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

The CompTox Dashboard: Data and Tools to Support Chemical and Environmental Risk Assessment and the ENTACT project

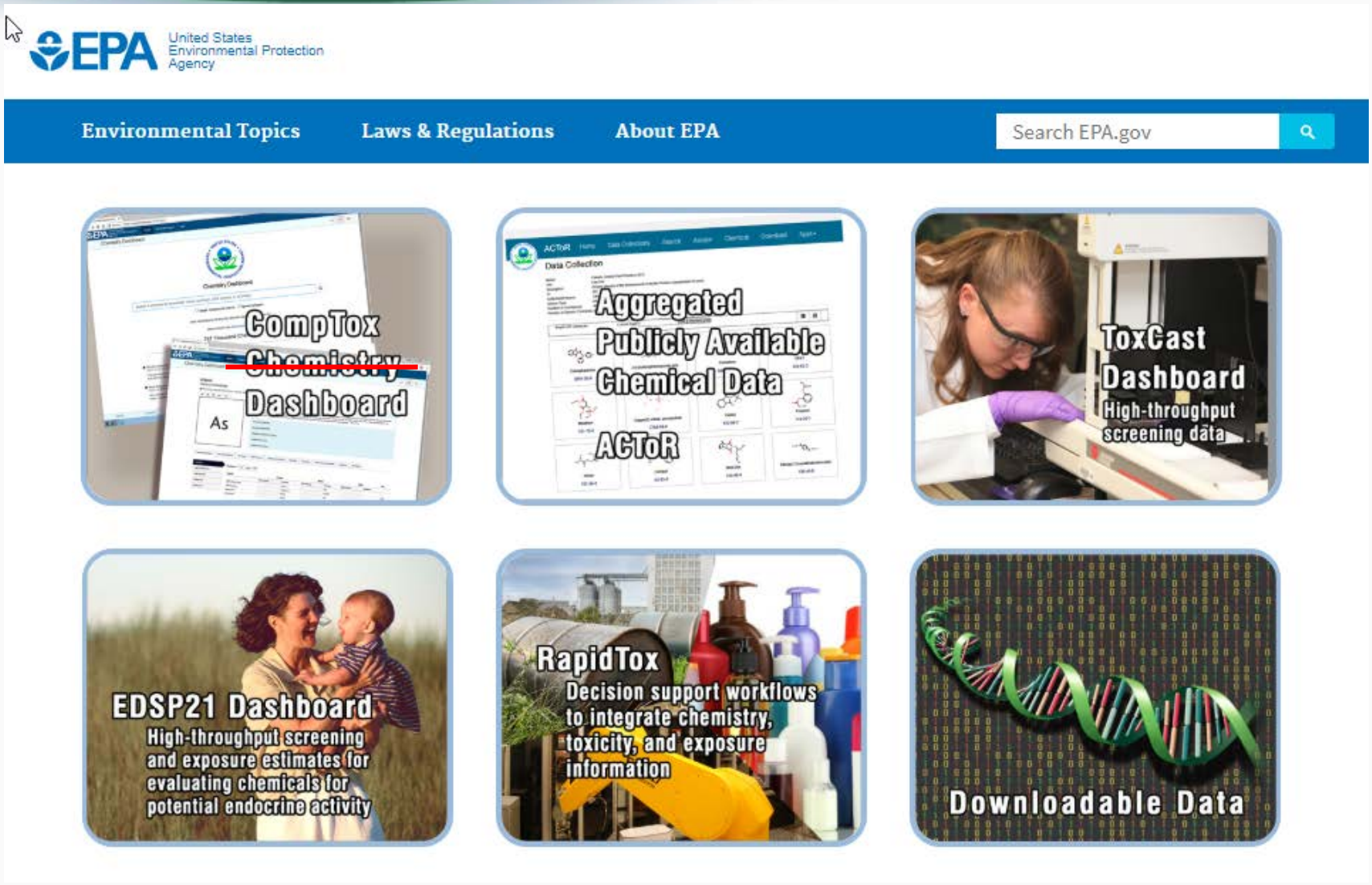
Antony Williams[1], Chris Grulke[1], Andrew McEachran[2], Emma Schymanski[3], Jon Sobus[4], Elin Ulrich[4], Ann Richard[1], Jeremy Dunne[1] and Jeff Edwards[1]

1) *National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC*
2) *Oak Ridge Institute of Science and Education (ORISE) Research Participant, RTP, NC*
3) *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, Luxembourg*
4) *National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC*

August 2018
ACS Fall Meeting, Boston

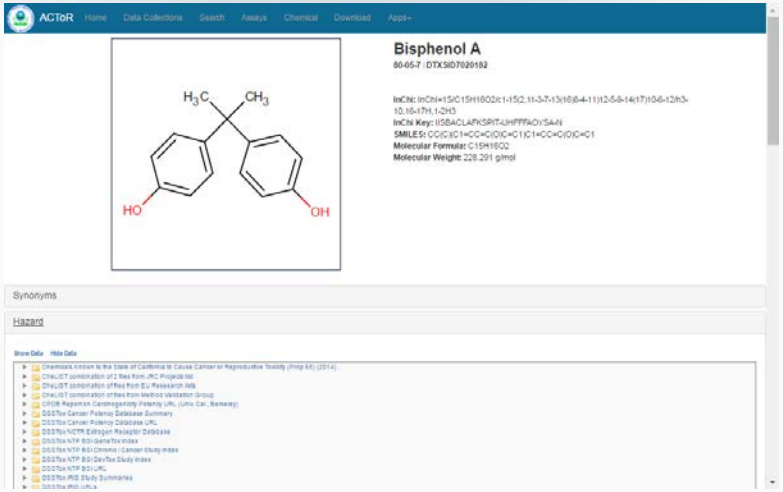
Coming soon to a screen near you: The CompTox Portal

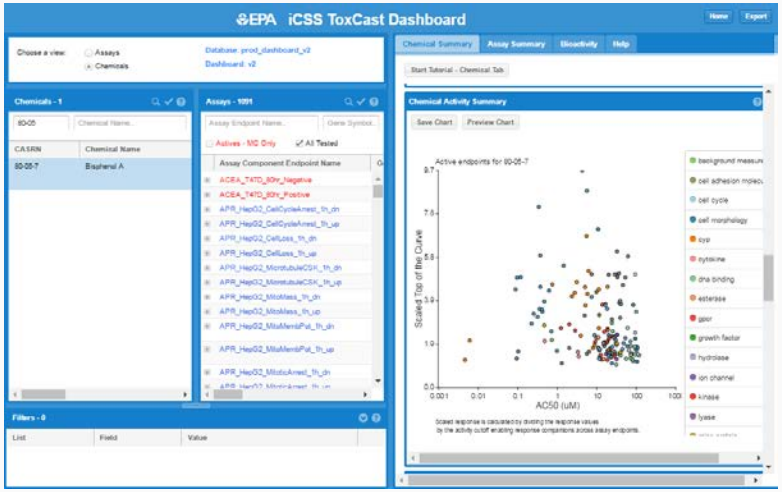
EPA United States Environmental Protection Agency

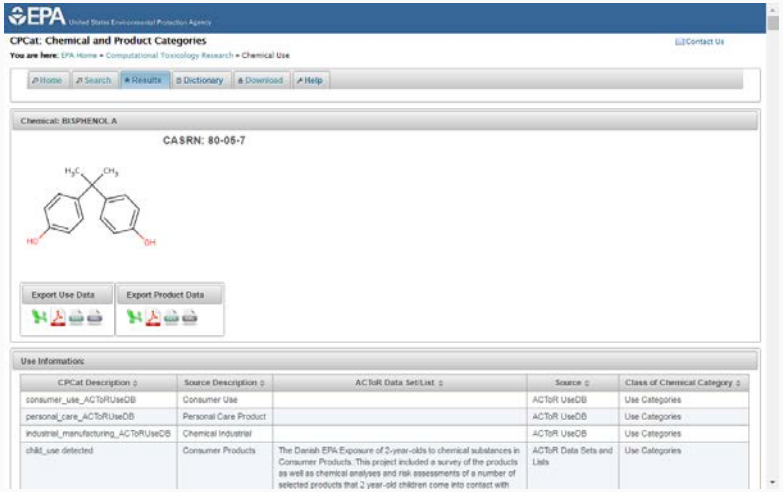


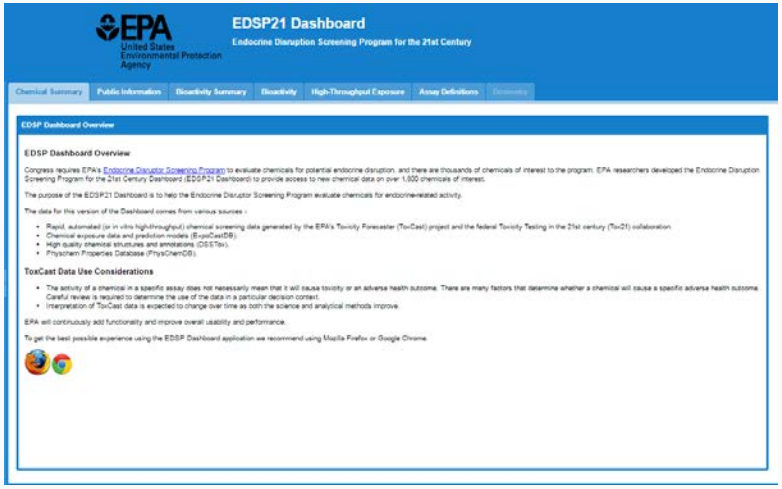
Some History

Early Dashboard Applications









The concept of DSSTox

- DSSTox was a concept in 2002
- Today it underpins the Dashboard

ELSEVIER

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

Volume 499, Issue 1, 29 January 2002, Pages 27-52

Mutation Research Frontiers

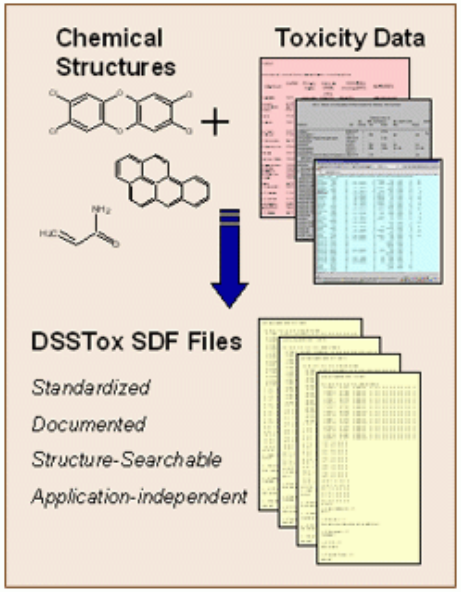
Distributed structure-searchable toxicity (DSSTox) public database network: a proposal

Ann M. Richard [a], ClarLynda R. Williams [a, b]

DSSTox Datafiles 2004

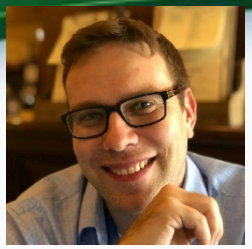
GOAL: Link chemical structures to data for SAR

- First release of data files in 2004
- Focused on high impact sets of data
  - Carcinogenic Potency Database
  - Drinking water disinfection by-products
  - EPA's Integrated Risk Information System
  - FDA's Maximum Daily Dose dataset
  - EPA's Fat Head Minnow Toxicity dataset
  - etc...



- Managed chemical registration for ToxCast/Tox21 chemicals (our HTS screening research)

And then came Chris Grulke…
ChemReg registration for DSSTox



View/Edit a Single Record | Structure Search | Browse/Curate Records | Export DSSTox | Chemotypes | Manage Chemical Lists | Manage Property Data | Add Deleted Casrns

Preferred Name matched null
You are viewing the record associated with
DTXSID80198757
CASRN: 62885-41-0

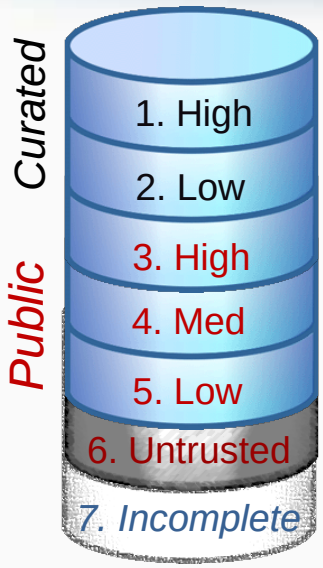
4-Hydroxy-3-methox

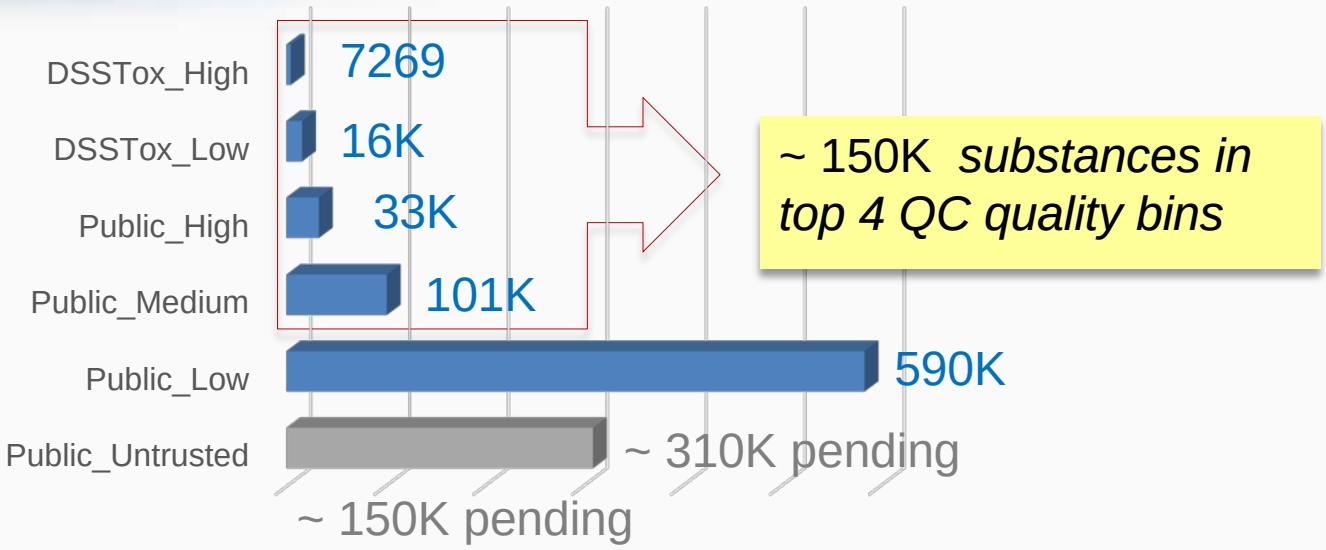
Valid license cannot be found

Calculate from Structure

Substance_ID:	DTXSID80198757	Compound_ID:	DTXCID40121248
CAS:	62885-41-0	Chemical Shown:	Tested Chemical
Name:	4-Hydroxy-3-methoxypyridine	Private Notes:	
Substance Type:	Single Compound		
QC Level:	DSSTox_High		
Data Source:	STN(DSSTox)	Source of CAS-Compound:	STN(DSSTox)
QC Notes:	CAS [50700-60-2] assigned by DSSTox to pyridin-one tautomer form, which resolves to hydroxy form thru InChI	Double Stereo:	None
		Chiral Stereo:	None
		Chemical Form:	Organic
		Organic Form:	Parent

Distribution of curated data





QC Levels

DSSTox_High:	Hand curated and validated
DSSTox_Low:	Hand curated and confirmed using multiple public sources
Public_High:	Extracted from EPA SRS and confirmed to have no conflicts in ChemID and PubChem
Public_Medium:	Extracted from ChemID and confirmed to have no conflicts in PubChem
Public_Low:	Extracted from ACToR or PubChem
Public_Untrusted:	Postulated, but found to have conflicts in public sources

Internal List Conflicts: ChemReg List Curation

EPA United States Environmental Protection Agency

View/Edit a Single Record | Structure Search | Browse/Curate Records | Export DSSTox | Chemotypes | **Manage Chemical Lists** | Manage Property Data | Add Deleted Casrns

Welcome cgrulke

Editing Listname: ALANWOOD

Internal Check Results

Description	Records
All	1718
Duplicate CASRN	2
Duplicate STRUCTURE,Duplicate STRUCTURE_INCHIKEY	2
Invalid casrn	2
NONE	1712

List: ALANWOOD

(1 of 69) 1 2 3 4 5 6 7 8 9 10 25

Record ID	External ID	1st Identifier	Warning
DTXRID303936283	(3-ethoxypropyl)mercury bromide	6012-84-6	NONE

Identifier
(3-ethoxypropyl)mercury bromide
6012-84-6
InChI=1S/C5H11O.BrH.Hg/c1-3-5-6-4-2;;/h1,3-5H2,2H3;1H;/q;;+1/p-1
BWUIOGHGUVLNSX-UHFFFAOYSA-M
BWUIOGHGUVLNSX-UHFFFAOYSA-M

DTXRID003936284	1,2-dichloropropane	78-87-5	NONE
DTXRID703936285	1,3-dichloropropene	542-75-6	NONE
DTXRID403936286	1-methylcyclopropene	3100-04-7	NONE
DTXRID103936287	1-naphthol	90-15-3	NONE
DTXRID803936288	2-(octylthio)ethanol	3547-33-9	NONE
DTXRID503936289	2,3,5-tri-iodobenzoic acid	88-82-4	NONE
DTXRID803936290	2,3,6-TBA	50-31-7	NONE
DTXRID503936291	2,4,5-T	93-76-5	NONE
DTXRID203936292	2,4,5-TB	93-80-1	NONE
DTXRID903936293	2,4-D	94-75-7	NONE

Chemical Collections (e.g. PCBs)

EPA United States Environmental Protection Agency

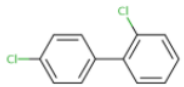
Successor Substances (209)

	CAS-RN	Relationship	Source
	32774-16-6	is a Representative Isome	
	2051-60-7	is a Representative Isome	
	2051-61-8	is a Representative Isome	
	2051-62-9	is a Representative Isome	
	13029-08-8	is a Representative Isome	
	16605-91-7	is a Representative Isome	
	25569-80-6	is a Representative Isome	
	33284-50-3	is a Representative Isome	
	34883-43-7	is a Representative Isome	
	34883-39-1	is a Representative Isome	
	33146-45-1	is a Representative Isome	
	2050-67-1	is a Representative Isome	
	2974-92-7	is a Representative Isome	
	2974-90-5	is a Representative Isome	
	34883-41-5	is a Representative Isomer of this	Public
	2050-68-2	is a Representative Isomer of this	Public

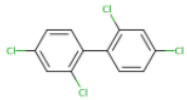
Related Chemicals

Found 209 chemicals

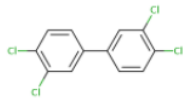
Download as: TSV | Excel | SDF



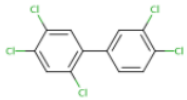
2,4'-Dichlorobiphenyl
34883-43-7



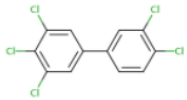
2,2',4,4'-Tetrachlorobiphenyl
2437-79-8



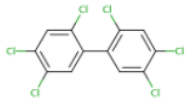
3,3',4,4'-Tetrachlorobinphenyl
32598-13-3



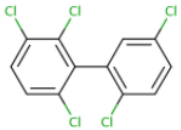
2,3',4,4',5-Pentachlorobiphenyl
31508-00-6



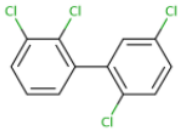
3,3',4,4',5-Pentachlorobiphenyl
57465-28-8



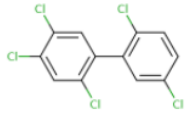
2,2',4,4',5,5'-Hexachlorobiphenyl
35065-27-1



2,2',3,5',6-Pentachlorobiphenyl
38379-99-6



2,2',3,5'-tetrachlorobiphenyl
41464-39-5



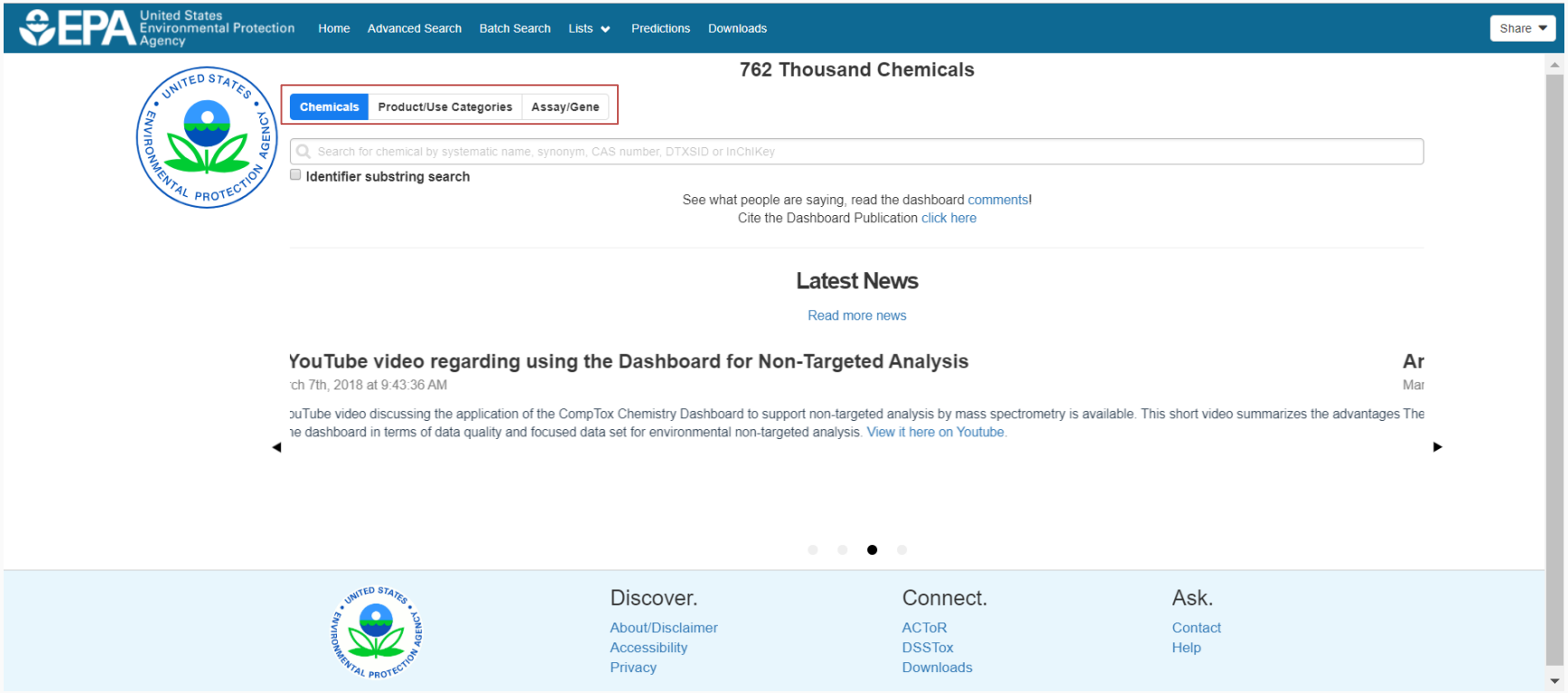
2,2,4,5,5'-Pentochlorobiphenyl
37680-73-2

The CompTox Dashboard
General Overview and **New Features 3.0**

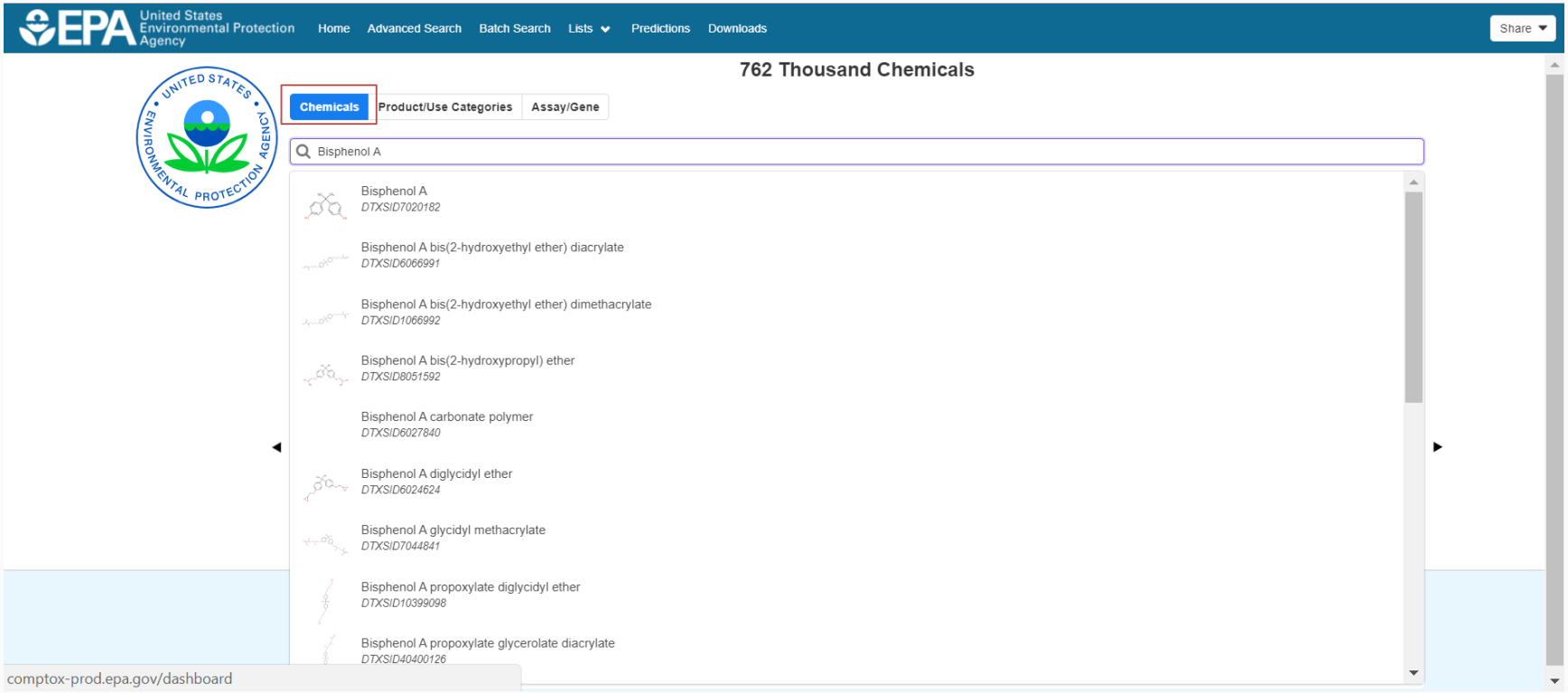
- Version 3.0 release, August 2018:
  - New user interface design
  - ~762,000 chemicals with related property data
  - New searches – product and use categories, assays and genes
  - More data – physchem, hazard data, bioactivity curves
  - Expanded number of chemical lists to review
  - Enhanced batch search capabilities
  - New chemical lists added (including PFAS chemical lists)
  - Handling of UVCB chemicals – many Markush

