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

The US EPA CompTox Chemicals Dashboard and using InChI as a mapping identifier

Antony John Williams

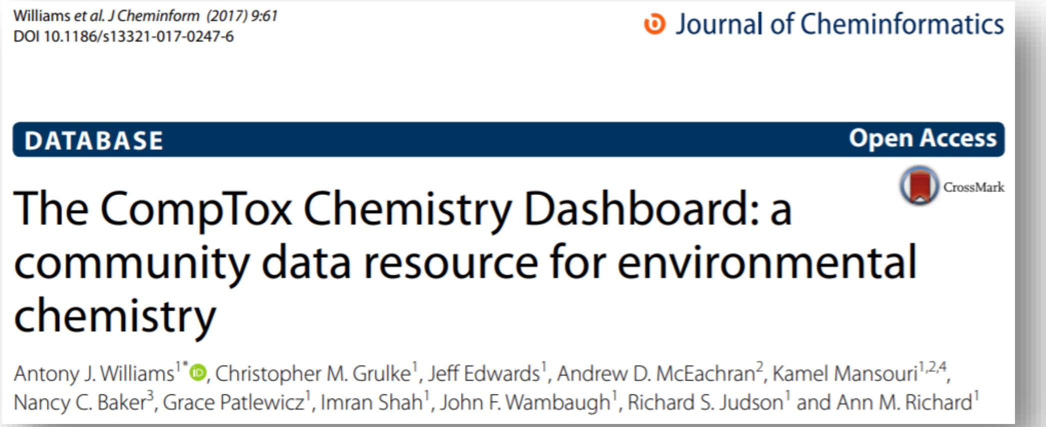
williams.antony@epa.gov

Center for Computational Toxicology and Exposure, US-EPA, RTP, NC

March 23^{3d} 2021 : Virtual Workshop on International Chemical Identifiers (InChI)

The intention for the Dashboard

- Develop a “**first-stop-shop**” for environmental chemical data to support US-EPA and partner decision making:
  - Centralized location for relevant chemical data
  - Chemistry, exposure, hazard and dosimetry
  - Combination of existing data and predictive models
  - Publicly accessible, periodically updated, curated



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

Journal of Cheminformatics

DATABASE **Open Access**

CrossMark

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]

