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

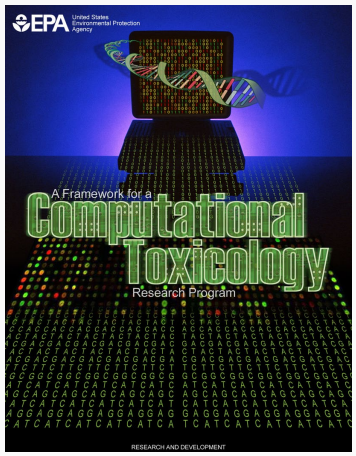
The EPA Comptox Chemicals Dashboard as a Data Integration Hub for Environmental Chemistry Data

Antony Williams

U.S. Environmental Protection Agency, RTP, NC

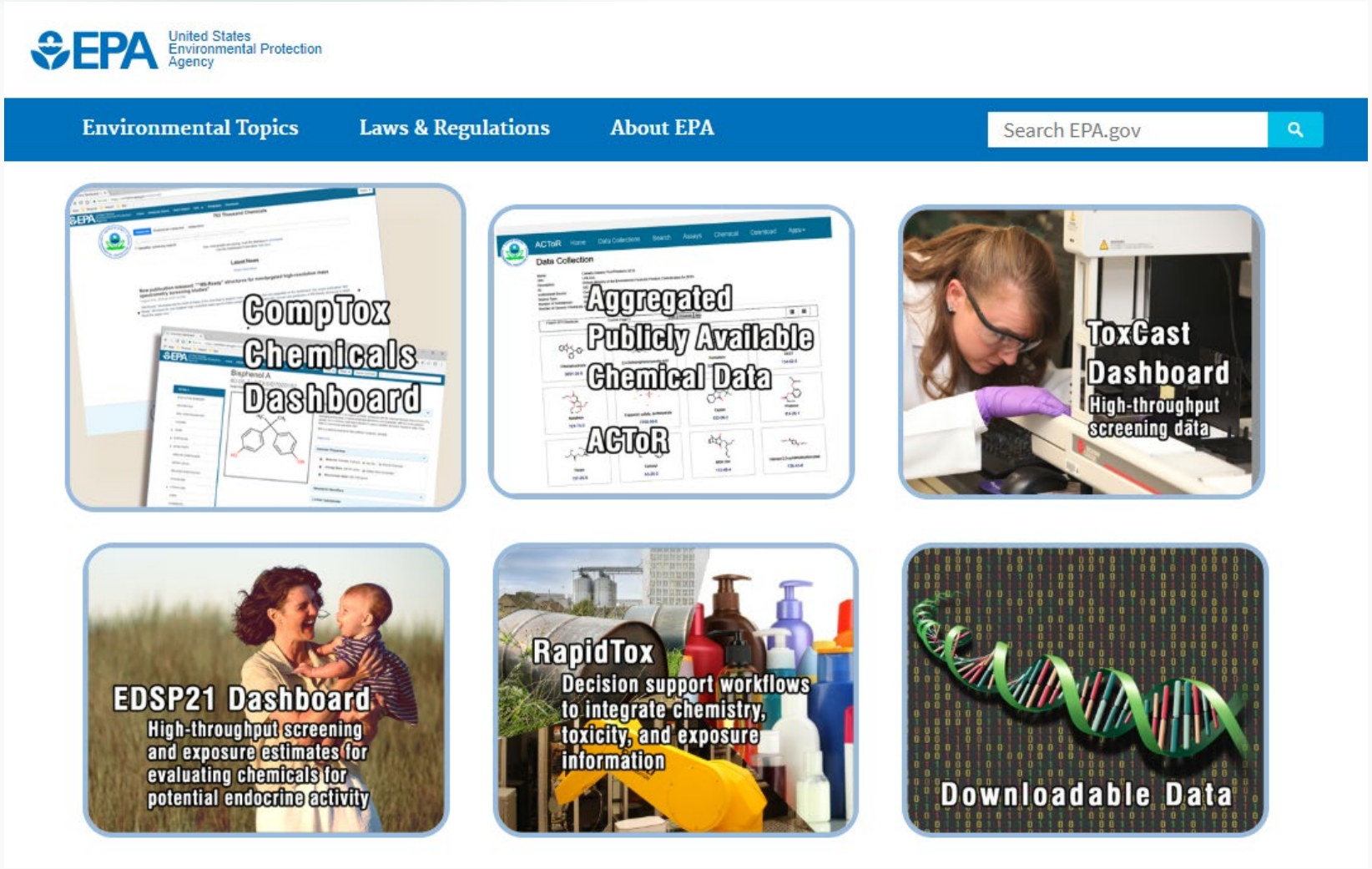
November 5th 2018

National Center for Computational Toxicology

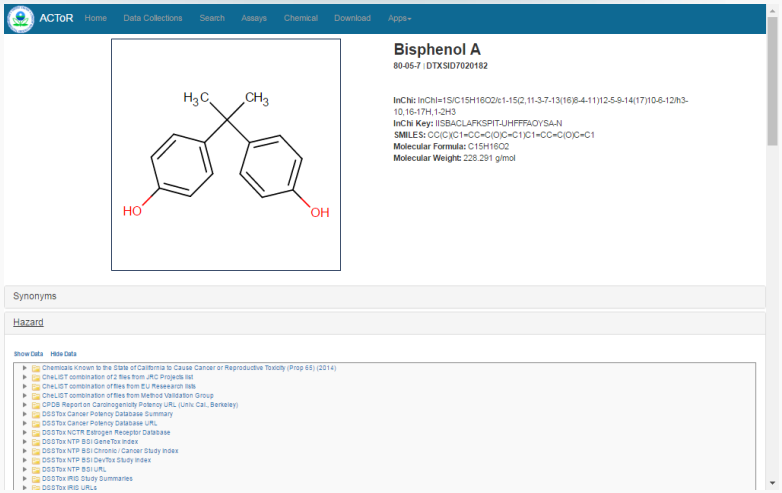


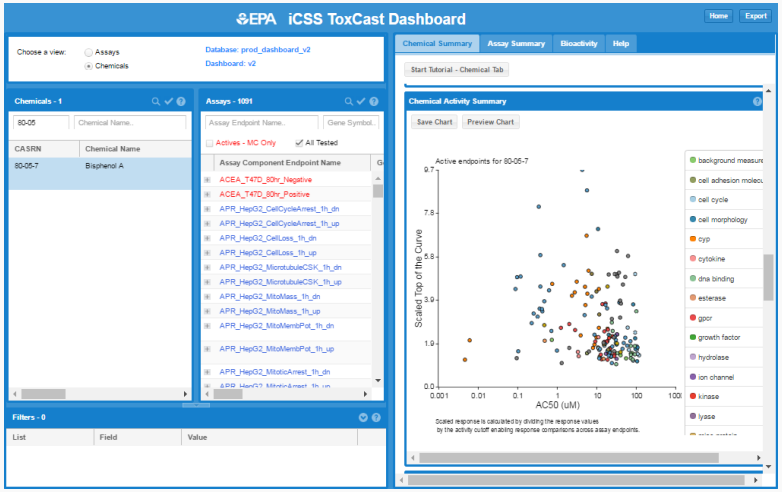
- National Center for Computational Toxicology established in 2005 to integrate:
  - High-throughput and high-content technologies
  - Modern molecular biology
  - Data mining and statistical modeling
  - Computational biology and chemistry
- Currently staffed by ~60 employees as part of EPA's Office of Research and Development
- Home of ToxCast & ExpoCast research efforts
- Key partner in U.S. Tox21 federal consortium
- Multiple cross-division collaborations (e.g. NERL, OPP, OPPT)

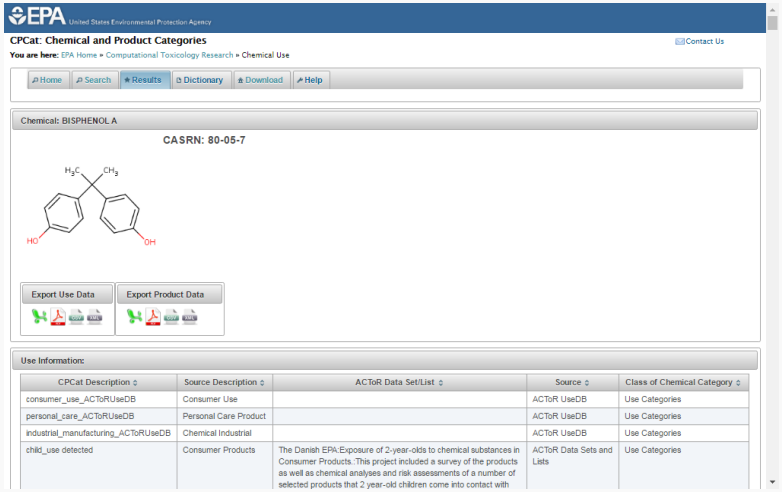
The CompTox Portal
https://comptox.epa.gov/

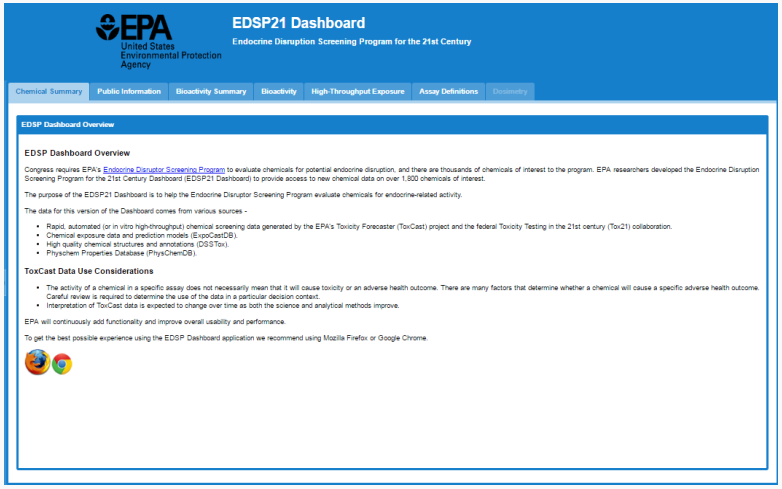


Earlier Dashboard Applications: Single architecture in development

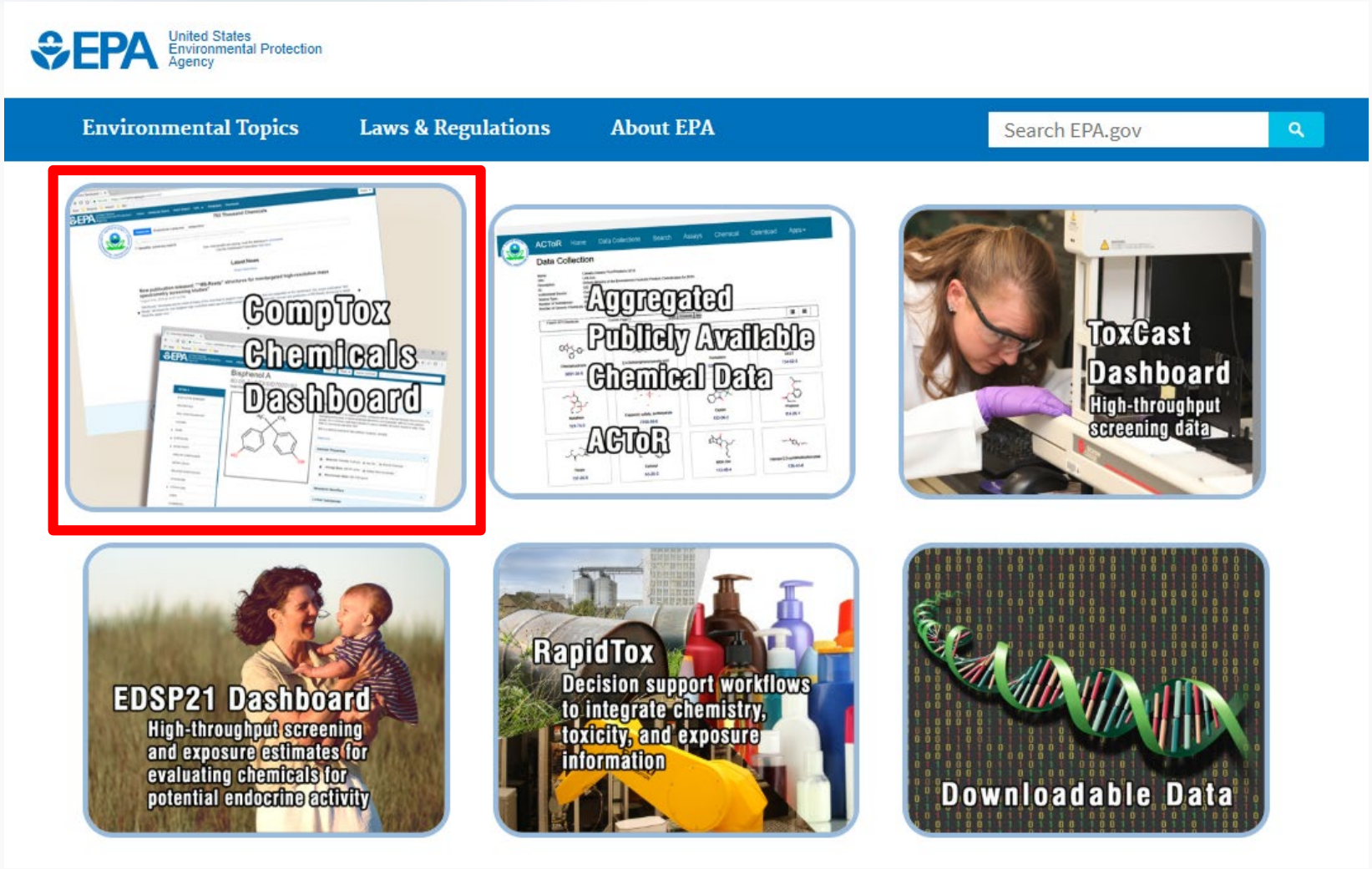








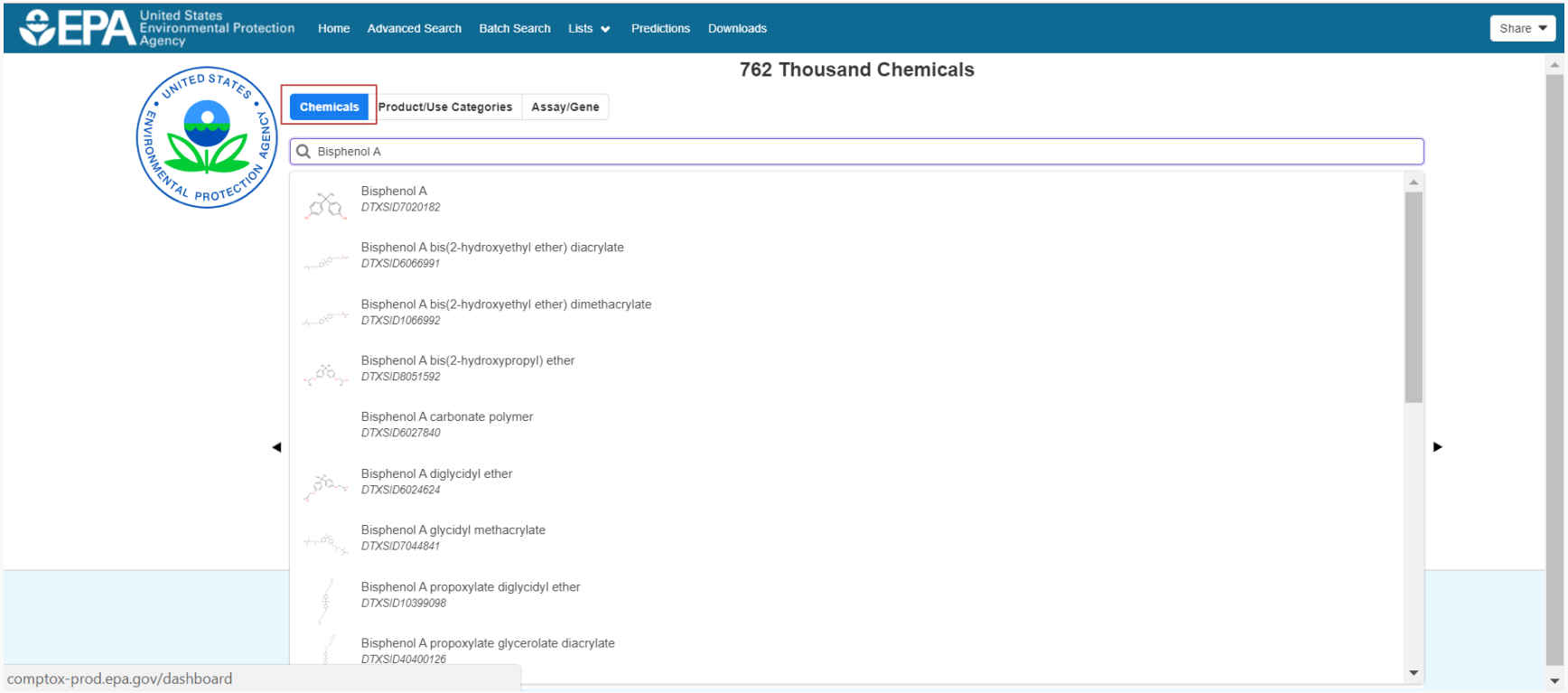
The CompTox Portal
https://comptox.epa.gov/



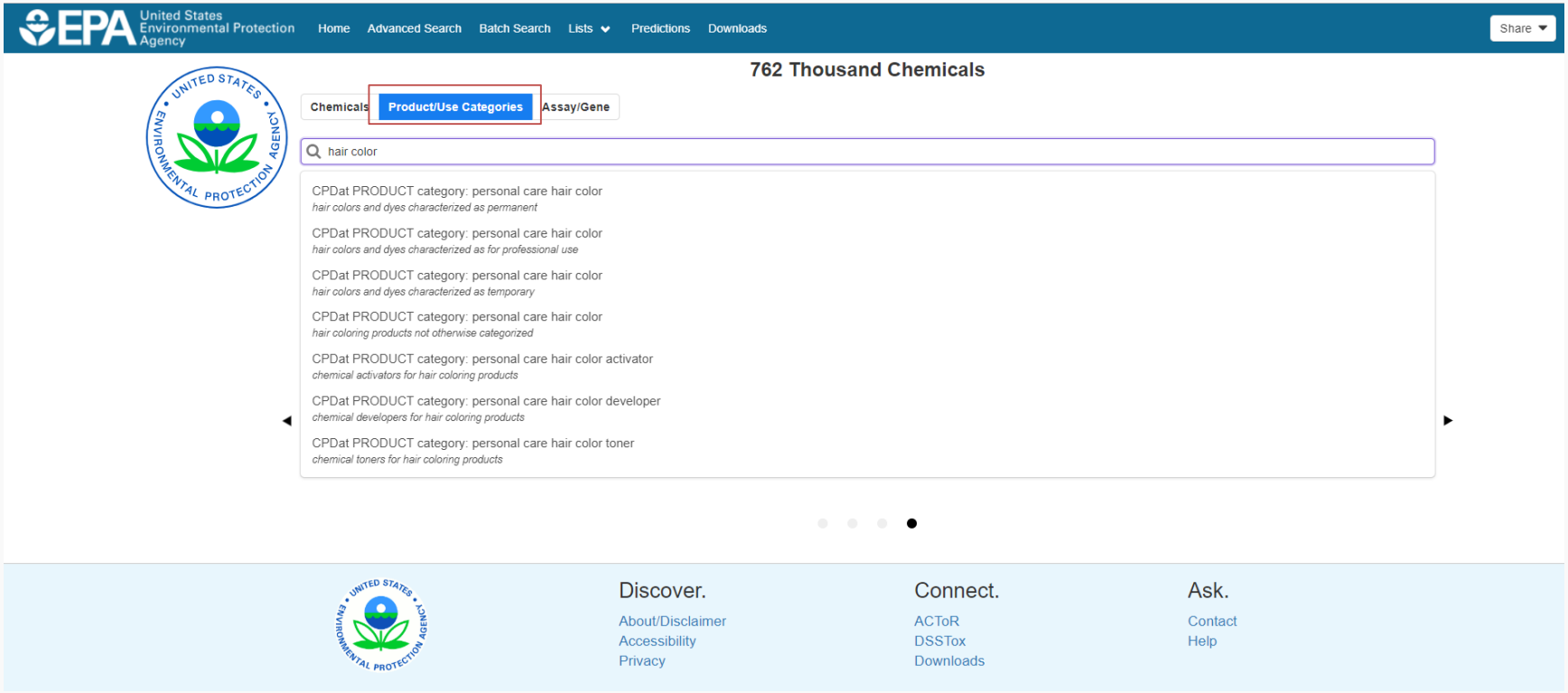
The CompTox Chemicals Dashboard

- A **publicly accessible website** delivering access:
  - ~765,000 chemicals with related property data
  - Experimental and predicted physicochemical property 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**

