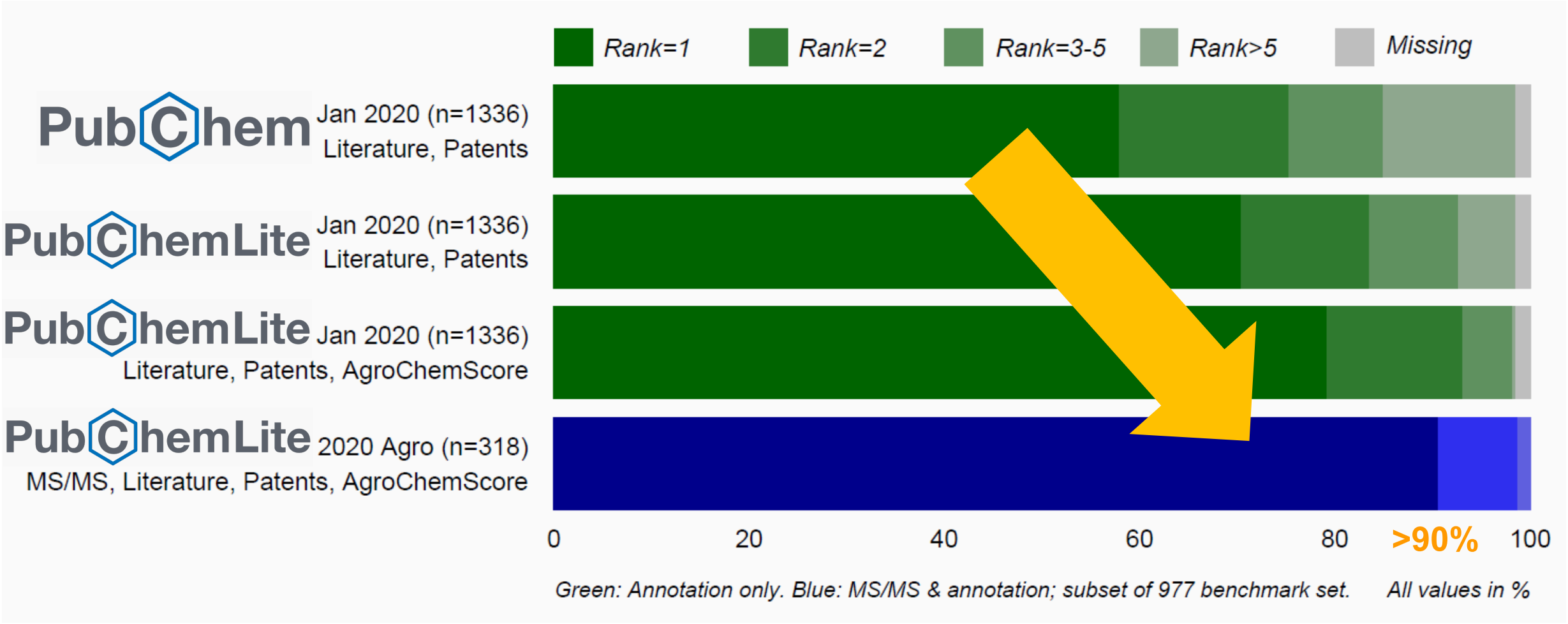
U.S. National Library of Medicine
National Center for Biotechnology Information

Influence of the Annotation Content in PubChemLite



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

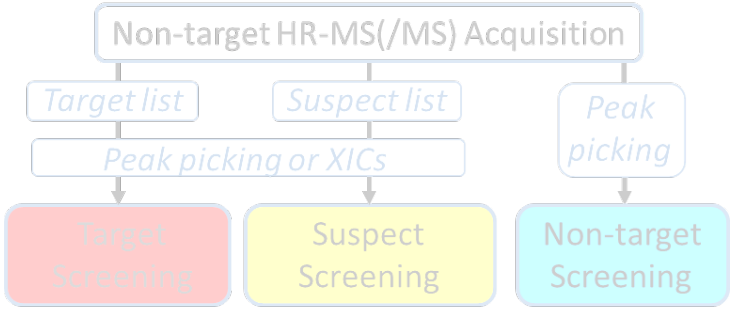
Influence of the Annotation Content in PubChemLite



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

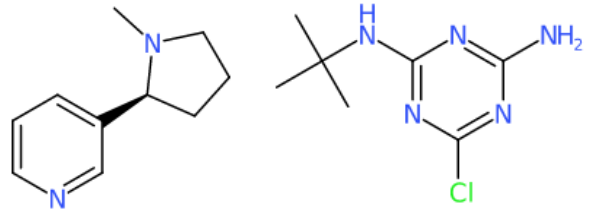
Outline for today

- Background
  - Environmental cheminformatics & HR-MS
  - Non-target screening and identification confidence
  - Compound databases and spectral libraries
- Introduction to PubChemLite and MetFrag
- Examples (MassBank+PubChemLite+MetFrag)
  - Nicotine – relating identifications to disease outcomes
  - Desethylterbutylazine – isobars and the role of metadata
- Closing – Perspectives and Community Contribution



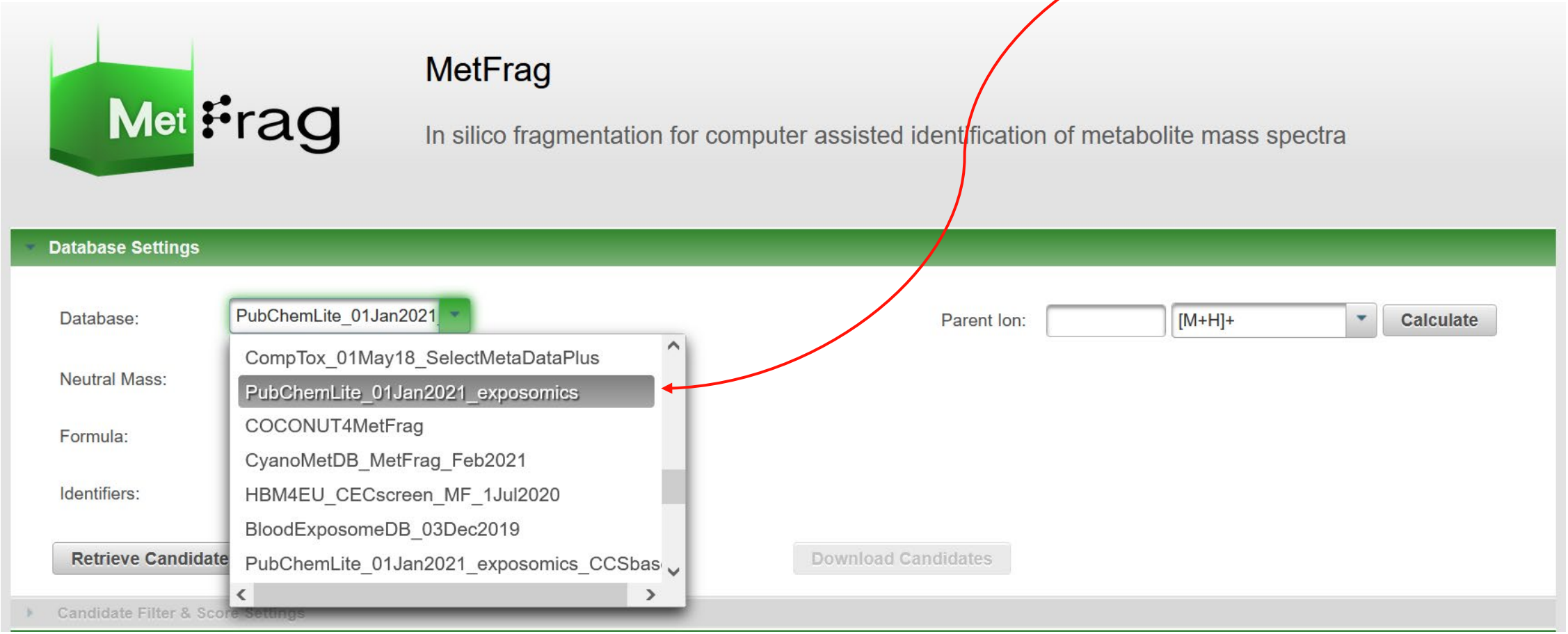
PubChemLite EXPOSOMICS

MetFrag



MetFrag Example: Nicotine and Disease Associations

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



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

MetFrag Example: Nicotine and Disease Associations

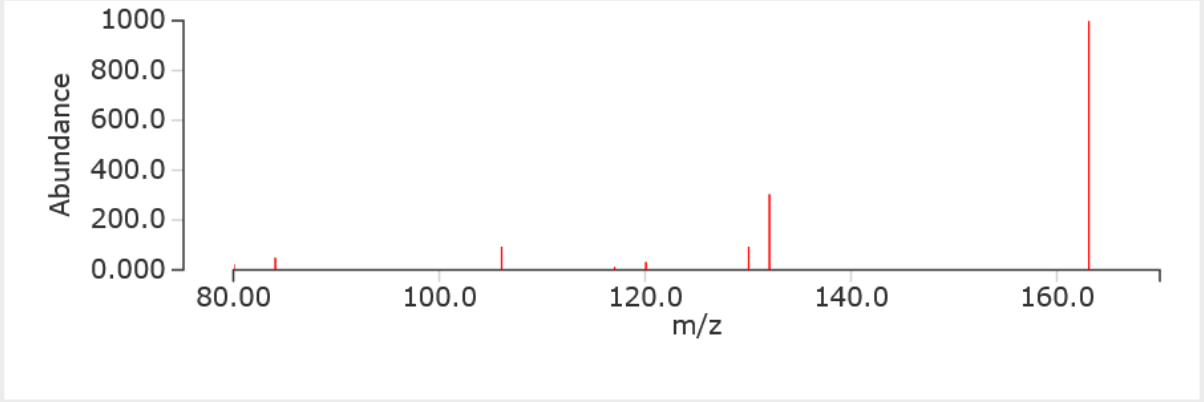
MassBank
High Quality Mass Spectral Database

