http://www.orcid.org/0000-0002-2668-4821

Using the US EPA's CompTox Chemistry Dashboard for structure identification and non-targeted analyses

Antony Williams[1], Andrew D. McEachran[3], Seth Newton[2], Kristin Isaacs[2], Katherine Phillips[2], Nancy Baker[1], Chris Grulke[1] and Jon R. Sobus[2]

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC
2) National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC
3) Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC

March 2018
ACS Spring Meeting, New Orleans

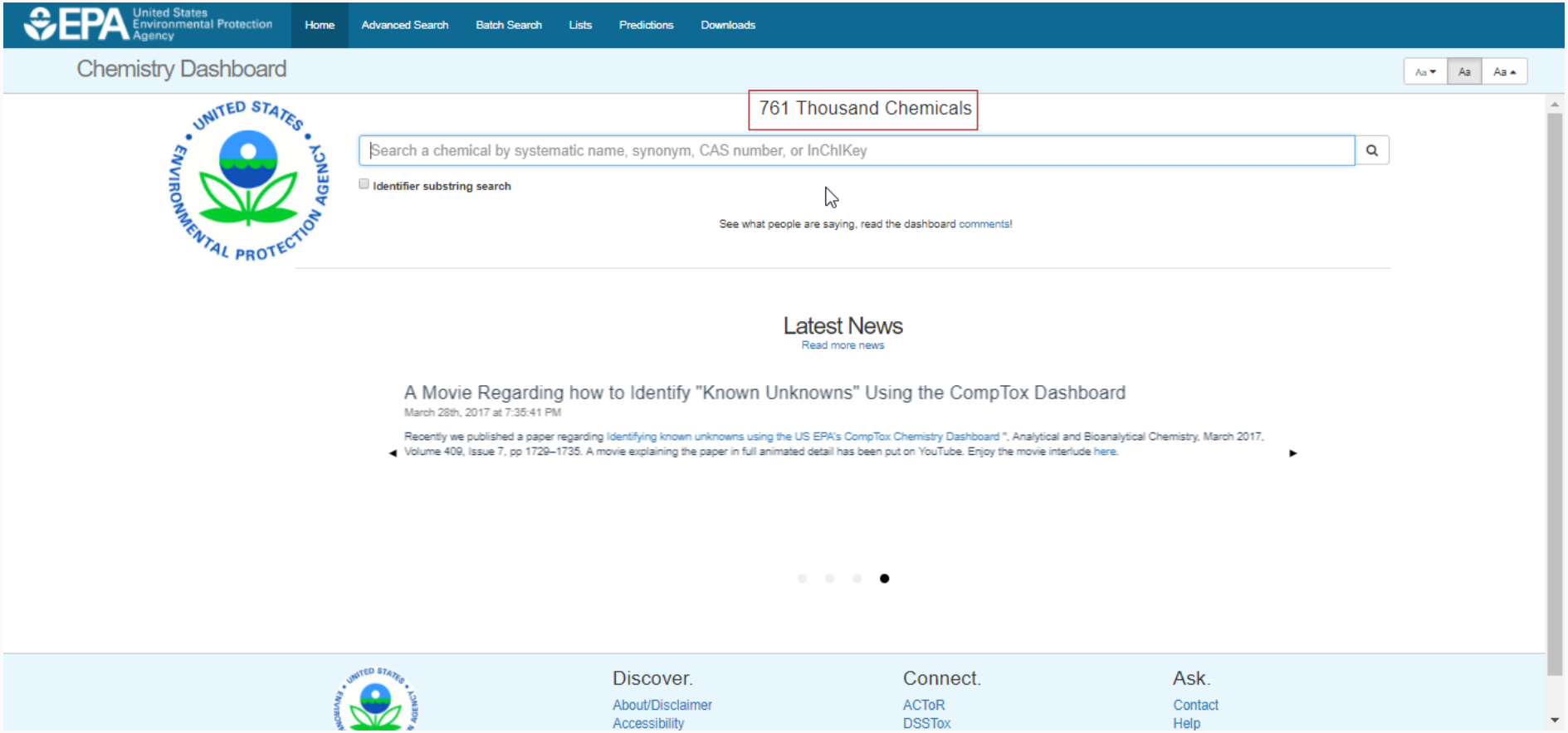
The CompTox Chemistry Dashboard

- A **publicly accessible website** delivering access:
  - ~760,000 chemicals with related property data
  - Experimental and predicted physicochemical property data
  - Experimental Human and Ecological hazard data
  - Integration to “biological assay data” for 1000s of chemicals
  - Information regarding consumer products containing chemicals
  - Links to other agency websites and public data resources
  - “Literature” searches for chemicals using public resources
  - “Batch searching” for thousands of chemicals

  - **DOWNLOADABLE Open Data for reuse and repurposing**

CompTox Chemistry Dashboard

https://comptox.epa.gov/dashboard



1 of ~761,000 Chemical Pages

EPA United States Environmental Protection Agency

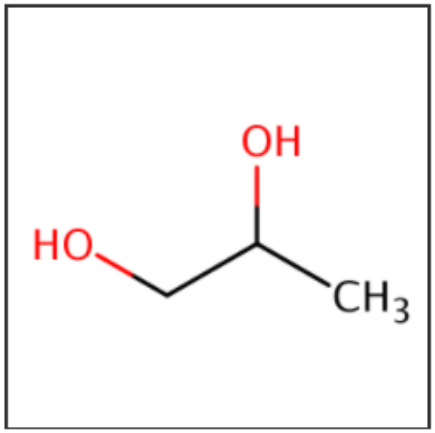
EPA United States Environmental Protection Agency | Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | Search All Data

Chemistry Dashboard | EPAHFR | Submit Comment | Copy | Aa | Aa | Aa

1,2-Propylene glycol

57-55-6 | DTXSID0021206

Searched by DSSTox_Substance_Id: Found 1 result for 'DTXSID0021206'.



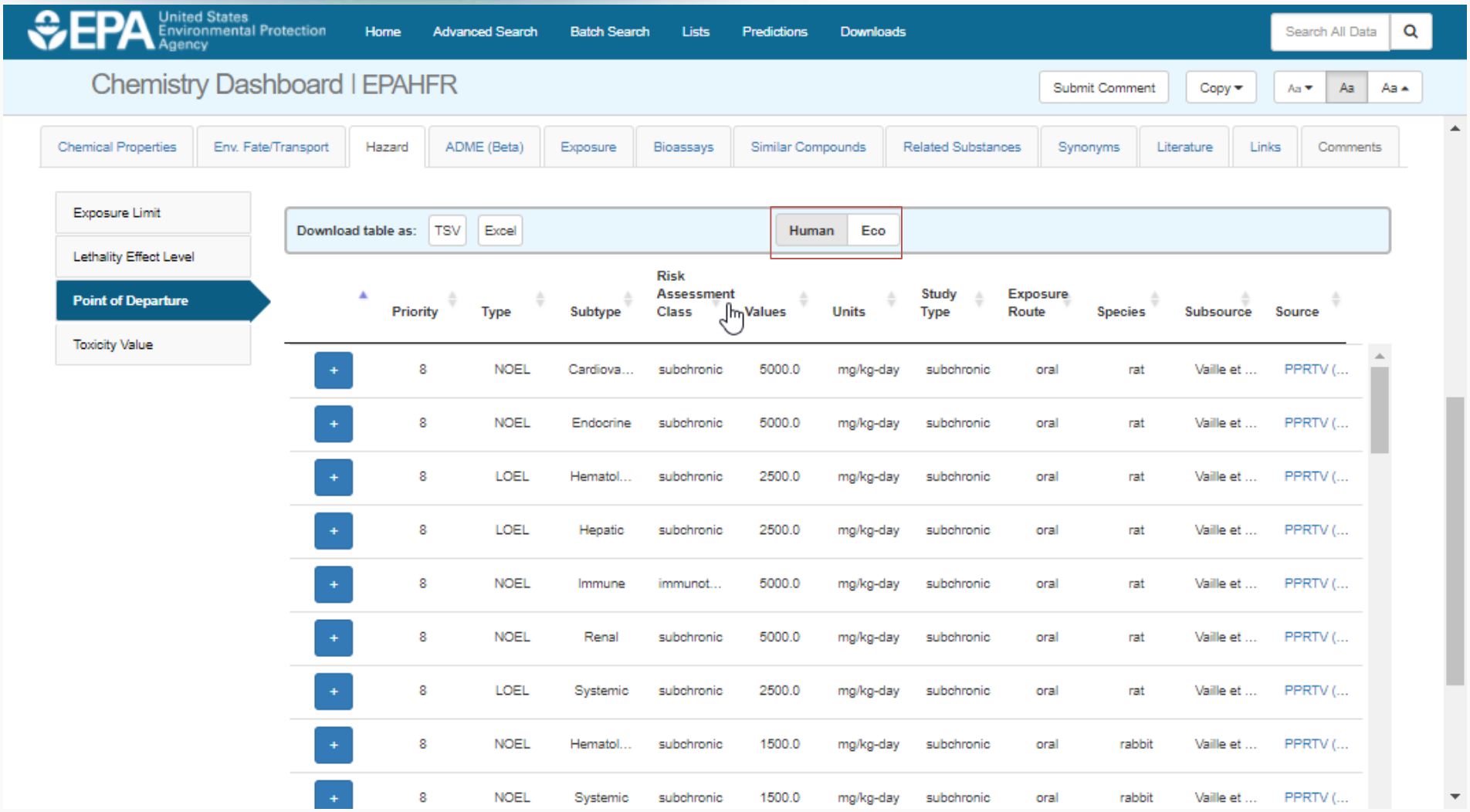
Wikipedia

Propylene glycol (IUPAC name: propane-1,2-diol) is a synthetic organic compound with the chemical formula C3H8O2. It is a viscous colorless liquid which is nearly odorless but possesses a faintly sweet taste. Chemically it is classed as a diol and is miscible with a broad range of solvents, including water, acetone, and chloroform. It is produced on a large scale and is primarily used in the production of polymers, but also sees use in food...Read more

- Intrinsic Properties
- Structural Identifiers
- Linked Substances
- Presence in Lists
- Record Information
- Quality Control Notes

Chemical Properties | Env. Fate/Transport | Hazard | ADME (Beta) | Exposure | Bioassays | Similar Compounds | Related Substances | Synonyms | Literature | Links | Comments

Access to Chemical Hazard Data



	Priority	Type	Subtype	Risk Assessment Class	Values	Units	Study Type	Exposure Route	Species	Subsource	Source
+	8	NOEL	Cardiova...	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Endocrine	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Hematol...	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Hepatic	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Immune	immunot...	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Renal	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Systemic	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Hematol...	subchronic	1500.0	mg/kg-day	subchronic	oral	rabbit	Vaille et ...	PPRTV (...
+	8	NOEL	Systemic	subchronic	1500.0	mg/kg-day	subchronic	oral	rabbit	Vaille et ...	PPRTV (...

