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

EPA United States Environmental Protection Agency

Chemistry data: Distortion and dissemination in the Internet Era

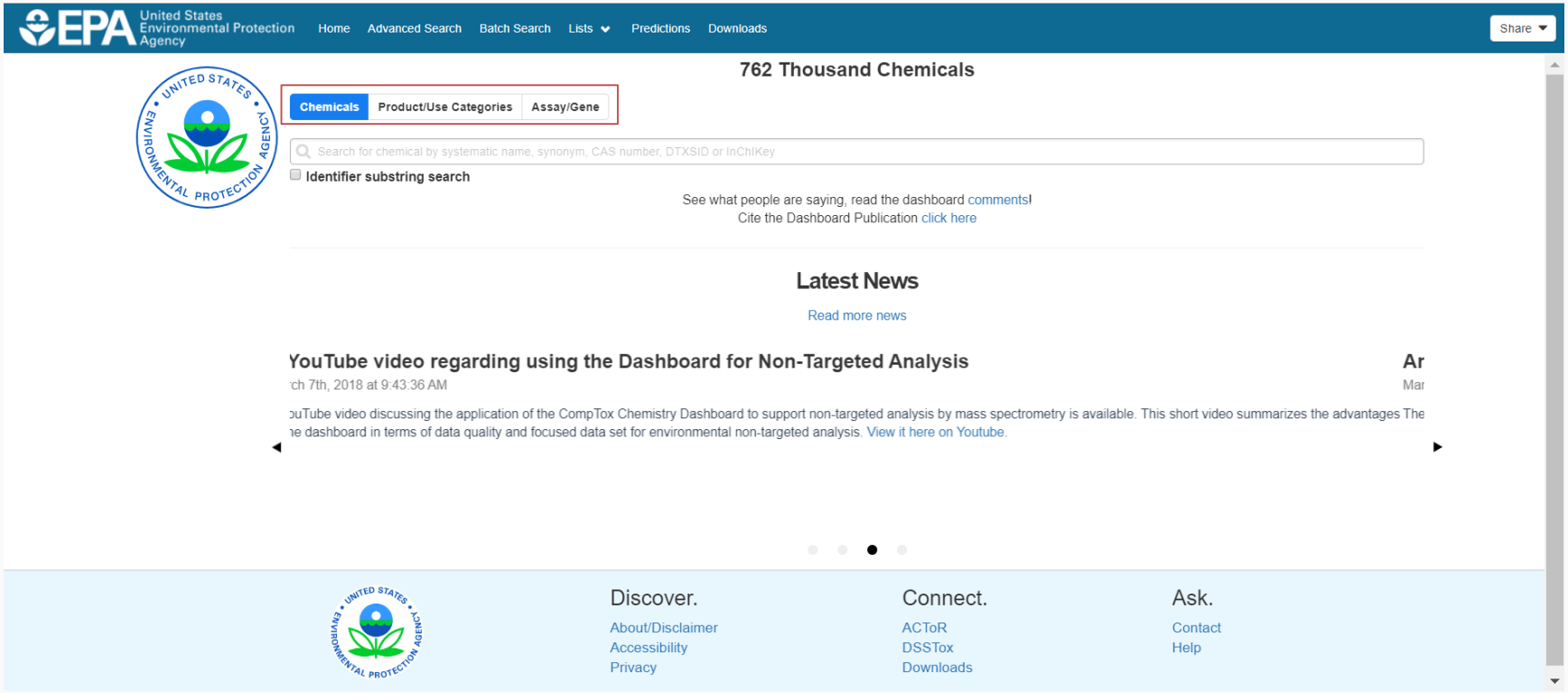
Antony Williams[1], Chris Grulke[1], Katie Paul-Friedman[1], and Emma Schymanski[2]

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC
2) Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, Luxembourg