MassBank Search Contents Download Accession Go More▼

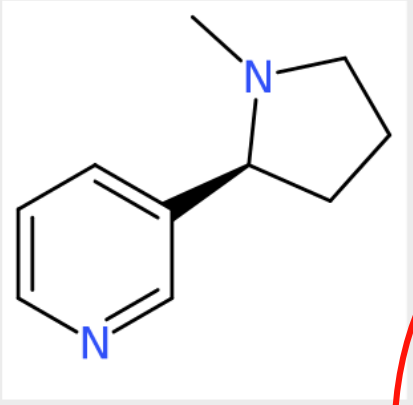
MassBank Record: EQ300801

Nicotine; LC-ESI-QFT; MS2; CE: 15; R=35000; [M+H]+

Mass Spectrum



Chemical Structure



metabolomics-usi visualisation

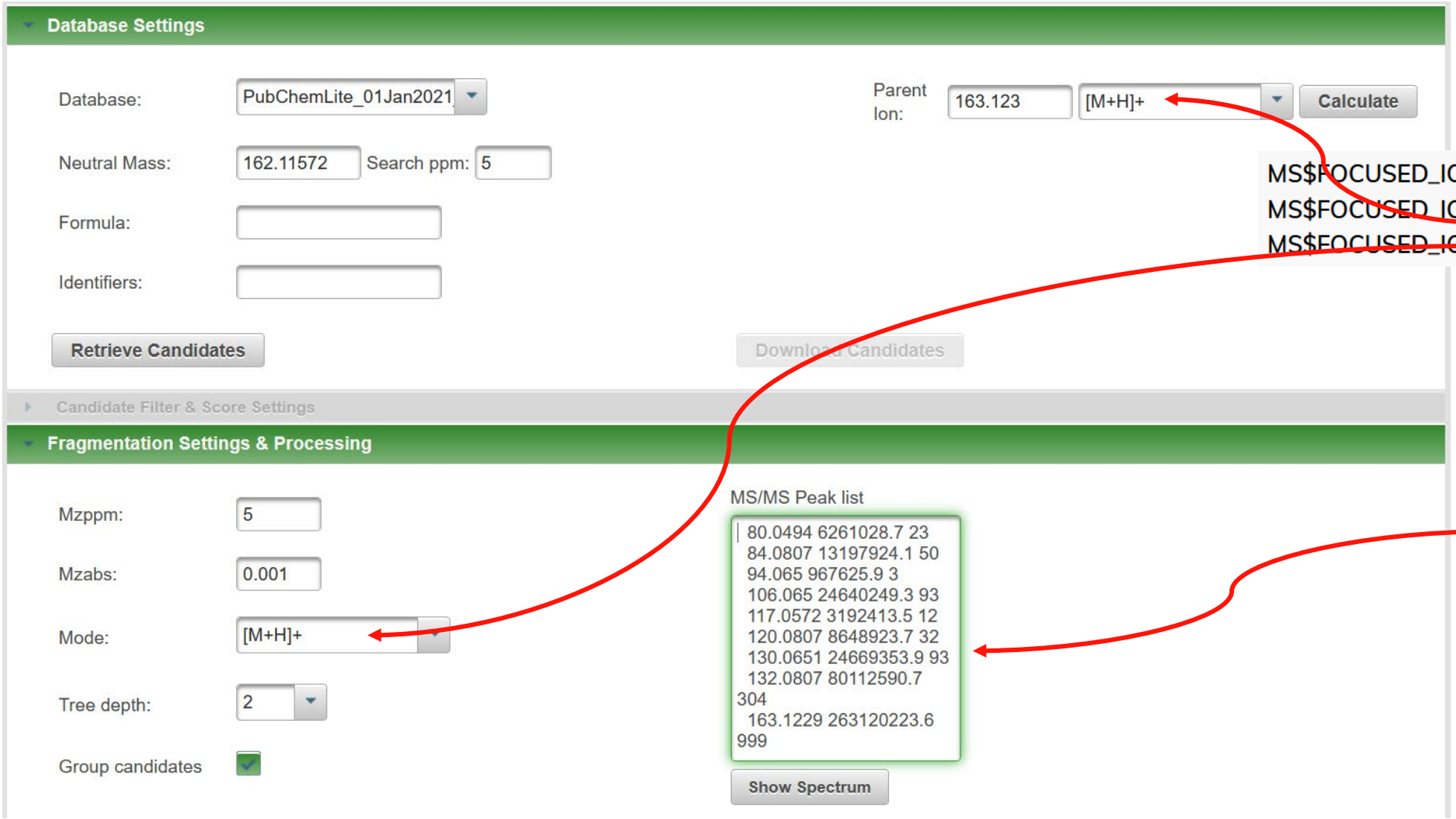
ACCESSION: EQ300801
RECORD_TITLE: Nicotine; LC-ESI-QFT; MS2; CE: 15; R=35000; [M+H]+
DATE: 2015.08.25
AUTHORS: Beck B, Stravs M, Schymanski E, Singer H, Department of Environmenta

```
PK$NUM_PEAK: 9
PK$PEAK: m/z int. rel.int.
80.0494 6261028.7 23
84.0807 13197924.1 50
94.065 967625.9 3
106.065 24640249.3 93
117.0572 3192413.5 12
120.0807 8648923.7 32
130.0651 24669353.9 93
132.0807 80112590.7 304
163.1229 263120223.6 999
//
```

MetFrag

https://massbank.eu/MassBank/RecordDisplay?id=EQ300801

MetFrag Example: Nicotine and Disease Associations



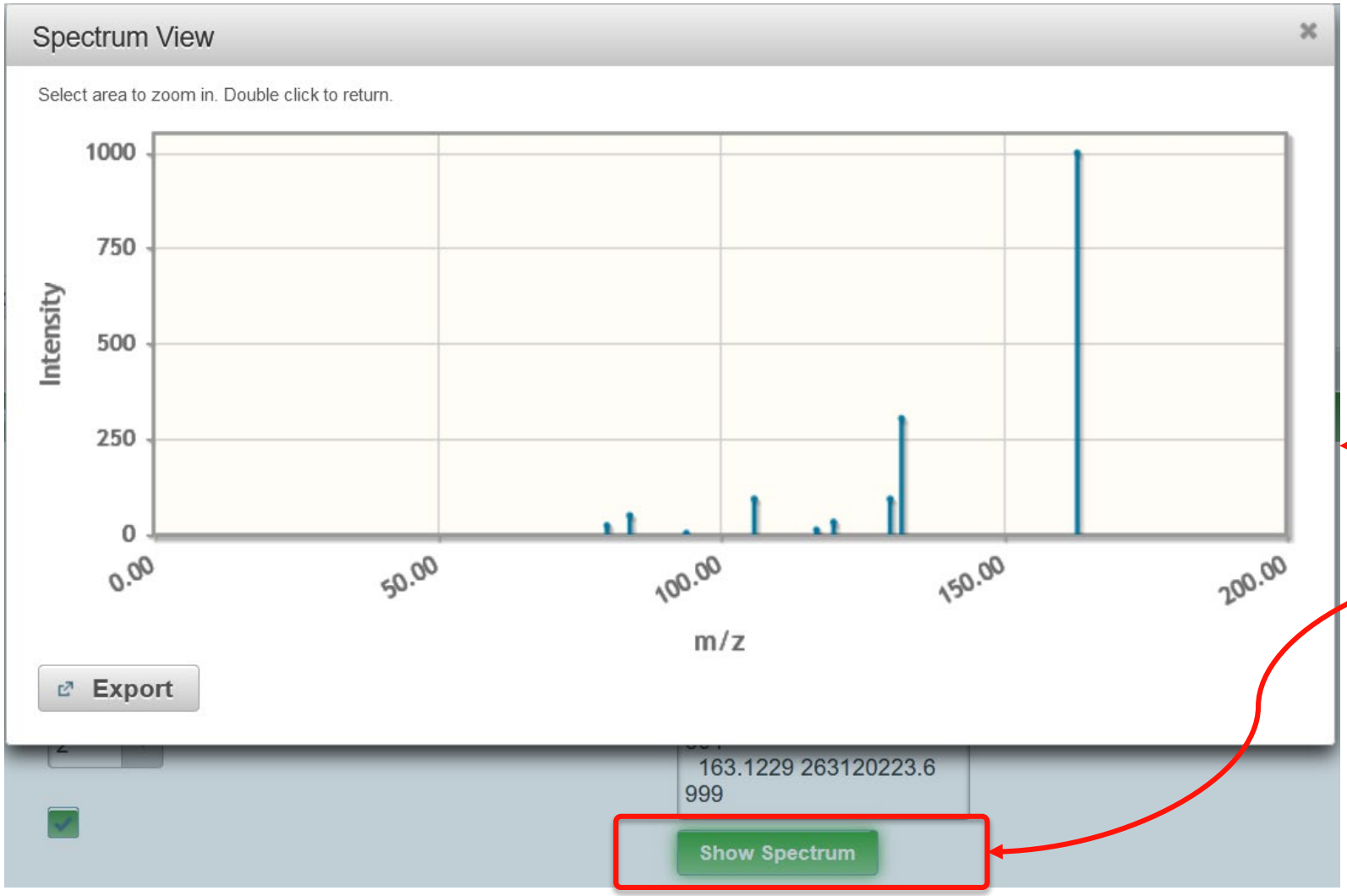
MS$FOCUSED_ION: BASE_PEAK 163.1229
MS$FOCUSED_ION: PRECURSOR_M/Z 163.123
MS$FOCUSED_ION: PRECURSOR_TYPE [M+H]+

PK$NUM_PEAK: 9
PK$PEAK: m/z int. rel.int.
80.0494 6261028.7 23
84.0807 13197924.1 50
94.065 967625.9 3
106.065 24640249.3 93
117.0572 3192413.5 12
120.0807 8648923.7 32
130.0651 24669353.9 93
132.0807 80112590.7 304
163.1229 263120223.6 999
//

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

MetFrag Example: Nicotine and Disease Associations

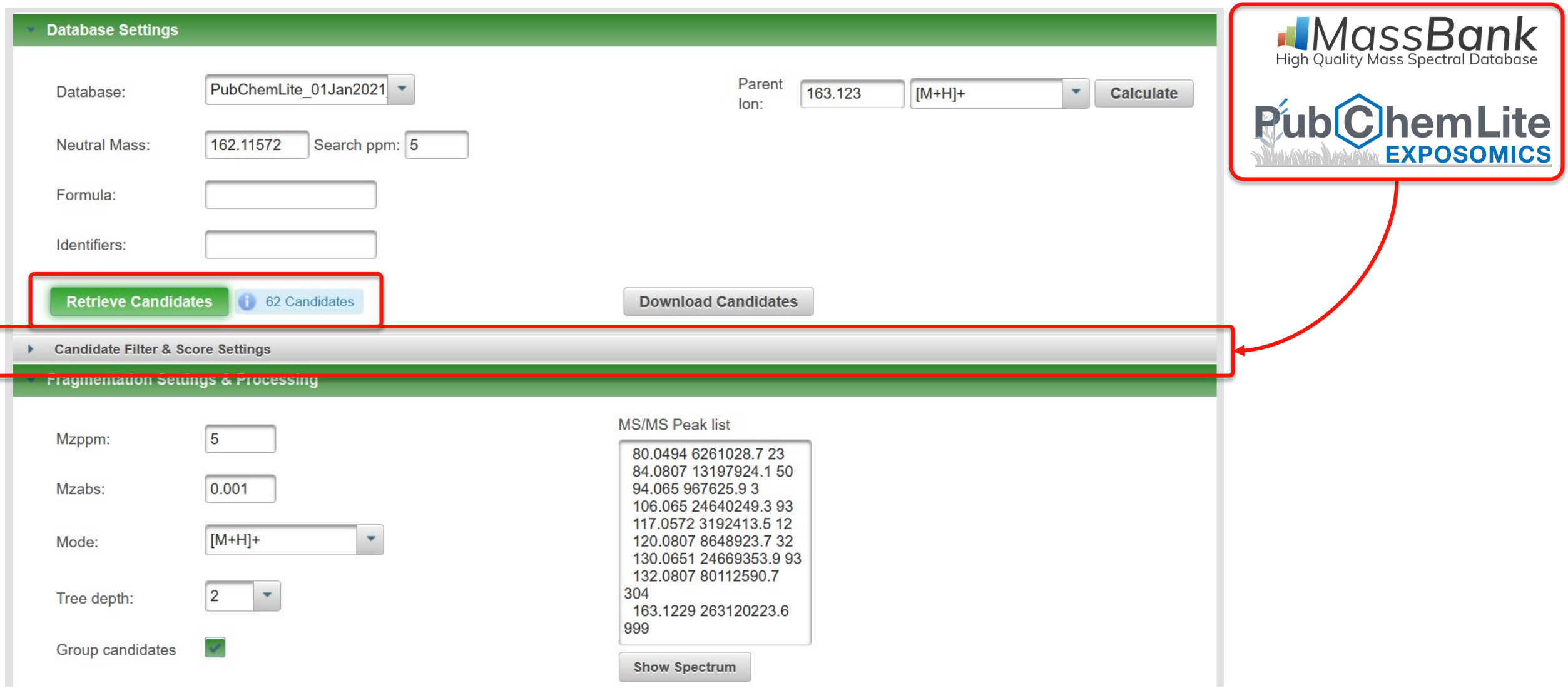
Check the spectrum (to see it copied well)