CompTox Dashboard

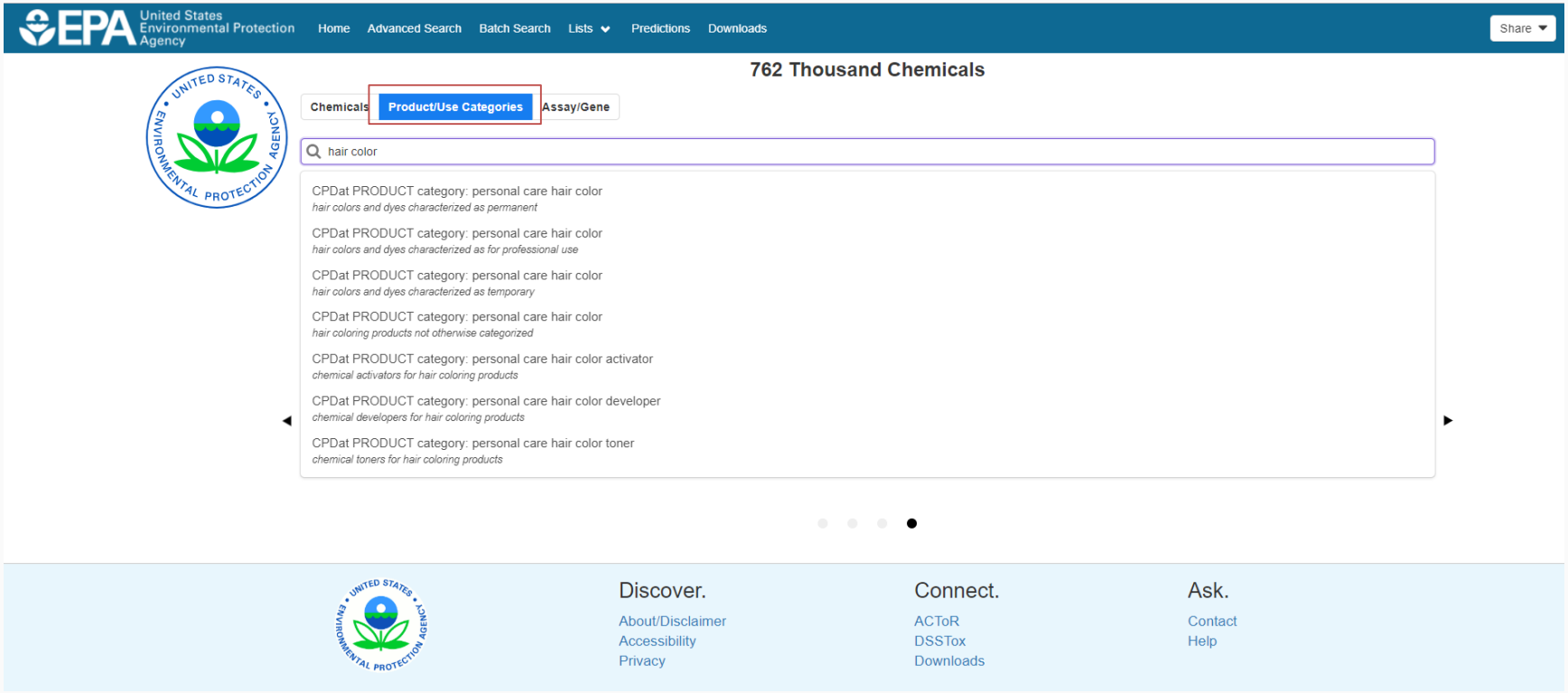
https://comptox.epa.gov/dashboard



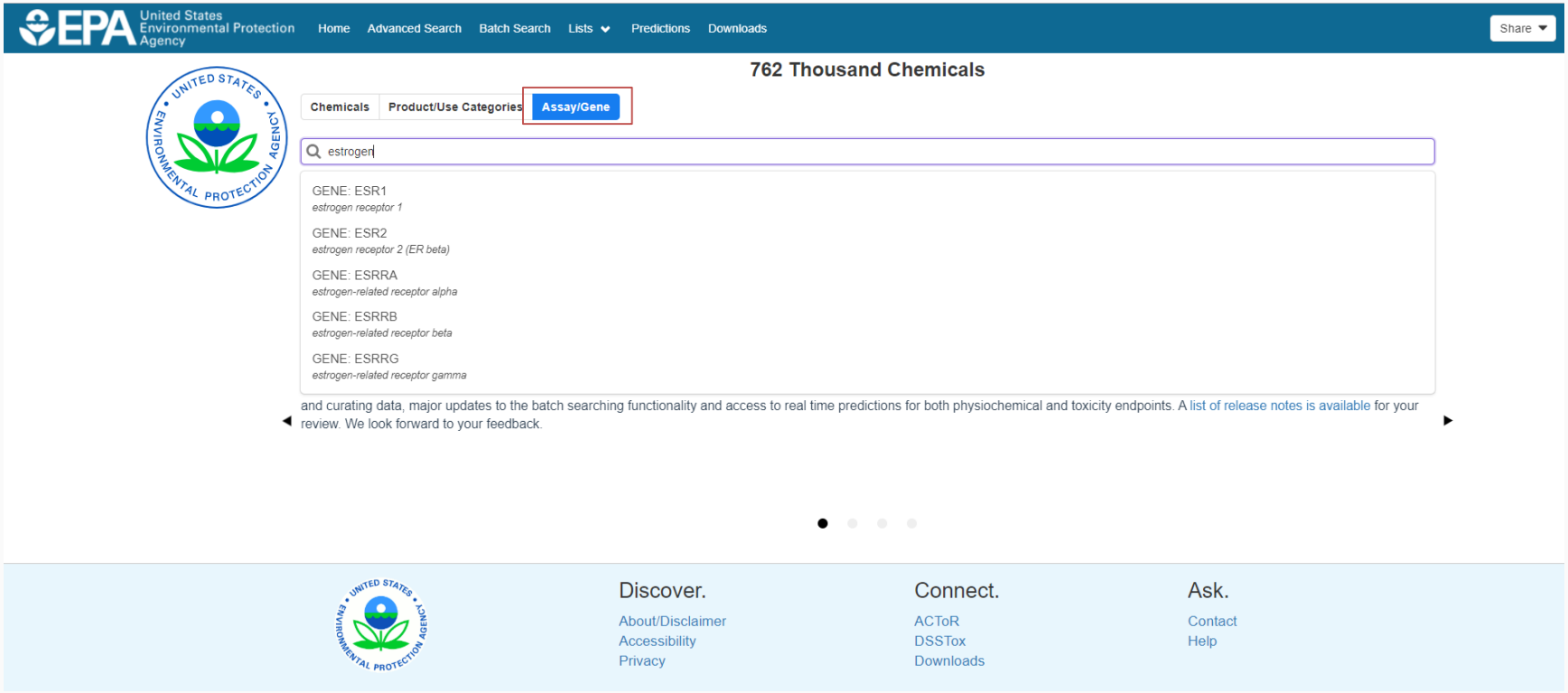
CompTox Dashboard Chemicals



CompTox Dashboard
Products and Use Categories



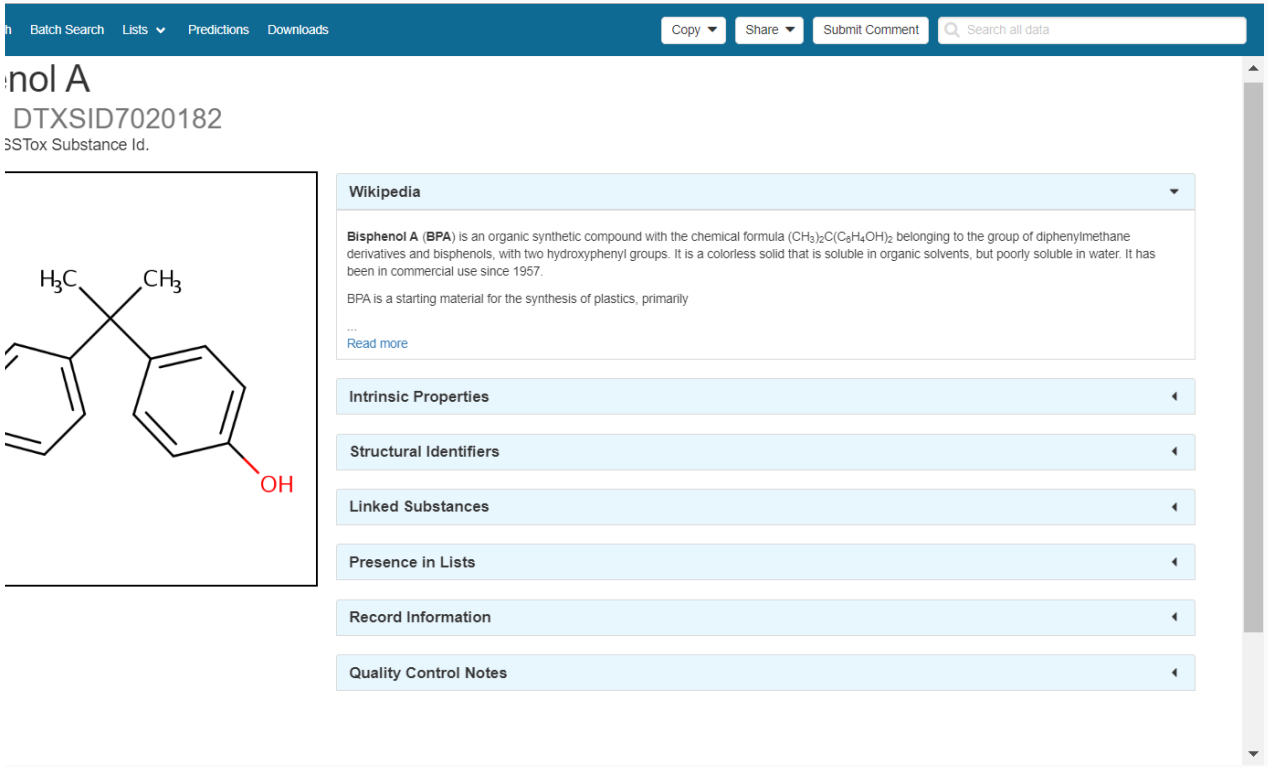
CompTox Dashboard
Assays and Genes



Detailed Chemical Pages Redesign

EPA United States Environmental Protection Agency

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS



Physicochemical properties

Property

Summary

Summary

Download

Columns

Search query

Property	Experimental average	Predicted average	Experimental median	Predicted median	Experimental range	Predicted range	Unit
LogP: Octanol-Water	3.32 (1)	3.29		3.43	3.32	2.40 to 3.64	
Melting Point	155 (7)	139	156	138	153 to 156	125 to 157	°C
Boiling Point	200 (1)	363		360	200	343 to 401	°C
Water Solubility	5.26e-4 (1)	9.64e-4		1.00e-3	5.26e-4	5.44e-4 to 1.31e-3	mol/L
Vapor Pressure	-	8.37e-7		3.43e-7	-	6.83e-8 to 2.59e-6	mmHg
Flash Point	-	190		190	-	188 to 192	°C
Surface Tension	-	46.0			-	46.0	dyn/cm
Index of Refraction	-	1.60			-	1.60	
Molar Refractivity	-	68.2			-	68.2	cm^3
Polarizability	-	27.0			-	27.0	Å^3
Density	-	1.17		1.17	-	1.14 to 1.20	g/cm^3
Molar Volume	-	200			-	200	cm^3
Thermal Conductivity	-	150			-	150	mW/(m*K)

Detailed OPERA Prediction Reports

EPA United States Environmental Protection Agency

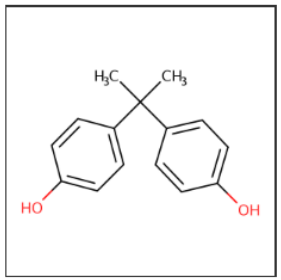
Source	Result	Calculation Details
Experimental Values		
PhysPropNCCT		
Predicted Values		
EPISUITE		
NICEATM		
ACD/Labs Conse		
ACD/Labs		
OPERA		

OPERA Models: LogP: Octanol-Water

Save PDF

Bisphenol A

80-05-7 | DTXSID7020182



Model Results

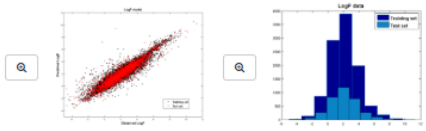
Predicted value: 3.35

Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.75

Model Performance

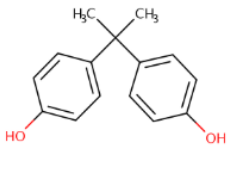


Weighted KNN model

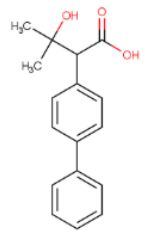
QMRF

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.86	0.67	0.86	0.78

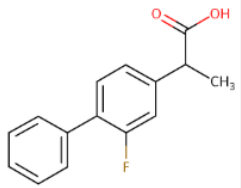
Nearest Neighbors from the Training Set



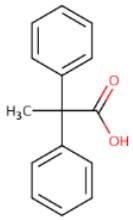
Bisphenol A
Measured: 3.32
Predicted: 3.35



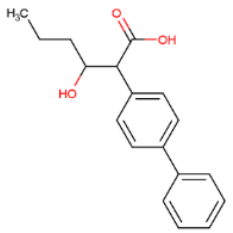
BUTANOIC ACID,2-(4-BIPHENYLYL)-3-HYDROX...
Measured: 3.25
Predicted: 3.45



Flurbiprofen
Measured: 4.16
Predicted: 3.83



2,2-Diphenylpropionic acid
Measured: 2.69
Predicted: 2.93



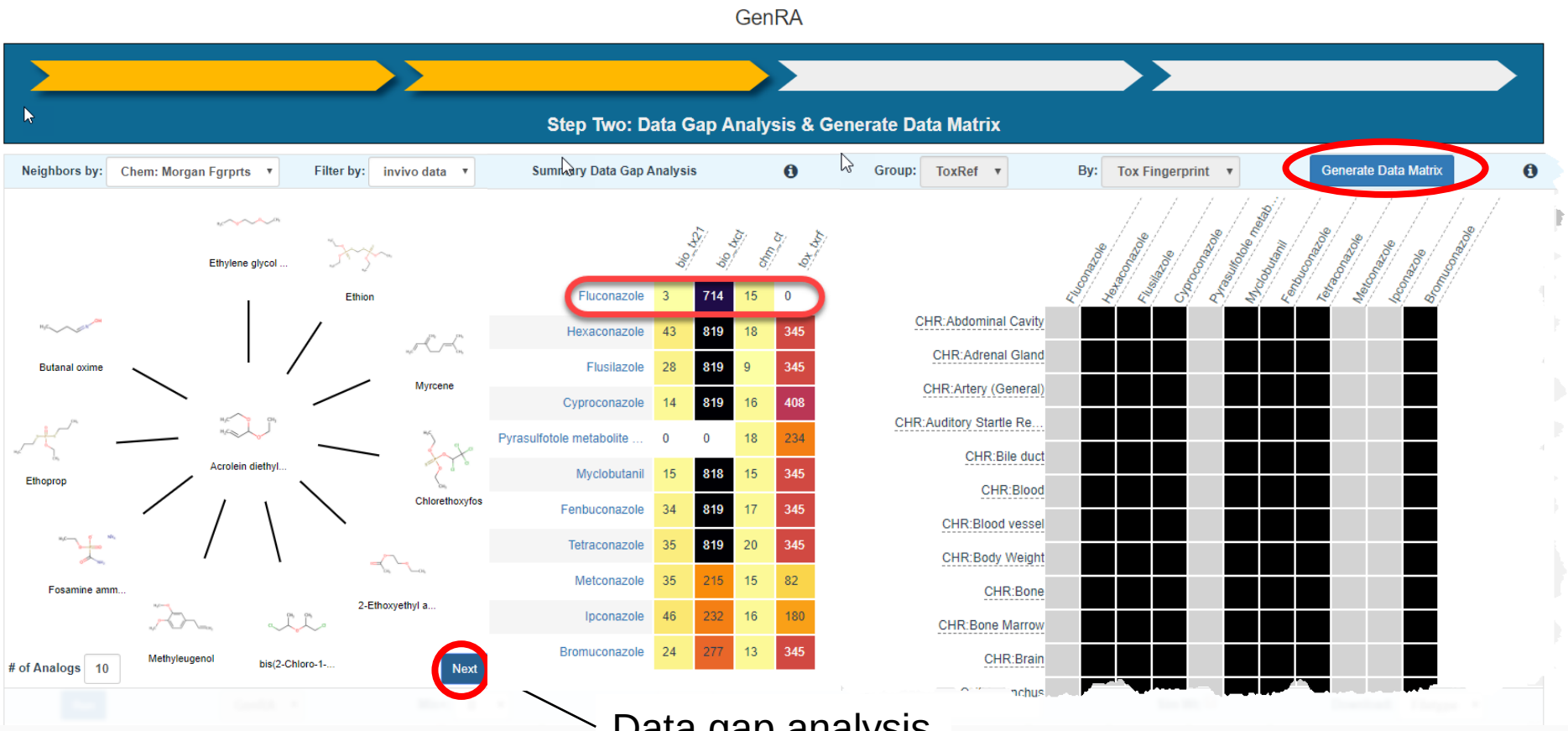
3-OH-2-(4-BIPHENYLYL)HEXANOIC ACID
Measured: 3.75
Predicted: 3.68

Other Dashboard Predictions

- Predictions and models expand outside of simply physicochemical and environmental fate and transport

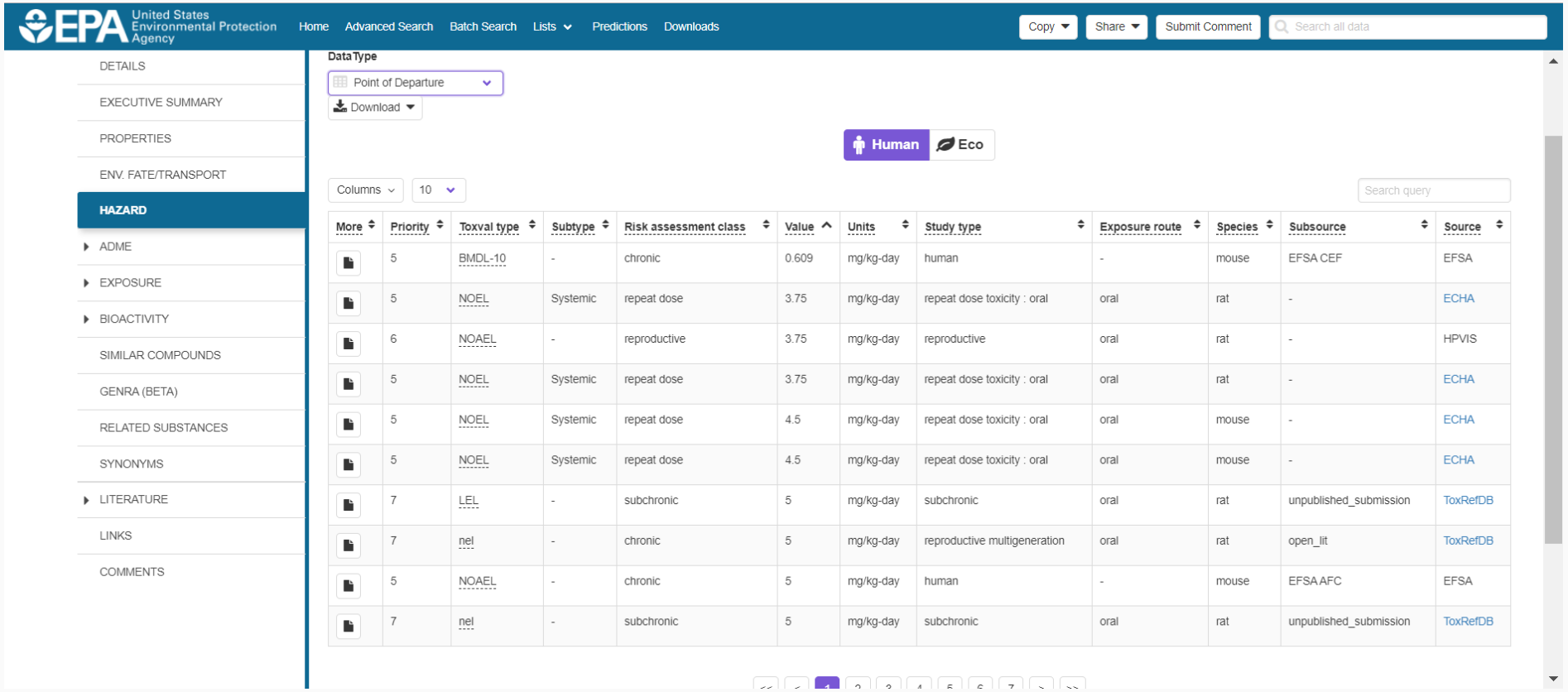
- Examples
  - Read-across for Toxicity Endpoints
  - Quantitative Structure–Use Relationship (QSUR) models
  - High-Throughput ToxicoKinetics (HTTK)
  - Models based on high throughput bioactivity data

GenRA (Generalised Read-Across)



Data gap analysis

Access to Chemical Hazard Data



More	Priority	Toxval type	Subtype	Risk assessment class	Value	Units	Study type	Exposure route	Species	Subsource	Source
	5	BMDL-10	-	chronic	0.609	mg/kg-day	human	-	mouse	EFSA CEF	EFSA
	5	NOEL	Systemic	repeat dose	3.75	mg/kg-day	repeat dose toxicity : oral	oral	rat	-	ECHA
	6	NOAEL	-	reproductive	3.75	mg/kg-day	reproductive	oral	rat	-	HPVIS
	5	NOEL	Systemic	repeat dose	3.75	mg/kg-day	repeat dose toxicity : oral	oral	rat	-	ECHA
	5	NOEL	Systemic	repeat dose	4.5	mg/kg-day	repeat dose toxicity : oral	oral	mouse	-	ECHA
	5	NOEL	Systemic	repeat dose	4.5	mg/kg-day	repeat dose toxicity : oral	oral	mouse	-	ECHA
	7	LEL	-	subchronic	5	mg/kg-day	subchronic	oral	rat	unpublished_submission	ToxRefDB
	7	nel	-	chronic	5	mg/kg-day	reproductive multigeneration	oral	rat	open_lit	ToxRefDB
	5	NOAEL	-	chronic	5	mg/kg-day	human	-	mouse	EFSA AFC	EFSA
	7	nel	-	subchronic	5	mg/kg-day	subchronic	oral	rat	unpublished_submission	ToxRefDB

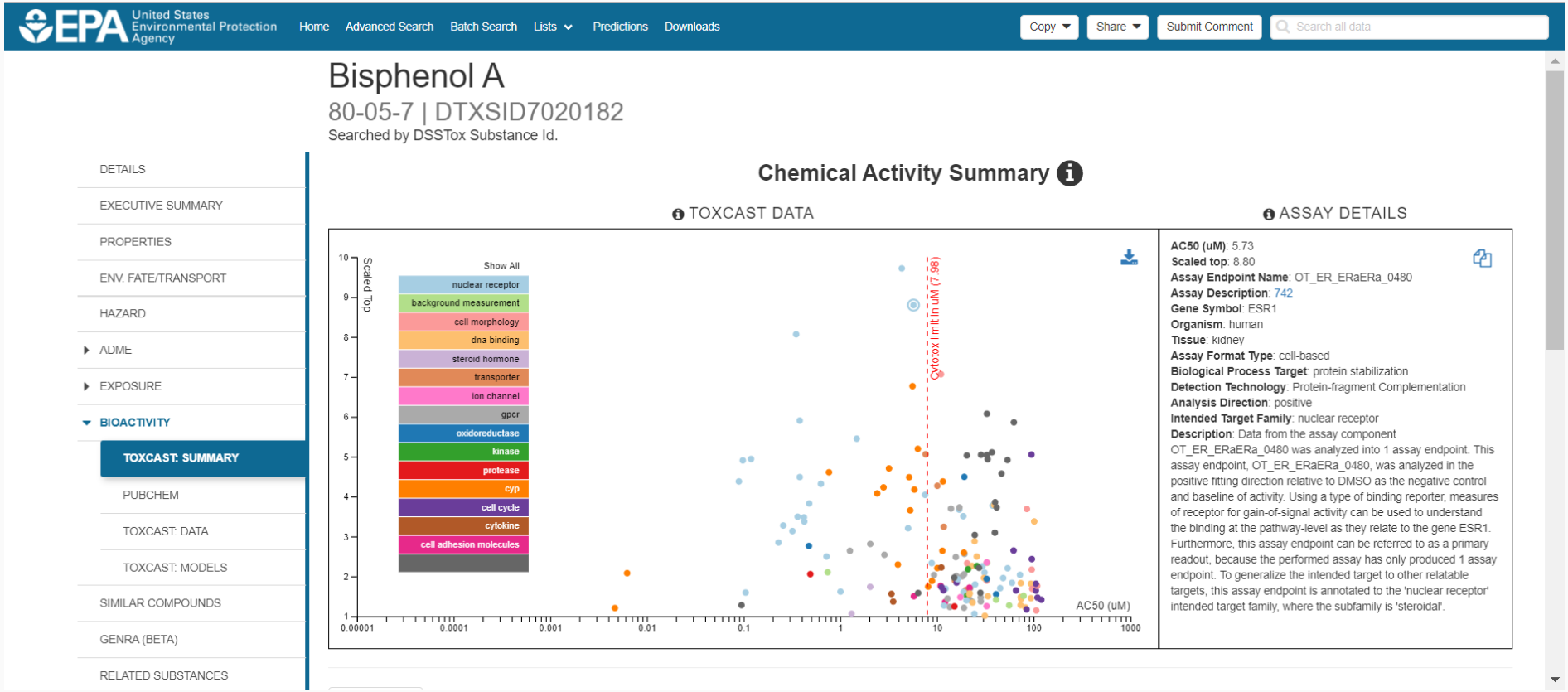
Hazard Data from “ToxVal_DB”