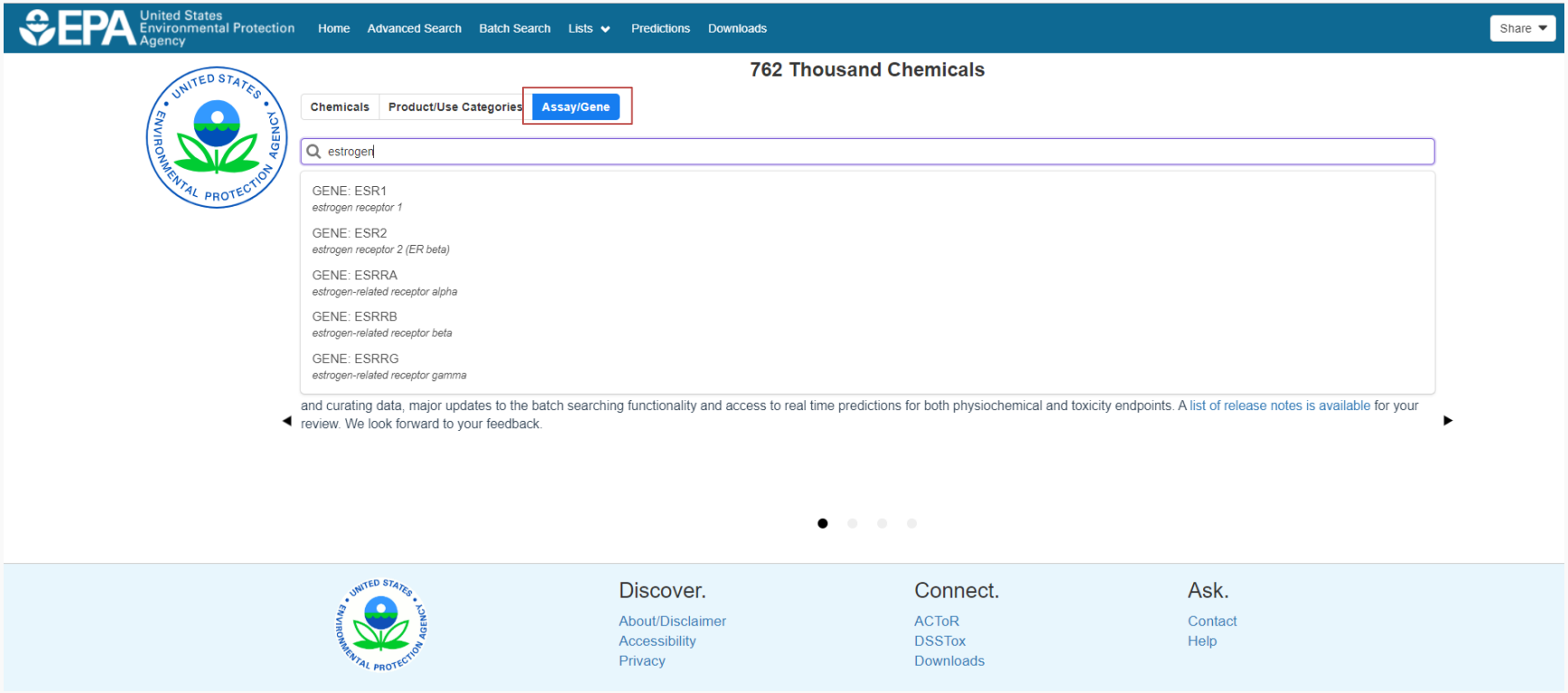
CompTox Chemicals Dashboard
Chemicals



CompTox Chemicals Dashboard
Products and Use Categories

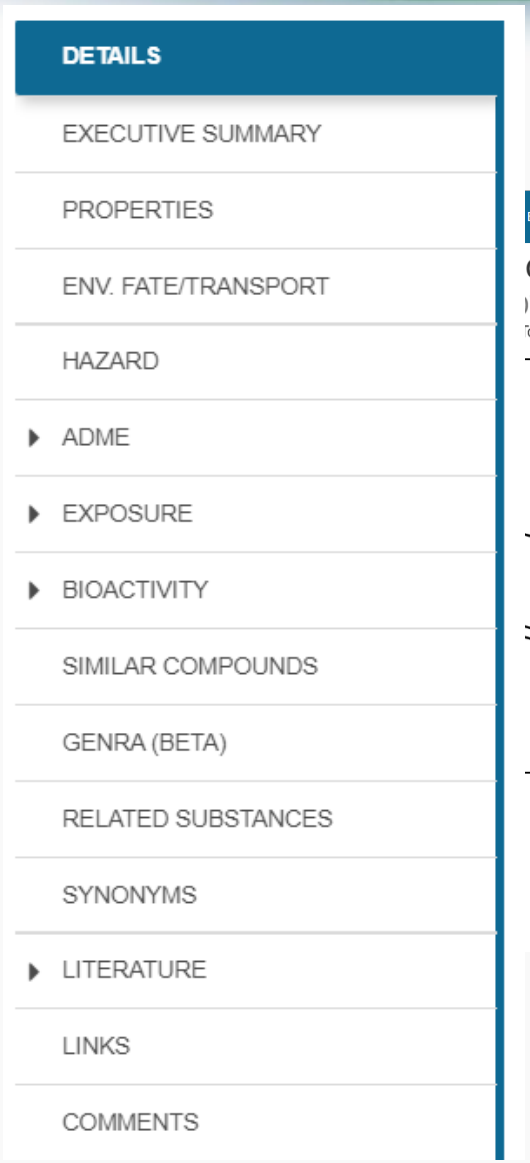


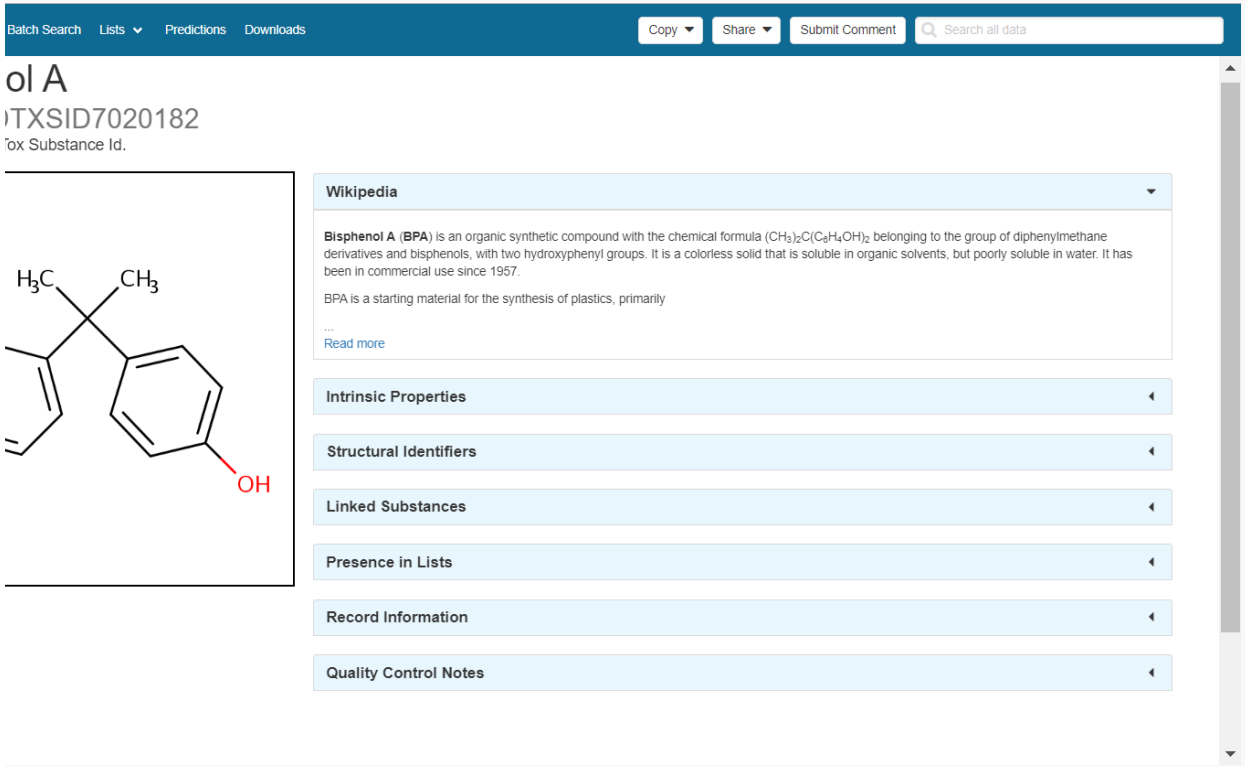
CompTox Chemicals Dashboard
Assays and Genes



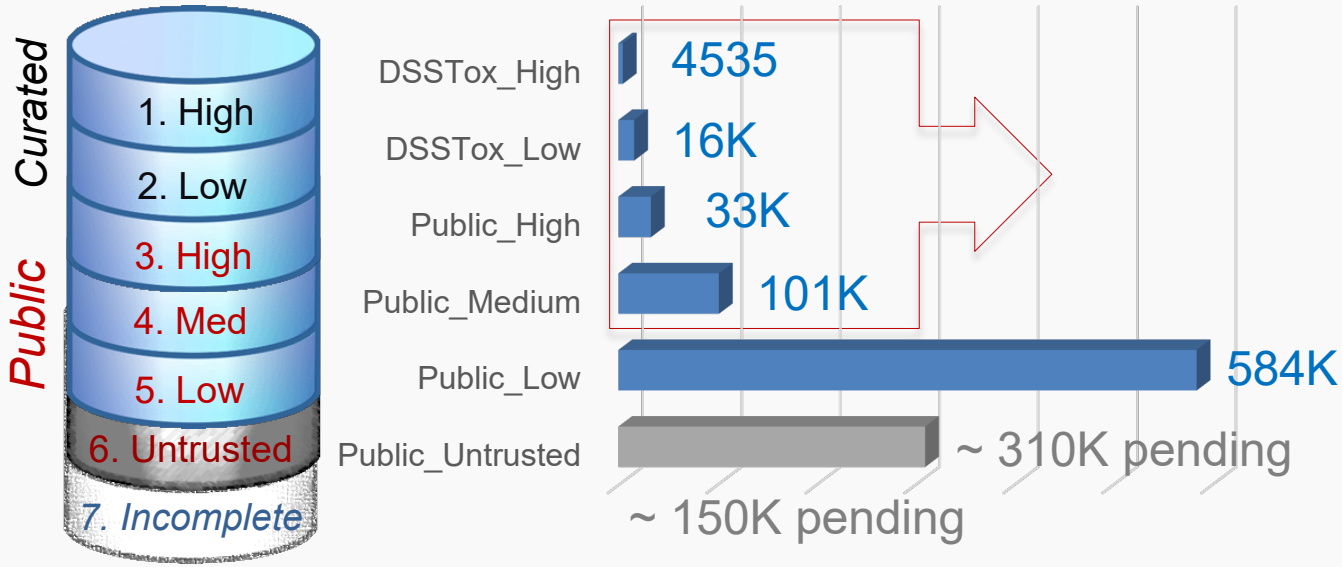
Detailed Chemical Pages

EPA United States Environmental Protection Agency





Almost 20 Years of Data…
Growing with daily curation



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

Physicochemical and Environmental Fate/Transport Properties

- Solubility
- Melting Point
- Boiling Point
- LogP (Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOA (Octanol/Air Partition Coefficient)
- LogKOC (Soil Adsorption Coefficient)
- Vapor Pressure

- And more…

Full transparency for predictions

EPA United States Environmental Protection Agency

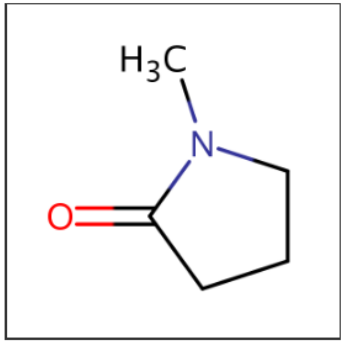
Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | Share | Search all data

OPERA Models: LogP: Octanol-Water

N-Methyl-2-pyrrolidone

872-50-4 | DTXSID6020856

Save PDF



Model Results

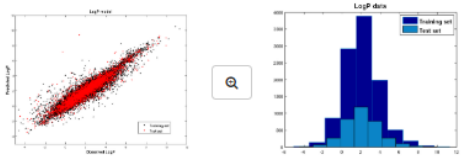
Predicted value: -0.300

Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.81

Model Performance

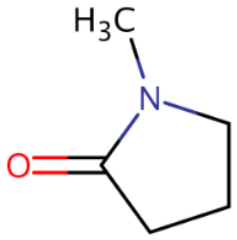


Weighted KNN model

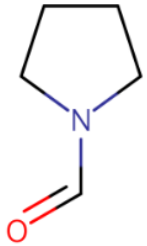
QMRF

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

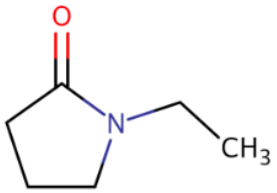
Nearest Neighbors from the Training Set



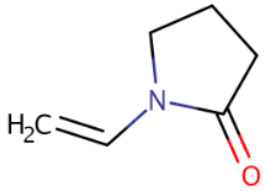
N-Methyl-2-pyrrolidone
Measured: -0.380
Predicted: -0.300



Prolinal
Measured: -0.320
Predicted: -0.226

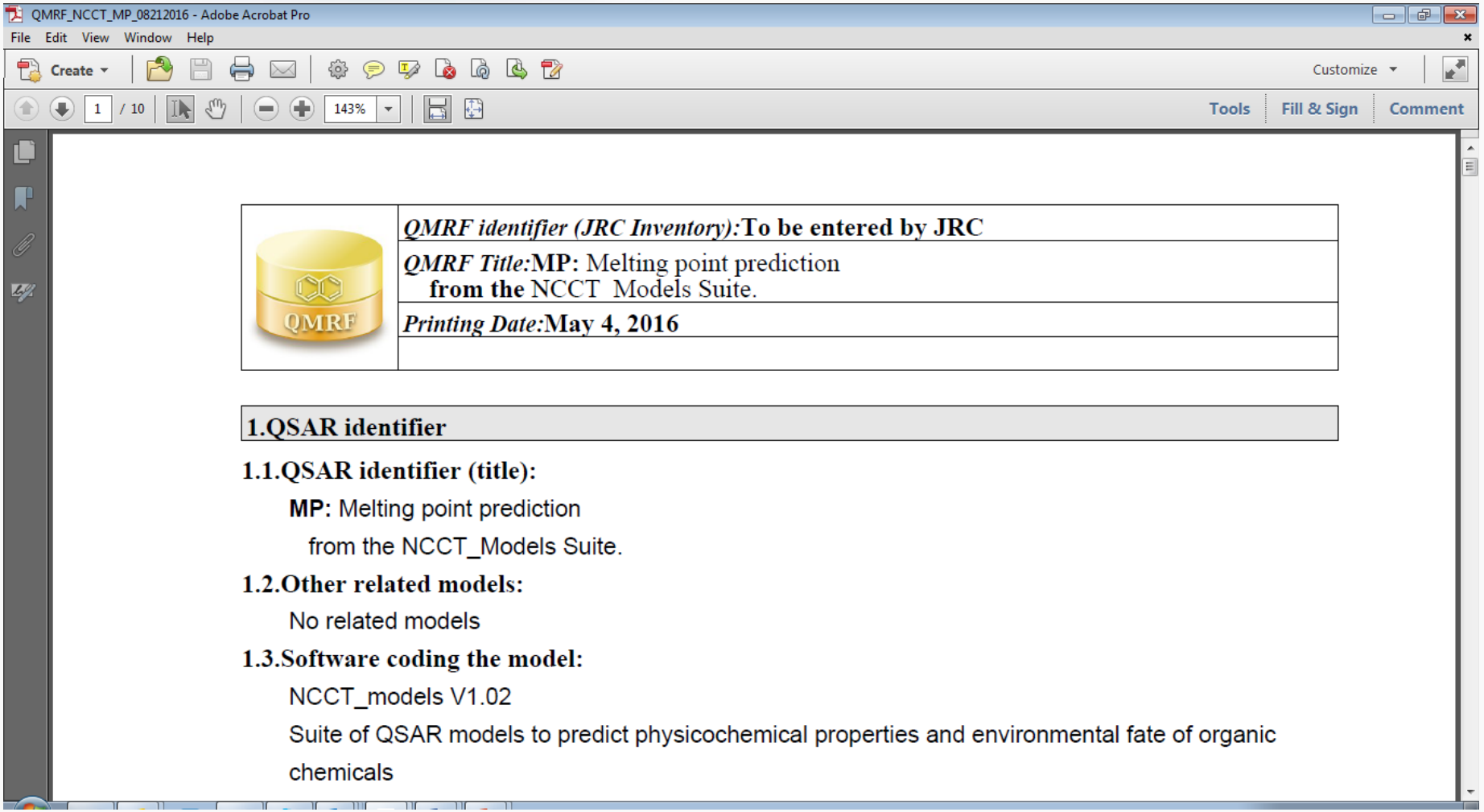


1-Ethyl-2-pyrrolidinone
Measured: -0.0400
Predicted: -0.211



N-Vinyl-2-pyrrolidone
Measured: 0.37
Predicted: 2.65e-01

Transparency: QMRF Report



Developing “NCCT Models”

- Our approach to modeling:
  - Obtain **high quality** training sets
  - Apply appropriate modeling approaches
  - **Validate** performance of models
  - Define the applicability domain and model limitations
  - Use models to predict properties across our full datasets
  - Release as **Open Data and Open Models**

Available online

EPI Suite Data - ISIS/Base & SDF

The downloaded files are provided in "zip" format ... the downloaded file must be "un-zipped" with common utility programs such as WinZip.

... *Updated September 15, 2010*

Basic Instructions:

(1) Download the zip file
(2) Un-Zip the file

NOTE ... zipped files extract to **Folders** containing the individual data files ... Folders named **EPI_ISIS_Data** and **EPI_SDF_Data**

Substructure Searching Files:

ISIS™/Base & SD Files of the EPI Suite Program Experimental Data Files are now available ... The ISIS™/Base files require the commercial program for use ... The SD Files can be imported into other commercial chemical structure programs (such as ChemFinder).