PK$NUM_PEAK: 9
PK$PEAK: m/z int. rel.int.
80.0494 6261028.7 23
84.0807 13197924.1 50
94.065 967625.9 3
106.065 24640249.3 93
117.0572 3192413.5 12
120.0807 8648923.7 32
130.0651 24669353.9 93
132.0807 80112590.7 304
163.1229 263120223.6 999
//

https://massbank.eu/MassBank/RecordDisplay?id=EQ300801

MetFrag Example: Nicotine and Disease Associations

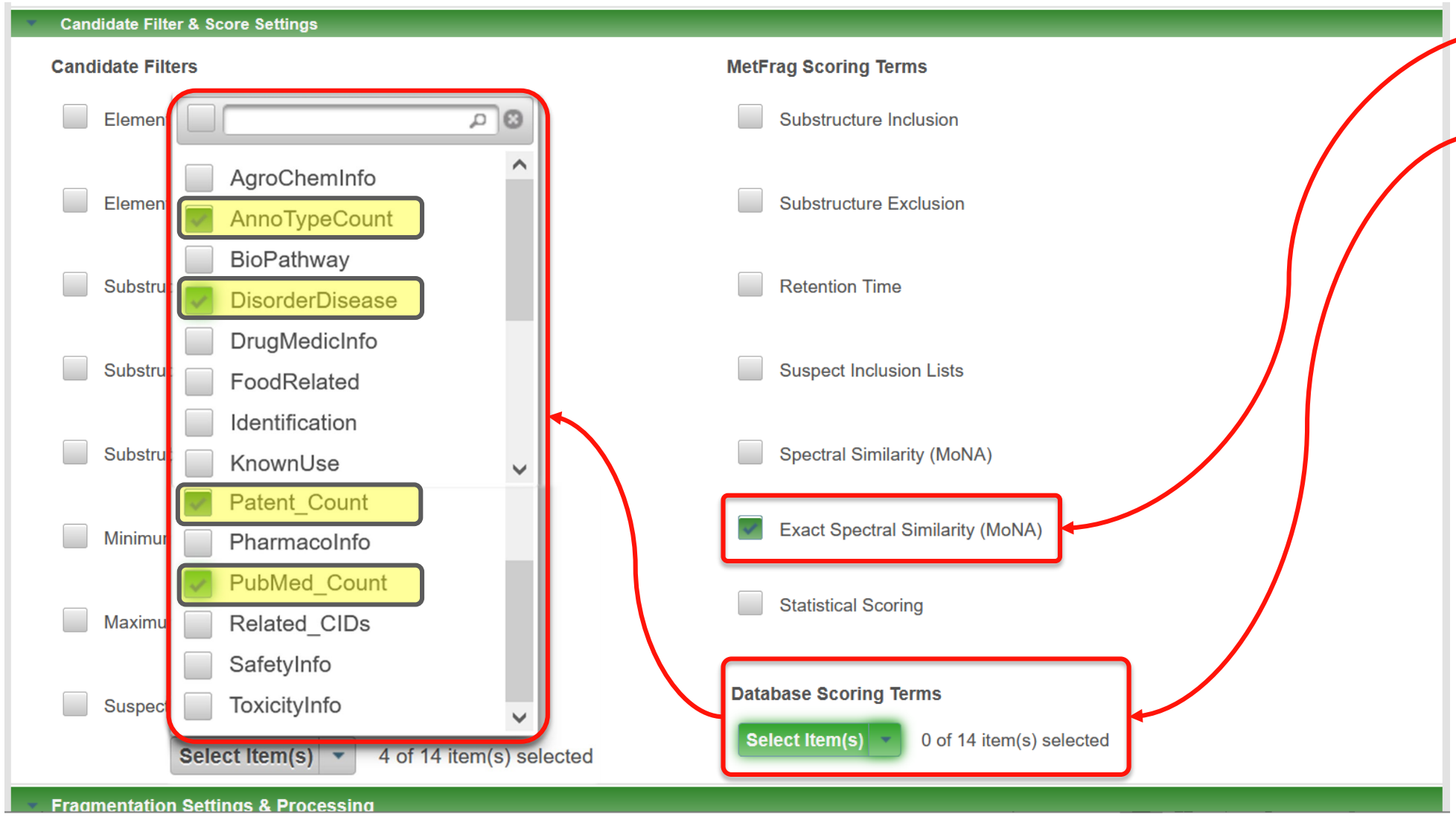


MassBank
High Quality Mass Spectral Database

PubChemLite
EXPOSOMICS

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

MetFrag Example: Nicotine and Disease Associations



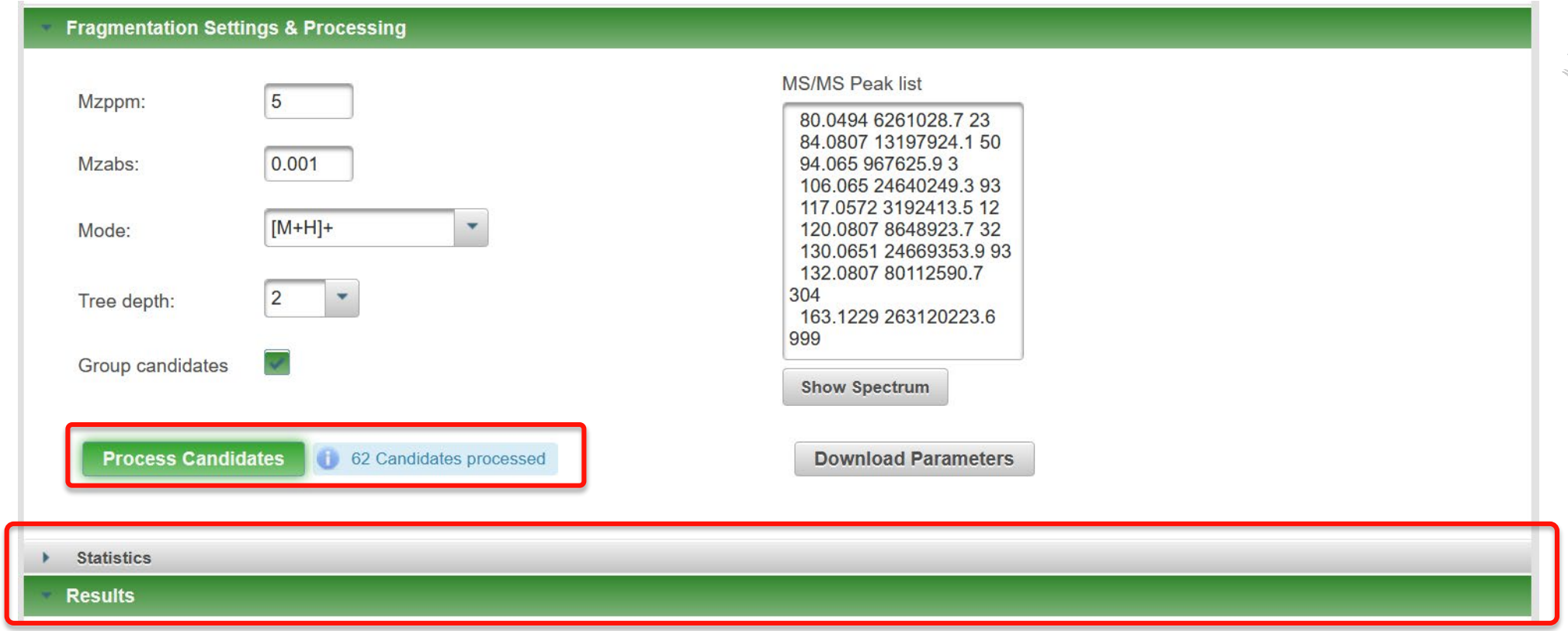
MassBank
High Quality Mass Spectral Database

PubChemLite
EXPOSOMICS

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

MetFrag Example: Nicotine and Disease Associations

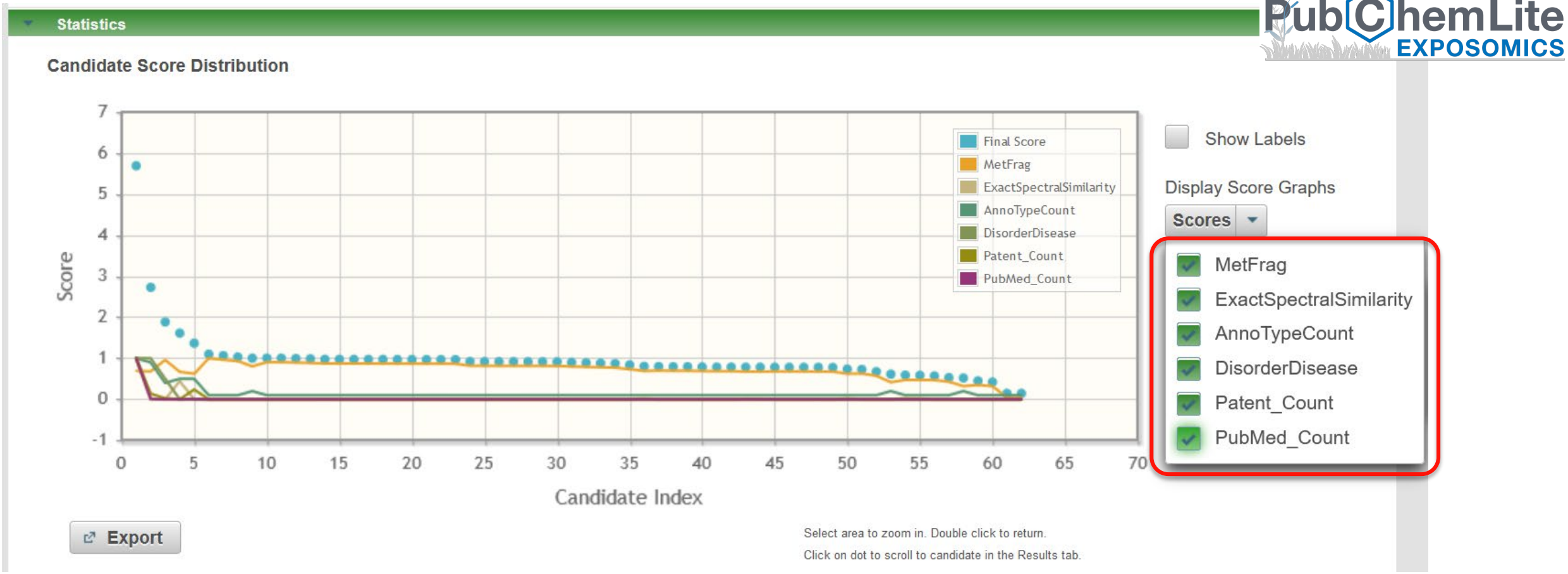
Process the candidates!



