MetFrag: Annotating "Unknowns"

Assoc. Prof. Dr. Emma L. Schymanski

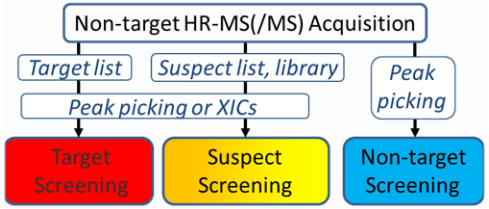
FNR ATTRACT Fellow and PI in Environmental Cheminformatics
Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg
Email: emma.schymanski@uni.lu and @ESchymanski

…and many colleagues who contributed to my science over the years!

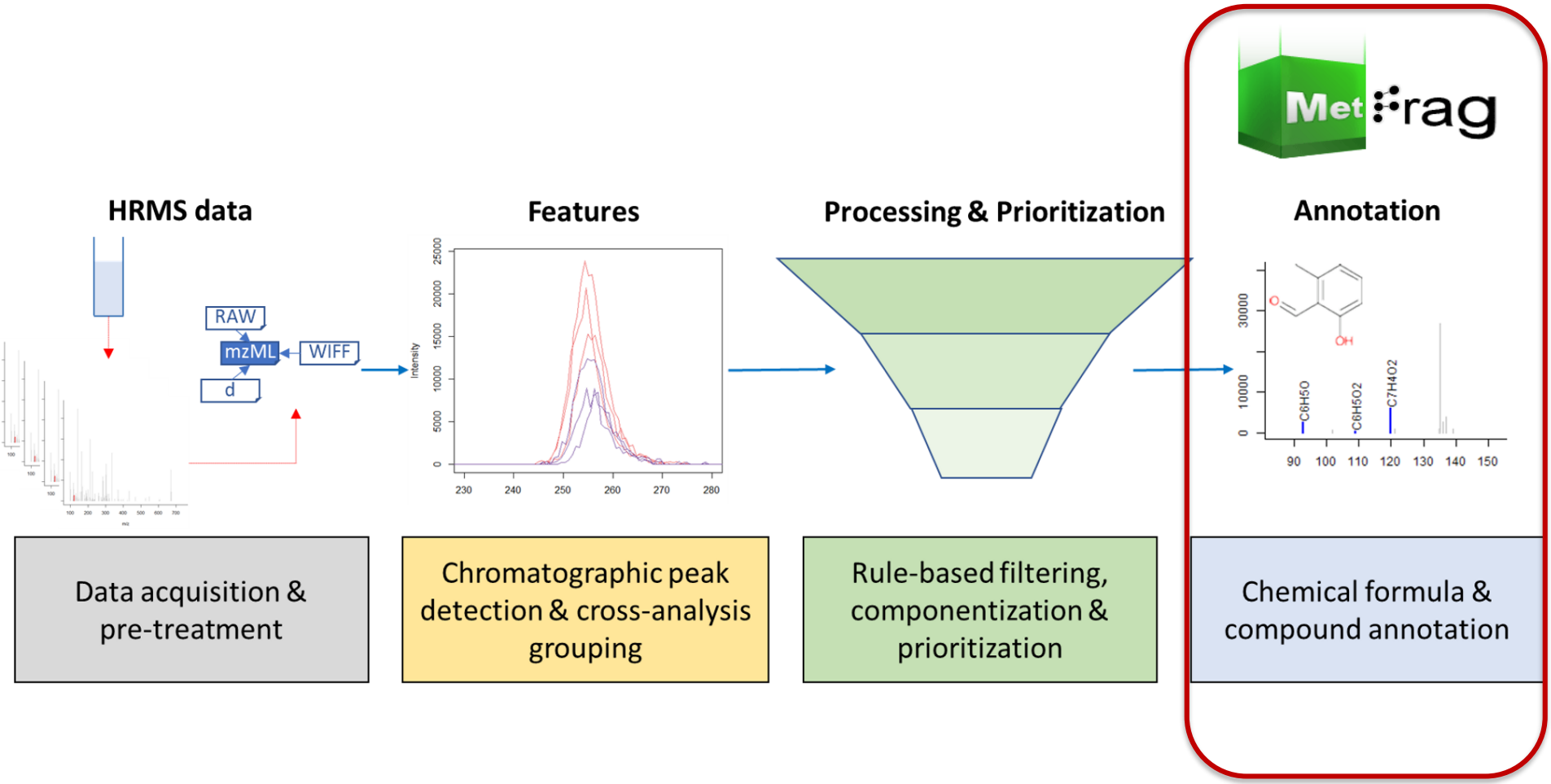
Talk available under DOI: 10.5281/zenodo.3953683

Plan for Today

- Background
  - Non-target Screening & Identification Confidence
  - Compound Databases and Spectral Libraries
- Introduction to MetFrag: MS/MS and Metadata
- MetFrag + CompTox + MassBank + Nicotine
  - MS-ready groupings … or not
- MetFrag + PubChem + MassBank + Isobars
  - The relevance of spectral matching!
- Future: MetFrag for High Throughput Exposomics
  - Build your own databases … PubChemLite



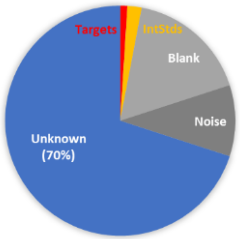
Overview: Identification with HR-MS

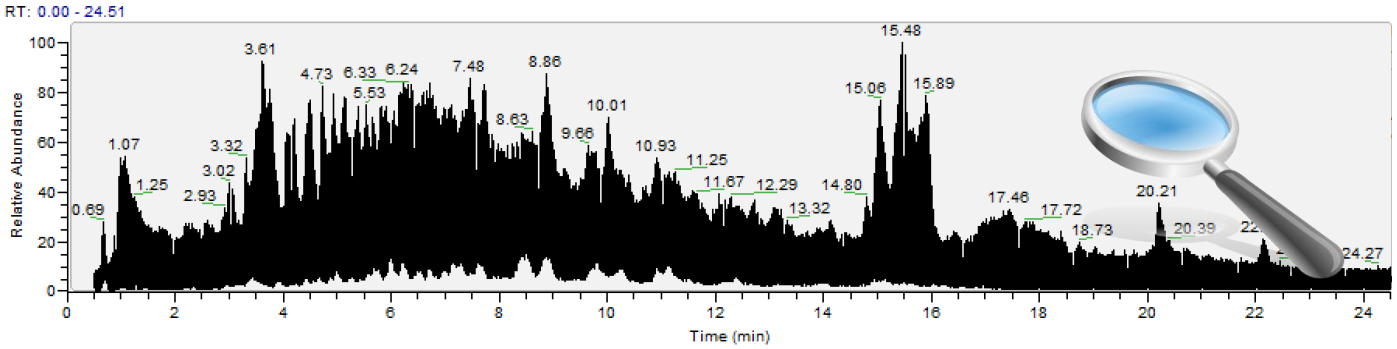


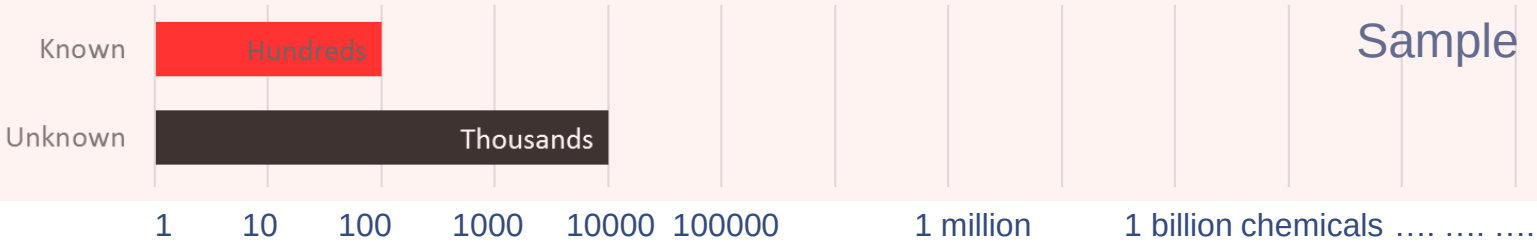
Helmus *et al.* submitted; preprint available from https://www.researchsquare.com/article/rs-36675/v1 DOI: 10.21203/rs.3.rs-36675/v1

Our (Community) Challenge: Identifying Chemicals

High resolution mass spectrometry





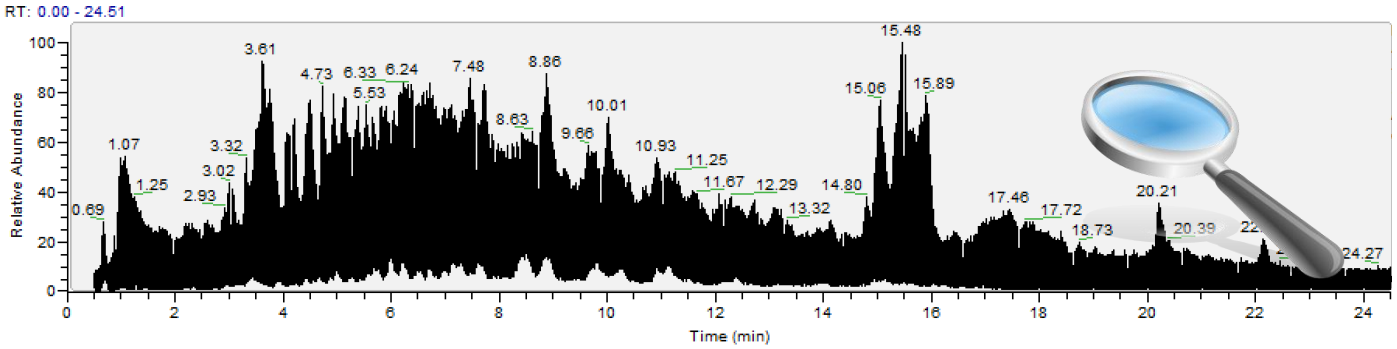


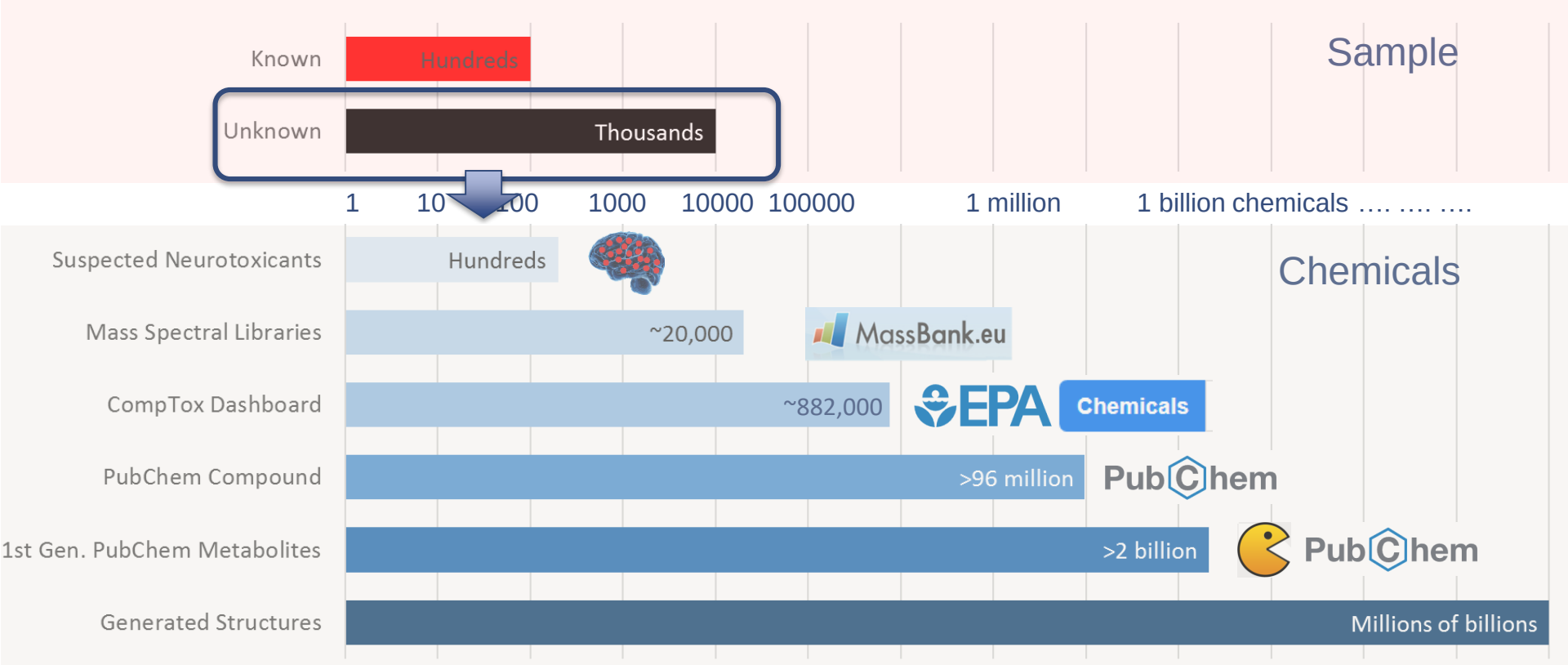
Data: Schymanski *et al* 2014, Environ. Sci. Technol. DOI: 10.1021/es4044374; Hollender *et al* 2017 DOI: 10.1021/acs.est.7b02184

Our (Community) Challenge: Identifying Chemicals

High resolution mass spectrometry

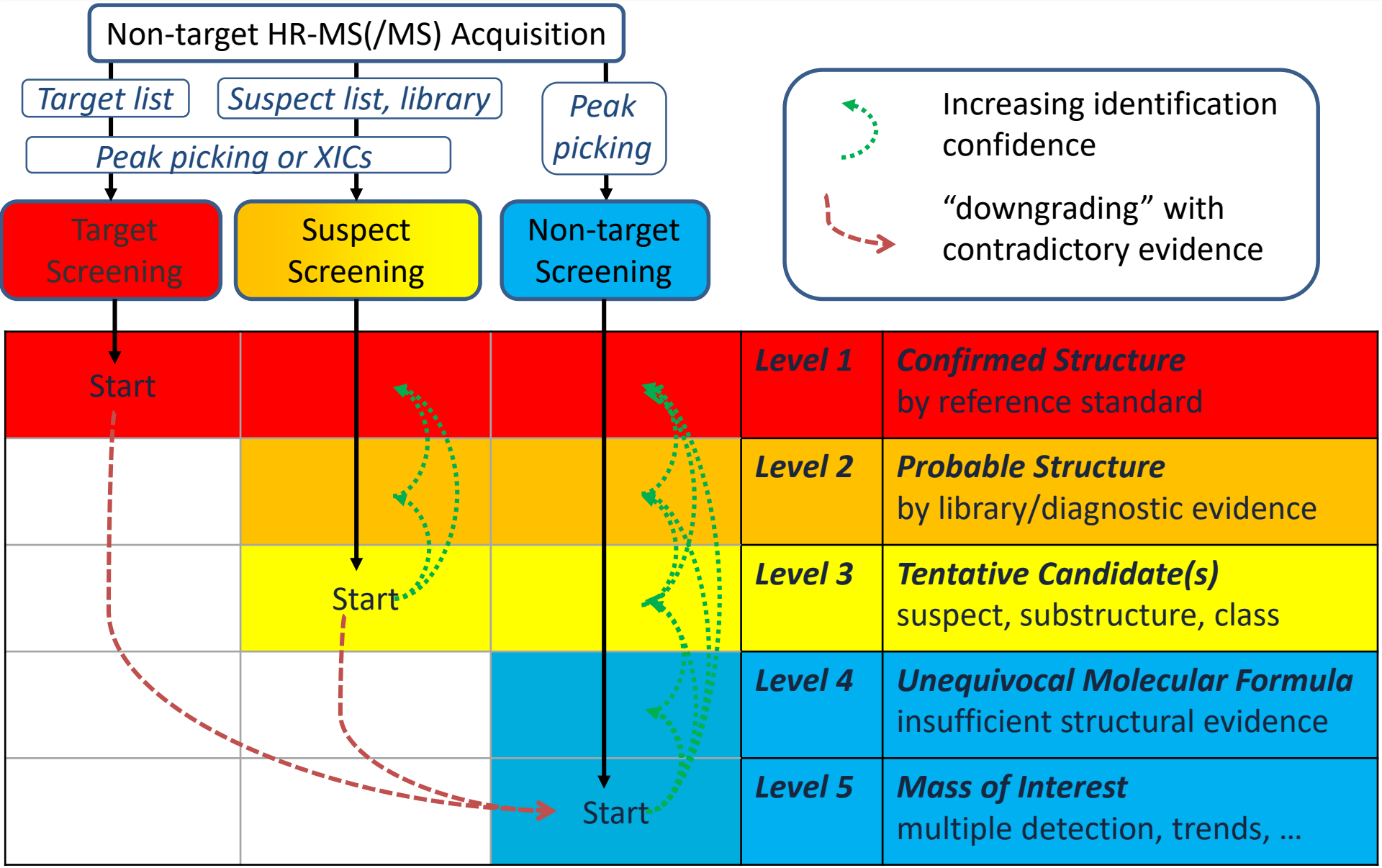
AND connecting chemical knowledge





Schymanski *et al*, DOI: 10.1021/es4044374; Schymanski & Williams, DOI: 10.5281/zenodo.3242674; Vermeulen *et al* DOI: 10.1126/science.aay3164

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

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!!!!