Sources of Exposure to Chemicals

- **Product & Use Categories**
- Chemical Weight Fraction
- Chemical Functional Use
- Monitoring Data
- Exposure Predictions
- Production Volume

osure | Bioassays | Similar Compounds | Related Substances | Synonyms | Literature | Links | Comments

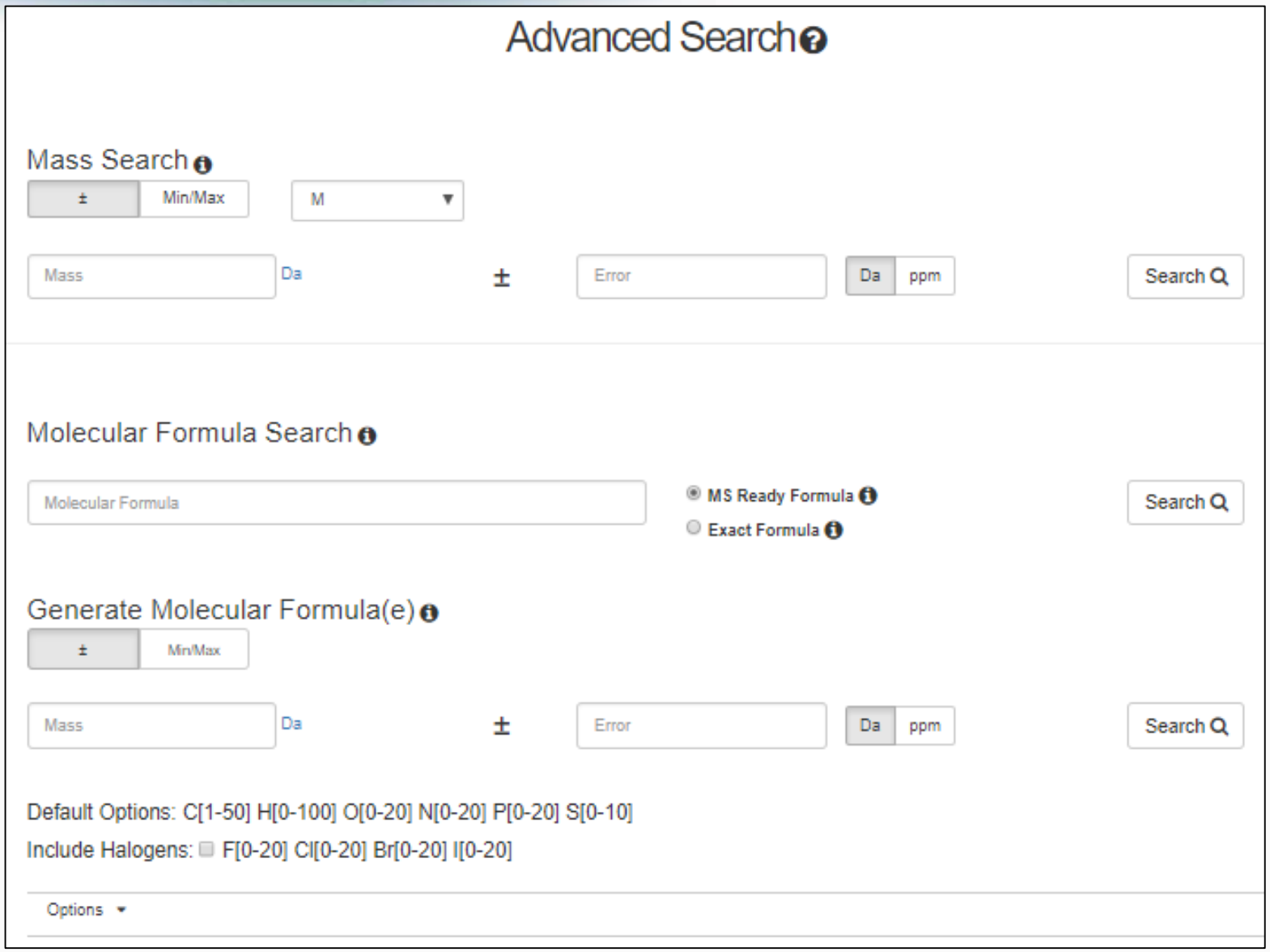
Product & Use Categories (PUCs)

Categorization type	Number of Unique Products
PUC	288
PUC	206
PUC	117
PUC	107
PUC	107
PUC	101
PUC	101
PUC	90
PUC	89

Dashboard for Structure ID

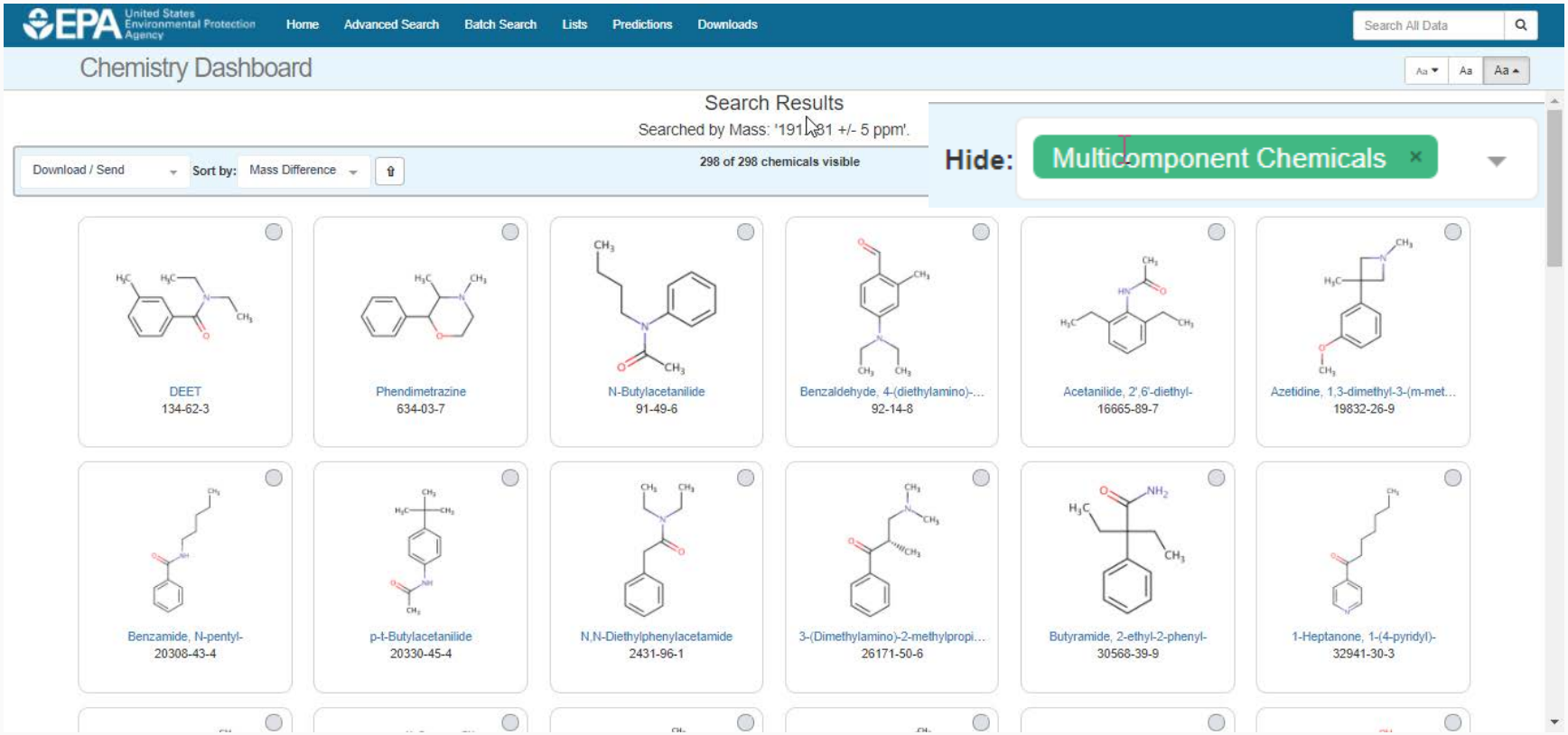
- Structure Identification using the dashboard
  - Formula/mass-based searching – 1 chemical at a time