Computational Toxicology 12 (2019) 100096

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

ELSEVIER
COMPUTATIONAL TOXICOLOGY
Check for updates

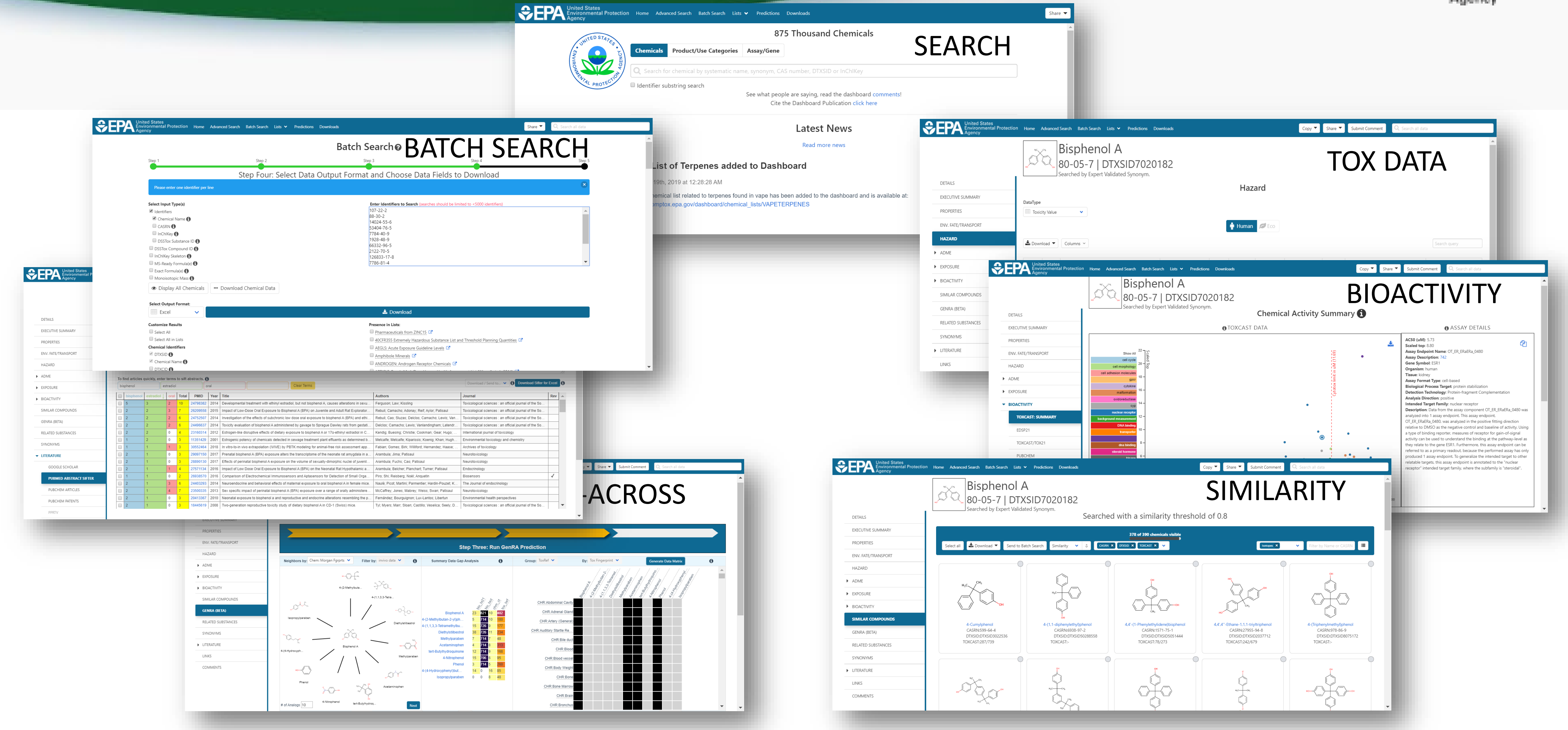
EPA’s DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research

Christopher M. Grulke[a], Antony J. Williams[a], Inthirany Thillanadarajah[b], Ann M. Richard[a,*]

[a] *National Center for Computational Toxicology, Office of Research & Development, US Environmental Protection Agency, Mail Drop D143-02, Research Triangle Park, NC 27711, USA*
[b] *Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA*

CompTox Chemicals Dashboard

https://comptox.epa.gov/dashboard



SEARCH

875 Thousand Chemicals

Chemicals | Product/Use Categories | Assay/Gene

☐ Identifier substring search

See what people are saying, read the dashboard comments!
Cite the Dashboard Publication click here

Latest News

Read more news

List of Terpenes added to Dashboard

BATCH SEARCH

Batch Search

Step Four: Select Data Output Format and Choose Data Fields to Download

TOX DATA

Bisphenol A
80-05-7 | DTXSID7020182
Searched by Expert Validated Synonym.

Hazard

BIOACTIVITY

Bisphenol A
80-05-7 | DTXSID7020182
Searched by Expert Validated Synonym.