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

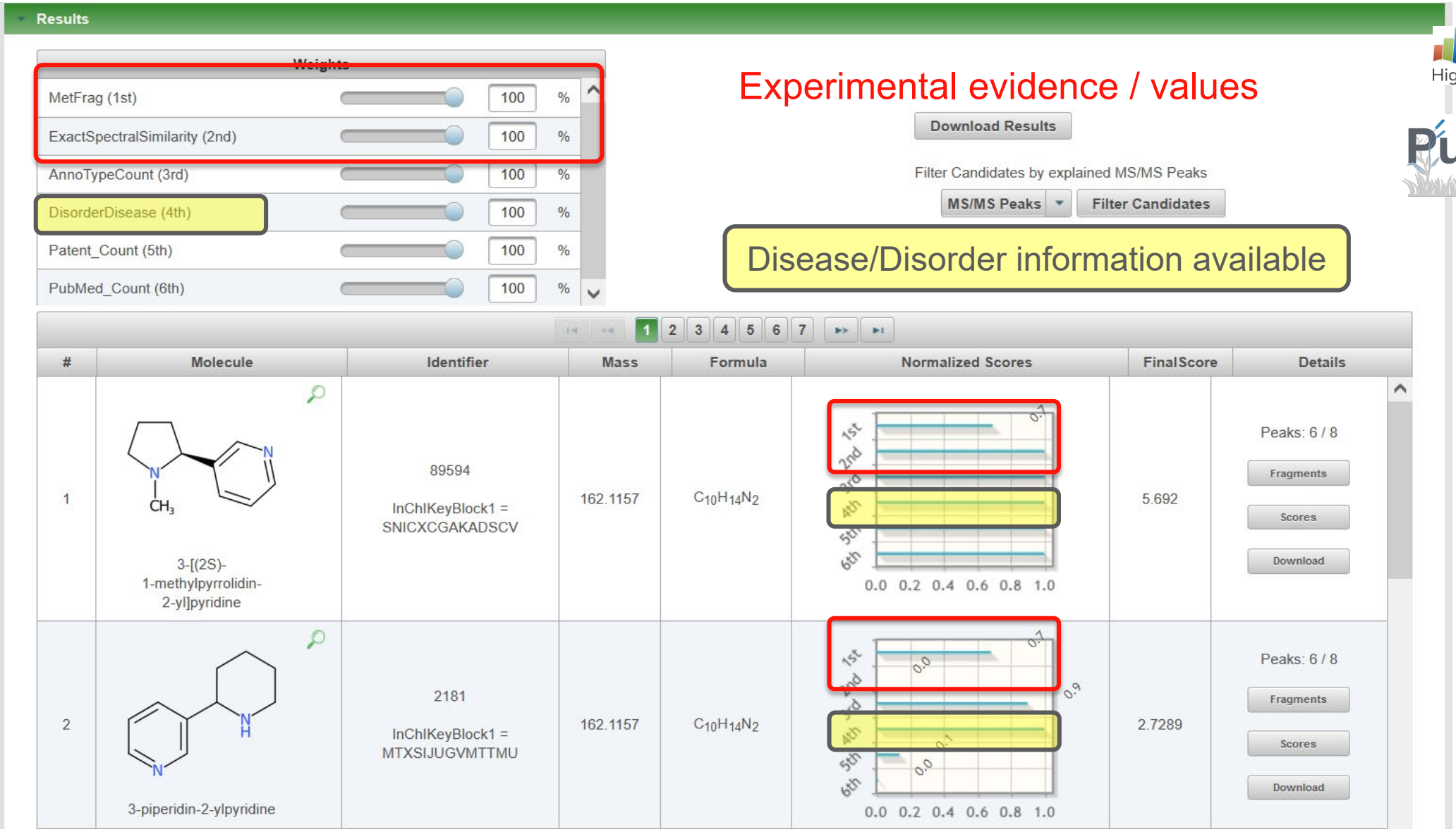
MetFrag Example: Nicotine and Disease Associations

Overview of all candidates: “Statistics” Plot



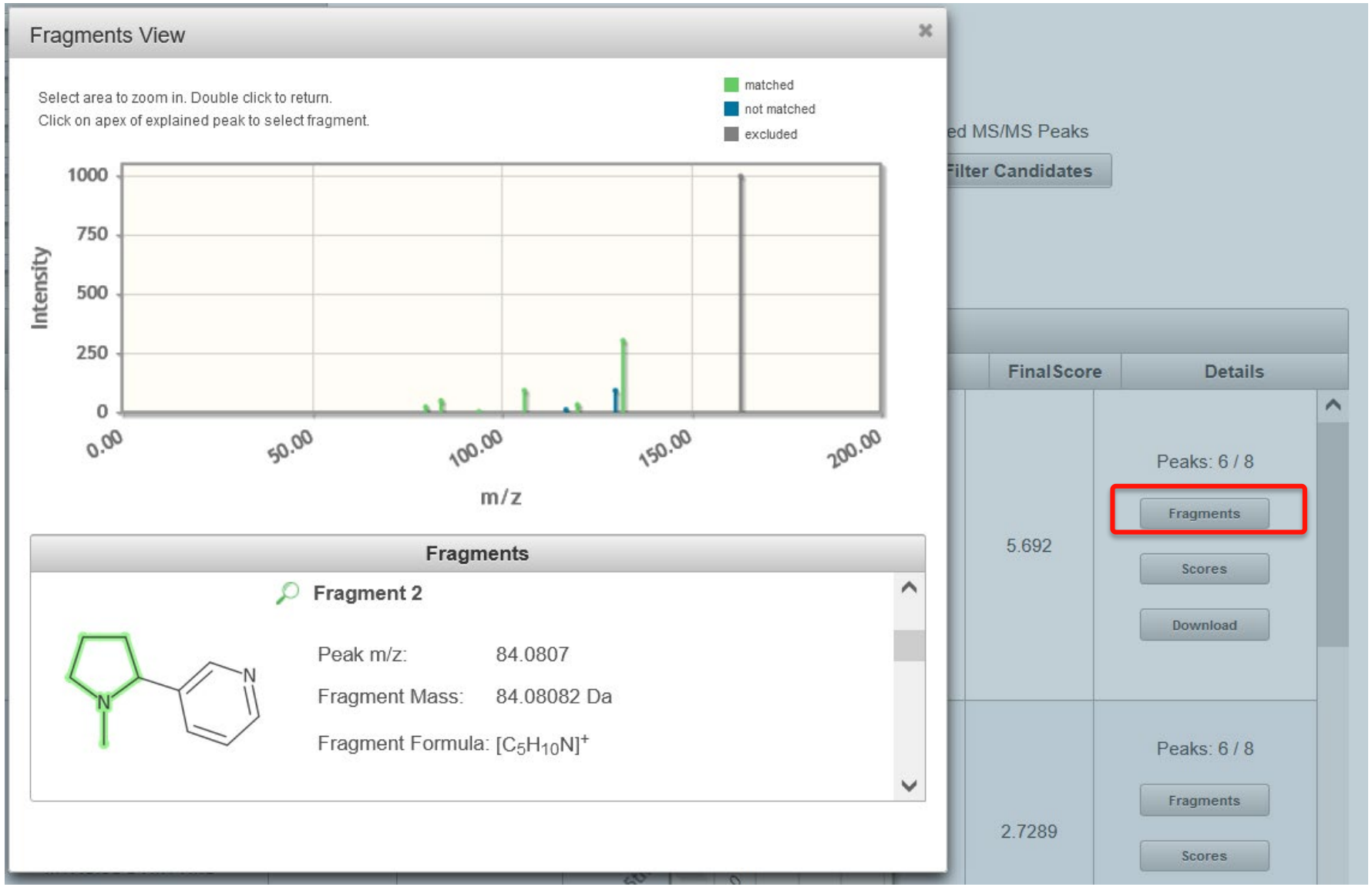
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9

MetFrag Example: Nicotine and Disease Associations



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

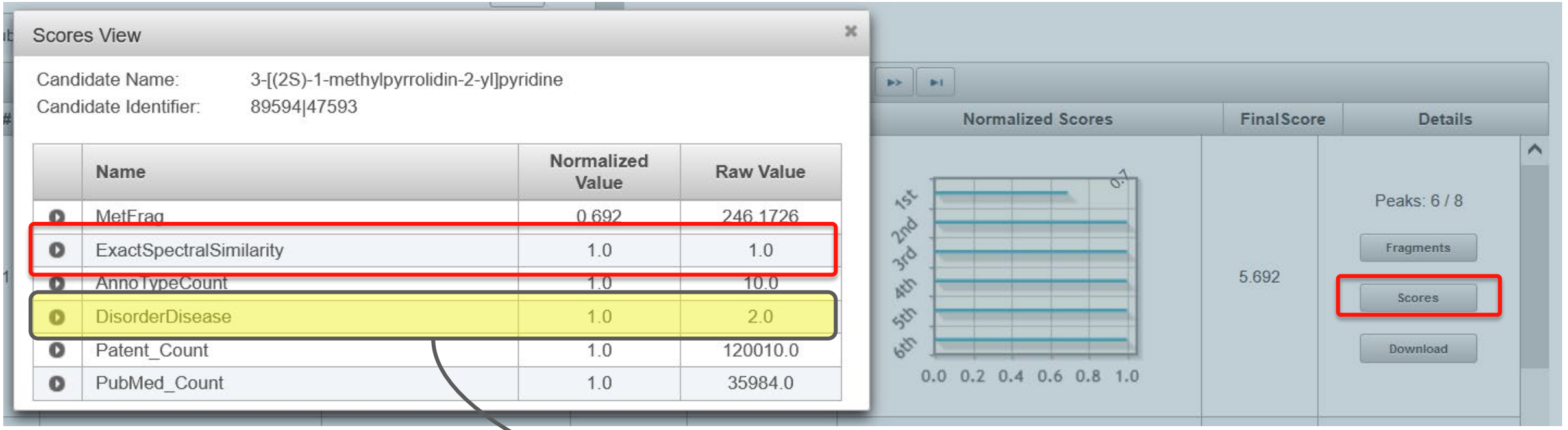
MetFrag Example: Nicotine and Disease Associations



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

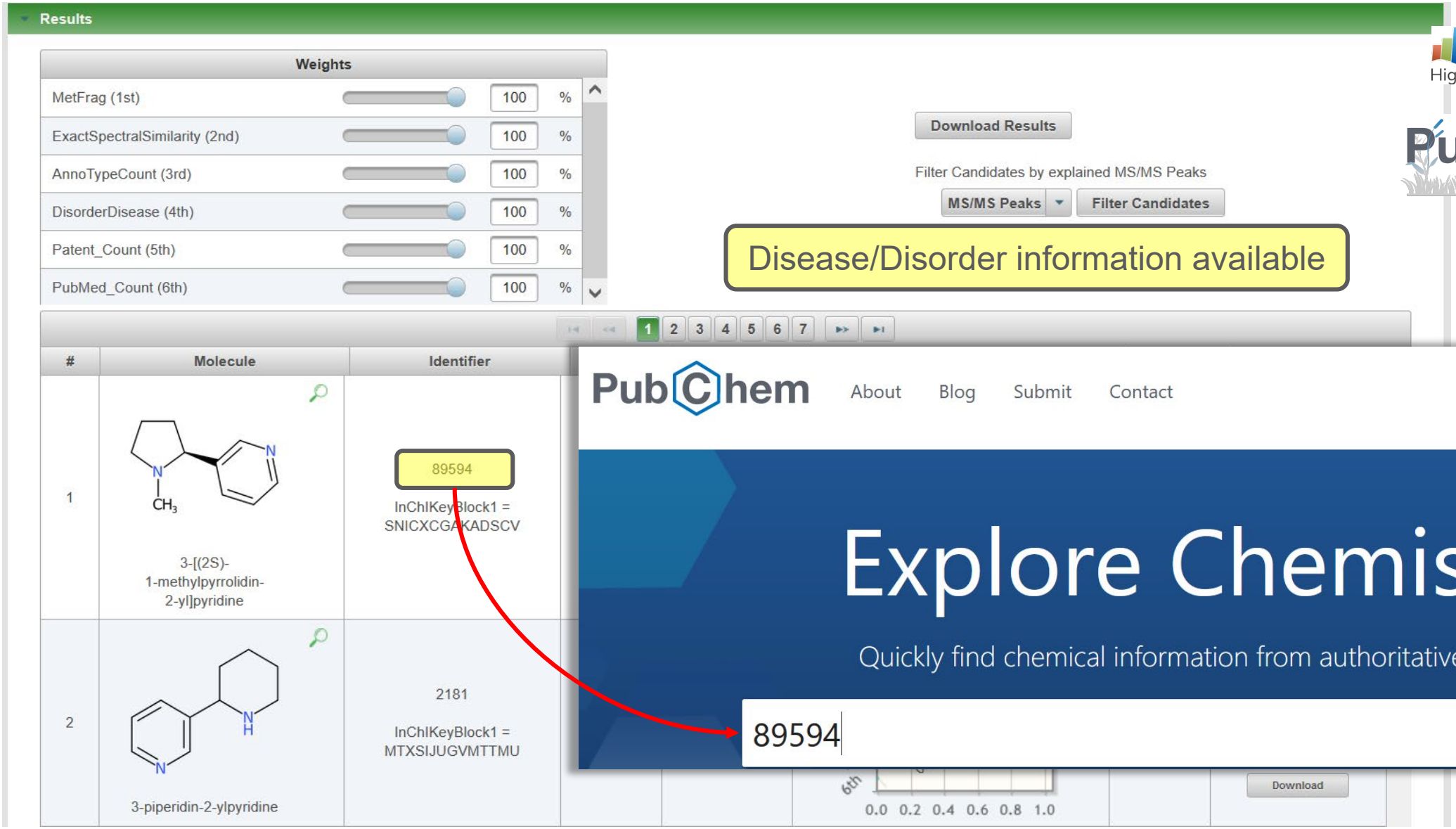
MetFrag Example: Nicotine and Disease Associations