- Key examples mentioned today:
  - PubChem: https://pubchem.ncbi.nlm.nih.gov/
  - CompTox: https://comptox.epa.gov/dashboard/

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)

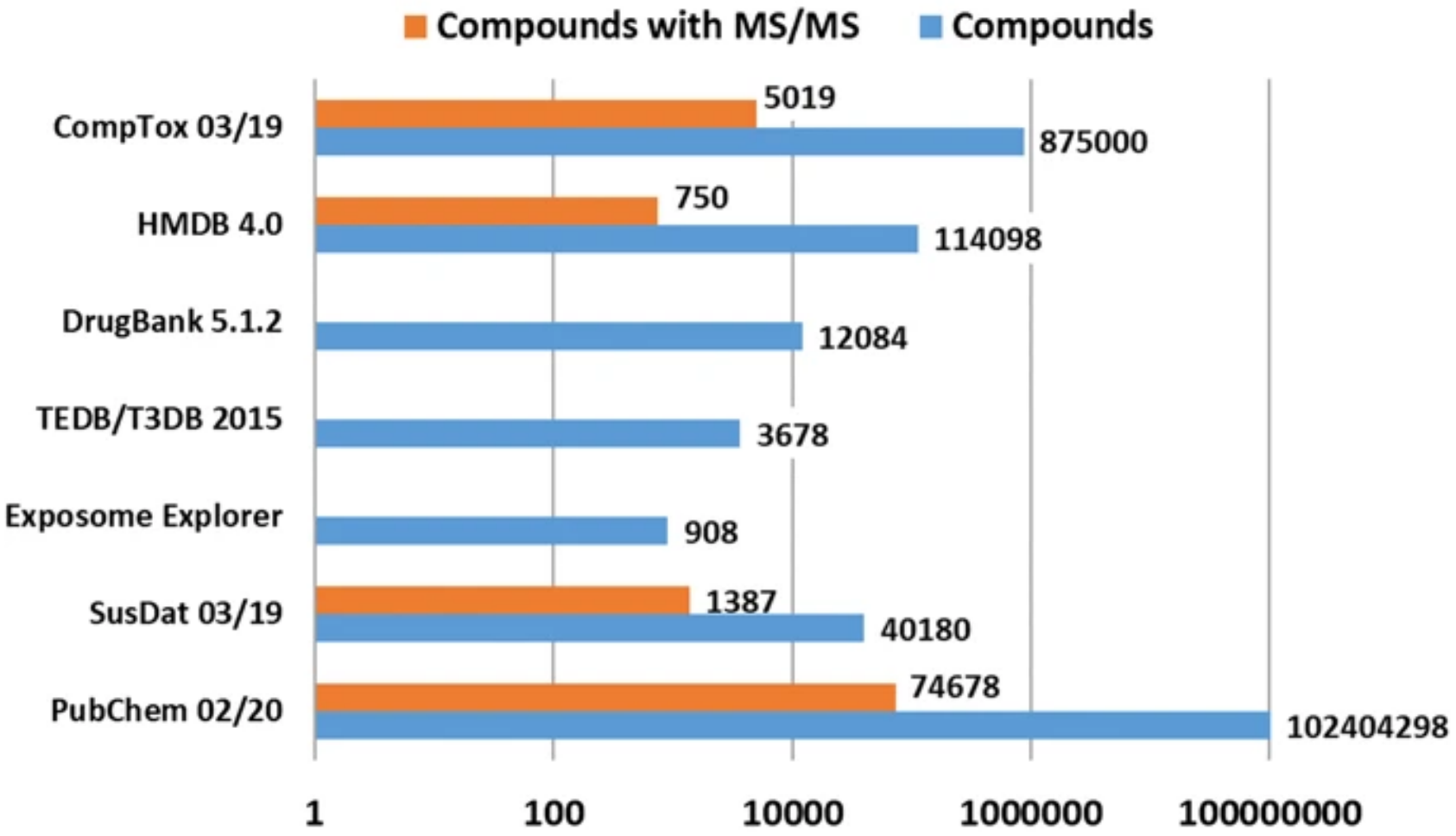
- Key resource(s) mentioned today:
  - MassBank EU: http://massbank.eu/MassBank

  - MoNA: https://mona.fiehnlab.ucdavis.edu/

Scarcity of MS/MS Spectra

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

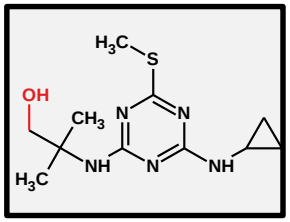
- MS/MS avail. for ~0.1-4 % of relevant exposomics resources



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

Confidence Levels for Tentative Structures

Example | *Identification confidence* | *Minimum data requirements*



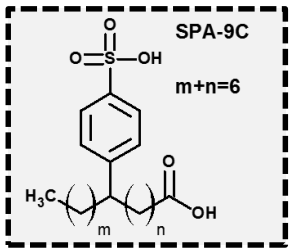
Level 1: Confirmed structure
by reference standard

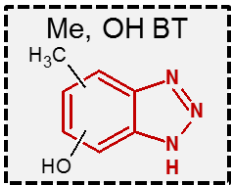
MS, MS^2, RT, **Reference Std.**

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

MS, MS^2, **Library MS^2 >0.9 sim**

MassBank.eu





Level 3: Tentative candidate(s)
structure, substituent, class

MS, MS^2, MetFrag

$C_6H_5N_3O_4$

Level 4: Unequivocal molecular formula

MS isotope/adduct

192.0757

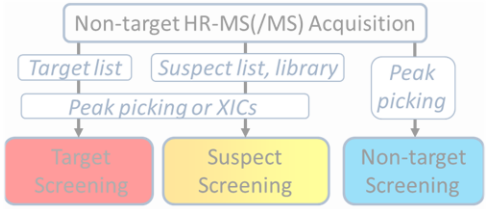
Level 5: Exact mass of interest

MS

Schymanski, Jeon, Gulde, Fenner, Ruff, Singer & Hollender (2014) ES&T, 48 (4), 2097-2098. DOI: 10.1021/es5002105

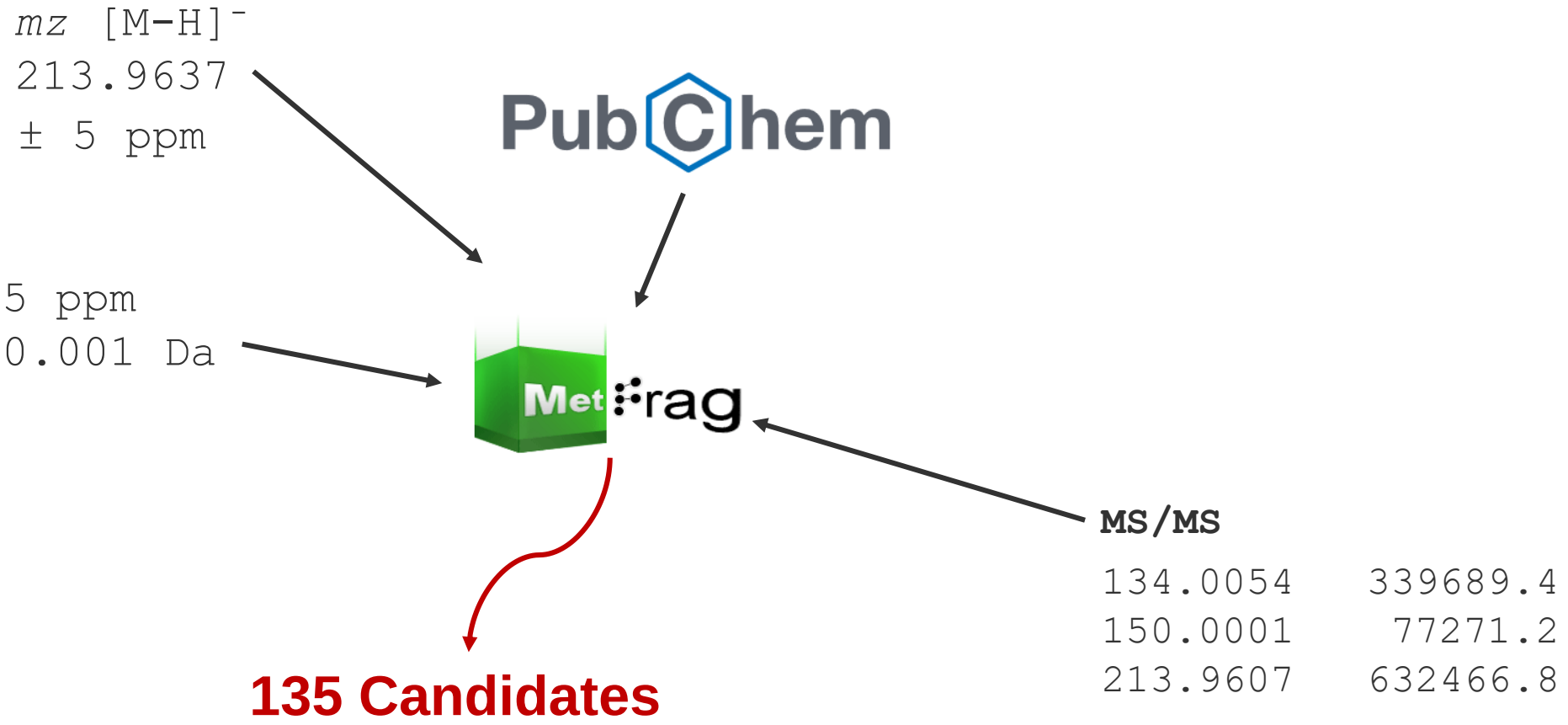
Plan for Today

- Background
  - Non-target Screening & Identification Confidence
  - Compound Databases and Spectral Libraries
- Introduction to MetFrag: MS/MS and Metadata
- MetFrag + CompTox + MassBank + Nicotine
  - MS-ready groupings … or not
- MetFrag + PubChem + MassBank + Isobars
  - The relevance of spectral matching!
- Future: MetFrag for High Throughput Exposomics
  - Build your own databases … PubChemLite



MetFrag: *In silico* non-target identification

Status: 2010



Wolf et al, 2010, *BMC Bioinf.* 11:148, DOI: 10.1186/1471-2105-11-148

MetFrag 2010 vs MetFrag Relaunched

Test set of 473 Eawag Target Substances

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

	MetFrag 2010	New MetFrag Fragments only	
ChemSpider[1]			
Top 1 Ranks	73	105	
% Top 1 Ranks	15 %	22 %	
PubChem[2]			
Top 1 Ranks	-	30	
% Top 1 Ranks		6 %	

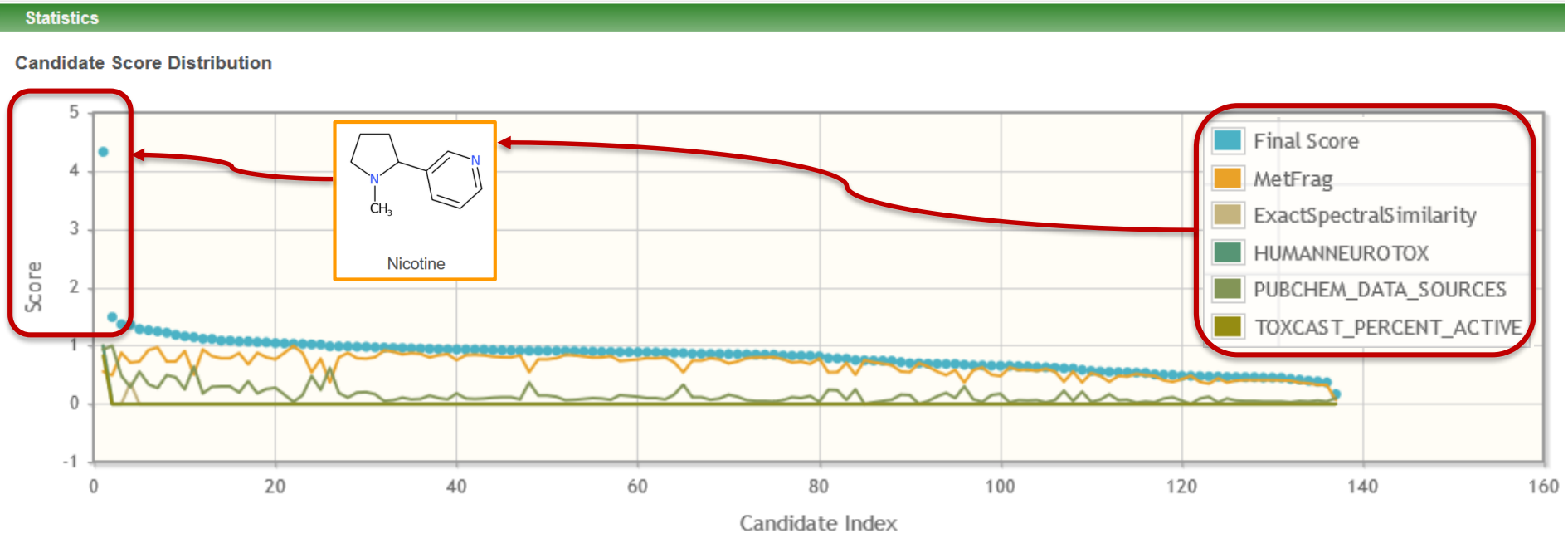
Still some room for improvement!

[1]www.chemspider.com; ~34 million entries
[2]https://pubchem.ncbi.nlm.nih.gov/; ~74 million entries

Ruttkies, Schymanski, Wolf, Hollender, Neumann, J. Chem. Inf., 2016, http://jcheminf.com/content/8/1/3

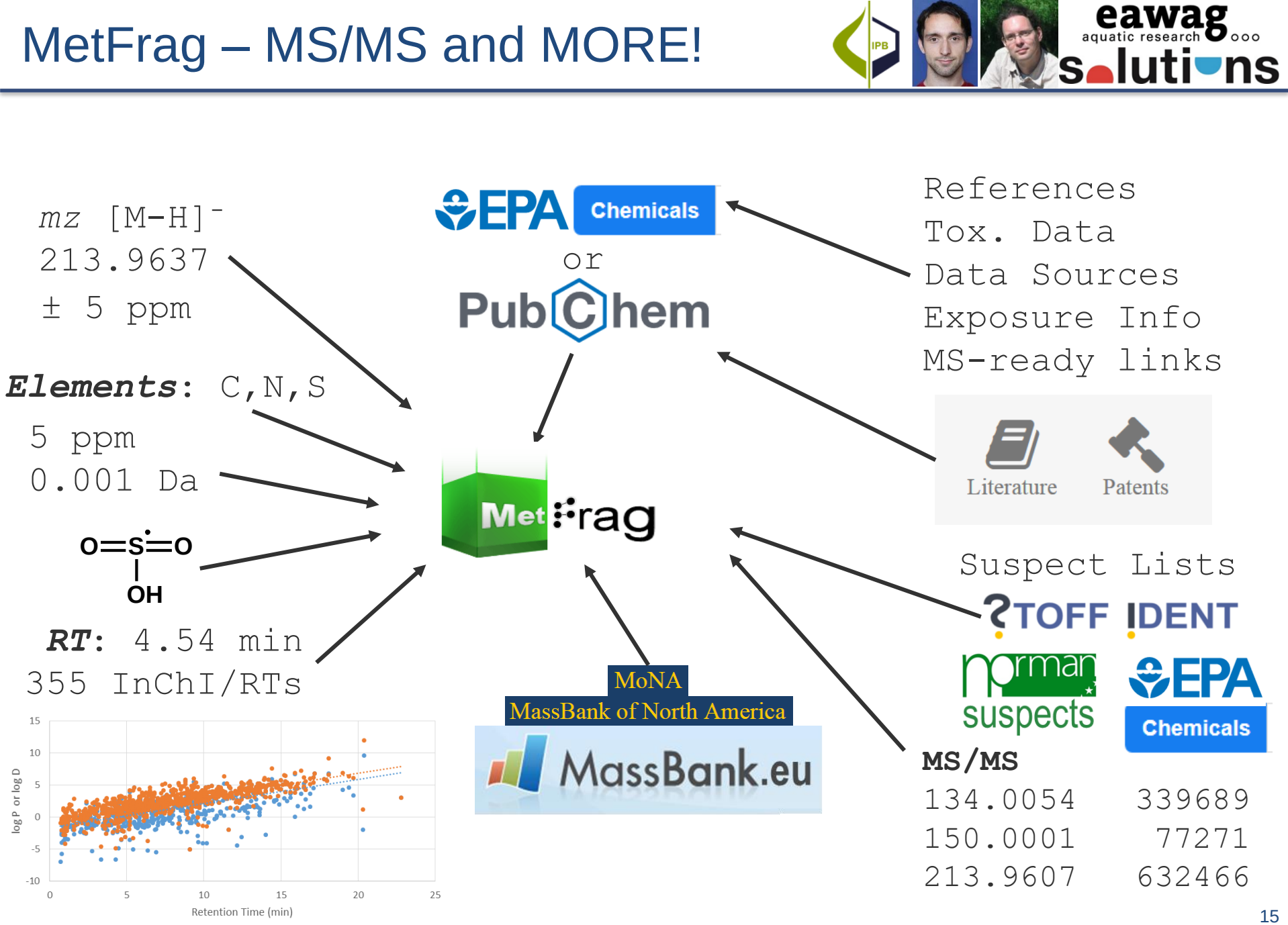
Connecting multiple lines of evidence for identification

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

MetFrag – MS/MS and MORE!



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

MetFrag2.2 & Non-target Identification

Test set of 473 Eawag Target Substances

	MetFrag 2010	New MetFrag Fragments only	***New MetFrag +References +Retention time***
ChemSpider[1]			
Top 1 Ranks	73	105	***420***
% Top 1 Ranks	15 %	22 %	***89 %***
PubChem[2]			
Top 1 Ranks	-	30	***336***
% Top 1 Ranks	-	6 %	***71 %***

Similar results with 3 independent datasets of 310, 289 and 225 substances from Eawag and UFZ (www.massbank.eu)

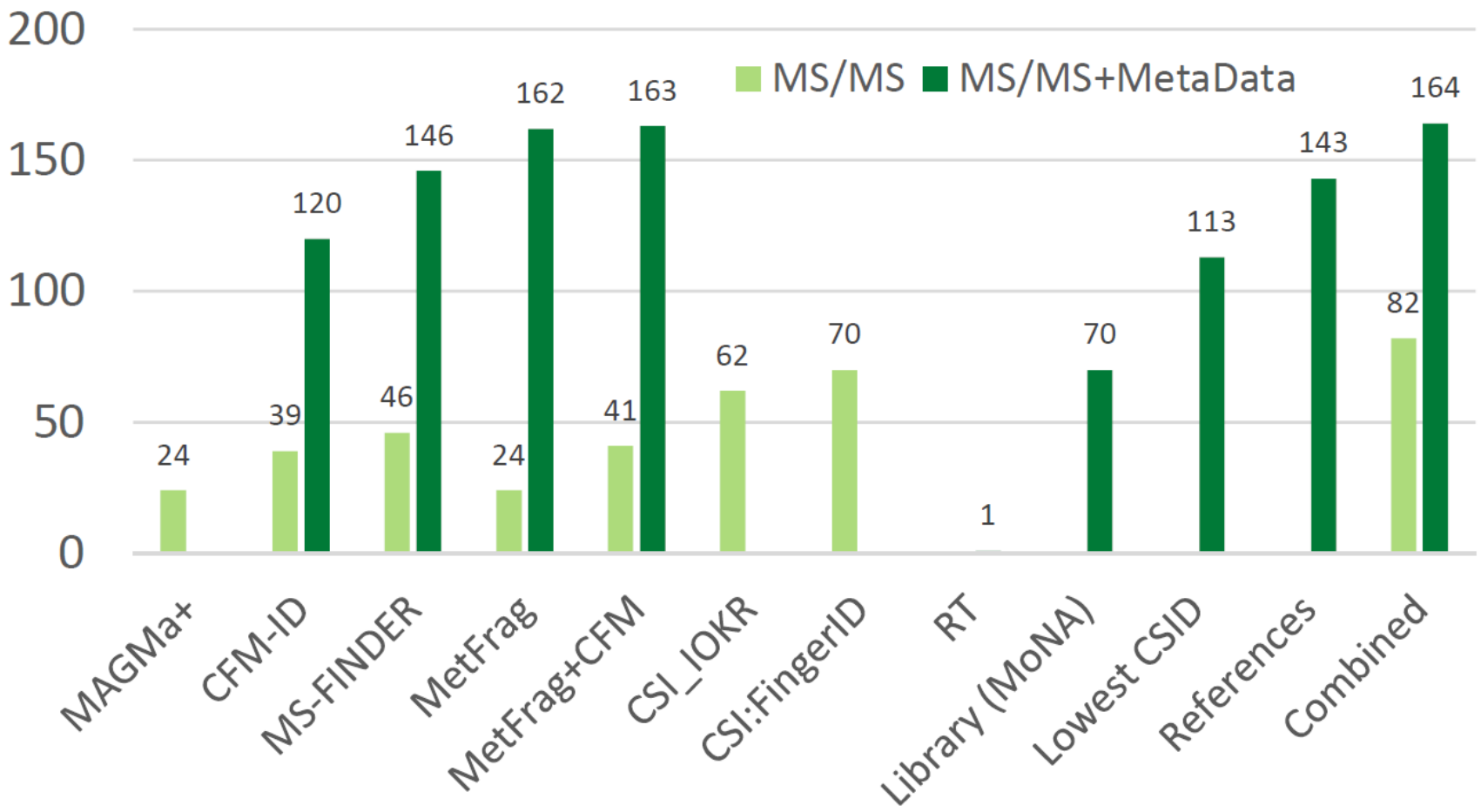
[1]www.chemspider.com; ~34 million entries
[2]https://pubchem.ncbi.nlm.nih.gov/; ~74 million entries

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

Ruttkies, Schymanski, Wolf, Hollender, Neumann, *J. Chem. Inf.*, 2016, http://jcheminf.com/content/8/1/3

State of the Art in Small Molecule Identification

Metadata is critical to improving annotation of known unknowns!