... **Click here to download EPI_ISIS_Data.zip** ... (about 11 MB)

... **Click here to download EPI_SDF_Data.zip** ... (about 10 MB)

NOTE ... **EPI Suite Data Files (some in Excel, Text, Word format) available at:**

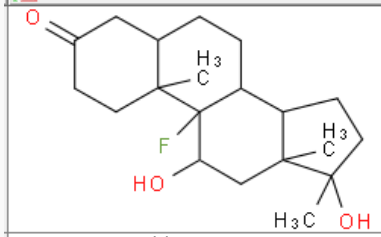
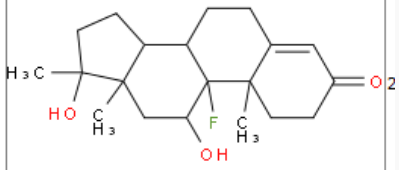
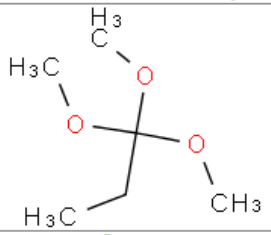
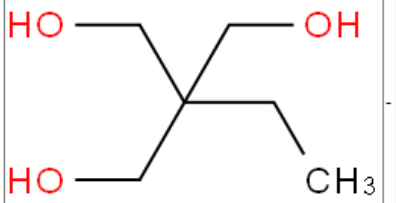
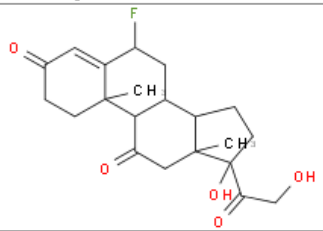
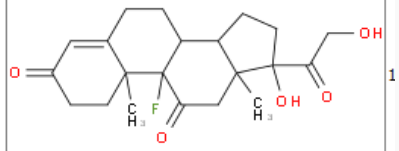
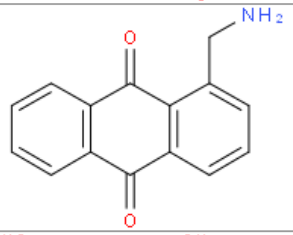
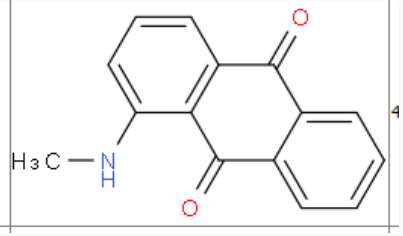
http://esc.syrres.com/interkow/EpiSuiteData.htm

We Curated These Public Data to Build Prediction Models

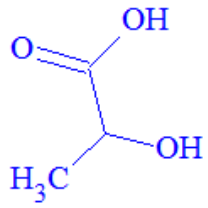
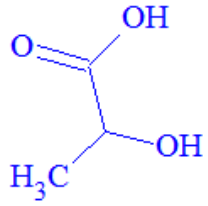
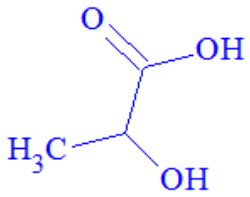
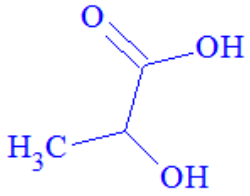
EPA United States Environmental Protection Agency

Public data should be curated prior to modeling

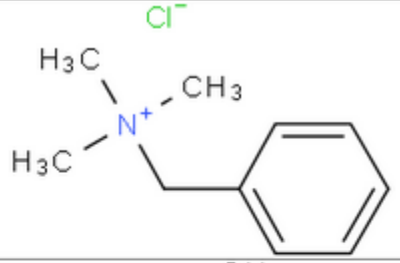
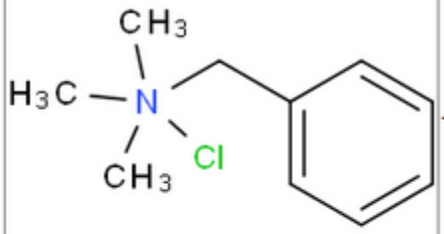
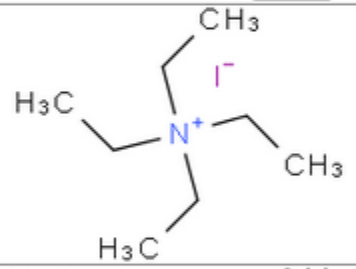
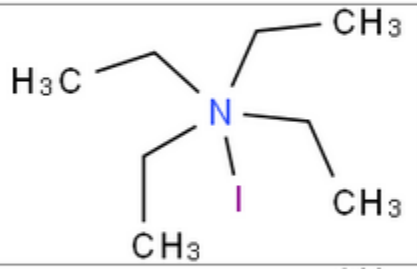
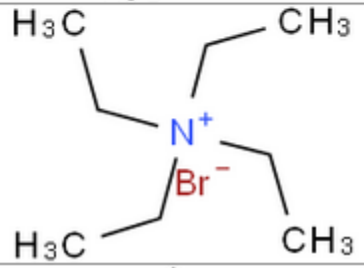
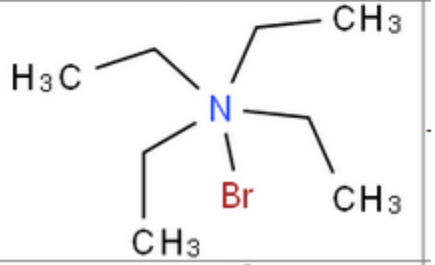
Different Compounds

Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Duplicate Structures

Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_3H_6O_3$	90.0779	000050-21-5	LACTIC ACID	1.68000000000000e+001	2.26600000000000e+001	5.86000000000000e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.30000000000000e+001	2.26600000000000e+001	-3.03400000000000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.80000000000000e+001	2.26600000000000e+001	4.66000000000000e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.28000000000000e+001	2.26600000000000e+001	-3.01400000000000e+001

Covalent Halogens

Mol Block	CAS	NAME	Smiles
	000056-93-9	BENZYL TRIMETHYL AMMONIUM CHLORIDE	
	000068-05-3	TETRAETHYL AMMONIUM IODIDE	
	000071-91-0	TETRAETHYL AMMONIUM BROMIDE	

Curation to QSAR Ready Files

Property	Initial file	Curated Data	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	~~14544~~	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	2348	2046	2010

LogP dataset:15,809 structures

- CAS Checksum: 12163 valid, 3646 invalid **(>23%)**
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES **(>24%)**
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry **(~8%)**
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds **(~3%)**

OPERA Predicted Properties

EPA United States Environmental Protection Agency

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling

K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

Mansouri et al. J Cheminform (2018) 10:10
https://doi.org/10.1186/s13321-018-0263-1

Journal of Cheminformatics

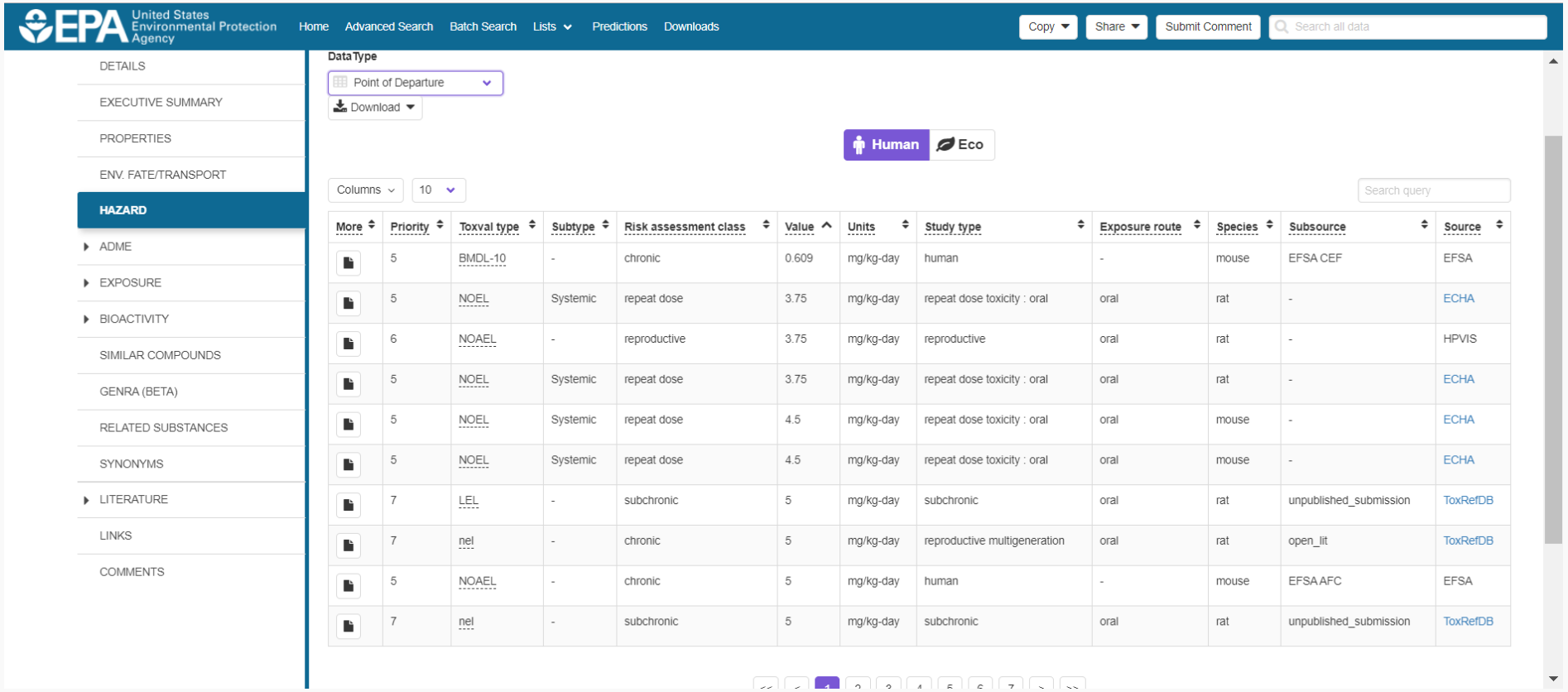
RESEARCH ARTICLE Open Access

OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri[1,2,3*], Chris M. Grulke[1], Richard S. Judson[1] and Antony J. Williams[1]

OPERA Models: https://github.com/kmansouri/OPERA

Other Data: Human and Ecological 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

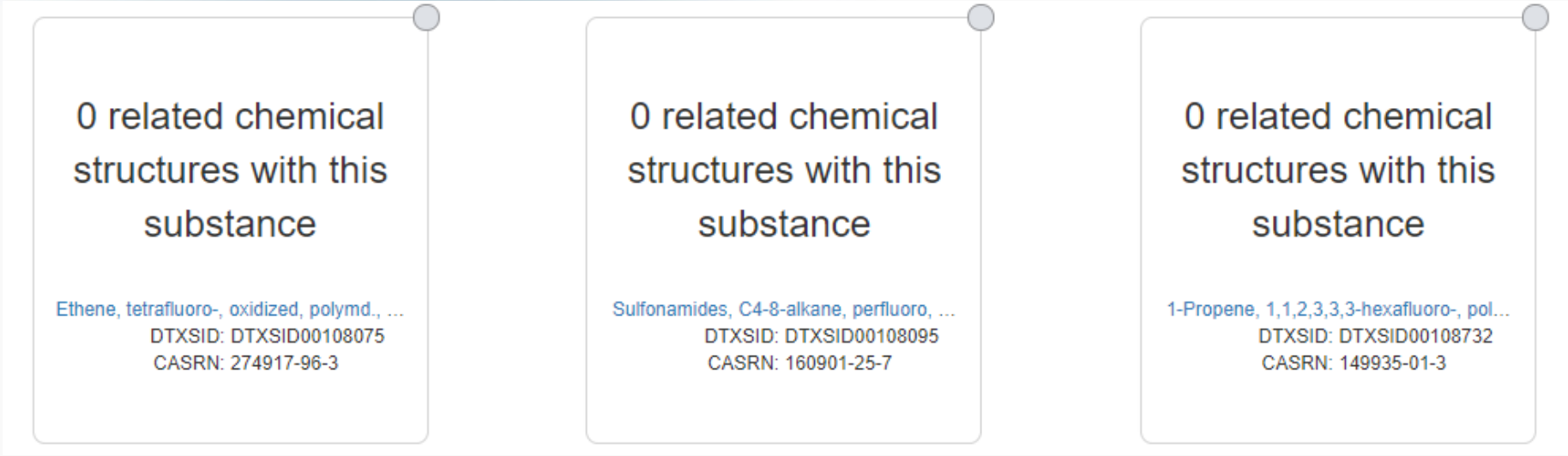
- If only the data were easy to extract…

Not just chemical “structures”

- Chemicals in commerce, of interest to the EPA, are not all easily represented by structures

- Different chemical substances supported
  - “UVCB chemicals” - Unknown or Variable Composition, Complex Reaction Products and Biological Materials
  - Homologous series as Markush Structures

Example PFAS-UVCBs



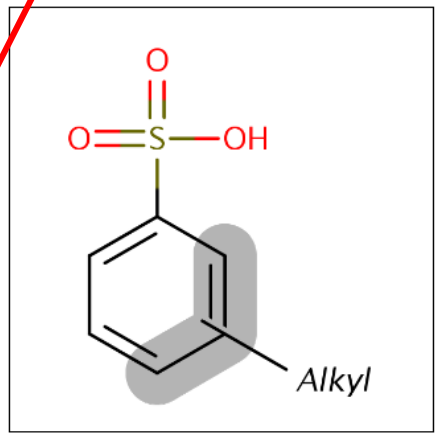
Ethene, tetrafluoro-, oxidized, polymd., reduced, decarboxylated, C6 fraction
274917-96-3 | DTXSID00108075

1-Propene, 1,1,2,3,3,3-hexafluoro-, polymer with 1,1-difluoroethene, ethene, 1,1,2,2-tetrafluoroethene and 1,1,2-trifluoro-2-(trifluoromethoxy)ethene
149935-01-3 | DTXSID00108732

UVCB Chemicals

(C10-C16) Alkylbenzenesulfonic acid
68584-22-5 | DTXSID2028723
Searched by DSSTox Substance Id.

DETAILS
RELATED SUBSTANCES
PROPERTIES
COMMENTS



Intrinsic Properties

Presence in Lists

Federal
Safer Choice Chemical List
TOX21SL: Tox21 Screening Library
TSCA Workplan Step 2 Chemicals
EPAHFR - EPA Chemicals associated with hydraulic fracturing

US State
None.

International
OLEM RapidTox Chemicals
REACH Dossier Chemicals

Other
Surfactant List Screened in Swiss Wastewater (2014)
EDSP Universe
TSCAACTIVE
EPA Chemicals associated with hydraulic fracturing

Markush Structures

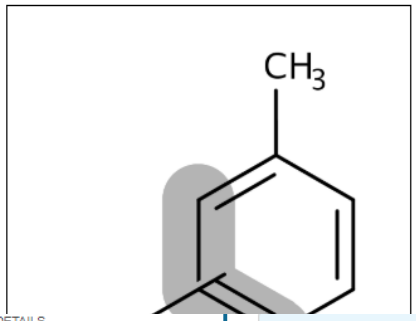
EPA United States Environmental Protection Agency

Xylenes

1330-20-7 | DTXSID2021446

Searched by DSSTox Substance Id.

DETAILS
RELATED SUBSTANCES
PROPERTIES
COMMENTS



Wikipedia

Xylene (from Greek ξύλο, *xylo*, "wood"), **xylol** or **dimethylbenzene** is any one of three isomers of dimethylbenzene, or a combination thereof. With the formula $(CH_3)_2C_6H_4$, each of the three compounds has a central benzene ring with two methyl groups attached at substituents. They are all colorless, flammable liquids, some of which are of great industrial value. The mixture is referred to as both xylene and, more precisely, xylenes

...
Read more

Intrinsic Properties

Presence in Lists

Record Information

DETAILS
RELATED SUBSTANCES
PROPERTIES
COMMENTS

8 chemicals

Download / Send

Show info: DTXSID × CASRN ×

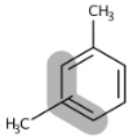
Select all

Sort by: Relationship

Filter by: Name or CASRN

Hide

Searched Chemical



Xylenes
DTXSID: DTXSID2021446
CASRN: 1330-20-7

Predecessor: Component

4 related chemical structures with this substance

Xylenes; defined mixture 1
DTXSID: DTXSID9021421
CASRN: NOCAS_21421

Predecessor: Component

4 related chemical structures with this substance

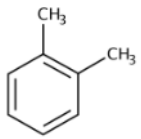
Xylenes; defined mixture 2
DTXSID: DTXSID7021447
CASRN: NOCAS_21447

Predecessor: Component

5 related chemical structures with this substance

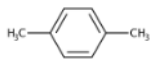
Total Petroleum Hydrocarbons (TPH)
DTXSID: DTXSID30891529
CASRN: NOCAS_891529

Markush Child



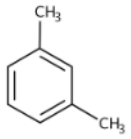
o-Xylene
DTXSID: DTXSID3021807
CASRN: 95-47-6

Markush Child



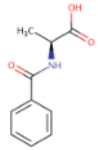
p-Xylene
DTXSID: DTXSID2021868
CASRN: 106-42-3

Markush Child



m-Xylene
DTXSID: DTXSID6026298
CASRN: 108-38-3

Transformation Product



N-Benzoylalanine
DTXSID: DTXSID40176394
CASRN: 2198-64-3

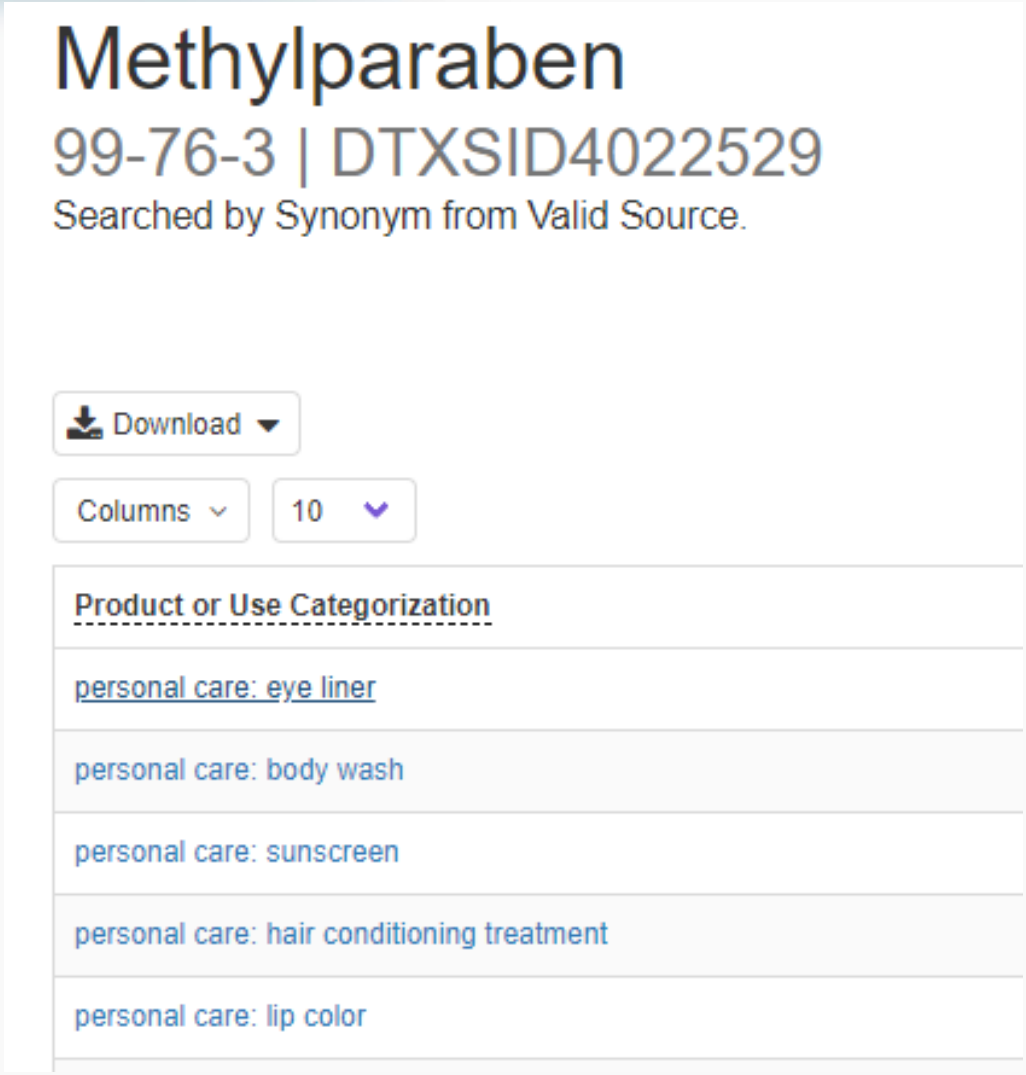
Environmental Chemistry: More about Names and CASRNs

- For EPA most chemicals are reported as text – chemical names and CAS Numbers

Attachment D (Method 3)
SIM quantitation ions and qualifiers for internal standards, references method analysis, and surrogates