Chemical Activity Summary

READ-ACROSS

Step Three: Run GenRA Prediction

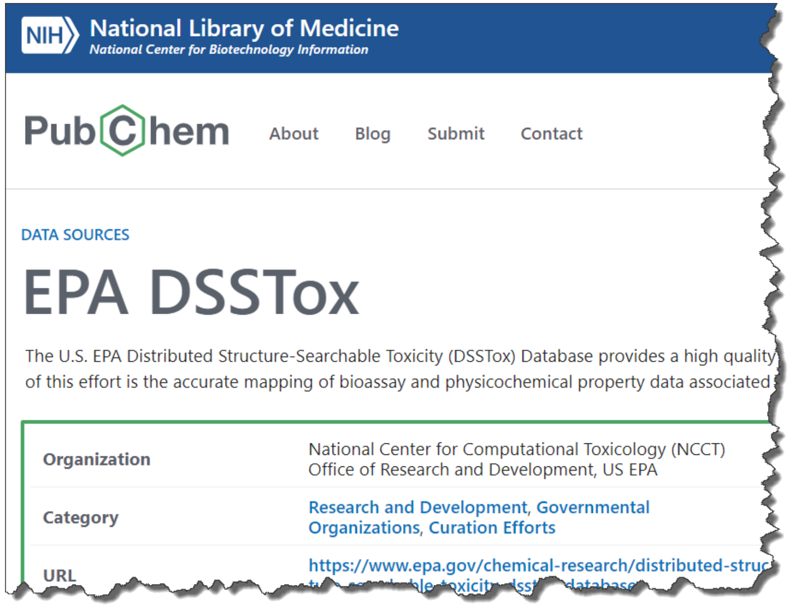
SIMILARITY

Bisphenol A
80-05-7 | DTXSID7020182
Searched by Expert Validated Synonym.

Searched with a similarity threshold of 0.8

The Dashboard data collection

- Relative to PubChem & ChemSpider our collection is small (but open)

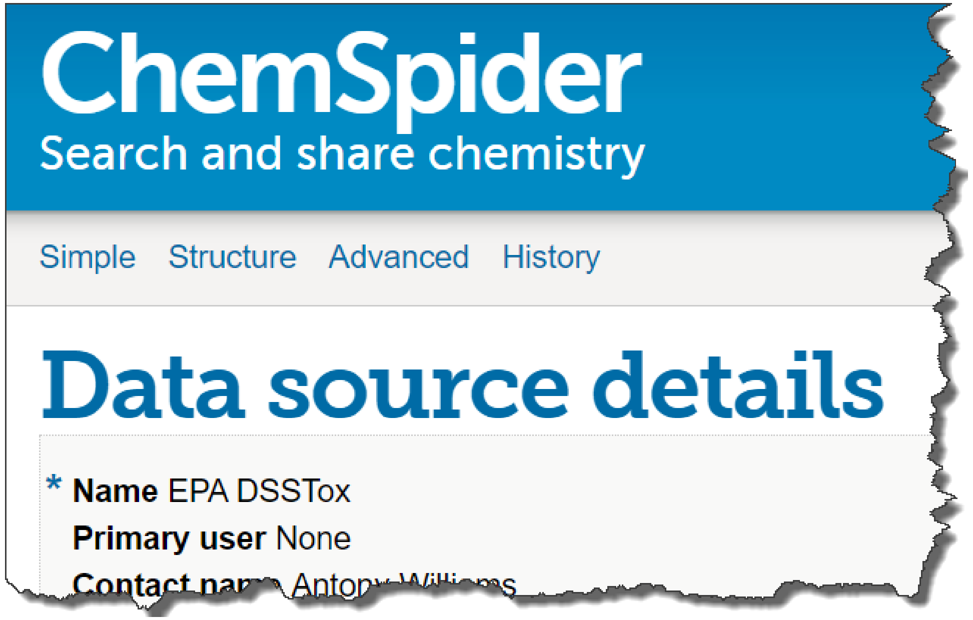

NIH
National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Blog
Submit
Contact

DATA SOURCES

EPA DSSTox

The U.S. EPA Distributed Structure-Searchable Toxicity (DSSTox) Database provides a high quality
of this effort is the accurate mapping of bioassay and physicochemical property data associated

Organization	National Center for Computational Toxicology (NCCT) Office of Research and Development, US EPA
Category	Research and Development, Governmental Organizations, Curation Efforts
URL	https://www.epa.gov/chemical-research/distributed-struc