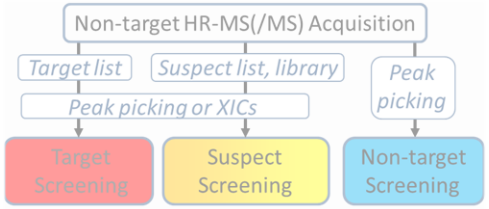


Schymanski *et al*, 2017, J Cheminf., DOI: 10.1186/s13321-017-0207-1

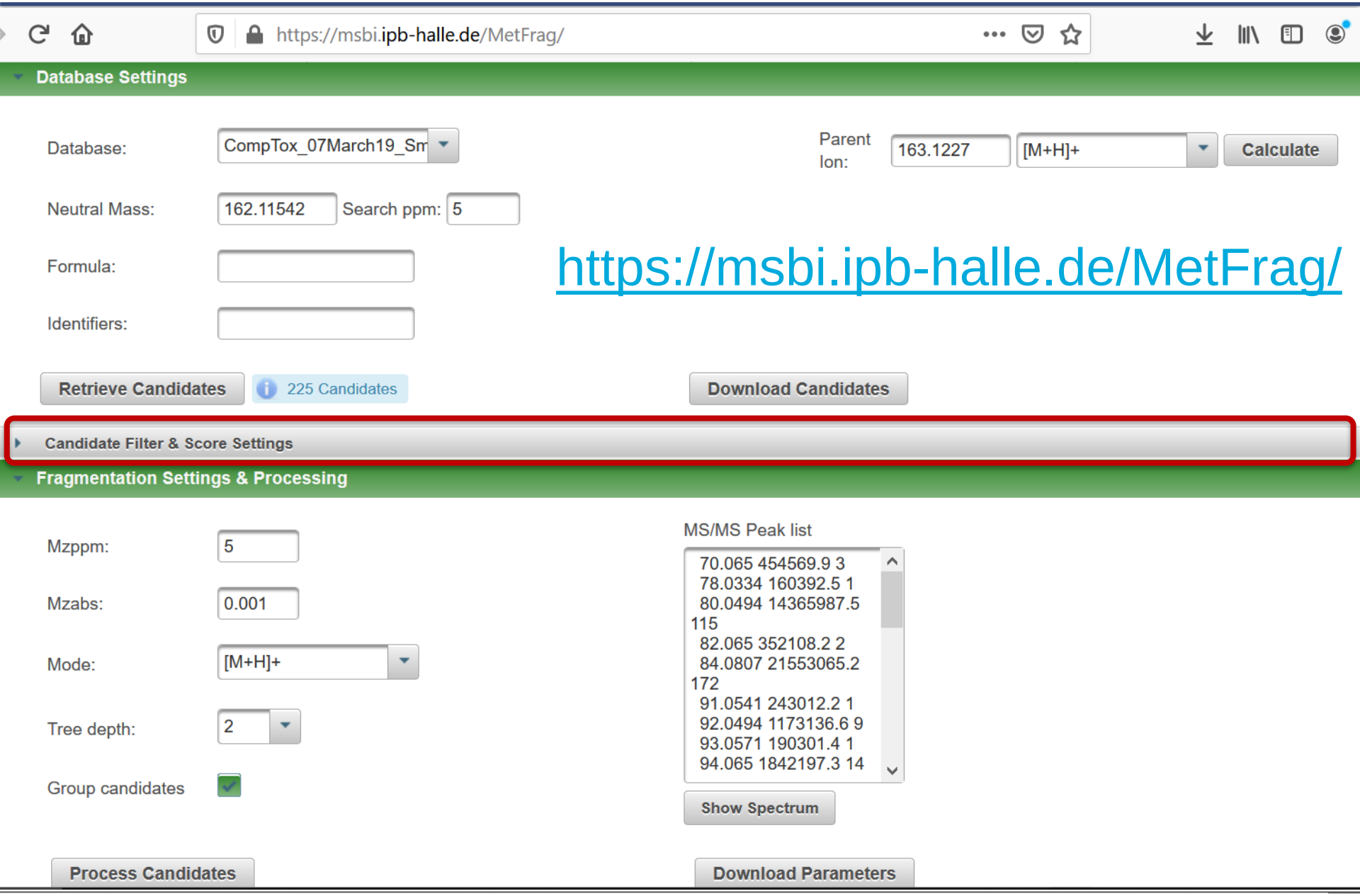
www.casmi-contest.org

Plan for Today

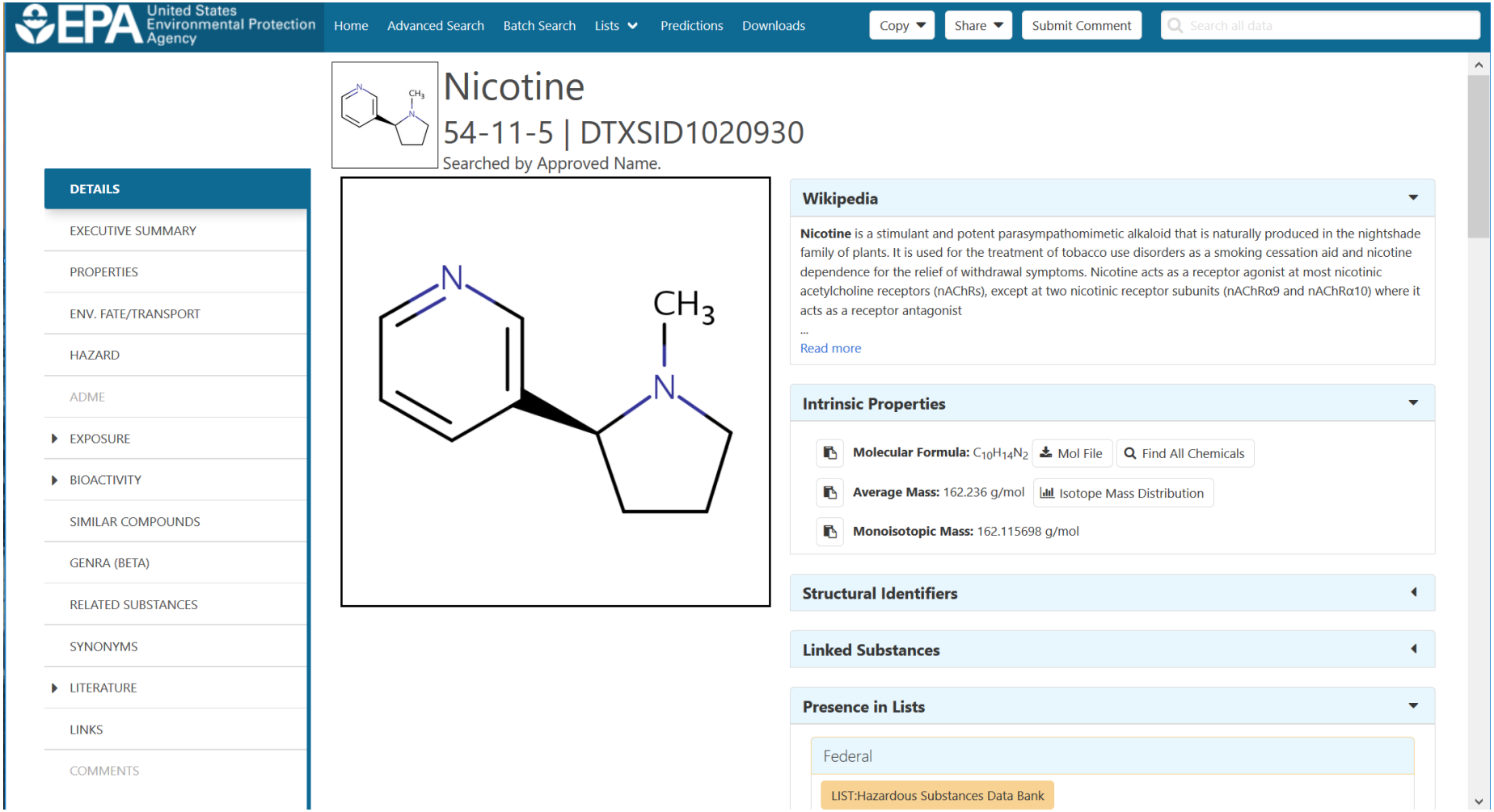
- Background
  - Non-target Screening & Identification Confidence
  - Compound Databases and Spectral Libraries
- Introduction to MetFrag: MS/MS and Metadata
- MetFrag + CompTox + MassBank + Nicotine
  - MS-ready groupings … or not
- MetFrag + PubChem + MassBank + Isobars
  - The relevance of spectral matching!
- Future: MetFrag for High Throughput Exposomics
  - Build your own databases … PubChemLite



MetFrag + CompTox + MassBank + Nicotine



MetFrag + CompTox + MassBank + Nicotine



https://comptox.epa.gov/dashboard/dsstoxdb/results?search=nicotine

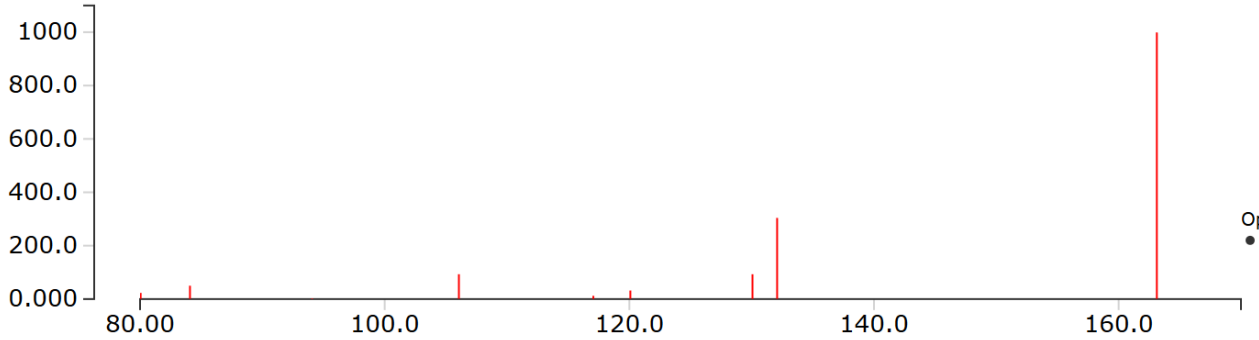
MetFrag + CompTox + MassBank + Nicotine

MassBank Record: EQ300801

Home | Search | Record Index | Data Privacy | Imprint | MassBank ID: Go

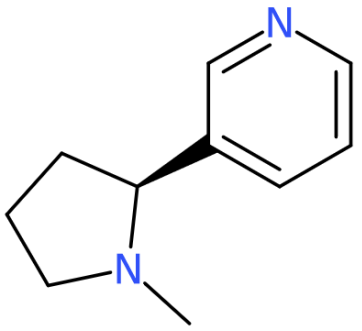
Nicotine; LC-ESI-QFT; MS2; CE: 15; R=35000; [M+H]+

Mass Spectrum



Chemical Structure

Options
● Labels



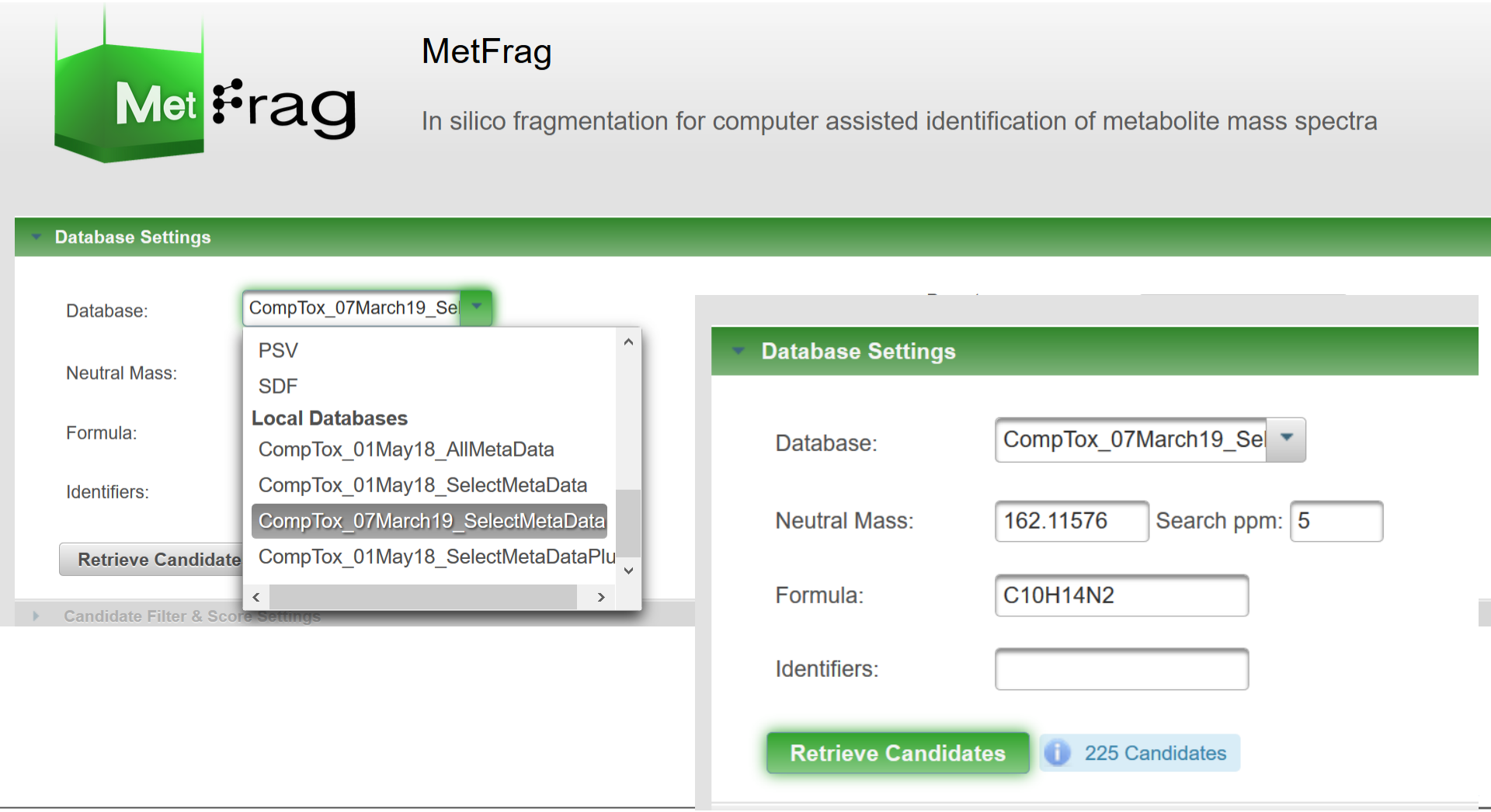
ACCESSION: EQ300801
RECORD_TITLE:
Nicotine; LC-ESI-QFT; MS2; CE: 15; R=35000; [M+H]+

```
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
//
```