- ToxVal Database contains following data:
  - 30,050 chemicals
  - 772,721 toxicity values
  - 29 sources of data
  - 21,507 sub-sources
  - 4585 journals cited
  - 69,833 literature citations

In Vitro Bioassay Screening

ToxCast and Tox21

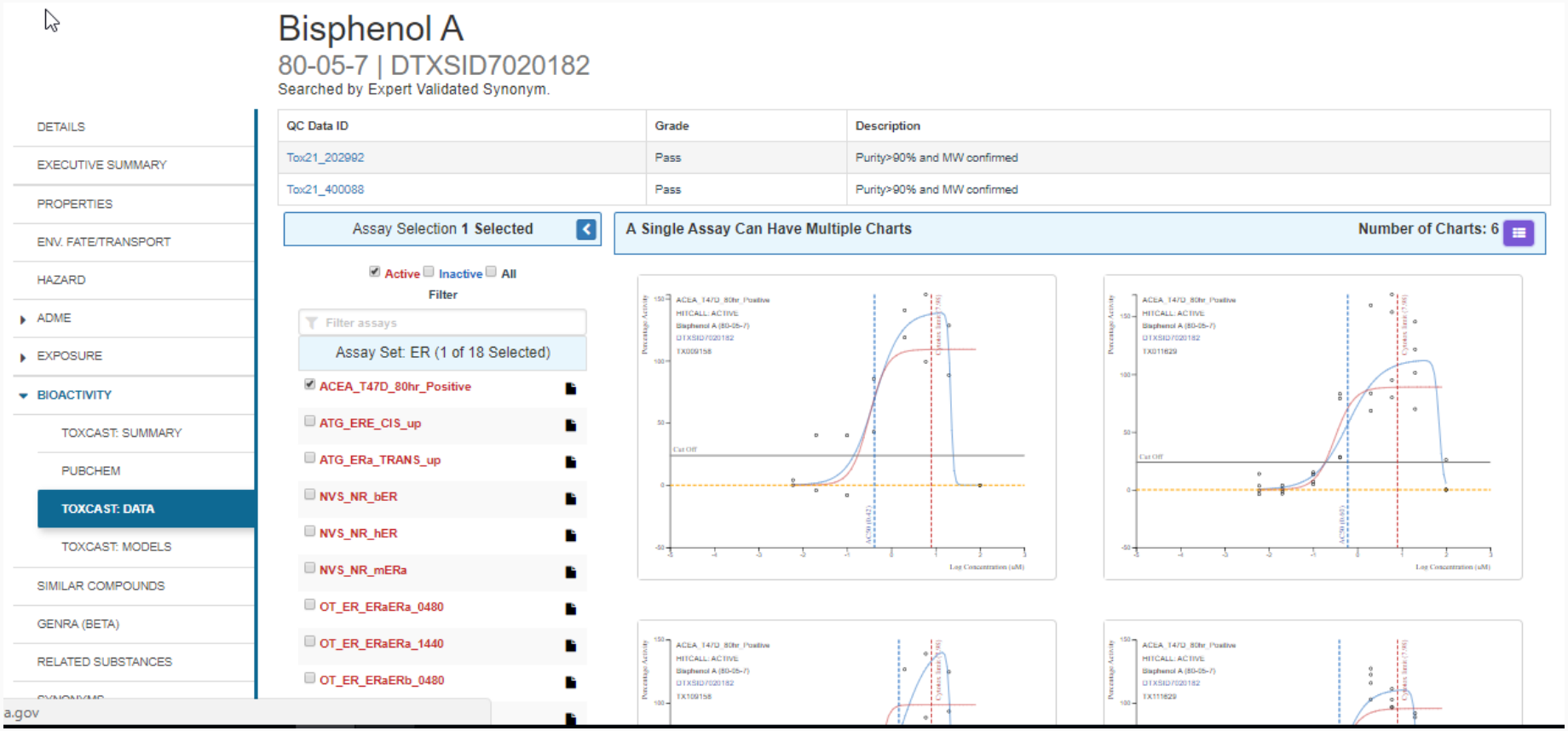
EPA United States Environmental Protection Agency



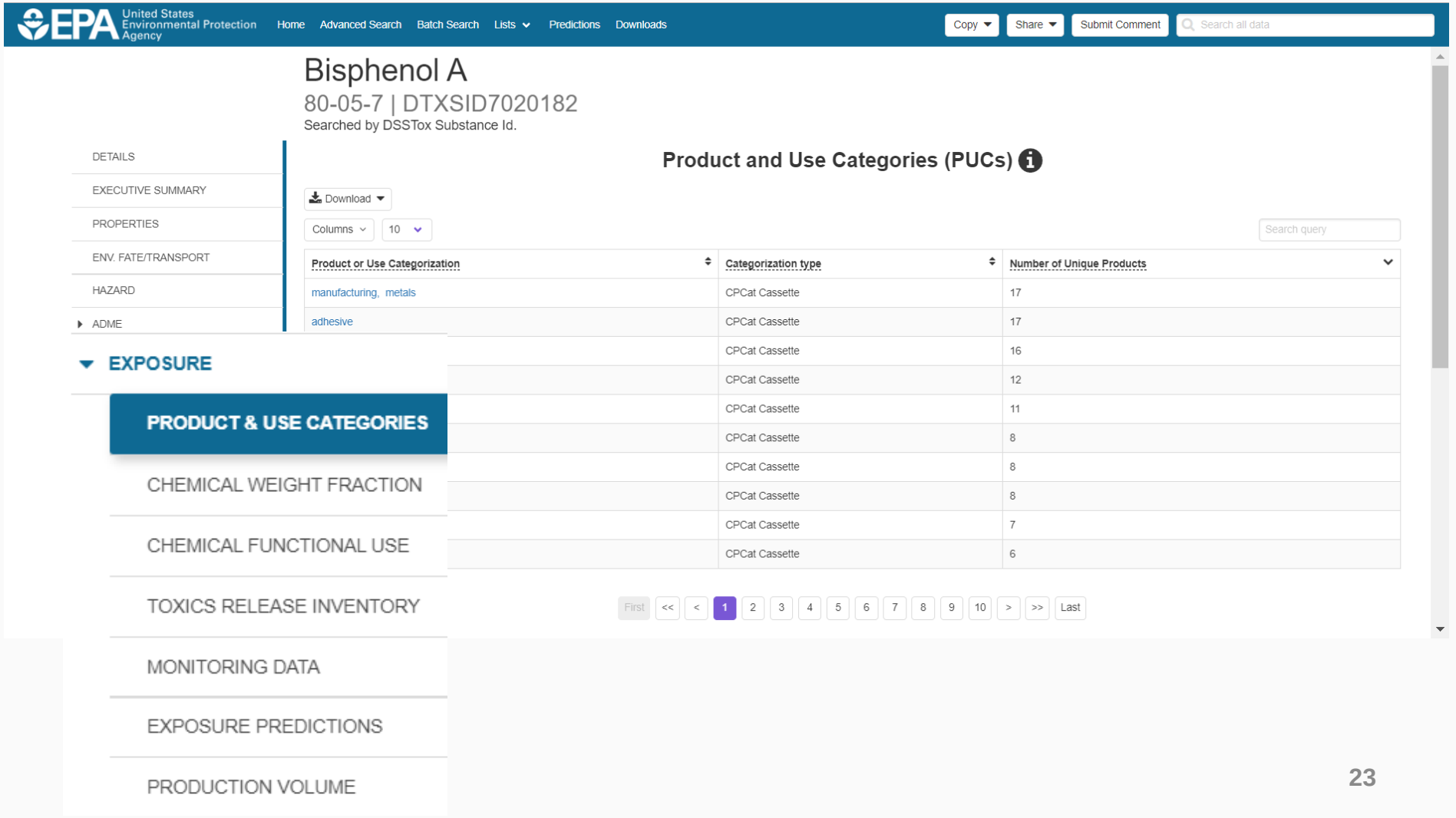
In Vitro Bioassay Screening

ToxCast and Tox21

EPA United States Environmental Protection Agency

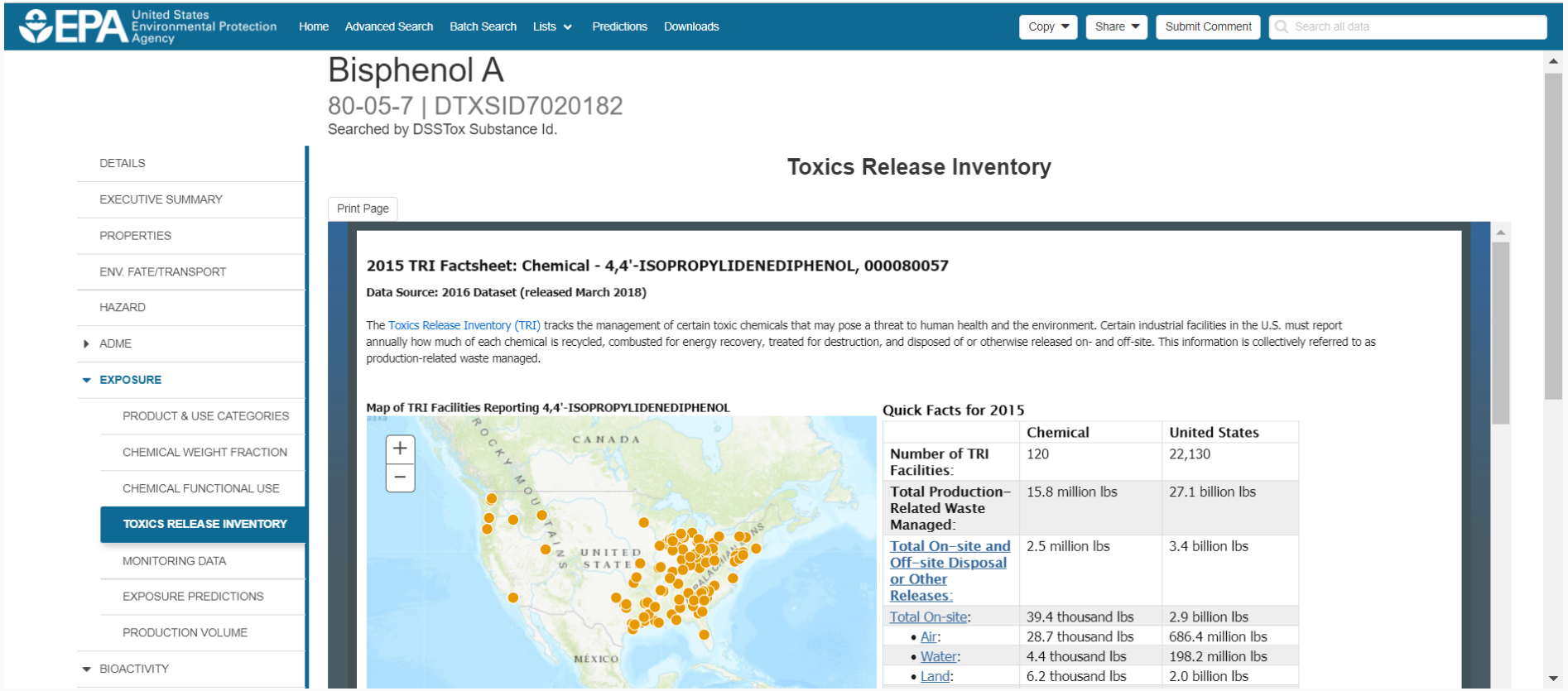


Sources of Exposure to Chemicals

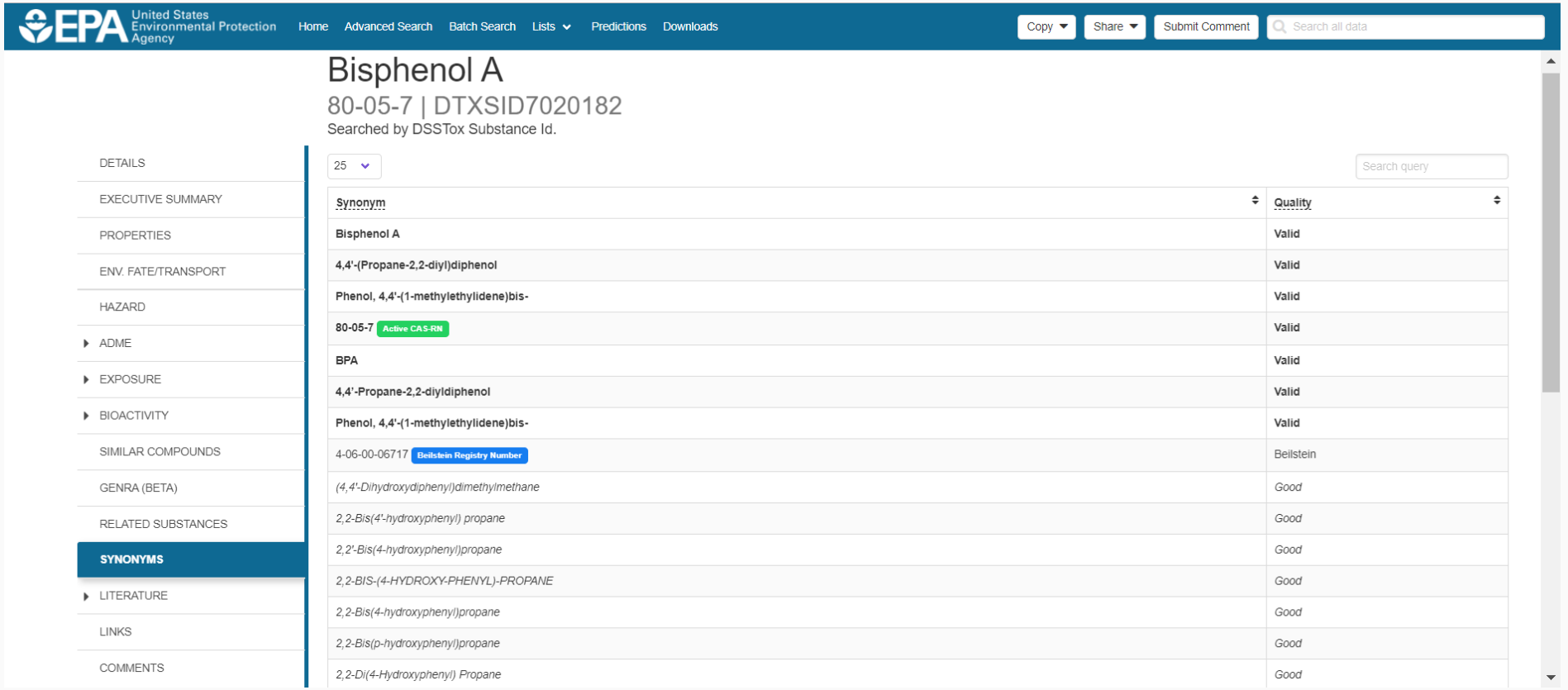


Sources of Exposure to Chemicals

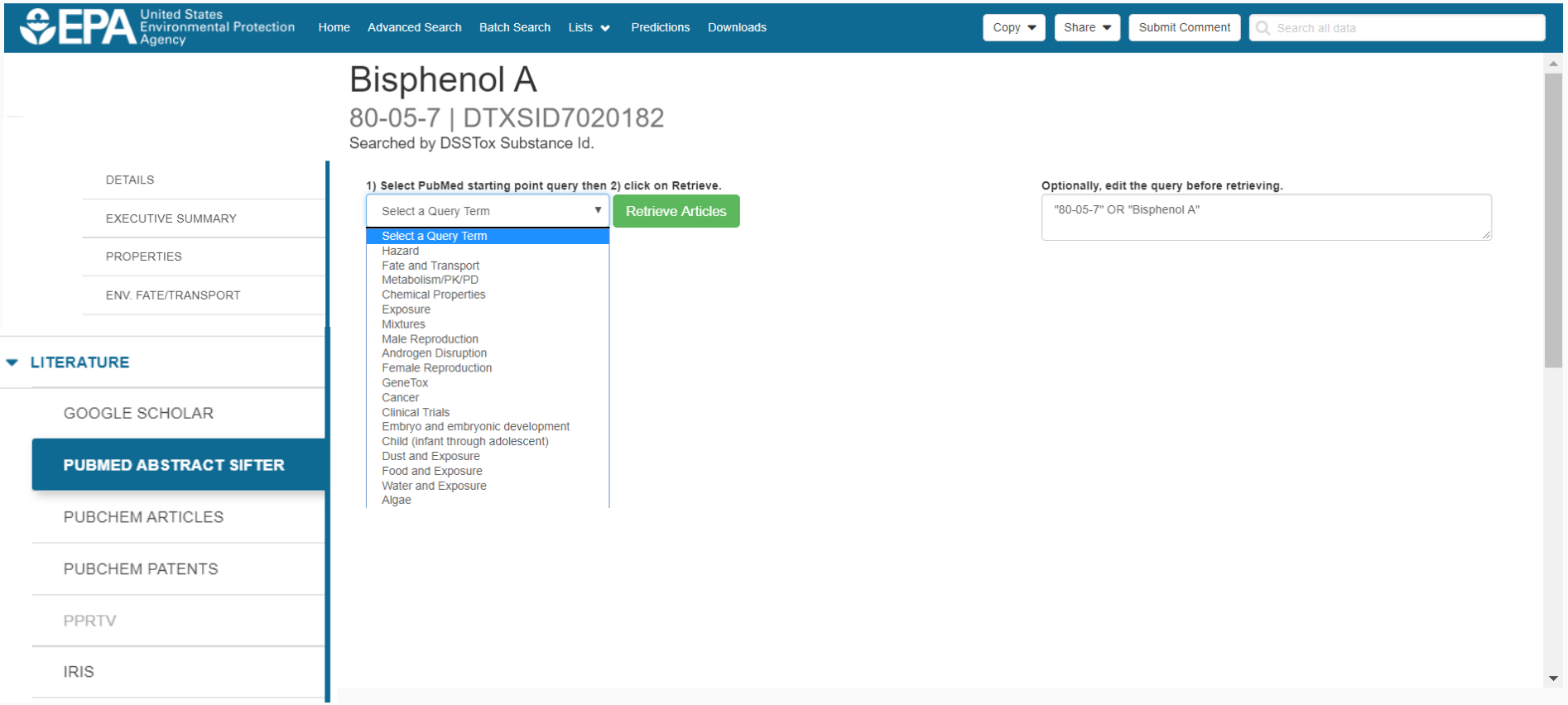
EPA United States Environmental Protection Agency



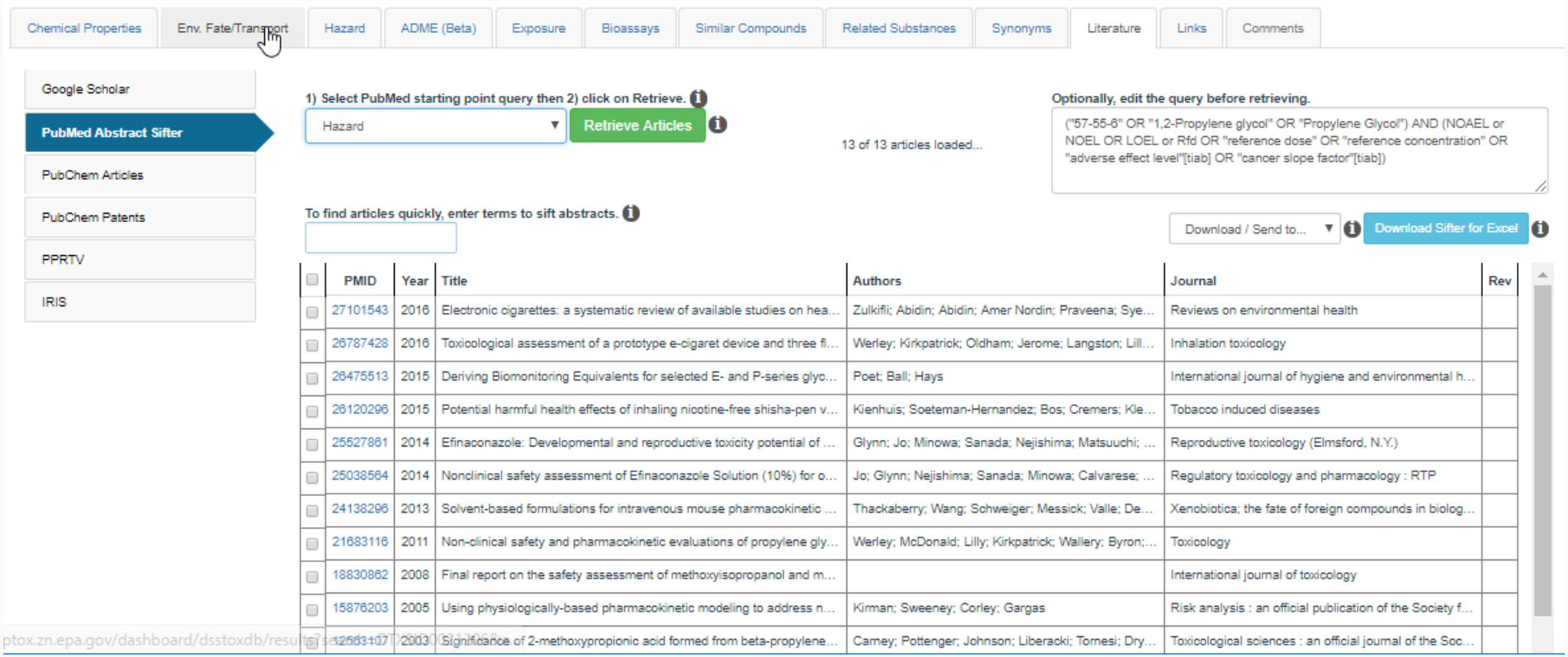
Identifiers to Support Searches



Literature Searches and Links

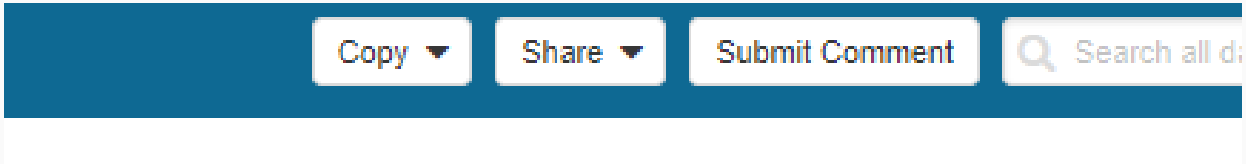


Literature Searches and Links



How can we curate our data?