Advanced Searches

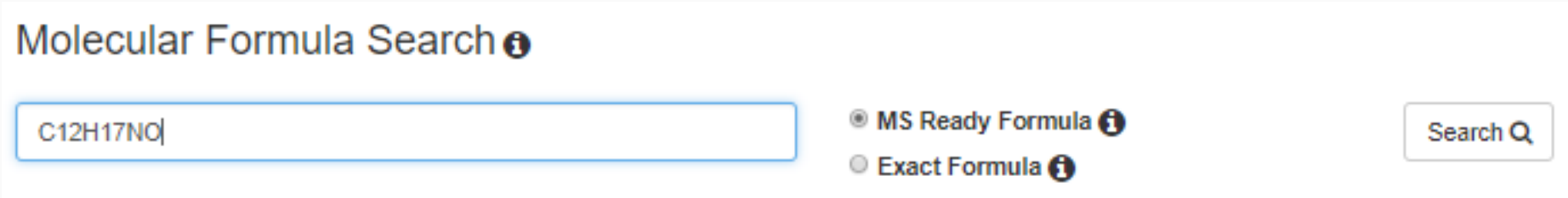


Advanced Searches
Mass Based Search

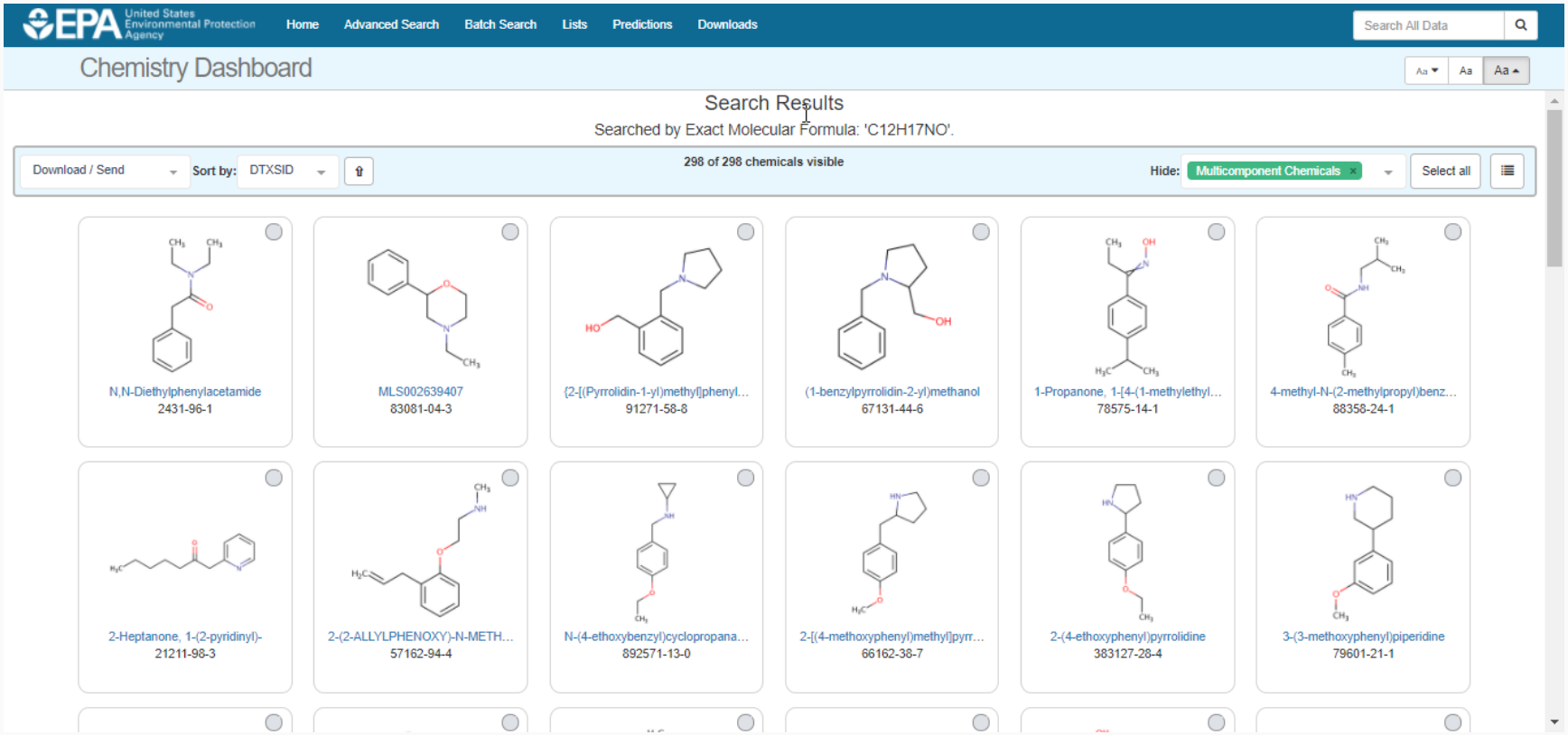
Advanced Searches



Formula Searches



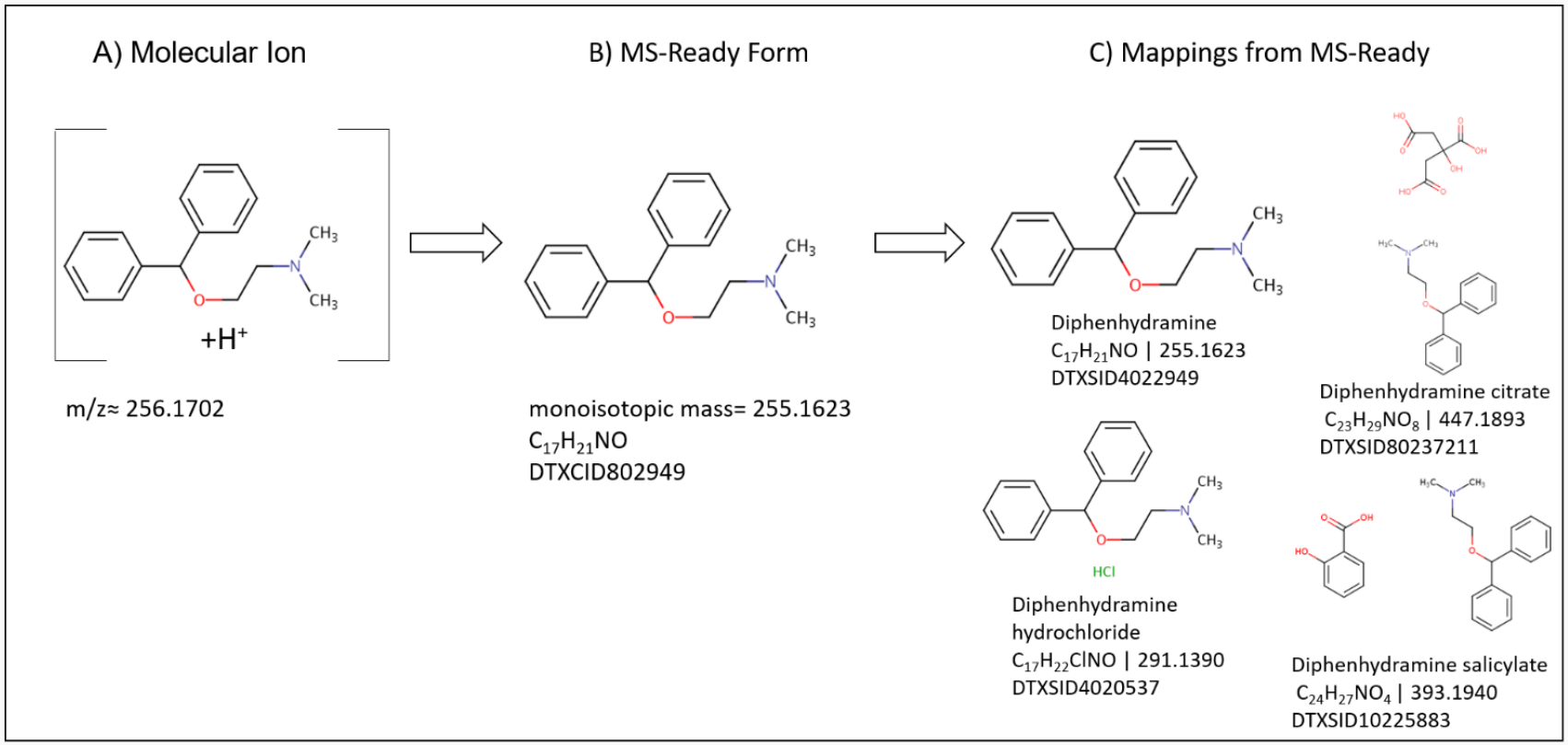
Exact Formula Search: C12H17NO
298 Chemicals



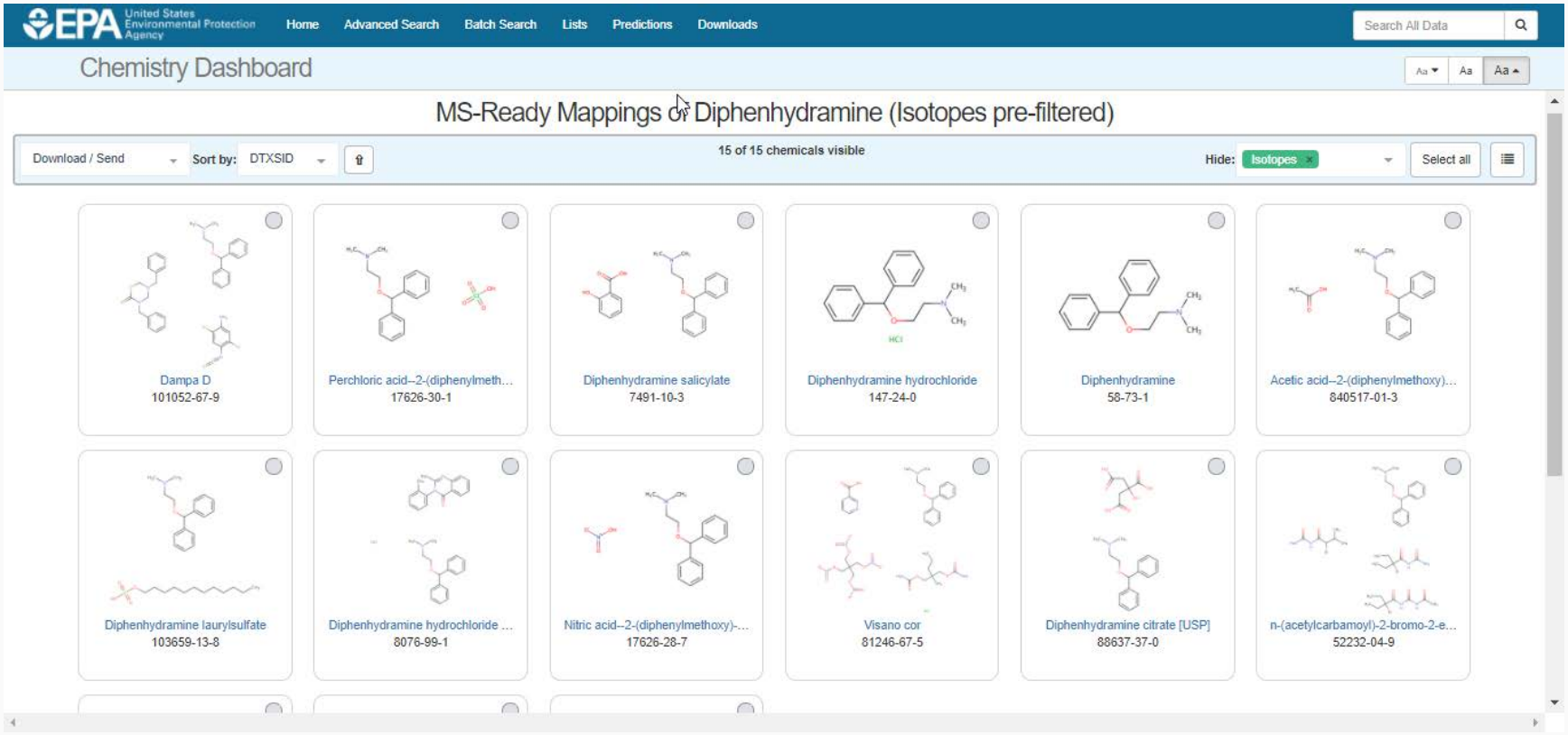
Dashboard for Structure ID

- Structure Identification using the dashboard
  - Formula/mass-based searching – 1 chemical at a time
  - Distilling structures into “MS-Ready form”

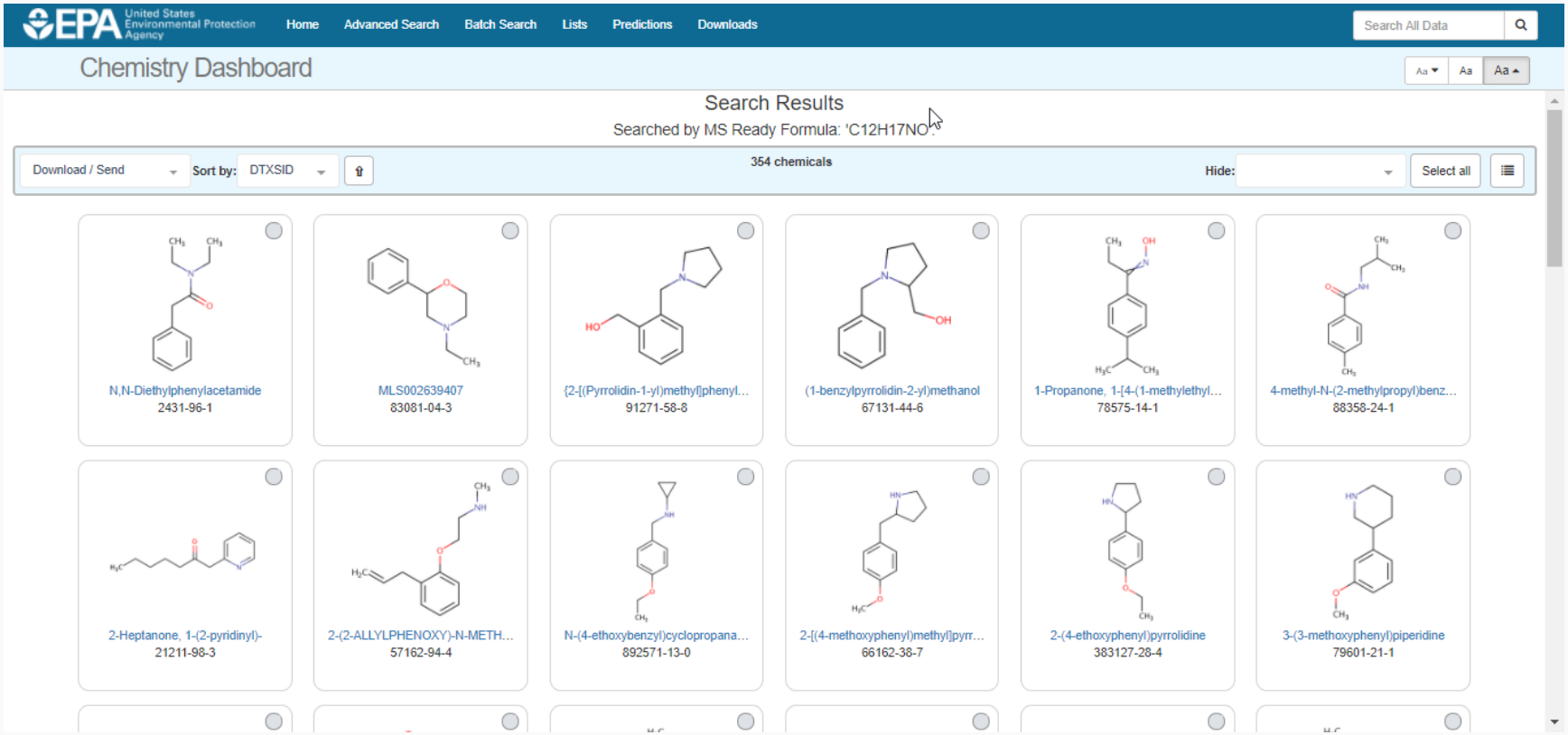
Specific Data-Mappings "MS-Ready Structures"