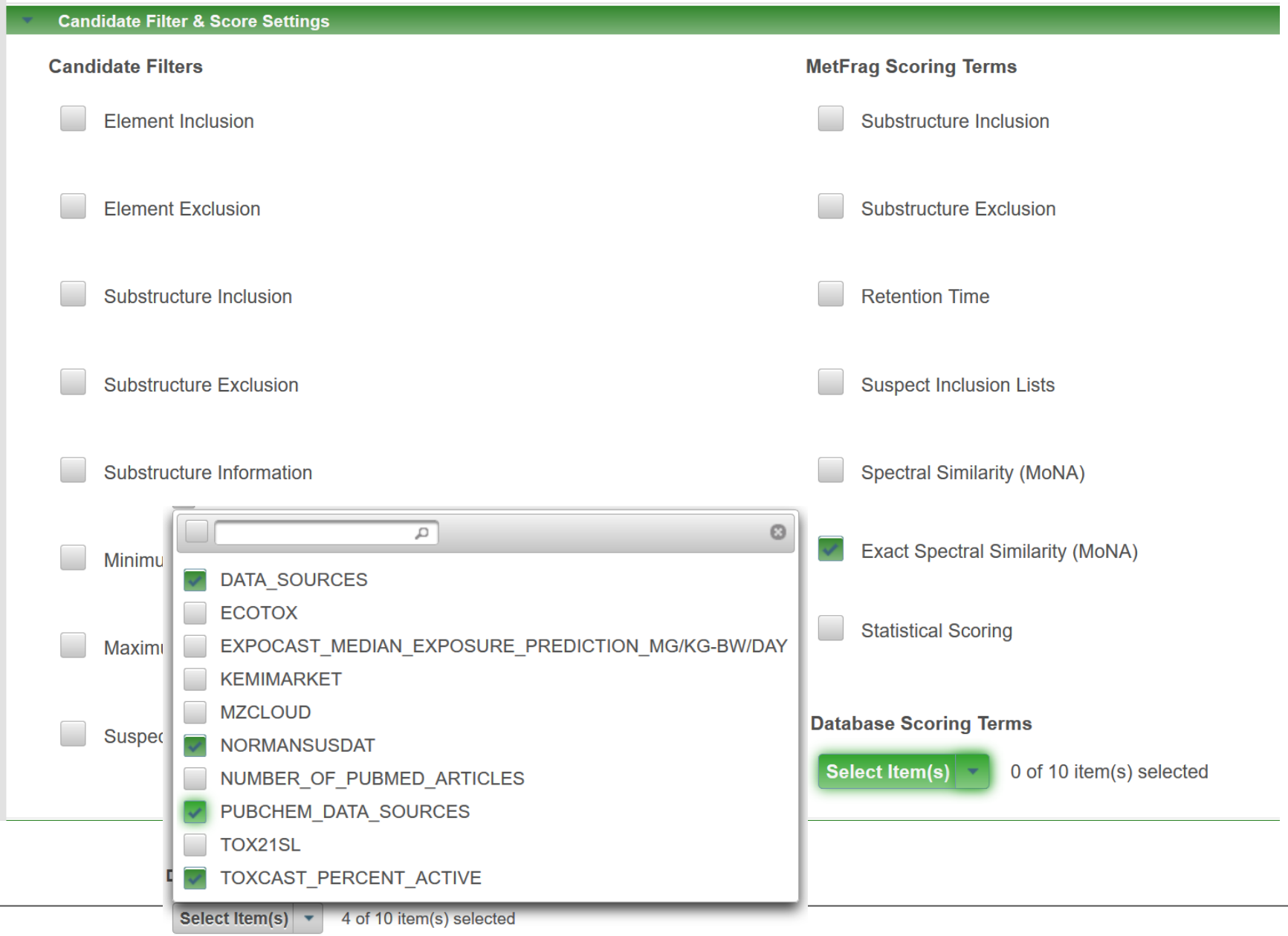

https://massbank.eu/MassBank/RecordDisplay.jsp?id=EQ300801&dsn=Eawag

MetFrag + CompTox + MassBank + Nicotine

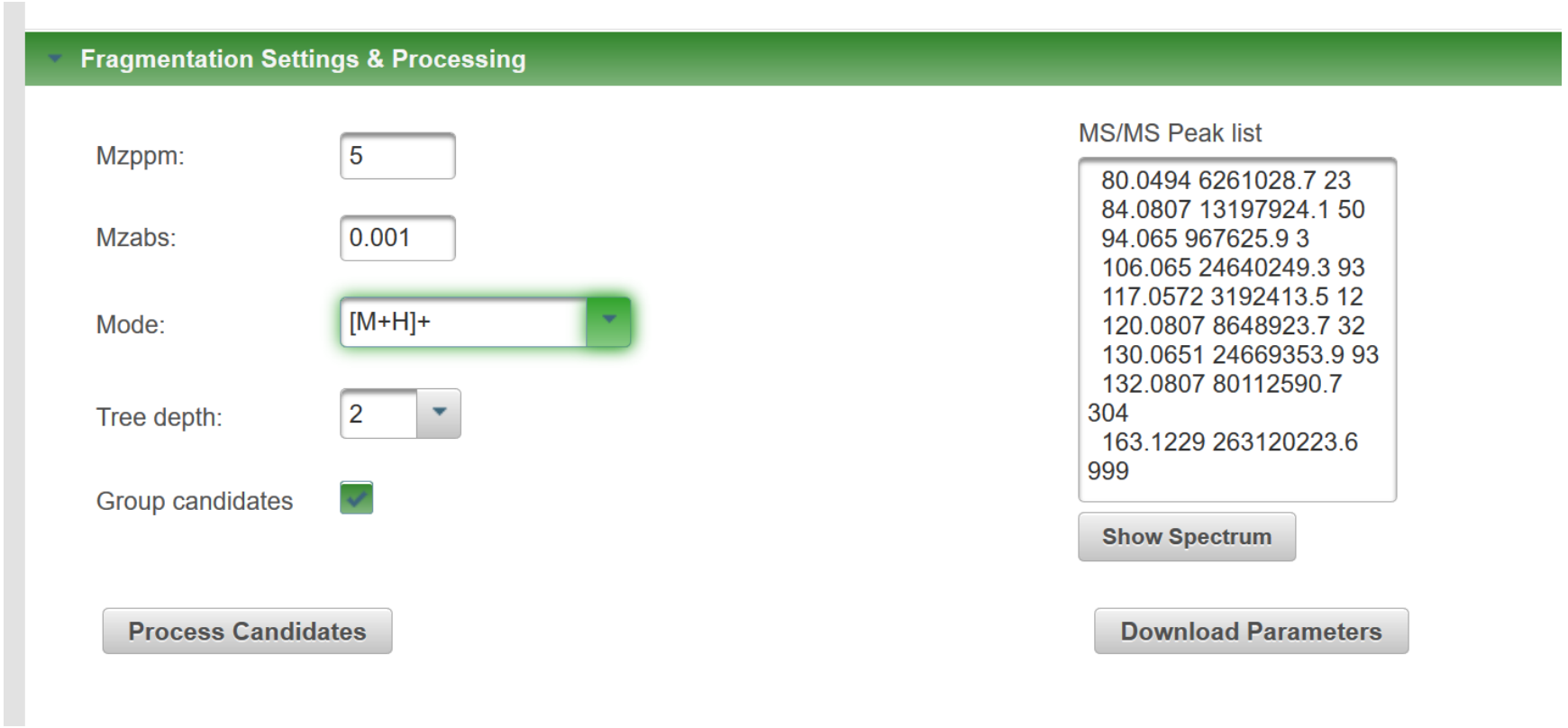
- https://msbi.ipb-halle.de/MetFrag/
- CompTox is integrated as a “Local Database”



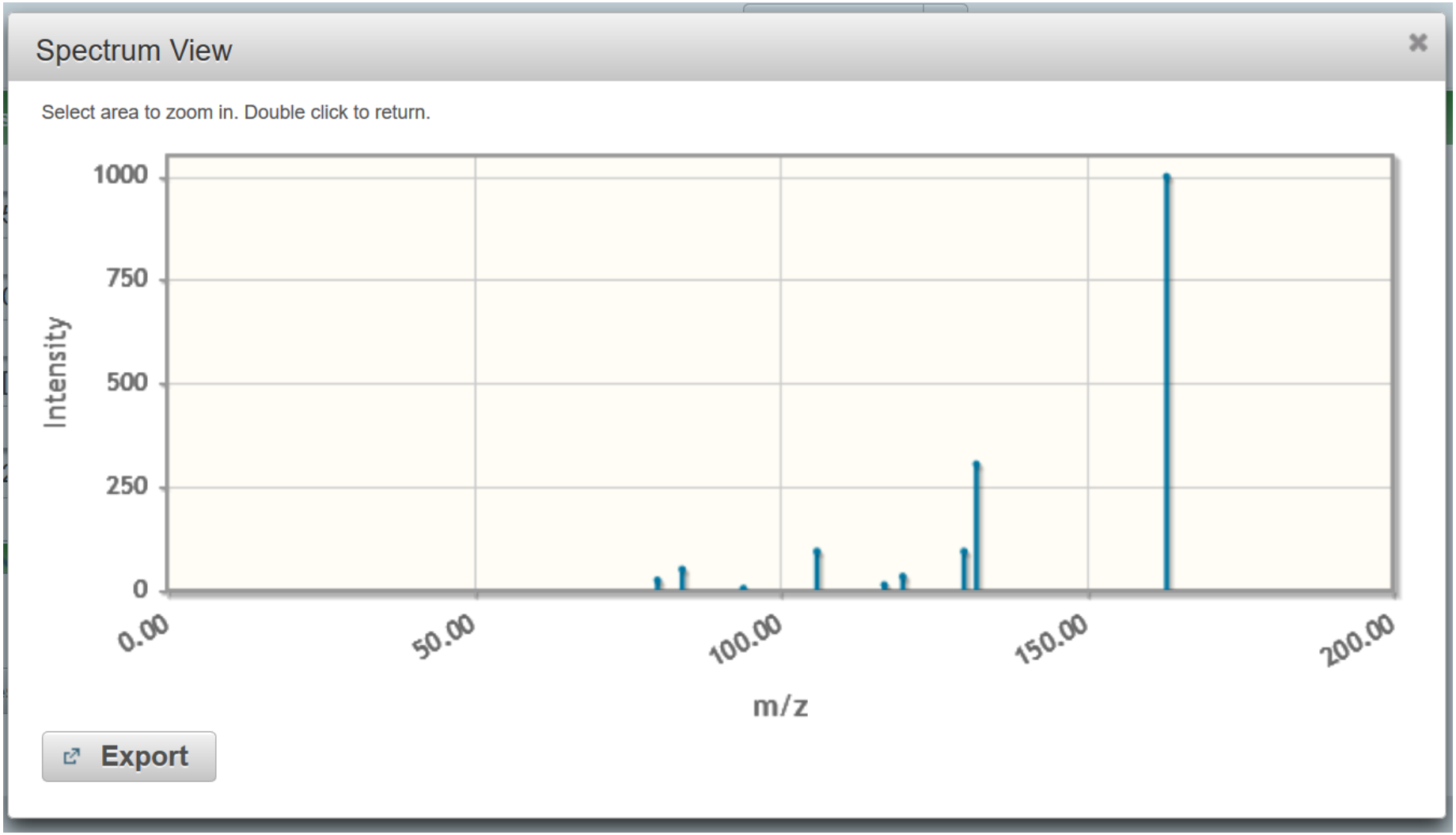
MetaData is included in CompTox



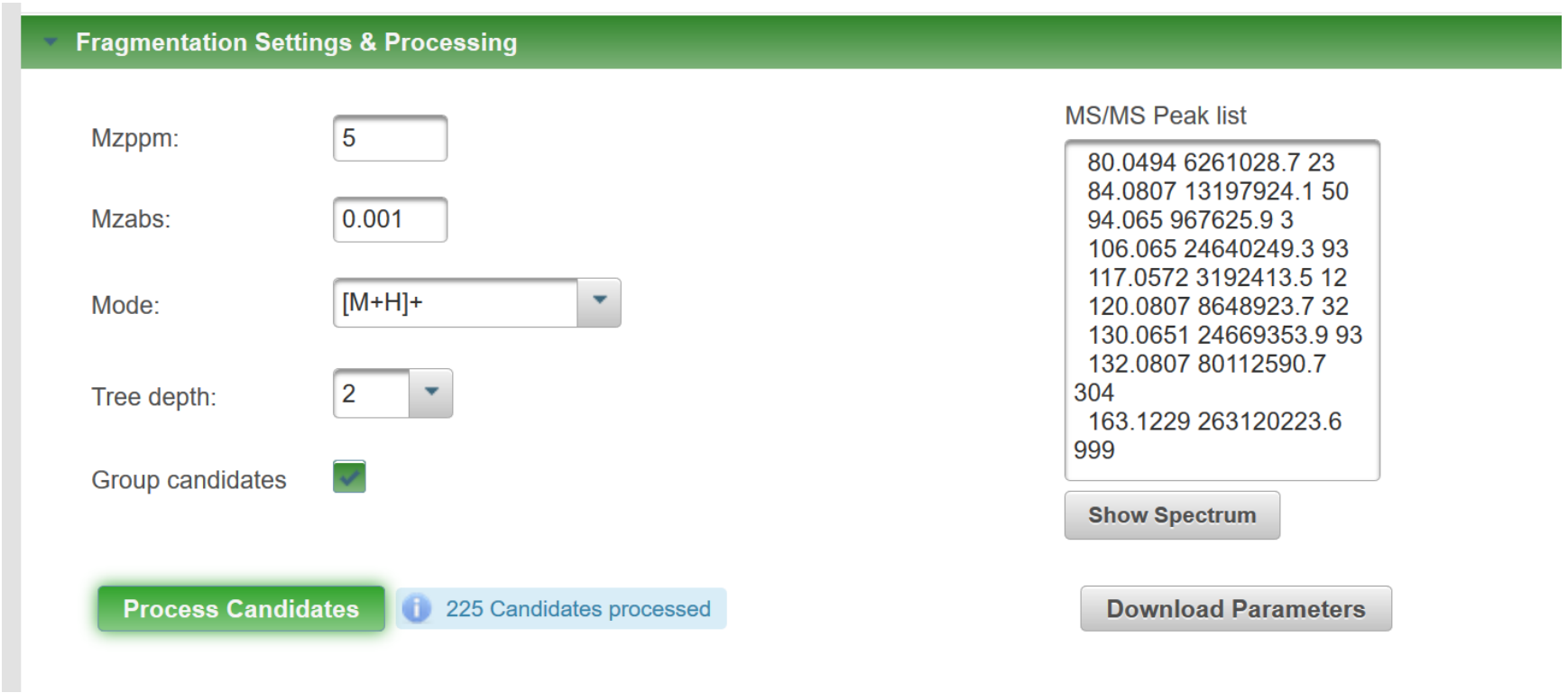
Include MS/MS and mode from MassBank Record



Check Spectrum



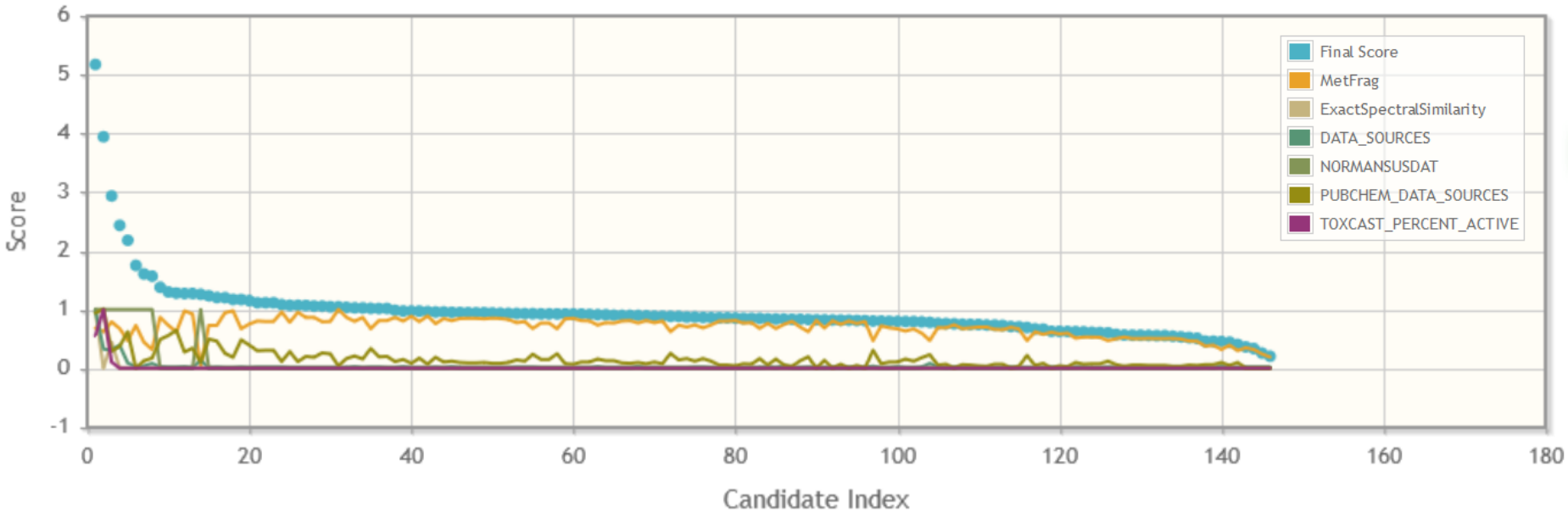
Process Candidates (Grouped)



Process Candidates (Grouped)

Statistics

Candidate Score Distribution



Export

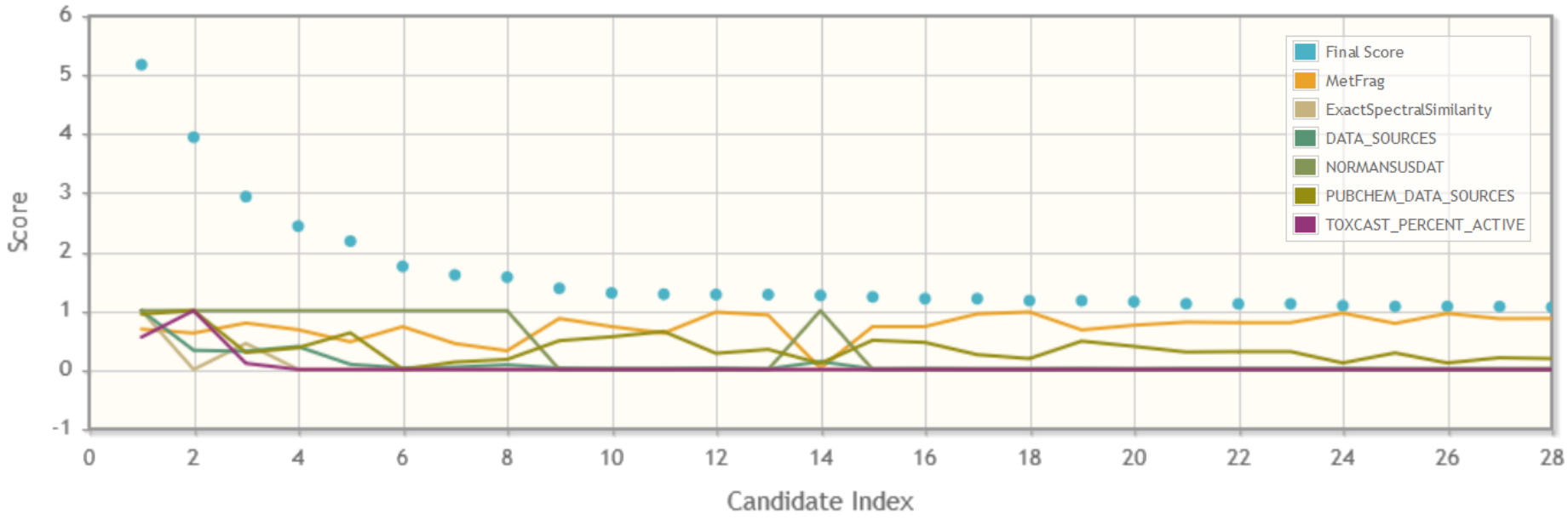
Select area to zoom in. Double click to return.

Click on dot to scroll to candidate in the Results tab.

Process Candidates (Grouped) – Zoomed In

Statistics

Candidate Score Distribution



Export

Select area to zoom in. Double click to return.
Click on dot to scroll to candidate in the Results tab.