- Crowdsourcing is well proven nowadays
- Comments can be added at a record level



- Submitted comments are reviewed by administrators and responded to

Public Crowdsourced Comments

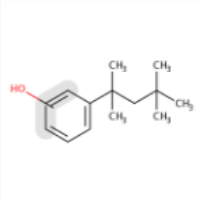
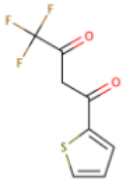
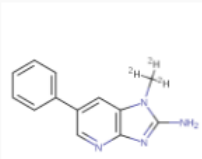
https://comptox.epa.gov/dashboard/comments/public_index

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency | Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | Share | Search all data

Crowdsourced Comments

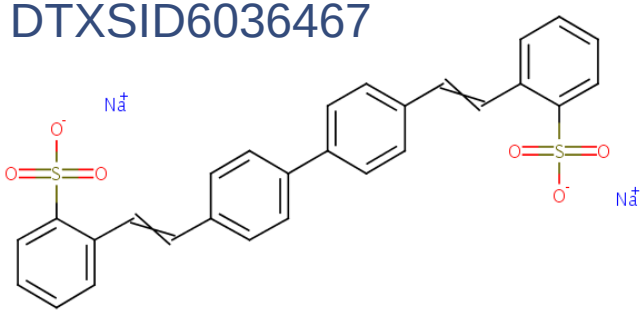
Show 10 entries | Search:

Chemical	Structure	Date	Comment	Status
(1,1,3,3-Tetramethylbutyl)phenol		2017-07-15	Octylphenol redirects here, yet the name and related chemicals are 1,1,3,3-tetramethylbutylphenol - which is only a subset of all octylphenol isomers? Is this CAS only for these alkyl isomers?	★
1,3-Butanedione, 4,4,4-trifluoro-1-(2-thienyl)-		2017-03-30	Synonym: TTFA (Any way to bank these reCAPTCHAs so I don't have to do it everytime?)	★
1-(²H₃)Methyl-6-phenyl-1H-imidazo[4,5-b]pyridin-2-amine		2017-05-06	1-(2H3)Methyl-6-phenyl-1H-imidazo[4,5-b]pyridin-2-amine 210049-13-1 | DTXSID70670097 contains an error in the empirical formula due to an error in the deuterium representation and subsequent counting	★

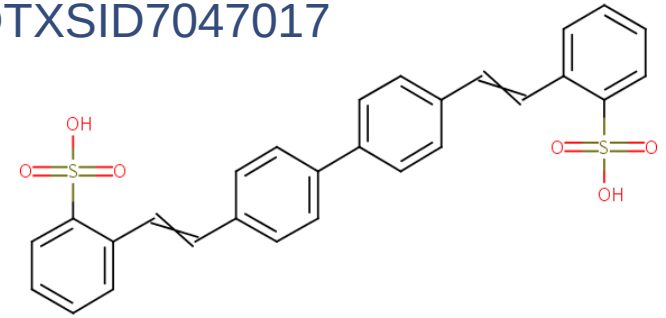
MassBank/CompTox Curation of External Data

- A “nice” example: 4-4'-Bis(2sulfostyryl)biphenyl

Purchased: CAS: 27344-41-8
DTXSID6036467



Registered: CAS: 38775-22-3 (UFZ)
DTXSID7047017



Comment from structure source: to my knowledge the stilbene-derived fluorescent whitening agents are all trans (E) isomers, as the cis (Z) isomers are not fluorescent (although they might undergo photo-isomerisation to the cis isomers under UV light, and clothing gets yellowish again then...) . Thus I would consider the E,E form the correct one, although I don't know whether it is synthesized in a way that x % of the technical product are actually the inactive E,Z (dunno if this is still active?) or Z,Z forms and thus the undefined stereo would be correct. In the environment E,E is partly transformed to E,Z (maybe also Z,Z) and thus both isomers occur, see: http://pubs.acs.org/doi/pdf/10.1021/es960748a CAS number on record was 27344-41-8 which is DTXSID6036467.

Add A Comment

Related to comment just added to DTXSID7047017 - it would be useful for us to know if this CAS number is undefined stereochemistry (a technical mix) or defined ... we would update our records accordingly.

User comment posted 4 months ago

ed to the mixed E/Z-form of the chemical. The E/E form is
6036467 is ALSO the mixed EZ form

I have a stereo-defined version of this on record as Fluorescent brightener 351 c1cc(c(cc1)S(=O)(=O)O)/C=C/c2ccc(cc2)c3ccc(cc3)/C=C/c4c(cccc4)S(=O)(=O)O SQAKQVFOMMLRPR-IWGRKNQJSA-N

User comment posted 4 months ago

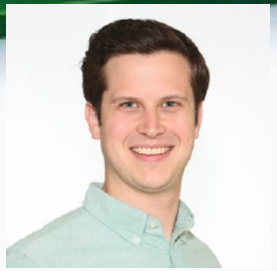
The CAS Number is indeed for the undefined stereochemistry form.

Admin reply posted 4 months ago

There is NO indication of the E-defined stereochemistry in the registry and it appears to be a mix of isomers.

Admin reply posted 4 months ago

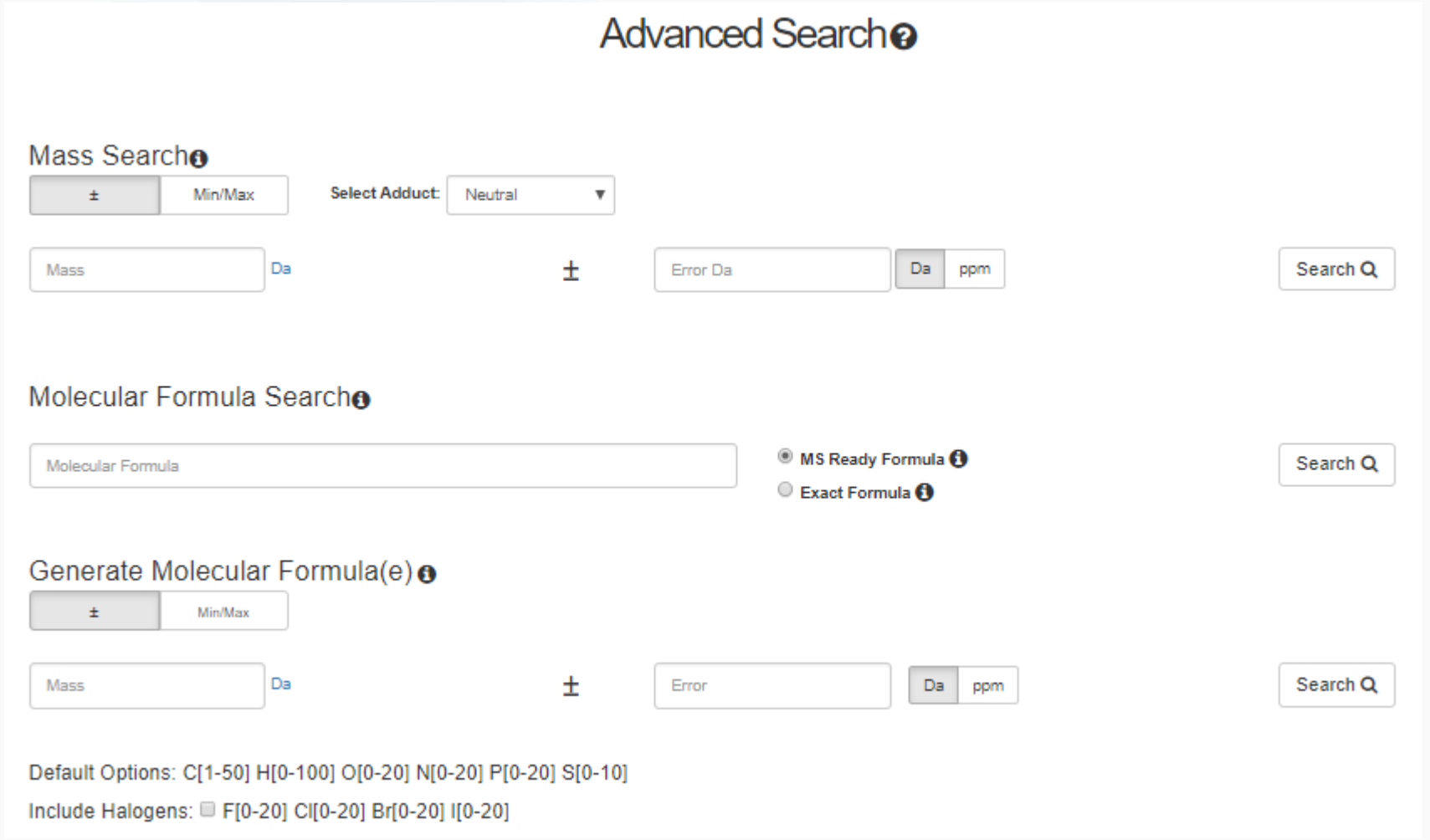
Mass Spec Specific Functionality



- The ENTACT project has driven specific functionality Mass Spectrometry needs
  - Specific searches within the dashboard
  - Mappings between chemicals in the database
  - Making use of specific representations for UVCBs
  - Addition of specific lists of chemicals
  - Collaborative cross-linking between sites
  - Research into best approaches for candidate ranking

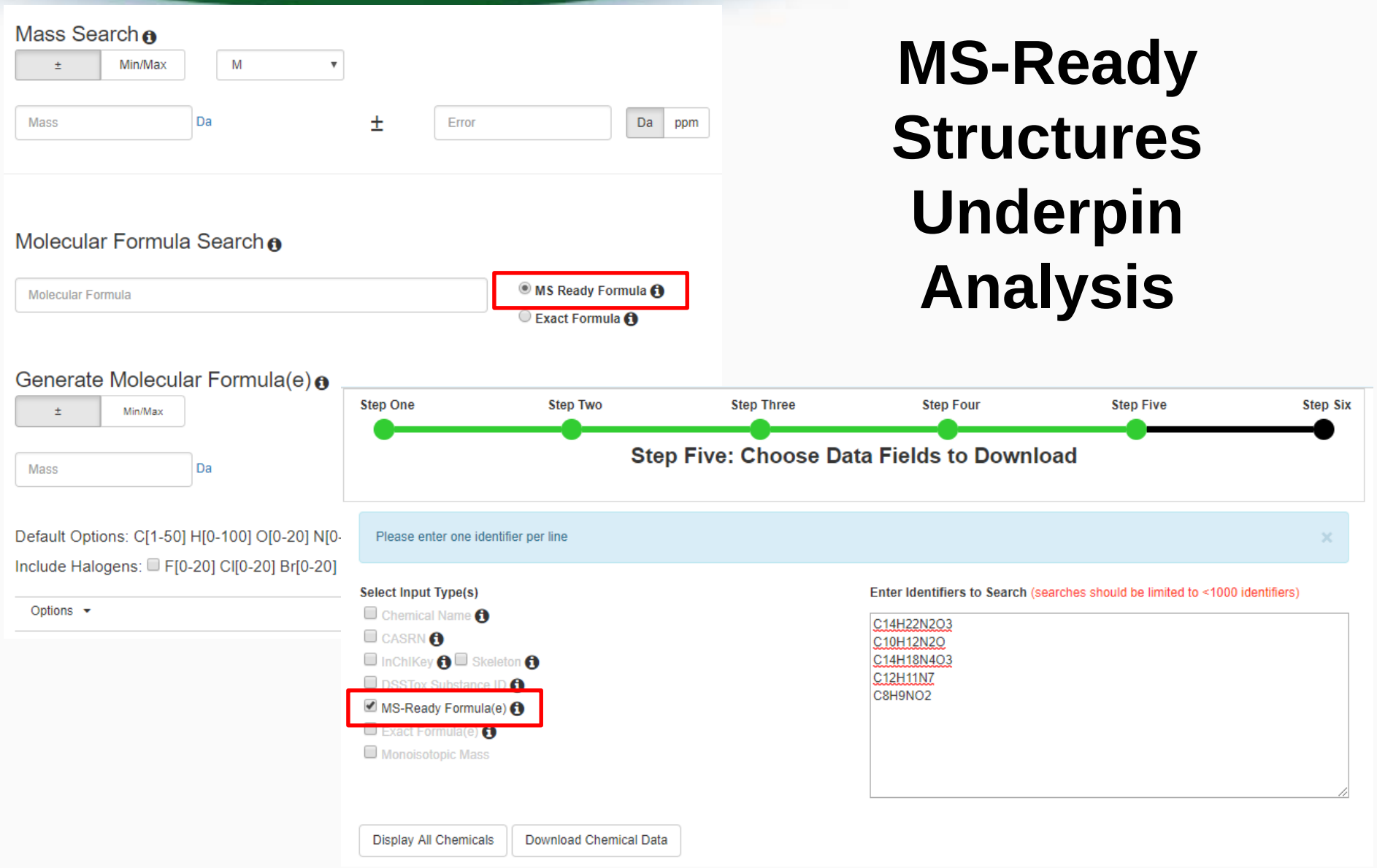
Mass and Formula Searches Supporting Mass Spectrometry

EPA United States Environmental Protection Agency



The Dashboard to Support MS-Analysis

EPA United States Environmental Protection Agency

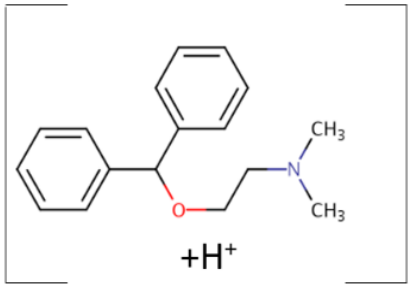


MS-Ready Structures Underpin Analysis

Specific Data-Mappings "MS-Ready Structures"

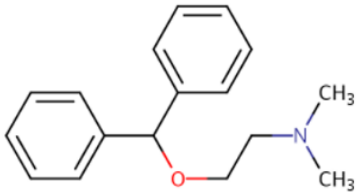
EPA United States Environmental Protection Agency

A) Molecular Ion



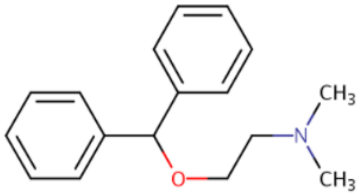
m/z≈ 256.1702

B) MS-Ready Form

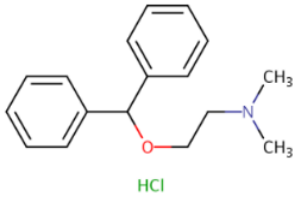


monoisotopic mass= 255.1623
$C_{17}H_{21}NO$
DTXCID802949

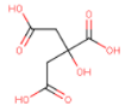
C) Mappings from MS-Ready

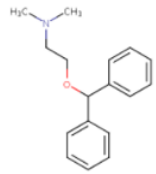


Diphenhydramine
$C_{17}H_{21}NO$ | 255.1623
DTXSID4022949

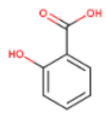


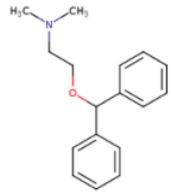
Diphenhydramine hydrochloride
$C_{17}H_{22}ClNO$ | 291.1390
DTXSID4020537





Diphenhydramine citrate
$C_{23}H_{29}NO_8$ | 447.1893
DTXSID80237211





Diphenhydramine salicylate
$C_{24}H_{27}NO_4$ | 393.1940
DTXSID10225883

Journal of Cheminformatics

- Already through peer review

Title: "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

Andrew D. McEachran[1,2*], Kamel Mansouri[1,2,3], Chris Grulke[2], Emma L. Schymanski[4], Christoph Ruttkies[5] and Antony J. Williams[2*]

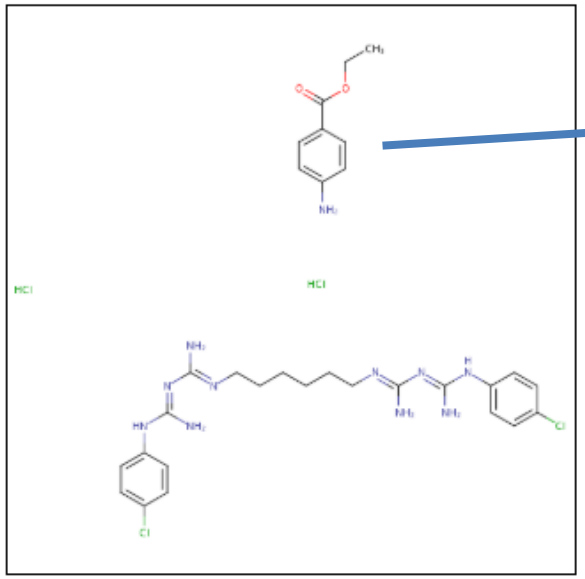
MS-Ready Mappings

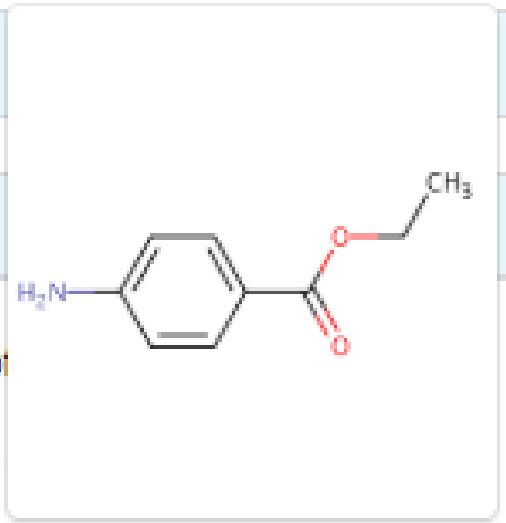
EPA United States Environmental Protection Agency

Progaron

108532-15-6 | DTXSID20148579

Searched by DSSTox Substance Id.

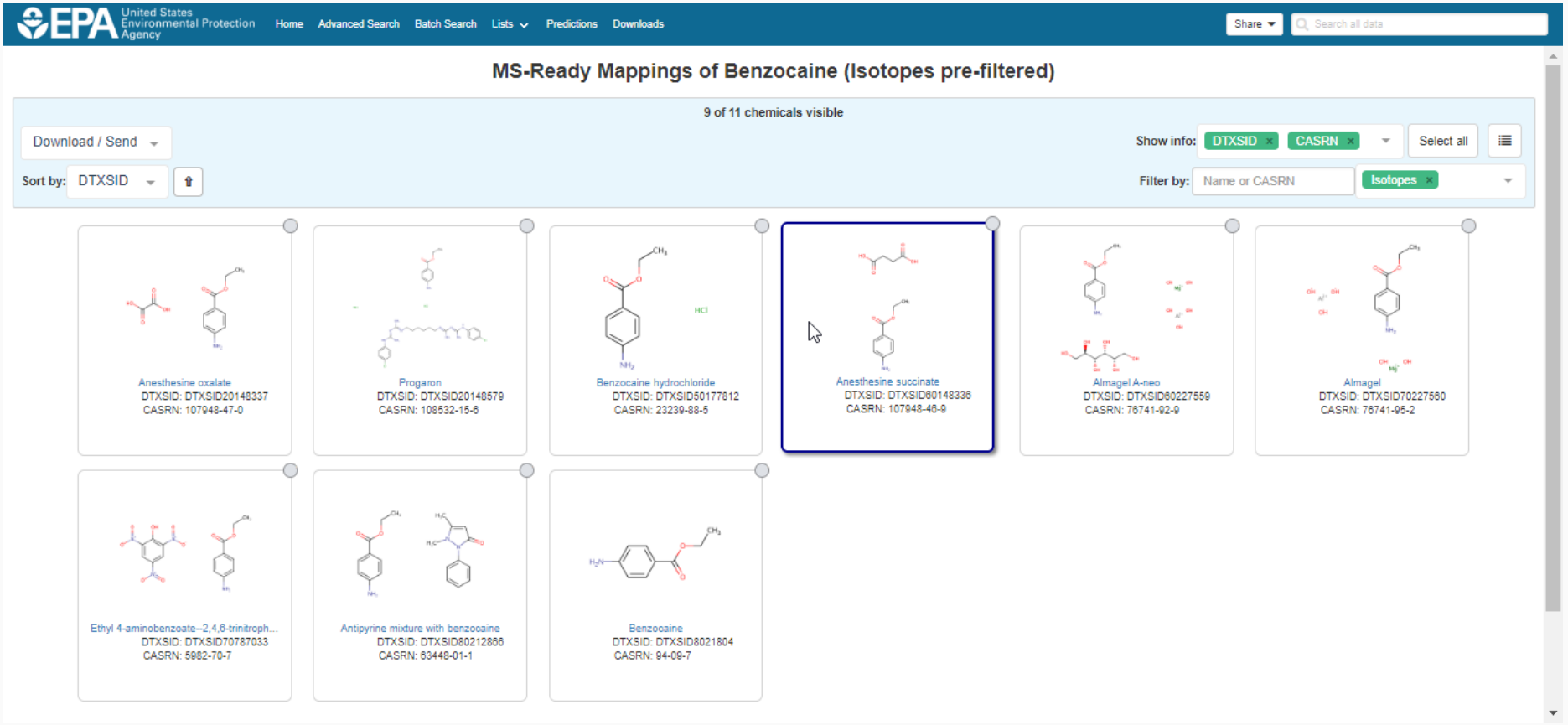




- Intrinsic Properties
- Structural Identifiers
- Linked Substances
  - Same Connectivity: 1 record (based on
  - Mixtures, Components and Neutralize
  - MS-Ready Mappings: DTXCID0013314
  - Similar Compounds: 0 records
- Presence in Lists
- Record Information
- Quality Control Notes

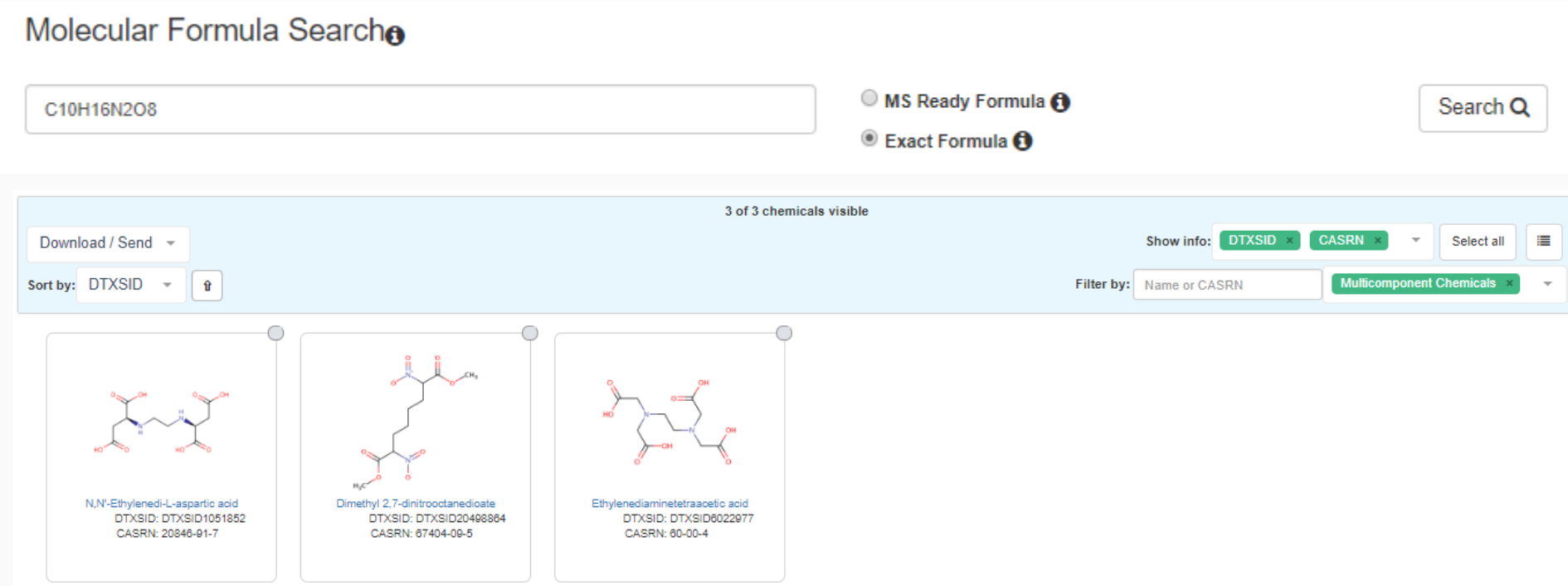
DTXCID301804:11 records;

MS-Ready Mappings Set



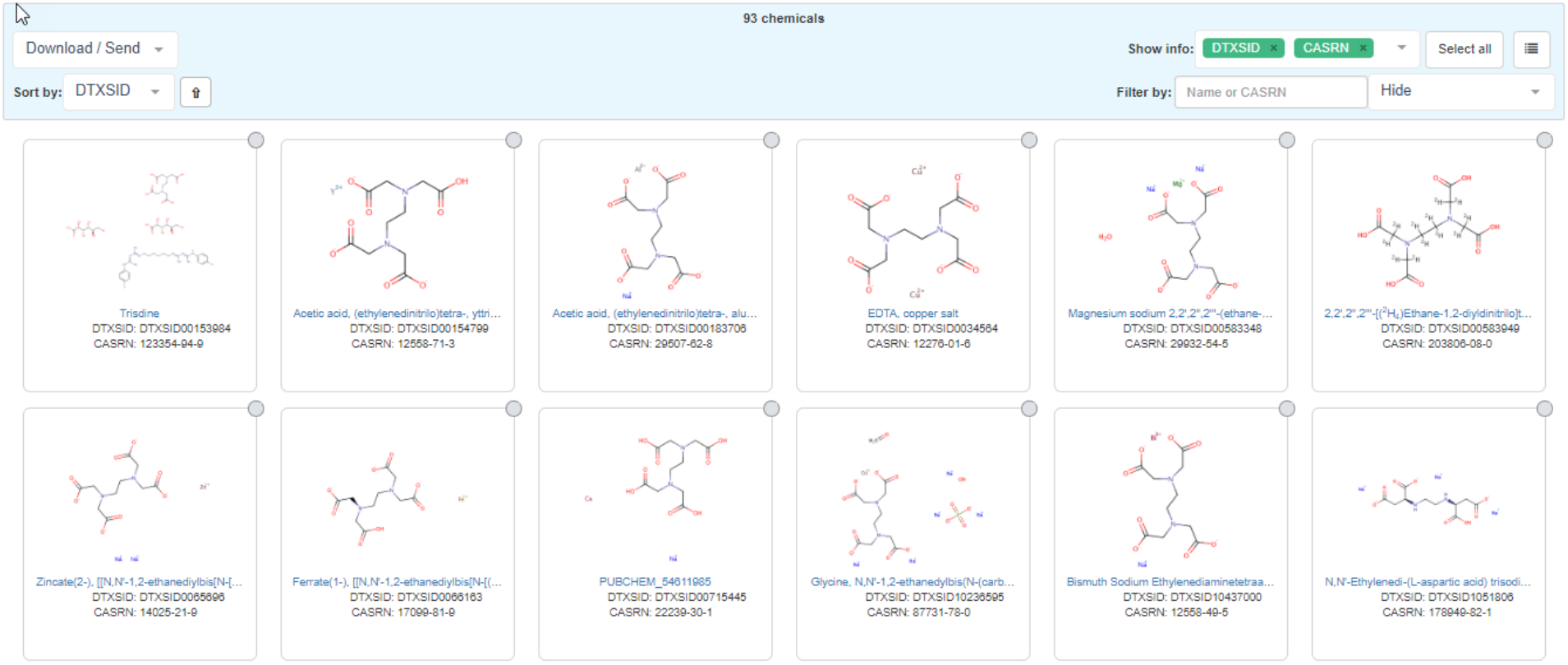
MS-Ready Mappings

- Input Formula: C10H16N2O8



MS-Ready Mappings

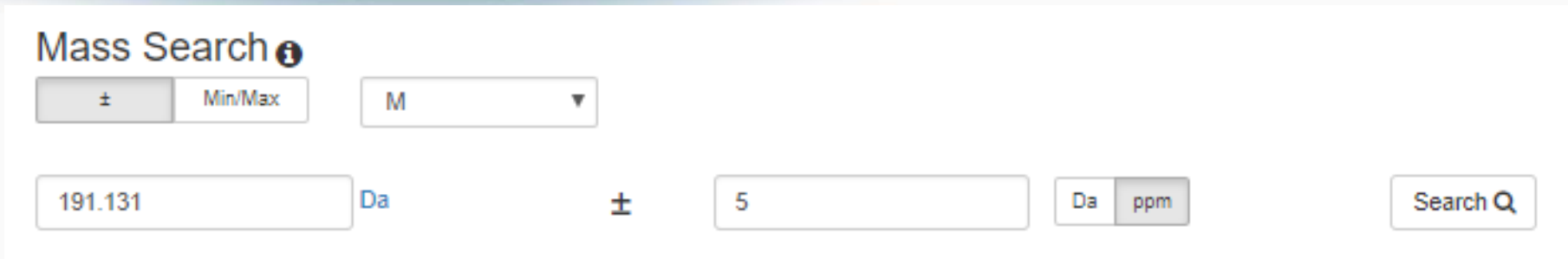
- **Same** Input Formula: C10H16N2O8
- MS Ready Formula Search: 93 Chemicals