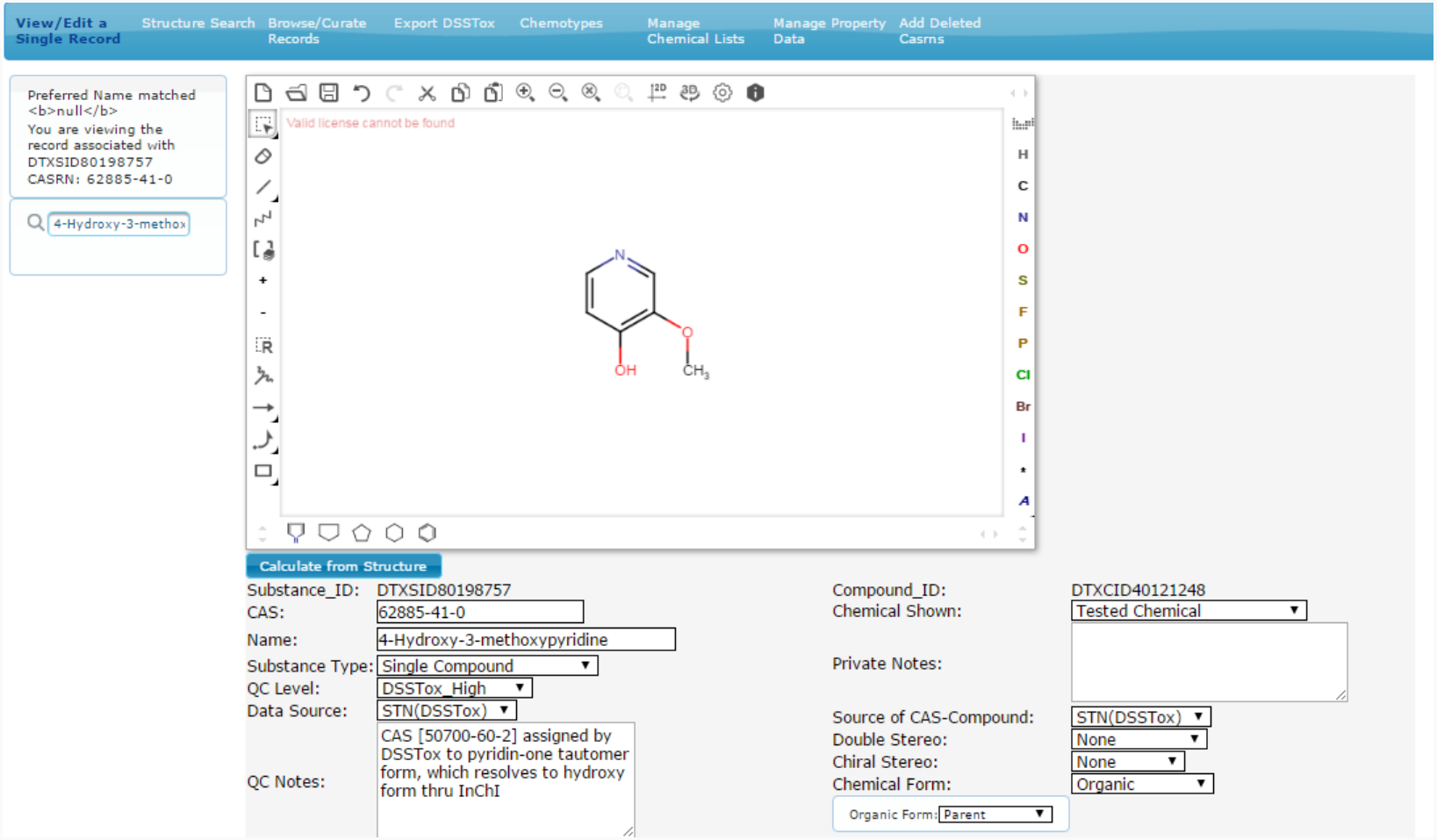
August 2018
ACS Fall Meeting, Boston

CompTox Dashboard

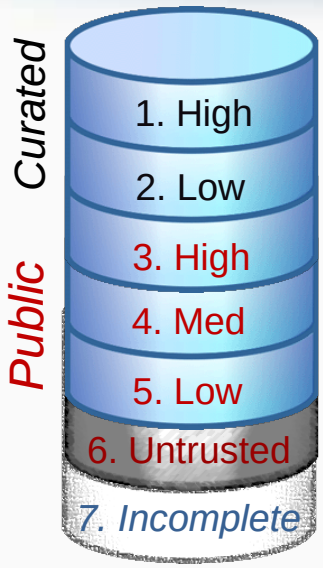
https://comptox.epa.gov/dashboard

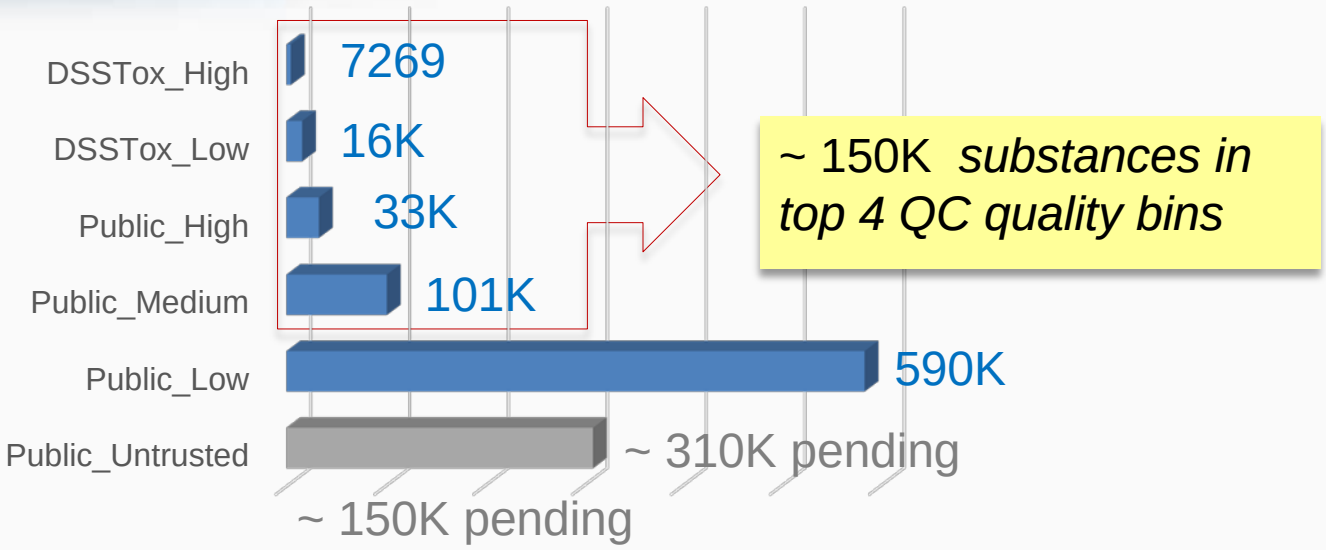


Underneath the Dashboard



Distribution of curated data

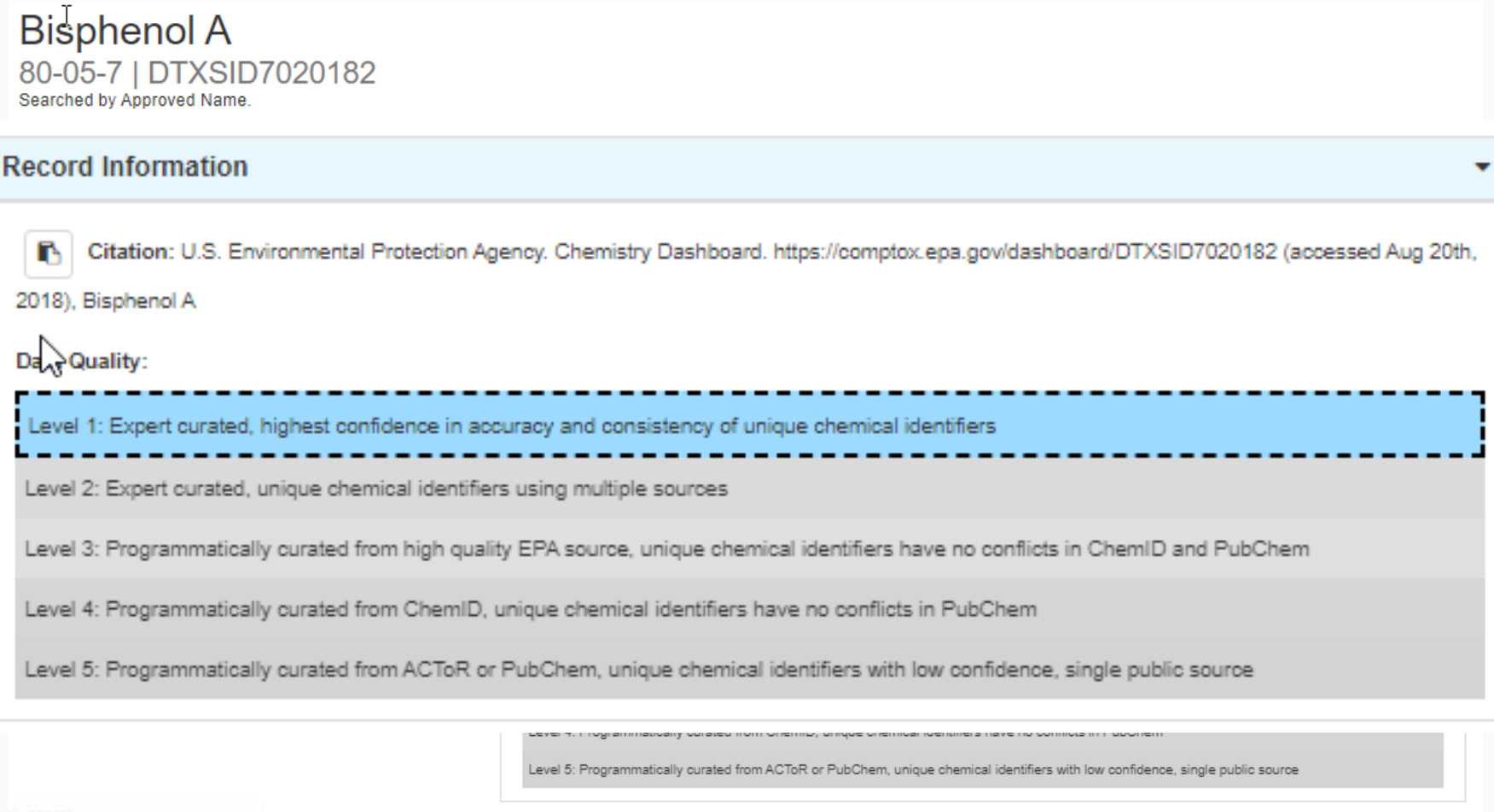




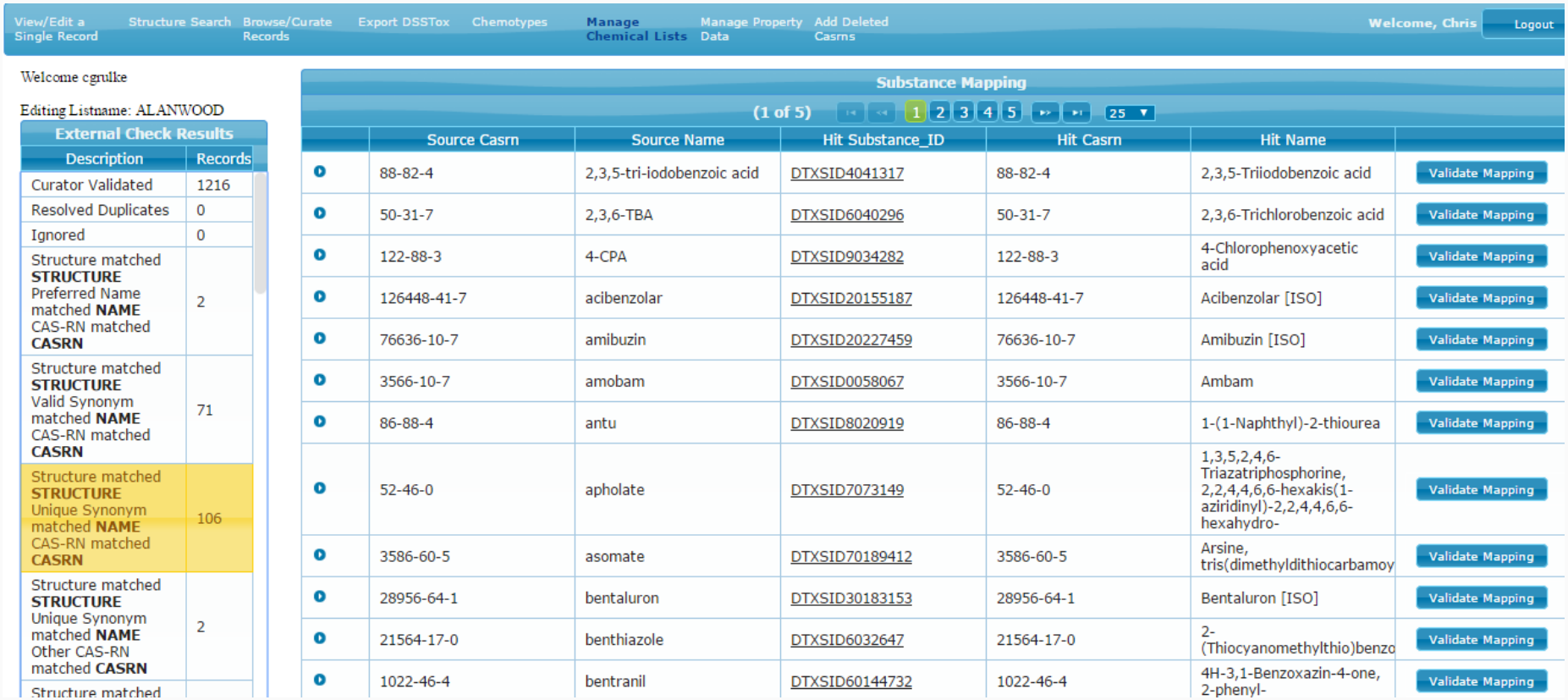
QC Levels

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

Record Information Quality Flags



ChemReg Curation
The ALANWOOD Pesticide Set



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

Welcome cgrulke

Editing Listname: ALANWOOD

External Check Results

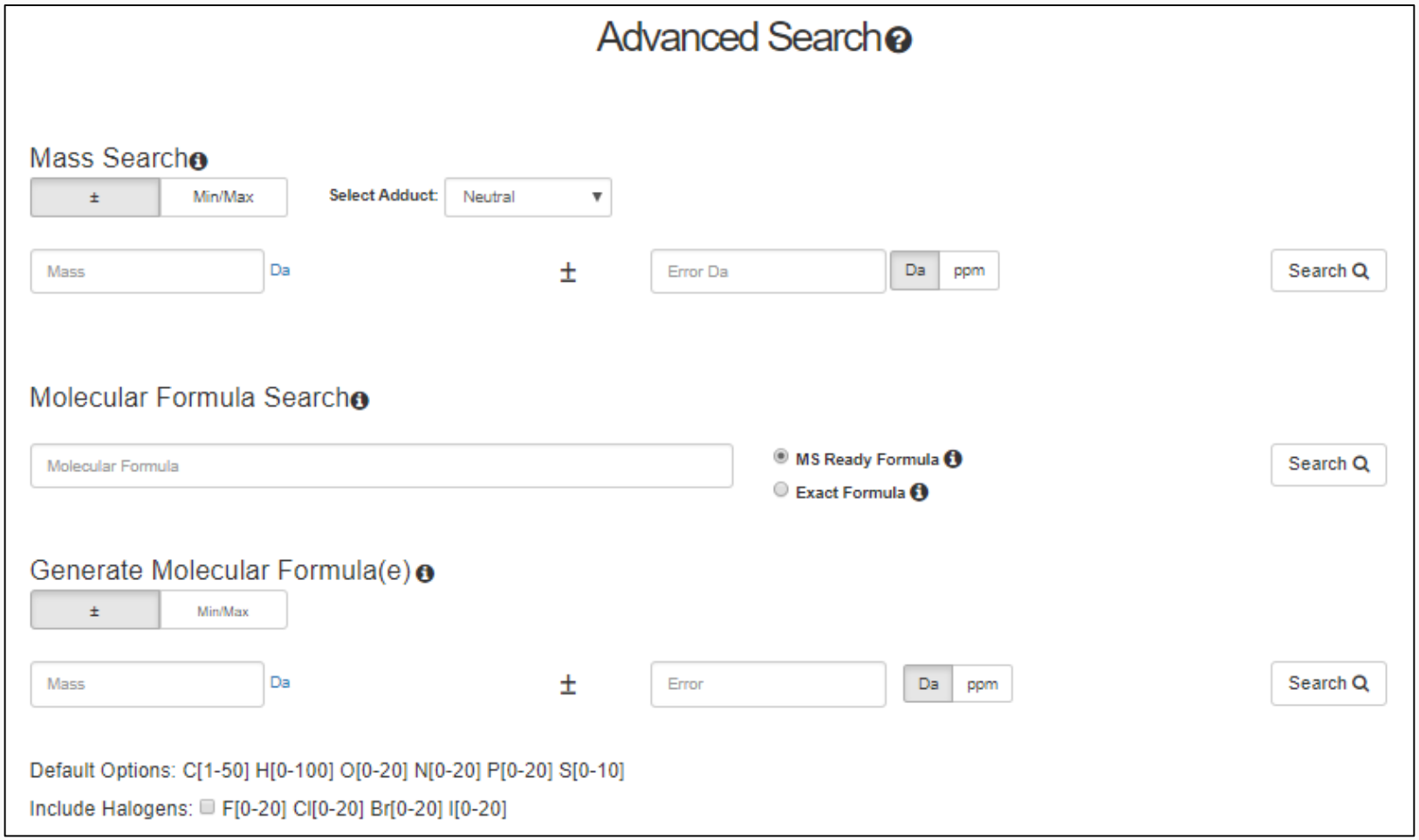
Description	Records
Curator Validated	1216
Resolved Duplicates	0
Ignored	0
Structure matched **STRUCTURE** Preferred Name matched **NAME** CAS-RN matched **CASRN**	2
Structure matched **STRUCTURE** Valid Synonym matched **NAME** CAS-RN matched **CASRN**	71
Structure matched **STRUCTURE** Unique Synonym matched **NAME** CAS-RN matched **CASRN**	106
Structure matched **STRUCTURE** Unique Synonym matched **NAME** Other CAS-RN matched **CASRN**	2
Structure matched	

Substance Mapping

(1 of 5) 1 2 3 4 5 25

Source Casrn	Source Name	Hit Substance_ID	Hit Casrn	Hit Name	
88-82-4	2,3,5-tri-iodobenzoic acid	DTXSID4041317	88-82-4	2,3,5-Triiodobenzoic acid	Validate Mapping
50-31-7	2,3,6-TBA	DTXSID6040296	50-31-7	2,3,6-Trichlorobenzoic acid	Validate Mapping
122-88-3	4-CPA	DTXSID9034282	122-88-3	4-Chlorophenoxyacetic acid	Validate Mapping
126448-41-7	acibenzolar	DTXSID20155187	126448-41-7	Acibenzolar [ISO]	Validate Mapping
76636-10-7	amibuzin	DTXSID20227459	76636-10-7	Amibuzin [ISO]	Validate Mapping
3566-10-7	amobam	DTXSID0058067	3566-10-7	Ambam	Validate Mapping
86-88-4	antu	DTXSID8020919	86-88-4	1-(1-Naphthyl)-2-thiourea	Validate Mapping
52-46-0	apholate	DTXSID7073149	52-46-0	1,3,5,2,4,6-Triazatriphosphorine, 2,2,4,4,6,6-hexakis(1-aziridinyl)-2,2,4,4,6,6-hexahydro-	Validate Mapping
3586-60-5	asomate	DTXSID70189412	3586-60-5	Arsine, tris(dimethyldithiocarbamoy	Validate Mapping
28956-64-1	bentaluron	DTXSID30183153	28956-64-1	Bentaluron [ISO]	Validate Mapping
21564-17-0	benthiazole	DTXSID6032647	21564-17-0	2-(Thiocyanomethylthio)benzo	Validate Mapping
1022-46-4	bentranil	DTXSID60144732	1022-46-4	4H-3,1-Benzoxazin-4-one, 2-phenyl-	Validate Mapping

The CompTox Dashboard for Structure Identification by MS



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
- It has highlighted data quality issues – and these are not unexpected!