Results Overview (Grouped)

Results

Weights		
MetFrag (1st)	100	%
ExactSpectralSimilarity (2nd)	100	%
DATA_SOURCES (3rd)	100	%
NORMANSUSDAT (4th)	100	%
PUBCHEM_DATA_SOURCES (5th)	100	%
TOXCAST_PERCENT_ACTIVE (6th)	100	%

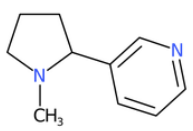
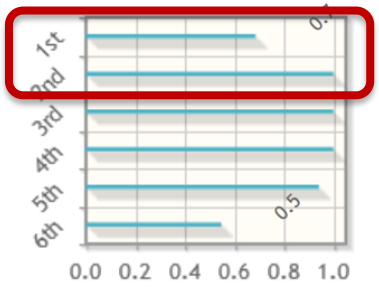
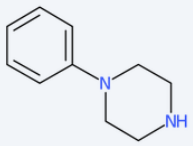
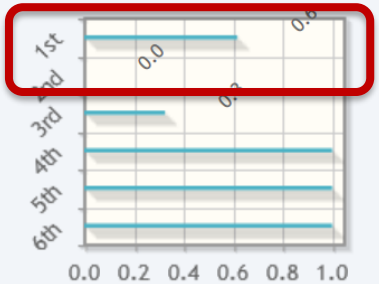
Experimental evidence / values

Download Results

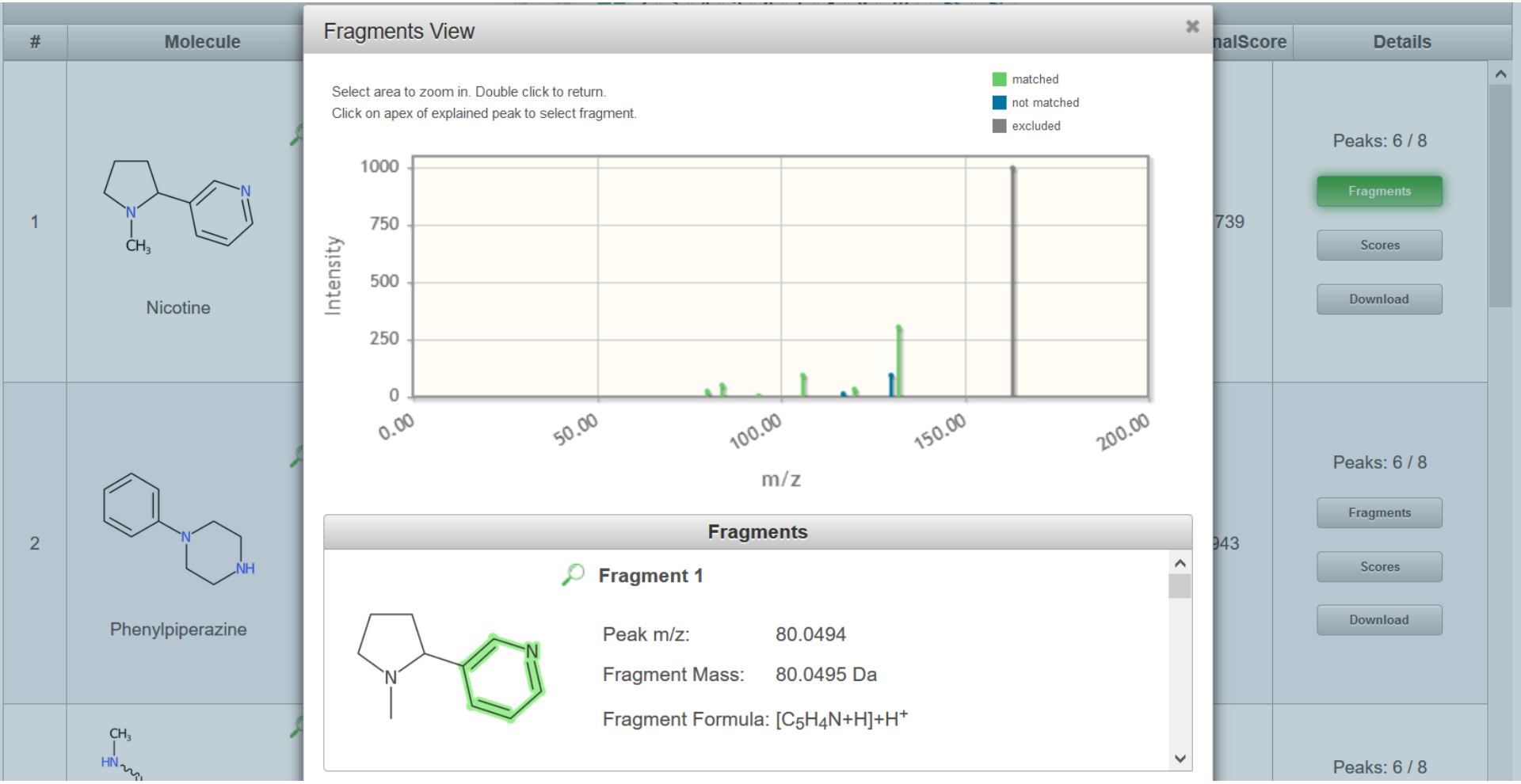
Filter Candidates by explained MS/MS Peaks

MS/MS Peaks | Filter Candidates

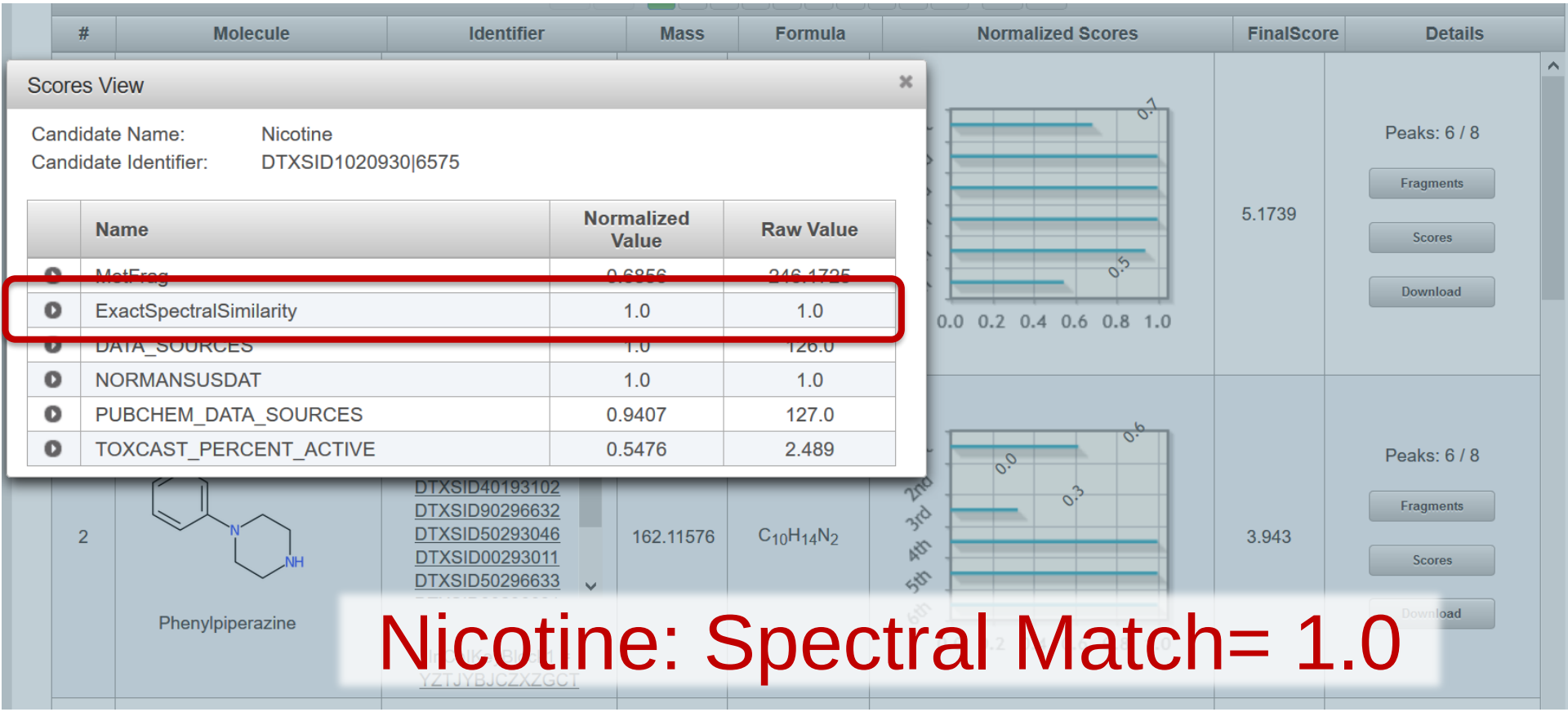
1 2 3 4 5 6 7 8 9 10

#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1	 Nicotine	DTXSID1020930 DTXSID3048154 DTXSID20237235 DTXSID8021725 DTXSID0046351 DTXSID6020931 DTXSID80442666 InChIKeyBlock1 = SNICXCGAKADSCV	162.11576	$C_{10}H_{14}N_2$		5.1739	Peaks: 6 / 8 Fragments Scores Download
2	 Phenylpiperazine	DTXSID8057855 DTXSID40176612 DTXSID40193102 DTXSID90296632 DTXSID50293046 DTXSID00293011 DTXSID50296633 InChIKeyBlock1 = YZTJYBJCZXZGCT	162.11576	$C_{10}H_{14}N_2$		3.943	Peaks: 6 / 8 Fragments Scores Download