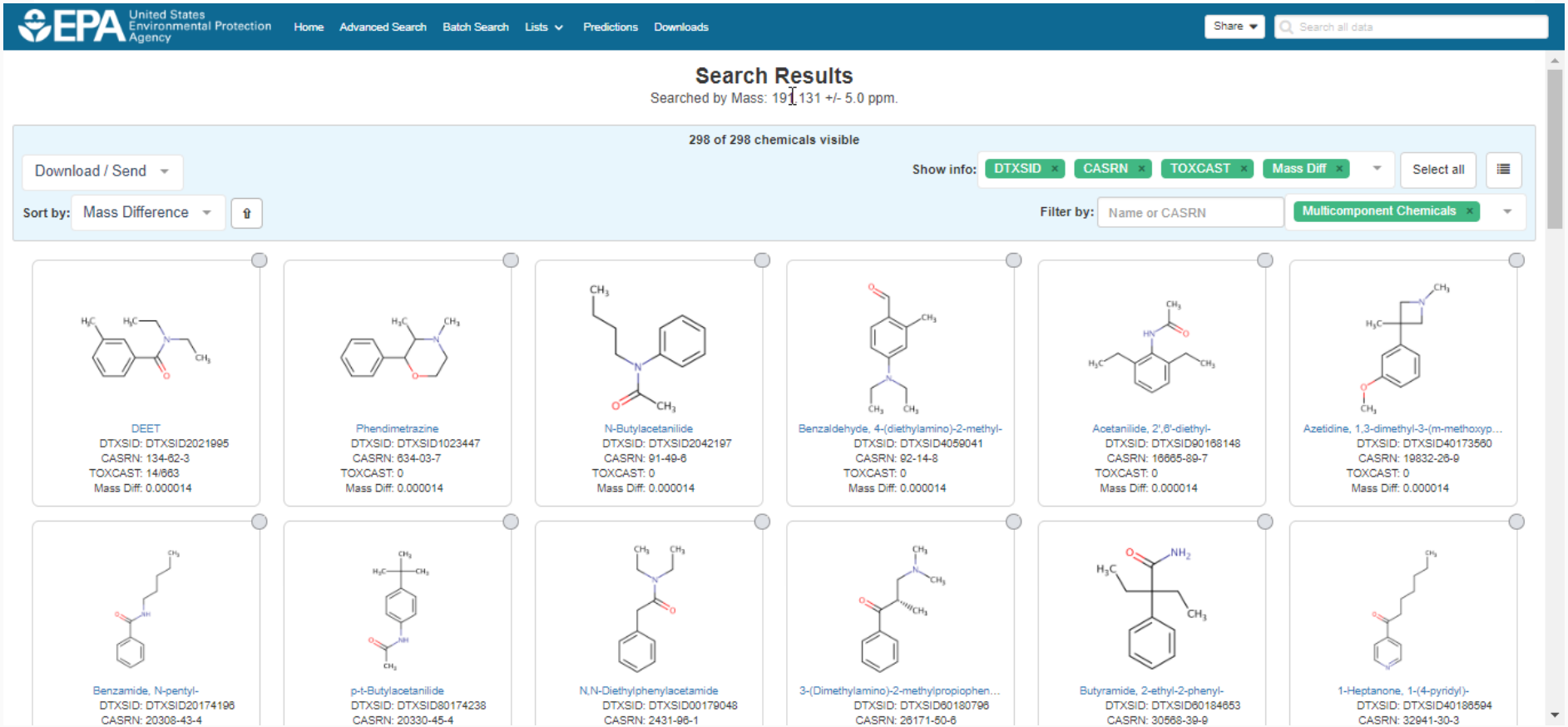


MS-Ready Mappings

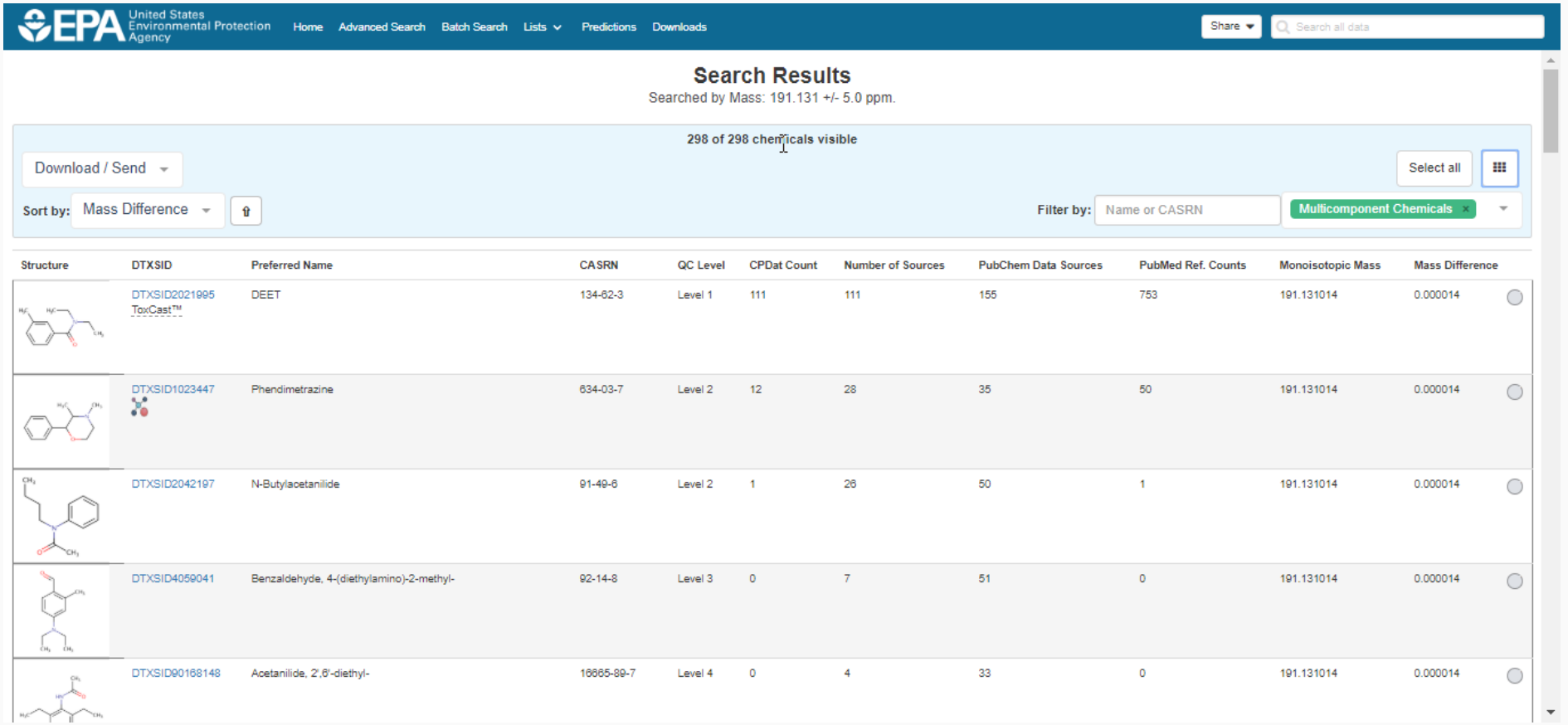
- 93 chemicals returned in total
  - Only 7 of the 93 are single component chemicals
  - 3 are neutral compounds and 1 is charged

Advanced Searches
Mass Search: Tile View





Advanced Searches
Mass Search: Table View

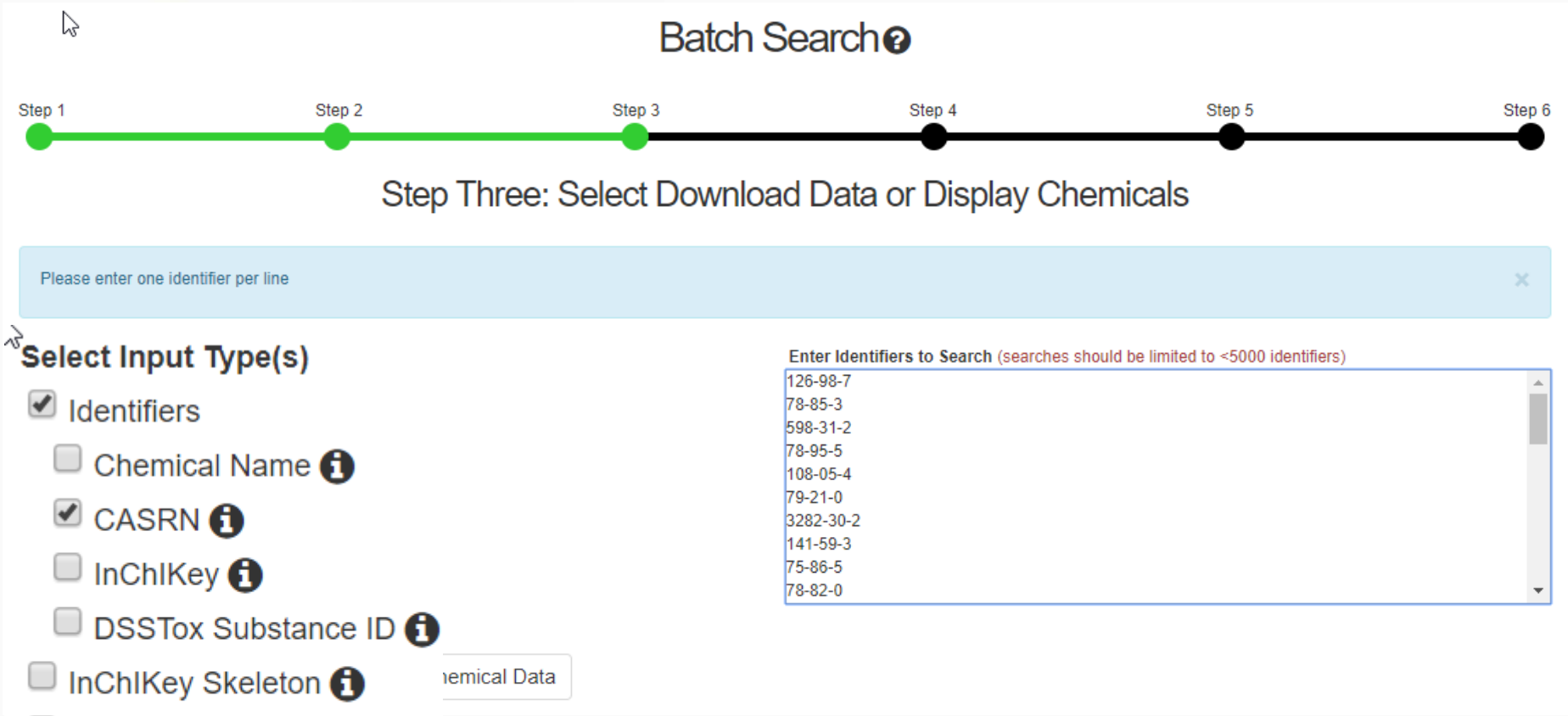


Batch Searching

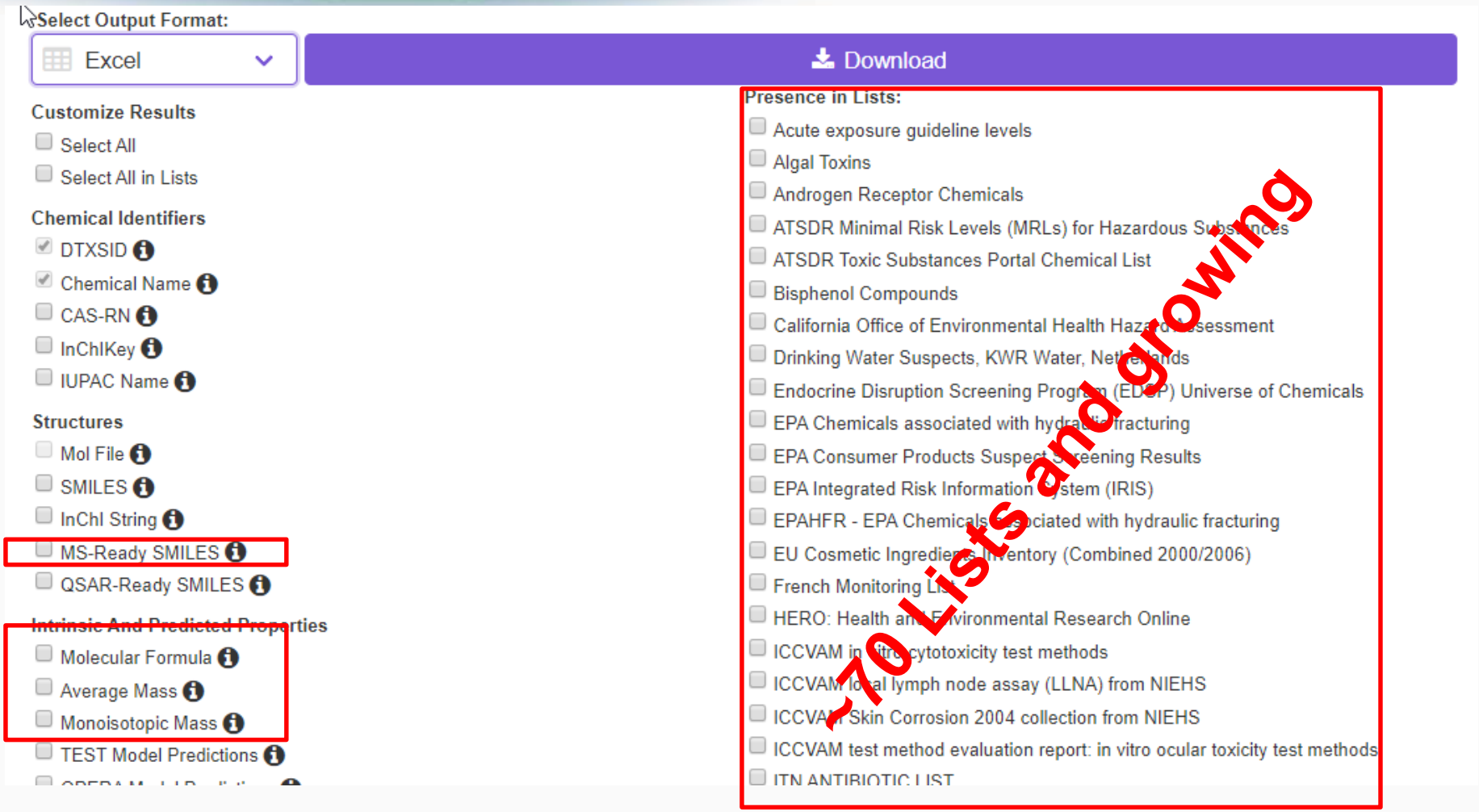
- Singleton searches are useful but we work with thousands of chemicals!

- Typical questions
  - What are the SMILES strings for a list of 1000 chemicals?
  - Do any of this list of chemicals have XXX type of data?
  - What are the predicted logP values for a list of chemicals?
  - Can I get lists of predicted properties in Excel files? In SDF files?
  - Can I get chemical lists in Excel files? In SDF files?
  - What is the list of chemicals for the formula $C_xH_yO_z$
  - What is the list of chemicals for a mass +/- error

Batch Searching



Batch Searching



Excel Output

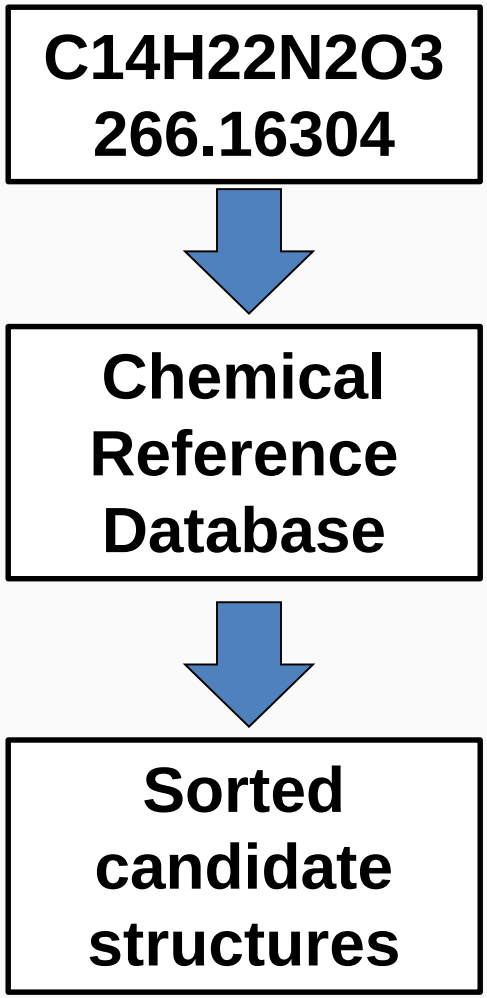
INPUT	FOUND_BY	DTXCID_IN	DATA_SOU	TOXVAL_D	TOXCAST_	TOXCAST_	NUMBER_C	PUBCHEM_	STO
C6H12O3	MS Ready I	DTXCID701	91	Y	0.36	2/562	24	83	Y
C6H12O3	MS Ready I	DTXCID003	67	Y	0.36	1/276	376	80	Y
C6H12O3	MS Ready I	DTXCID106	65	Y	4.42	5/113	6	77	Y
C6H12O3	MS Ready I	DTXCID105	45	Y	0.0	0/163	3	94	-
C6H12O3	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	-

Searching batches
Formula (or mass) searching

	A	B	C	D	E	F	G
1	INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS	DATA SOURCES
2	C14H22N2O3	DTXSID2022628	29122-68-7	Atenolol	C14H22N2O3	266.163042576	46
3	C14H22N2O3	DTXSID0021179	6673-35-4	Practolol	C14H22N2O3	266.163042576	32
4	C14H22N2O3	DTXSID4048854	841-73-6	Bucolome	C14H22N2O3	266.163042576	20
5	C14H22N2O3	DTXSID1045407	13171-25-0	Trimetazidine dihydrochloride	C14H24Cl2N2O3	338.116398	19
6	C14H22N2O3	DTXSID0045753	56715-13-0	R-(+)-Atenolol	C14H22N2O3	266.163042576	19
7	C14H22N2O3	DTXSID2048531	5011-34-7	Trimetazidine	C14H22N2O3	266.163042576	14
8	C14H22N2O3	DTXSID10239405	93379-54-5	Esatenolol	C14H22N2O3	266.163042576	12
9	C14H22N2O3	DTXSID50200634	52662-27-8	N-(2-Diethylaminoethyl)-2-(4-hydroxyphenoxy)acetamide	C14H22N2O3	266.163042576	7
10	C14H22N2O3	DTXSID4020111	51706-40-2	dl-Atenolol hydrochloride	C14H23ClN2O3	302.1397203	6
11	C14H22N2O3	DTXSID1068693	51963-82-7	Benzenamine, 2,5-diethoxy-4-(4-morpholinyl)-	C14H22N2O3	266.163042576	5
12	C18H34N2O6S	DTXSID3023215	154-21-2	Lincomycin	C18H34N2O6S	406.213757997	35
13	C18H34N2O6S	DTXSID7047803	859-18-7	Lincomycin hydrochloride	C18H35ClN2O6S	442.1904357	22
14	C18H34N2O6S	DTXSID20849438	1398534-62-7	PUBCHEM_71432748	C18H35ClN2O6S	442.1904357	1
15	C10H12N2O	DTXSID1047576	486-56-6	Cotinine	C10H12N2O	176.094963014	40
16	C10H12N2O	DTXSID8075330	50-67-9	Serotonin	C10H12N2O	176.094963014	22
17	C10H12N2O	DTXSID8044412	2654-57-1	4-Methyl-1-phenylpyrazolidin-3-one	C10H12N2O	176.094963014	18
18	C10H12N2O	DTXSID80165186	153-98-0	Serotonin hydrochloride	C10H13ClN2O	212.0716407	11
19	C10H12N2O	DTXSID2048870	29493-77-4	(4R,5S)-4-methyl-5-phenyl-4,5-dihydro-1,3-oxazol-2-amine	C10H12N2O	176.094963014	10
20	C10H12N2O	DTXSID10196105	443-31-2	6-Hydroxytryptamine	C10H12N2O	176.094963014	9
21	C10H12N2O	DTXSID90185693	31822-84-1	1,4,5,6-Tetrahydro-5-phenoxypyrimidine	C10H12N2O	176.094963014	7
22	C10H12N2O	DTXSID40178777	2403-66-9	2-Benzimidazolepropanol	C10H12N2O	176.094963014	7
23	C10H12N2O	DTXSID80157026	13140-86-8	N-Cyclopropyl-N'-phenylurea	C10H12N2O	176.094963014	6
24	C10H12N2O	DTXSID30205607	570-14-9	4-Hydroxytryptamine	C10H12N2O	176.094963014	6
25	C14H18N4O3	DTXSID5023900	17804-35-2	Benomyl	C14H18N4O3	290.137890456	68
26	C14H18N4O3	DTXSID3023712	738-70-5	Trimethoprim	C14H18N4O3	290.137890456	51
27	C14H18N4O3	DTXSID40209671	60834-30-2	Trimethoprim hydrochloride	C14H19ClN4O3	326.1145682	8
28	C14H18N4O3	DTXSID70204210	55687-49-5	Benzenemethanol, 4-((2,4-diamino-5-pyrimidinyl)methyl)-2,	C14H18N4O3	290.137890456	5
29	C14H18N4O3	DTXSID20152671	120075-57-2	6-Methoxy-4-(3-(N,N-dimethylamino)propylamino)-5,8-quin	C14H18N4O3	290.137890456	4
30	C14H18N4O3	DTXSID30213742	63931-79-3	1H-1,2,4-Benzotriazepine-3-carboxylic acid, 4,5-dihydro-4-	C14H18N4O3	290.137890456	3
31	C14H18N4O3	DTXSID30219608	69449-07-6	2,4-Pyrimidinediamine, 5-((3,4,5-trimethoxyphenyl)methyl)-	C14H20N4O4	308.14845514	3
32	C14H18N4O3	DTXSID20241155	94232-27-6	L-Aspartic acid, compound with 5-((3,4,5-trimethoxyphenyl	C18H25N5O7	423.175398165	3
33	C14H18N4O3	DTXSID80241156	94232-28-7	L-Glutamic acid, compound with 5-((3,4,5-trimethoxypheny	C19H27N5O7	437.191048229	3
34	C14H18N4O3	DTXSID20143781	101204-93-7	1H-Pyrido(2,3-e)-1,4-diazepine-2,3,5-trione, 4-(2-(diethylam	C14H18N4O3	290.137890456	3
35	C12H11N7	DTXSID6021373	396-01-0	Triamterene	C12H11N7	253.107593382	52
36	C12H11N7	DTXSID00204465	5587-93-9	Ampyrimine	C12H11N7	253.107593382	7
37	C12H11N7	DTXSID5064621	7300-26-7	Benzenamine, 4-azido-N-(4-azidophenyl)-	C12H9N7	251.091943318	4
38	C12H11N7	DTXSID00848025	90293-82-6	Sulfuric acid--6-phenylpteridine-2,4,7-triamine (1/1)	C12H13N7O4S	351.074973101	1
39	C12H11N7	DTXSID50575293	92310-83-3	(1E)-N-Phenyl-1,2-bis(1H-1,2,4-triazol-1-yl)ethan-1-imine	C12H11N7	253.107593382	1
40	C8H9NO2	DTXSID2020006	103-90-2	Acetaminophen	C8H9NO2	151.063328534	75
41	C8H9NO2	DTXSID6026567	134-20-3	Methyl 2-aminobenzoate	C8H9NO2	151.063328534	50

Data Source Ranking for Identification in SSA/NTA

- Mass and/or formula unknown to a researcher, contained within a reference database
- Most likely candidate chemicals have the most references/sources



Comparing Dashboard with ChemSpider

- On same 162 chemicals, Dashboard outperforms ChemSpider

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]