Diphenhydramine
15 Total MS-Ready Mappings



“MS Ready” Formula Search C12H17NO 354 Chemicals



Dashboard for Structure ID

- Structure Identification using the dashboard
  - Formula/mass-based searching – 1 chemical at a time
  - Distilling structures into “MS-Ready form”
  - Ranking based on metadata

Identifying Known Unknowns by reference ranking

Journal of The American Society for Mass Spectrometry
January 2012, Volume 23, Issue 1, pp 179–185 | Cite as

Table 1

Searching ChemSpider by Elemental Composition then Sorting by Number of Associated References

Class of compounds	Number compounds in class	Position of compound sorted in descending order by number of references					
		#1	#2	#3	#4	#5	>#5
Drugs	45	43	1	1			
Pesticides	8	7	1				
Toxins	2	2					
Polymer antioxidants	15	15					
Polymer UV stabilizers	10	8	1	1			

Data source ranking using the Dashboard

Anal Bioanal Chem
DOI 10.1007/s00216-016-0139-z

RAPID COMMUNICATION

Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard

Andrew D. McEachran[1] · Jon R. Sobus[2] · Antony J. Williams[3]

	Mass-based searching		Formula-based searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average rank position	1.3	2.2[a]	1.2	1.4
Percent in #1 position	85%	70%	88%	80%

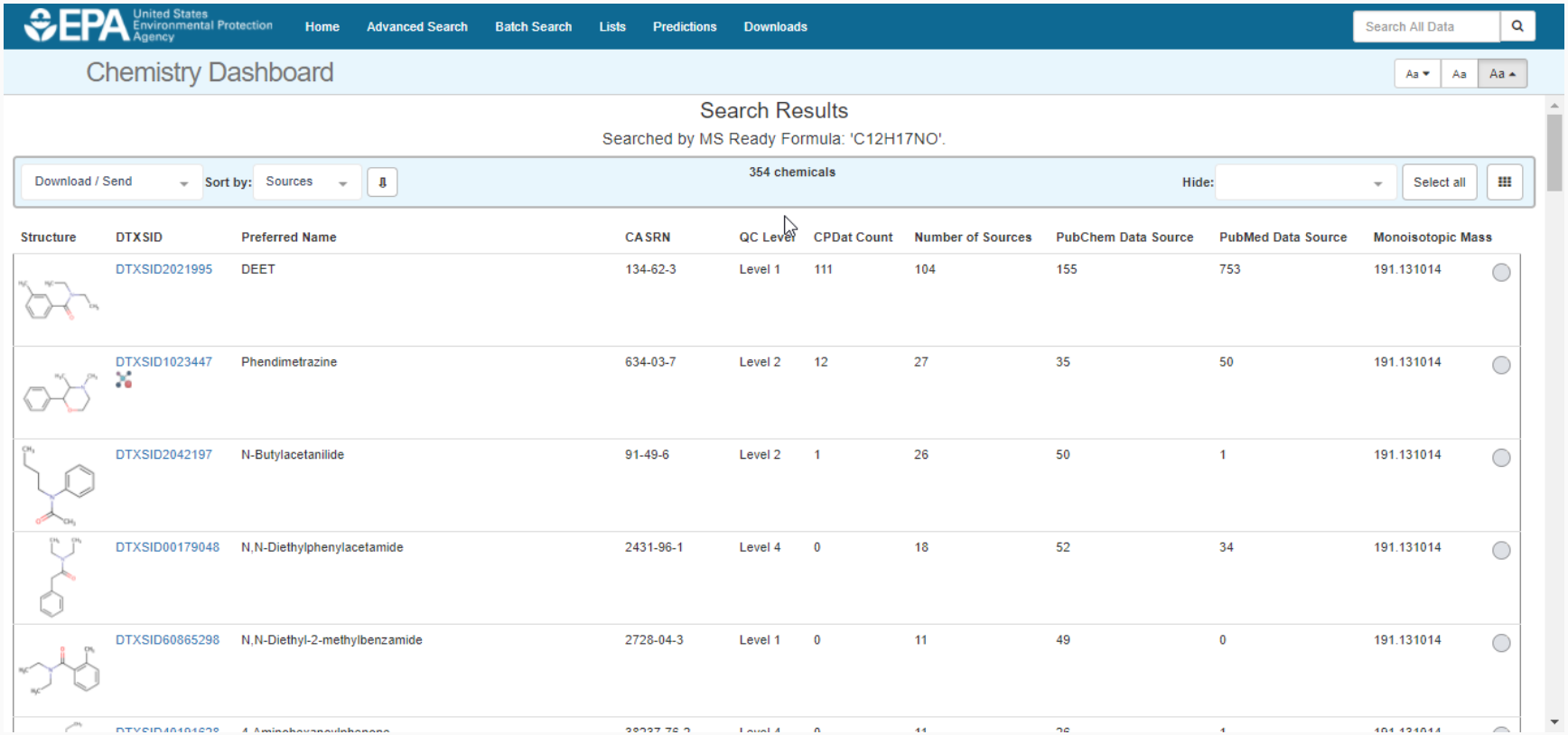
[a] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5

DOI: 10.1007/s00216-016-0139-z

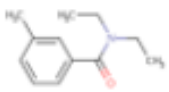
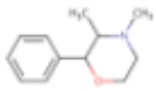
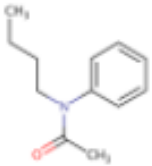
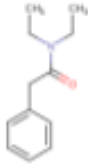
Additional Metadata Ranking

- US EPA CompTox Chemistry Dashboard Data Sources
- “CPDat” Consumer Product Database
- PubChem Data Source Count
- PubMed Reference Count

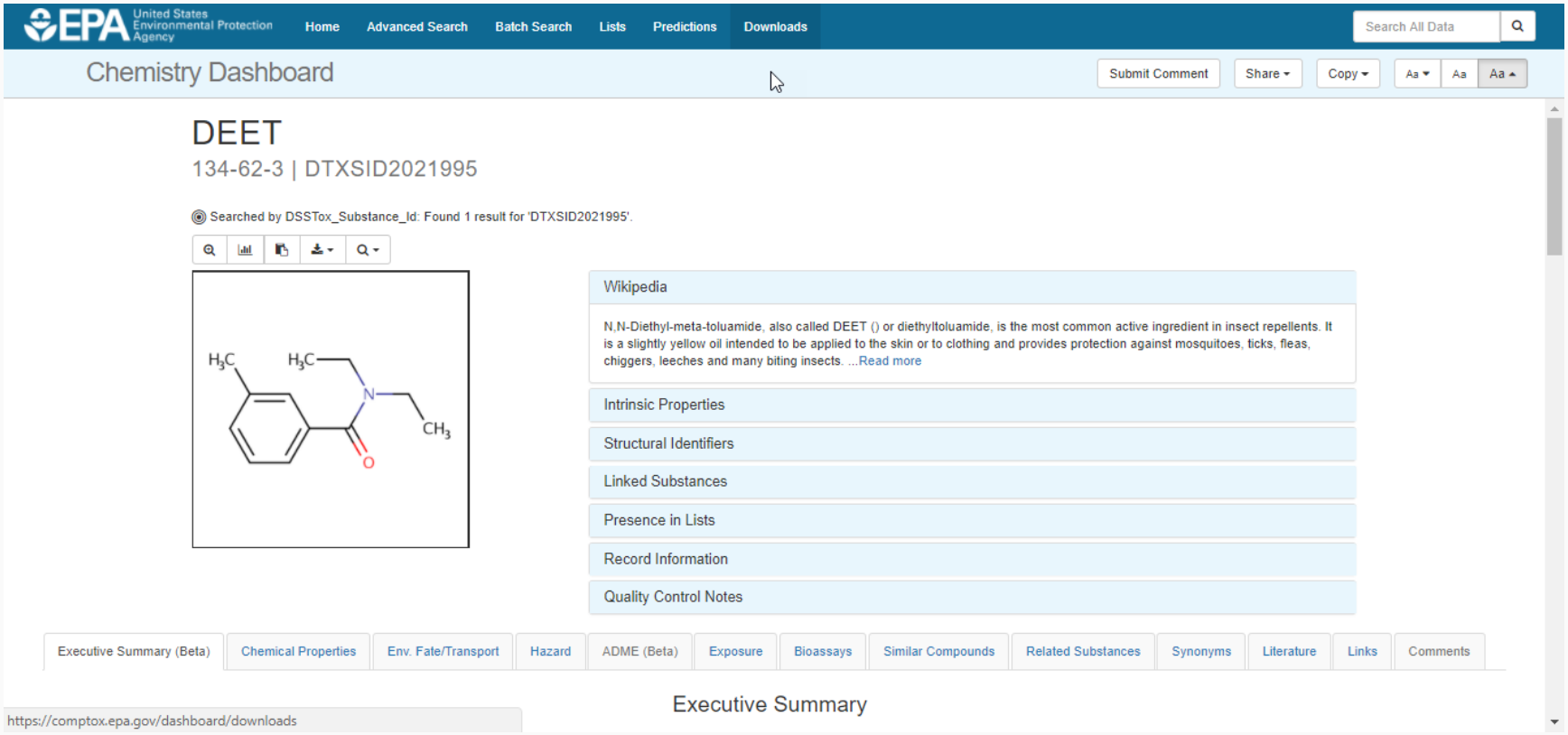
Additional Metadata Ranking C12H17NO: 354 Chemicals