Name of Compound	CAS No.	Quantitation Ion	Qualifier Ions
Phenol-d6 (SS)	13187-88-3	99	71, 42
Phenol	108-95-2	94	66
1,4-Dichlorobenzene	106-46-0	146	111. 75, 50
Acetophenone	98-86-2	105	77, 51, 120
Acenaphthene-d10 (IS)	15067-26-2	162	160, 80
p-Cresol	106-44-5	107	108, 77
Isophorone	78-59-1	82	138, 54
Camphor	76-22-2	95	81, 108, 152
Isoborneol	124-76-5	95	110, 121, 136
Menthol	89, 78, 1	71	81, 123, 138
Naphthalene	91-20-3	128	102, 51
Methyl salicilate	119-36-8	120	92, 152, 65

Chemicals in product SDS sheets are commonly UVCBs



182 chemicals in Personal Care: Eye Liner Category

EPA United States Environmental Protection Agency

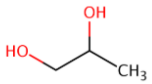
Searched by Product & Use Categories

Results for CPDat Product Category: Personal Care: Eye Liner

182 chemicals

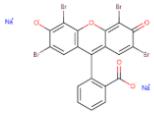
Download / Send | Show info: DTXSID × CASRN × TOXCAST × | Select all

Sort by: DTXSID | Filter by: Name or CASRN | Hide



1,2-Propylene glycol
DTXSID: DTXSID0021206
CASRN: 57-55-6
TOXCAST: 11/539

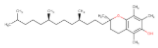
cis-Oleyl alcohol
DTXSID: DTXSID0022010
CASRN: 143-28-2
TOXCAST: 0



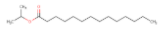
Eosin
DTXSID: DTXSID0025234
CASRN: 17372-87-1
TOXCAST: 45/302

1 related chemical structure with this substance

Polyvinylpyrrolidone
DTXSID: DTXSID0025941
CASRN: 9003-39-8
TOXCAST: 0



alpha-Vitamin E
DTXSID: DTXSID0026339
CASRN: 59-02-9
TOXCAST: 0



Isopropyl tetradecanoate
DTXSID: DTXSID0026838
CASRN: 110-27-0
TOXCAST: 1/299

0 related chemical structures with this substance

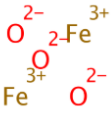
Paraffin waxes and Hydrocarbon waxes...
DTXSID: DTXSID0028115
CASRN: 63231-60-7
TOXCAST: 0

1 related chemical structure with this substance

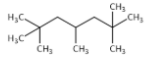
Alcohols, C16-18
DTXSID: DTXSID0028323
CASRN: 67762-27-0
TOXCAST: 1/163

0 related chemical structures with this substance

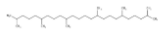
Fats and Glyceridic oils, vegetable, hyd...
DTXSID: DTXSID0028454
CASRN: 68334-28-1
TOXCAST: 0



Iron(III) oxide
DTXSID: DTXSID0029632
CASRN: 1309-37-1
TOXCAST: 0



2,2,4,6,6-Pentamethylheptane
DTXSID: DTXSID0042034
CASRN: 13475-82-6
TOXCAST: 0



Squalane
DTXSID: DTXSID0046513
CASRN: 111-01-3
TOXCAST: 2/109

CASRNs can be problematic…

Name of Compound	CAS No.
Phenol-d6 (SS)	13187-88-3
Phenol	108-95-2
1,4-Dichlorobenzene	106-46-0
Acetophenone	98-86-2
Acenaphthene-d10 (IS)	15067-26-2
p-Cresol	106-44-5
Isophorone	78-59-1
Camphor	76-22-2
Isoborneol	124-76-5
Menthol	89, 78, 1
Naphthalene	91-20-3
Methyl salicilate	119-36-8

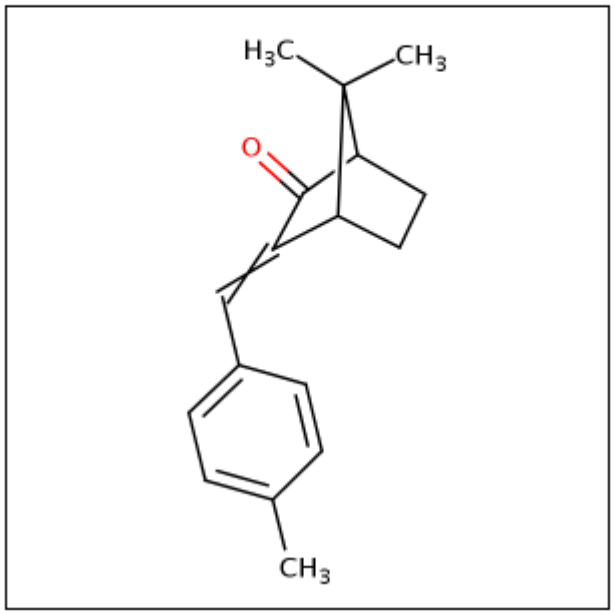
Active vs Deleted CASRN Also “Alternates”

EPA United States Environmental Protection Agency

Enzacamene

36861-47-9 | DTXSID8047896

Searched by Approved Name.



Synonym	Quality
Enzacamene	Valid
7,7-Dimethyl-3-[(4-methylphenyl)methylidene]bicyclo[2.2.1]heptan-2-one	Valid
Bicyclo[2.2.1]heptan-2-one, 7,7-dimethyl-3-[(4-methylphenyl)methylene]-	Valid
36861-47-9 Active CASRN	Valid
Bicyclo[2.2.1]heptan-2-one, 1,7,7-trimethyl-3-[(4-methylphenyl)methylene]-	Valid
EINECS 253-242-6	Other
Eusolex 6300	Other
Uvinul MBC 95	Other
Parsol 5000	Other
UNII-8I3XWY40L9	Other
4-Methylbenzylidenecamphor	Other
p-Methylbenzylidenecamphor	Other
38102-62-4 Deleted CASRN	Deleted
84055-65-2 Deleted CASRN	Deleted

Tricky mapping by CASRN
This one has 316 Deleted CASRN

EPA United States Environmental Protection Agency

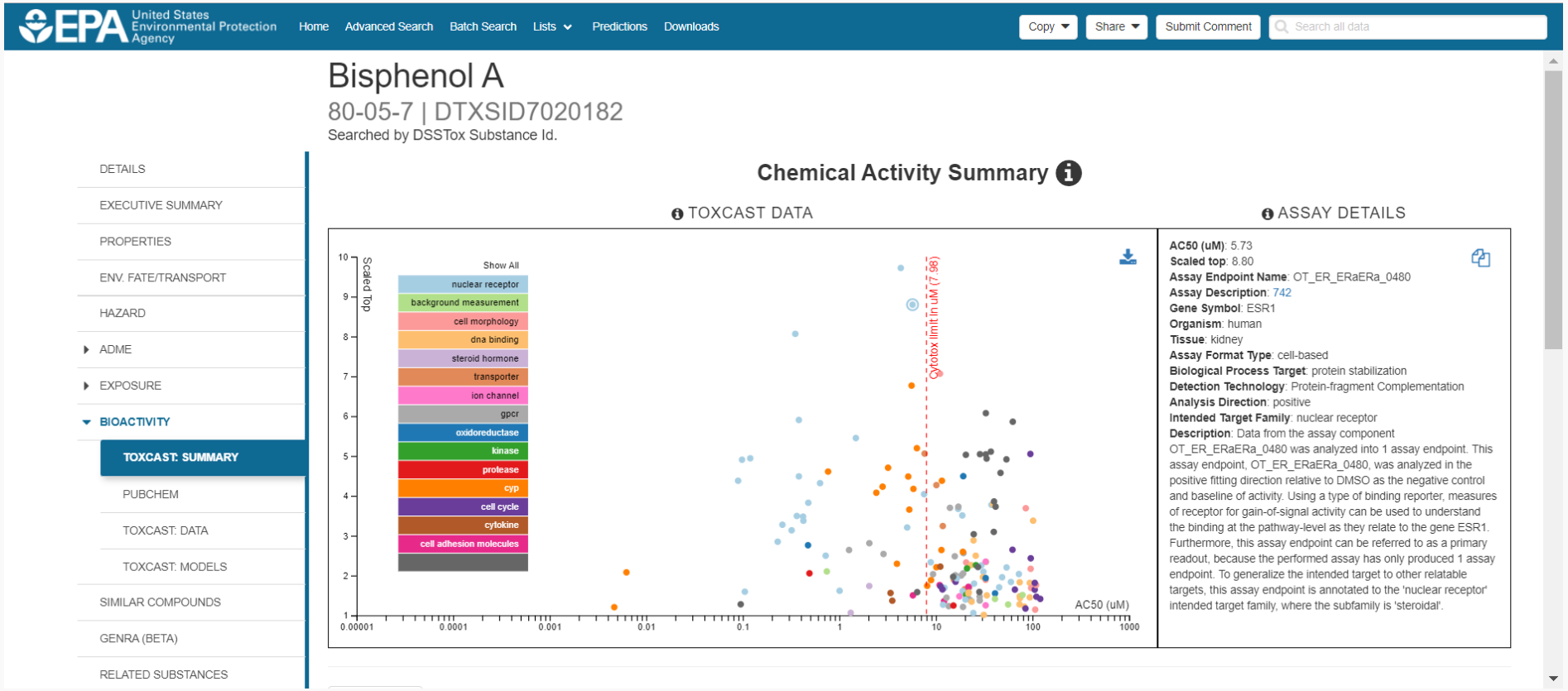
CAS Registry Number: 25068-38-6
$(C_{15}H_{16}O_2 . C_3H_5ClO)_x$
Phenol, 4,4'-(1-methylethylidene)bis-, polymer with 2-(chloromethyl)oxirane

Polymer
Polymer Class Terms: Epoxy resin
Alternate CAS Registry Numbers: 26402-79-9
Deleted CAS Registry Numbers: 1336-88-5, 1337-15-1, 8000-31-5, 9015-99-0, 9049-54-1, 9050-21-9, 9081-91-8, 9083-76-5, 9084-94-0, 9086-62-8, 9087-26-7, 9087-76-7, 11097-80-6, 11098-13-8, 11098-40-1, 11100-23-5, 11108-41-1, 11120-31-3, 11121-19-0, 11126-36-6, 20232-24-0, 35038-60-9, 36344-96-4, 36347-13-4, 36347-14-5, 37184-50-2, 37184-52-4, 37208-29-0, 37217-92-8, 37230-74-3, 37243-66-6, 37243-67-7, 37251-33-5, 37265-21-7, 37270-82-9, 37291-75-1, 37293-07-5, 37294-18-1, 37305-82-1, 37307-45-2, 37317-45-6, 37325-21-6, 37338-63-9, 37342-17-9, 37345-34-9, 37348-56-4, 37348-57-5, 37357-73-6, 37360-93-3, 39277-59-3, 39288-99-8, 39296-08-7, 39296-09-8, 39296-11-2, 39296-15-6, 39315-77-0, 39349-91-2, 39354-86-4, 39362-25-9, 39362-45-3, 39373-81-4, 39378-29-5, 39378-55-7, 39389-49-6, 39405-18-0, 39412-57-2, 39419-66-4, 39453-22-0, 39454-54-1, 39454-69-8, 39470-62-7, 42612-34-0, 42618-03-1, 50642-36-9, 50642-55-2, 50642-78-9, 51158-20-4, 51273-81-5, 51329-73-8, 51393-99-8, 51394-03-7, 51553-00-5, 52011-87-7, 52038-45-6, 52051-70-4, 52051-82-8, 52052-16-1, 52232-05-0, 52232-75-4, 52276-55-8, 52365-33-0, 52519-66-1, 52519-67-2, 52627-94-8, 52907-38-7, 53027-88-6, 53127-14-3, 53200-30-9, 53238-86-1, 53238-87-2, 53239-67-1, 53239-68-2, 53570-97-1, 53570-98-2, 53681-78-0, 53858-93-8, 54018-73-4, 54352-05-5, 55464-96-5, 55584-55-9, 55585-07-4, 55818-73-0, 56258-35-6, 56449-43-5, 56509-48-9, 57107-66-1, 57284-90-9, 57534-21-1, 57693-04-6, 58052-05-4, 58128-38-4, 58392-89-5, 58392-92-0, 58516-14-6, 58572-71-7, 59029-19-5, 59459-14-2, 59473-30-2, 59948-36-6, 60202-19-9, 60267-31-4, 60382-89-0, 60606-56-6, 60800-54-6, 60831-77-8, 60894-16-8, 61036-82-6, 61287-42-1, 61356-27-2, 61711-38-4, 61763-30-2, 61991-18-2, 62169-28-2, 62169-29-3, 62601-75-6, 62601-76-7, 62887-23-4, 63055-40-3, 63172-55-4, 63799-24-6, 63993-57-7, 63993-58-8, 64086-14-2, 64086-16-4, 64176-52-9, 64176-61-0, 64176-66-5, 64177-03-3, 65233-49-0, 65931-38-6, 65931-39-7, 66995-96-8, 67185-62-0, 68821-97-6, 69899-40-7, 70179-83-8, 70213-44-4, 70726-45-3, 71965-91-8, 72514-40-0, 73413-19-1, 74504-20-4, 74564-76-4, 75831-44-6, 78564-77-9, 79585-43-6, 80702-61-0, 81458-12-0, 81843-57-4, 81843-58-5, 81855-87-0, 82197-12-4, 82197-46-4, 83202-85-1, 84286-97-5, 84683-04-5, 84931-29-3, 85537-69-5, 86090-60-0, 88385-37-9, 88528-19-2, 88651-18-7, 89750-00-5, 91727-28-5, 91727-29-6, 92481-37-3, 95327-25-6, 96420-31-4, 96510-68-8, 97568-16-6, 97709-01-8, 99400-50-7, 101027-12-7, 102256-87-1, 103599-13-9, 103599-14-0, 104364-97-8, 104491-99-8, 105521-57-1, 106207-08-3, 106856-89-7, 107991-47-9, 108556-05-4, 108728-21-8, 110158-22-0, 111367-08-9, 111517-59-0, 114013-37-5, 115902-32-4, 117216-90-7, 117313-45-8, 117786-92-2, 118340-04-8, 120146-74-9, 120797-43-5, 121181-85-9, 121273-37-8, 121547-73-7, 123939-44-6, 125147-87-7, 127176-80-1, 127176-81-2, 128281-71-0, 132822-20-9, 132893-73-3, 135976-90-8, 137545-29-0, 138157-20-7, 138361-18-9, 139554-29-3, 142540-11-2, 144046-24-2, 144046-25-3, 144855-66-3, 149013-58-1, 150825-32-4, 157321-42-1, 157481-46-4, 158725-45-2, 160674-45-3, 161937-12-8, 162031-55-2, 167972-06-7, 168042-08-8, 179607-24-0, 183581-68-2, 183890-12-2, 187619-11-0, 188448-56-8, 189282-49-3, 191606-83-4, 220090-06-2, 222835-65-6, 222835-66-7, 222835-68-9, 222835-69-0, 222835-70-3, 222835-72-5, 222835-74-7, 222835-77-0, 309945-96-8, 339530-81-3, 353239-57-3, 367523-08-8, 383889-26-7, 383889-27-8, 395069-05-3, 470462-49-8, 681001-41-2, 848887-61-6, 913745-83-2, 917483-69-3, 922728-11-8, 934588-09-7, 945610-97-9, 950907-45-6, 1033821-54-3, 1034342-45-4, 1068160-75-7, 1082736-74-0, 1096473-97-0, 1114797-08-8, 1189565-70-5, 1190235-62-1, 1190729-68-0, 1192045-32-1, 1195324-26-5, 1196030-95-1, 1198291-96-1, 1199811-18-1, 1203835-26-0, 1206700-05-1, 1228639-00-6, 1245563-83-0, 1271727-39-9, 1300093-58-6, 1300102-07-1, 1305321-17-8, 1338071-08-1, 1446691-72-0, 1450839-98-1, 1620807-39-7, 1641551-32-7, 1807886-28-7, 1815624-46-4, 1815624-47-5

In Vitro Bioassay Screening

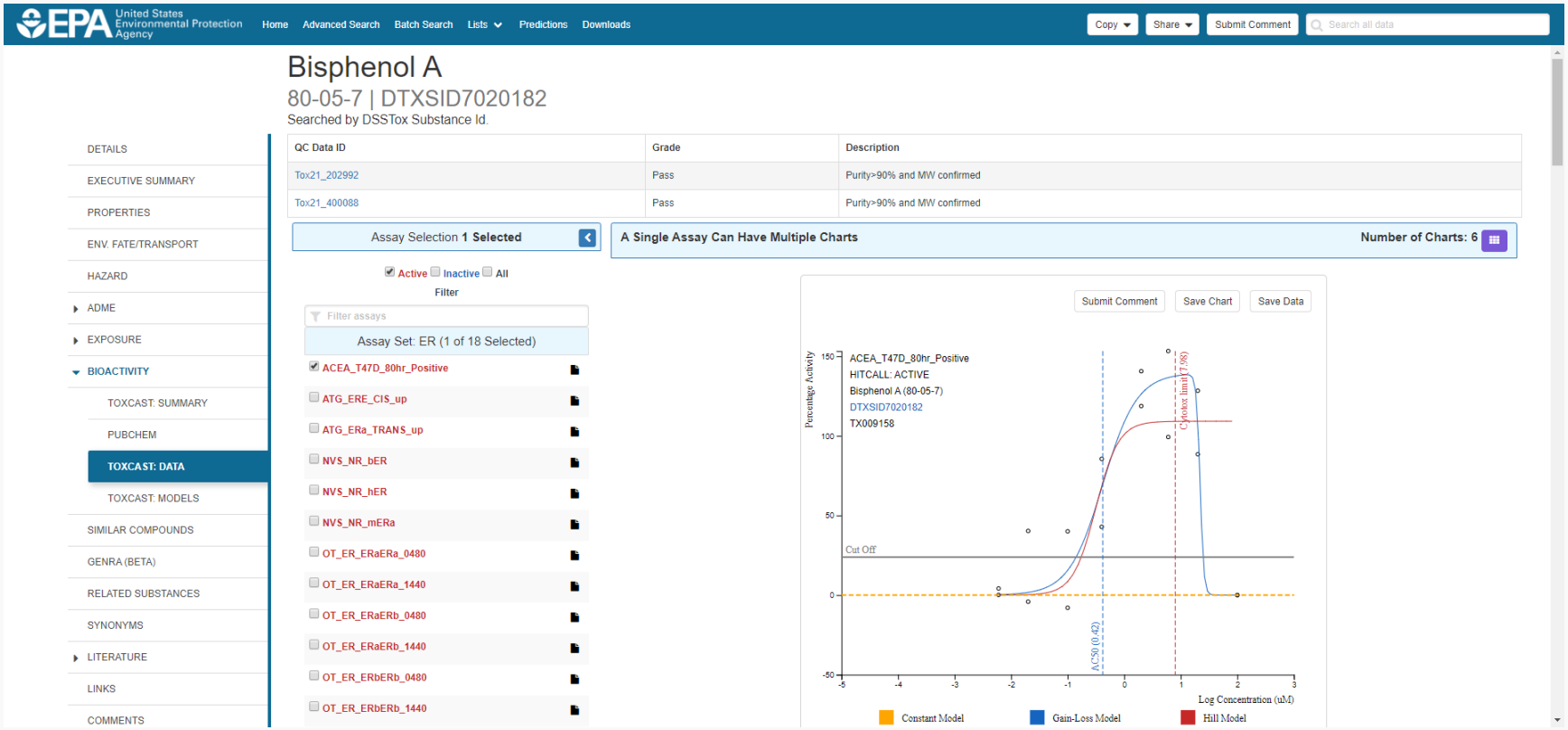
ToxCast and Tox21

EPA United States Environmental Protection Agency



In Vitro Bioassay Screening

ToxCast and Tox21



Bioactivity: Downloadable Data

https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data

EPA United States Environmental Protection Agency

Exploring ToxCast Data: Downloadable Data

The results after processing through the Pipeline are available on the ToxCast Dashboard, and for most users EPA recommends accessing the data there.