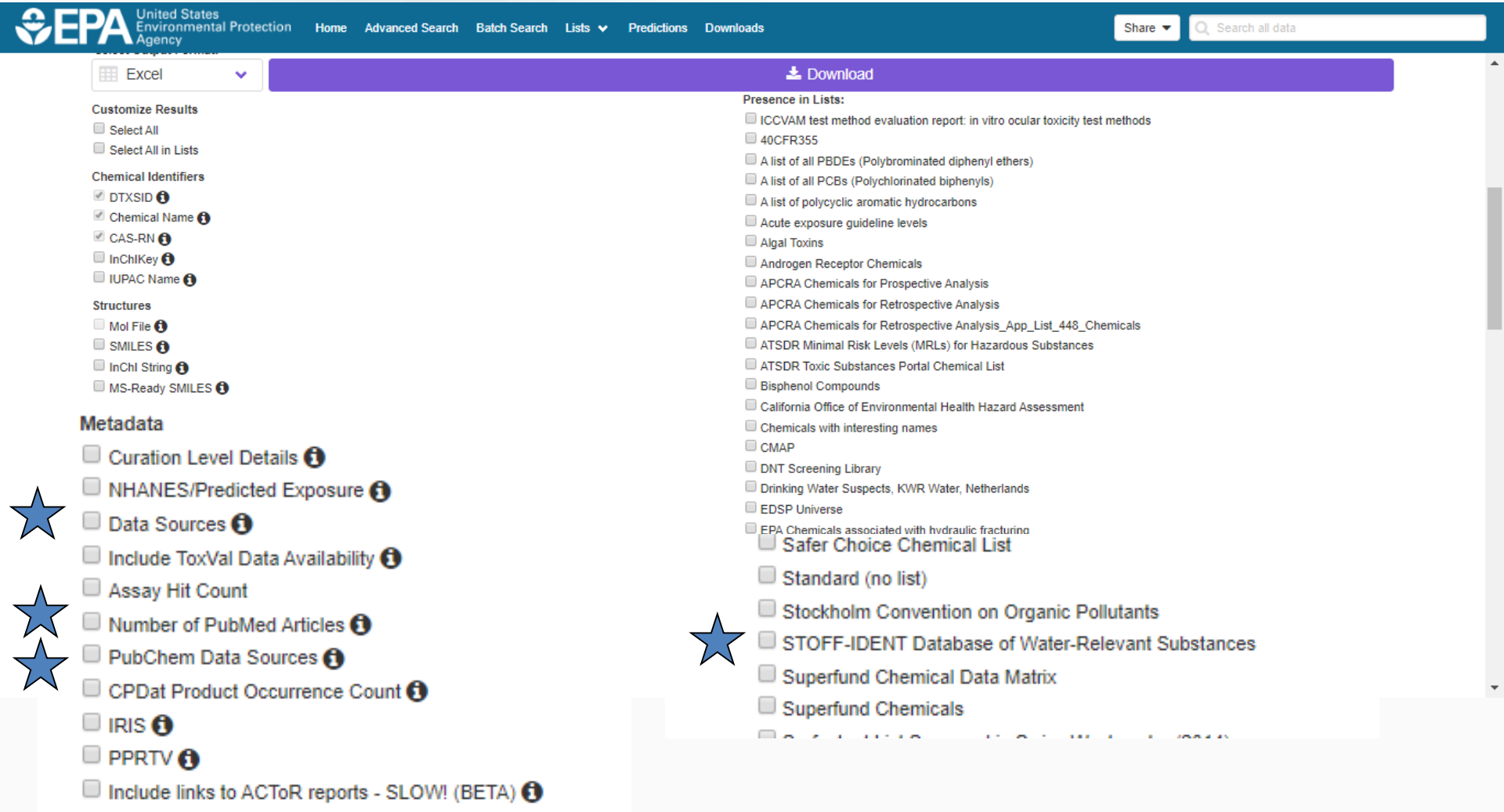
	Mass-based searching		Formula-based searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average rank position	1.3	2.2[a]	1.2	1.4
Percent in #1 position	85%	70%	88%	80%

[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

Data Streams to Improve Identifications

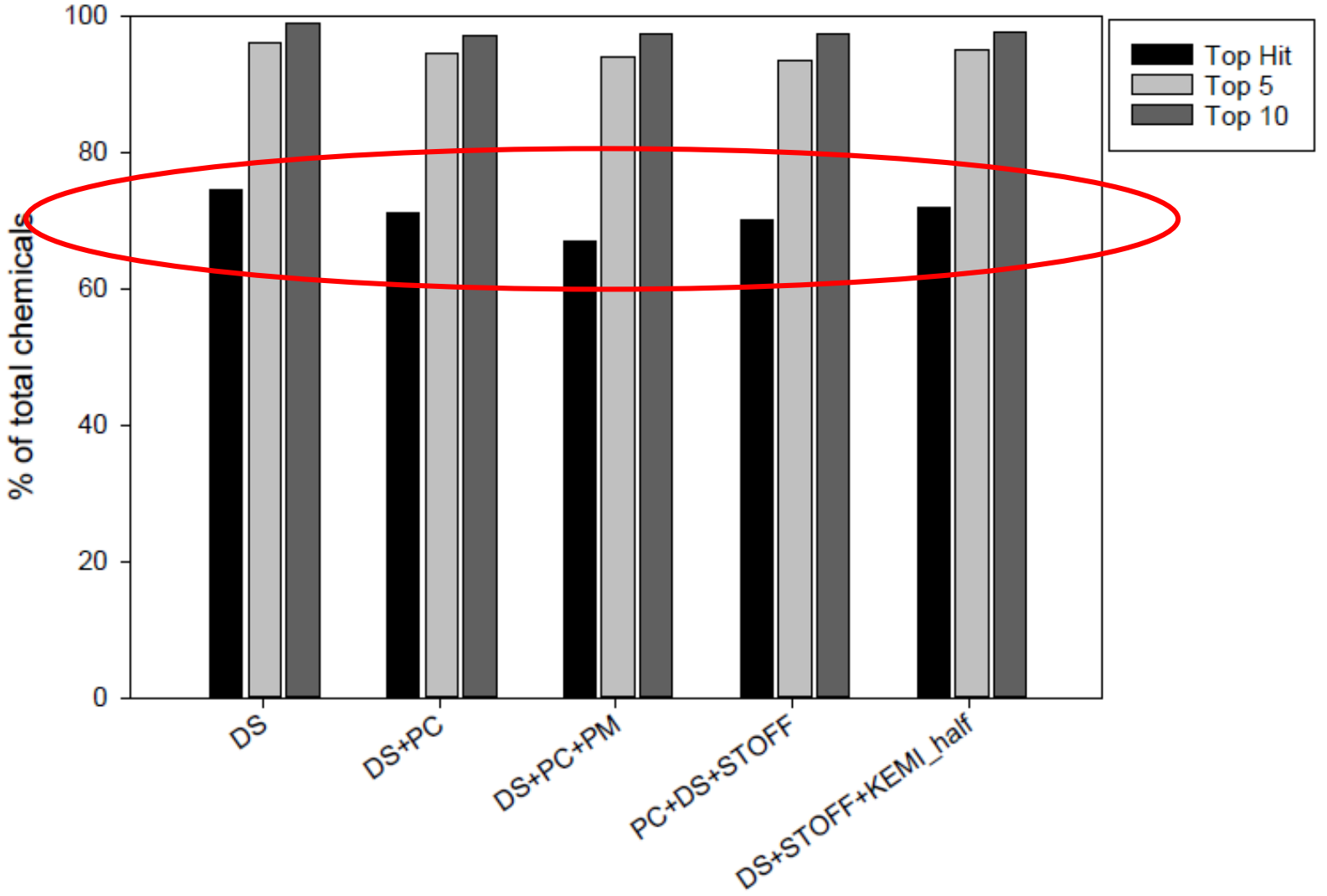
- US EPA CompTox Dashboard Data Sources (DS)
- PubChem Data Source Count
- PubMed.gov Reference Count
- Presence in ?TOFF !DENT Database
- Predicted Environmental Media Occurrence
- CPDat Product Occurrence Count
- OPERA PhysChem Properties
- NORMAN Network Priority List

All available via Batch Search

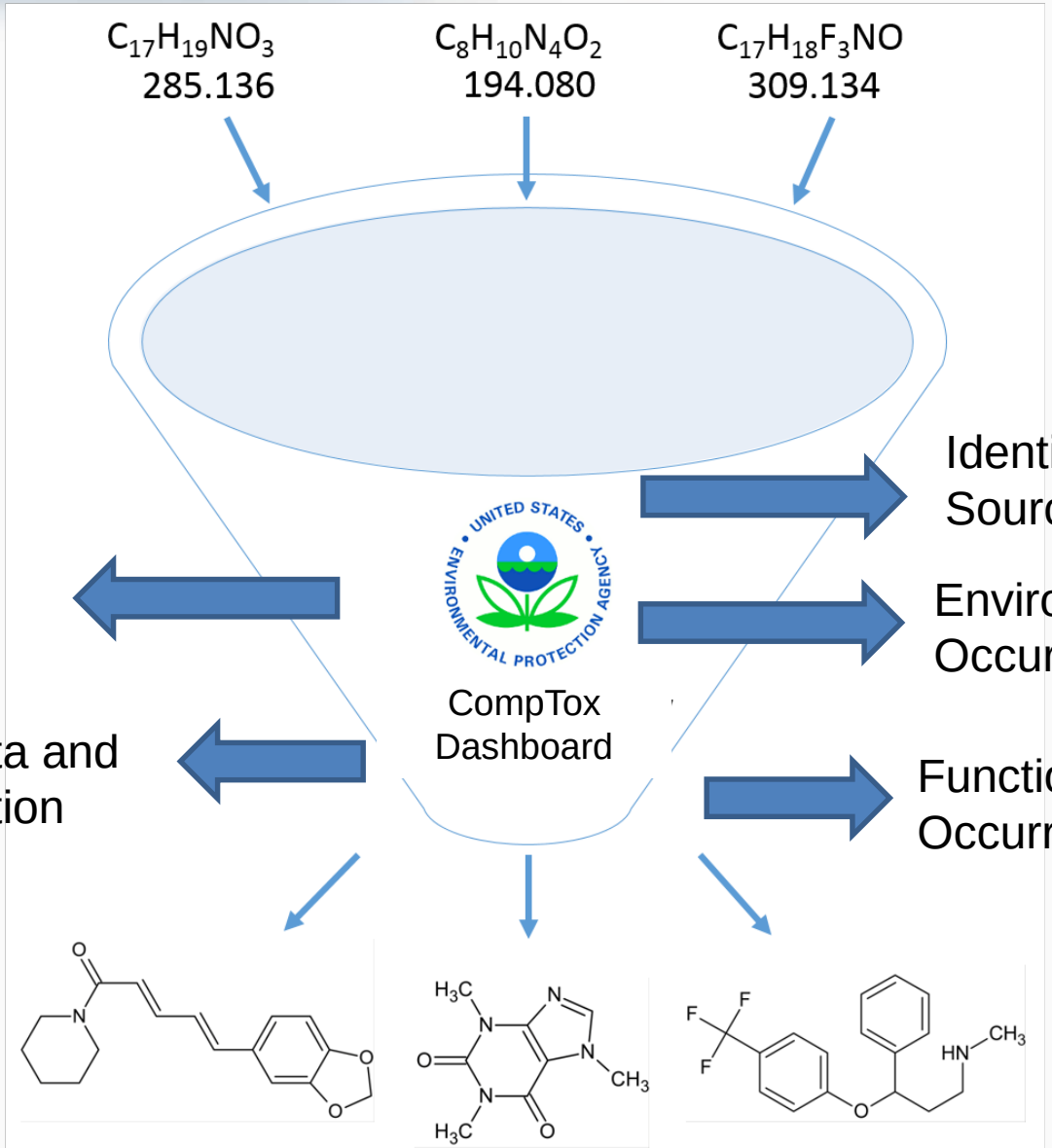


Identification ranking: 1783 chemicals with multiple data streams

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



Dashboard in NTA Workflows



Retention Time Prediction

MS/MS Data and Fragmentation Prediction

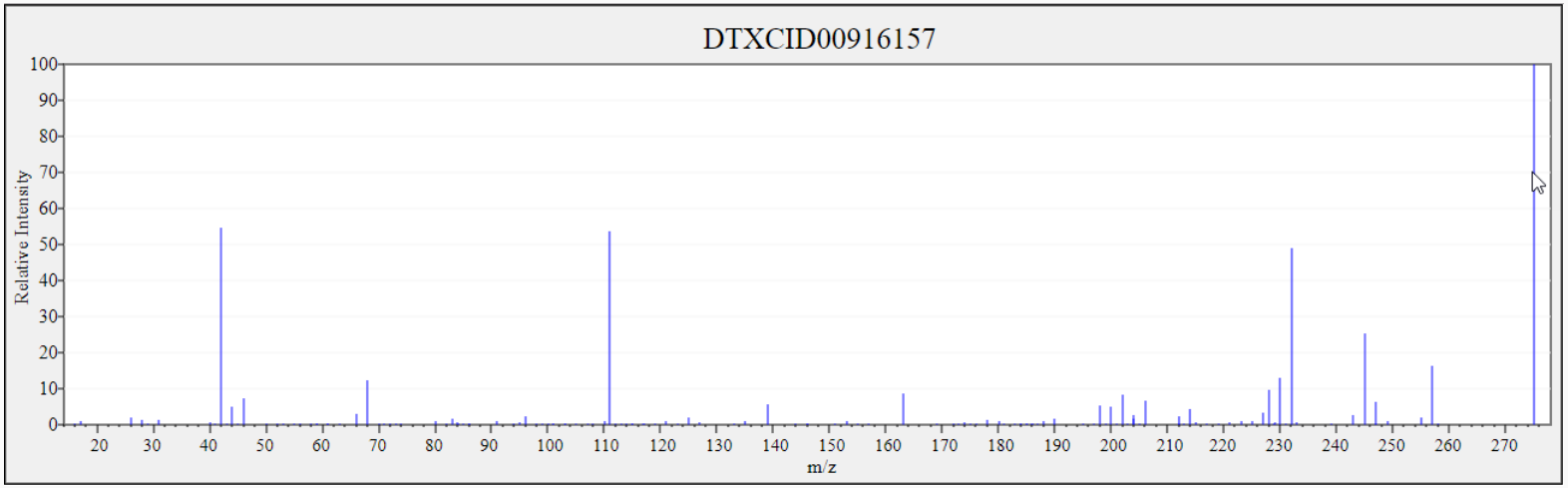
Identification by Data Source Ranking

Environmental Media Occurrence

Functional Use/Product Occurrence

Work in Progress: Incorporating MS/MS data via the Dashboard

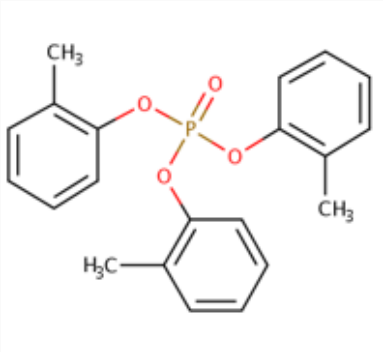
- MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >700,000 structures, to be accessible via Dashboard



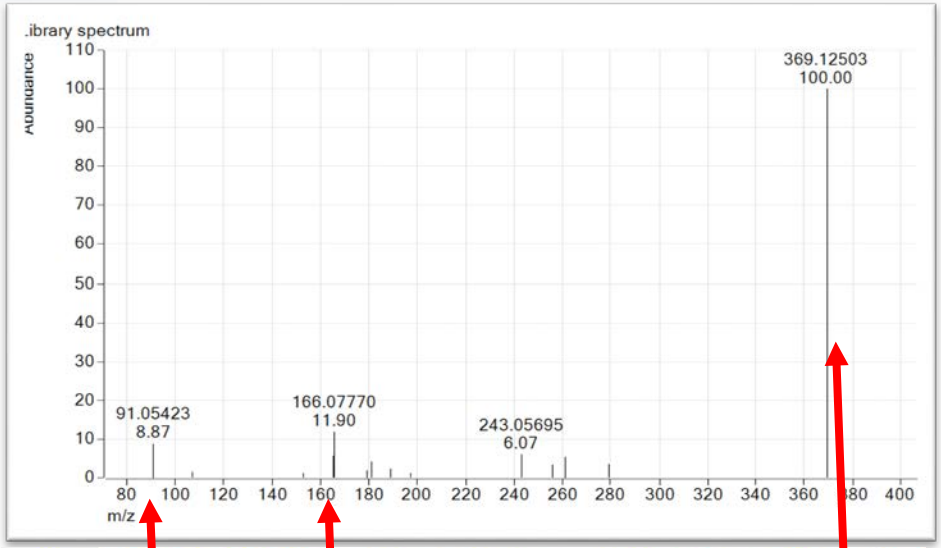
http://cfmid.wishartlab.com/
Allen, *et al* 2014, 2015, 2016

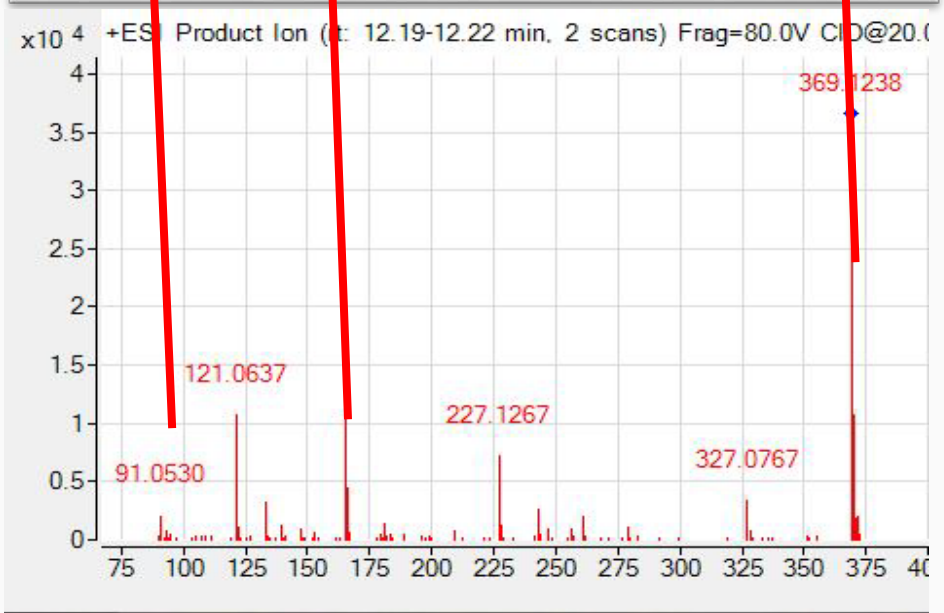
Predicted Mass Spectra

Library Fragmentation Spectra (20eV)



Observed Fragmentation Spectra (20eV)





Match Score

Evaluating on CASMI 2016

- Critical Assessment of Small Molecule Identification
  - Training data= 312 peak lists (from 285 substances)
    - 234 MS/MS in positive mode
    - 58 in negative mode
  - Challenge Data= 208 peak lists (from 188 substances)
    - 127 in positive mode
    - 81 in negative mode
- Precursor ion search window= 15 ppm
- Fragment ion match threshold= 0.02 Da
- Candidates limited to Dashboard results within precursor ion search window

http://www.casmi-contest.org/2016/index.shtml

Work in progress

CASMI 2016 Contest Challenge Set (n=208)