Fragments View



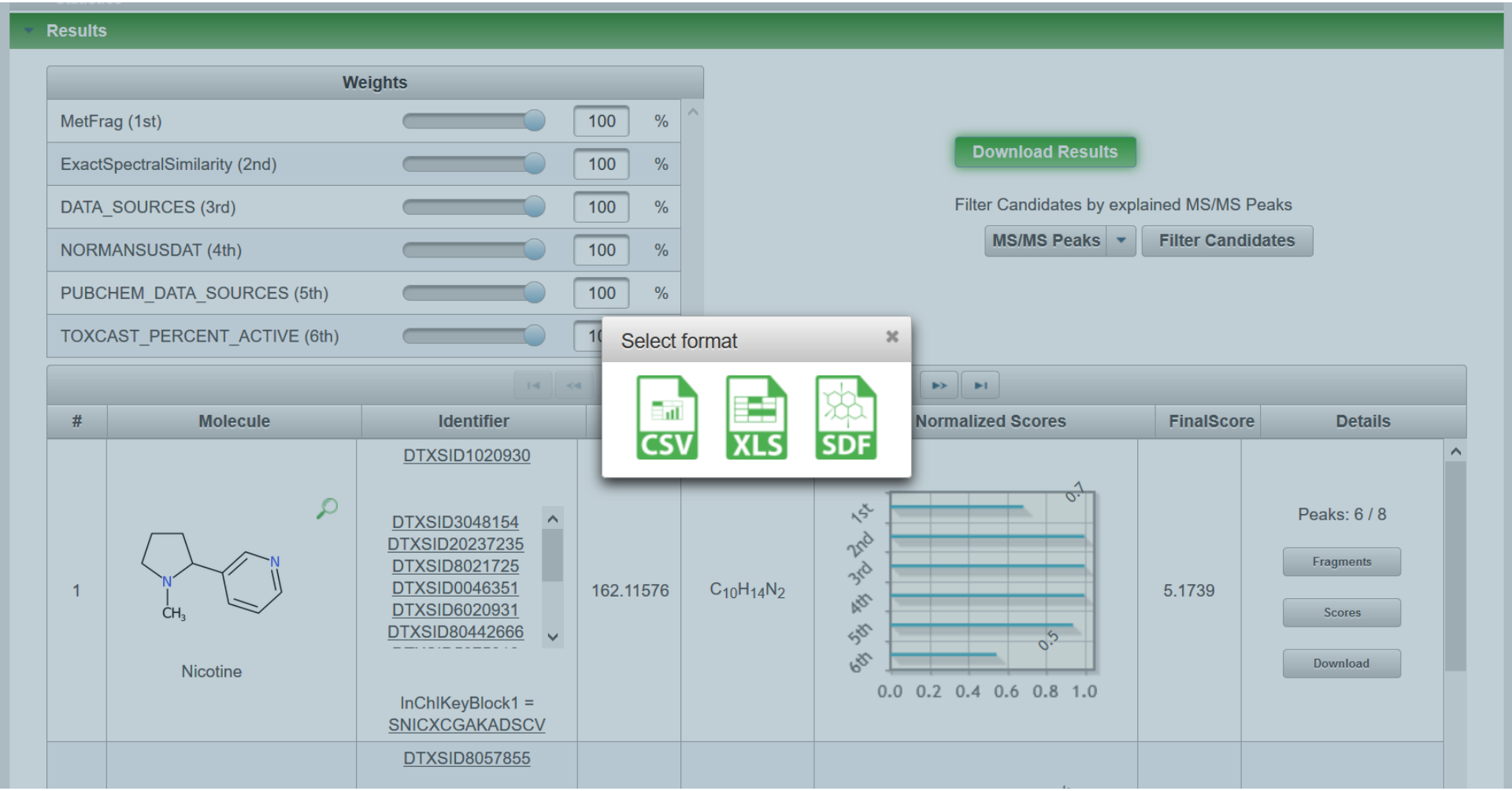
Scores View



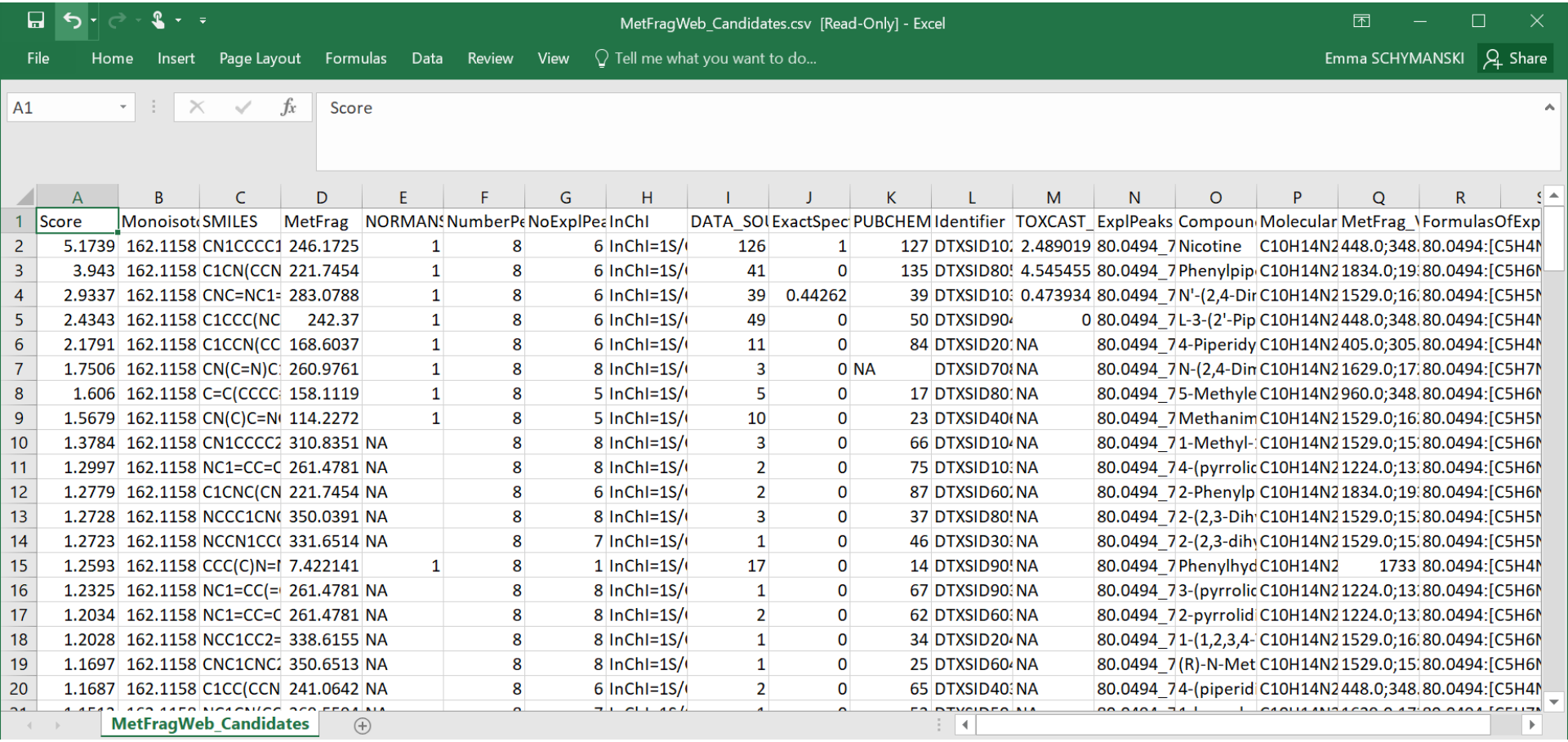
Nicotine: Spectral Match= 1.0

Level 2a

Download Results



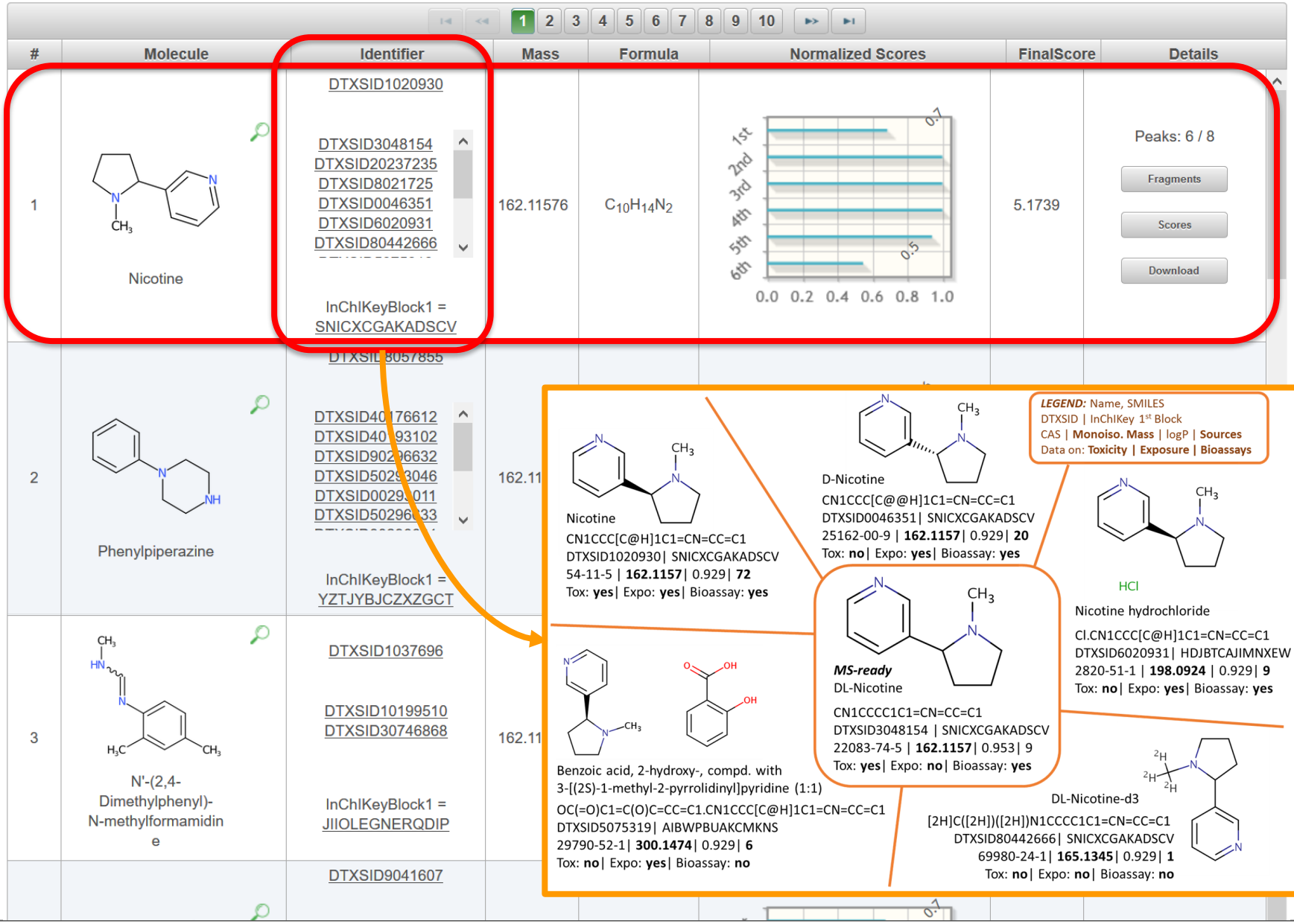
Download Results



Score	Monoisot	SMILES	MetFrag	NORMANS	NumberPe	NoExplPea	InChI	DATA_SO	ExactSpec	PUBCHEM	Identifier	TOXCAST_	ExplPeaks	Compoun	Molecular	MetFrag_	FormulasOfExp
5.1739	162.1158	CN1CCCC1	246.1725	1	8	6	InChI=1S/	126	1	127	DTXSID10	2.489019	80.0494_7	Nicotine	C10H14N2	448.0;348.	80.0494:[C5H4N
3.943	162.1158	C1CN(CCN	221.7454	1	8	6	InChI=1S/	41	0	135	DTXSID80	4.545455	80.0494_7	Phenylpip	C10H14N2	1834.0;19	80.0494:[C5H6N
2.9337	162.1158	CNC=NC1=	283.0788	1	8	6	InChI=1S/	39	0.44262	39	DTXSID10	0.473934	80.0494_7	N'-(2,4-Dir	C10H14N2	1529.0;16	80.0494:[C5H5N
2.4343	162.1158	C1CCC(NC	242.37	1	8	6	InChI=1S/	49	0	50	DTXSID90	0	80.0494_7	L-3-(2'-Pip	C10H14N2	448.0;348.	80.0494:[C5H4N
2.1791	162.1158	C1CCN(CC	168.6037	1	8	6	InChI=1S/	11	0	84	DTXSID20	NA	80.0494_7	4-Piperidy	C10H14N2	405.0;305.	80.0494:[C5H4N
1.7506	162.1158	CN(C=N)C	260.9761	1	8	8	InChI=1S/	3	0	NA	DTXSID70	NA	80.0494_7	N-(2,4-Dir	C10H14N2	1629.0;17	80.0494:[C5H7N
1.606	162.1158	C=C(CCCC	158.1119	1	8	5	InChI=1S/	5	0	17	DTXSID80	NA	80.0494_7	5-Methyle	C10H14N2	960.0;348.	80.0494:[C5H6N
1.5679	162.1158	CN(C)C=N(	114.2272	1	8	5	InChI=1S/	10	0	23	DTXSID40	NA	80.0494_7	Methanim	C10H14N2	1529.0;16	80.0494:[C5H5N
1.3784	162.1158	CN1CCCC2	310.8351	NA	8	8	InChI=1S/	3	0	66	DTXSID10	NA	80.0494_7	1-Methyl-	C10H14N2	1529.0;15	80.0494:[C5H6N
1.2997	162.1158	NC1=CC=C	261.4781	NA	8	8	InChI=1S/	2	0	75	DTXSID10	NA	80.0494_7	4-(pyrrolic	C10H14N2	1224.0;13	80.0494:[C5H6N
1.2779	162.1158	C1CNC(CN	221.7454	NA	8	6	InChI=1S/	2	0	87	DTXSID60	NA	80.0494_7	2-Phenylp	C10H14N2	1834.0;19	80.0494:[C5H6N
1.2728	162.1158	NCCC1CN(	350.0391	NA	8	8	InChI=1S/	3	0	37	DTXSID80	NA	80.0494_7	2-(2,3-Dih	C10H14N2	1529.0;15	80.0494:[C5H5N
1.2723	162.1158	NCCN1CC(	331.6514	NA	8	7	InChI=1S/	1	0	46	DTXSID30	NA	80.0494_7	2-(2,3-dih	C10H14N2	1529.0;15	80.0494:[C5H5N
1.2593	162.1158	CCC(C)N=I	7.422141	1	8	1	InChI=1S/	17	0	14	DTXSID90	NA	80.0494_7	Phenylhyd	C10H14N2	1733	80.0494:[C5H4N
1.2325	162.1158	NC1=CC(=	261.4781	NA	8	8	InChI=1S/	1	0	67	DTXSID90	NA	80.0494_7	3-(pyrrolic	C10H14N2	1224.0;13	80.0494:[C5H6N
1.2034	162.1158	NC1=CC=C	261.4781	NA	8	8	InChI=1S/	2	0	62	DTXSID60	NA	80.0494_7	2-pyrrolid	C10H14N2	1224.0;13	80.0494:[C5H6N
1.2028	162.1158	NCC1CC2=	338.6155	NA	8	8	InChI=1S/	1	0	34	DTXSID20	NA	80.0494_7	1-(1,2,3,4-	C10H14N2	1529.0;16	80.0494:[C5H6N
1.1697	162.1158	CNC1CNC2	350.6513	NA	8	8	InChI=1S/	1	0	25	DTXSID60	NA	80.0494_7	(R)-N-Met	C10H14N2	1529.0;15	80.0494:[C5H6N
1.1687	162.1158	C1CC(CCN	241.0642	NA	8	6	InChI=1S/	2	0	65	DTXSID40	NA	80.0494_7	4-(piperid	C10H14N2	448.0;348.	80.0494:[C5H4N

You can try this entire example offline – see Example 1 in documentation

“MS-ready” Form: Grouped vs Ungrouped…



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

Process Candidates (Ungrouped)

Fragmentation Settings & Processing

Mzppm: 5

Mzabs: 0.001

Mode: [M+H]+

Tree depth: 2

Group candidates

MS/MS Peak list

```
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
```

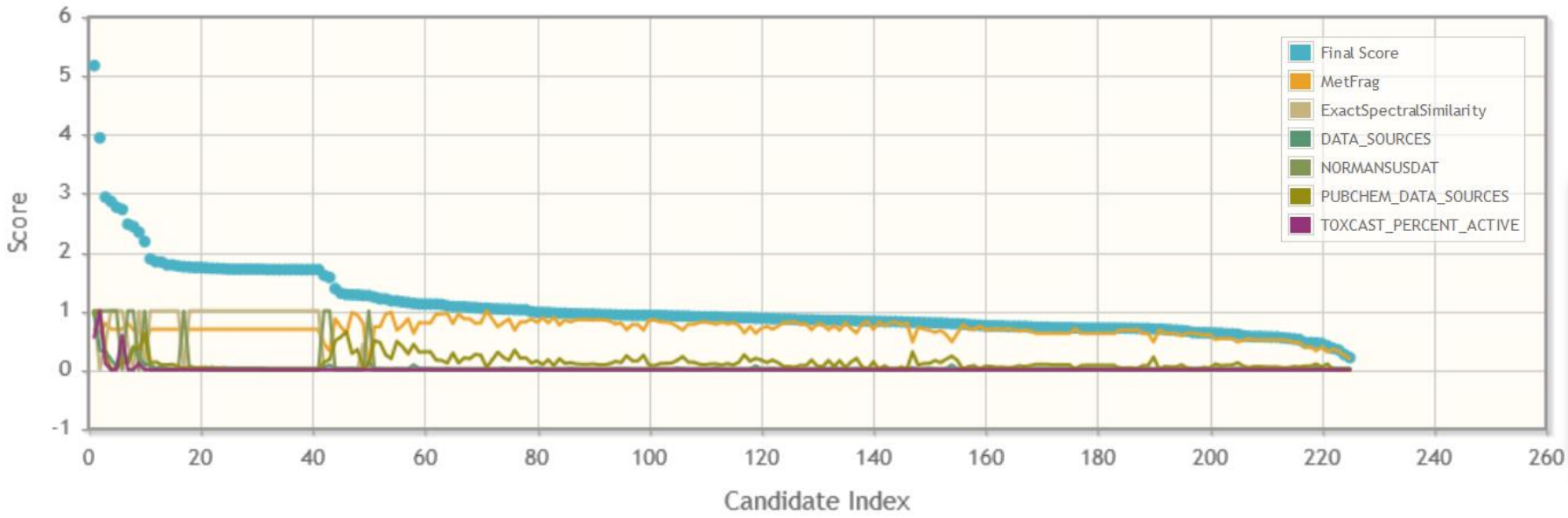
Show Spectrum

Process Candidates | 225 Candidates processed

Download Parameters

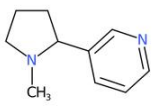
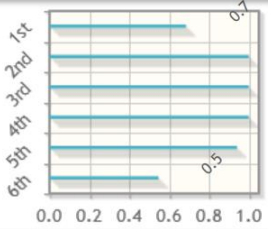
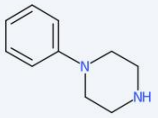
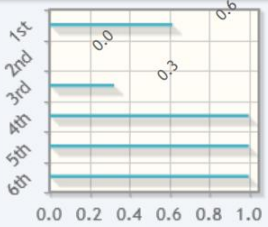
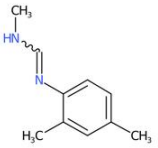
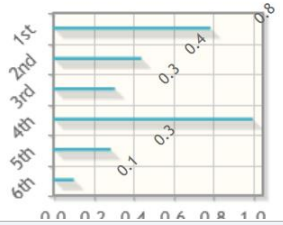
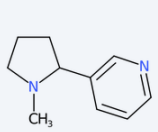
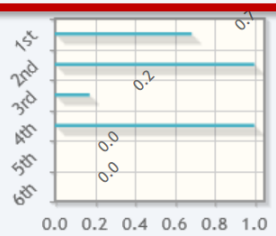
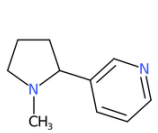
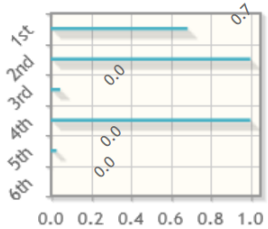
Statistics

Candidate Score Distribution



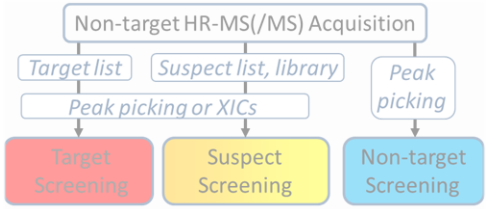
Process Candidates (Ungrouped)

1 2 3 4 5 6 7 8 9 10

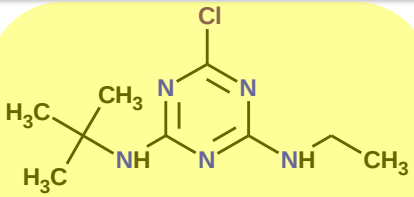
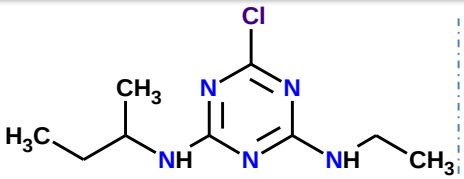
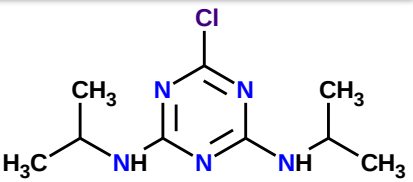
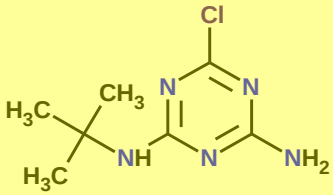
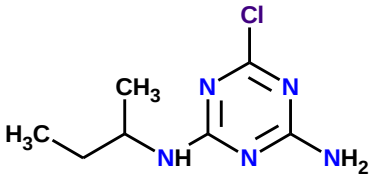
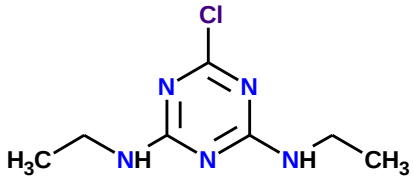
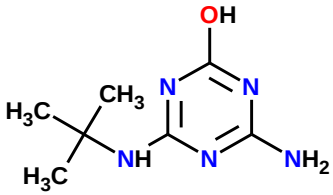
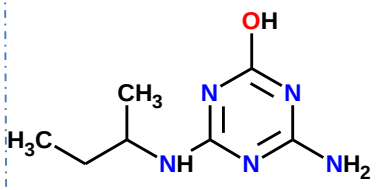
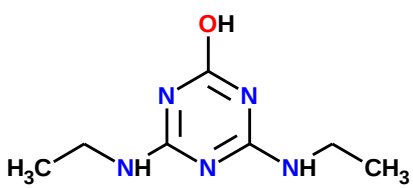
#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1	 Nicotine	DTXSID1020930 InChIKeyBlock1 = SNICXCGAKADSCV	162.11576	$C_{10}H_{14}N_2$		5.1739	Peaks: 6 / 8 Fragments Scores Download
2	 Phenylpiperazine	DTXSID8057855 InChIKeyBlock1 = YZTJYBJCZXZGCT	162.11576	$C_{10}H_{14}N_2$		3.943	Peaks: 6 / 8 Fragments Scores Download
3	 N'-(2,4-Dimethylphenyl)-N-methylformamidin	DTXSID1037696 InChIKeyBlock1 = JIIOLEGNERQDIP	162.11576	$C_{10}H_{14}N_2$		2.9337	Peaks: 6 / 8 Fragments Scores Download
4	 DL-Nicotine	DTXSID3048154 InChIKeyBlock1 = SNICXCGAKADSCV	162.11576	$C_{10}H_{14}N_2$		2.8602	Peaks: 6 / 8 Fragments Scores Download
5	 (S)-Nicotine dicitrate	DTXSID20237235 InChIKeyBlock1 = SNICXCGAKADSCV	162.11576	$C_{10}H_{14}N_2$		2.7628	Peaks: 6 / 8 Fragments Scores Download

Plan for Today

- Background
  - Non-target Screening & Identification Confidence
  - Compound Databases and Spectral Libraries
- Introduction to MetFrag: MS/MS and Metadata
- MetFrag + CompTox + MassBank + Nicotine
  - MS-ready groupings … or not
- MetFrag + PubChem + MassBank + Isobars
  - The relevance of spectral matching!
- Future: MetFrag for High Throughput Exposomics
  - Build your own databases … PubChemLite



Isobars: The tricky cases!

$C_9H_{16}ClN_5$ *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

NORMAN Digital Sample Freezing Platform

Contributed Samples | Results | Chromatograms | Interactive Map | Help

Joint Black Sea Survey 2016

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

MetFrag + PubChem + MS/MS + Metadata

MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Database Settings

Database: PubChem Include references: Parent Ion:

Neutral Mass: 201.07816 Search ppm: 5

Formula: C7H12ClN5

Identifiers:

Retrieve Candidates 73 Candidates Download Candidates

Schymanski (2019). ChemInChIformatics: IUPAC's Role in the Environmental Monitoring Revolution. DOI: 10.5281/zenodo.3270900

MetFrag Web Interface – Add MS/MS Details

MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Fragmentation Settings & Processing

Mzppm: 5

Mzabs: 0.001

Mode: [M+H]+

Tree depth: 2

Group candidates

MS/MS Peak list

```
57.0699 34824.5 11
61.9792 27260 8
68.0244 499632.7 162
79.0059 755457.7 245
104.0011 904831.9 293
110.0462 250173.3 81
128.0568 46548.7 15
146.023 3077763.2 999
```

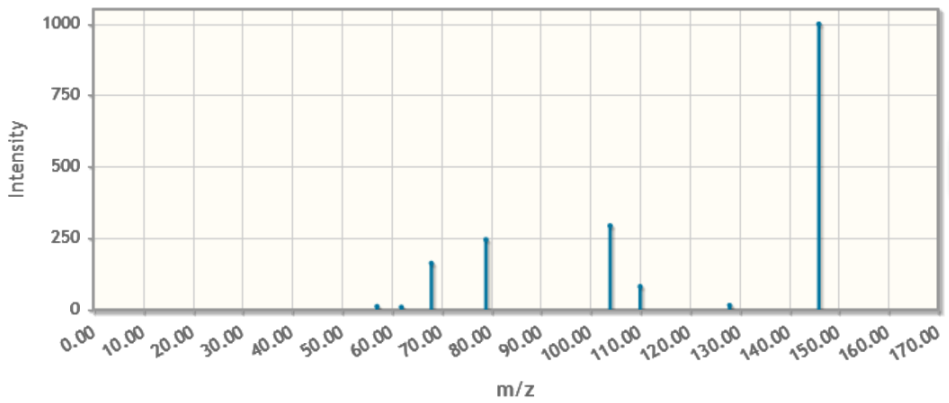
Show Spectrum

Process Candidates

Download Parameters

Spectrum View

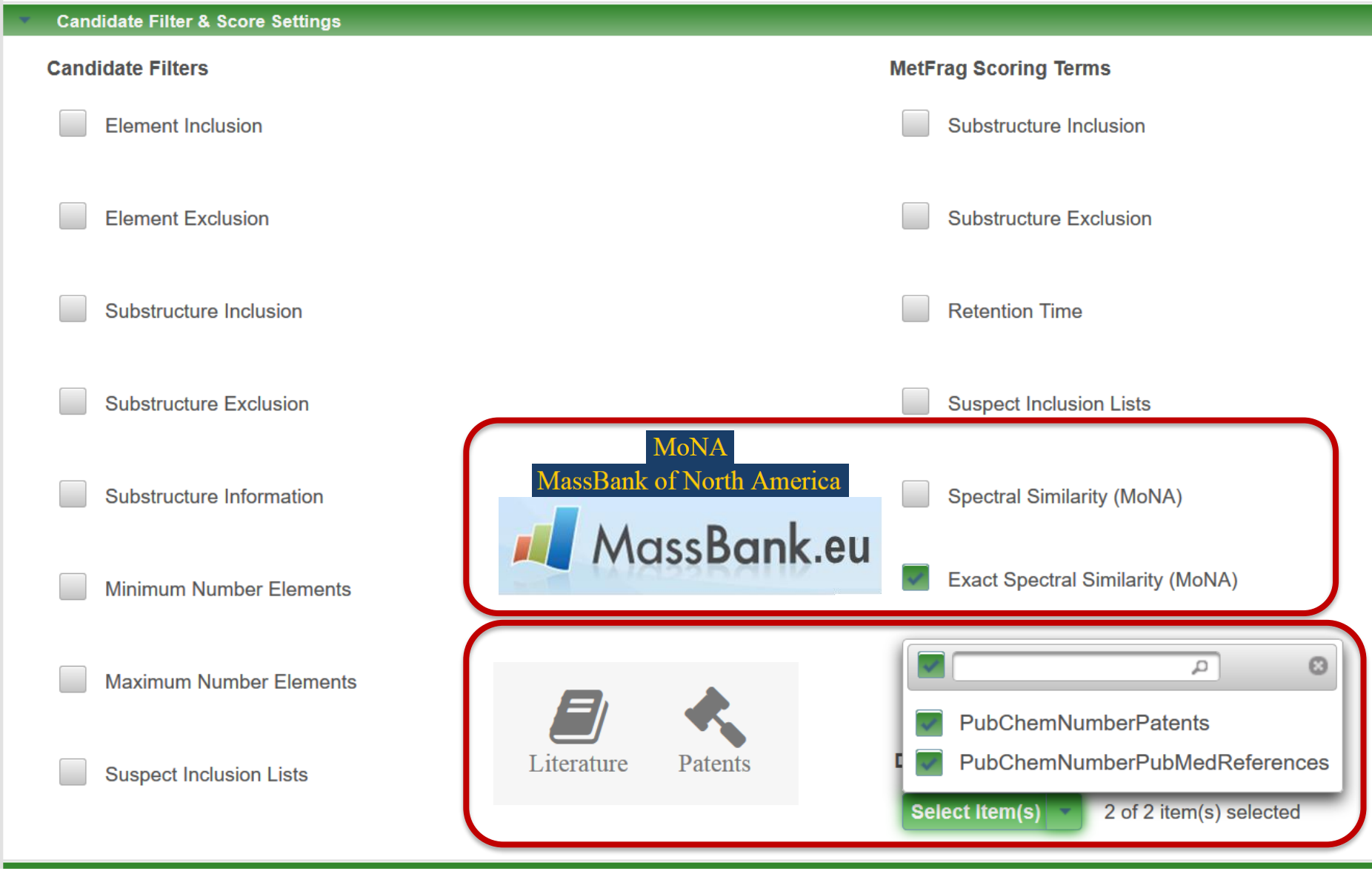
Select area to zoom in. Double click to return.