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.

mol_ID	Name	CAS_RN	SMILES
StNorman2	Microcystin-RR	CAS_RN: 111755-37-4	C[C@H]1[C@@H](NC(=O)[C@@H](NC(=O)[C@
StNorman3	Microcystin-YR	CAS_RN: 101064-48-6	C[C@H]1[C@@H](NC(=O)[C@@H](NC(=O)[C@
StNorman5	2,6-Di-tert-butylphenol	CAS_RN: 128-39-2	CC(C)(C)c1cccc(c1O)C(C)(C)C
StNorman6	Butylated hydroxyanisole	CAS_RN: 25013-16-5	COc1ccc(O)c(c1)C(C)(C)C
StNorman7	tert-Butylhydroquinone	CAS_RN: 1948-33-0	CC(C)(C)c1cc(O)ccc1O
StNorman8	Butylated hydroxytoluene	CAS_RN: 128-37-0	Cc1cc(c(O)c(c1)C(C)(C)C)C(C)(C)C
StNorman15	Diethylenetriaminepentaace	CAS_RN: 67-43-6	OC(=O)CN(CCN(CC(O)=O)CC(O)=O)CCN(CC(O)=
StNorman16	Ethylenediaminetetraacetic	CAS_RN: 60-00-4	OC(=O)CN(CCN(CC(O)=O)CC(O)=O)CC(O)=O
StNorman17	Nitrilotriacetic acid	CAS_RN: 139-13-9	OC(=O)CN(CC(O)=O)CC(O)=O
StNorman18	Oxadixyl	CAS_RN: 77732-09-3	COCC(=O)N(N1CCOC1=O)c2c(C)cccc2C
StNorman19	Tetraacetylethylenediamine	CAS_RN: 10543-57-4	CC(=O)N(CCN(C(C)=O)C(C)=O)C(C)=O
StNorman26	2-(2-(4-Nonylphenoxy)etho	CAS_RN: 106807-78-7	CCCCCCCCCc1ccc(OCCOCC(O)=O)cc1
StNorman29	4-Octylphenoxy acetic acid	CAS_RN: 15234-85-2	CCCCCCCCc1ccc(OCC(O)=O)cc1
StNorman30	Cyanoformaldehyde	CAS_RN: 4471-47-0	NCC=O
StNorman31	Decabromodiphenyl ethane	CAS_RN: 84852-53-9	C(CC1=C(C(=C(C(=C1Br)Br)Br)Br)Br)C2=C(C(=C
StNorman32	Hexabromocyclododecane	CAS_RN: 25637-99-4	BrC1CC(Br)CC(Br)CC(Br)CC(Br)CC(Br)C1
StNorman33	n-Nitrosodimethylamine	CAS_RN: 62-75-9	CN(C)N=O
StNorman34	Benzylbutylphthalate	CAS_RN: 85-68-7	CCCCOC(=O)c1ccccc1C(=O)OCc2ccccc2
StNorman35	Diethyl phthalate	CAS_RN: 84-66-2	CCOC(=O)c1ccccc1C(=O)OCC
StNorman37	Di-n-butylphthalate	CAS_RN: 84-74-2	CCCCOC(=O)c1ccccc1C(=O)OCCCC
StNorman39	Bisphenol A	CAS_RN: 80-05-7	CC(C)(c1ccc(O)cc1)c2ccc(O)cc2

Mapping on Two Identifiers
Lookup Based on Name

EPA United States Environmental Protection Agency

LOOKUP BASED ON CHEMICAL NAME							
DTXSID	PREFERRED NAME	CASRN	INCHI KEY	IUPAC NA	SMILES	MOL FORMULA	MONOISC
DTXSID40880085	Microcystin RR	111755-37-4	JIGDOBKZI	(5R,8S,11R	CO[C@@H	C49H75N13O12	1037.5658
DTXSID00880086	Microcystin YR	101064-48-6	OWHASZC	(5R,8S,11R	CO[C@@H	C52H72N10O13	1044.5280
DTXSID6027052	2,6-Di-tert-butylphen	128-39-2	DKCPKDPY	2,6-Di-tert	CC(C)(C)C1	C14H22O	206.16706
DTXSID7020215	Butylated hydroxyani	25013-16-5	CZBZUDVE	2-tert-But	COC1=CC=	C22H32O4	360.23005
DTXSID6020220	tert-Butylhydroquino	1948-33-0	BGNXCDM	2-tert-But	CC(C)(C)C1	C10H14O2	166.09937
DTXSID2020216	Butylated hydroxytol	128-37-0	NLZUEZXR	2,6-Di-tert	CC1=CC(=C	C15H24O	220.18271
DTXSID2023434	Pentetic acid	67-43-6	QPCDCPDI	N,N-Bis{2-	OC(=O)CN	C14H23N3O10	393.13834
DTXSID6022977	Ethylenediaminetetra	60-00-4	KCXVZYZYI	2,2',2'',2'''-	OC(=O)CN	C10H16N2O8	292.09066
DTXSID6020939	Nitrilotriacetic acid	139-13-9	MGFYIUFZ	2,2',2''-Nit	OC(=O)CN	C6H9NO6	191.04298
DTXSID2032631	Oxadixyl	77732-09-3	UWVQIRO	N-(2,6-Din	COCC(=O)	C14H18N2O4	278.12665
DTXSID5040752	Tetraacetylethylened	10543-57-4	BGRWYDH	N,N'-(Etha	CC(=O)N(C	C10H16N2O4	228.11100
DTXSID70147779	2-(2-(4-Nonylphenox	106807-78-7	RAQHOBR	[2-(4-Nony	CCCCCCCC	C19H30O4	322.21440
DTXSID60165003	Acetic acid, (4-octylph	15234-85-2	DWUYSEN	(4-Octylph	CCCCCCCC	C16H24O3	264.17254
DTXSID0021549	Cyanoformaldehyde	4471-47-0	TUHMQDC	Oxoacetor	O=CC#N	C2HNO	55.005813
DTXSID2052732	1,1'-Ethane-1,2-diylbi	84852-53-9	BZQKBFHE	1,1'-(Ethar	BrC1=C(Br	C14H4Br10	961.21468
DTXSID8025383	Hexabromocyclodod	25637-99-4		-	-	-	-
DTXSID7021029	N-Nitrosodimethylan	62-75-9	UMFJAHH	N,N-Dimet	CN(C)N=O	C2H6N2O	74.048012
DTXSID3020205	Benzyl butyl phthalat	85-68-7	IRIAEXORF	Benzyl but	CCCCOC(=	C19H20O4	312.13615
DTXSID7021780	Diethyl phthalate	84-66-2	FLKPEMZC	Diethyl be	CCOC(=O)	C12H14O4	222.08920
DTXSID2021781	Dibutyl phthalate	84-74-2	DOIRQSBF	Dibutyl be	CCCCOC(=	C16H22O4	278.15180
DTXSID7020182	Bisphenol A	80-05-7	IISBACLAF	4,4'-(Propa	CC(C)(C1=	C15H16O2	228.11502

Mapping on Two Identifiers
Lookup Based on CASRN

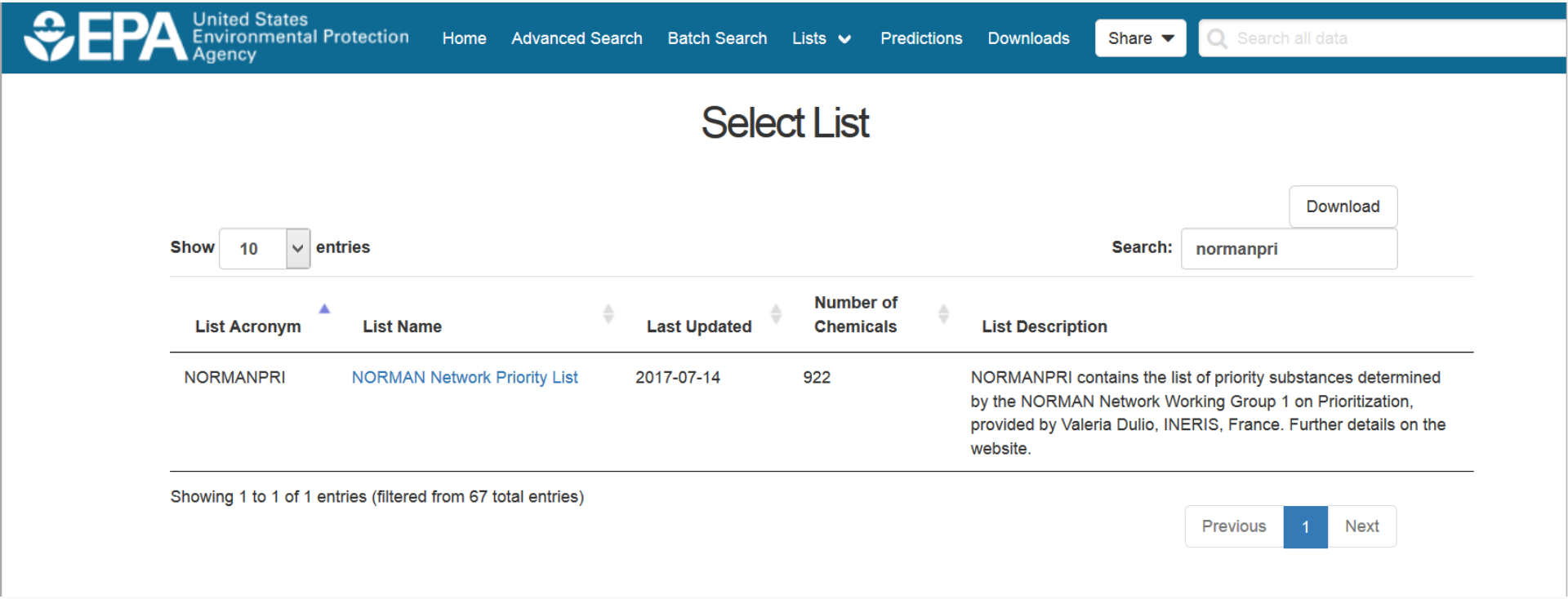
			LOOKUP BASED ON CASRN				
DTXSID	PREFERRED NAME	CASRN	INCHI KEY	IUPAC NA	SMILES	MOL FORMULA	MONOISC
DTXSID40880085	Microcystin RR	111755-37-4	JIGDOBKZI	(5R,8S,11R	CO[C@@H	C49H75N13O12	1037.5658
DTXSID00880086	Microcystin YR	101064-48-6	OWHASZC	(5R,8S,11R	CO[C@@H	C52H72N10O13	1044.5280
DTXSID6027052	2,6-Di-tert-butylpheno	128-39-2	DKCPKDPY	2,6-Di-tert	CC(C)(C)C1	C14H22O	206.16706
DTXSID7020215	Butylated hydroxyanis	25013-16-5	CZBZUDVE	2-tert-But	COC1=CC=	C22H32O4	360.23005
DTXSID6020220	tert-Butylhydroquinon	1948-33-0	BGNXCDM	2-tert-But	CC(C)(C)C1	C10H14O2	166.09937
DTXSID2020216	Butylated hydroxytolue	128-37-0	NLZUEZXR	2,6-Di-tert	CC1=CC(=C	C15H24O	220.18271
DTXSID2023434	Pentetic acid	67-43-6	QPCDCPDI	N,N-Bis{2-	OC(=O)CN	C14H23N3O10	393.13834
DTXSID6022977	Ethylenediaminetetraa	60-00-4	KCXVZYZYI	2,2',2'',2'''-	OC(=O)CN	C10H16N2O8	292.09066
DTXSID6020939	Nitrilotriacetic acid	139-13-9	MGFYIUFZ	2,2',2''-Nit	OC(=O)CN	C6H9NO6	191.04298
DTXSID2032631	Oxadixyl	77732-09-3	UWVQIRO	N-(2,6-Din	COCC(=O)	C14H18N2O4	278.12665
DTXSID5040752	Tetraacetylethylenedia	10543-57-4	BGRWYDH	N,N'-(Etha	CC(=O)N(C	C10H16N2O4	228.11100
DTXSID70147779	2-(2-(4-Nonylphenoxy)	106807-78-7	RAQHOBR	[2-(4-Non	CCCCCCCC	C19H30O4	322.21440
DTXSID60165003	Acetic acid, (4-octylphe	15234-85-2	DWUYSEM	(4-Octylph	CCCCCCCC	C16H24O3	264.17254
DTXSID0021549	Cyanoformaldehyde	4471-47-0	TUHMQDC	Oxoacetor	O=CC#N	C2HNO	55.005813
DTXSID2052732	1,1'-Ethane-1,2-diylbis(	84852-53-9	BZQKBFHE	1,1'-(Ethar	BrC1=C(Br	C14H4Br10	961.21468
DTXSID8025383	Hexabromocyclododec	25637-99-4		-	-	-	-
DTXSID7021029	N-Nitrosodimethylamir	62-75-9	UMFJAHH	N,N-Dimet	CN(C)N=O	C2H6N2O	74.048012
DTXSID3020205	Benzyl butyl phthalate	85-68-7	IRIAEXORF	Benzyl but	CCCCOC(=	C19H20O4	312.13615
DTXSID7021780	Diethyl phthalate	84-66-2	FLKPEMZC	Diethyl be	CCOC(=O)	C12H14O4	222.08920
DTXSID2021781	Dibutyl phthalate	84-74-2	DOIRQSBF	Dibutyl be	CCCCOC(=	C16H22O4	278.15180
DTXSID7020182	Bisphenol A	80-05-7	IISBACLAF	4,4'-(Propa	CC(C)(C1=	C15H16O2	228.11502

Mapping Quality Control

DTXSID_Name	DTXSID_CAS	DTXSID	DTXSID Equal	Name InChIKey	CAS InChIKey	Name Block1	CAS Block1	Name WCAS	Name WName	KeyOR Name	CAS Match	Num NoMatch
DTXSID2047064	DTXSID0020151	DTXSID0020151	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	0
DTXSID7042481	DTXSID2032683	DTXSID2032683	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID5022853	DTXSID60110018	DTXSID60110018	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID3023215	DTXSID7047803	DTXSID7047803	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID1045033	DTXSID8044545	DTXSID8044545	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID3041083	DTXSID9023386	DTXSID9023386	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID8021480	DTXSID3030636	DTXSID3030636	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID80185521	DTXSID90205325	DTXSID90205325	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID8022826	DTXSID40872344	DTXSID40872344	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID4021509	DTXSID7061277	DTXSID7061277	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	0
DTXSID20213251	DTXSID5060936	DTXSID5060936	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	0
DTXSID3021803	DTXSID20220667	DTXSID20220667	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	0
DTXSID80109343	DTXSID1034715	DTXSID1034715	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	0
DTXSID2023359	DTXSID9041073	DTXSID9041073	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	0
DTXSID6029094	DTXSID8073471	DTXSID8073471	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	0
DTXSID30860093	DTXSID7025219	DTXSID7025219	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0
DTXSID1047524	DTXSID2041171	DTXSID2041171	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0
DTXSID7041099	DTXSID2035726	-	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0
-	DTXSID7027041	DTXSID7027041	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	1
-	DTXSID3032416	DTXSID3032416	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	1