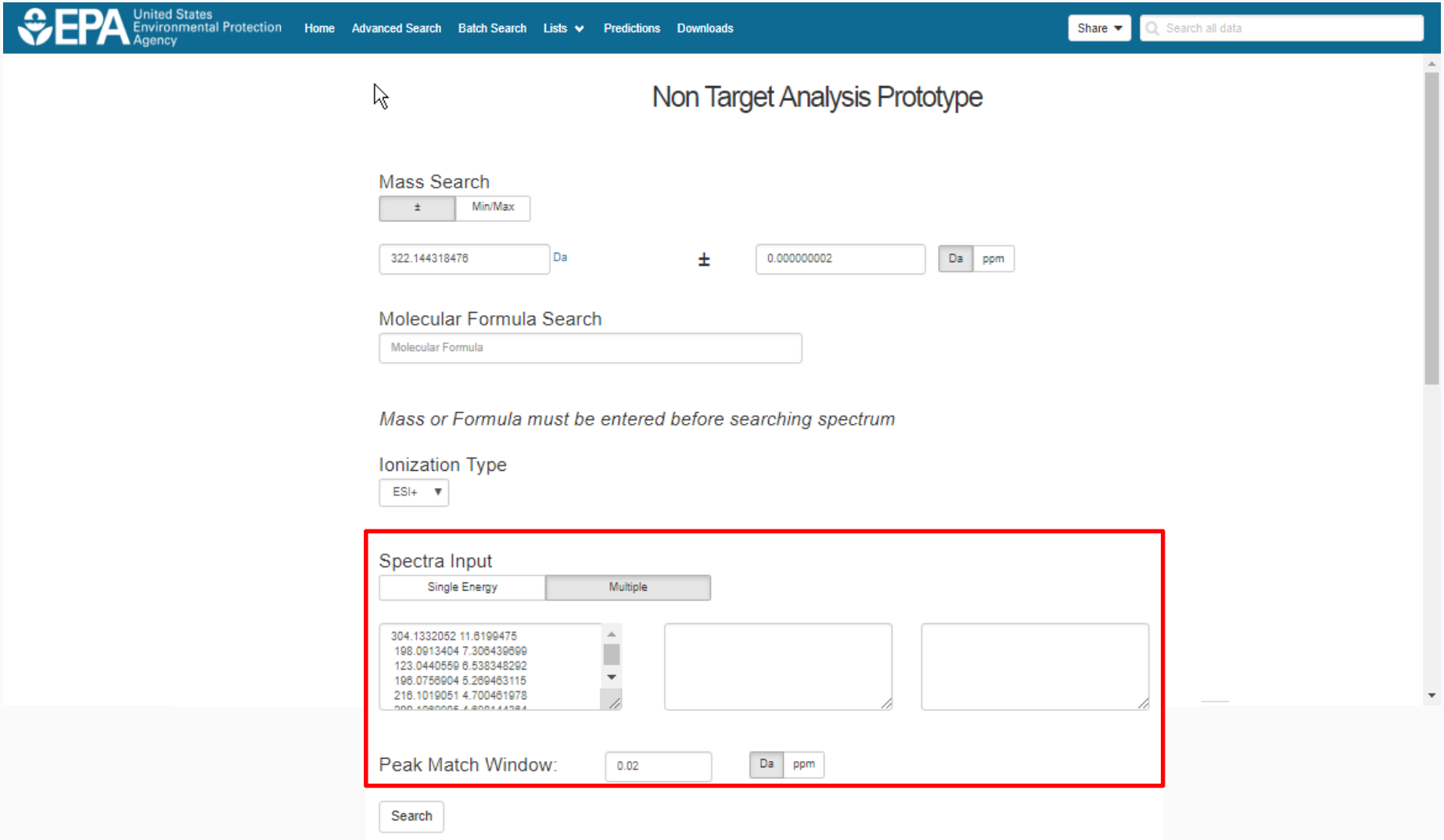
CFM-ID only

	# Identified	% of Total
#1 Hits	89	43%
Top 5	154	74%
Top 10	174	84%
Top 20	190	91%

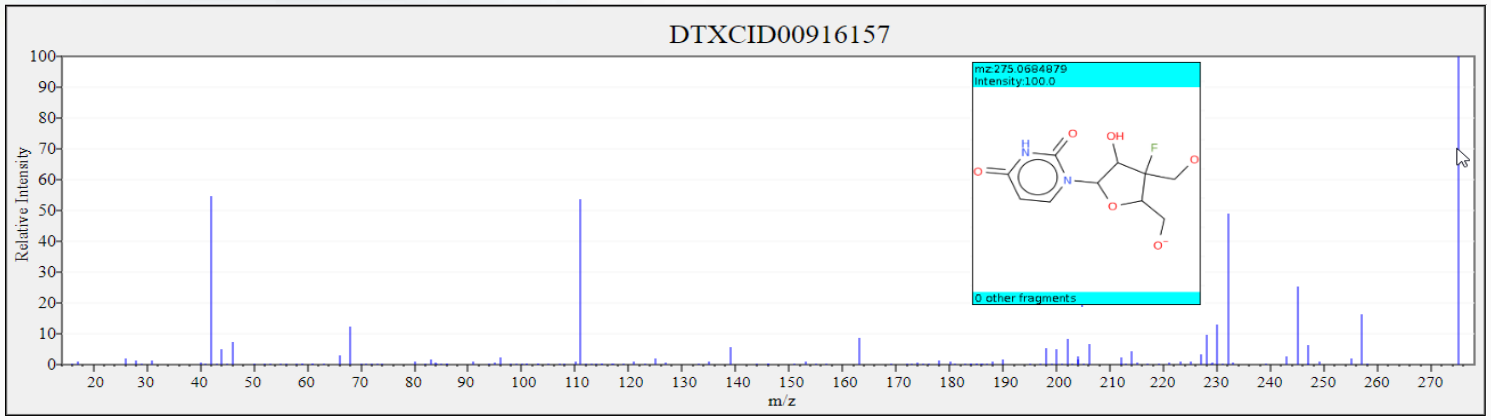
CFM-ID +DSSTox Data Sources

	# Identified	% of Total
#1 Hits	154	74%
Top 5	195	94%
Top 10	198	95%
Top 20	202	97%

Search Expt. vs. Predicted Spectra

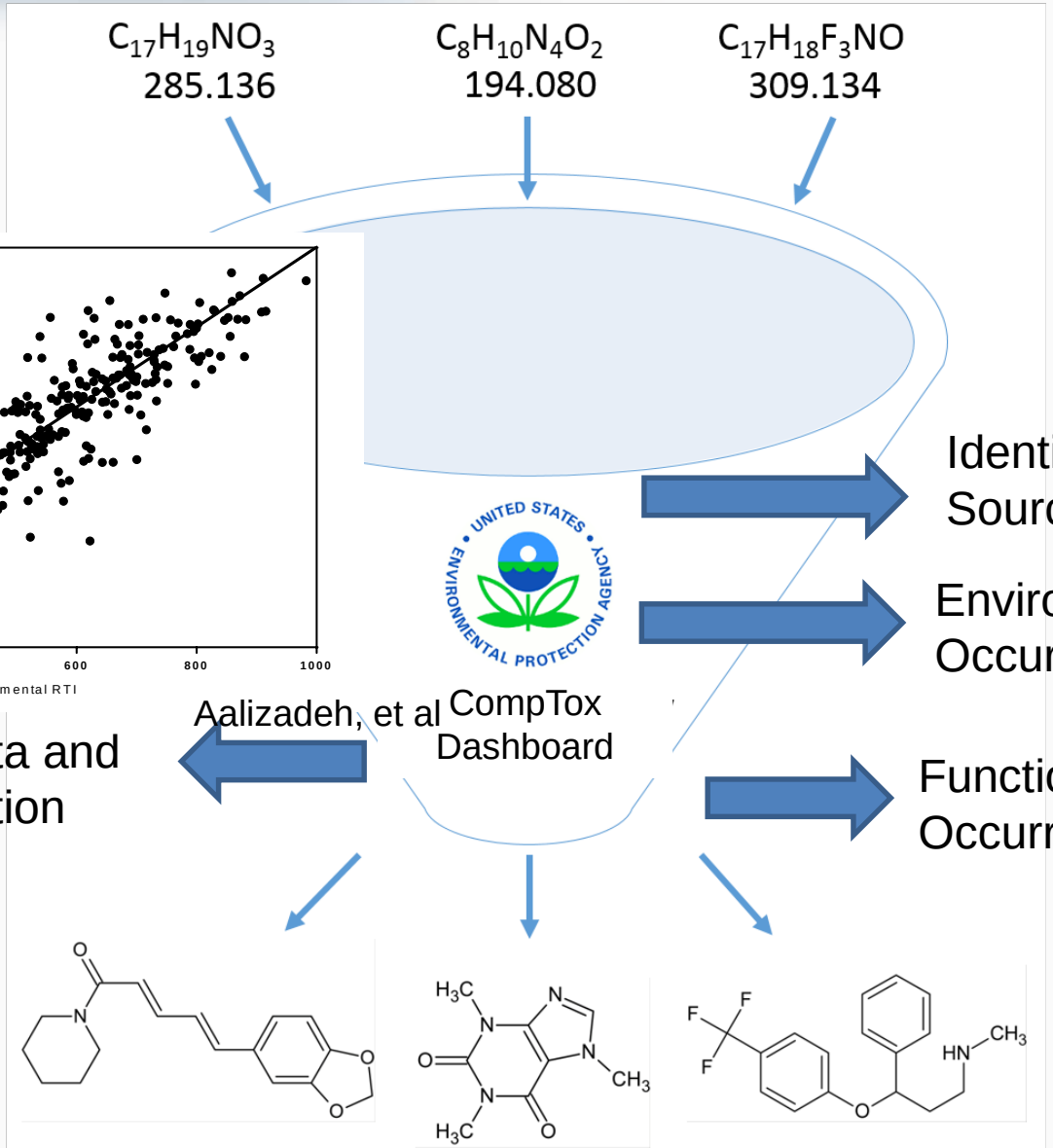


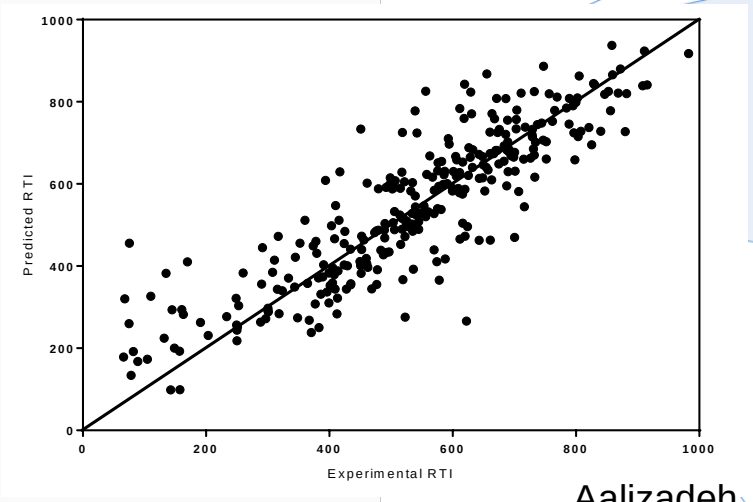
Future Access via the Dashboard



- Manuscript in Preparation for Nature Scientific Data
- Data will be available for download with publication
- (NO timeline…)

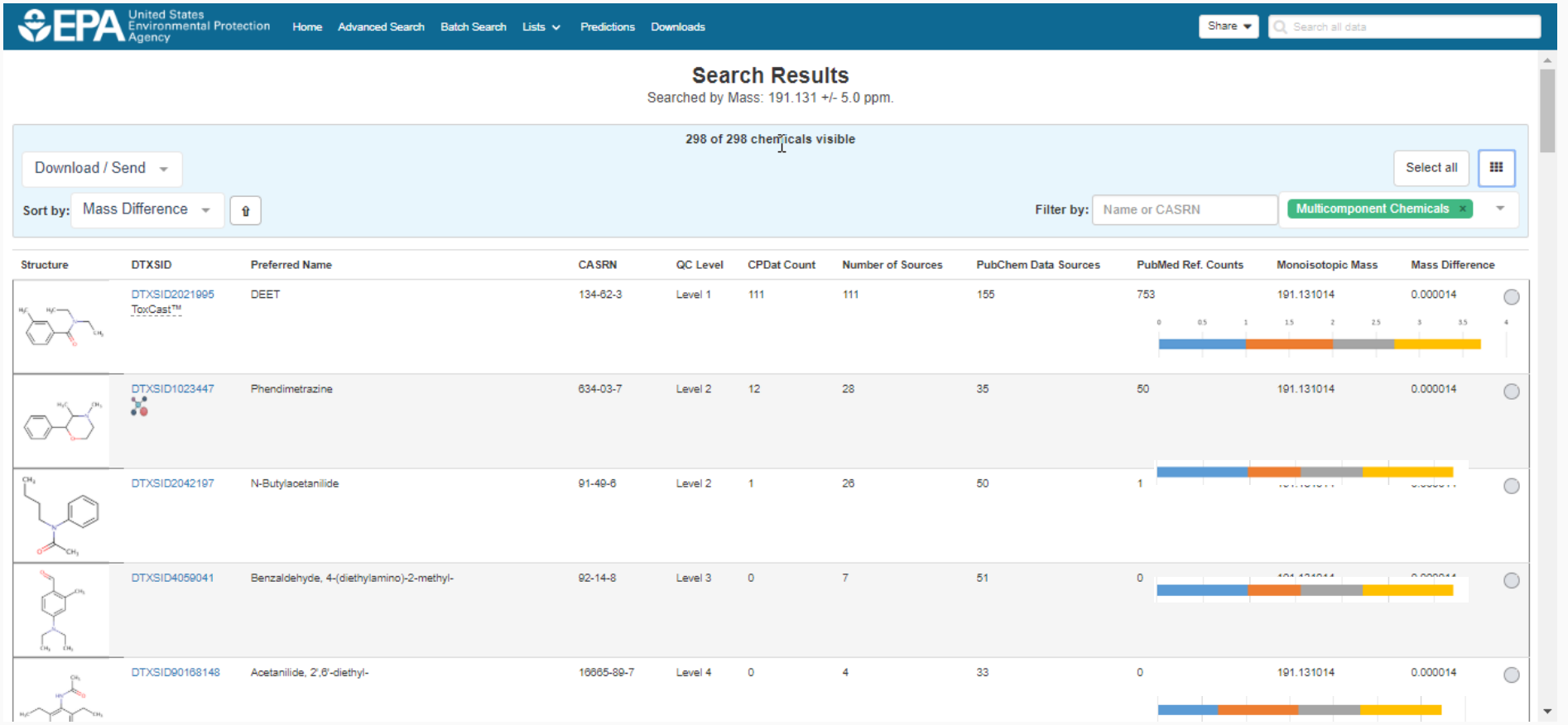
Dashboard in NTA Workflows



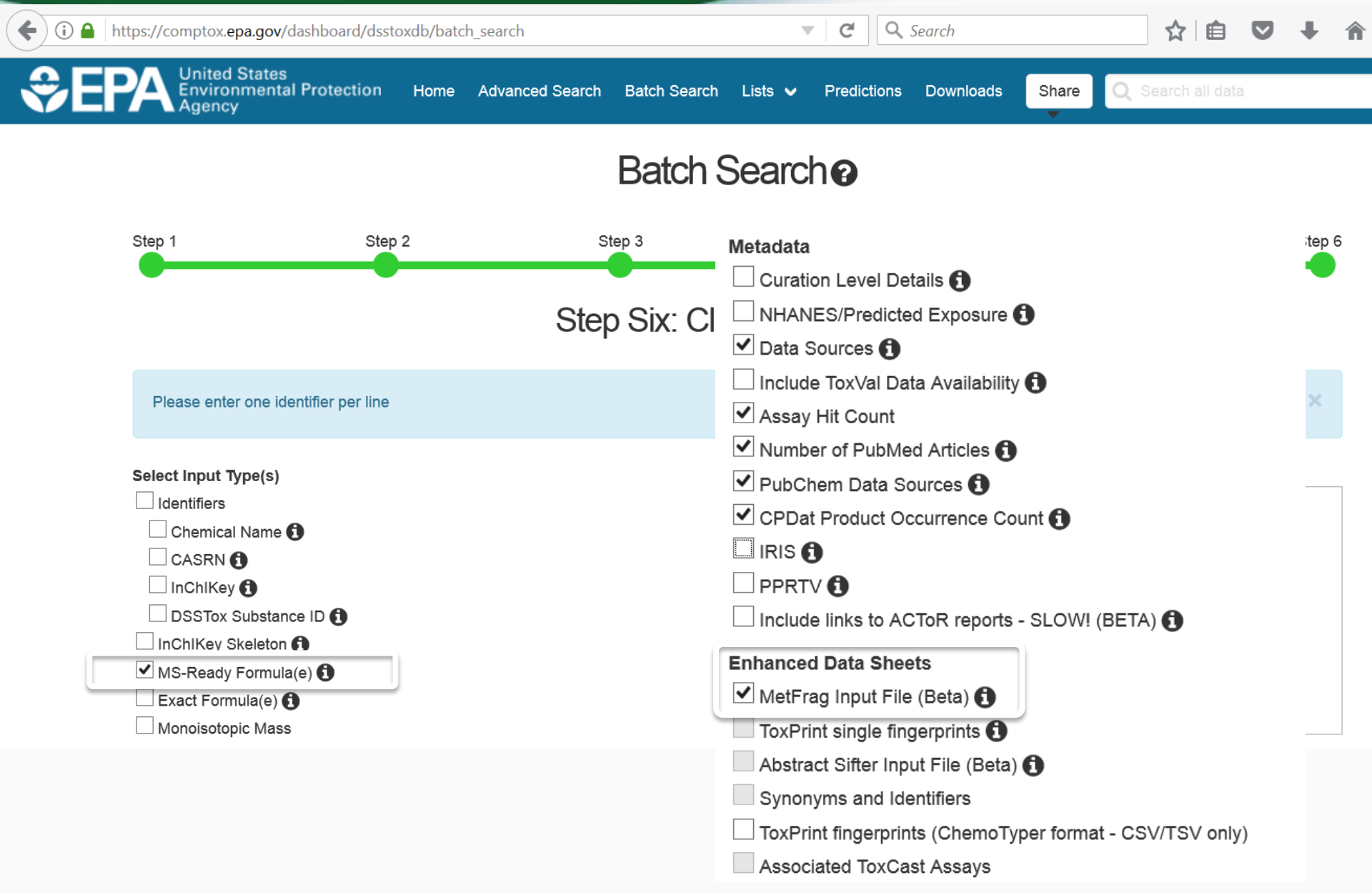


Future directions
Intuitive Visualization of Results

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



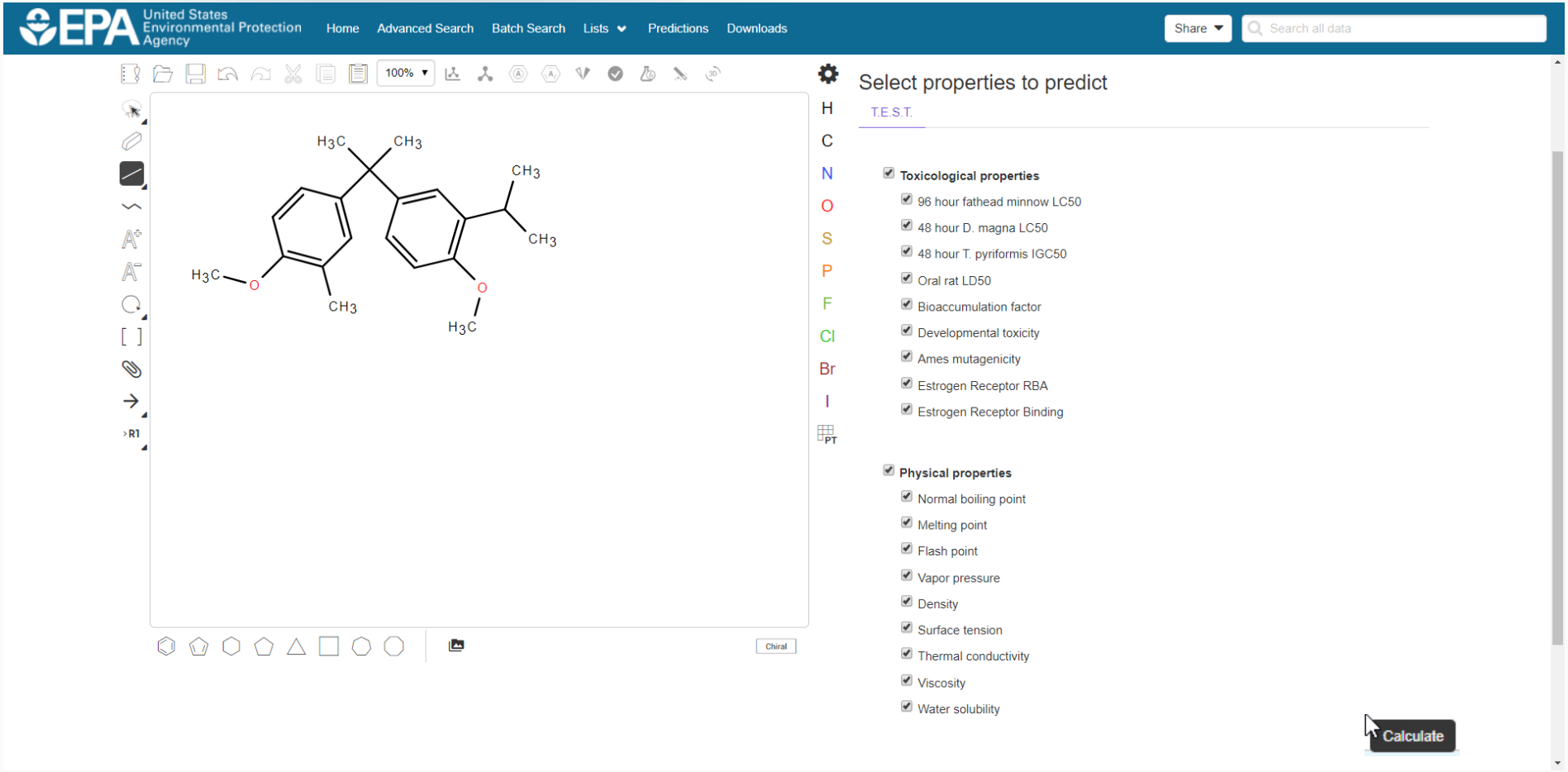
MetFrag Integration (more from Emma)



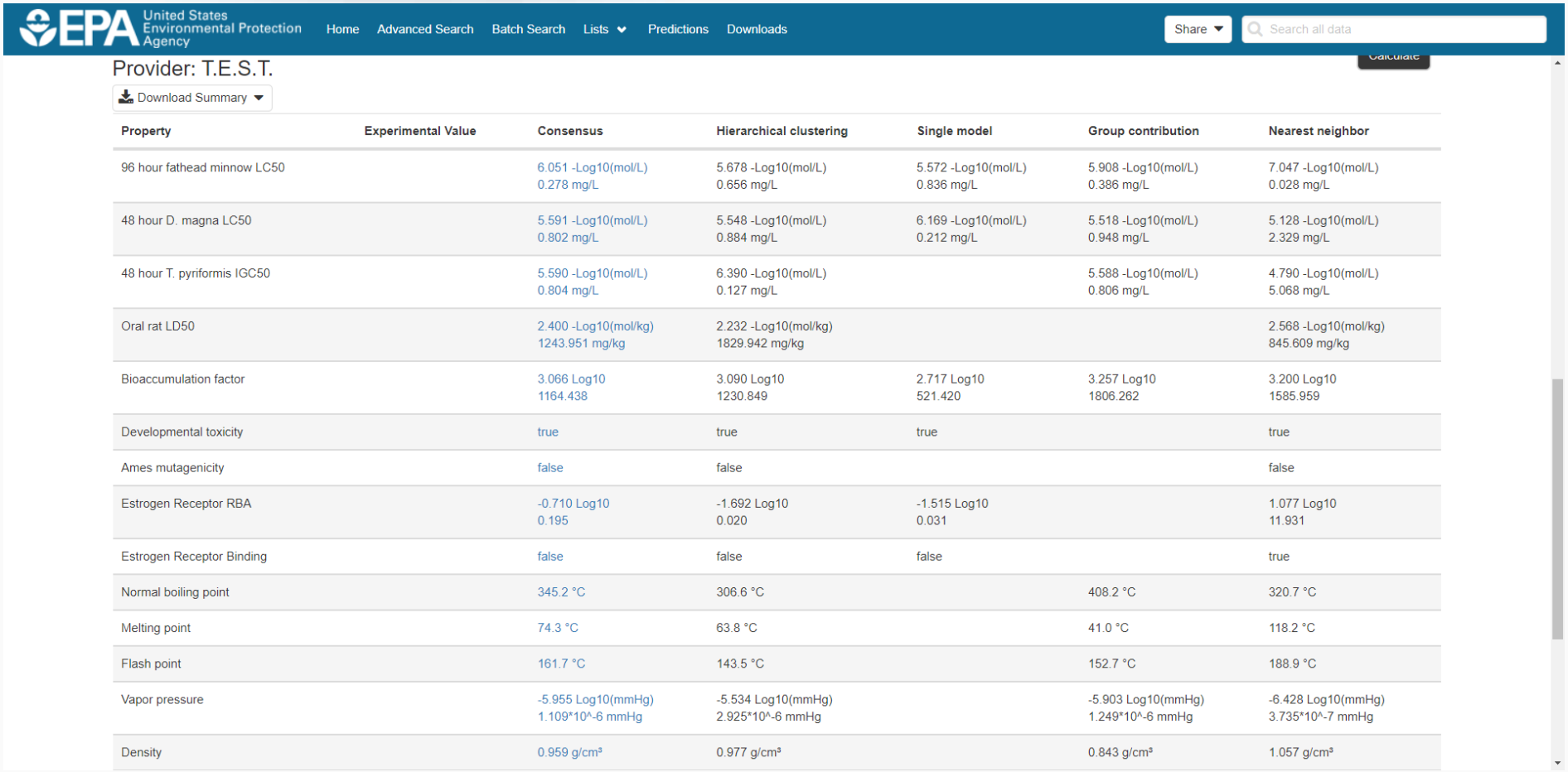
Work in Progress

- For every chemical registered into DSSTox we need to:
  - Generate QSAR-ready and MS-Ready forms
  - Pass the QSAR-ready form through predictors
    - OPERA PhysChem and Env. Fate and Transport
    - TEST PhysChem and Toxicity Models
    - Bioactivity models
  - More prediction algorithms are in development
- All predictions should be service-based
- We will deliver these services to users (eventually)

Real-Time Predictions



Real-Time Predictions



Property	Experimental Value	Consensus	Hierarchical clustering	Single model	Group contribution	Nearest neighbor
96 hour fathead minnow LC50		6.051 -Log10(mol/L) 0.278 mg/L	5.678 -Log10(mol/L) 0.656 mg/L	5.572 -Log10(mol/L) 0.836 mg/L	5.908 -Log10(mol/L) 0.386 mg/L	7.047 -Log10(mol/L) 0.028 mg/L
48 hour D. magna LC50		5.591 -Log10(mol/L) 0.802 mg/L	5.548 -Log10(mol/L) 0.884 mg/L	6.169 -Log10(mol/L) 0.212 mg/L	5.518 -Log10(mol/L) 0.948 mg/L	5.128 -Log10(mol/L) 2.329 mg/L
48 hour T. pyriformis IGC50		5.590 -Log10(mol/L) 0.804 mg/L	6.390 -Log10(mol/L) 0.127 mg/L		5.588 -Log10(mol/L) 0.806 mg/L	4.790 -Log10(mol/L) 5.068 mg/L
Oral rat LD50		2.400 -Log10(mol/kg) 1243.951 mg/kg	2.232 -Log10(mol/kg) 1829.942 mg/kg			2.568 -Log10(mol/kg) 845.609 mg/kg
Bioaccumulation factor		3.066 Log10 1164.438	3.090 Log10 1230.849	2.717 Log10 521.420	3.257 Log10 1806.262	3.200 Log10 1585.959
Developmental toxicity		true	true	true		true
Ames mutagenicity		false	false			false
Estrogen Receptor RBA		-0.710 Log10 0.195	-1.692 Log10 0.020	-1.515 Log10 0.031		1.077 Log10 11.931
Estrogen Receptor Binding		false	false	false		true
Normal boiling point		345.2 °C	306.6 °C		408.2 °C	320.7 °C
Melting point		74.3 °C	63.8 °C		41.0 °C	118.2 °C
Flash point		161.7 °C	143.5 °C		152.7 °C	188.9 °C
Vapor pressure		-5.955 Log10(mmHg) 1.109*10^-6 mmHg	-5.534 Log10(mmHg) 2.925*10^-6 mmHg		-5.903 Log10(mmHg) 1.249*10^-6 mmHg	-6.428 Log10(mmHg) 3.735*10^-7 mmHg
Density		0.959 g/cm³	0.977 g/cm³		0.843 g/cm³	1.057 g/cm³

API in development
Prototype services available

https://comptox.epa.gov/dashboard/web-test/WS?smiles=CCO&method=hc

JSON Raw Data Headers

Save Copy

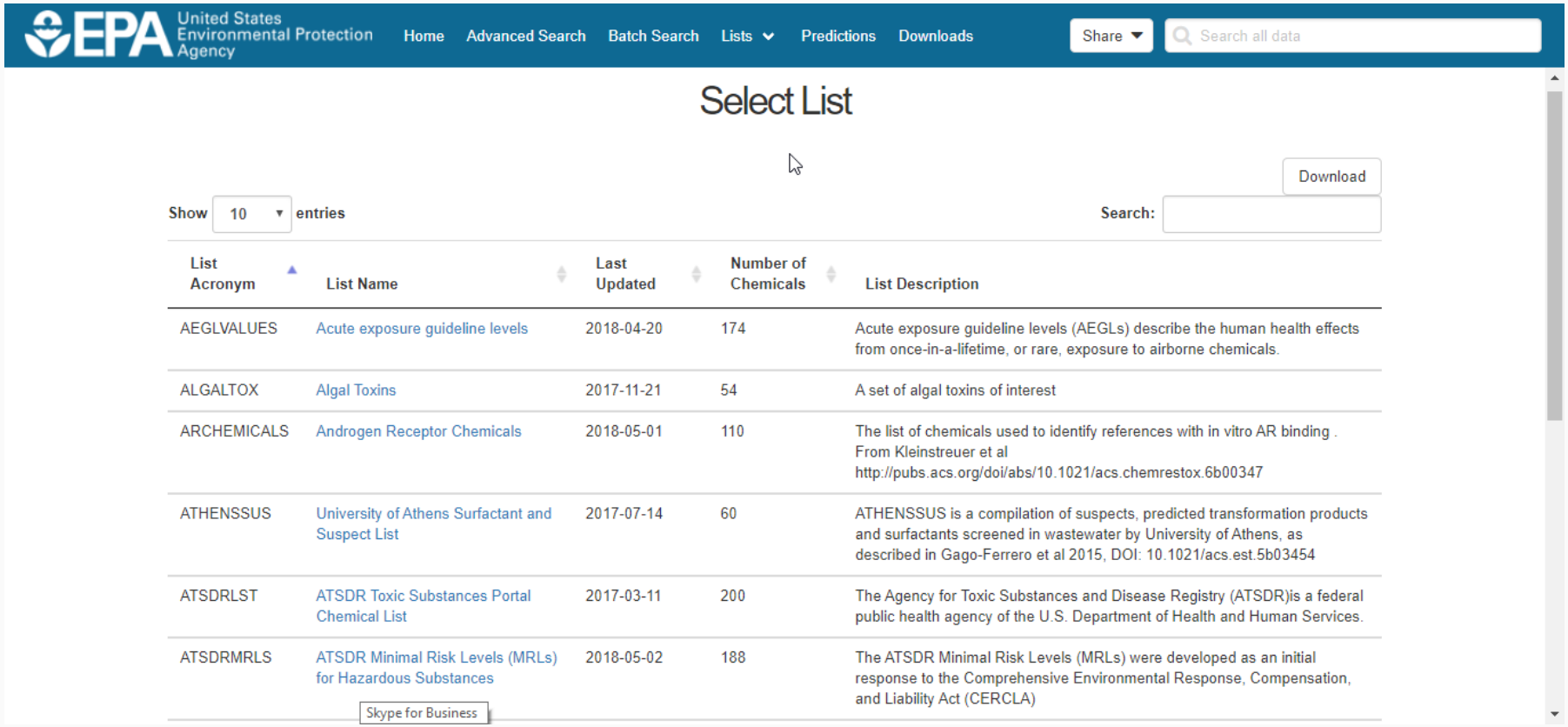

```
uuid:                "55547f4f-f966-48e8-b831-a0d217998064"
predictionTime:      1520539090089
software:            "T.E.S.T (Toxicity Estimation Software Tool)"
softwareVersion:     "5.01"
condition:           "25°C"
endpoint:            "Water solubility at 25°C"
method:              "Hierarchical clustering"
predictions:
  0:
    id:              "C_1520539090089"
    smiles:          "OCC"
    expValMolarLog:  "-1.337"
    expValMass:      "1001180.703"
    predValMolarLog: "-1.338"
    predValMass:     "1002625.241"
    molarLogUnits:   "-Log10(mol/L)"
    massUnits:       "mg/L"
```