- ToxCast Chemicals
- ToxCast Assays

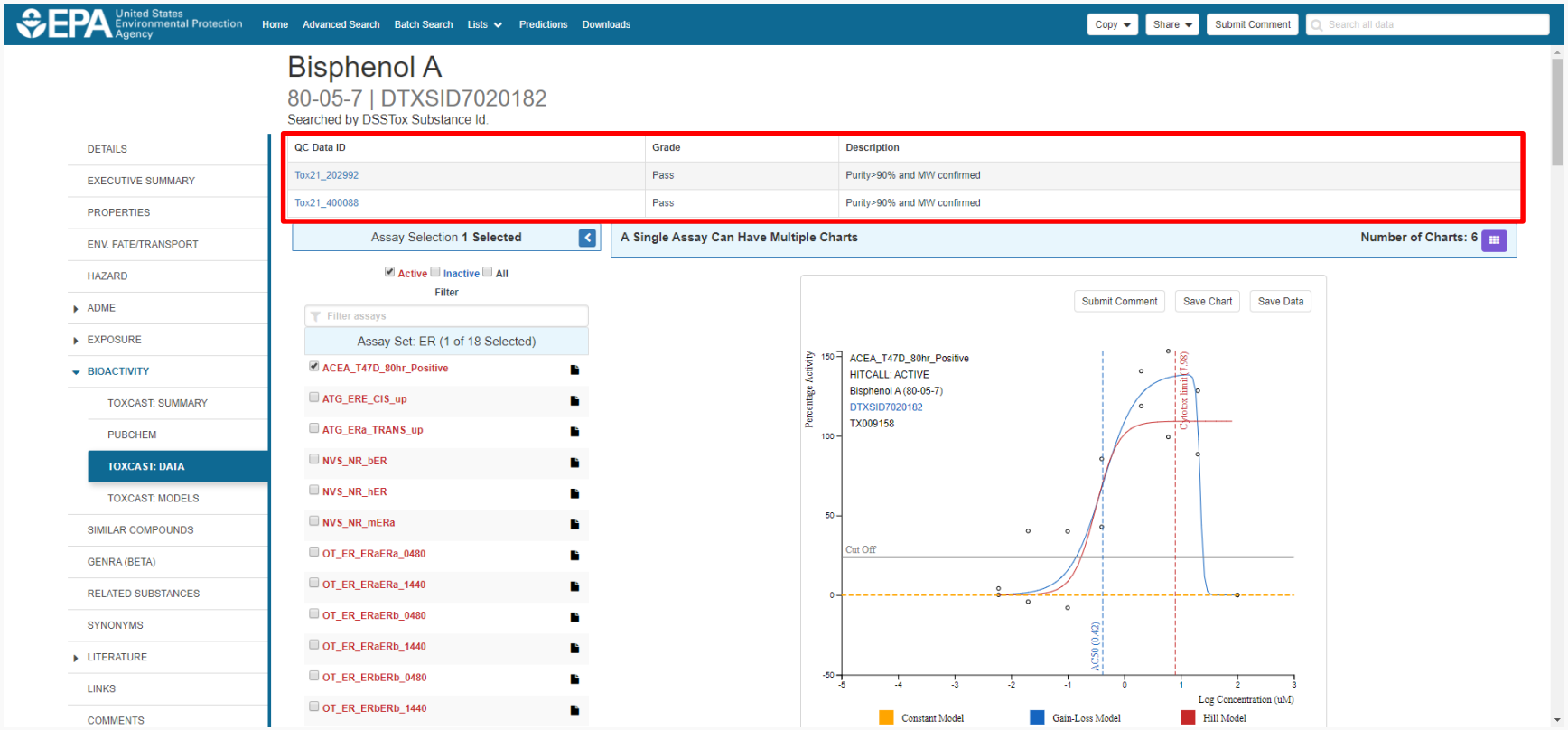
Resources

- Toxicity Forecaster (ToxCast) Fact Sheet
- ToxCast Publications
- ToxCast Citation
- About ToxCast

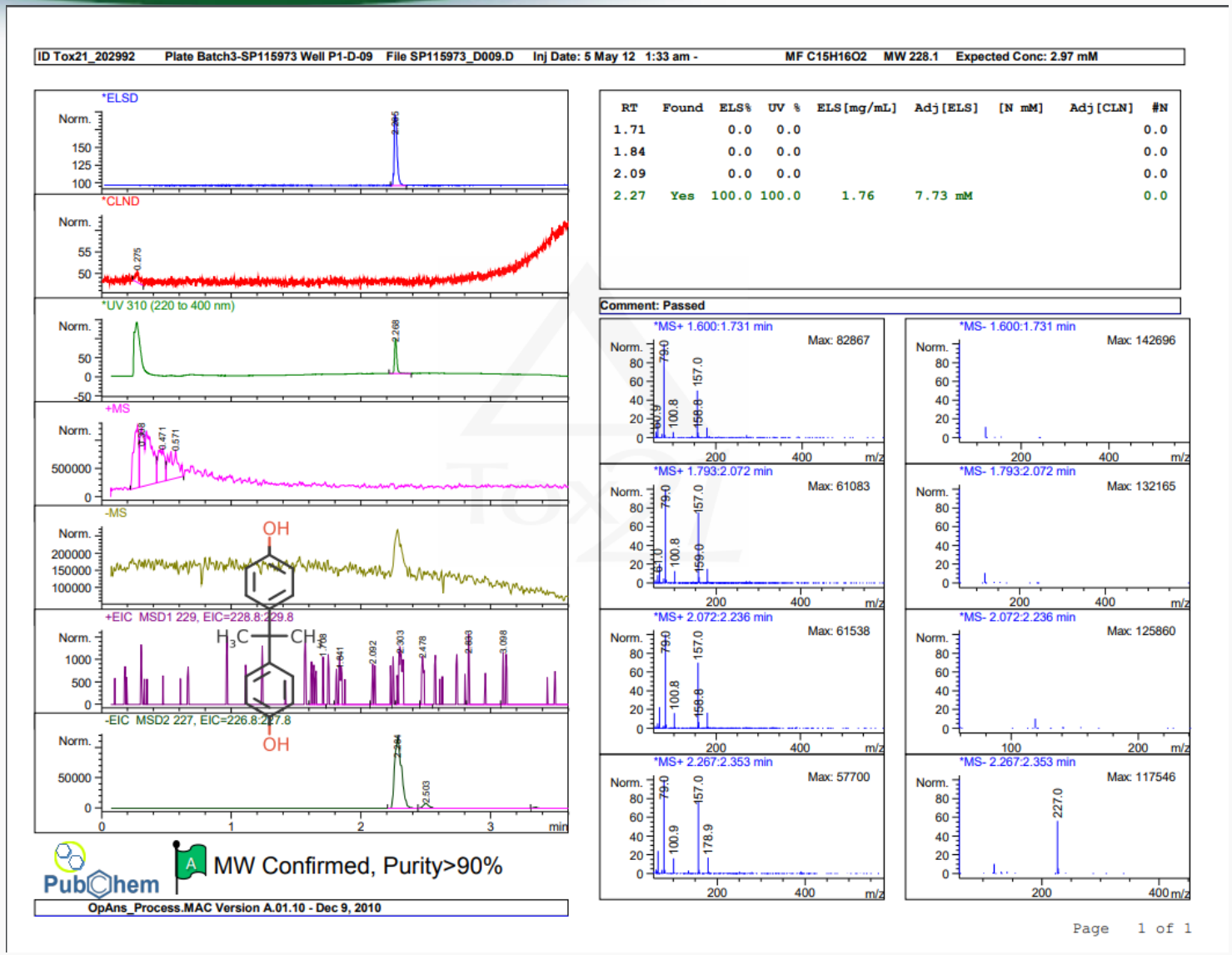
ToxCast Data and Information

- **ToxCast & Tox21 Summary Files.** Data for a single chemical endpoint pair for thousands of chemicals and assay endpoints for 20 variables such as the activity or hit call, activity concentrations, whether the chemical was tested in a specific assay, etc.
  - Download ToxCast Summary Information
  - Download ReadMe
- **ToxCast & Tox21 Data Spreadsheet.** A spreadsheet of EPA's analysis of the chemicals screened through ToxCast and the Tox21 collaboration which includes EPA's activity calls from the screening of over 1,800 chemicals.
  - Download Data
  - Download ReadMe
- **ToxCast Data Pipeline R Package.** The R computer programming package used to process and model all EPA ToxCast and Tox21 chemical screening data. The files include the R programming package as well as documents that provide overviews of the data analysis pipeline used and the R package. Users will need experience with R to use these files.
  - Download Package
  - TCPL Overview

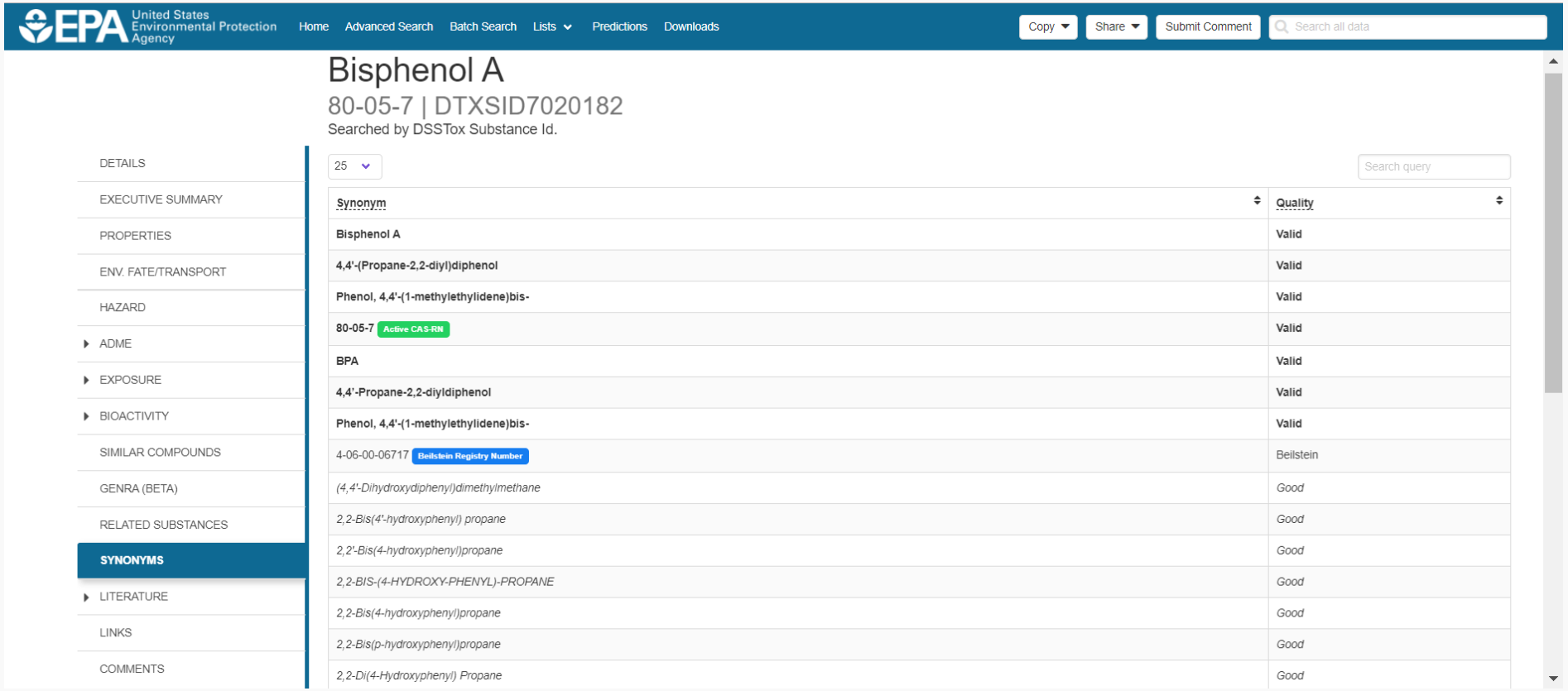
ToxCast/Tox21 Data
Analytical QC of the chemicals