Level 2: Probable structure	
a) by library spectrum match	MS, MS^2, **Library MS^2 >0.9 sim**
b) by diagnostic evidence	MS, MS^2, Exp. data



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

MetFrag Example: Nicotine and Disease Associations



Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://pubchem.ncbi.nlm.nih.gov/#query=89594

PubChem Annotations: Disease Associations

PubChem Nicotine (Compound)

14 Associated Disorders and Diseases

246 items View More Rows & Details

Download

Disease	Evidence Type	Evidence PMID
Abdominal Pain	marker/mechanism	9179092
Abnormalities, Drug-Induced	marker/mechanism	16193507
Acute Kidney Injury	marker/mechanism	21511693
Airway Obstruction	marker/mechanism	33082140
Akinetic Mutism	therapeutic	19455735

1 2 3 ... 50 Next >

Comparative Toxicogenomics Database (CTD)

Disease	References
	PubMed: 9137998, 8294547, 10636262, 6890513,

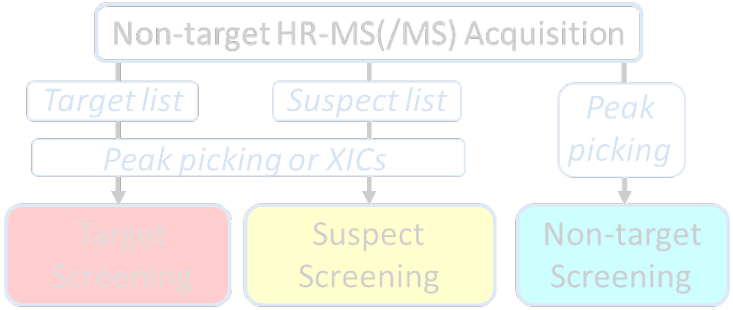
Cite Download

CONTENTS

https://pubchem.ncbi.nlm.nih.gov/compound/89594#section=Associated-Disorders-and-Diseases

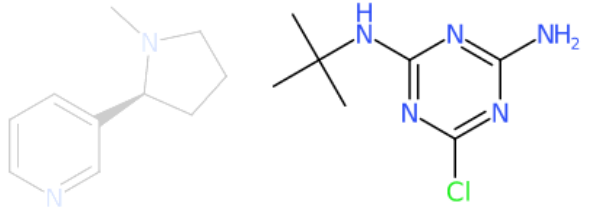
Outline for today

- Background
  - Environmental cheminformatics & HR-MS
  - Non-target screening and identification confidence
  - Compound databases and spectral libraries
- Introduction to PubChemLite and MetFrag
- Examples (MassBank+PubChemLite+MetFrag)
  - Nicotine – relating identifications to disease outcomes
  - Desethylterbutylazine – isobars and the role of metadata
- Closing – Perspectives and Community Contribution



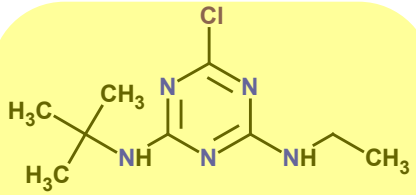
PubChemLite EXPOSOMICS

MetFrag

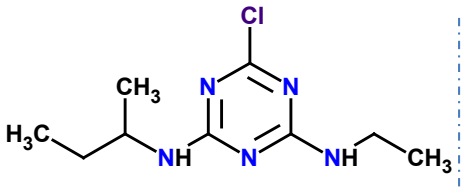


MetFrag Example: Isobars and Metadata!

$C_9H_{16}ClN_5$ *m/z 229.1094 Da*

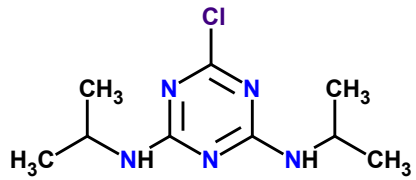


Terbutylazine
Detects: 12; # Refs: 220



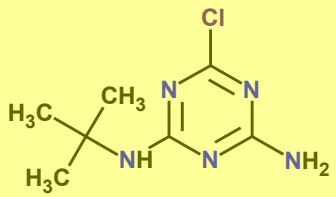
Sebutylazine
Detects: 3; # Refs: 51

(no related compound at this mass)

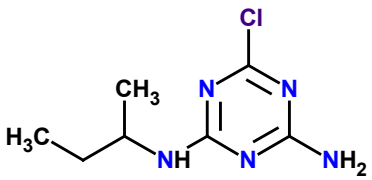


Propazine
Detects: 3; # Refs: 201

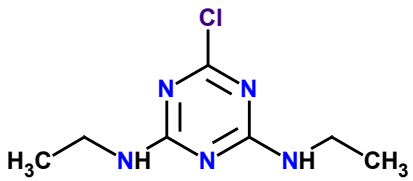
$C_7H_{12}ClN_5$ *m/z 201.0781 Da*



Terbutylazine-desethyl
Detects: 9; # Refs: 92



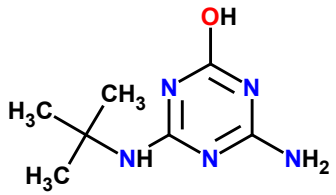
Sebutylazine-desethyl
Detects: 1; # Refs: 14



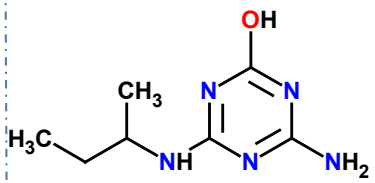
Simazine
Detects: 4; # Refs: 518

(no related compound at this mass)

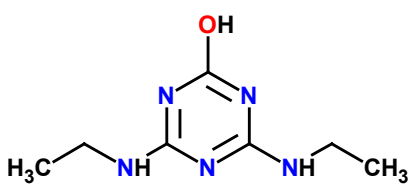
$C_7H_{13}N_5O$ *m/z 183.1120 Da*



Terbutylazine-desethyl-2-hydroxy
Detects: 2; # Refs: 57



Sebutylazine-desethyl-2-hydroxy
Detects: 0; # Refs: 3



Simazine-2-hydroxy

Detects: 2; # Refs: 66

(no related compound at this mass)

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

MetFrag Example: Isobars and Metadata!