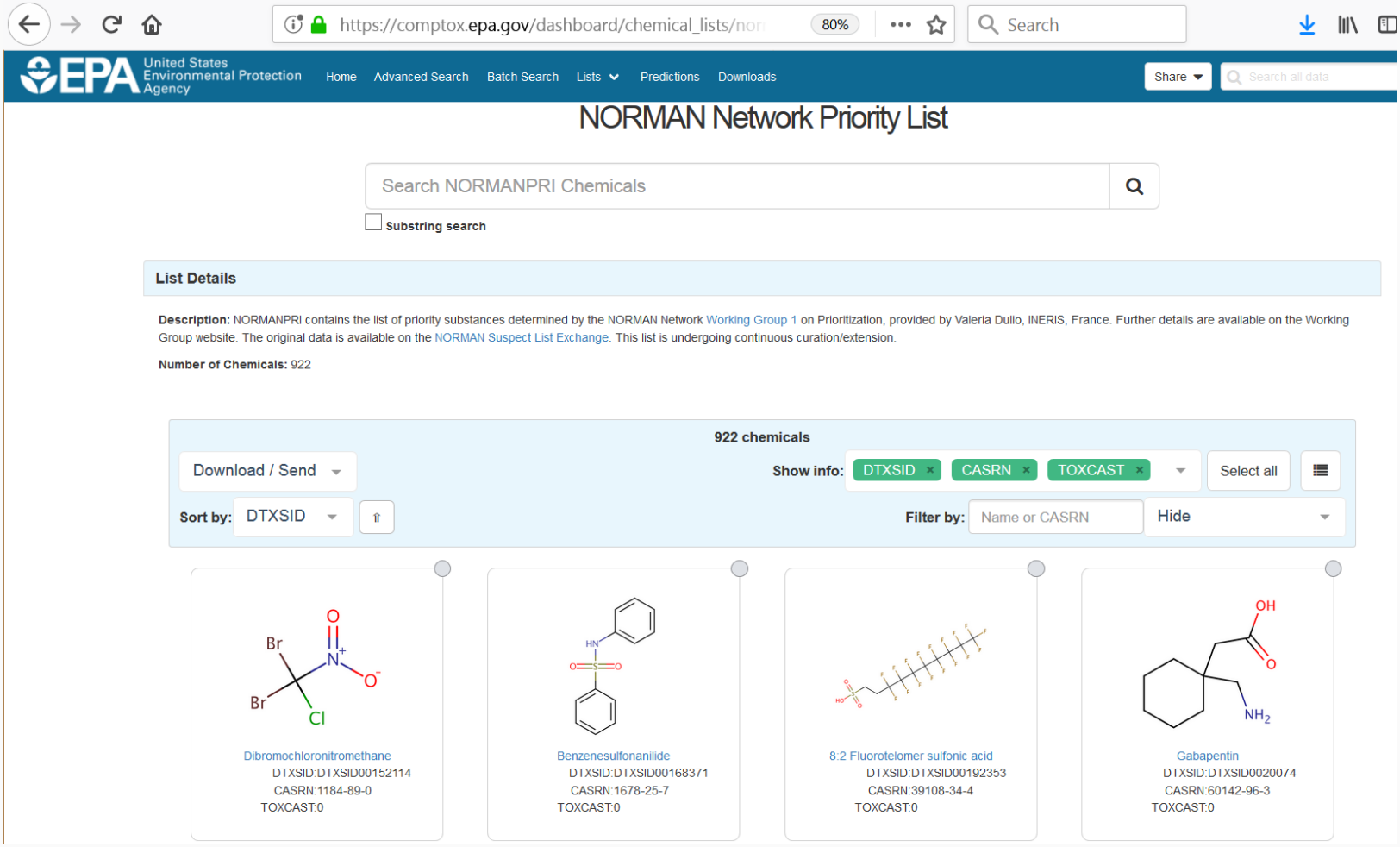
Example: NORMAN Priority List

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.



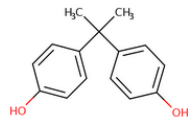
Example: NORMAN Priority List

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.

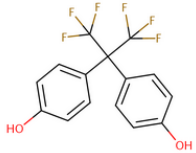


>23 NORMAN Lists Available

EPA United States Environmental Protection Agency

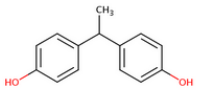


Bisphenol A
DTXSID:DTXSID7020182
CASRN:80-05-7
TOXCAST:166/812

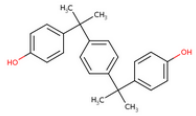


Bisphenol AF
DTXSID:DTXSID7037717
CASRN:1478-61-1
TOXCAST:248/639

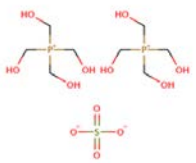
Bisphenol Compounds



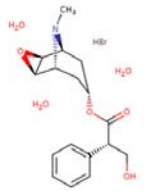
Bisphenol E
DTXSID:DTXSID3047891
CASRN:2081-08-5
TOXCAST:33/276



Bisphenol P
DTXSID:DTXSID0058693
CASRN:2167-51-3
TOXCAST:0

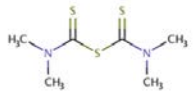


Tetrakis(hydroxymethyl)phosphonium su...
DTXSID:DTXSID0021331
CASRN:55566-30-8

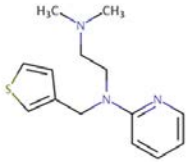


Scopolamine hydrobromide trihydrate
DTXSID:DTXSID0021258
CASRN:6533-68-2

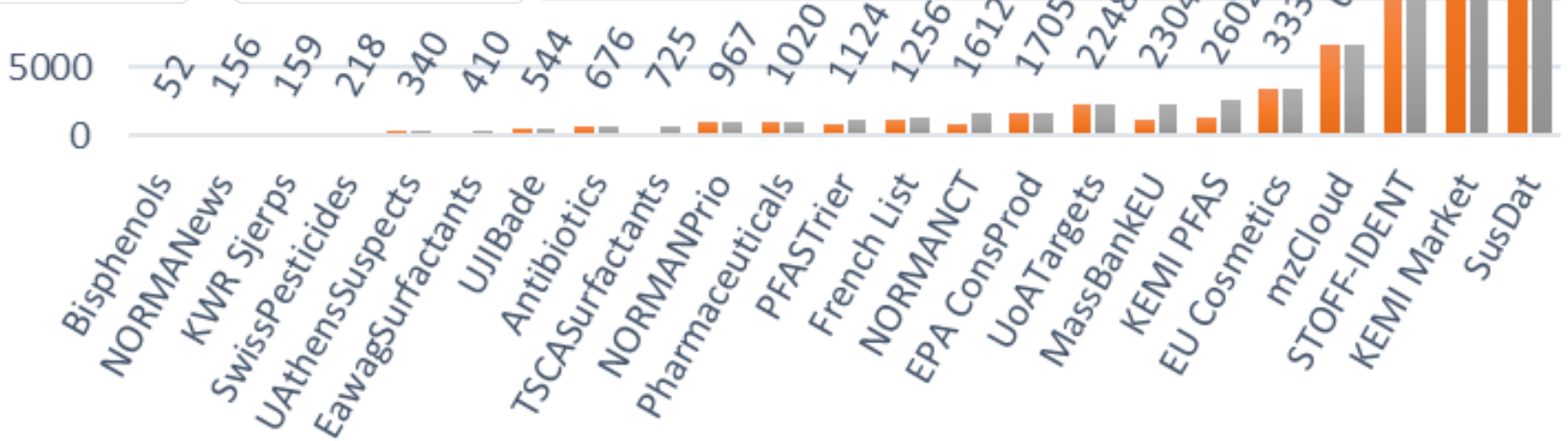
KEMI List of Substances on the Market



Tetramethylthiuram monosulfide
DTXSID:DTXSID0021333
CASRN:97-74-5
TOXCAST:74/453



Thenyldiamine
DTXSID:DTXSID0021335
CASRN:91-79-2
TOXCAST:0



Aggregating Data

- We aggregate data from various sources
- Downloading is “easy”
- Databasing is “easy”
- Data curation is a laborious and painful task

Online Today

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:**

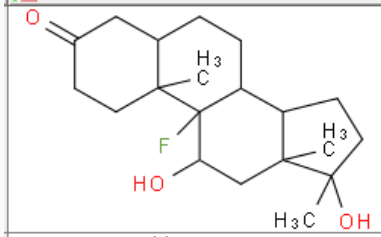
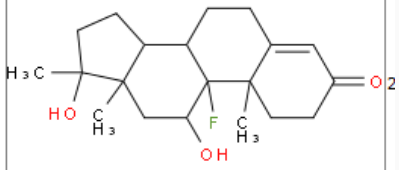
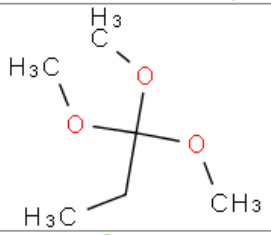
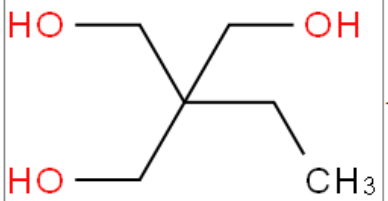
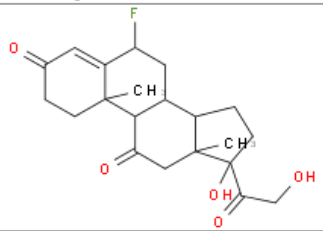
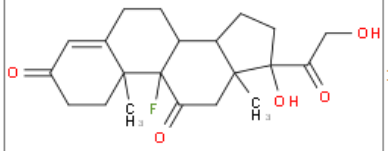
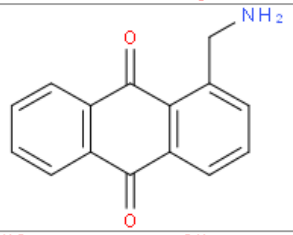
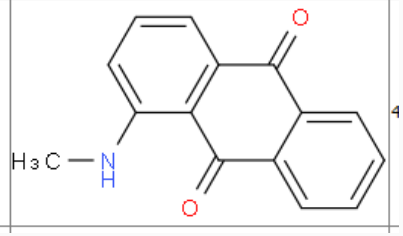
http://esc.syrres.com/interkow/EpiSuiteData.htm

We Curated These Public Data to Build Prediction Models

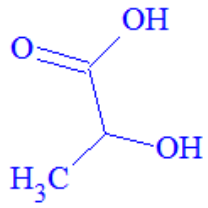
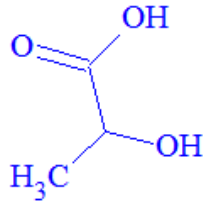
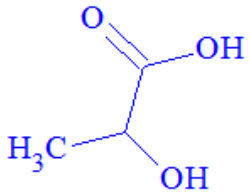
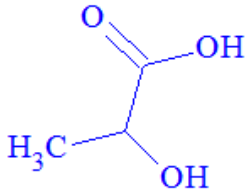
EPA United States Environmental Protection Agency

Public data should be curated prior to modeling

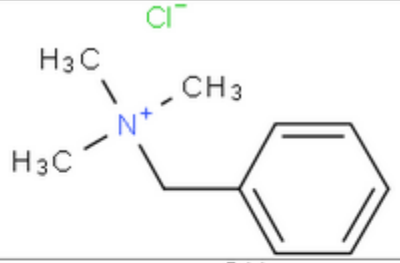
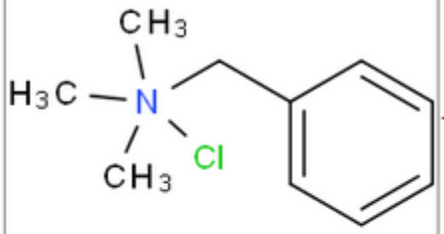
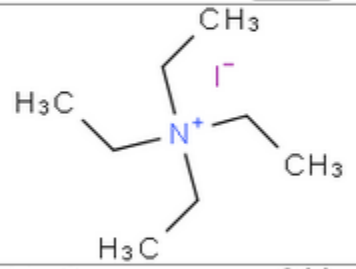
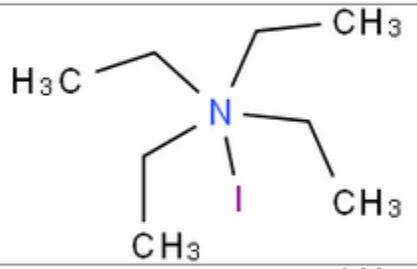
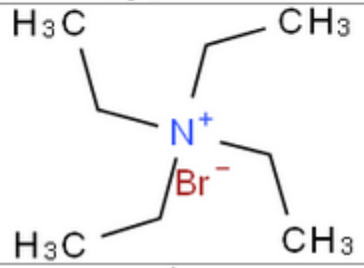
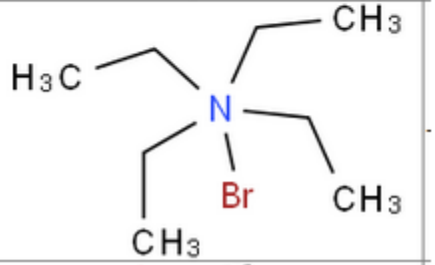
Different Compounds

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%)

Workflow Details and Data

Journal
SAR and QSAR in Environmental Research ›
Volume 27, 2016 - Issue 11: 17th International Conference on QSAR in Environmental and Health Sciences (QSAR 2016) - Part II. Guest Editors: C.G. Barber and G.J. Myatt

Enter keywords, authors, DOI etc.