Additional Metadata Ranking
C12H17NO: 354 Chemicals

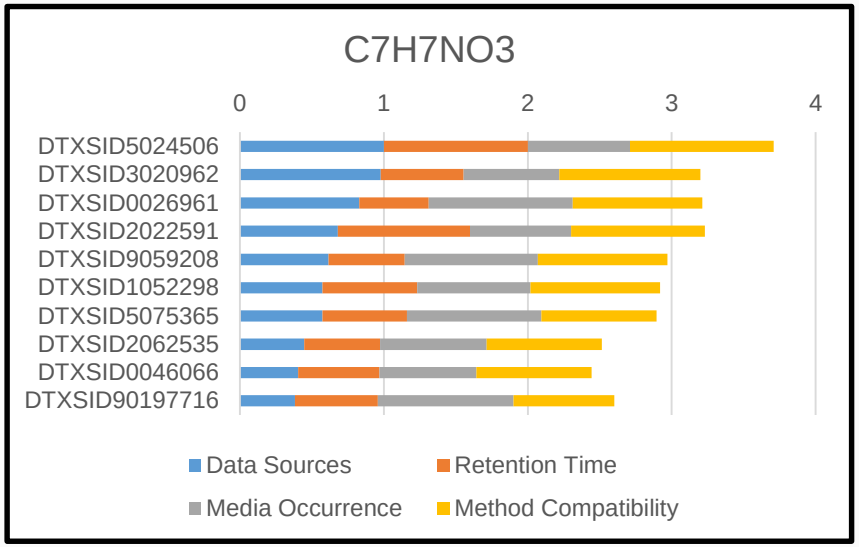
Structure	DTXSID	Preferred Name	CPDat Count	Number of Sources	PubChem Data Source	PubMed Data Source
	DTXSID2021995	DEET	111	104	155	753
	DTXSID1023447	Phendimetrazine	12	27	35	50
	DTXSID2042197	N-Butylacetanilide	1	26	50	1
	DTXSID00179048	N,N-Diethylphenylacetamide	0	18	52	34

Top Ranked Chemical



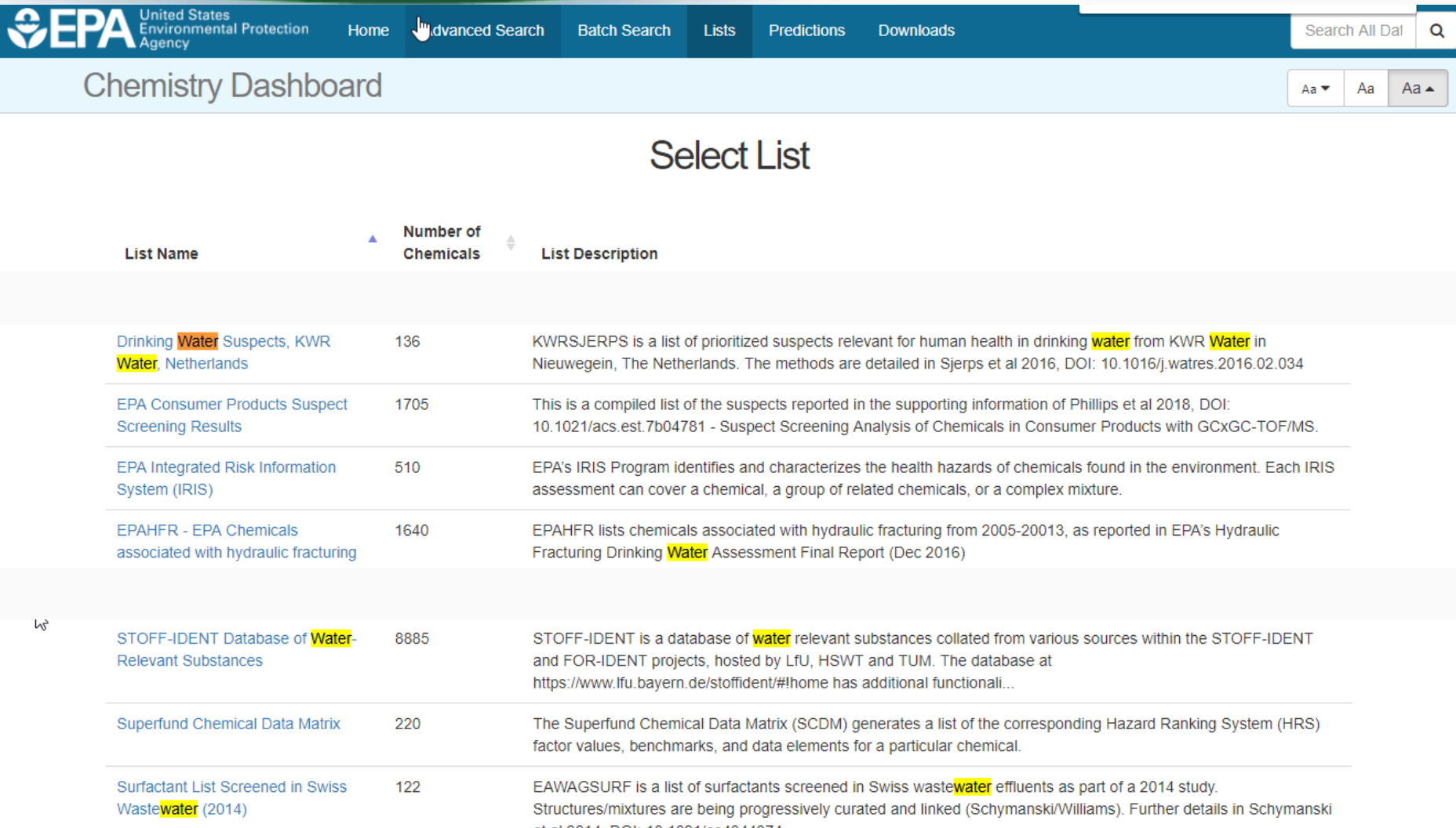
Additional data streams in development

- US EPA CompTox Chemistry Dashboard Data Sources
- “CPDat” Consumer Product Database
- PubChem Data Source Count
- PubMed Reference Count
- Retention Time Prediction $SC_{TOTAL} = SC_{DS} + SC_{PM} + SC_{RT} + SC_{MO} + \cdots$
- Predicted Environmental Media Occurrence
- Presence in Lists



“Chemicals Detected in Water”

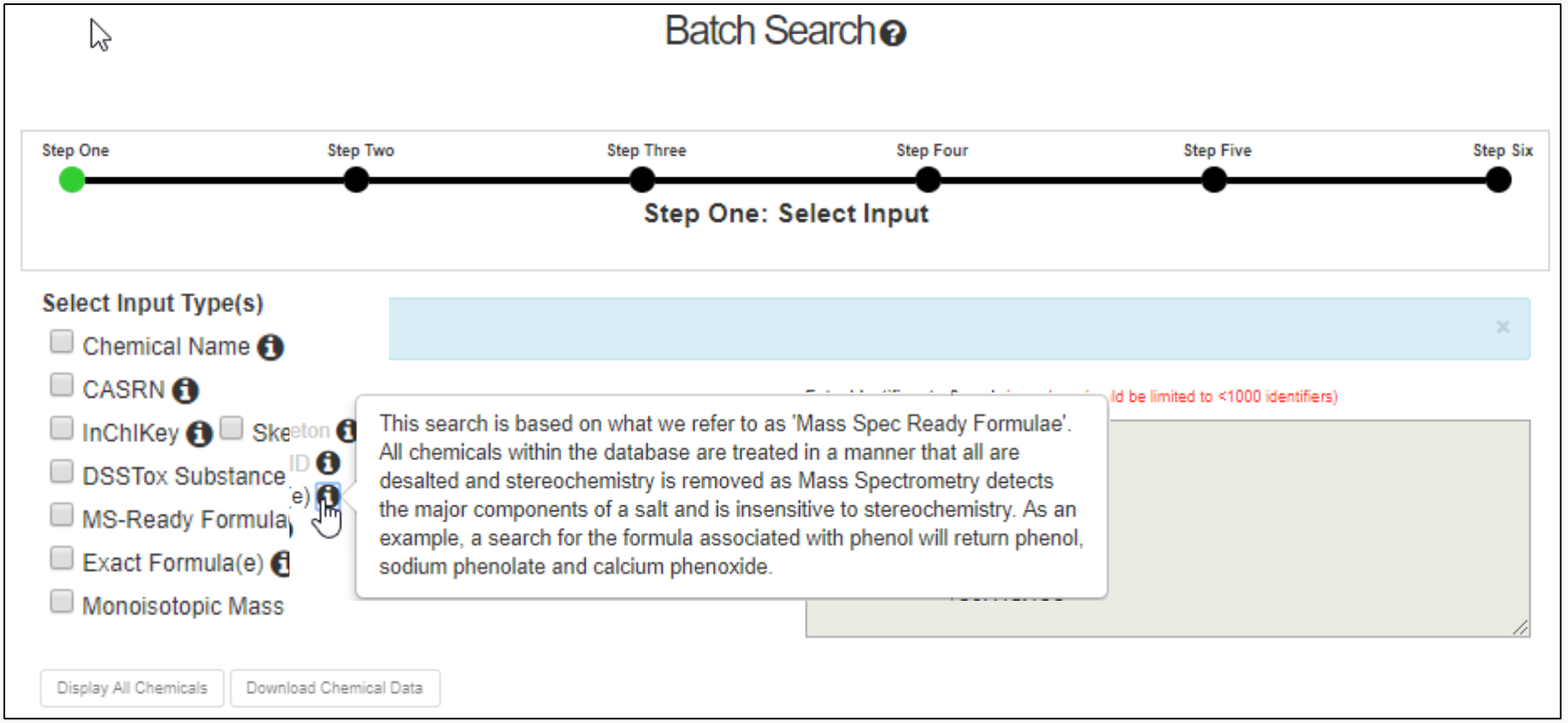
EPA United States Environmental Protection Agency



Dashboard for Structure ID

- Structure Identification using the dashboard
  - Formula/mass-based searching – 1 chemical at a time
  - Distilling structures into “MS-Ready form”
  - Ranking based on metadata
  - Batch searching of formulae and masses

Batch Search



Batch Search

Select Input Type(s)

- [ ] Chemical Name
- [ ] CASRN
- [ ] InChIKey
- [ ] Skeleton
- [ ] DSSTox Substance ID
- [x] MS-Ready Formula(e)
- [ ] Exact Formula(e)
- [ ] Monoisotopic Mass

Enter Identifiers to Search (searches should be limited to <1000 identifiers)

C6H12O3
C7H7N3
C8H11NO
C7H5NOS
C9H15NO
C11H12O
C9H8O3
C6H12O5
C9H15NO2

Metadata

- [ ] Curation Level Details
- [x] Data Sources
- [x] Assay Hit Count
- [ ] Include links to ACToR reports - SLOW! (BETA)
- [x] NHANES/Predicted Exposure
- [x] Include ToxVal Data Availability
- [x] Number of PubMed Articles
- [ ] Abstract Sifter Input File (Beta)
- [ ] MetFrag Input File(Beta)
- [x] IRIS
- [x] PPRTV
- [x] PubChem Data Sources
- [ ] ToxPrint fingerprints