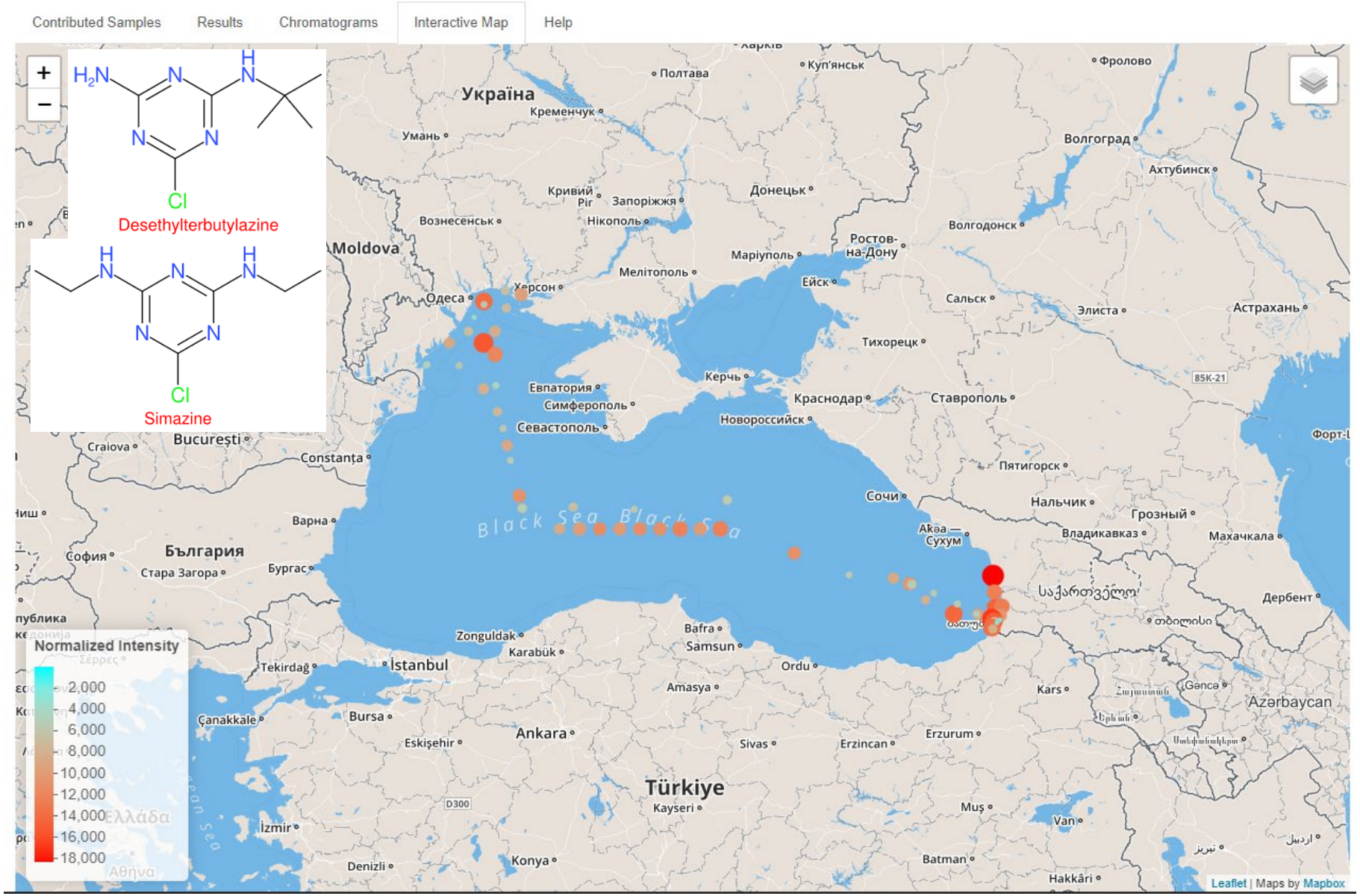


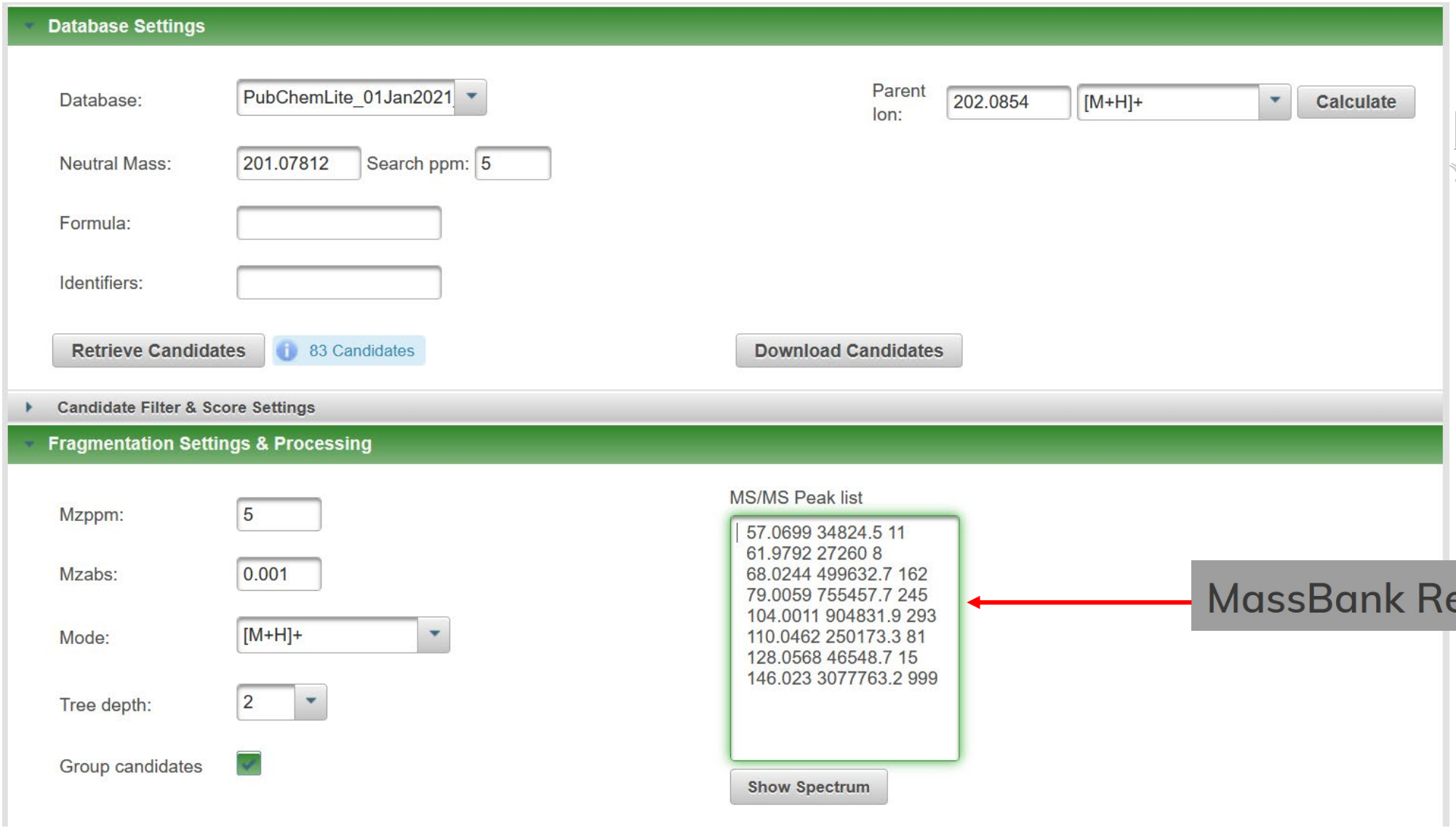
Image provided by Nikiforos Alygizakis. DSFP: Alygizakis *et al*, 2019, TrAC, DOI: 10.1016/j.trac.2019.04.008

Joint Black Sea Survey 2016

NORMAN Digital Sample Freezing Platform

Two very common isomers …

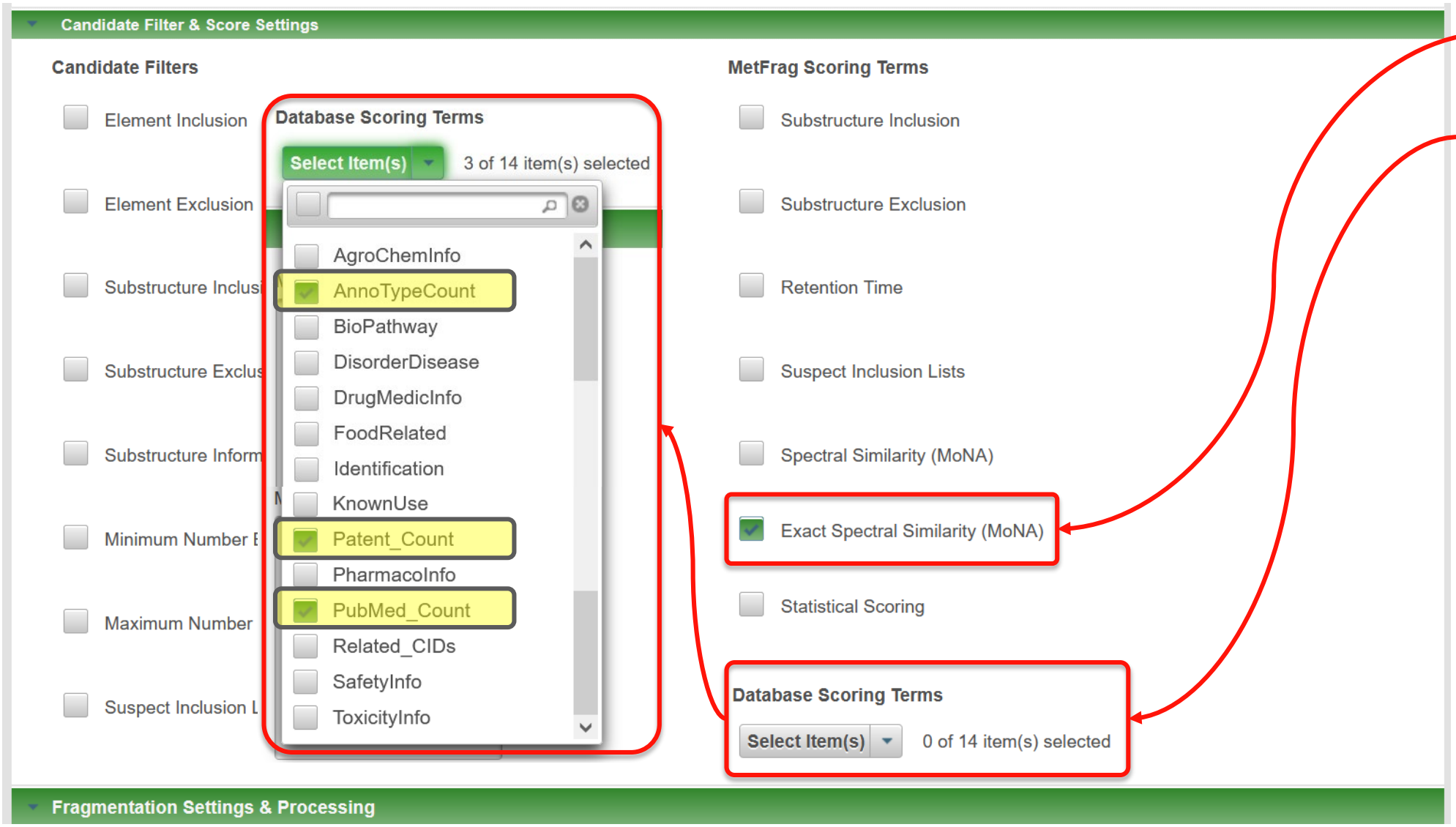
MetFrag Example: Isobars and Metadata!



MassBank Record: EA067112

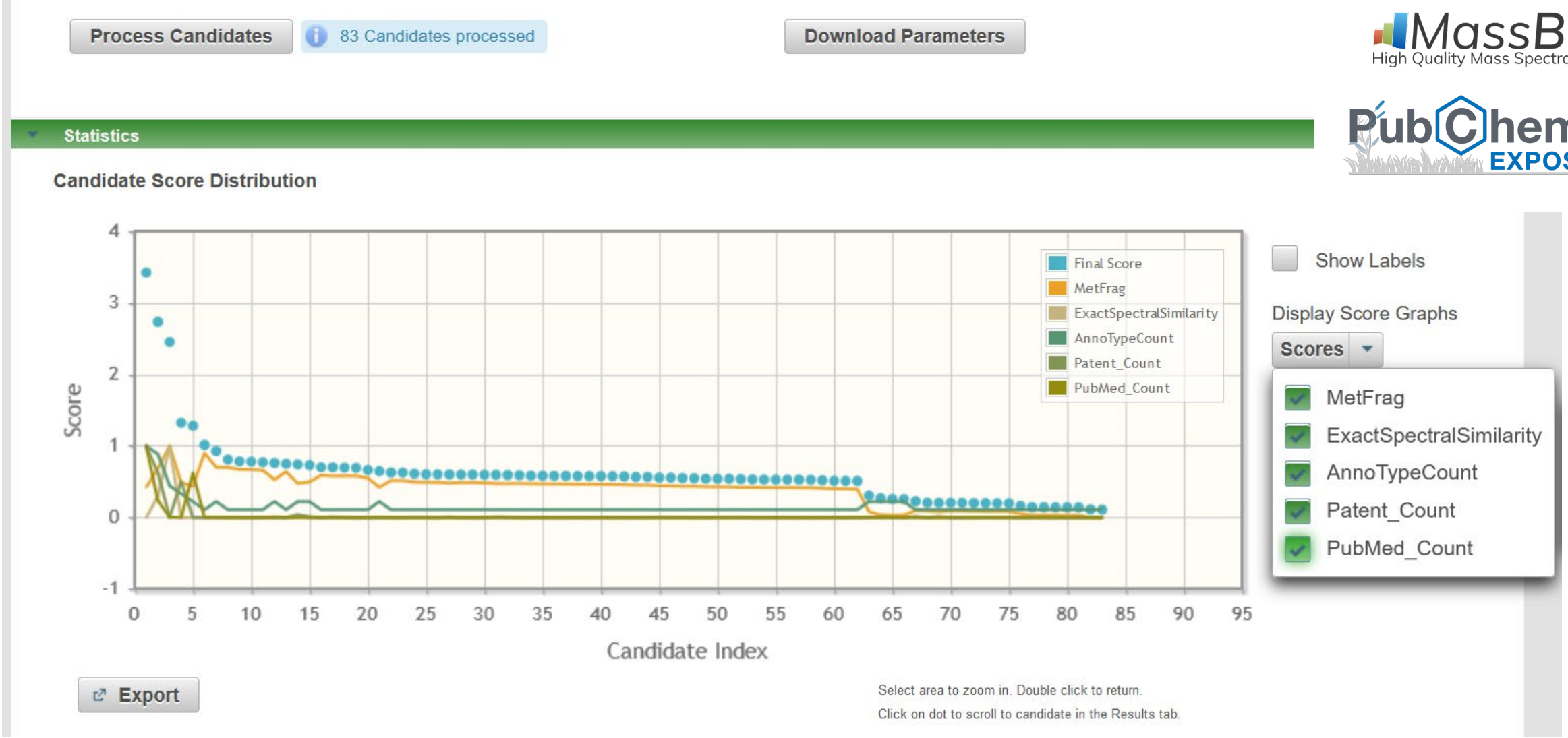
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

MetFrag Example: Isobars and Metadata!