Export

https://massbank.eu/MassBank/RecordDisplay.jsp?id=EA067112&dsn=Eawag

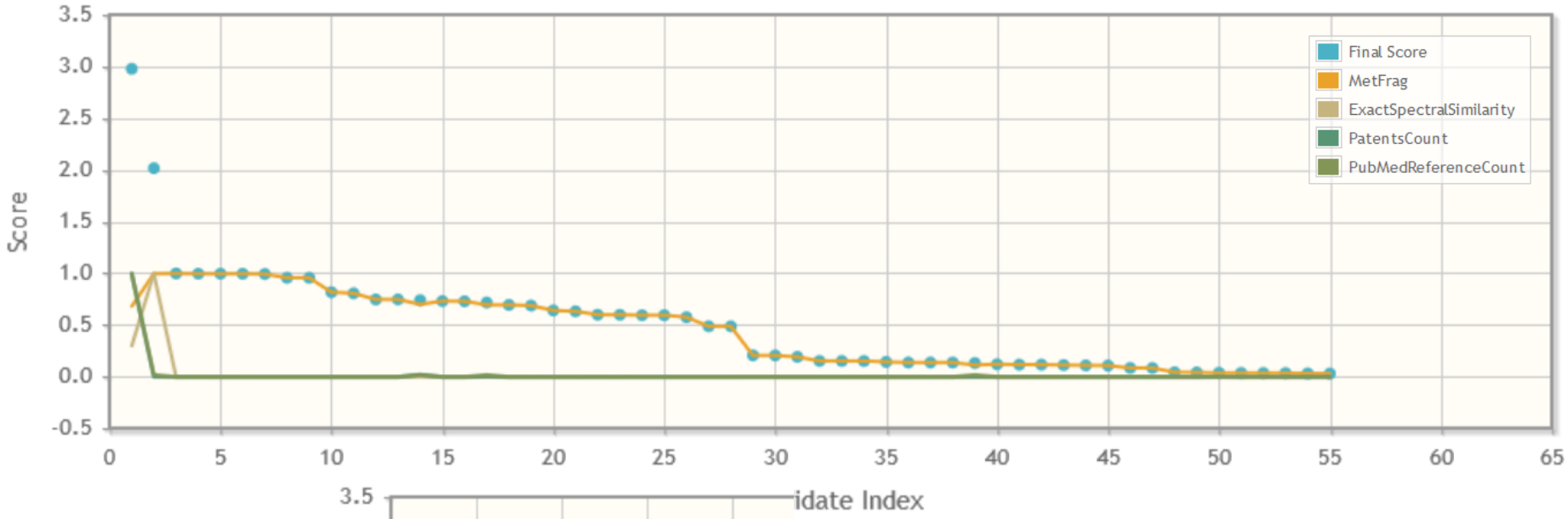
MetFrag + PubChem: add MS Library + Metadata



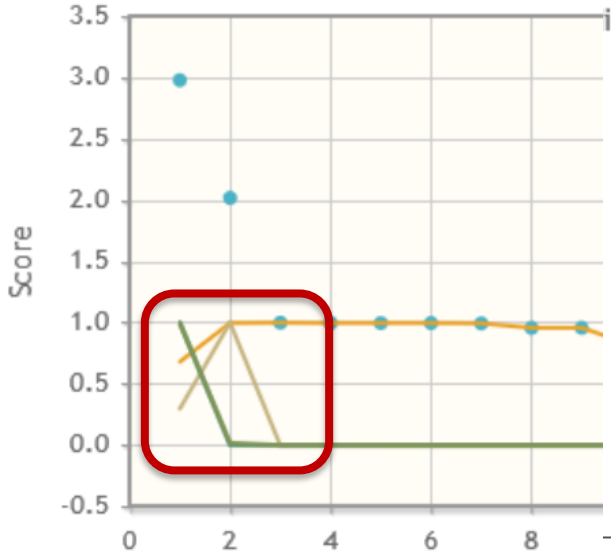
MetFrag Results

Statistics

Candidate Score Distribution

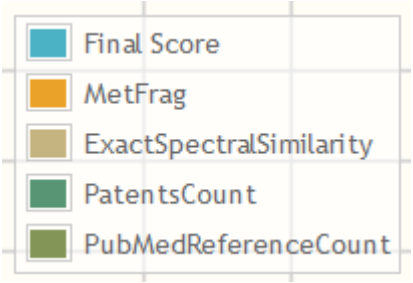


Export



Select area to zoom in. Double click to return.

Click on dot to scroll to candidate in the Results tab.



MetFrag Results – Metadata is good … but …

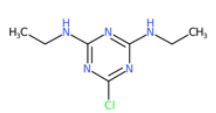
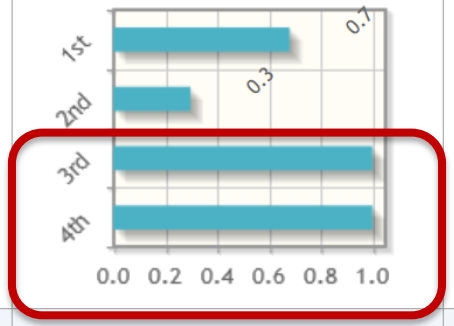
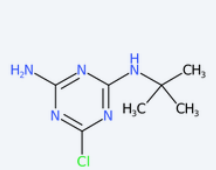
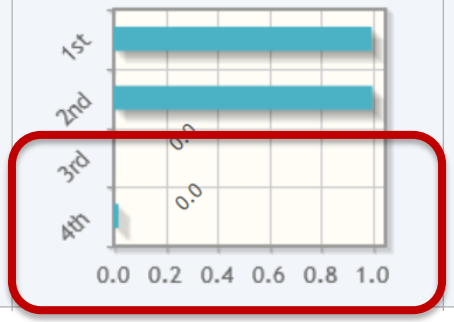
Weights		
MetFrag (1st)	100	%
ExactSpectralSimilarity (2nd)	100	%
PatentsCount (3rd)	100	%
PubMedReferenceCount (4th)	100	%

Download Results

Filter Candidates by explained MS/MS Peaks

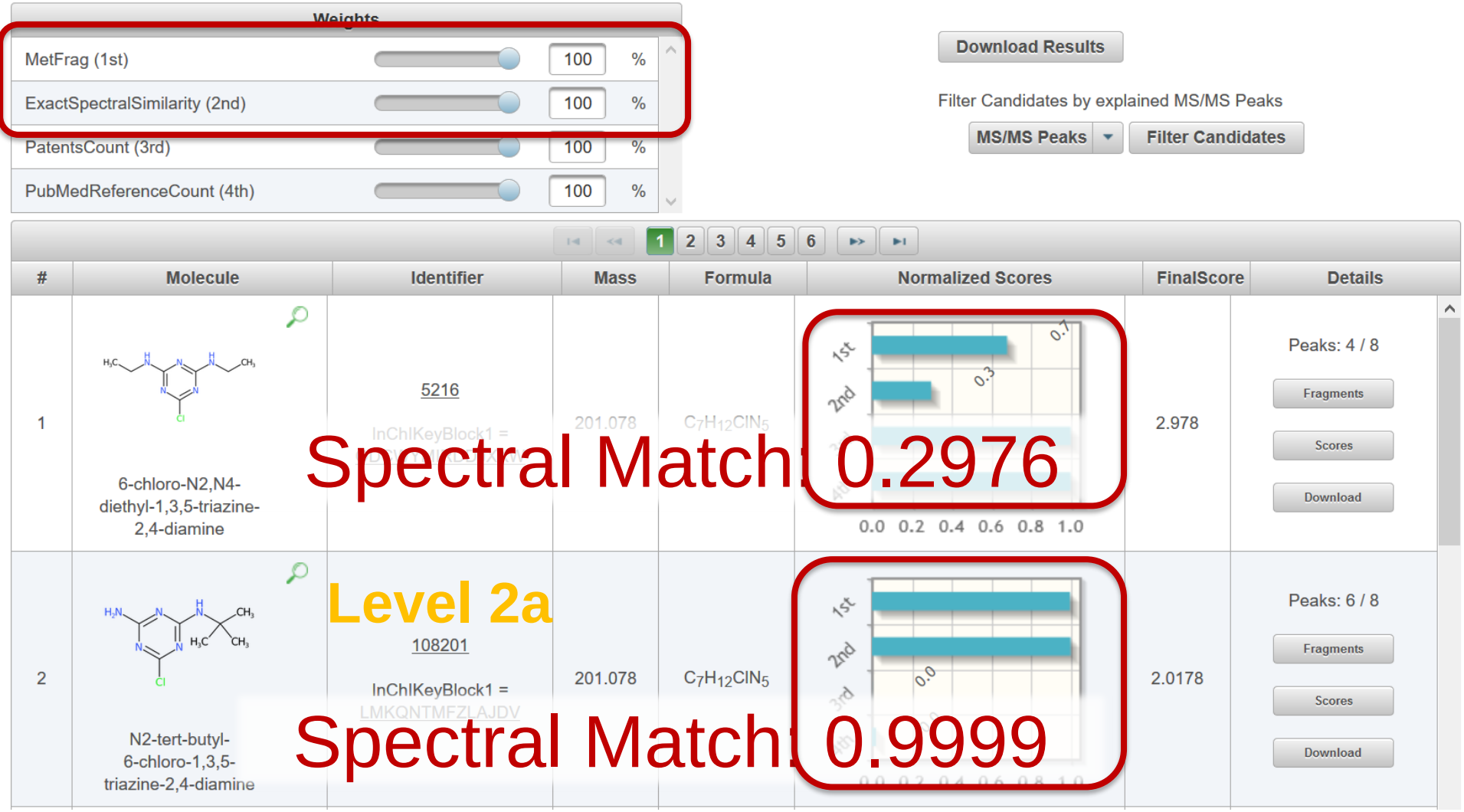
MS/MS Peaks | Filter Candidates

1 2 3 4 5 6

#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1	 6-chloro-N2,N4-diethyl-1,3,5-triazine-2,4-diamine	5216 InChIKeyBlock1 = ODCWYMIRDDJXKW	201.078	$C_7H_{12}ClN_5$		2.978	Peaks: 4 / 8 Fragments Scores Download
2	 N2-tert-butyl-6-chloro-1,3,5-triazine-2,4-diamine	108201 InChIKeyBlock1 = LMKQNTMFZLAJDV	201.078	$C_7H_{12}ClN_5$		2.0178	Peaks: 6 / 8 Fragments Scores Download

Schymanski (2019). ChemInChIformatics: IUPAC's Role in the Environmental Monitoring Revolution. DOI: 10.5281/zenodo.3270900

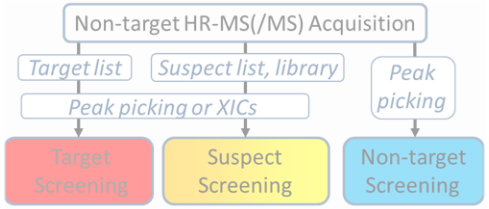
MetFrag Results - Consider Experimental Evidence!



Schymanski (2019). ChemInChIformatics: IUPAC's Role in the Environmental Monitoring Revolution. DOI: 10.5281/zenodo.3270900

Plan for Today

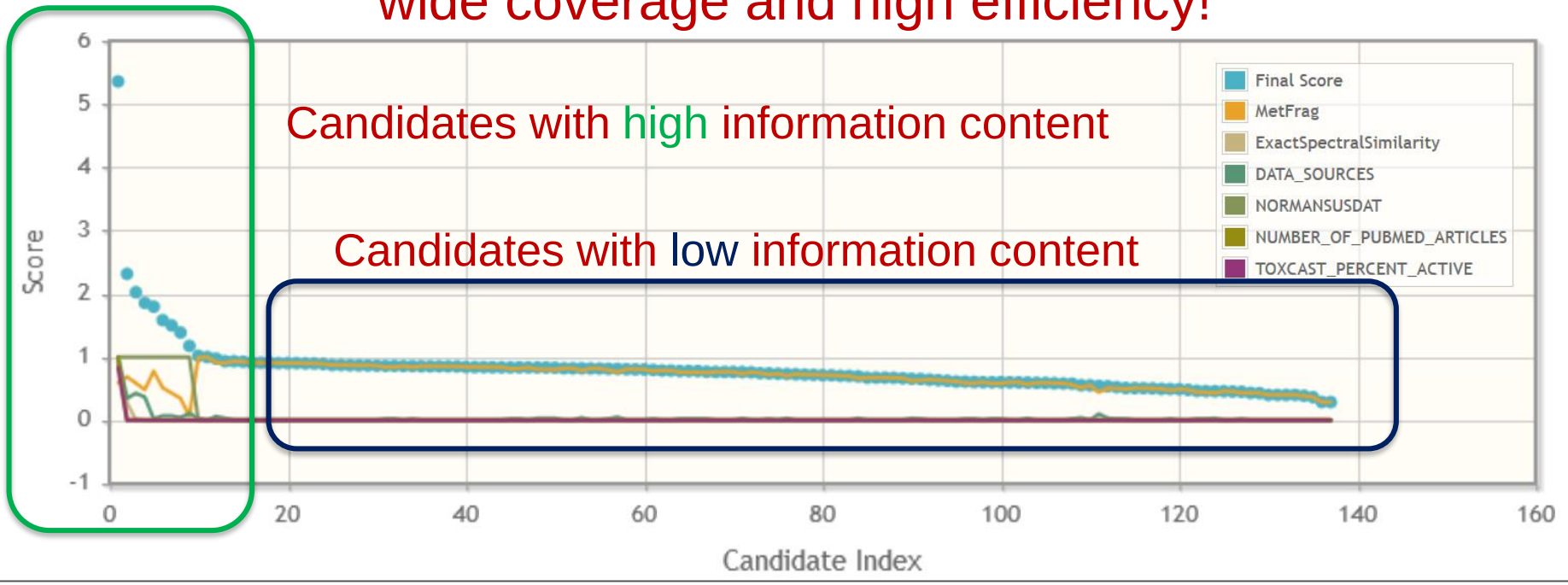
- Background
  - Non-target Screening & Identification Confidence
  - Compound Databases and Spectral Libraries
- Introduction to MetFrag: MS/MS and Metadata
- MetFrag + CompTox + MassBank + Nicotine
  - MS-ready groupings … or not
- MetFrag + PubChem + MassBank + Isobars
  - The relevance of spectral matching!
- Future: MetFrag for High Throughput Exposomics
  - Build your own databases … PubChemLite



Connecting multiple lines of evidence for identification

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

Challenge: the growing number of candidates …
High throughput exposomics needs both
wide coverage and high efficiency!



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

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

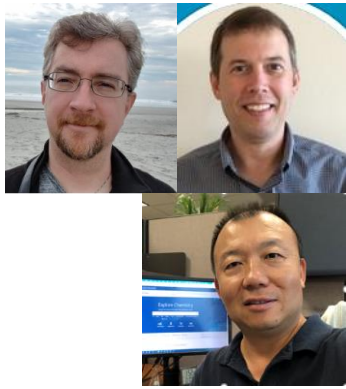
- PubChem Compound TOC ? 33,765,953
  - Agrochemical Information ? 2,002
  - Biologic Description ? 1,539,532
  - Biological Test Results ? 3,467,416
  - Biomolecular Interactions and Pathways ? 109,610
  - Chemical and Physical Properties ? 237,729
  - Classification ? 18,569,627
  - Diseases ?
  - 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:

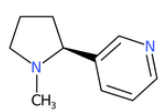
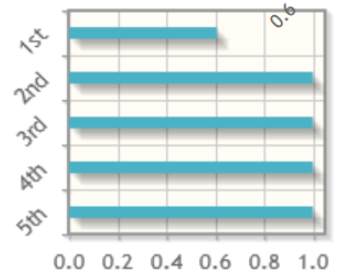
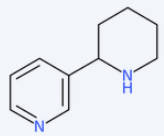
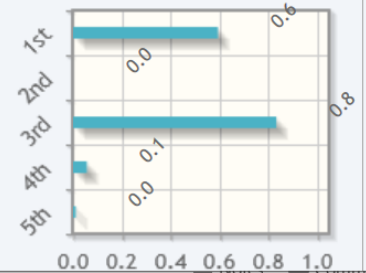
Identifiers:

Retrieve Candidates

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

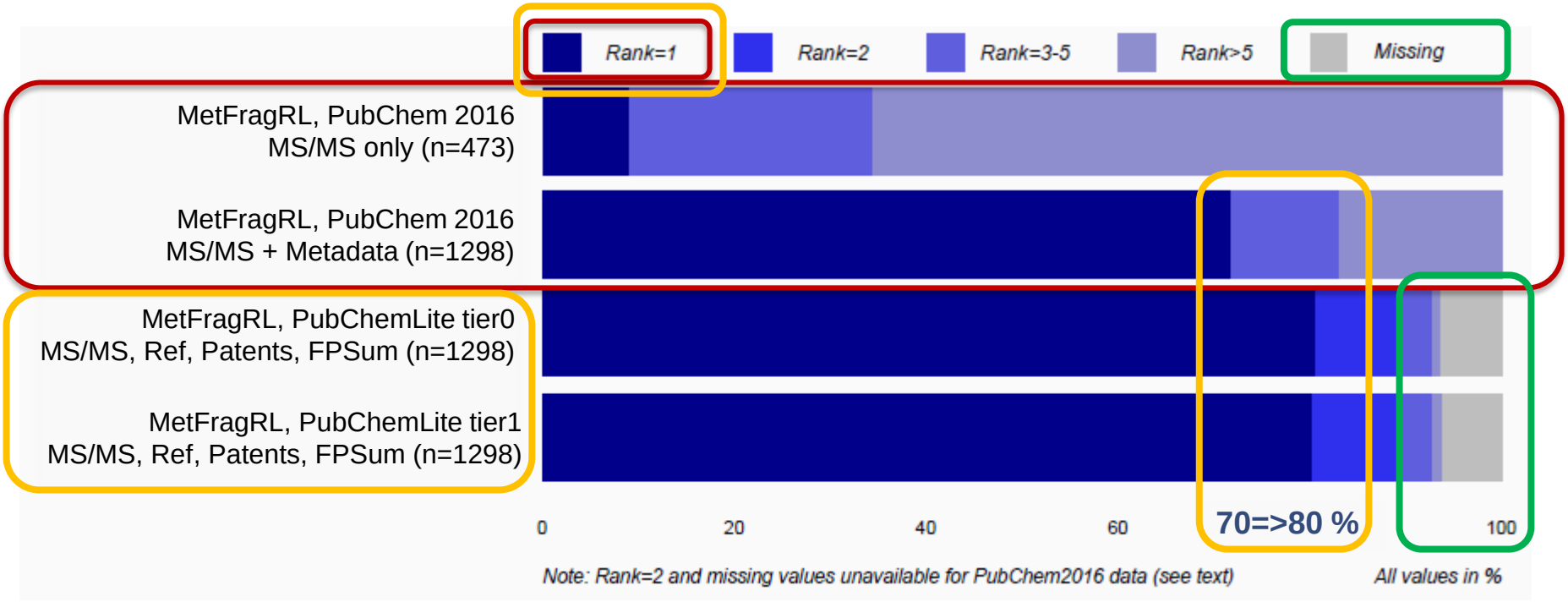
100 %

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?