258
Views
4
CrossRef citations
16
Altmetric

Articles

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

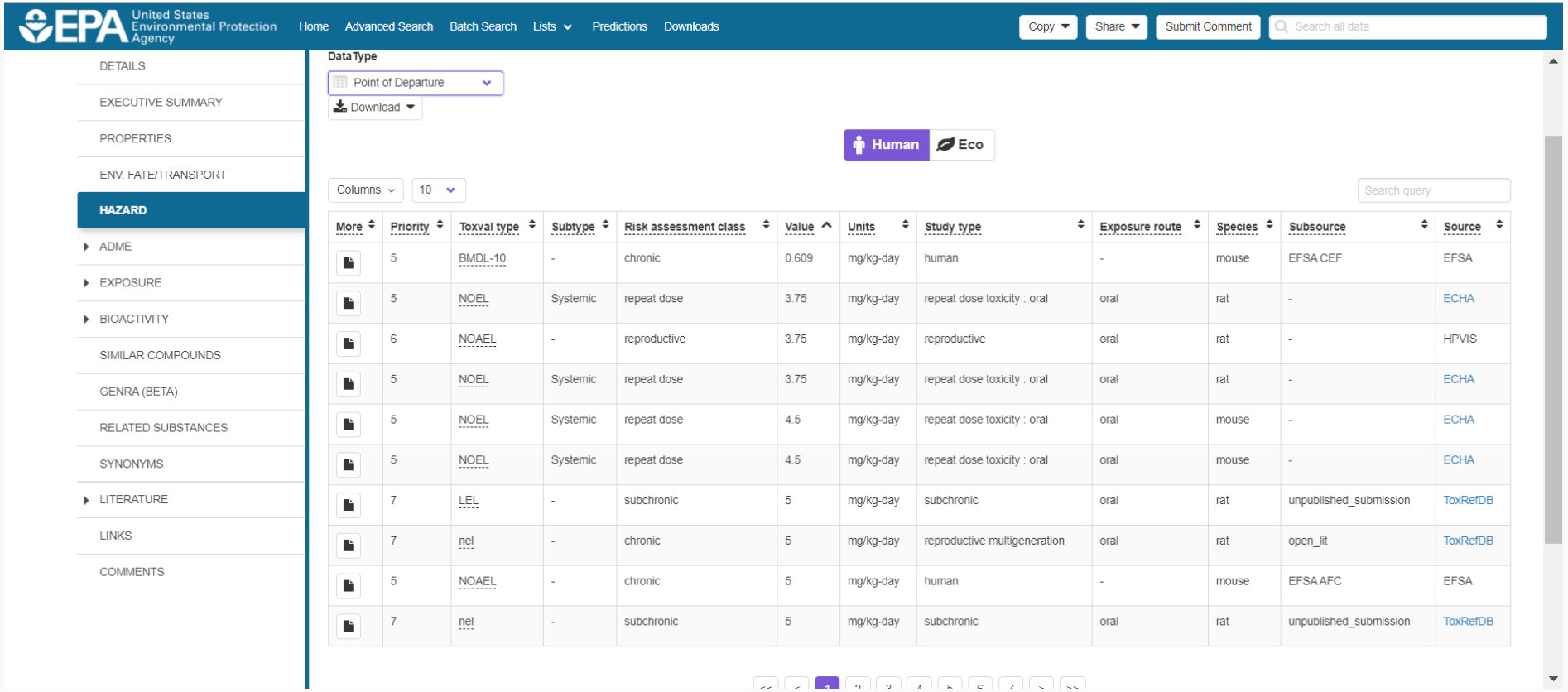
K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

Pages 911-937 | Received 03 Sep 2016, Accepted 24 Oct 2016, Published online: 25 Nov 2016

Download citation http://dx.doi.org/10.1080/1062936X.2016.1253611 Check for updates

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

Sourcing Chemical Hazard Data



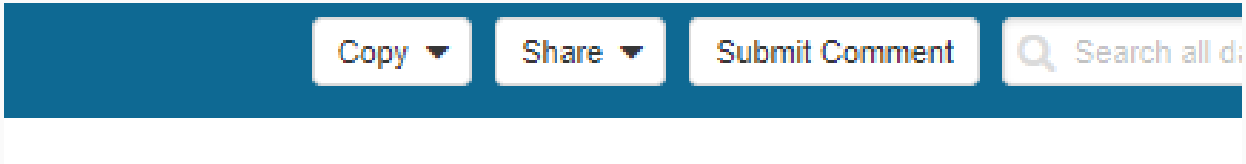
More	Priority	Toxval type	Subtype	Risk assessment class	Value	Units	Study type	Exposure route	Species	Subsource	Source
	5	BMDL-10	-	chronic	0.609	mg/kg-day	human	-	mouse	EFSA CEF	EFSA
	5	NOEL	Systemic	repeat dose	3.75	mg/kg-day	repeat dose toxicity : oral	oral	rat	-	ECHA
	6	NOAEL	-	reproductive	3.75	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

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

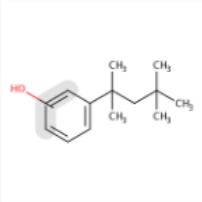
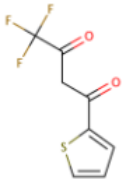
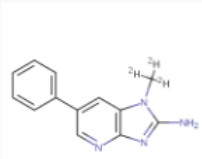
https://comptox.epa.gov/dashboard/comments/public_index

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	★

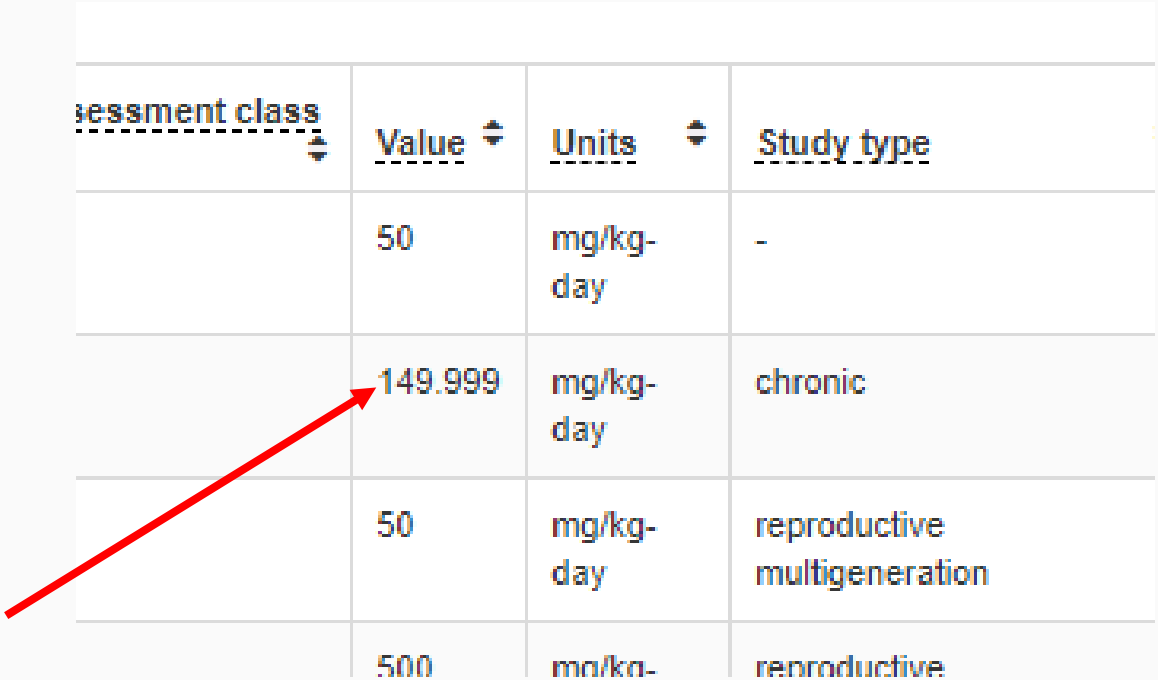
Reviewer comments are public

Comments to date

- The majority of comments to date:
  - Structure and names/CASRN do not match
  - Add additional synonyms
  - Request to add specific property data
  - Structure layout/depiction needs improving

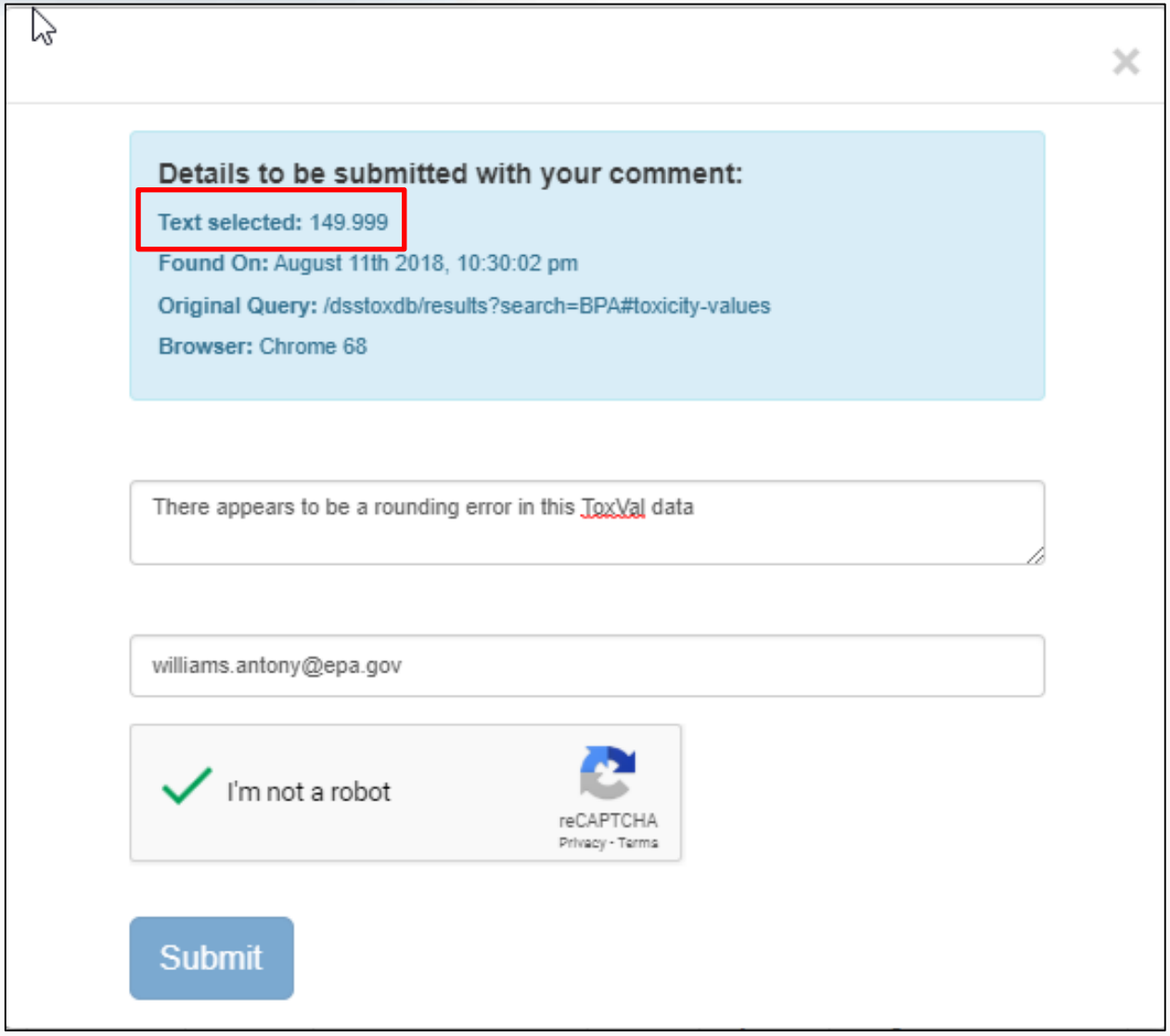
Crowdsourcing Comments
Single Cell Commenting added