Data source details

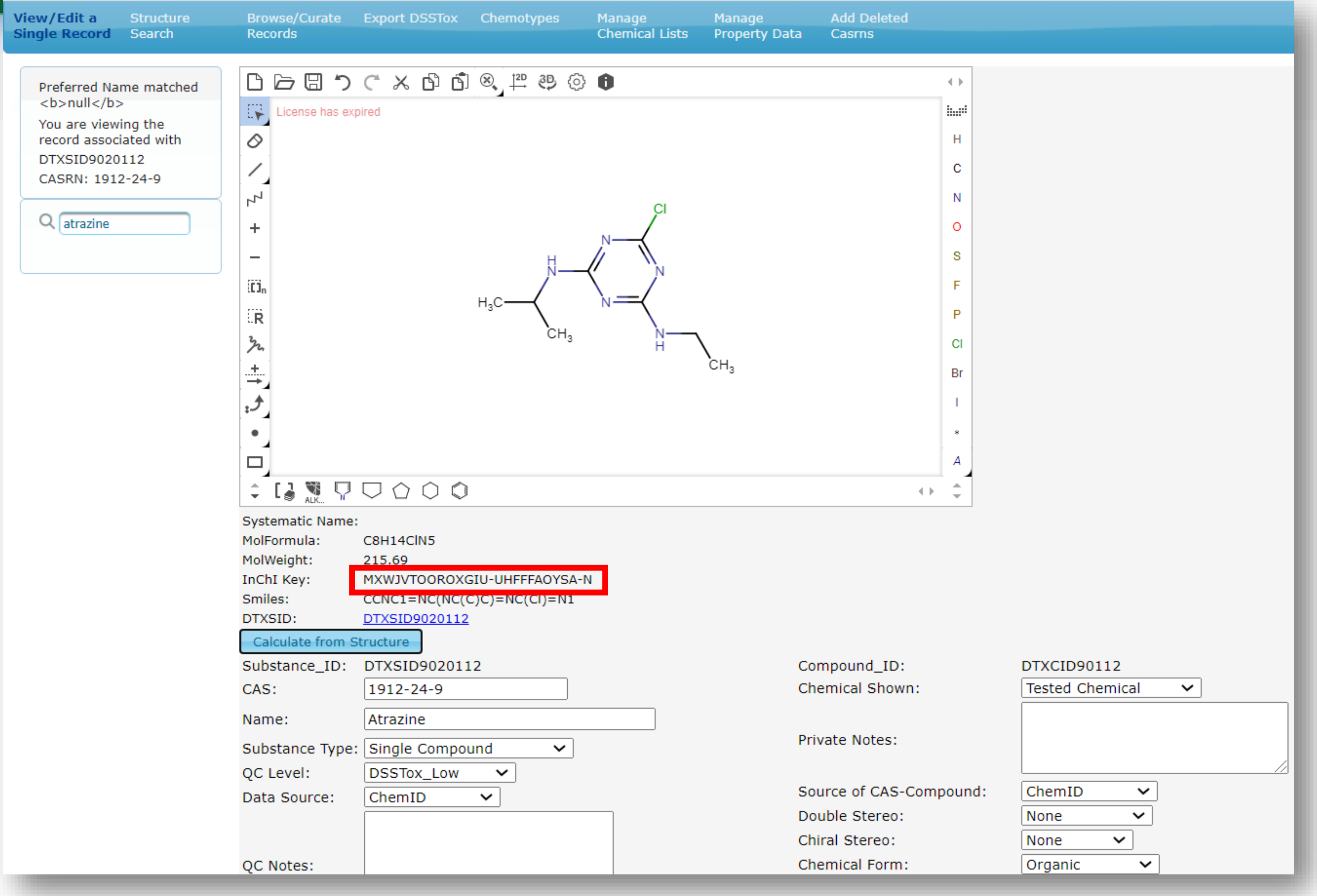
* **Name** EPA DSSTox
Primary user None
Contact name Antony Williams

- We are focused specifically on chemicals of interest to the agency and, increasingly, those that can be detected in the environment

We use InChI identifiers…

- …in our registration and curation processes
- …for searching single chemicals and batches of chemicals
- …for mappings within the application
- …for linking to third-party websites
- …for registration into resolver databases
- …to add to our data exports for database mapping
- …as default information in many of our download files

ChemReg Registration Tool



Preferred Name matched <b>null</b>

You are viewing the record associated with DTXSID9020112 CASRN: 1912-24-9

atrazine

Systematic Name:

MolFormula: C8H14ClN5

MolWeight: 215.69

InChI Key: MXWJVTOOROXGIU-UHFFFAOYSA-N

Smiles: CCNC1=NC(NC(C)C)=NC(Cl)=N1

DTXSID: DTXSID9020112

Substance_ID: DTXSID9020112

CAS: 1912-24-9

Name: Atrazine

Substance Type: Single Compound

QC Level: DSSTox_Low

Data Source: ChemID

QC Notes:

Compound_ID: DTXCID90112

Chemical Shown: Tested Chemical

Private Notes:

Source of CAS-Compound: ChemID

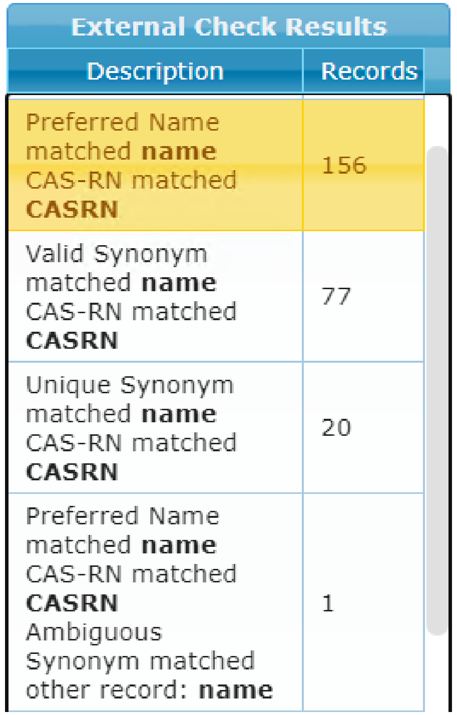
Double Stereo: None

Chiral Stereo: None

Chemical Form: Organic

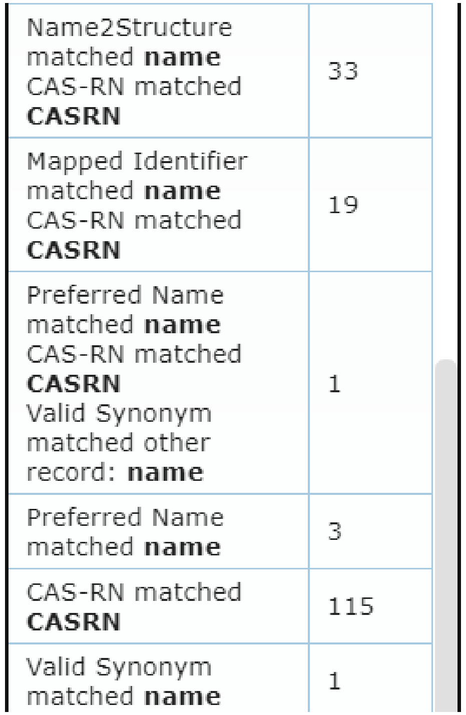
Using InChI during registrations

- List registrations can often include InChIs as one of the identifiers
- We use InChIs in tandem with names, CASRNs, SMILES etc to cross-reference with existing records in the database
- Our team of curators use InChIs for online searching and comparison

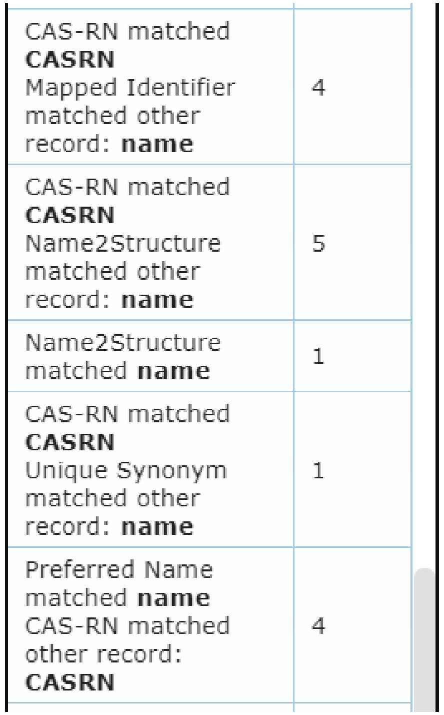


External Check Results

Description	Records
Preferred Name matched **name** CAS-RN matched **CASRN**	156
Valid Synonym matched **name** CAS-RN matched **CASRN**	77
Unique Synonym matched **name** CAS-RN matched **CASRN**	20
Preferred Name matched **name** CAS-RN matched **CASRN** Ambiguous Synonym matched other record: **name**	1