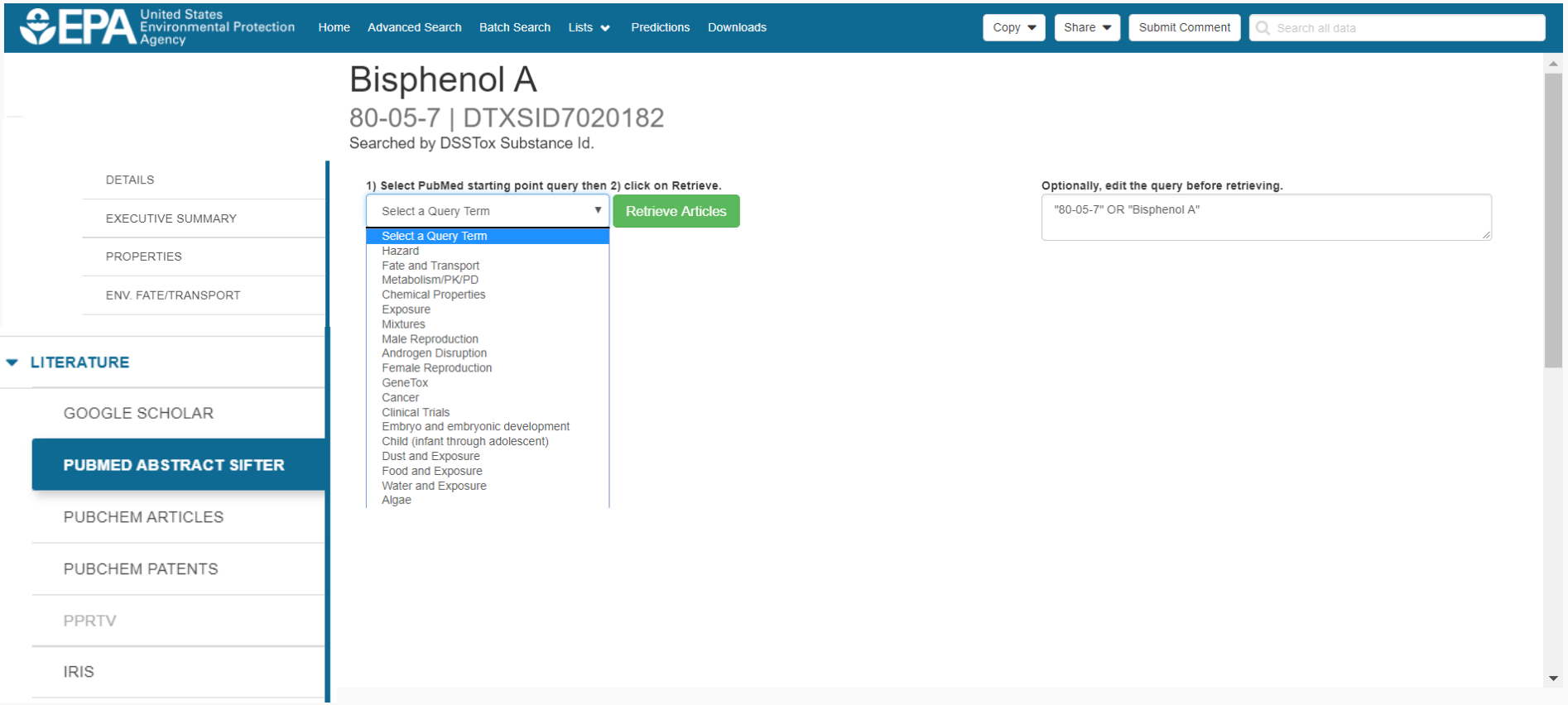
Access to Analytical QC Data



Names and CASRNs to Support Searches

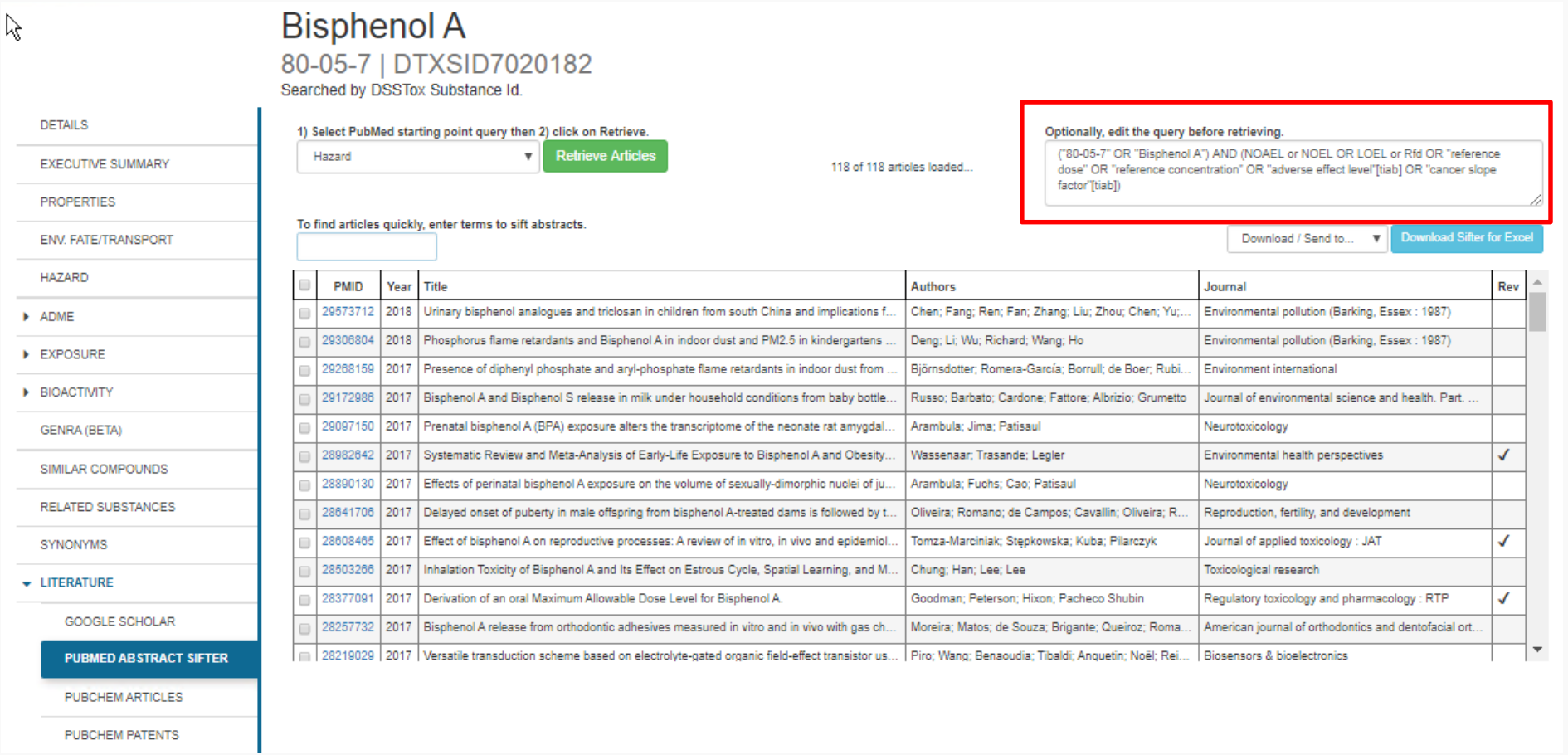


Literature Searches - Querying 28 Million PubMed abstracts

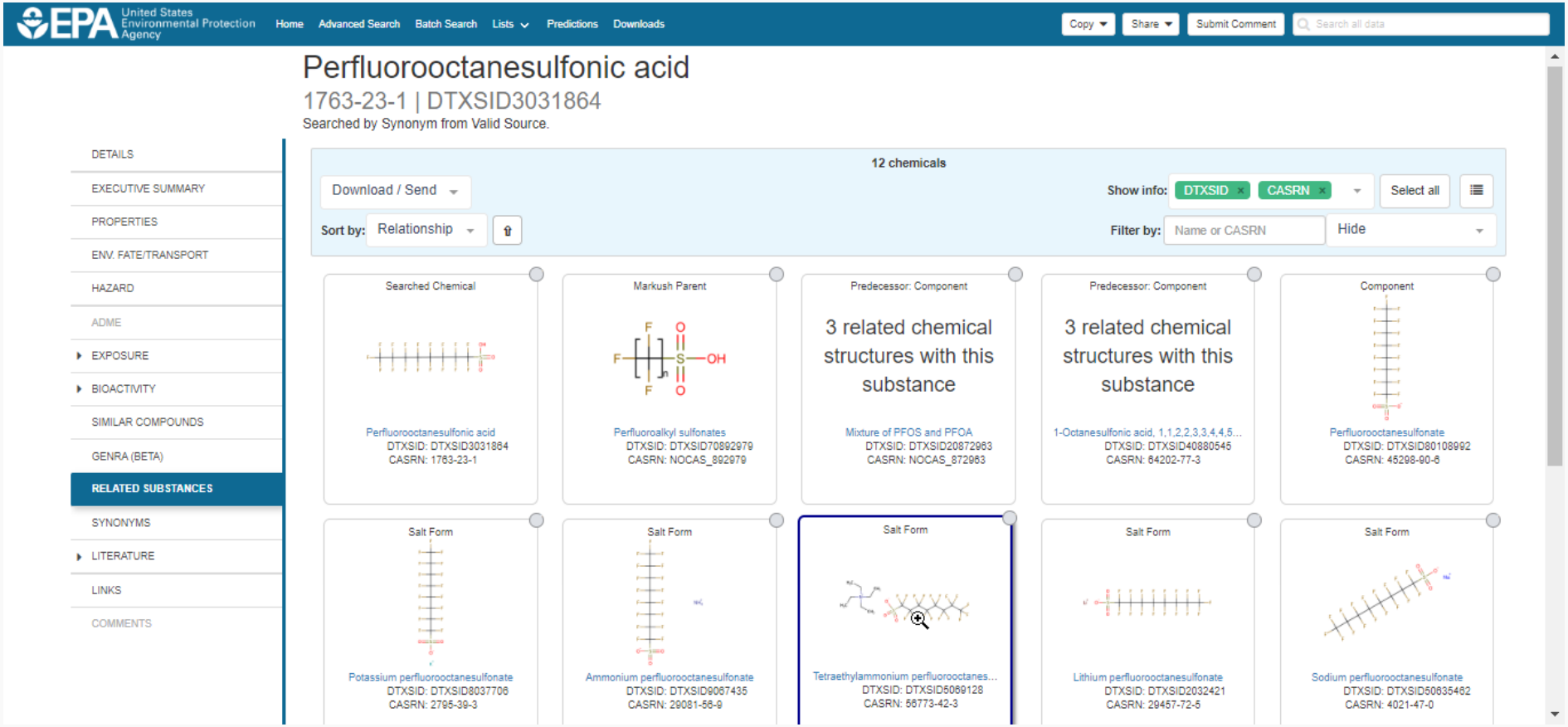


Abstract Sifter - Querying 28 Million PubMed abstracts

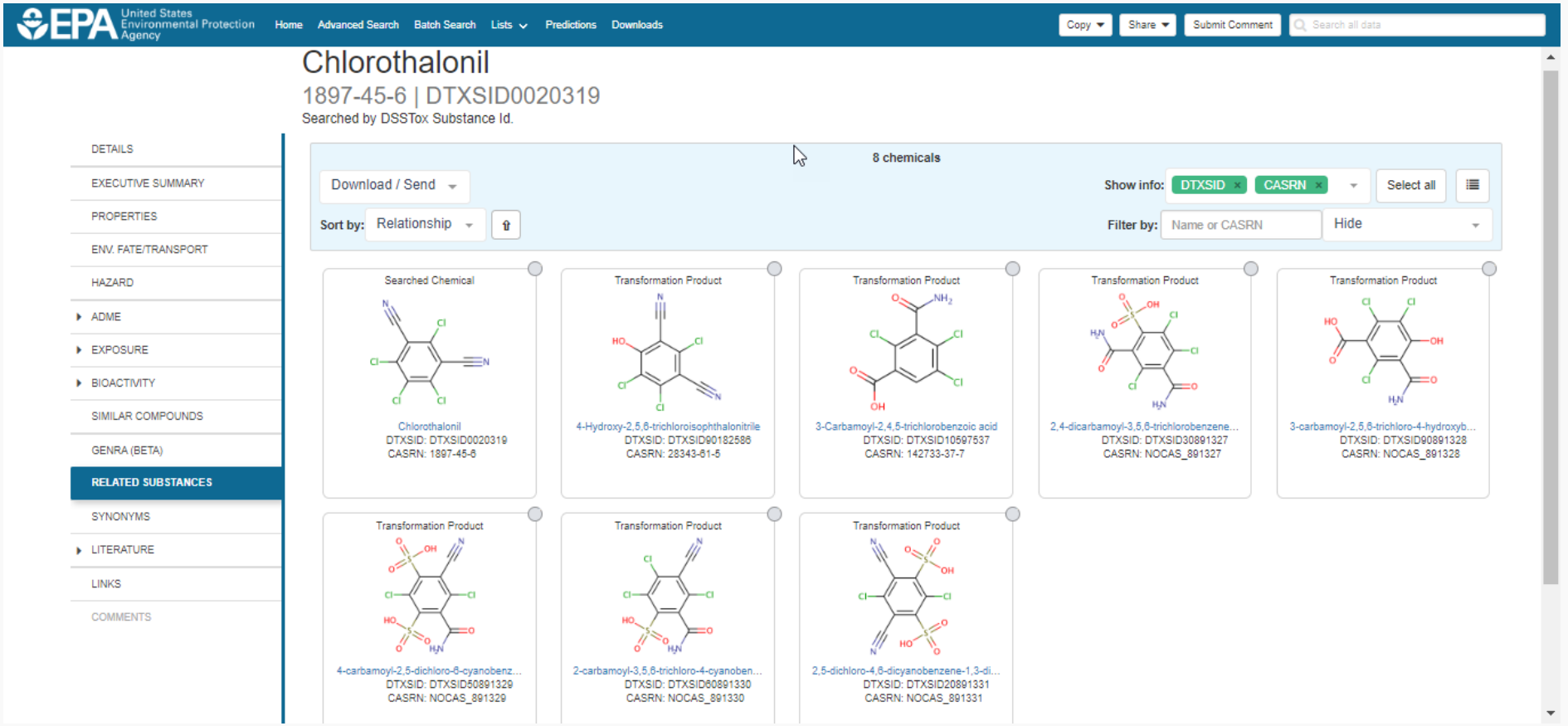
EPA United States Environmental Protection Agency



Relationships in the data



Transformation Products



7 salt forms of PFOS (and the ion itself)

EPA United States Environmental Protection Agency

Component

Perfluorooctanesulfonate
DTXSID: DTXSID80108992
CASRN: 45298-90-6

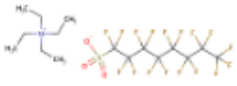
Salt Form

Potassium perfluorooctanesulfonate
DTXSID: DTXSID8037706
CASRN: 2795-39-3

Salt Form

Ammonium perfluorooctanesulfonate
DTXSID: DTXSID9067435
CASRN: 29081-56-9

Salt Form

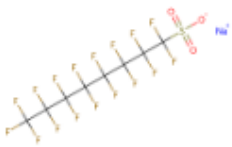


Tetraethylammonium perfluorooctanesu...
DTXSID: DTXSID5069128
CASRN: 56773-42-3

Salt Form

Lithium perfluorooctanesulfonate
DTXSID: DTXSID2032421
CASRN: 29457-72-5

Salt Form

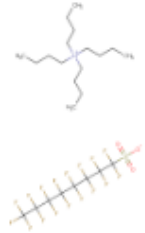


Sodium perfluorooctanesulfonate
DTXSID: DTXSID50635462
CASRN: 4021-47-0

Salt Form

Magnesium bis(heptadecafluorooctane...
DTXSID: DTXSID80881314
CASRN: 91036-71-4

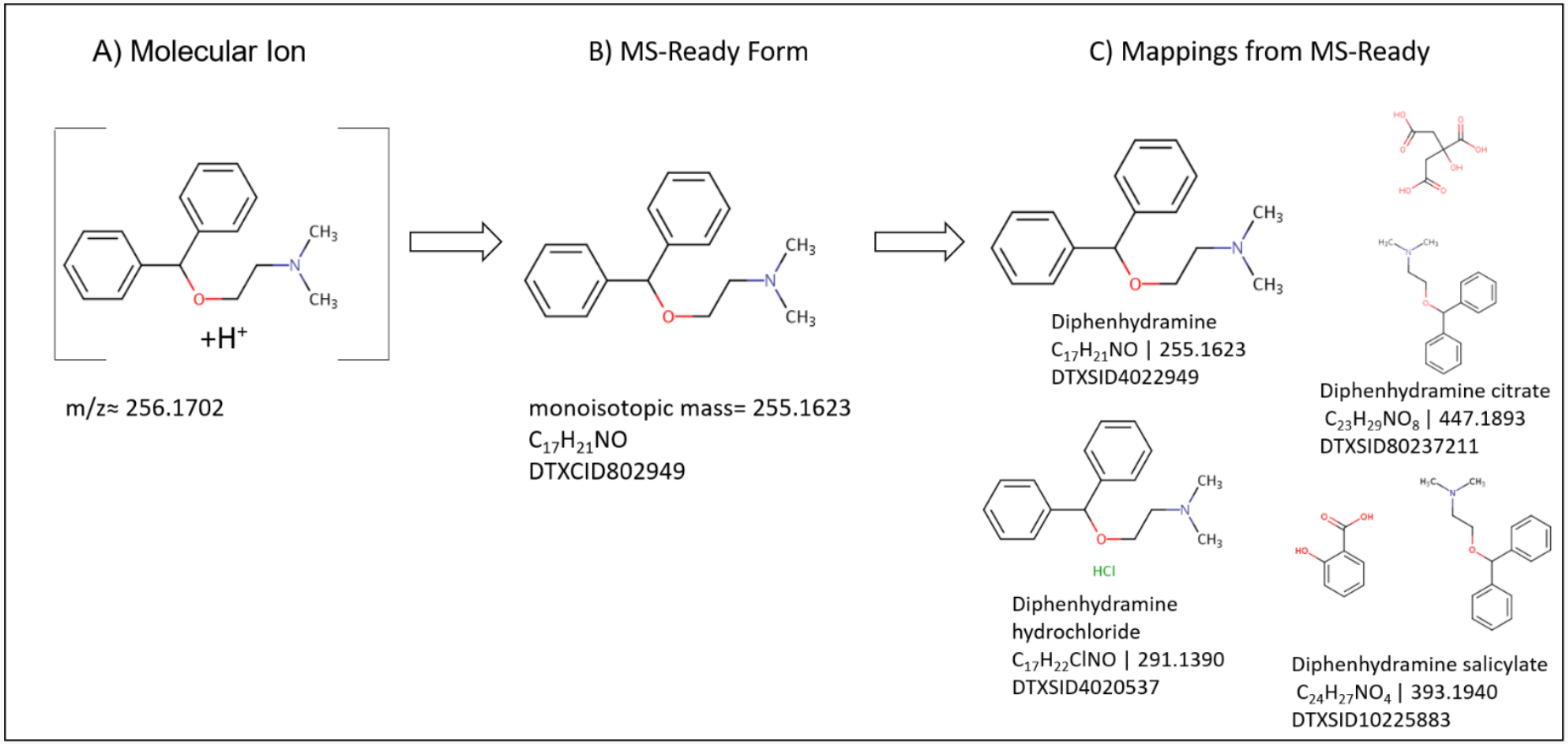
Salt Form



Tetrabutylammonium perfluorooctanesu...
DTXSID: DTXSID40584995
CASRN: 111873-33-7

Using data relationships

- We have purposely built relationships in the data. Specifically, “MS-Ready mappings”



“MS-Ready” : Suspect Screening and Non-Targeted Analysis Workflow

EPA United States Environmental Protection Agency

Suspect Screening

- Raw Samples
- Extracted Samples
- Raw Features
- “Molecular Features”
- DSSTox Chemical Database
- Matched Formulas
- Mapped Structures
- Prioritized Structures (using ToxPi)
- Confirmed Structures (using ToxCast standards)
- Predicted Concentrations

Non-Targeted Analysis

- Processed Features
- Prioritized Features
- Predicted Formulas
- Candidate Structures
- Sorted Structures
- Predicted Retention Times
- Predicted Mass Spectra
- Predicted/Observed Functional Use
- Predicted/Observed Media Occurrence
- Methodological Concordance
- Top Candidate Structure(s)

Color Key

Red = Analytical Chemistry

Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

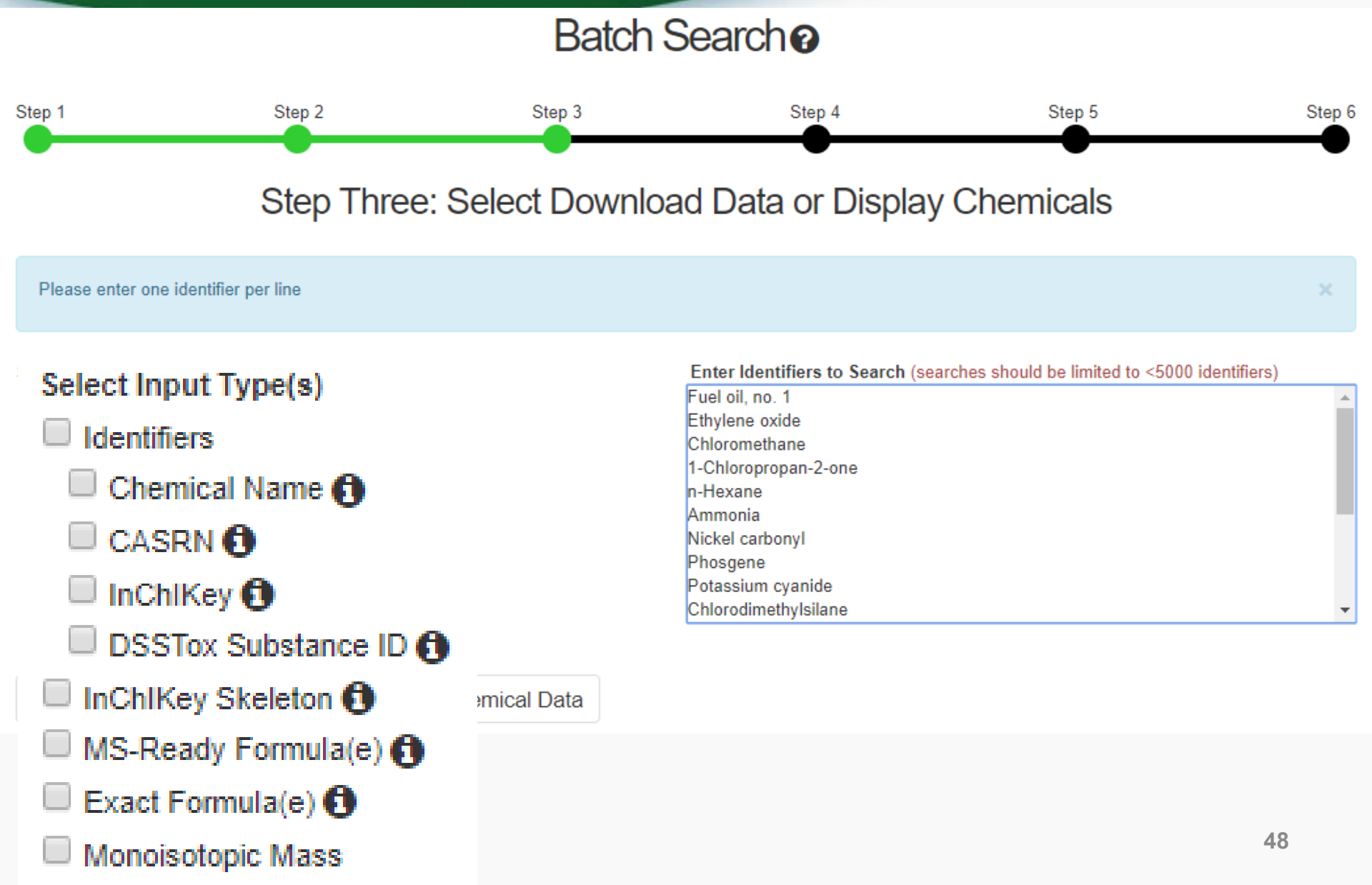
Green = Informatics & Web Services

Batch Searching

- Singleton searches are useful but we need to search thousands of masses, formulae, names, InChIs and CASRNs!

- Typical questions
  - What chemicals can I get for 5000 CAS Numbers?
  - Can I get predicted properties for 1000 chemicals?
  - What is the list of chemicals for the formula $C_xH_yO_z$?
  - What is the list of chemicals for a mass +/- error ?
  - Can I get chemical lists in Excel files? In SDF files?

Batch Searching



Batch Searching

EPA United States Environmental Protection Agency

Select Output Format:

Excel

Download

Customize Results

- Select All
- Select All in Lists

Chemical Identifiers

- DTXSID
- Chemical Name
- CAS-RN
- InChIKey
- IUPAC Name

Structures

- Mol File
- SMILES
- InChI String
- MS-Ready SMILES
- QSAR-Ready SMILES

Intrinsic And Predicted Properties

- Molecular Formula
- Average Mass
- Monoisotopic Mass
- TEST Model Predictions
- OPERA Model Predictions

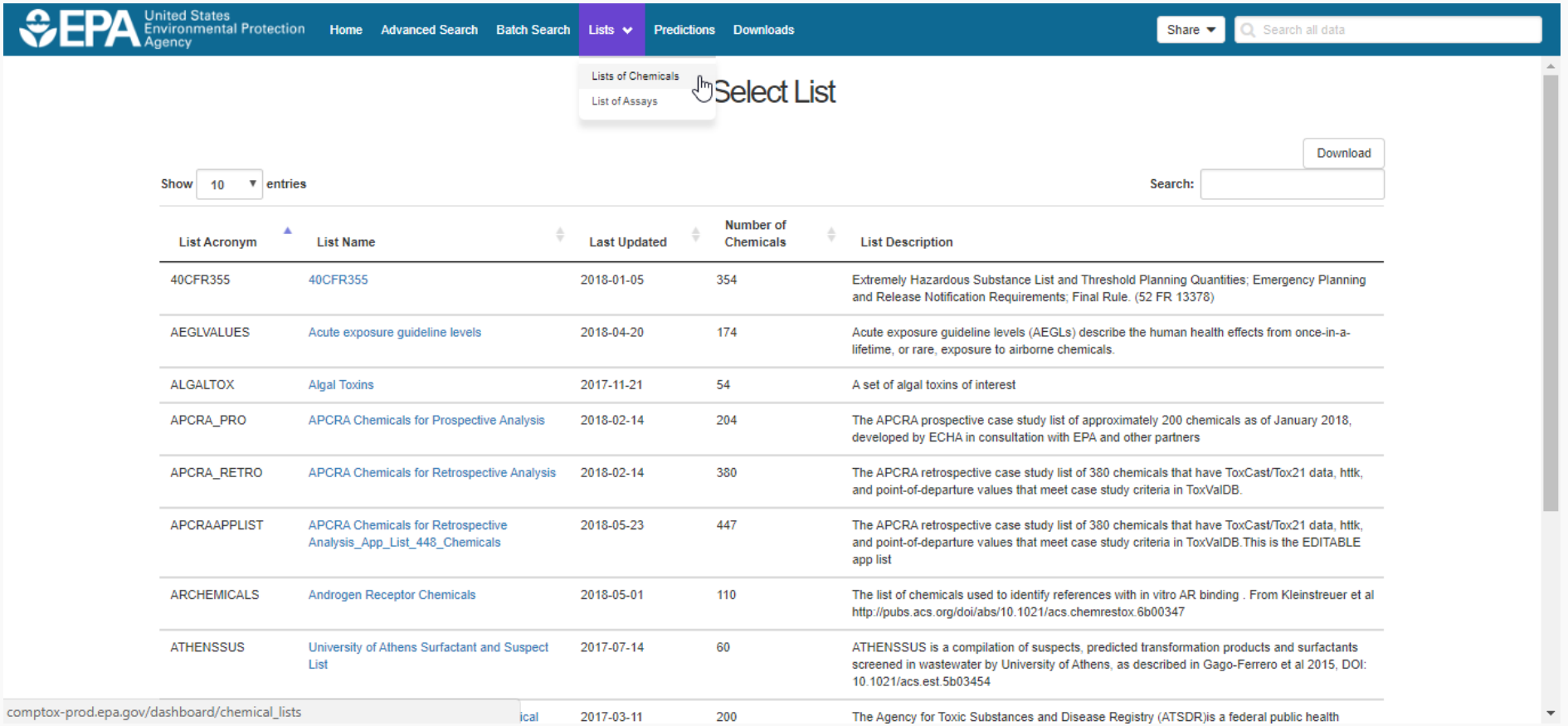
Presence in Lists:

- ICCVAM test method evaluation report: in vitro ocular toxicity test methods
- 40CFR355
- A list of all PBDEs (Polybrominated diphenyl ethers)
- A list of all PCBs (Polychlorinated biphenyls)
- A list of polycyclic aromatic hydrocarbons
- Acute exposure guideline levels
- Algal Toxins
- Androgen Receptor Chemicals
- APCRA Chemicals for Prospective Analysis
- APCRA Chemicals for Retrospective Analysis
- APCRA Chemicals for Retrospective Analysis_App_List_448_Chemicals
- ATSDR Minimal Risk Levels (MRLs) for Hazardous Substances
- ATSDR Toxic Substances Portal Chemical List
- Bisphenol Compounds
- California Office of Environmental Health Hazard Assessment
- Chemicals with interesting names
- CMAP
- DNT Screening Library
- Drinking Water Suspects, KWR Water, Netherlands
- EDSP Universe
- EPA Chemicals associated with hydraulic fracturing

Excel Output

	A	B	C	D	E	F	G	H
1	INPUT	FOUND_BY	DTXSID	PREFERRED_NAME	EXPOCAST	EXPOCAST	NHANES	TOXVAL_D
2	1445-75-6	CAS-RN	DTXSID5024051	Diisopropyl methylpho	2.09e-08	Y	-	Y
3	50-00-0	CAS-RN	DTXSID7020637	Formaldehyde	1.32e-06	Y	-	Y
4	107-06-2	CAS-RN	DTXSID6020438	1,2-Dichloroethane	4.9e-06	Y	-	Y
5	57-12-5	CAS-RN	DTXSID6023991	Cyanide	-	-	-	Y
6	7550-45-0	CAS-RN	DTXSID8042476	Titanium tetrachloride	-	-	-	Y
7	79-01-6	CAS-RN	DTXSID0021383	Trichloroethylene	7.27e-06	Y	-	Y
8	121-82-4	CAS-RN	DTXSID9024142	Cyclonite	6.72e-08	Y	-	Y
9	108-05-4	CAS-RN	DTXSID3021431	Vinyl acetate	8.3e-05	Y	-	Y
10	7803-51-2	CAS-RN	DTXSID2021157	Phosphine	-	-	-	Y
11	122-66-7	CAS-RN	DTXSID7020710	1,2-Diphenylhydrazine	1.49e-07	Y	-	Y
12	101-77-9	CAS-RN	DTXSID6022422	4,4'-Methylenedianiline	6.08e-06	Y	-	Y
13	14017-34-6	CAS-RN	DTXSID90161250	Selenium difluoride	-	-	-	-
14	75-44-5	CAS-RN	DTXSID0024260	Phosgene	-	-	-	Y
15	621-64-7	CAS-RN	DTXSID6021032	N-Nitrosodipropylamine	4.55e-07	Y	-	Y
16	75-09-2	CAS-RN	DTXSID0020868	Dichloromethane	2.02e-06	Y	-	Y
17	100-41-4	CAS-RN	DTXSID3020596	Ethylbenzene	8.32e-05	Y	-	Y
18	7440-28-0	CAS-RN	DTXSID2036035	Thallium	-	-	-	Y
19	108-88-3	CAS-RN	DTXSID7021360	Toluene	8.61e-05	Y	-	Y
20	111-44-4	CAS-RN	DTXSID9020168	Bis(2-chloroethyl) ethe	2.82e-07	Y	-	Y
21	7440-42-8	CAS-RN	DTXSID3023922	Boron	-	-	-	Y
22	7440-29-1	CAS-RN	DTXSID6049800	Thorium	-	-	-	Y

List of Chemicals



EPA activities around PFAS chemicals

https://www.epa.gov/pfas

EPA United States Environmental Protection Agency

Per- and Polyfluoroalkyl Substances (PFAS)

CONTACT US

SHARE

"The National Leadership Summit on PFAS provided an unprecedented opportunity for stakeholders to share vital information and best practices regarding PFAS." - *Former Administrator Pruitt*

- Community Events
- Infographic

Basic Information

- What are PFAS?
- Why are PFAS important?
- How people are exposed?

EPA Actions to Address PFAS

- EPA actions
- National leadership summit and engagement

Tools and Resources

- EPA data and tools
- State information
- Site-specific resources

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

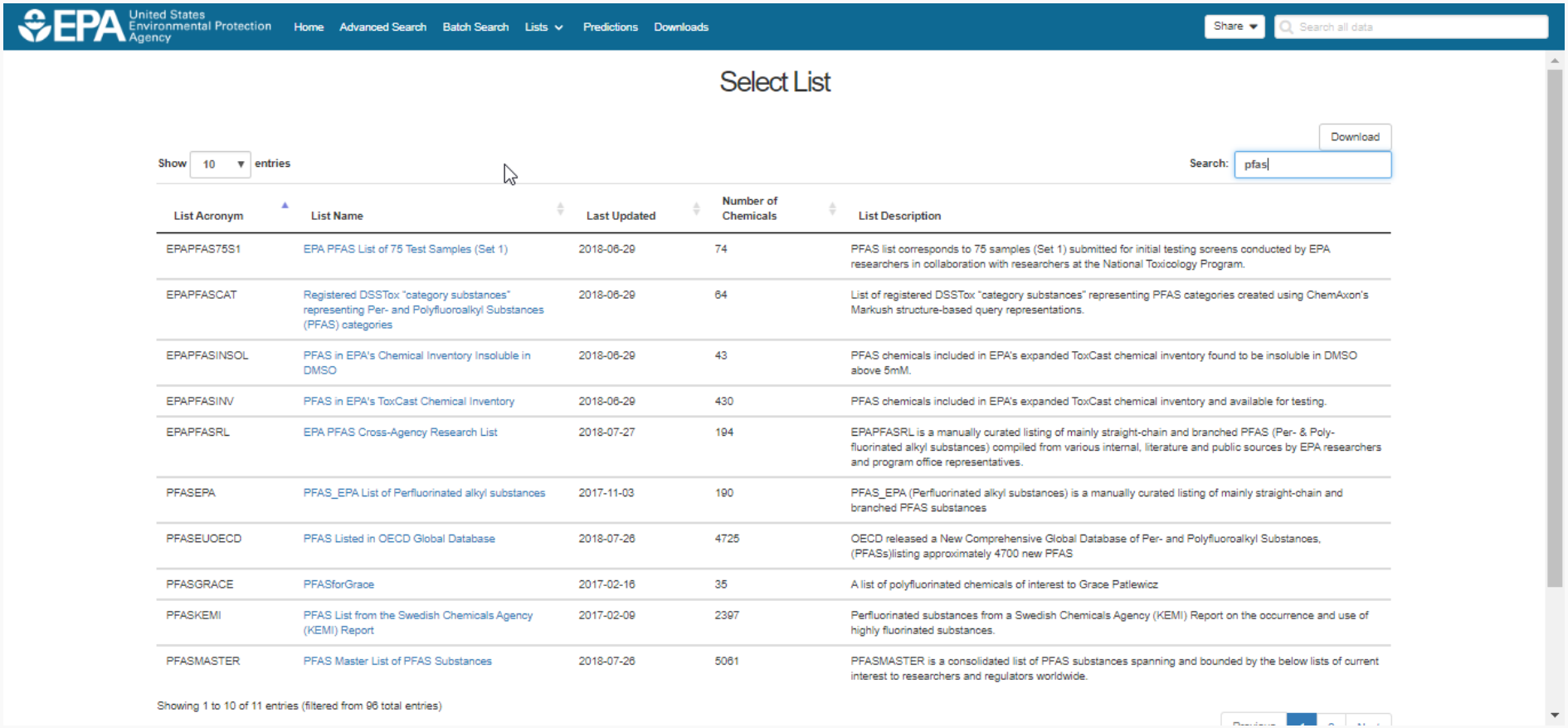
The OECD releases a new list of Per- and Polyfluoroalkyl Substances (PFASs) based on a comprehensive analysis of information available in the public domain. In total, 4730 PFAS-related CAS numbers have been identified and categorised in this study, including several new groups of PFASs that fulfil the 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.

11 PFAS Lists

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



PFAS Categories in Development

Registered DSSTox "category substances" representing Per- and Polyfluoroalkyl Substances (PFAS) categories

Search EPAPFASCAT Chemicals

Substring search

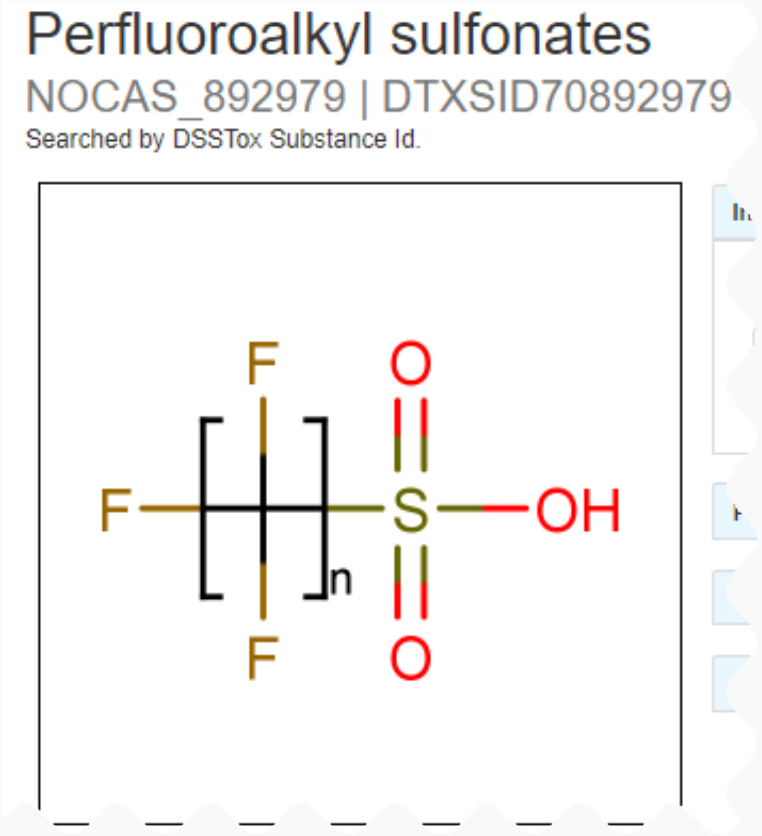
List Details

Description: List of registered DSSTox "category substances" representing Per- and Polyfluoroalkyl Substances (PFAS) categories created using ChemAxon's Markush structure-based query representations. Markush categories can be broad and inclusive of more specific categories, or can represent a unique category not overlapping with other registered categories. Each PFAS category registered with a unique DTXSID is considered a generalized substance or "parent ID" that can be associated with one or many "child IDs" (i.e. many parent-child mappings) within the full DSSTox database. These category DTXSIDs can be used to search and retrieve all currently registered DSSTox substances within the category group, and offer an objective, transparent and reproducible structure-based means of defining a category of chemicals. This list and the corresponding category mappings is undergoing continuous curation and expansion.

Number of Chemicals: 64

Markush Chemicals

- PFOS is a member of linear perfluoroalkyl sulfonates



Collaborative Data Curation

- Mapping between our data (and websites) has resulted in collaborative data curation
- Collaboration with Emma Schymanski re. the NORMAN Suspects Exchange https://www.norman-network.com/?q=node/236
- Multiple NORMAN lists now mapped

NORMAN Suspect Exchange

EPA United States Environmental Protection Agency

NORMAN

Network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances

Home | NORMAN Network | Working Groups | Membership | NORMAN Bulletin | Success Stories | Publications | Job opportunities | Contact | Gallery | Members' Area | NORMAN GA meetings

Menu

- Emerging Substances
- DATABASES
- Topics and Activities
- Workshops and Events
- QA/QC Issues
- Glossary
- Useful links
- Members' Area

Search

User login

Username *

Password *

Request new password

Log in

Home

http://www.norman-network.com/?q=node/236

NORMAN Suspect List Exchange

In September 2014, NORMAN members expressed the need to exchange various lists of substances to improve their suspect screening efforts. This website was established aspart of the 2015 Joint Programme of Activities as a central access point for NORMAN members (and others) to find suspect lists relevant for their environmental monitoring question. All suspect lists currently available are compiled in the table below and on the US EPA CompTox Chemistry Dashboard (**website**, **downloads**, **chemical lists**).

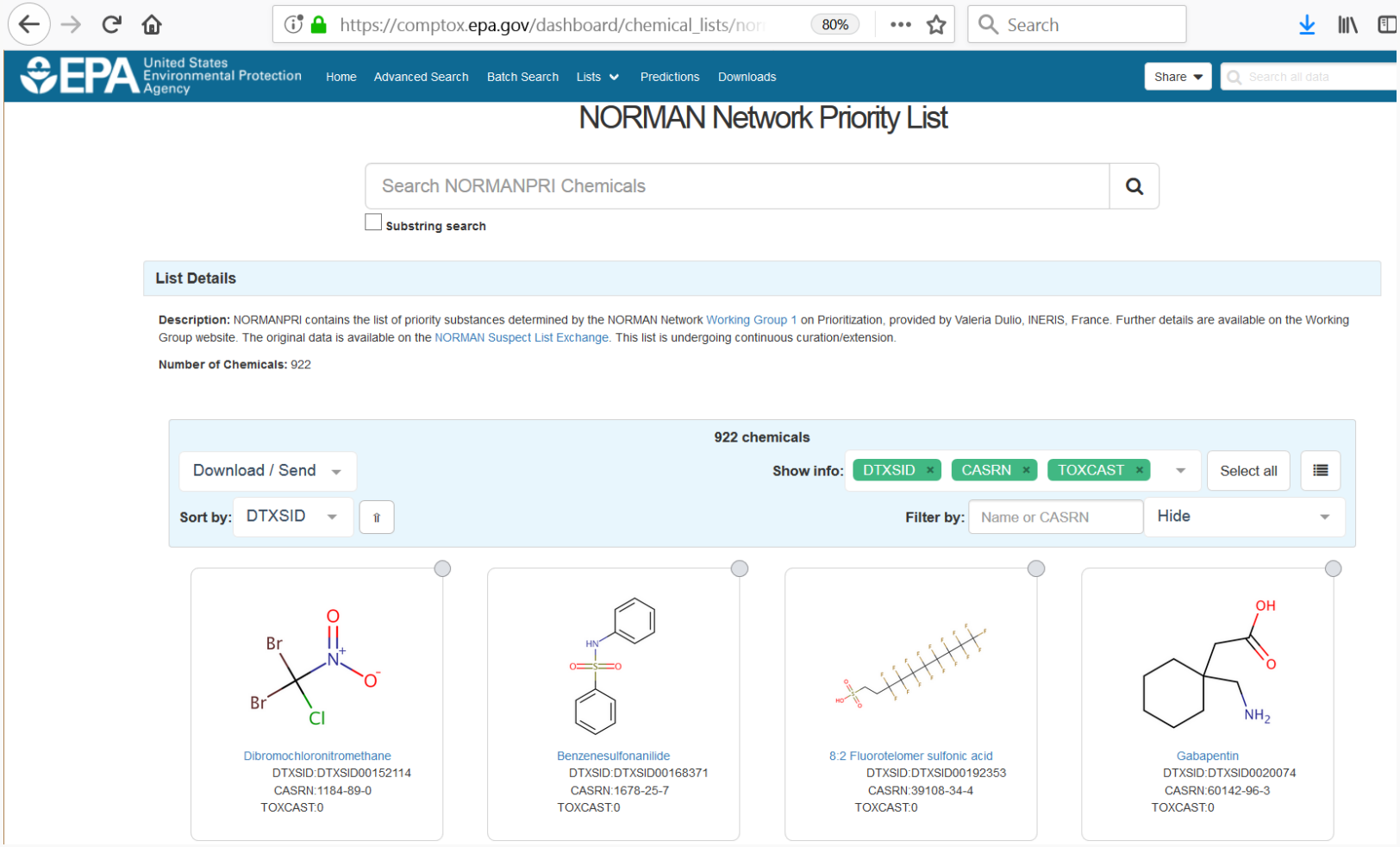
The "Link to full list" column below contains an excel or comma-separated file (csv) with all available information, e.g. as provided as supporting information for the publication, while the third column provides a list of the structures as InChIKeys only, which allows suspect searching using MetFrag or other workflows. The fourth column contains references for the data: please cite these references if you use the respective datasets.

Recent Suspect Exchange and Dashboard presentations/publications include: **ICCE Oslo 2017**: **NORMAN Suspects meet the Dashboard** and **NORMAN MassBank and Suspect Exchange**; SETAC **Mixtures** Denver: **Identifying Complex Mixtures with Cheminformatics and HR-MS**; ACS Fall 2017: **Markush Enumeration for UVCBs** and a **viewpoint article**.

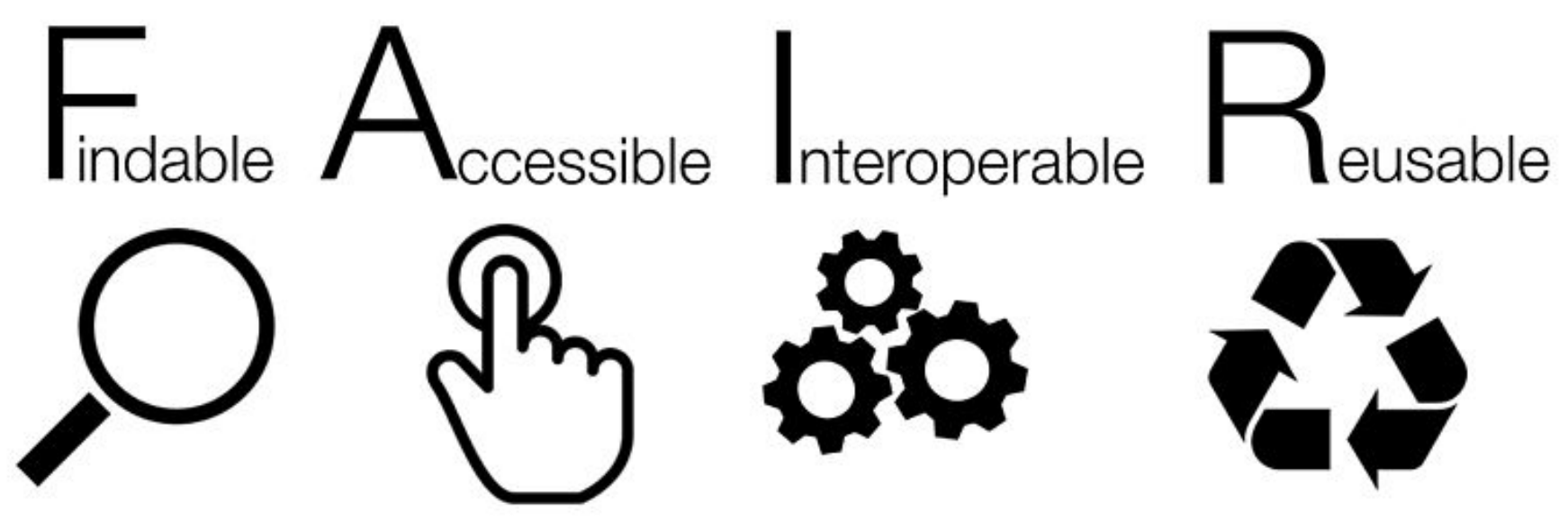
No.	Abbreviation	Description	Link to full list	Link to InChIKey list	References
	SUSDAT	**Merged NORMAN Suspect List: SusDat**	**Interactive Data table** (updating…)	**MS-ready InChIKeys** (1/03/2018)	A merged list of >40,000structures from suspect lists. See **interactive version**. Compiled by Reza Aalizadeh, University of Athens, including RTI and toxicity values, support by Nikiforos Alygizakis, EI. *Work in progress ... please report any issues!*
S1	MASSBANK	**NORMAN Compounds in MassBank**	**CSV**, **XLSX** with Fragments (3/10/2017) CompTox **MassBank EU Reference List** CompTox **MassBank EU Special Cases** CompTox **Fragment Download**	**MassBankEUInChIKeys** (11/04/2017)	**www.massbank.eu** Stravs *et al.* 2013. DOI: **10.1002/jms.3131**
S2	STOFFIDENT	**HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances**	**STOFF-IDENT Contents** (6/09/2017) CompTox STOFF-IDENT List Further curation in progress…	**STOFF-IDENT InChIKeys** (6/09/2017)	The database enables the search for exact masses from target or unknown lists and the automatic use of a Retention Time Index. See: **https://www.lfu.bayern.de/stoffident/#!home** (single search for free; batch search after free registration).
S3	NORMANCT15	**NORMAN Collaborative Trial Targets and Suspects**	LC-MS: **CSV**, **XLSX** (3/10/2017) GC-MS: **CSV**, **XLSX** (3/10/2017) CompTox **NORMANCT15 List**	**LC-MS InChIKeys** (31/10/2016) **GC-MS InChIKeys** (31/10/2016)	Schymanski *et al.* 2015. DOI: **10.1007/s00216-015-8681-7**

Example: NORMAN Priority List

S15	NORMANPRI	**NORMAN Priority List**	NORMAN Priority **CSV** (13/7/2017) CompTox **NORMAN Priority List** Further curation in progress…	**NORMAN Priority InChIKeys** (16/05/2017)	Priority substances from NORMAN WG-1 (Prioritisation), provided by Valeria Dulio.