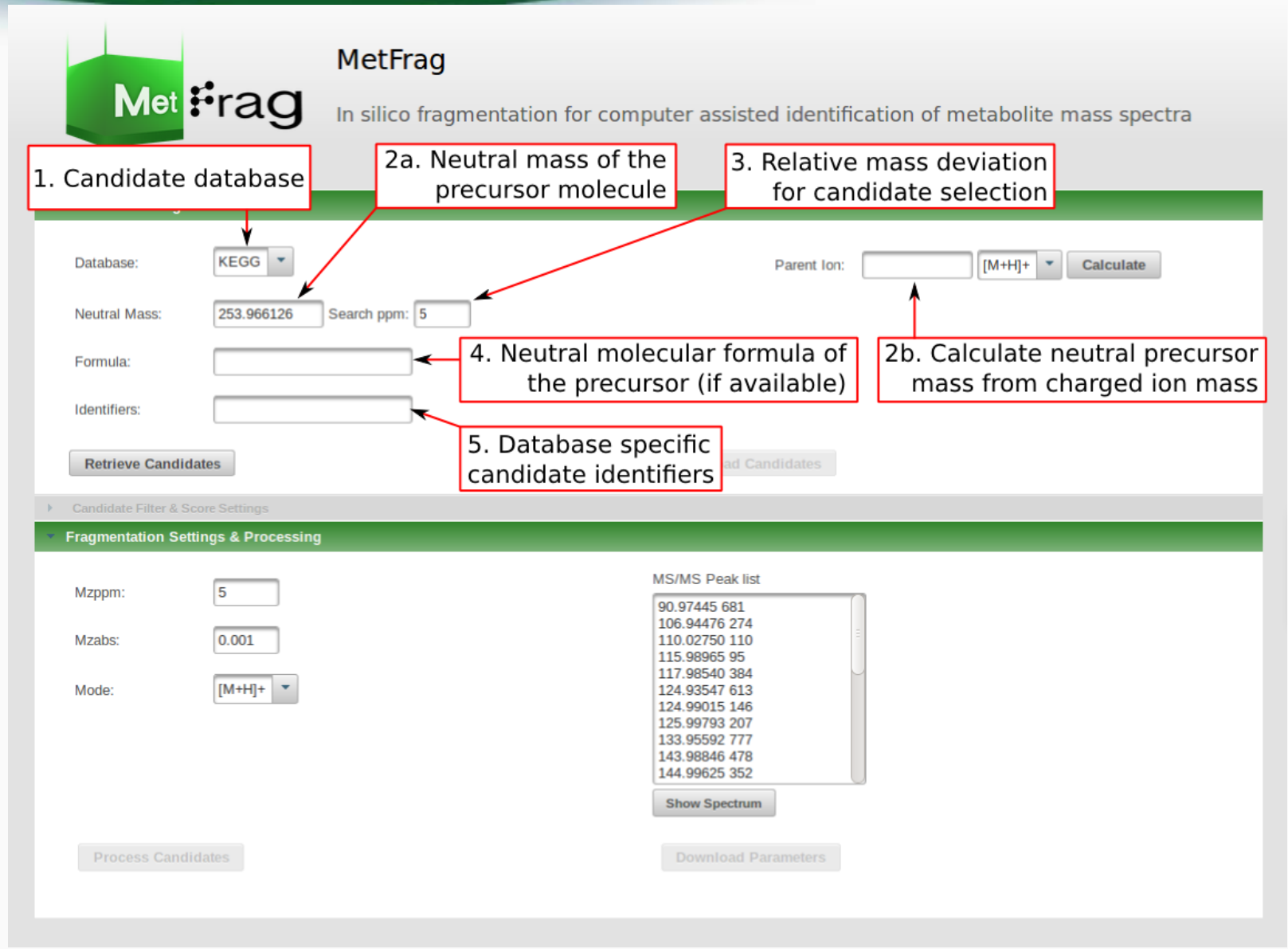
- [ ] NIOSH IDLH Values
- [ ] NIOSH International Chemical Safety Cards
- [ ] NIOSH Pocket Guide to Chemical Hazards
- [ ] NIOSH Skin Notation Profiles
- [ ] NORMAN Collaborative Trial 2015 Targets and Suspects
- [ ] Norman Network PFAS (KEMI Report)
- [ ] NORMAN Network Priority List
- [ ] NormaNEWS: Norman Early Warning System
- [ ] PFAS list provided by X.Trier et al
- [ ] Pharmaceutical List with EU, Swiss and US Consumption Data
- [ ] Provisional Peer Reviewed Toxicity Values
- [ ] Stockholm Convention on Organic Pollutants
- [x] STOFF-IDENT Database of Water-Relevant Substances
- [ ] Superfund Chemical Data Matrix
- [x] Surfactant List Screened in Swiss Wastewater (2014)

Excel Output

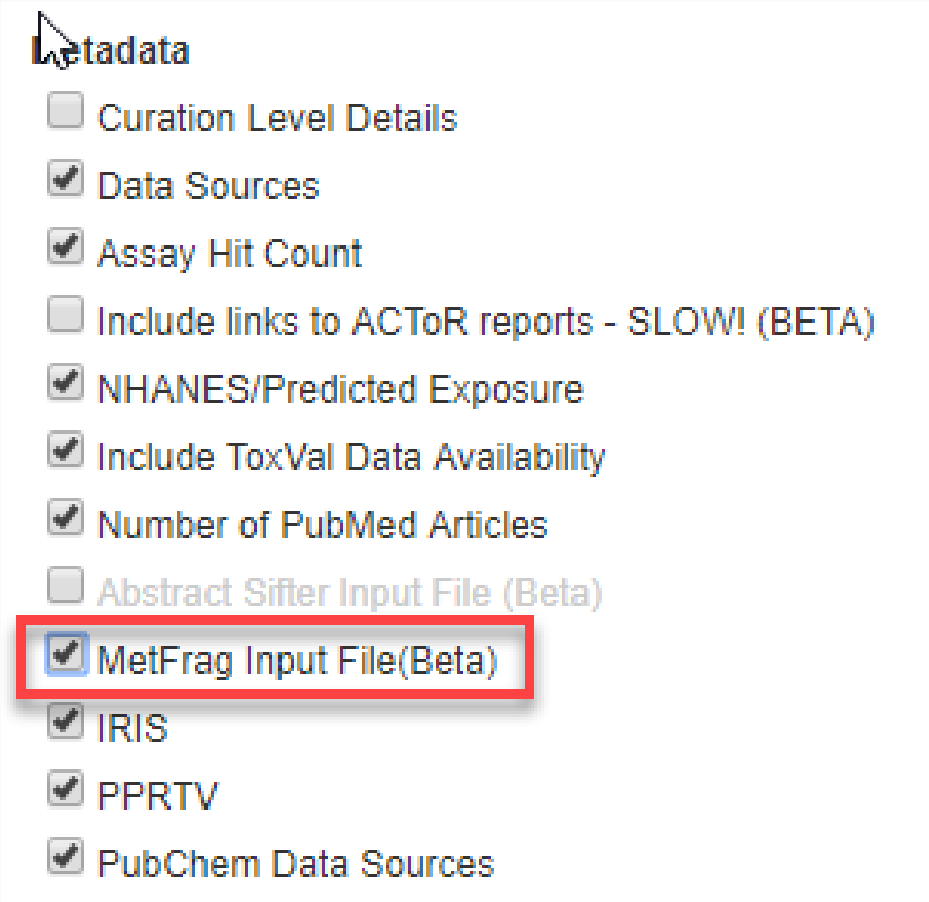
INPUT	FOUND_BY	DTXCID_IN	DATA_SOU	TOXVAL_D	TOXCAST_	TOXCAST_	NUMBER_C	PUBCHEM_	STO
C6H12O3	MS Ready I	DTXCID701	91	Y	0.36	2/562	24	83	Y
C6H12O3	MS Ready I	DTXCID003	67	Y	0.36	1/276	376	80	Y
C6H12O3	MS Ready I	DTXCID106	65	Y	4.42	5/113	6	77	Y
C6H12O3	MS Ready I	DTXCID105	45	Y	0.0	0/163	3	94	-
C6H12O3	MS Ready I	DTXCID901	38	Y	-	-	14	110	Y
C6H12O3	MS Ready I	DTXCID402	34	Y	0.0	0/113	-	53	Y
C6H12O3	MS Ready I	DTXCID202	31	Y	-	-	-	36	Y
C6H12O3	MS Ready I	DTXCID202	30	-	2.54	7/276	-	54	-
C6H12O3	MS Ready I	DTXCID109	26	Y	-	-	-	46	-
C6H12O3	MS Ready I	DTXCID202	24	Y	0.0	0/113	-	47	-
C6H12O3	MS Ready I	DTXCID303	22	Y	-	-	-	89	-
C6H12O3	MS Ready I	DTXCID302	20	Y	-	-	2	25	Y
C6H12O3	MS Ready I	DTXCID407	19	Y	-	-	12	62	-
C6H12O3	MS Ready I	DTXCID704	17	Y	-	-	-	64	-
C6H12O3	MS Ready I	DTXCID704	16	Y	-	-	3	49	-