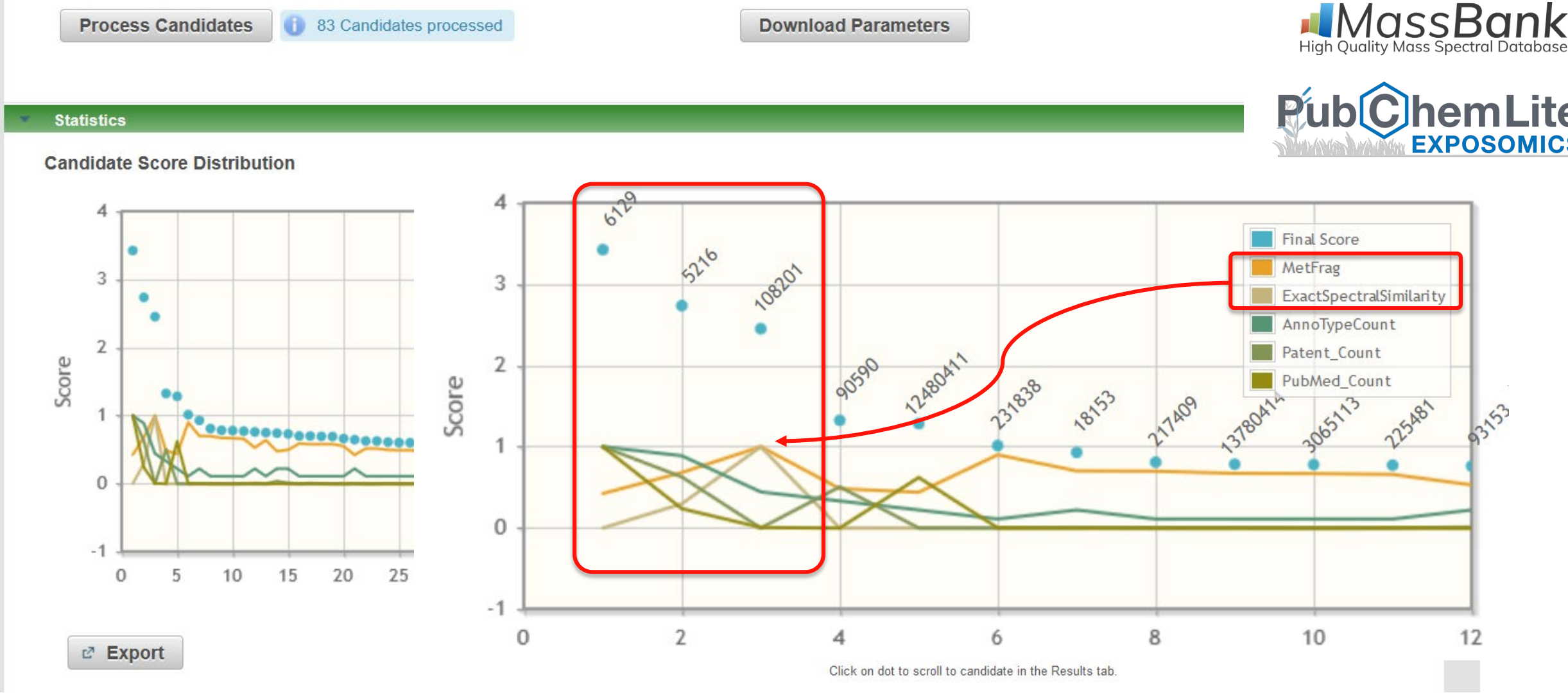
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

MetFrag Example: Isobars and Metadata!



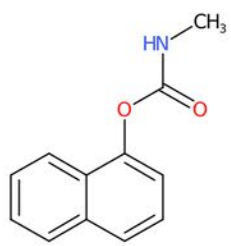
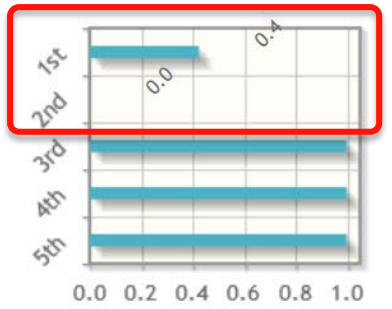
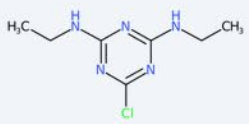
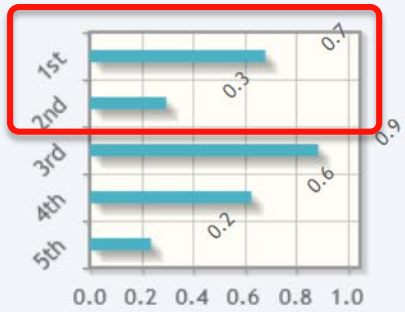
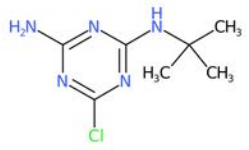
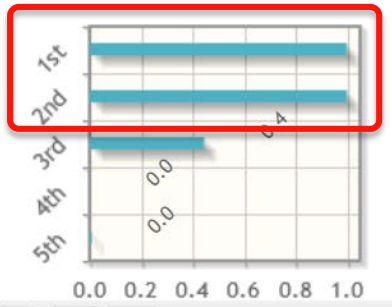
Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

MetFrag Example: Isobars and Metadata!



Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

MetFrag Example: Isobars and Metadata!

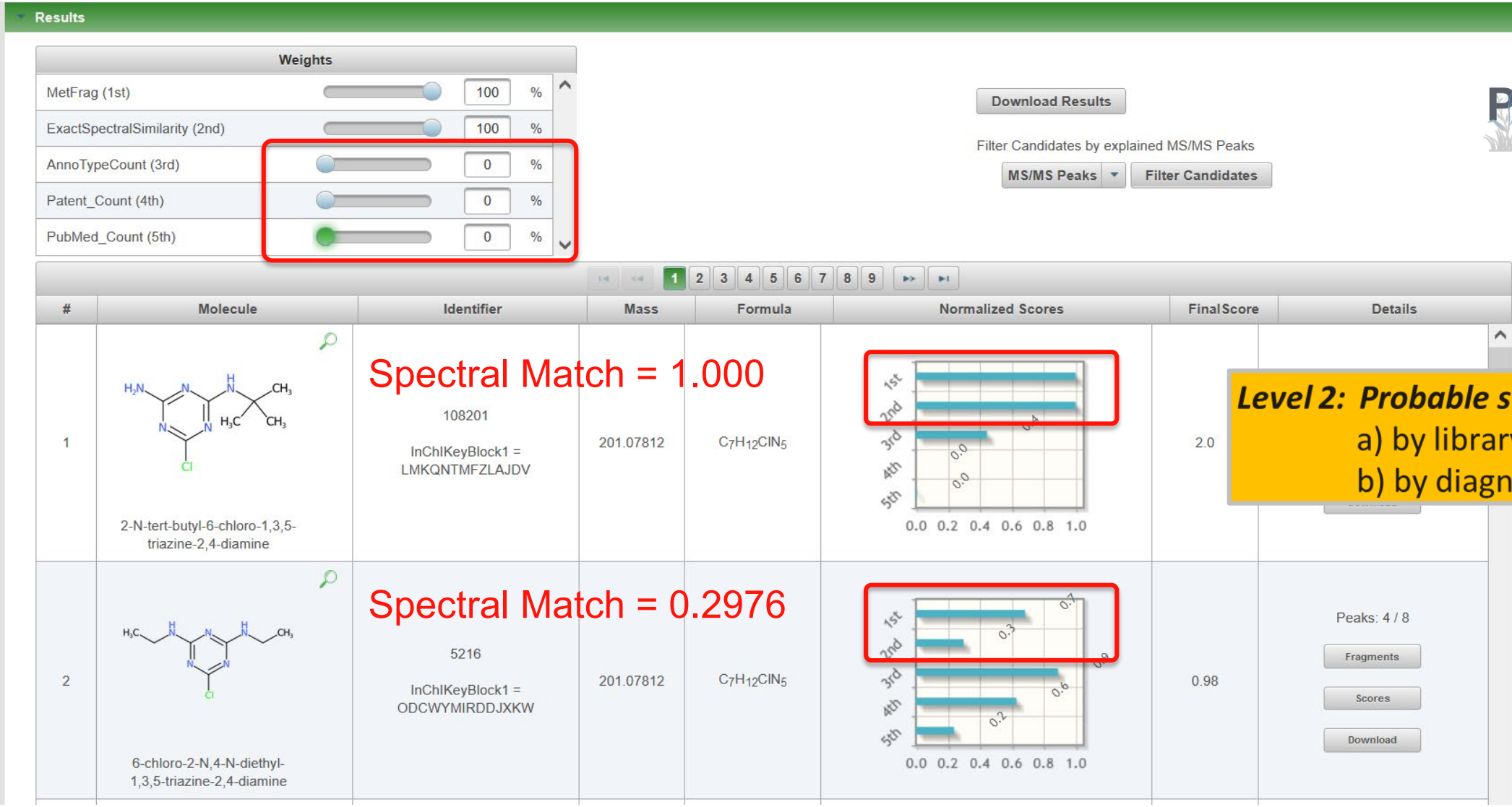
#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1	 naphthalen-1-yl N-methylcarbamate	Spectral Match = 0, Frag 0.4 6129 InChIKeyBlock1 = CVXBEEMKQHEXEN	201.07898	$C_{12}H_{11}NO_2$		3.4241	Peaks: 2 / 8 Fragments Scores Download
2	 6-chloro-2-N,4-N-diethyl-1,3,5-triazine-2,4-diamine	Spectral Match = 0.2976 5216 InChIKeyBlock1 = ODCWYMIRDDJXKW	201.07812	$C_7H_{12}ClN_5$		2.7357	Peaks: 4 / 8 Fragments Scores Download
3	 2-N-tert-butyl-6-chloro-1,3,5-triazine-2,4-	Spectral Match = 1.000 108201 InChIKeyBlock1 = LMKQNTMFZLAJDV	201.07812	$C_7H_{12}ClN_5$			Peaks: 6 / 8

1 2 3 4 5 6 7 8 9

Level 2: Probable structure
a) by library spectrum match
b) by diagnostic evidence

Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

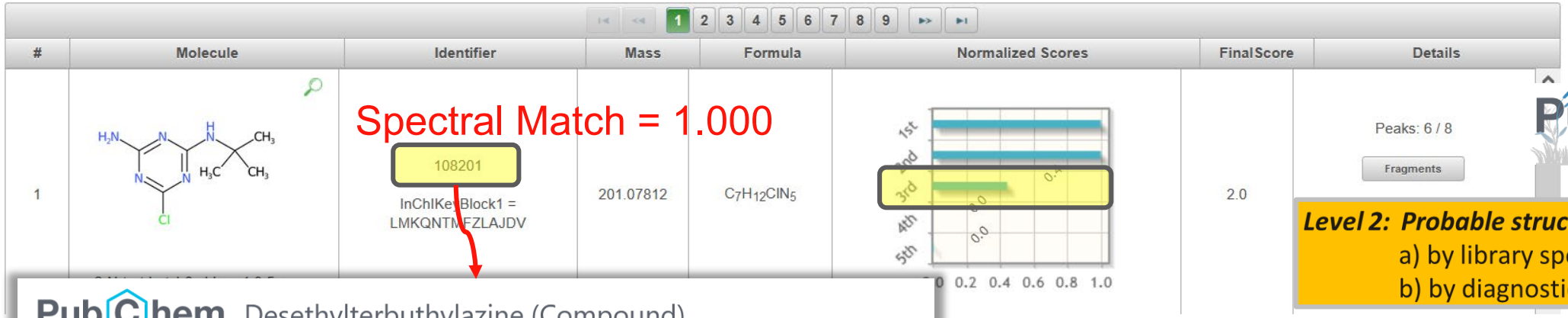
MetFrag Example: Isobars & Adjusted Weighting