- Highlight an alphanumeric text string



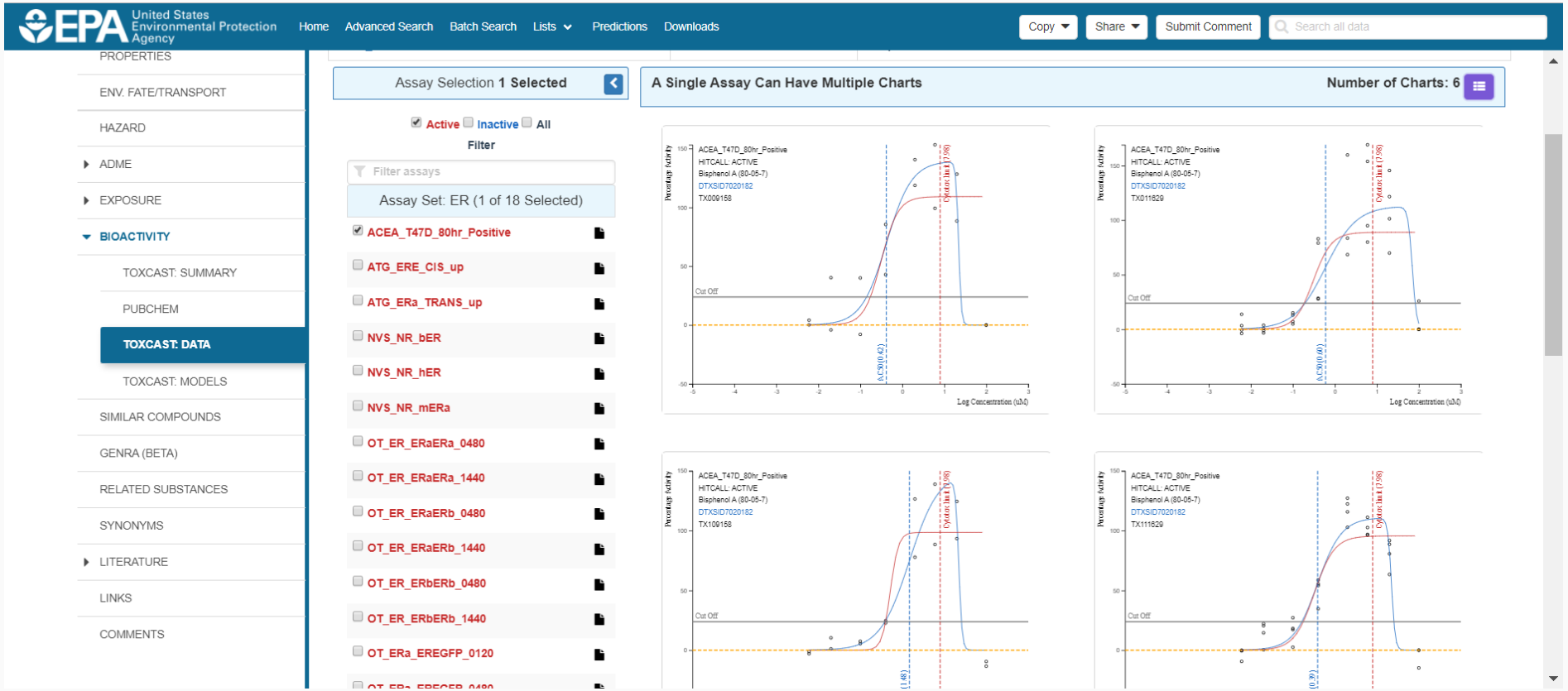
sessment class	Value	Units	Study type
	50	mg/kg-day	-
	149.999	mg/kg-day	chronic
	50	mg/kg-day	reproductive multigeneration
	500	mg/kg-	reproductive

Crowdsourcing Comments



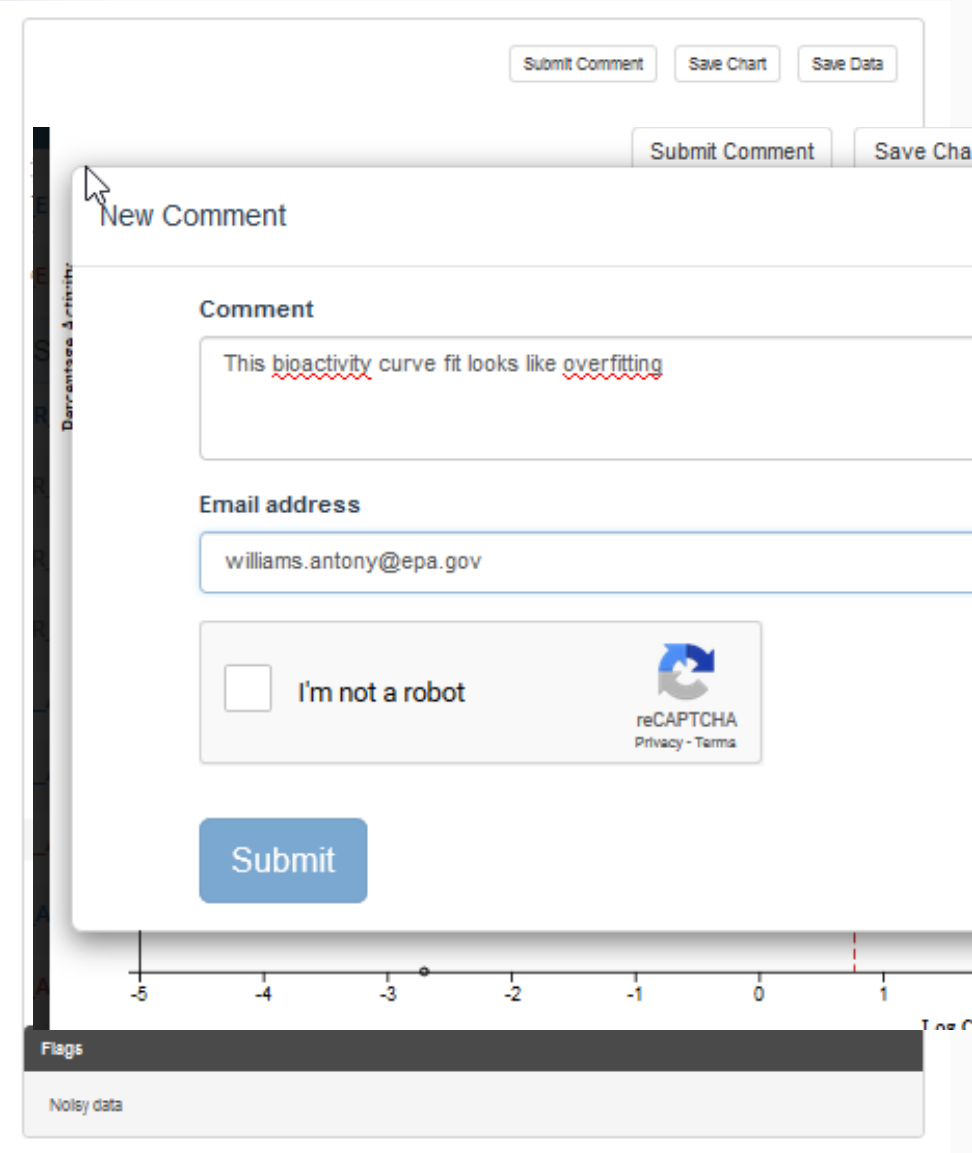
In Vitro Bioassay Screening

ToxCast and Tox21

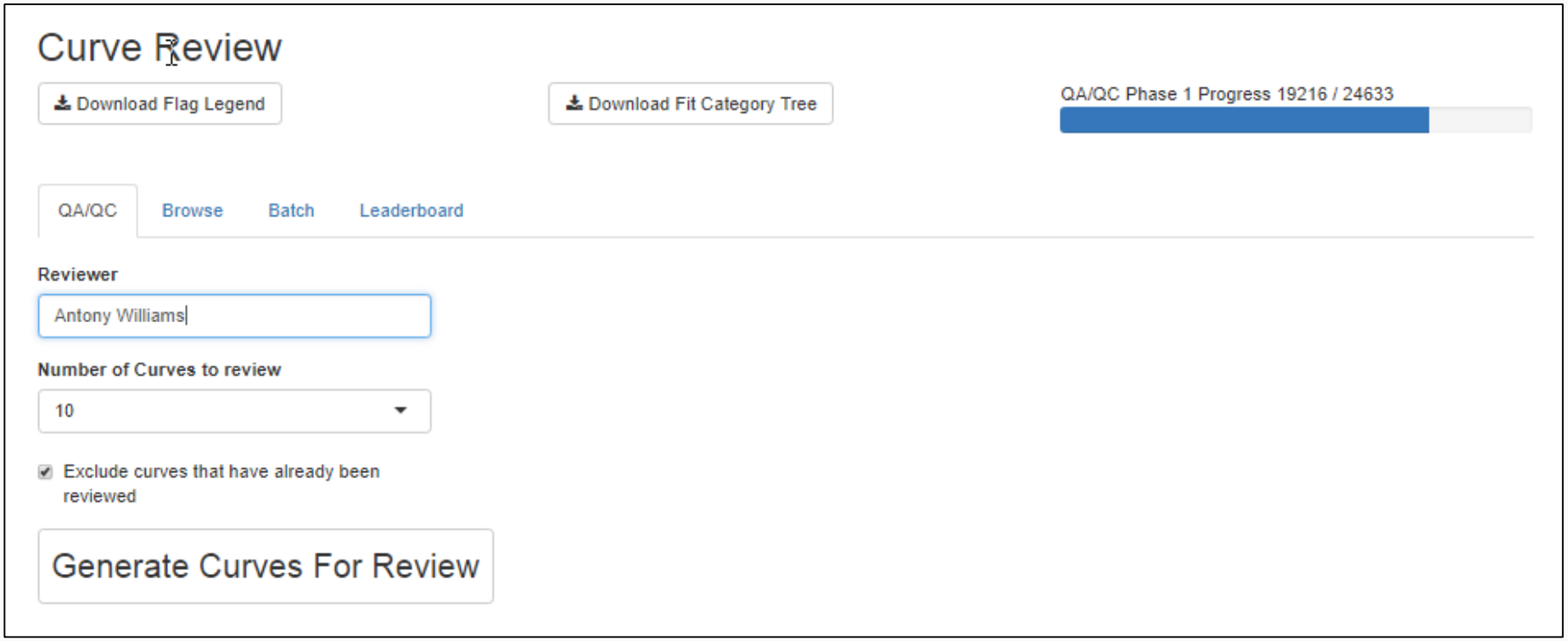


Bioactivity Data

- 100s of thousands of bioactivity curves to review
- Impossible to review every one manually
- Now accepting public Crowdsourced Comments
- **Public crowdsourcing will not suffice!!!**