Batch Search Integration to MetFrag

http://c-ruttkies.github.io/MetFrag/projects/metfragweb/

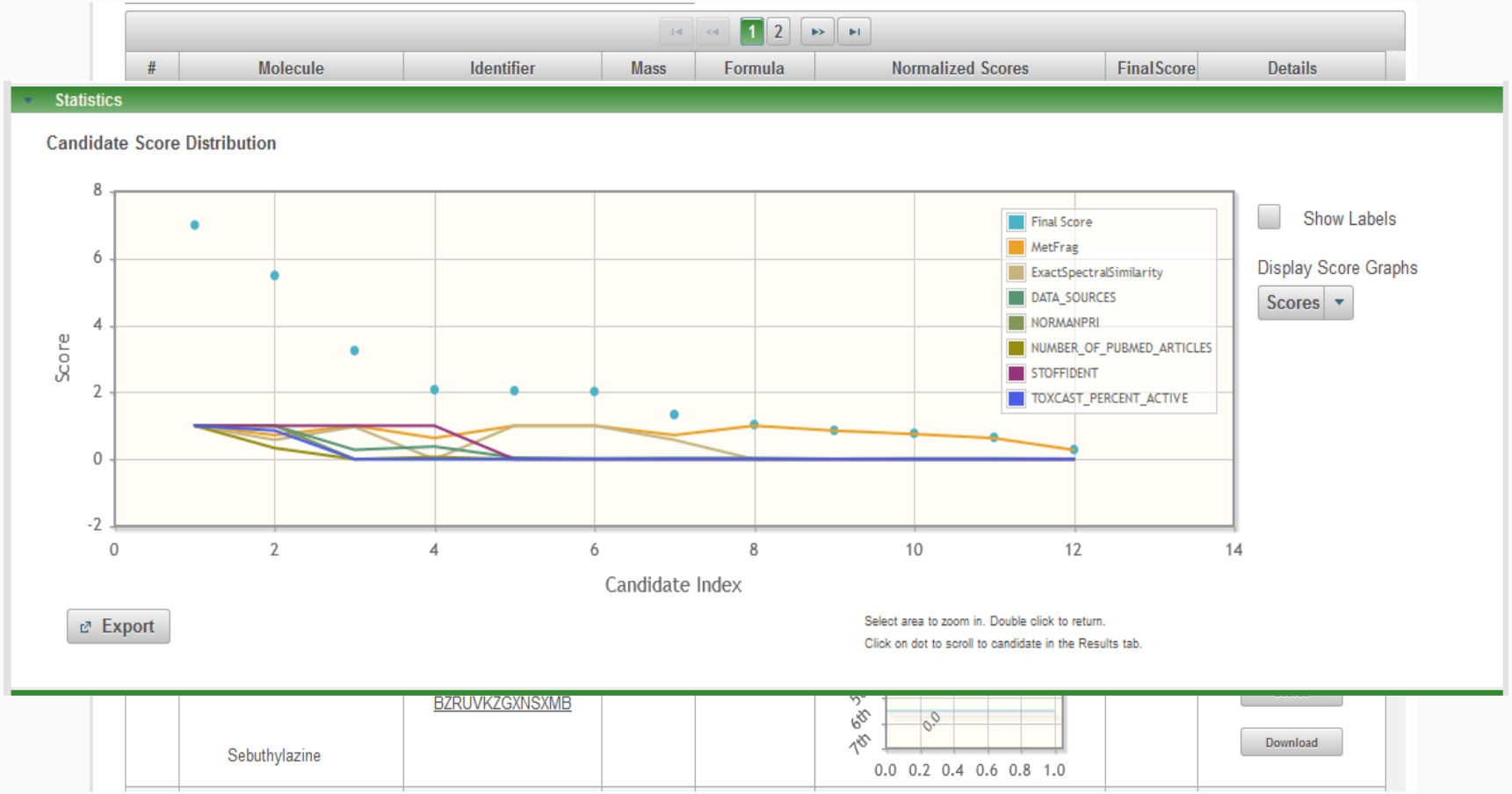


MetFrag Input File



Batch Search Integration to MetFrag

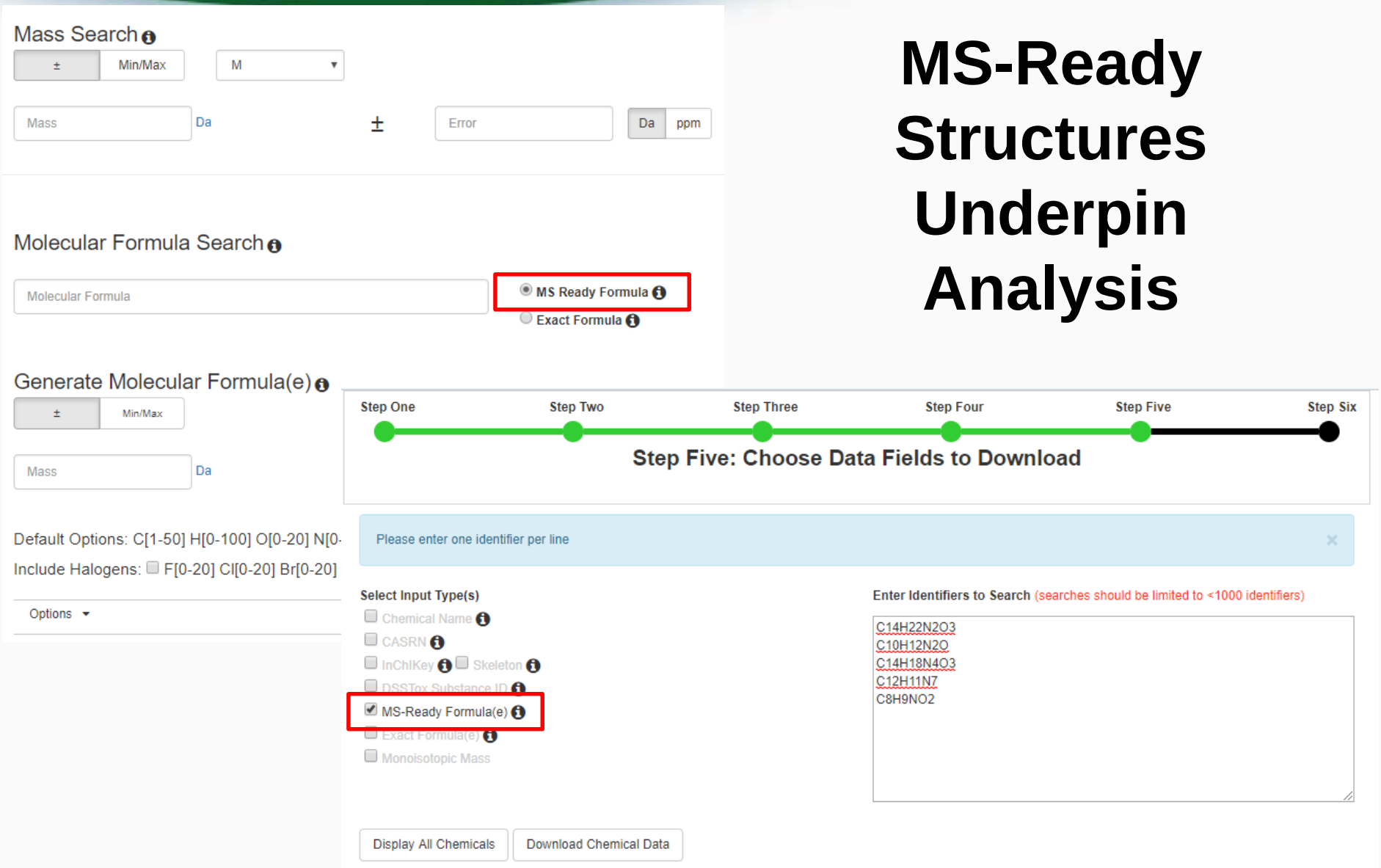
http://c-ruttkies.github.io/MetFrag/projects/metfragweb/