Level 2: Probable structure
a) by library spectrum match
b) by diagnostic evidence

Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://massbank.eu/MassBank/RecordDisplay?id=EA067112

MetFrag & Isobars: Supporting Annotation Content

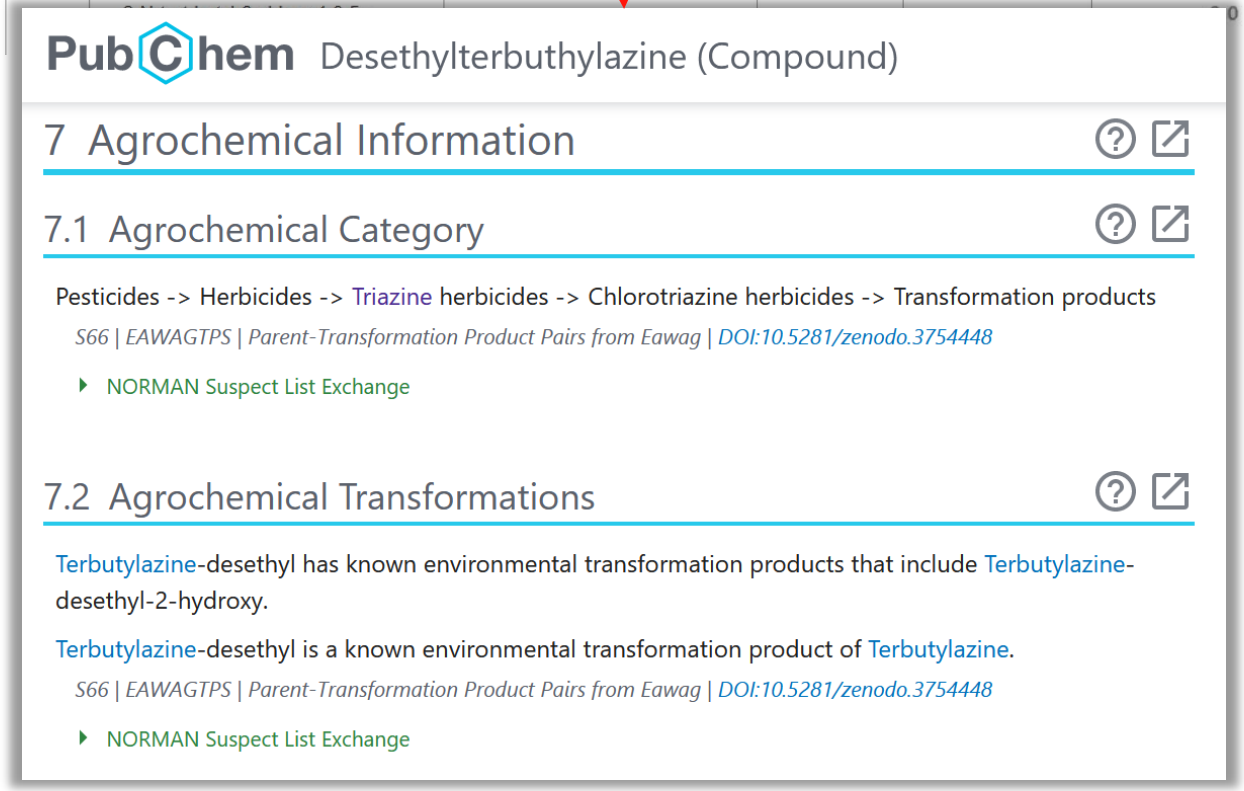


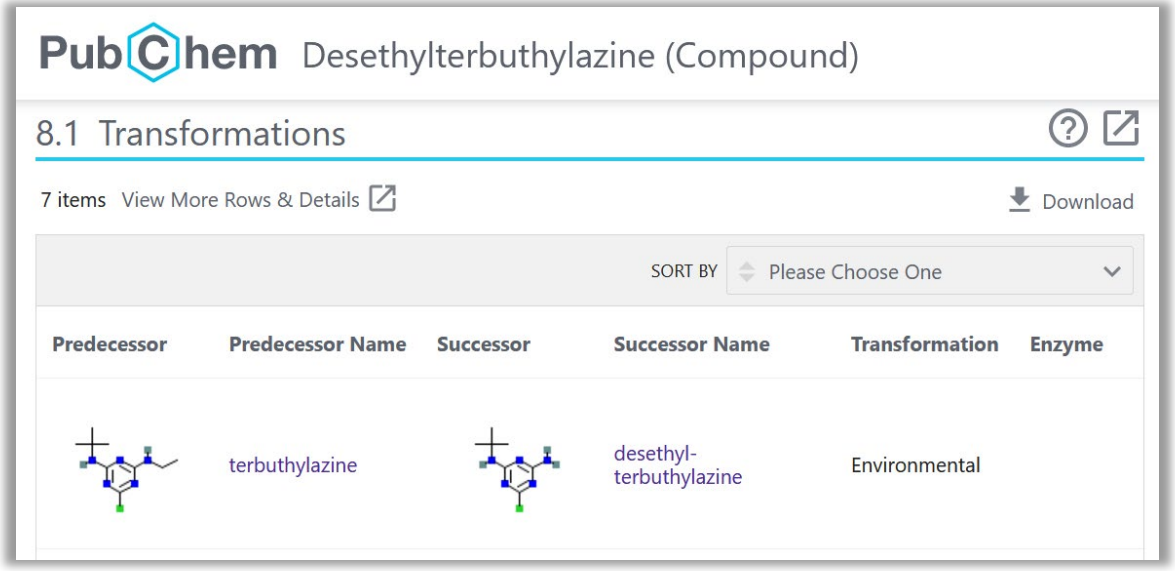
Spectral Match = 1.000

MassBank
High Quality Mass Spectral Database

PubChemLite
EXPOSOMICS

Level 2: Probable structure
a) by library spectrum match
b) by diagnostic evidence

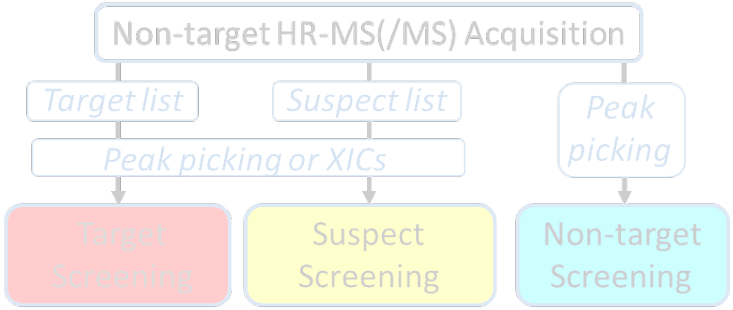




Schymanski *et al*. (2021) DOI: 10.1186/s13321-016-0115-9 and https://pubchem.ncbi.nlm.nih.gov/compound/108201#section=Transformations

Outline for today

- Background
  - Environmental cheminformatics & HR-MS
  - Non-target screening and identification confidence
  - Compound databases and spectral libraries
- Introduction to PubChemLite and MetFrag
- Examples (MassBank+PubChemLite+MetFrag)
  - Nicotine – relating identifications to disease outcomes
  - Desethylterbutylazine – isobars and the role of metadata
- Closing – Perspectives and Community Contribution



PubChemLite EXPOSOMICS

Metfrag

Expert Knowledge is YOUR Knowledge!

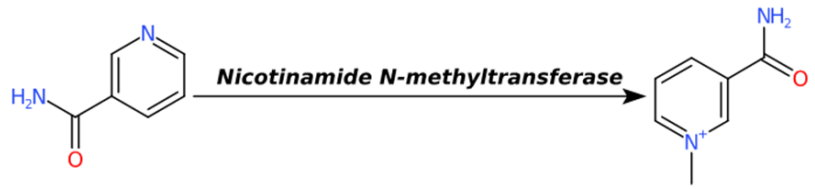
https://ftp.ncbi.nlm.nih.gov/pubchem/Other/Submissions/

- Help us help you! Add data with FAIR templates
  - Chemical Structures

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

  - Transformations

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



Nicotinamide to MNAM DOI:10.1124/dmd.112.049734

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

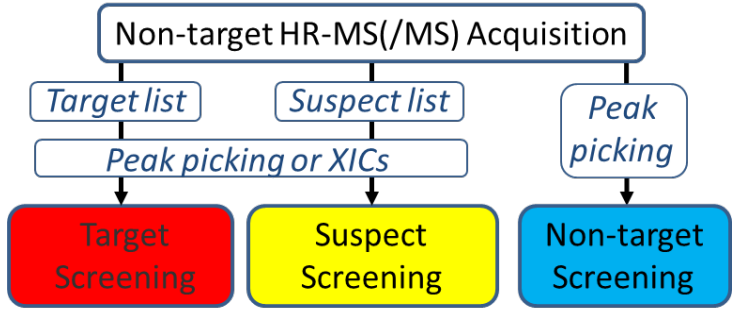
Schymanski & Bolton (2021) FAIR Chemical Structures. *J. Cheminform*. DOI: 10.1186/s13321-021-00520-4

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

“Take home” Messages

PubChemLite EXPOSOMICS

- Background of HR-MS and Cheminformatics

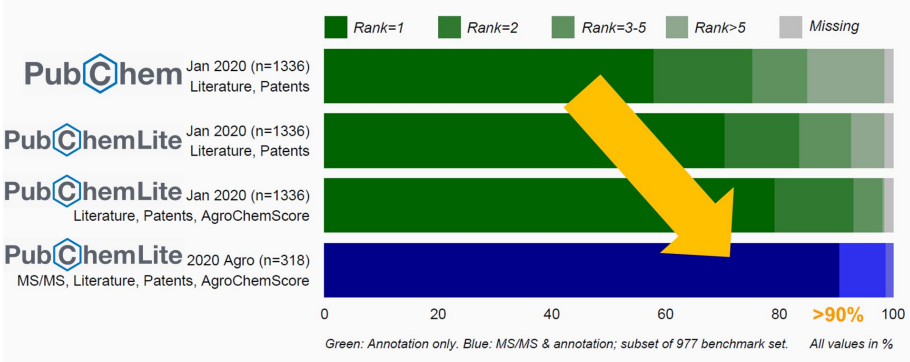


- Introduction to MetFrag and PubChemLite
  - Mass spectral libraries help deliver Level 2a IDs
  - Annotation content is extremely powerful

MassBank High Quality Mass Spectral Database

MetFrag

- Please try the two examples & interpretation
  - More examples in the additional exercises



- Consider contributing your knowledge!

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

Thank you!

PubChemLite EXPOSOMICS

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

Slides @ 10.5281/zenodo.5115498

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

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

https://massbank.eu/MassBank/

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

Genes
Innovation
Environment
Healthcare
Modeling

MassBank
High Quality Mass Spectral Database

PubChem

… and team

MetFrag

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

Luxembourg National Research Fund