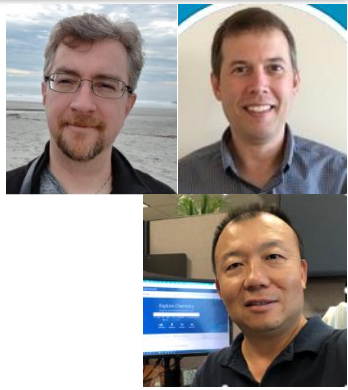


MetFragRL: Ruttkies et al. (2016) DOI: 10.1186/s13321-016-0115-9
Bolton, Schymanski et al., in prep. +
Bolton & Schymanski (2020). PubChemLite tier0 and tier1, DOI: 10.5281/zenodo.3611238

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

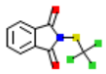
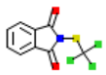
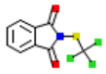


PubChem Folpet (Compound)

8.7 Transformations

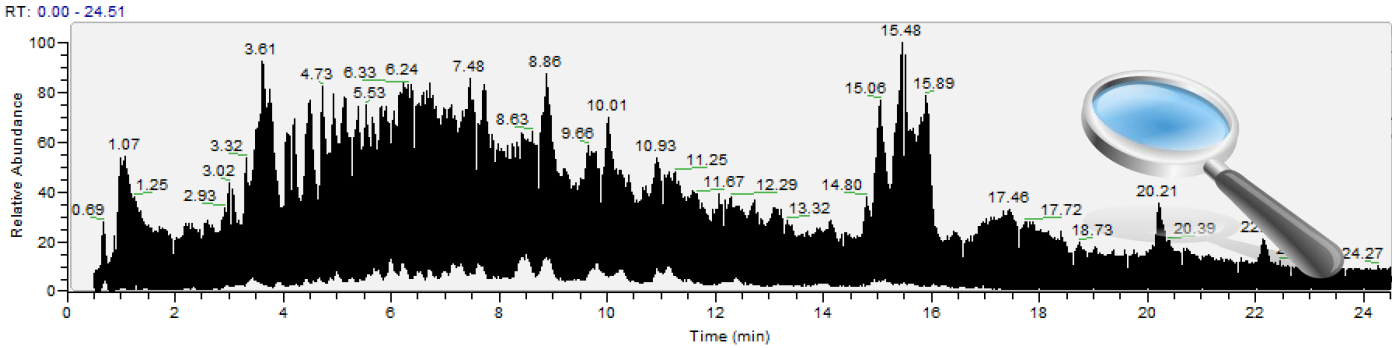
3 items View More Details Download

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***

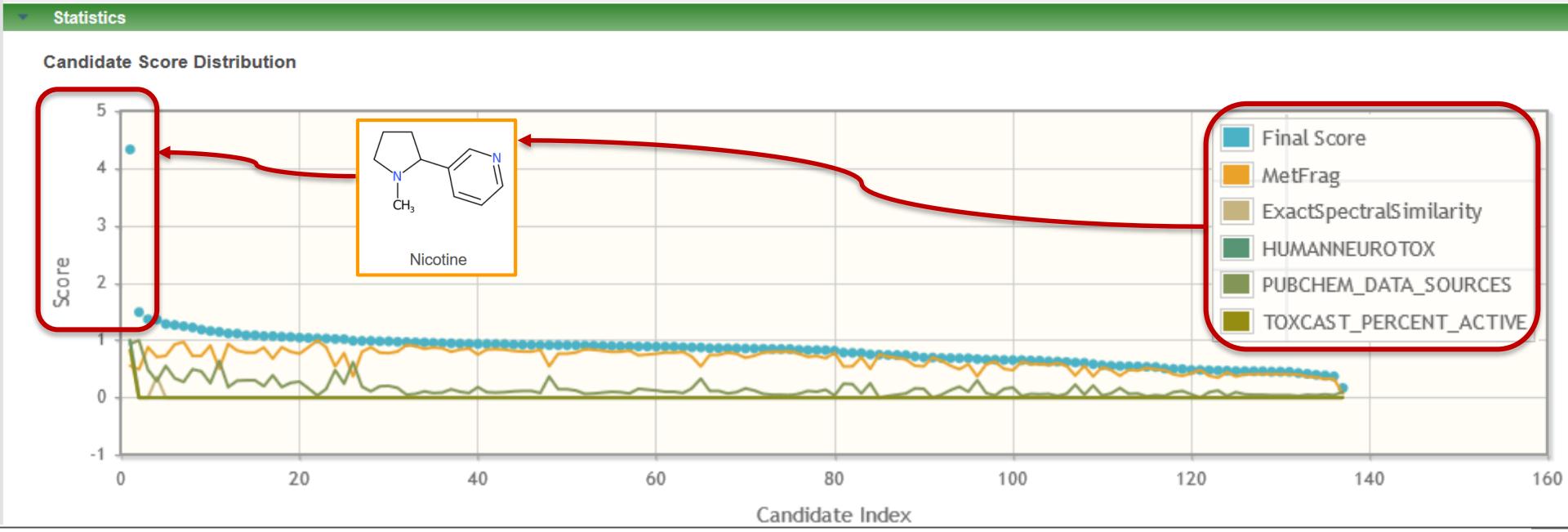


Take Home Messages

- 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 => ***70-80 %***!

Take Home Messages

- Over 60 % of HR-MS peaks are potentially ***relevant*** but ***unknown***
- ***Annotating unknowns*** requires data and evidence from ***many different sources***
- Annotating exposomics “known unknowns” with MetFrag 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/

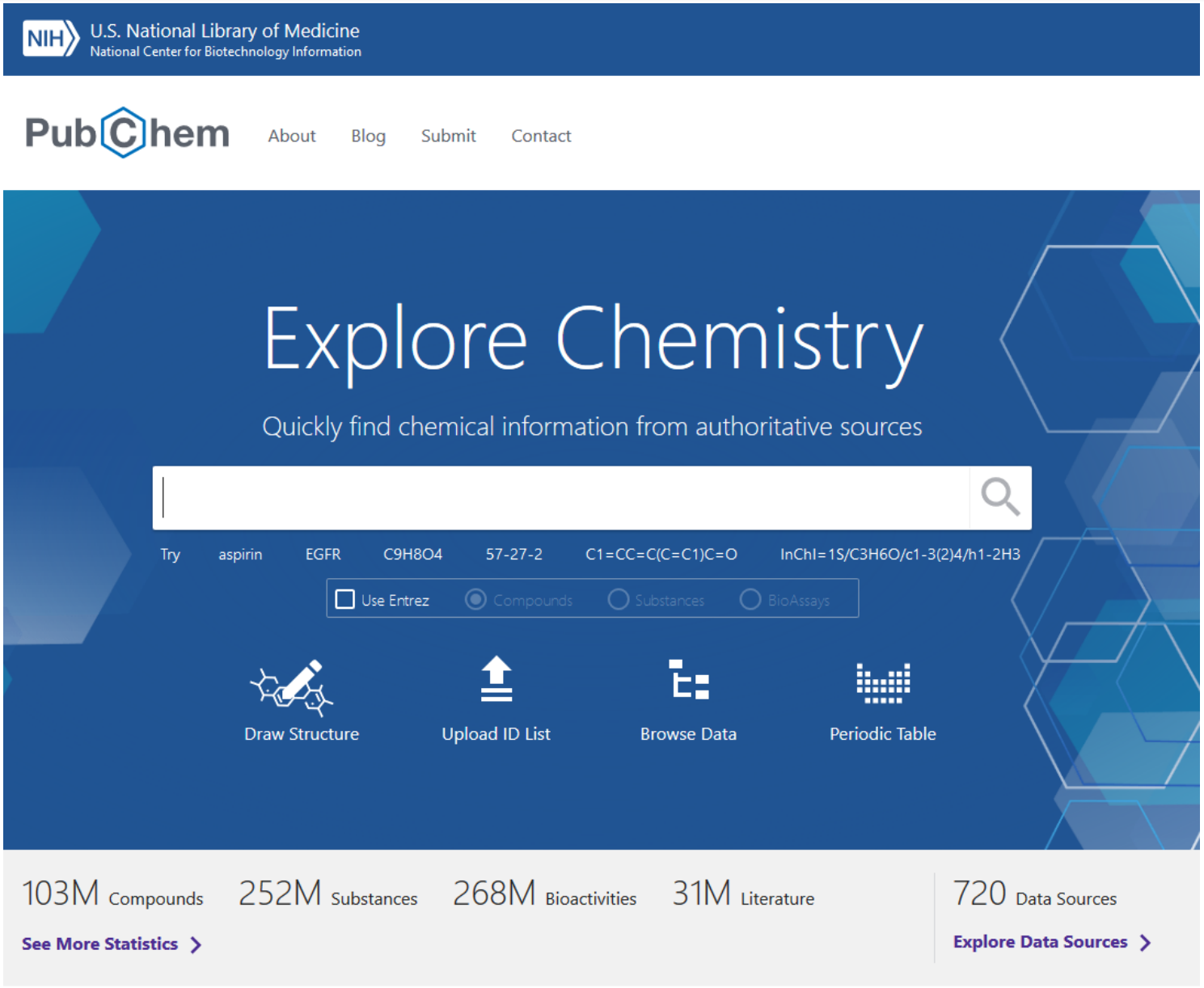
https://www.norman-network.com/nds/SLE/

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

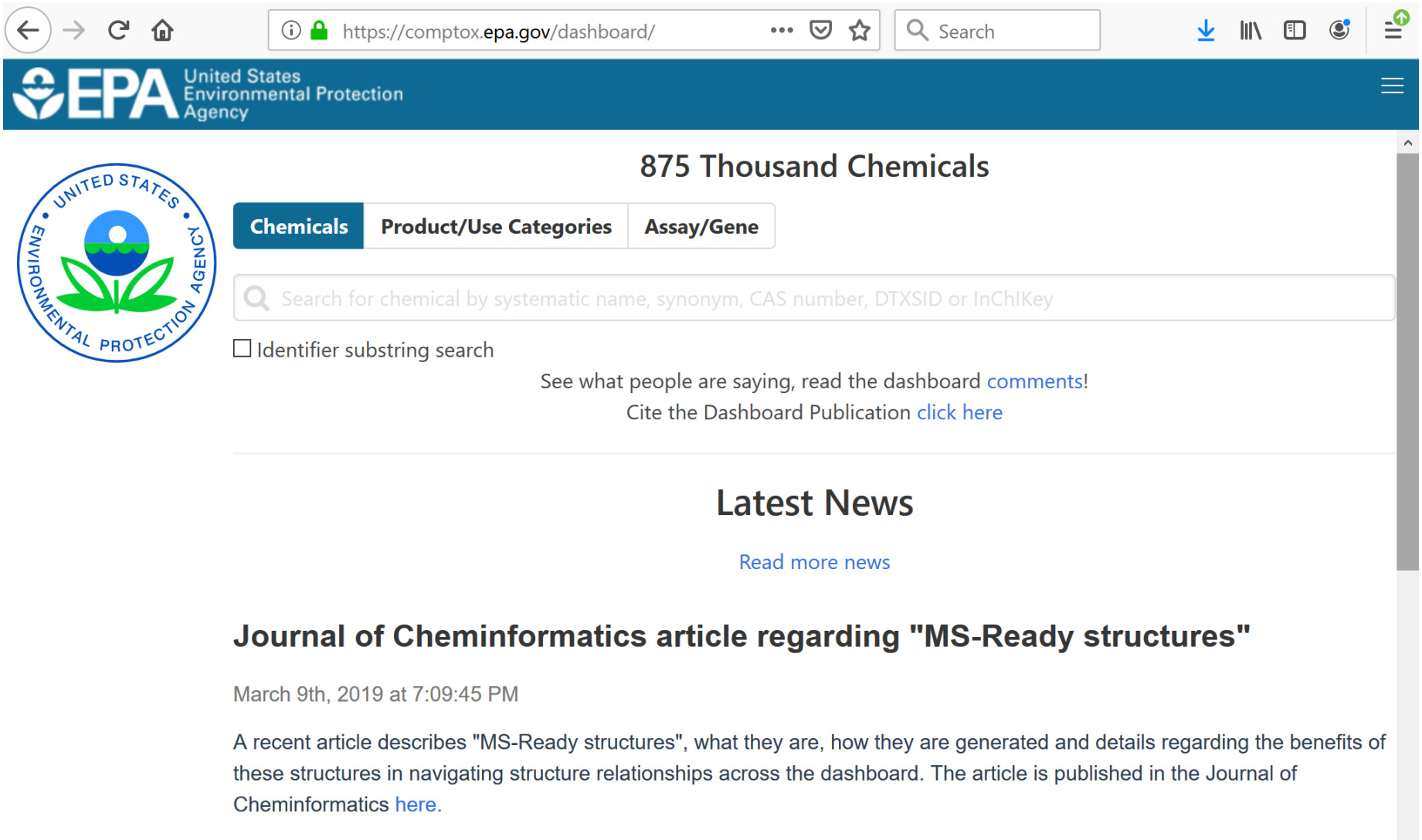
Slides and supplementary data available under DOI: 10.5281/zenodo.3953683


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

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

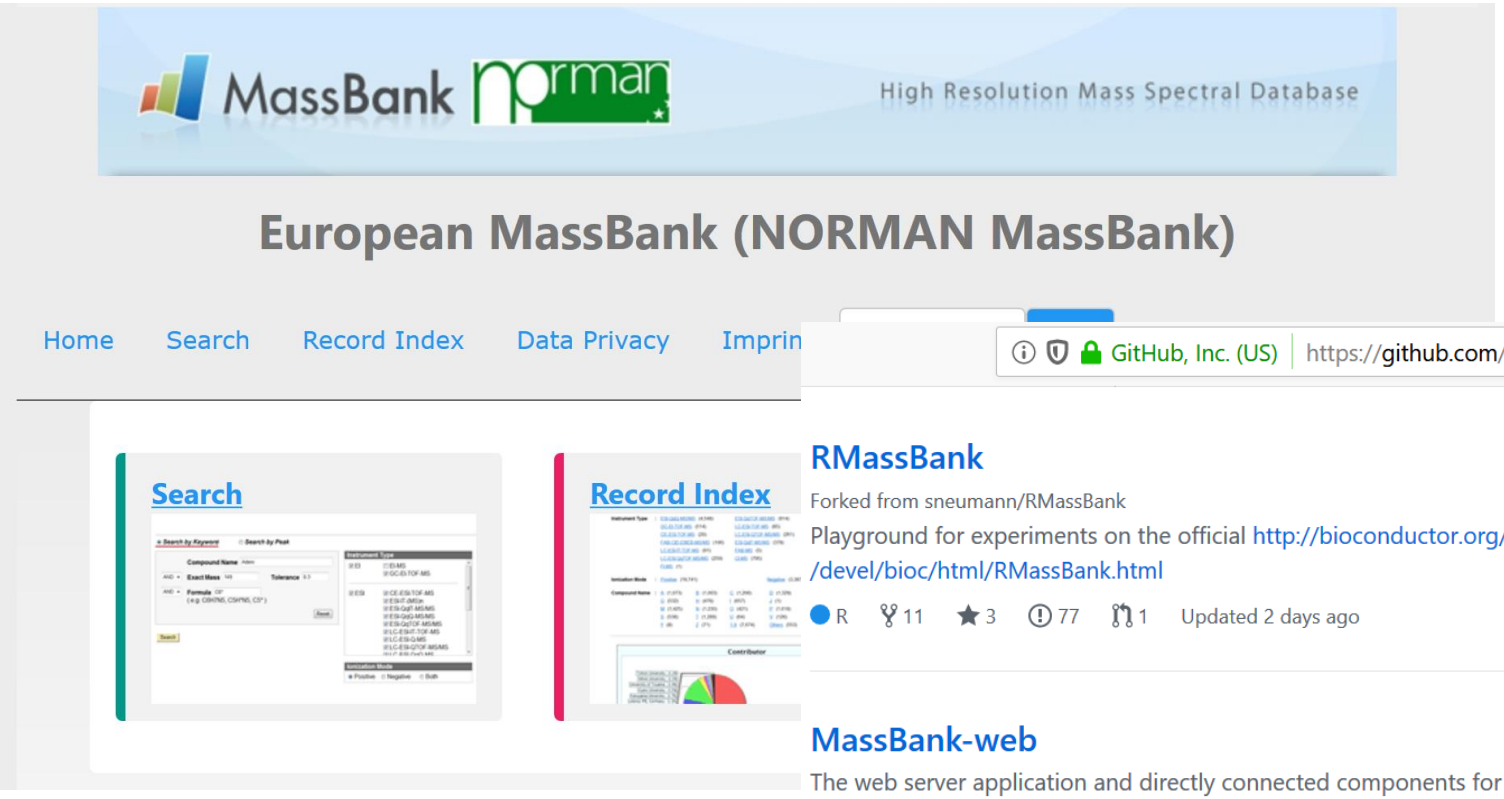


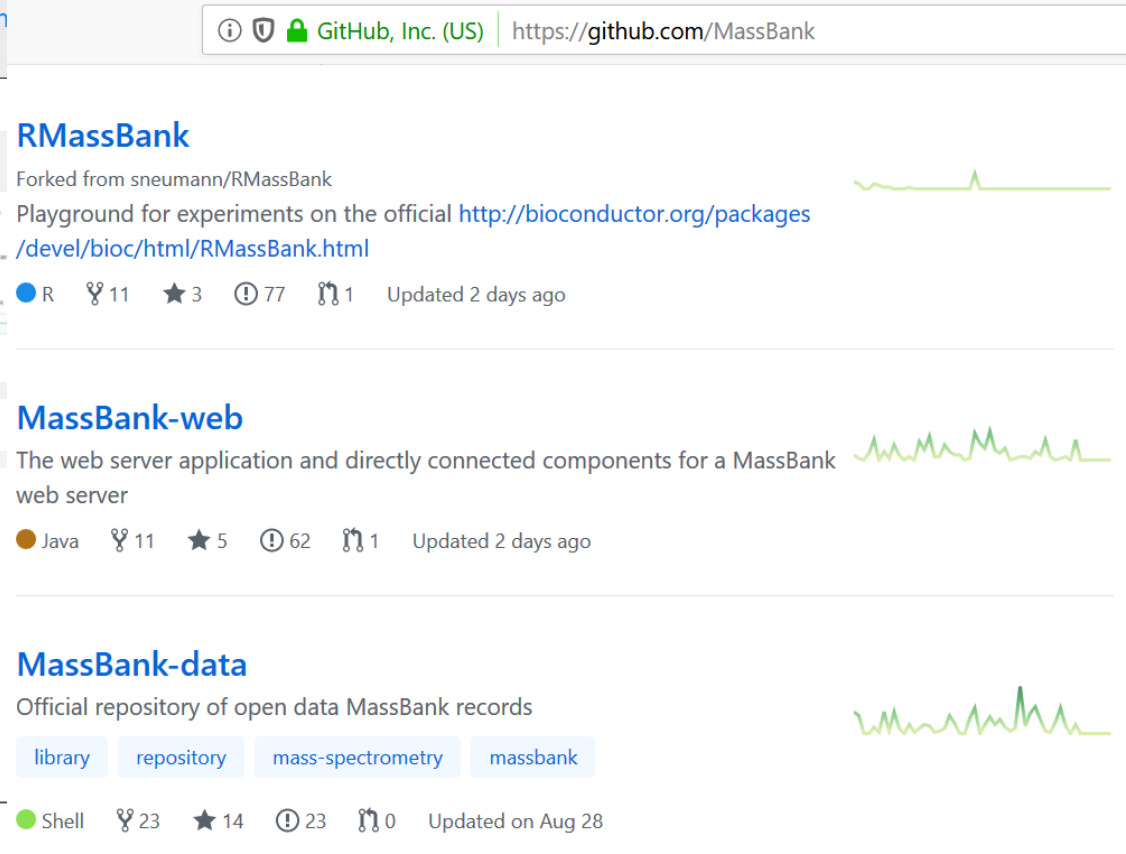
CompTox: https://comptox.epa.gov/dashboard/



MassBank EU

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





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