Future work

- Prototype asynchronous pipelining of data already developed – to be integrated
- First prototype of QSAR/MS-ready processing about to go into testing
- Future plans (no deadline)
  - Batch processing of files to produce QSAR/MS-Ready (we encourage standard processes for QSAR modeling)
  - Web service access to QSAR/MS-Ready file generation

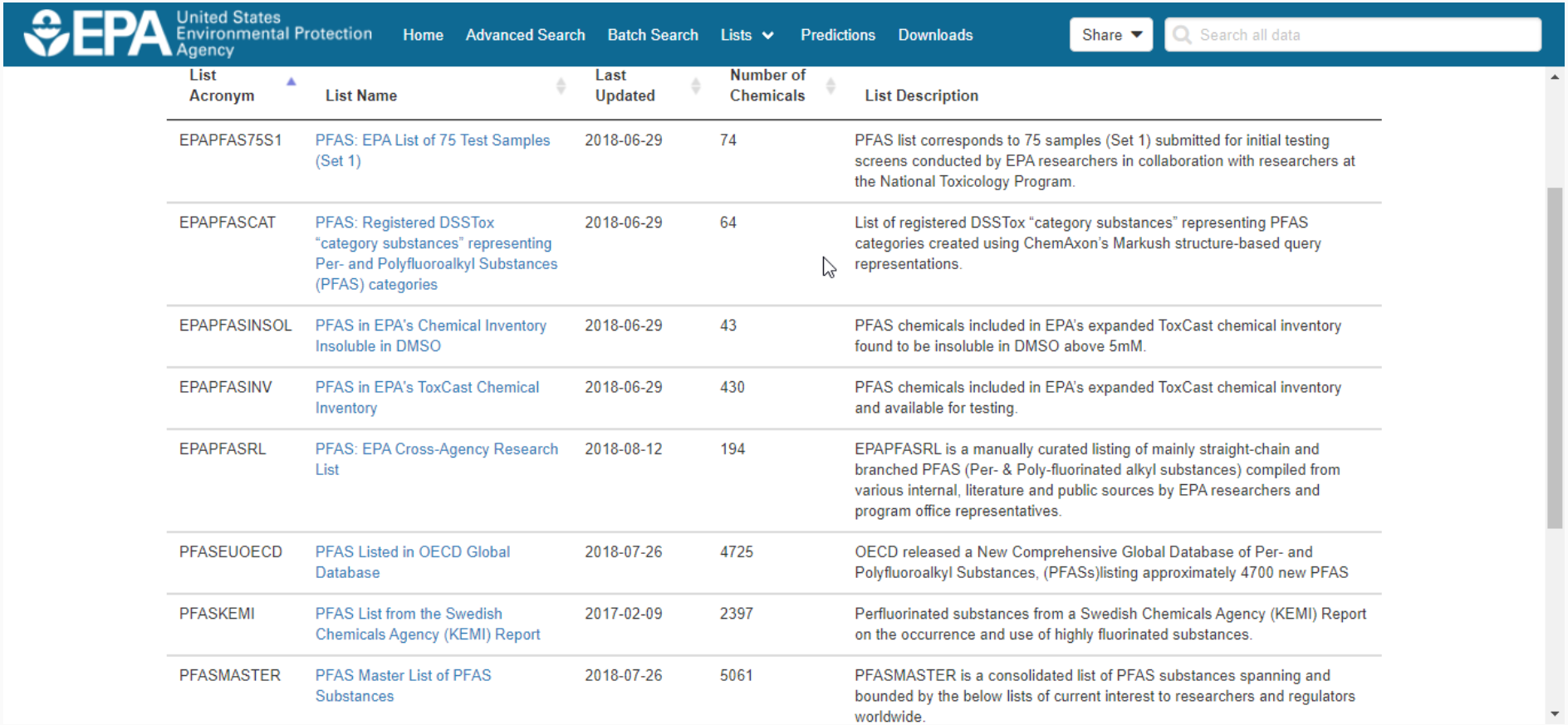
A List of Lists of Chemicals

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



Eleven PFAS Lists

http://comptox-prod.epa.gov/dashboard/chemical_lists



List Acronym	List Name	Last Updated	Number of Chemicals	List Description
EPAPFAS75S1	PFAS: EPA List of 75 Test Samples (Set 1)	2018-06-29	74	PFAS list corresponds to 75 samples (Set 1) submitted for initial testing screens conducted by EPA researchers in collaboration with researchers at the National Toxicology Program.
EPAPFASCAT	PFAS: Registered DSSTox "category substances" representing Per- and Polyfluoroalkyl Substances (PFAS) categories	2018-06-29	64	List of registered DSSTox "category substances" representing PFAS categories created using ChemAxon's Markush structure-based query representations.
EPAPFASINSOL	PFAS in EPA's Chemical Inventory Insoluble in DMSO	2018-06-29	43	PFAS chemicals included in EPA's expanded ToxCast chemical inventory found to be insoluble in DMSO above 5mM.
EPAPFASINV	PFAS in EPA's ToxCast Chemical Inventory	2018-06-29	430	PFAS chemicals included in EPA's expanded ToxCast chemical inventory and available for testing.
EPAPFASRL	PFAS: EPA Cross-Agency Research List	2018-08-12	194	EPAPFASRL is a manually curated listing of mainly straight-chain and branched PFAS (Per- & Poly-fluorinated alkyl substances) compiled from various internal, literature and public sources by EPA researchers and program office representatives.
PFASEUOECD	PFAS Listed in OECD Global Database	2018-07-26	4725	OECD released a New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances, (PFASs)listing approximately 4700 new PFAS
PFASKEMI	PFAS List from the Swedish Chemicals Agency (KEMI) Report	2017-02-09	2397	Perfluorinated substances from a Swedish Chemicals Agency (KEMI) Report on the occurrence and use of highly fluorinated substances.
PFASMASTER	PFAS Master List of PFAS Substances	2018-07-26	5061	PFASMASTER is a consolidated list of PFAS substances spanning and bounded by the below lists of current interest to researchers and regulators worldwide.

The OECD List of PFAS

http://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/

EPA United States Environmental Protection Agency

The OECD releases a new list of PFASs

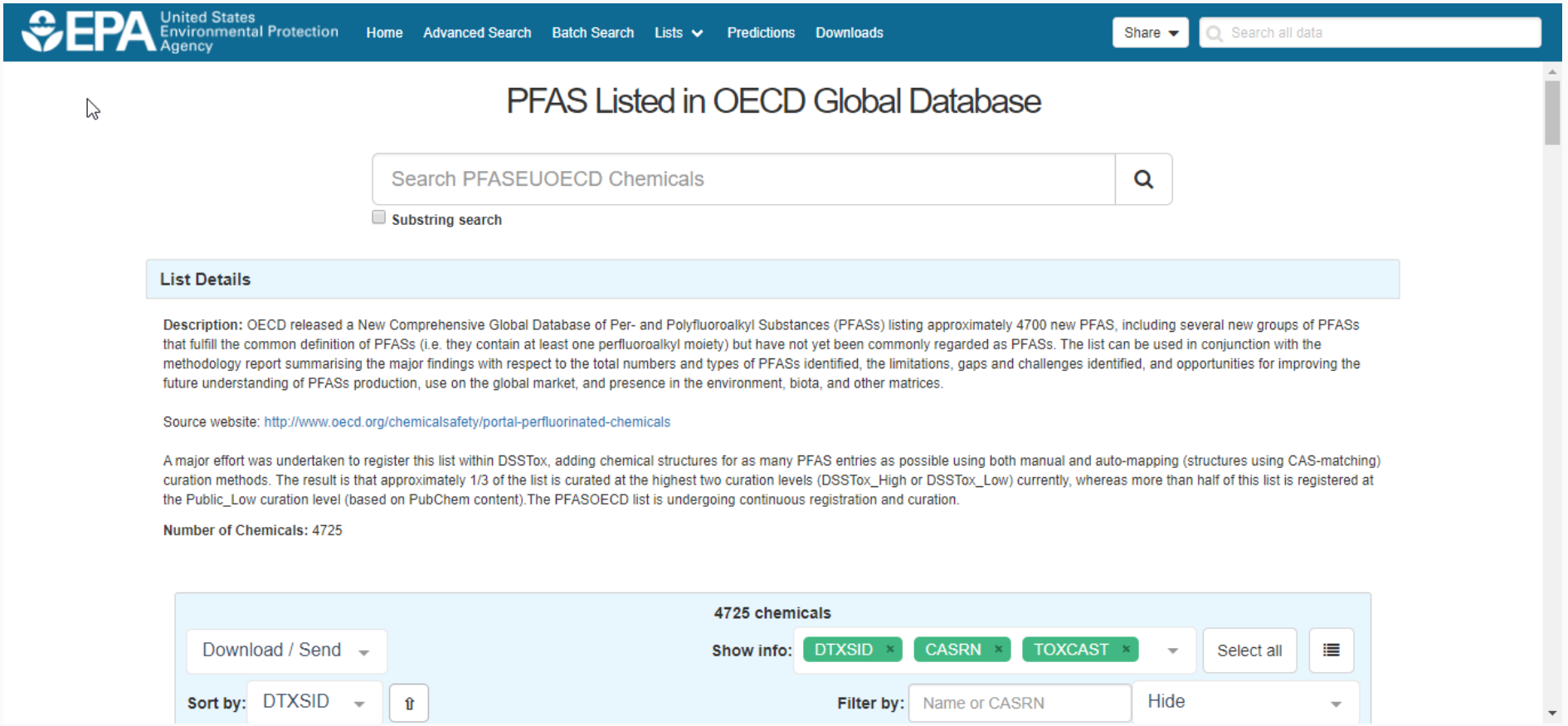
The OECD releases a new list of Per- and Polyfluoroalkyl Substances (PFASs) based on a comprehensive analysis of information available in the public domain. In total, 4730 PFAS-related CAS numbers have been identified and categorised in this study, including several new groups of PFASs that fulfil the common definition of PFASs (i.e. they contain at least one perfluoroalkyl moiety) but have not yet been commonly regarded as PFASs.

This work has been conducted under the OECD/UN Environment Global PFC Group in support of the Strategic Approach to International Chemicals Management (SAICM) and shifting to safer alternatives for PFASs.

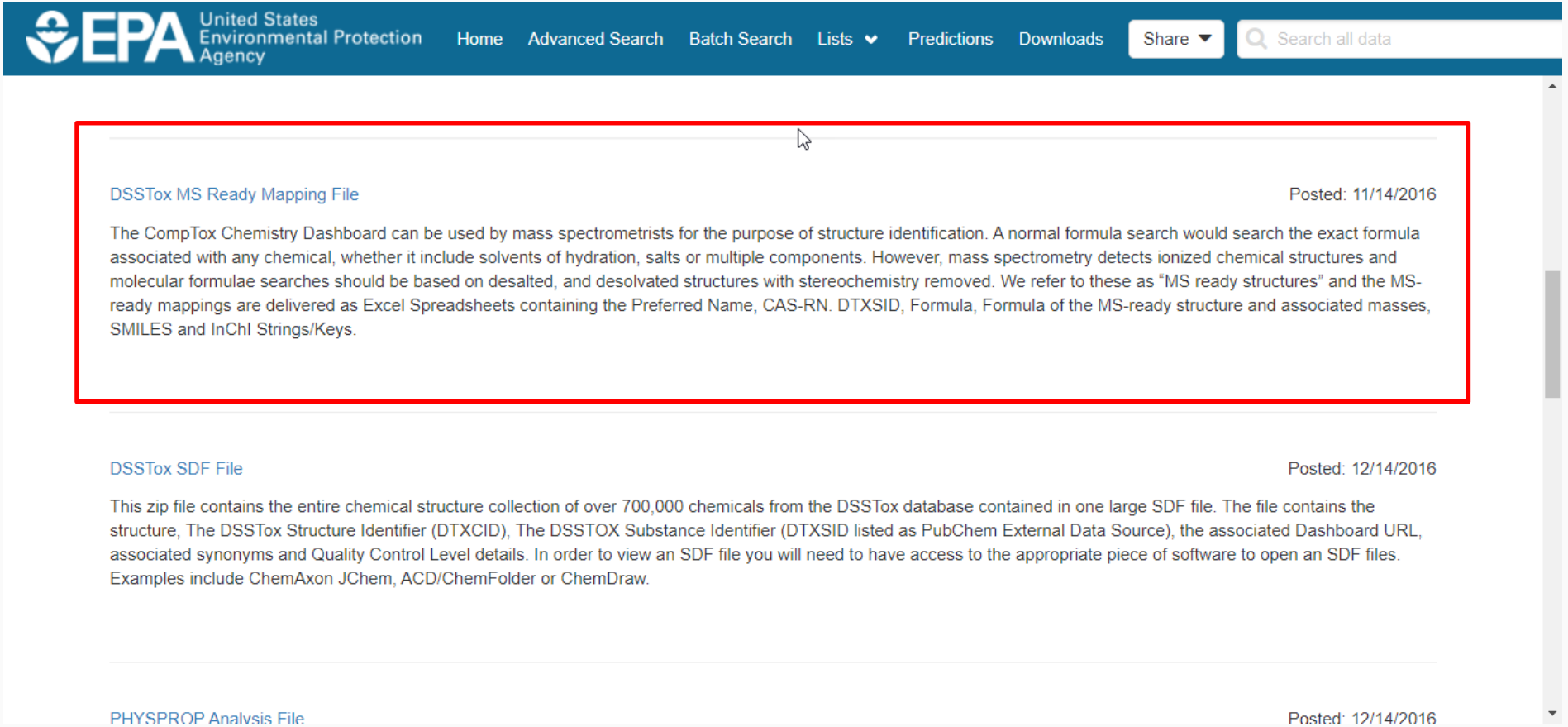
The New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances (PFASs) comes with a methodology report also detailing the major findings with respect to the total numbers and types of PFASs identified, the limitations, gaps and challenges identified in the development of the new list, and opportunities for improving the future understanding of PFASs production, use on the global market, and presence in the environment, biota, and other matrices.

The OECD List of PFAS

http://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/



Downloadable Data



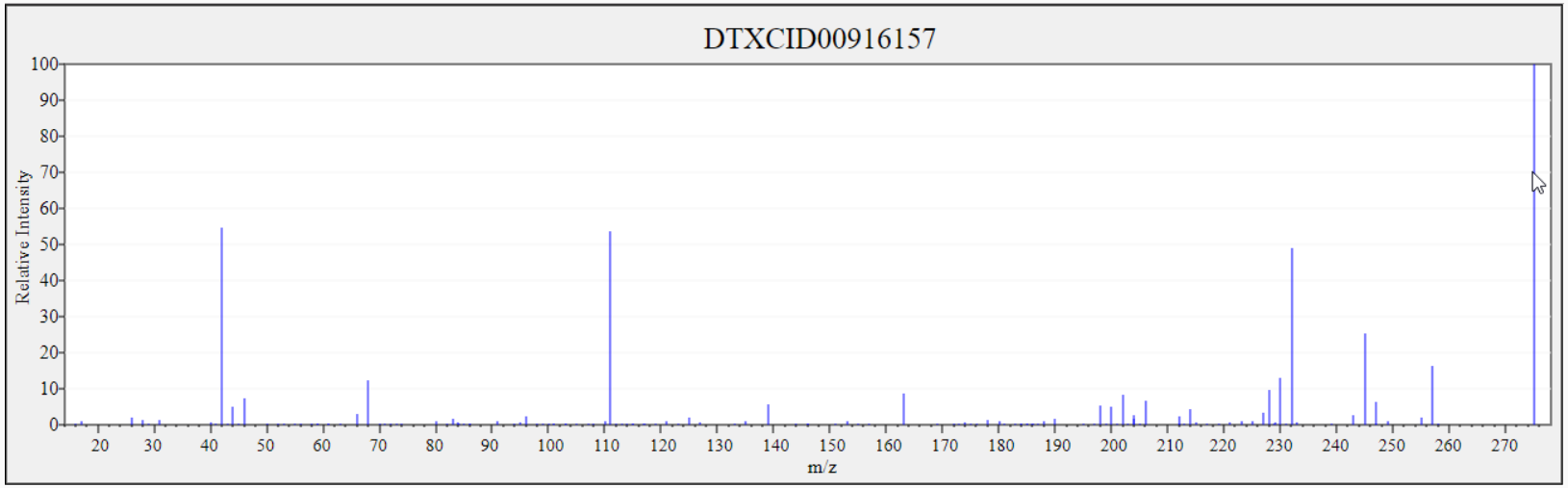
Work in Progress

- CFM-ID
  - Viewing and Downloading pre-predicted spectra
  - Search spectra against the database
- MetFrag integration (beta already done)
- Investigating Retention Time Index Prediction for candidate ranking
- Structure/substructure/similarity search

Predicted Mass Spectra

http://cfmid.wishartlab.com/

- MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >700,000 structures, to be accessible via Dashboard



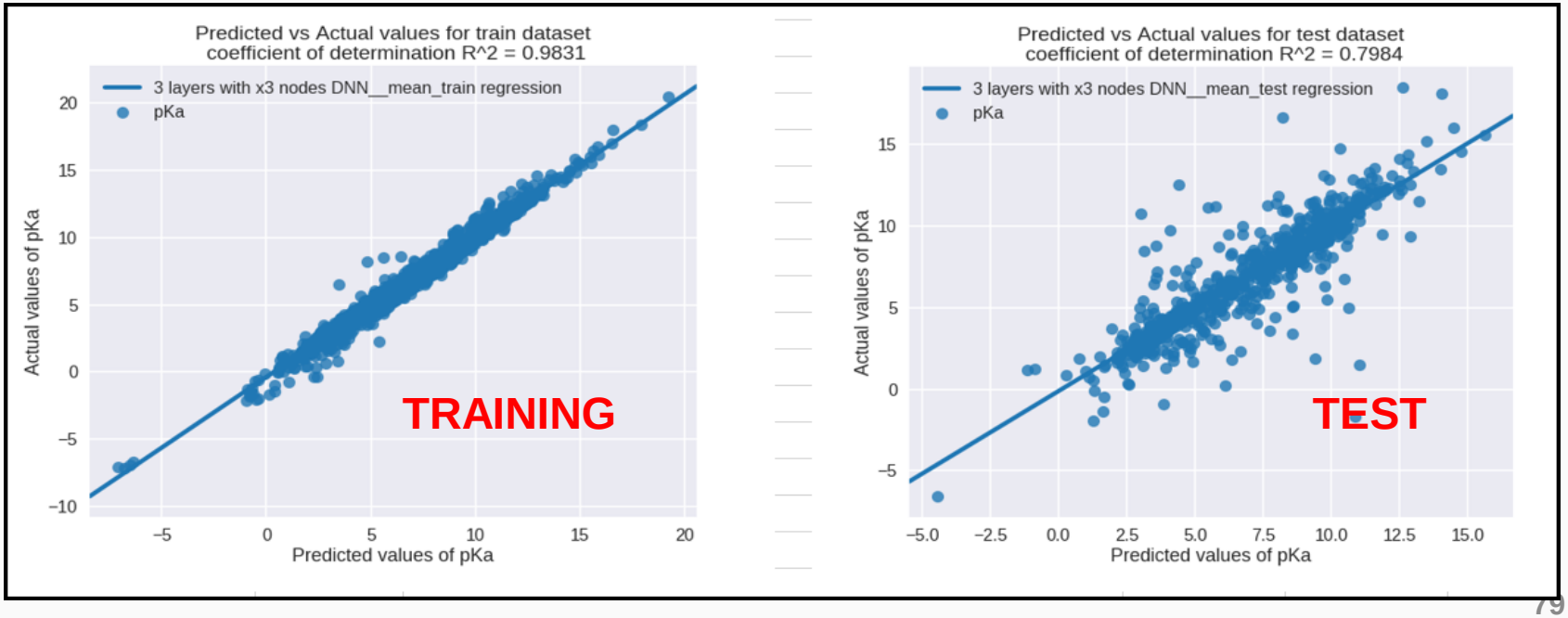
What structures are amenable to LC vs GC? Prediction models

- For many of our studies we would like to know whether chemicals are amenable to LC vs GC, +ve vs –ve ion etc?

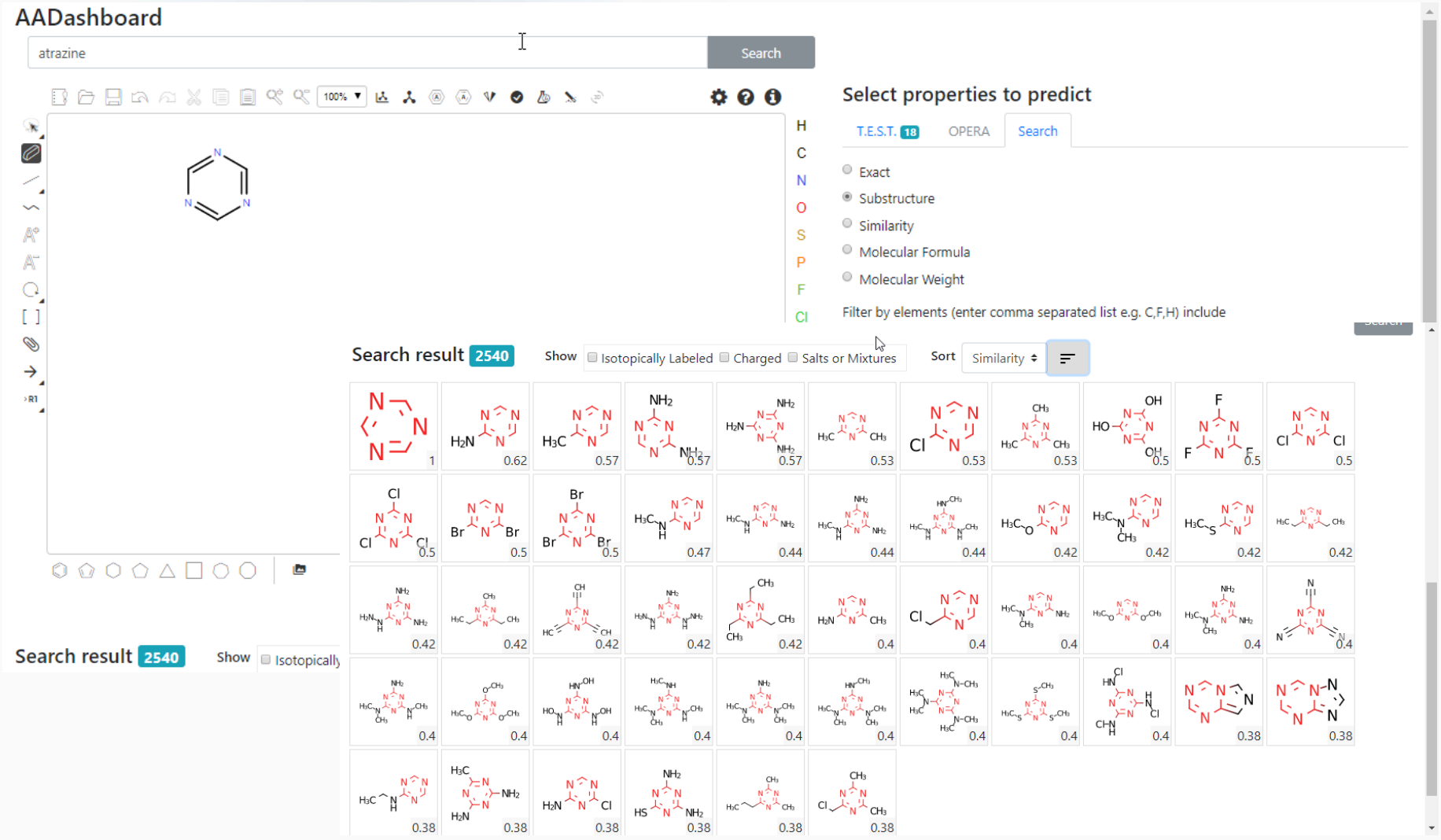
- Public databases are being harvested in order to build training sets for prediction models and availability of public data

pKa Prediction Model

- pKa prediction models based on Open Data Set of 8000 chemicals – acidic, basic and amphoteric chemicals
- pKa for prediction of LC/GC amenable??



Prototype Development



Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~762,000 chemicals
- High quality data from ongoing curation efforts
- An integration hub for multiple “modules”
- The dashboard serves many purposes but has functionality to support Mass Spectroscopy
- MS-Ready processing of chemicals has high value

- With much work to do, we are well on the way…

Acknowledgments

- The CompTox Dashboard Development Team
- NCCT scientists providing data and modules
- The NERL Mass Spectrometry Team
- Our wonderful team of curators
- Tommy Cathey, Tom Transue and Ilya Balyabin (CFM-ID fragmentation work)
- Multiple external collaborators via NORMAN
- Emma Schymanski (so many contributions)
- Christoph Ruttkies (MetFrag)

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