Our support for FAIR Data



- We're definitely not there yet…but are making progress at speed…

How we are serving FAIR

- Our data are licensed as public domain data
  - available from downloads page
  - registered to Figshare
  - SQL data dumps
- Collection of web services for old dashboards are available – API is being fully revamped

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"
```

How we are serving FAIR

- Our data are licensed as public domain data
  - available from downloads page
  - registered to Figshare
  - SQL data dumps
- Collection of web services for old dashboards are available – API is being fully revamped
- Models are published to Github

OPERA Models on Github
https://github.com/kmansouri

Kamel Mansouri
kmansouri

- Computational Chemistry/Toxicology - Cheminformatics and data mining - QSAR/QSPR and ADME-Tox properties modeling [orcid:0000-0002-6426-8036]

Block or report user

RTP, NC, USA
https://www.linkedin.com/in/ka...

Overview | Repositories 3 | Stars 3 | Followers 5 | Following 4

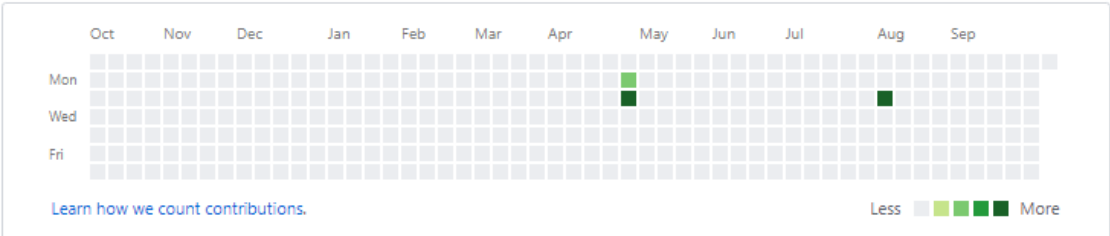
Popular repositories

OPERA
Command line application providing QSAR models predictions as well as applicability domain and accuracy assessment for physicochemical properties and environmental fate endpoints.
★3 1

QSAR-ready
Standardization workflow for QSAR-ready chemical structures pretreatment.

MS-ready
Standardization workflow for MS-ready chemical structures pretreatment.

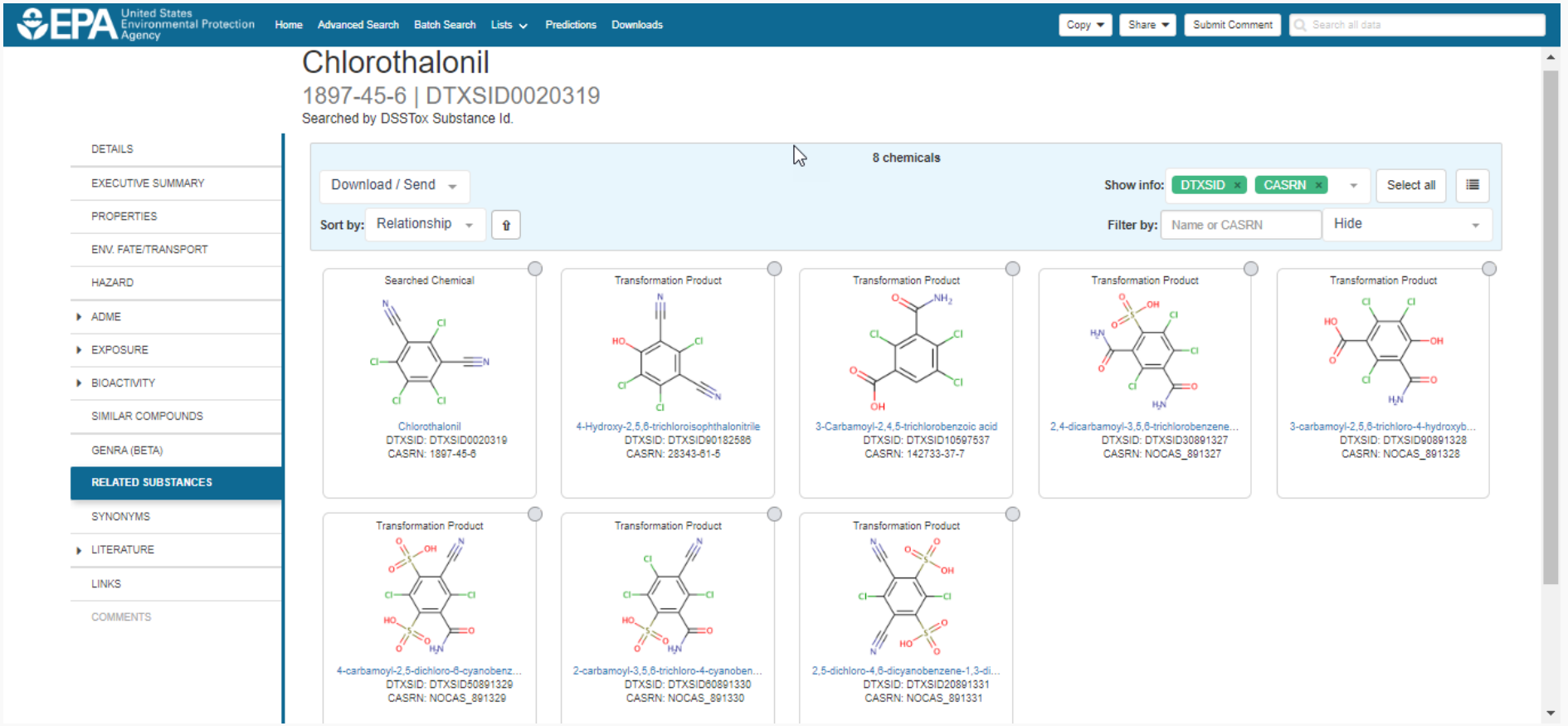
12 contributions in the last year



In Progress : “InvitroDB_v3”

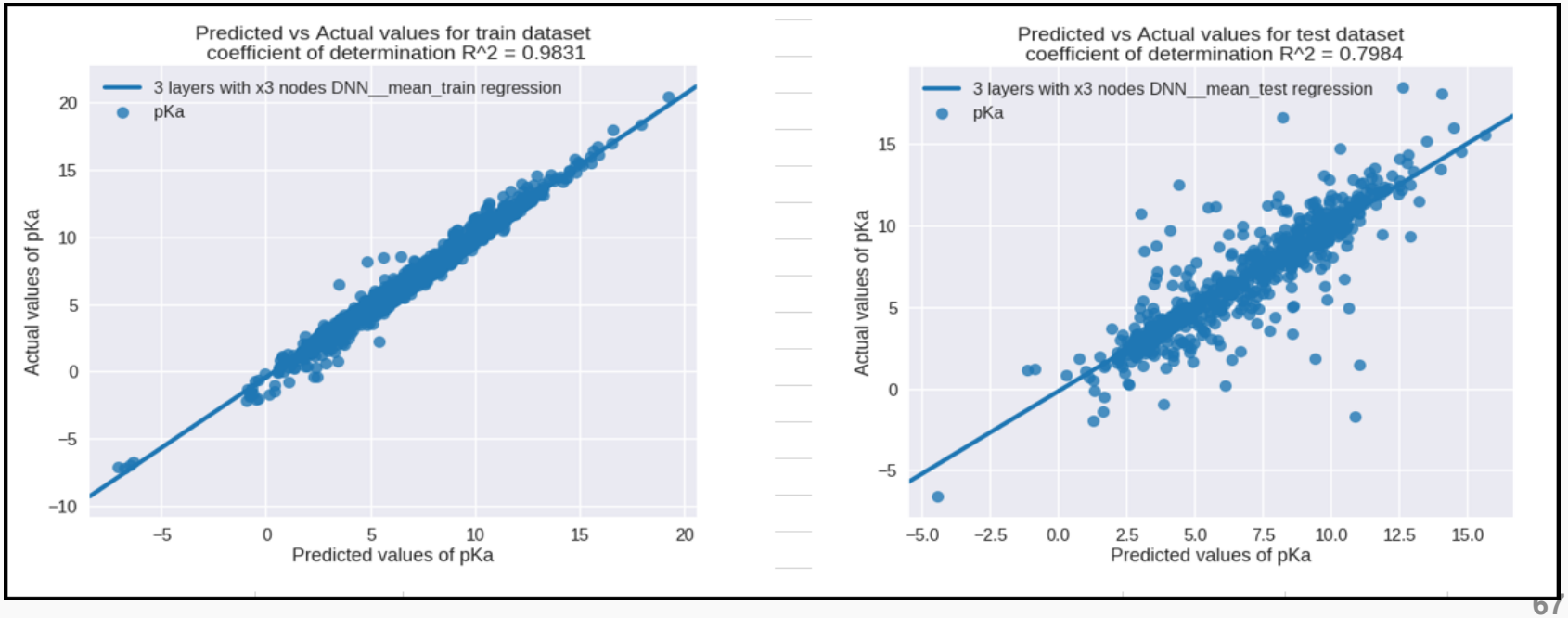
- The last public release of ToxCast data (invitroDB_v2) was in 3rd Quarter of 2015
- Next release of invitroDB_v3 is Fall 2018
- Data includes new assays, new chemicals, new pipelining, results of data curation
- Data will also release via the Dashboard
- Data available at https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data

In Progress: Metabolites extracted from literature and databases



In Progress : pKa Prediction Model

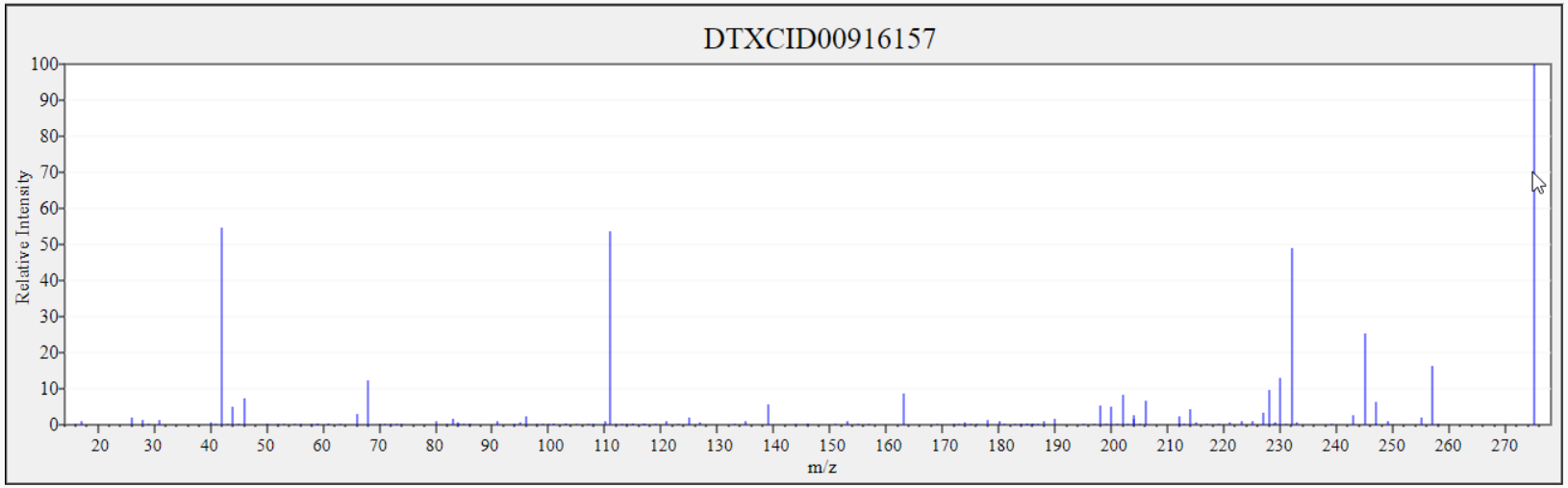
- pKa prediction models based on Open Data Set of 8000 chemicals – acidic, basic and amphoteric chemicals



In Progress: 700k Predicted MS

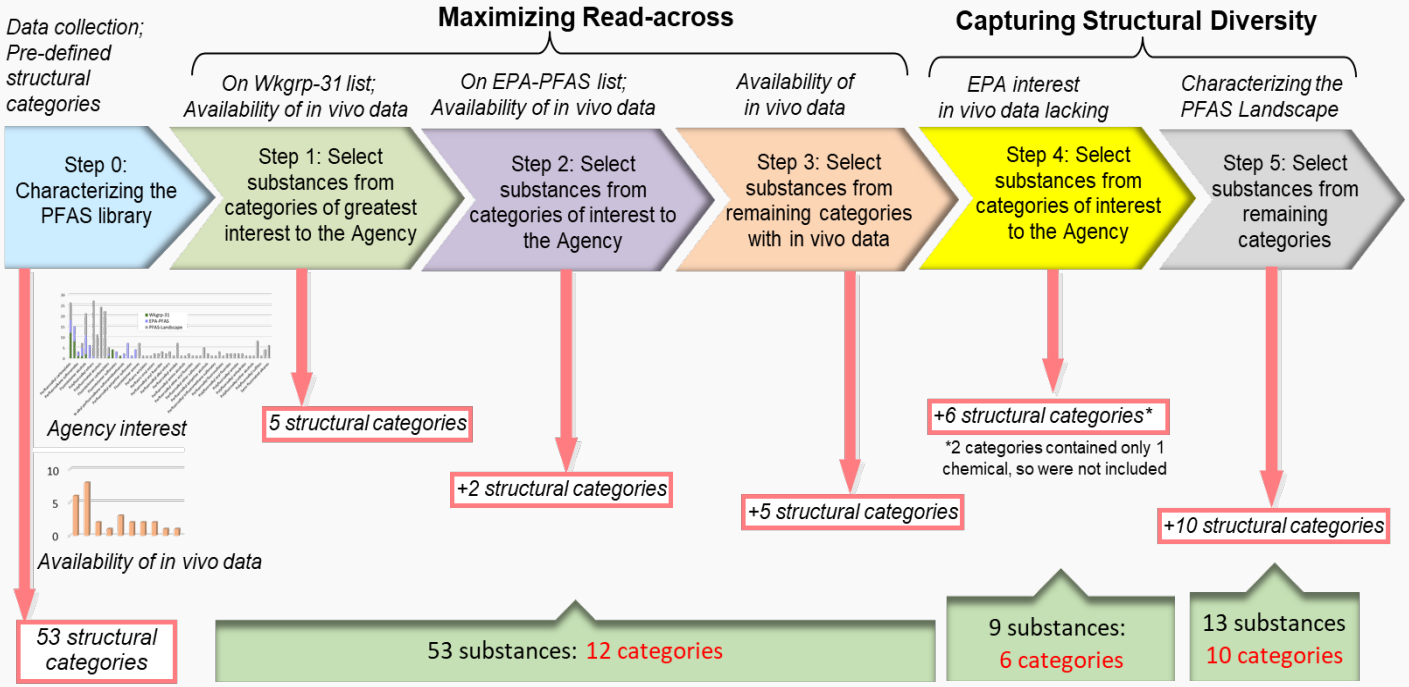
http://cfmid.wishartlab.com/

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



In progress: PFAS Chemical Library

- Development of a high-throughput screening library and collection of physical samples (~400)
- 75 PFAS chemicals for screening based on categories, diversity, exposure considerations, procurability and testability, availability of existing toxicity data



Conclusion

- Transparent access to data supporting computational toxicology – file downloads, SQL data dumps and web services
- CompTox Chemicals Dashboard provides access to data for ~765,000 chemicals
- Our publications provide chemistry data in usable formats – Excel, SDF (V2000/V3000)
- Web Services developing to serve API development
- Next release: Fall 2018 with InvitroDB_v3 data

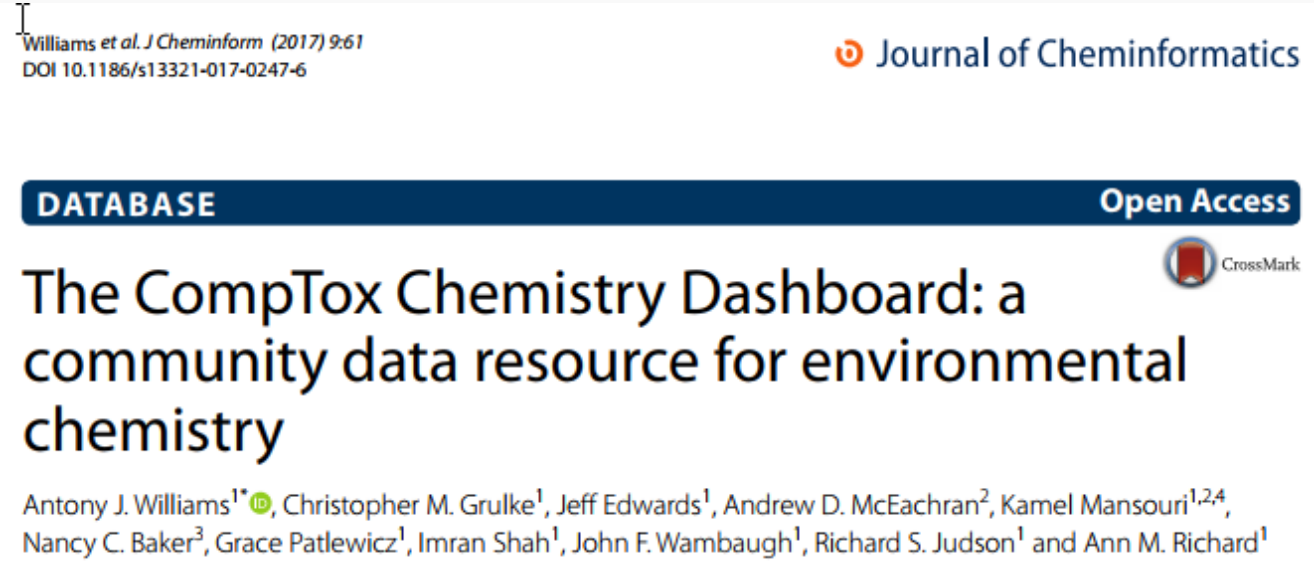
Acknowledgements

EPA-RTP

- *An enormous team of contributors from NCCT, especially the IT software development team*
- *Multiple centers and laboratories across the EPA*
- *Many public domain databases and open data contributors*

Contact

Antony Williams
NCCT, US EPA Office of Research and Development,
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821



Williams et al. J Cheminform (2017) 9:61
DOI 10.1186/s13321-017-0247-6

Journal of Cheminformatics

DATABASE Open Access

The CompTox Chemistry Dashboard: a community data resource for environmental chemistry

Antony J. Williams[1*], Christopher M. Grulke[1], Jeff Edwards[1], Andrew D. McEachran[2], Kamel Mansouri[1,2,4], Nancy C. Baker[3], Grace Patlewicz[1], Imran Shah[1], John F. Wambaugh[1], Richard S. Judson[1] and Ann M. Richard[1]

https://doi.org/10.1186/s13321-017-0247-6