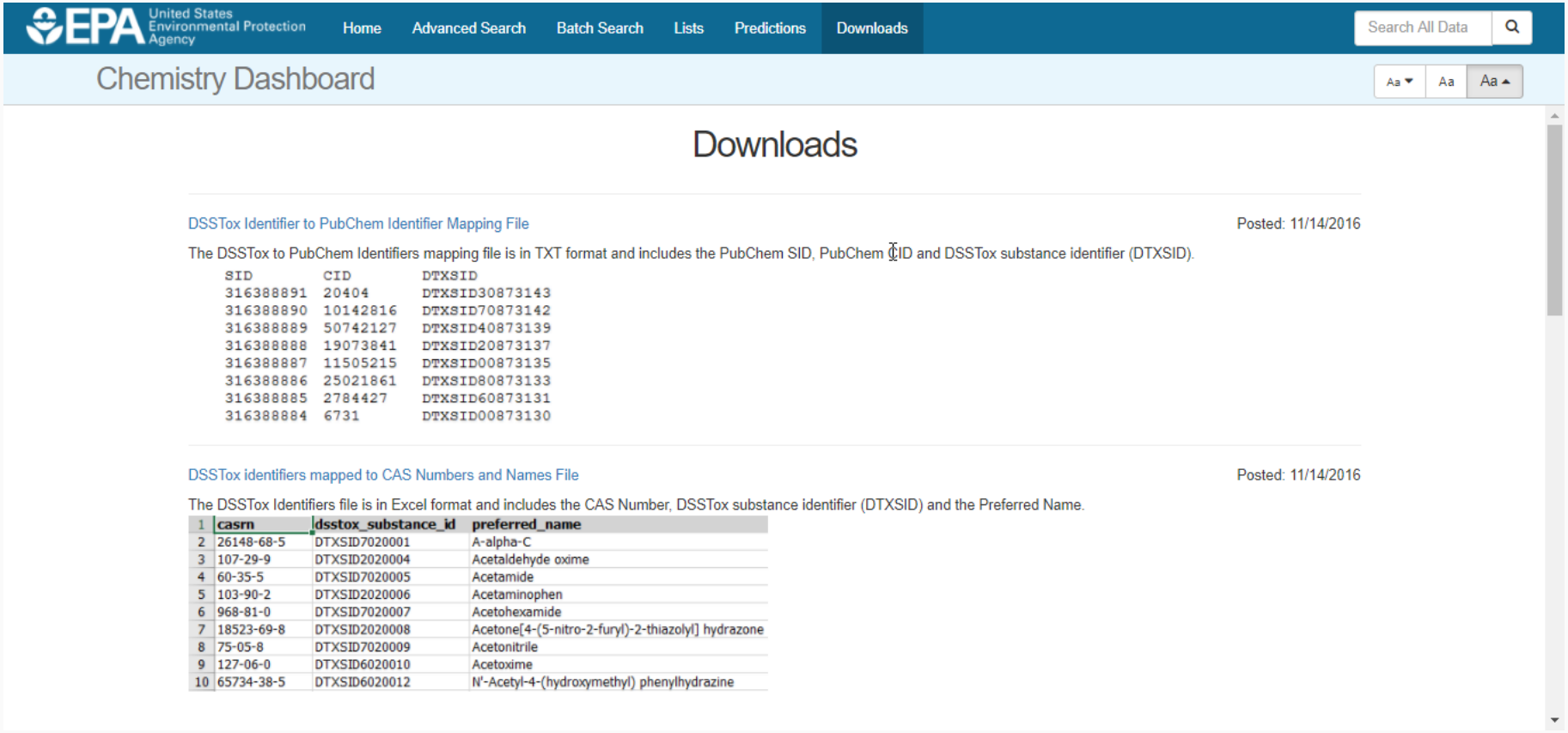
The Dashboard to Support MS-Analysis

EPA United States Environmental Protection Agency

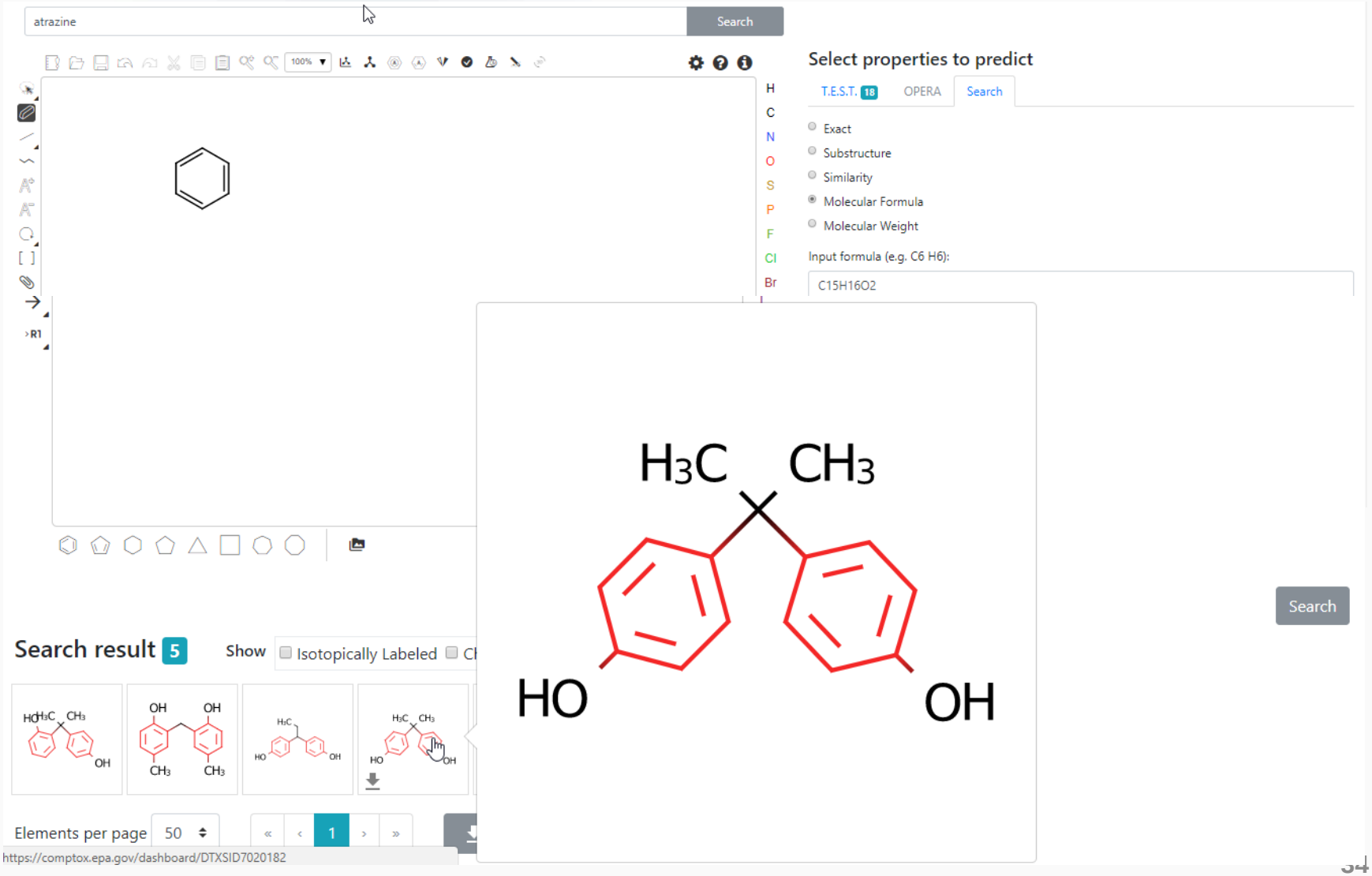
MS-Ready Structures Underpin Analysis



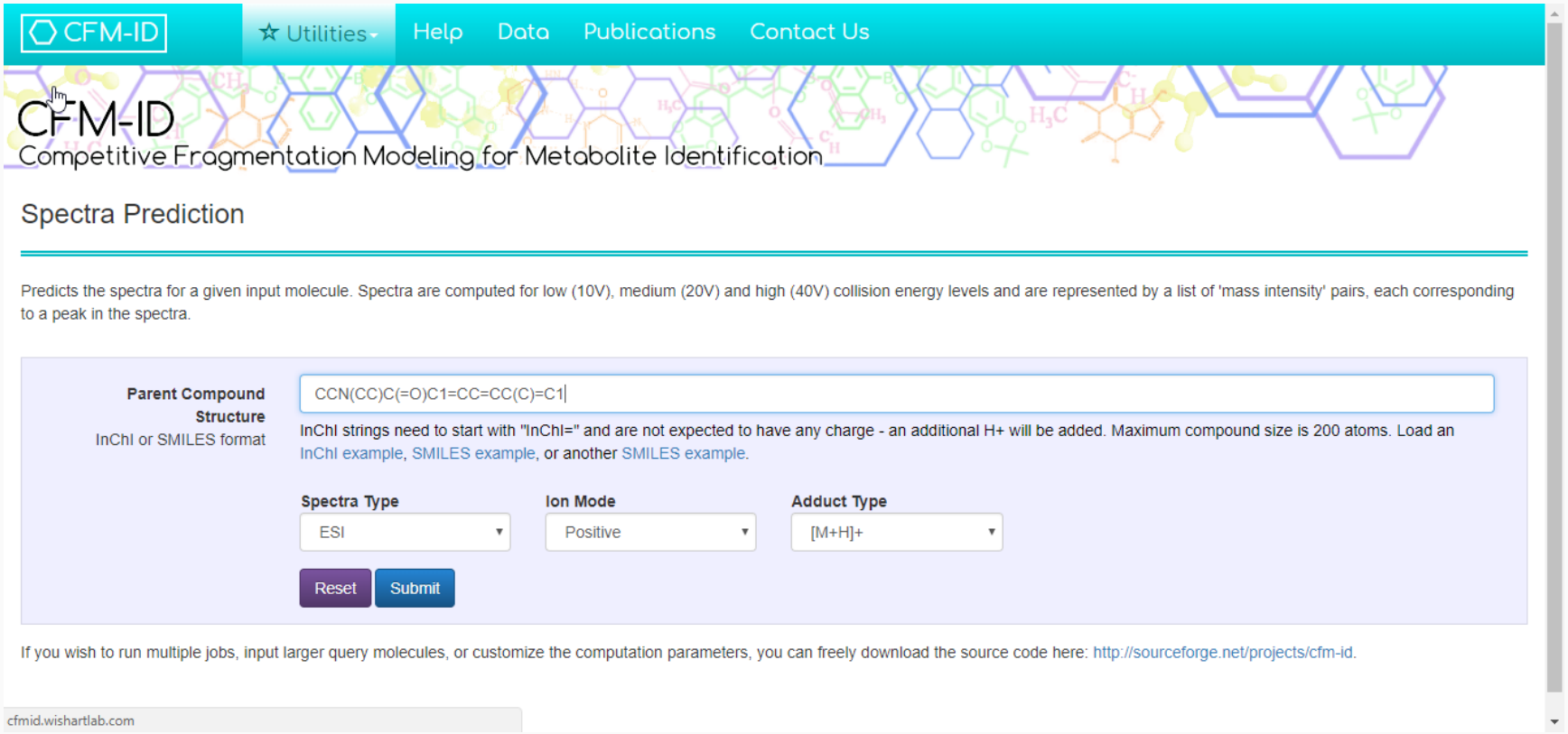
Downloadable Data



Future Work: Combined Substructure/Formula Searching

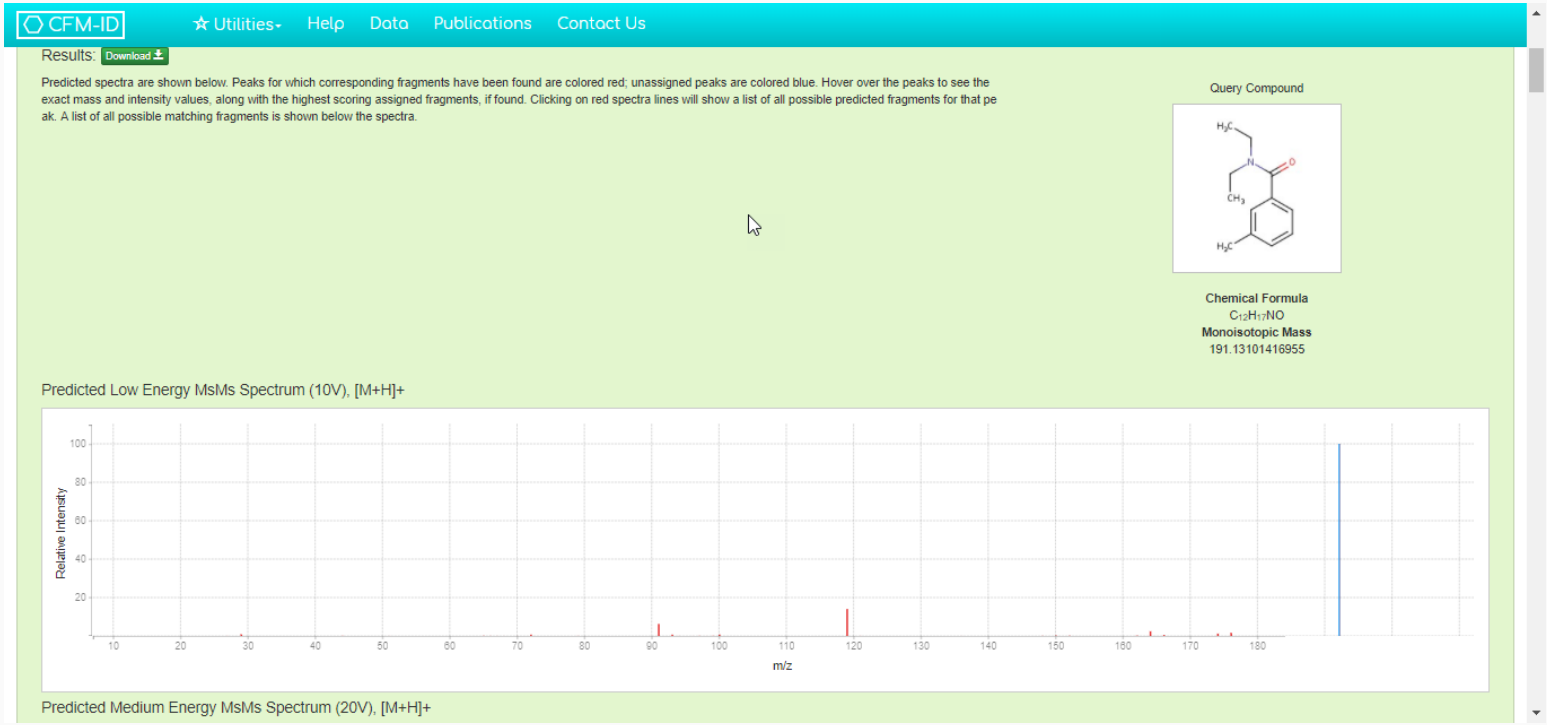


Future Work: Searching Against Predicted Spectra



Future Work: Searching Against Predicted Spectra

- CFM-ID predicted spectra generated for 700,000 chemicals
  - Positive ion, Negative ion, Electron Impact
  - Three energies



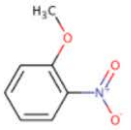
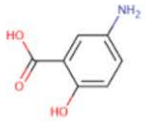
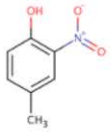
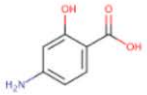
Future Work
Scoring scheme into results

EPA United States Environmental Protection Agency

Search Results

Searched by molecular formula: Found 188 results.

Download as: TSV Excel SDF

ID	Structure	Preferred Name	CAS-RN	QC Level	CPCat Count	Number of S...	PubChem D...	Monoisotopic Mass	Identification Score
DTXSID3020962 ToxCast™		1-Methoxy-2-nitrobenzene	91-23-6	Level 2: Expert curate...	0	58	111	153.042593	3.6
DTXSID5024506 ToxCast™		5-Aminosalicylic acid	89-57-6	Level 2: Expert curate...	2	57	193	153.042593	3.3
DTXSID0026961 ToxCast™		4-Methyl-2-nitrophenol	119-33-5	Level 2: Expert curate...	0	46	93	153.042593	3.3
DTXSID2022591		Aminosalicylic acid	65-49-6	Level 1: Expert curate...	0	40	177	153.042593	2.5

Data Sources PubMed Ct Media Occurr CPDat Ct

$$SC_{TOTAL} = SC_{DS} + SC_{PM} + SC_{RT} + SC_{MO} + \cdots$$

Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- High quality curated data and rich metadata facilitates mass spec analysis
- “MS-Ready” processed data enables structure identification

Acknowledgments

- The CompTox Chemistry Dashboard team
- NERL colleagues:
  - Jon Sobus, Elin Ulrich, Mark Strynar, Seth Newton (NTA Analysis)
  - Katherine Phillips, Kathie Dionisio, Kristin Isaacs (Consumer Products Database)
- Emma Schymanski – Luxembourg Center for Systems Biomedicine (MS-ready/NTA)

Contact

Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821