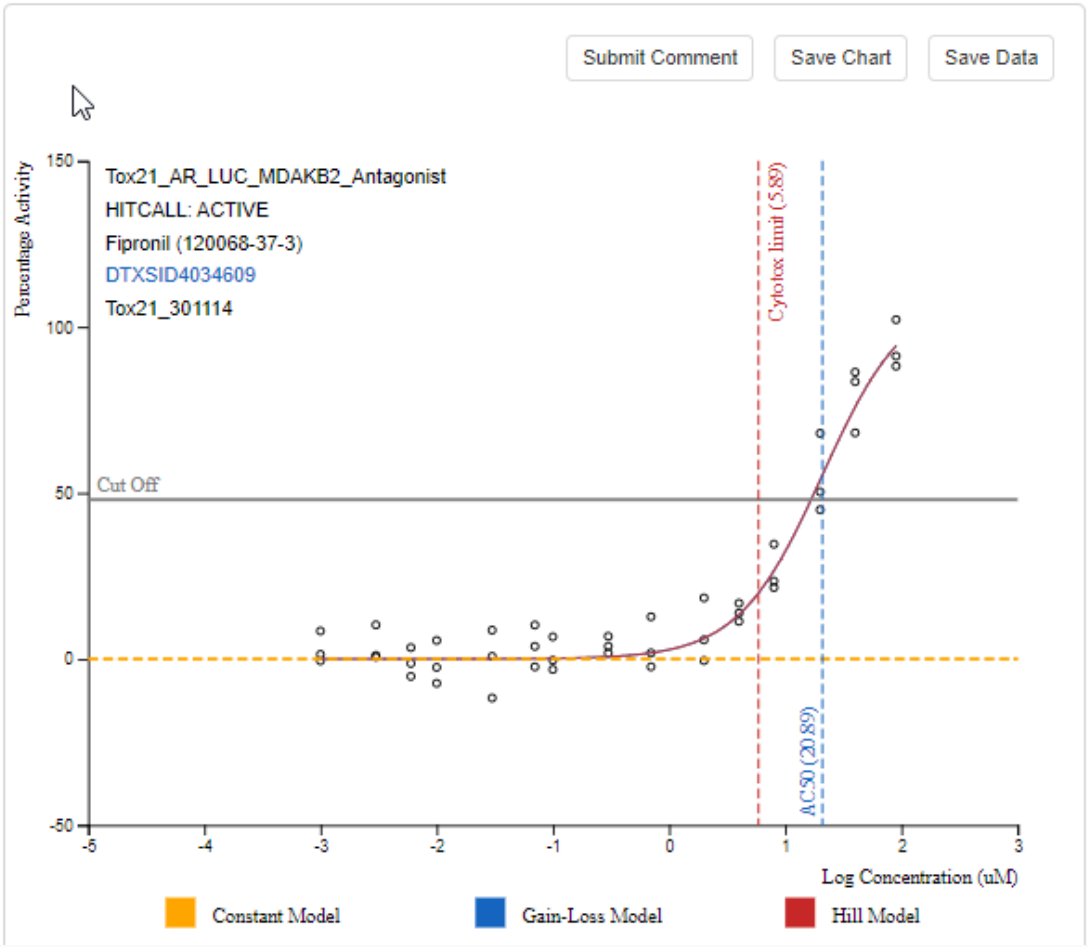


Internal Review of 25,000 curves



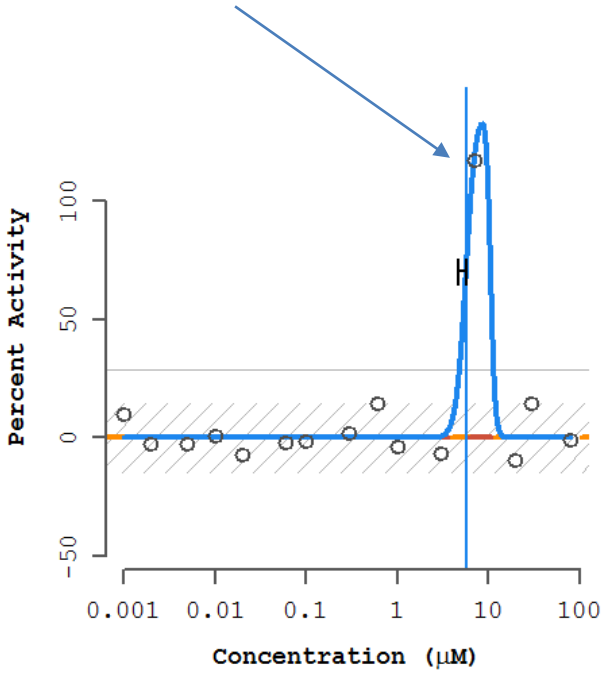
Screenshot of entry page for Beta R Shiny Application for NCCT users
Brown & Paul-Friedman

Internal Review of 25,000 curves
A “good fit” bioactivity curve



Internal Review of 25,000 curves

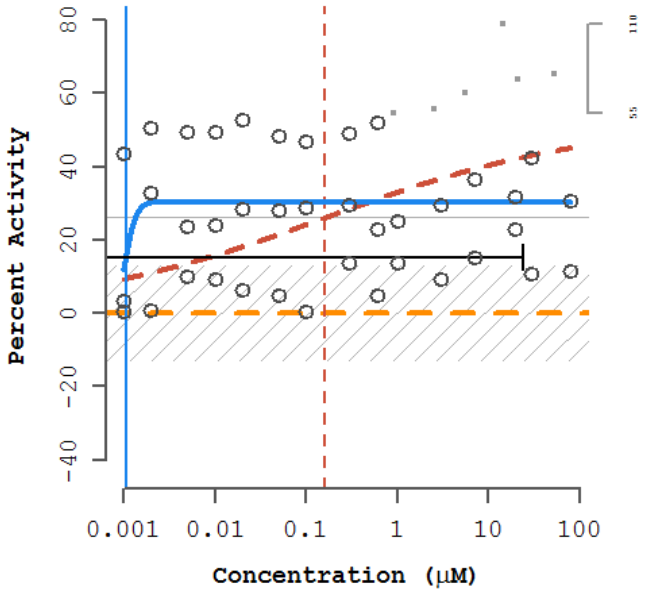
Single-Point in middle of concentration range drives ACTIVE Hit Call




```
ASSAY:    TOX21_RT_HEK293_FLO_40hr_viability

NAME:     Tributylmethylammonium bis(trifluoromethylsulfc
CHID:     47905    CASRN: 405514-94-5
SPID(S): Tox21_200582
M4ID:     19917619

HILL MODEL (in red):
       tp        ga        gw
val:   104       2.4       7.97
sd:    NaN       NaN       NaN

GAIN-LOSS MODEL (in blue):
       tp        ga        gw        la        lw
val:   140       0.758     8         1.03      18
sd:    56.9      0.105     3.92      0.822     64.2

       CNST         HILL         GNLS
AIC:   125.38       131.38       113.6
PROB:  0            0            1
RMSE:  31.04        31.04        7.1

MAX_MEAN: 117        MAX_MED: 117        BMAD: 4.76

COFF: 28.5    HIT-CALL: 1     FITC: 50    ACTP: 1

FLAGS:

HIT-PCT: 0.887  MED-GA: 0.7582  GA-CI: 0.0826
```

…and gain-loss fit with a cell viability assay (makes little sense)

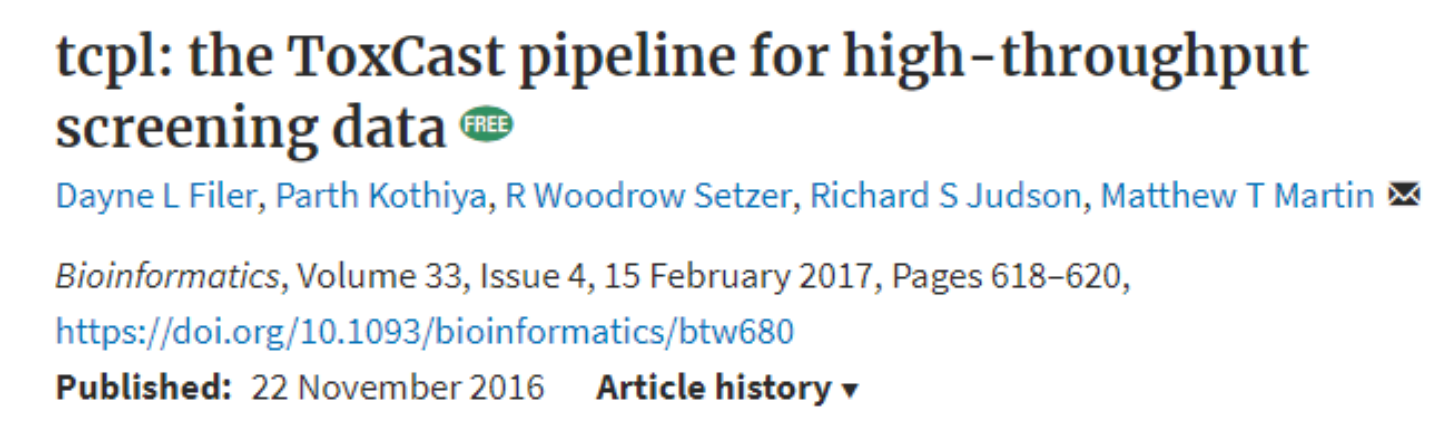
Internal Review of 25,000 curves Abnormally High-Noise




```
ASSAY:    TOX21_AR_BLA_Antagonist_ratio

NAME:     1-Nitropyrene
CHID:     20983    CASRN: 5522-43-0
SPID(S): Tox21_200066
M4ID:     18181279  BRK

HILL MODEL (in red):
        tp        ga        gw
val:    52        -0.797    0.3
sd:     NaN       NaN       NaN

GAIN-LOSS MODEL (in blue):
        tp        ga        gw        la        lw
val:    30.4      -2.97     7.76      3.77      4.44
sd:     3.58      0.211     59.3      7180      16800

        CNST          HILL          GNLS
AIC:    464.32        423.78        423.68
PROB:   0             0.49          0.51
RMSE:   39.92         23.61         23.62

MAX_MEAN: 55.2       MAX_MED: 42.4       BMAD: 4.35

COFF: 26.1    HIT-CALL: 1     FITC: 46    ACTP: 1

FLAGS: 17; 11

HIT-PCT: 0.971   MED-GA: -1.5919   GA-CI: 6.2178
```

Internal Curve Review Results

- Internal curve review has resulted in:
  - Instances of correction of fitting procedures in the ToxCast Pipeline
  - Identification of issues with source data
  - Identification of additional flags or filters that could be used, depending on the application of ToxCast data
  - a beta implementation of quality assurance for HTS data
  - Brown & Paul-Friedman, **Uncertainty in ToxCast Curve-Fitting: Quantitative and Qualitative Descriptors Inform a Model to Predict Reproducible Fits** *(in preparation)*



tcpl: the ToxCast pipeline for high-throughput screening data FREE

Dayne L Filer, Parth Kothiya, R Woodrow Setzer, Richard S Judson, Matthew T Martin

Bioinformatics, Volume 33, Issue 4, 15 February 2017, Pages 618–620,
https://doi.org/10.1093/bioinformatics/btw680

Published: 22 November 2016 **Article history**

Chemical structures as text

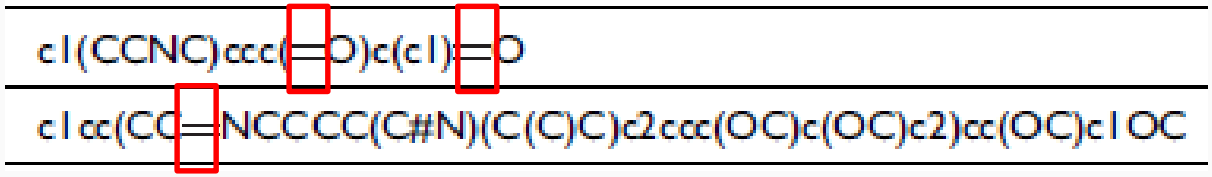
EPA United States Environmental Protection Agency

Table 1. Metabolites found in the different incubations tested

Name	RT	m/z	Formula	m/z Diff (ppm)	Mass score	SMILES
Parent	3.13	455.2926	C27H38N2O4	−3.48		N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2
M6 −164	2.26	291.2077	C17H26N2O2	−1.37	429	N(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1
M16 −14	3.06	441.2743	C26H36N2O4	2.41	534	c1cc(CCNCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
M14 +16	2.92	471.2866	C27H38N2O5	−1.59	476	N#CC(CCCN(C)CC(O)c1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2
M9 −14	2.78	441.2761	C26H36N2O4	−1.7	590	C(#N)C(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(O)c(OC)c2
M11 −14	2.84	441.2742	C26H36N2O4	2.57	473	Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC
M12 +2	2.87	457.2707	C26H36N2O5	−0.93	570	O(C)c1cc(ccc1OC)C(C#N)(CCCNCC(O)c2ccc(OC)c(OC)c2)C(C)C
M5 −178	2.2	277.1894	C16H24N2O2	7.84	419	C(C)(C)C(C#N)(CCCN)c1ccc(OC)c(OC)c1
M8 +2	2.67	457.2708	C26H36N2O5	−1.18	581	OC(CN(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1)c2ccc(O)c(OC)c2
M15 −14	2.92	441.2743	C26H36N2O4	2.23	614	Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC
M2 −259	0.73	196.1326	C11H17NO2	5.91	618	c1(CCNC)ccc(OC)c(c1)OC
M10 −28	2.8	427.2617	C25H34N2O4	−4.79	487	c1(OC)cc(ccc1OC)C(C#N)(CCCNCCc2ccc(O)c(OC)c2)C(C)C
M7 +2	2.46	457.2717	C26H36N2O5	−3.23	492	c1cc(CCN(O)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
M17 +16	3.21	471.2853	C27H38N2O5	1.19	534	N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(c2ccc(OC)c(OC)c2)C(C)(C)O
M4 −178	1.86	277.1927	C16H24N2O2	−3.8	444	COc1cc(ccc1O)C(C#N)(CCCNC)C(C)C
M1 −289	0.44	166.0858	C9H11NO2	5.93	136	c1(CCNC)ccc(=O)c(c1)=O
M13 −16	2.88	439.2603	C26H34N2O4	−1.43	367	c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC

How do I extract structures?

- Copy-Paste doesn't work



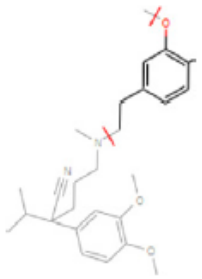
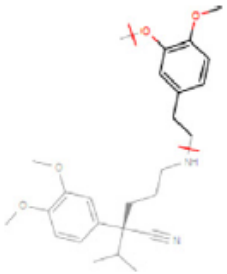
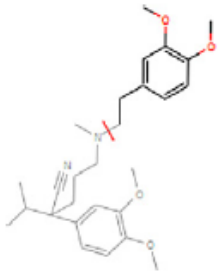
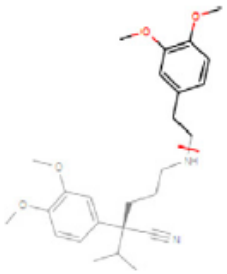
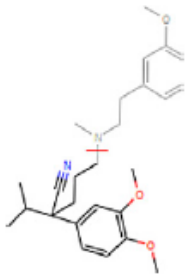
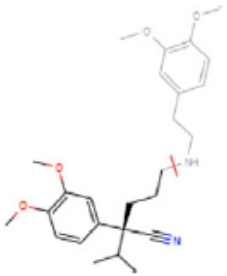

```
c1(CCNC)ccc( O)c(c1) O
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

- This is not the way a publisher should deliver chemistry. But this is on the AUTHOR!

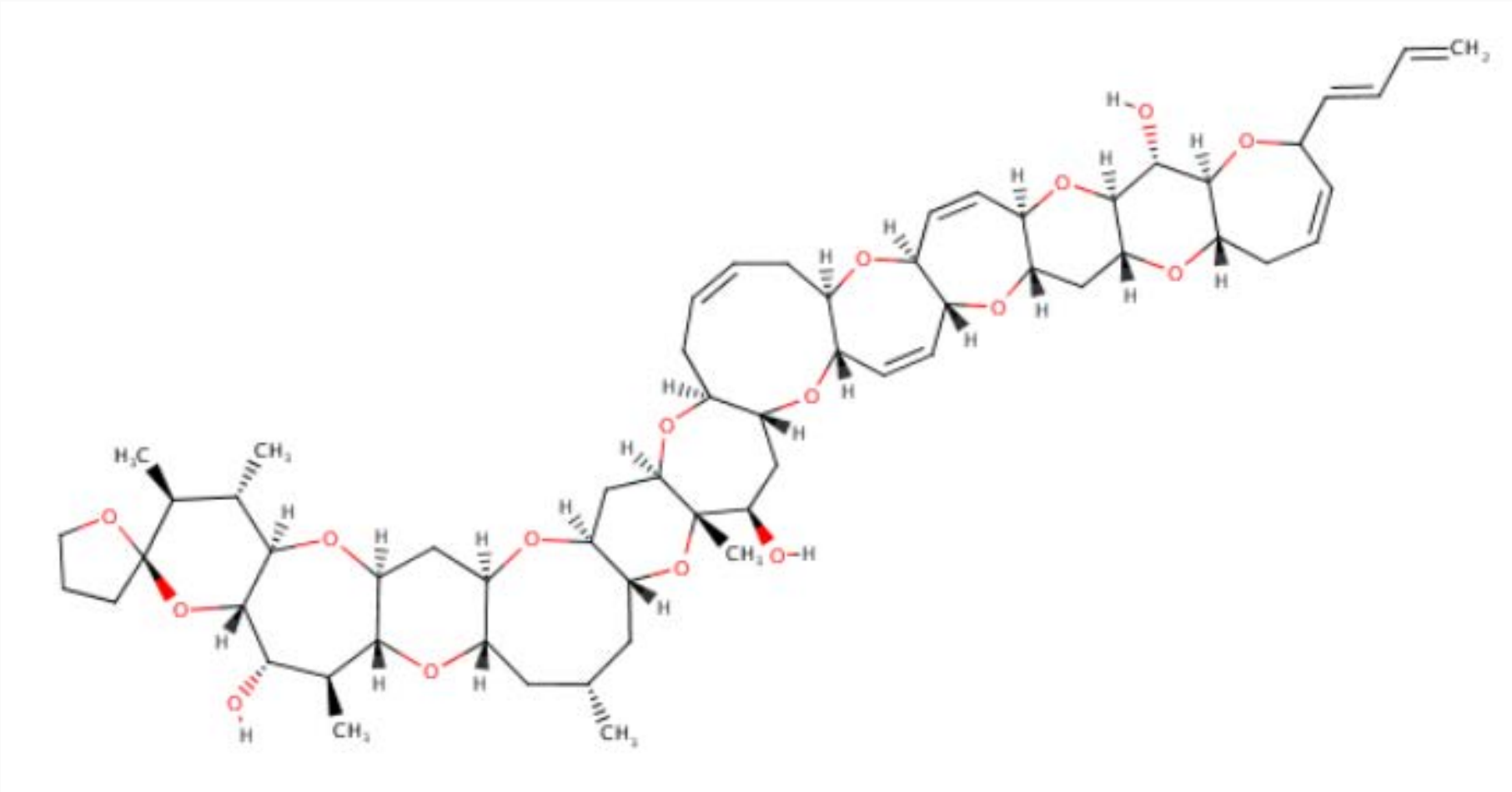
Extract Structure Drawings???

EPA United States Environmental Protection Agency

Table 2. Selection of fragments that help in the M16-16 metabolite structure elucidation

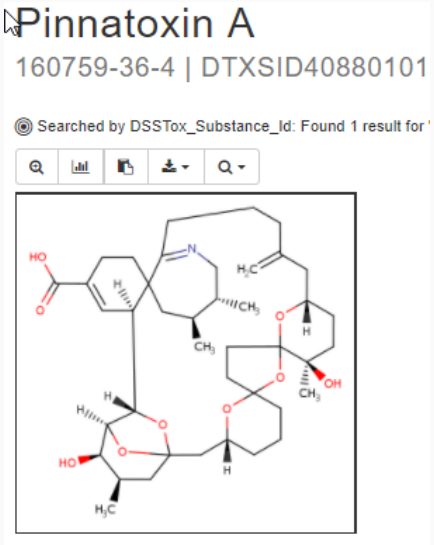
Sub. obs. m/z	Sub. cal. m/z	Sub. m/z diff. ppm	Substrate	Metabolite	Δ	Met. obs. m/z	Met. calc. m/z	Met. m/z diff. ppm
150.0664	150.0681	11.42			+0	150.0670	150.0681	7.25
165.0869	165.0916	28.22			+0	165.0892	165.0916	14.30
260.1637	260.1651	5.33			+0	260.1652	260.1651	−0.50

Try hand-drawing Algal Toxins!

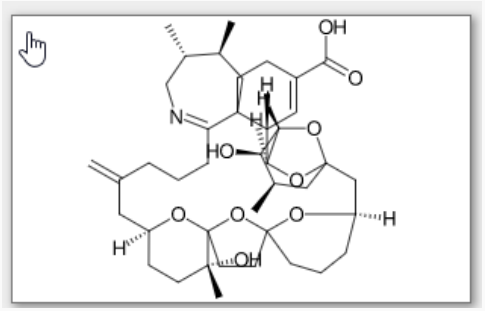


Think of files in multiple formats!

- SMILES are hyper-dependent on good layout algorithms. It's not easy!




```
[H][C@]12OC3(C[C@
@H](C)[C@H]1O)C[C
@@]1([H])CCCC4(CC
C5(O4)O[C@@]([H])(
CC[C@@]5(C)O)CC(=
C)CCCC4=NC[C@H](
C)[C@@H](C)CC44C
CC(=C[C@]4([H])[C@]
2([H])O3)C(O)=O)O1
```

Names and CASRNs are **NOT** structures

- In our domain most chemicals are 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

And generally problematic…

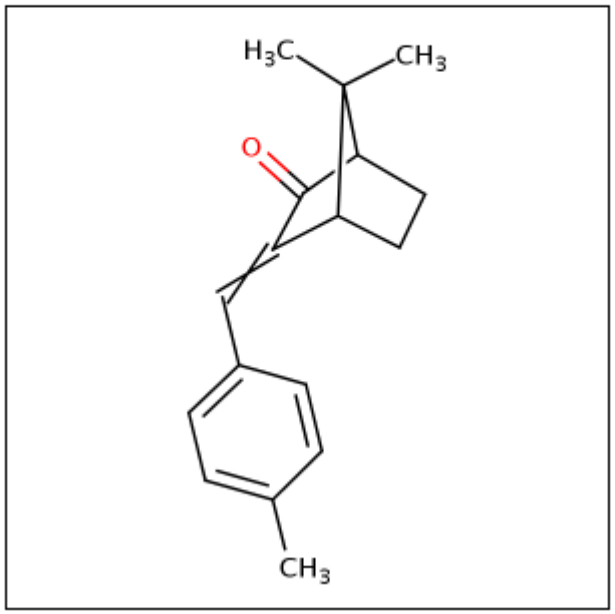
Name of Compound	CAS No.
Phenol-d6 (SS)	13187-88-3
Phenol	108-95-2
1,4-Dichlorobenzene	106-46-0
Acetophenone	98-86-2
Acenaphthene-d10 (IS)	15067-26-2
p-Cresol	106-44-5
Isophorone	78-59-1
Camphor	76-22-2
Isoborneol	124-76-5
Menthol	89, 78, 1
Naphthalene	91-20-3
Methyl salicilate	119-36-8

Active vs Deleted CASRN

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

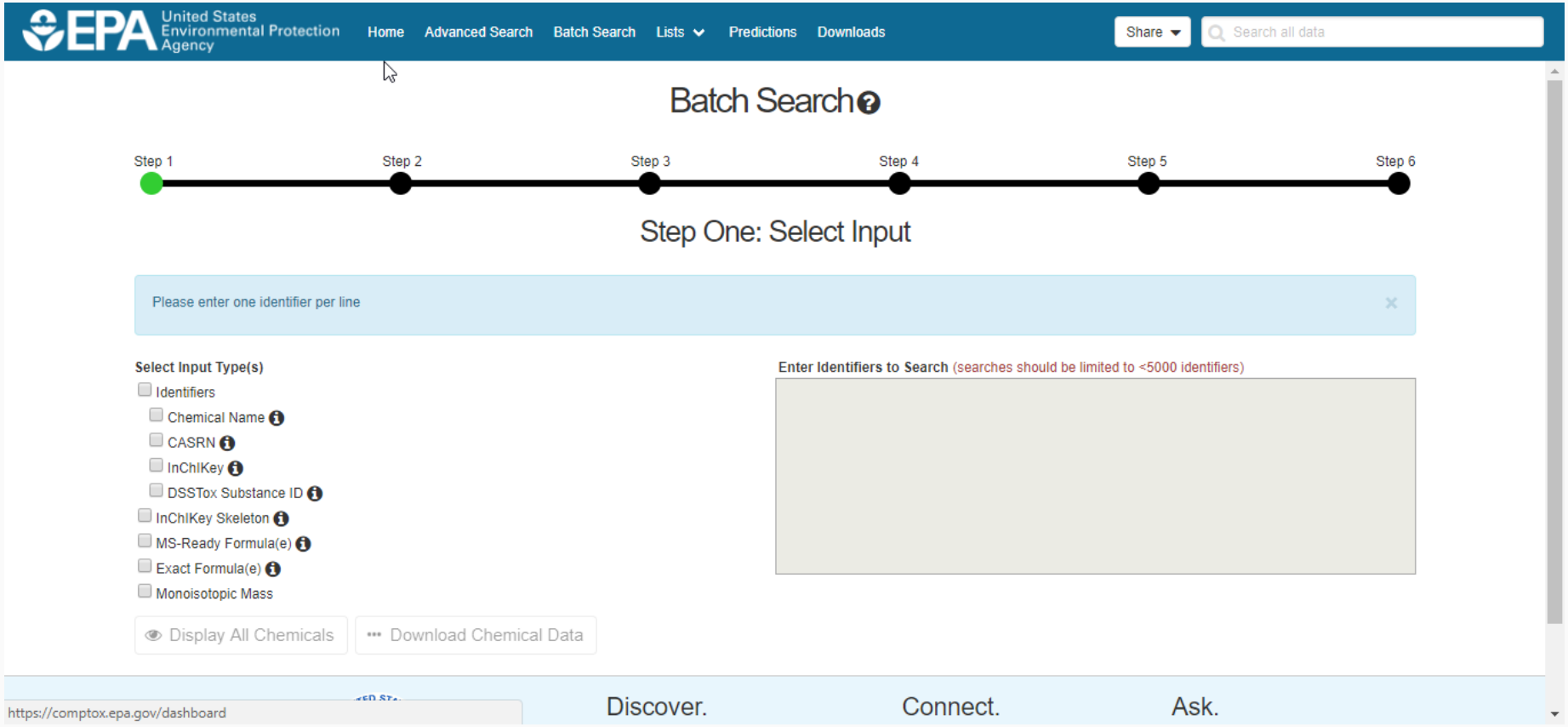
Tricky mapping by CASRN
This one has **316** Deleted CASRN

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

PLEASE use our curated data



Thirty Years of Experience

- What has changed?
  - Cheminformatics has progressed
  - The internet proliferates data access
  - Standards have progressed (InChI) and are improving
  - Anybody can access/download millions of structures!
- What hasn't changed?
  - Minor progress presenting structures in publications
  - Delivery via PDFs still dominates
  - Mandates on scientists are very unlikely

Conclusion

- The CompTox Dashboard provides access to data for ~760,000 chemicals
- Sourced and aggregated data from multiple resources
- Data quality is very random, source-to-source.
- We work hard to curate data - PLEASE USE IT!
- Our data is not perfect - we have a long way to go!

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