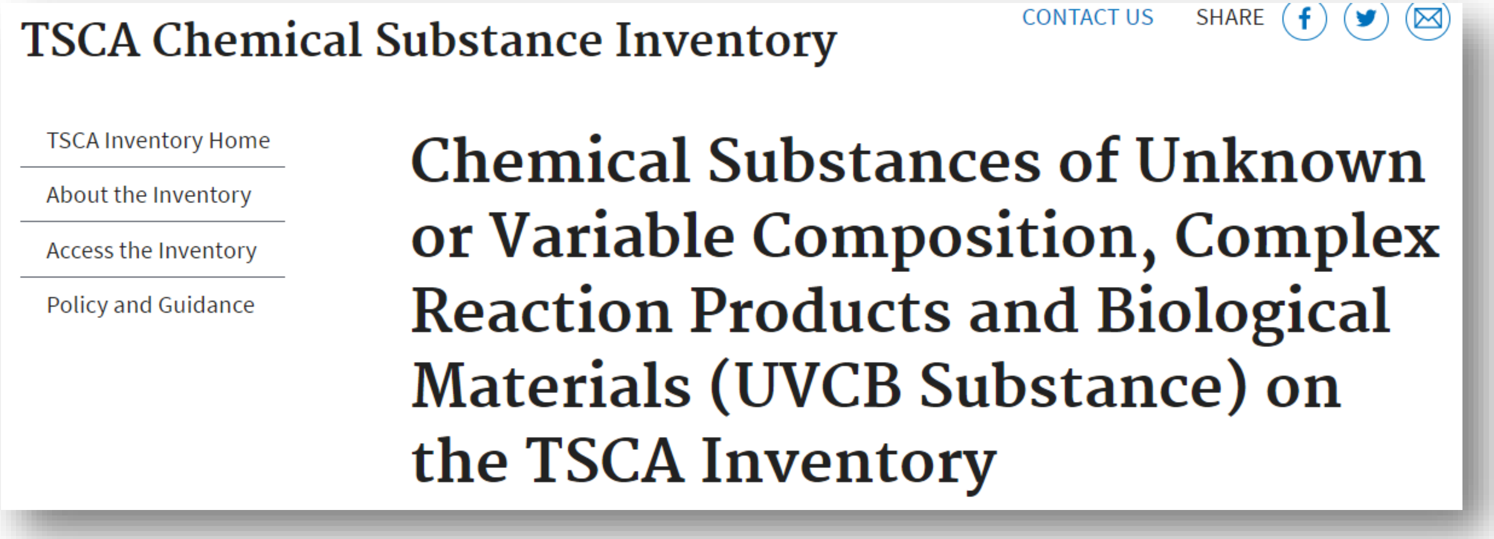
Name2Structure matched **name** CAS-RN matched **CASRN**	33
Mapped Identifier matched **name** CAS-RN matched **CASRN**	19
Preferred Name matched **name** CAS-RN matched **CASRN** Valid Synonym matched other record: **name**	1
Preferred Name matched **name**	3
CAS-RN matched **CASRN**	115
Valid Synonym matched **name**	1



CAS-RN matched **CASRN** Mapped Identifier matched other record: **name**	4
CAS-RN matched **CASRN** Name2Structure matched other record: **name**	5
Name2Structure matched **name**	1
CAS-RN matched **CASRN** Unique Synonym matched other record: **name**	1
Preferred Name matched **name** CAS-RN matched other record: **CASRN**	4

Examples of use in curation

- Where mappings against substances in our database collide
  - Polymers with InChIs
  - Ambiguous stoichiometry
  - UVCB chemicals

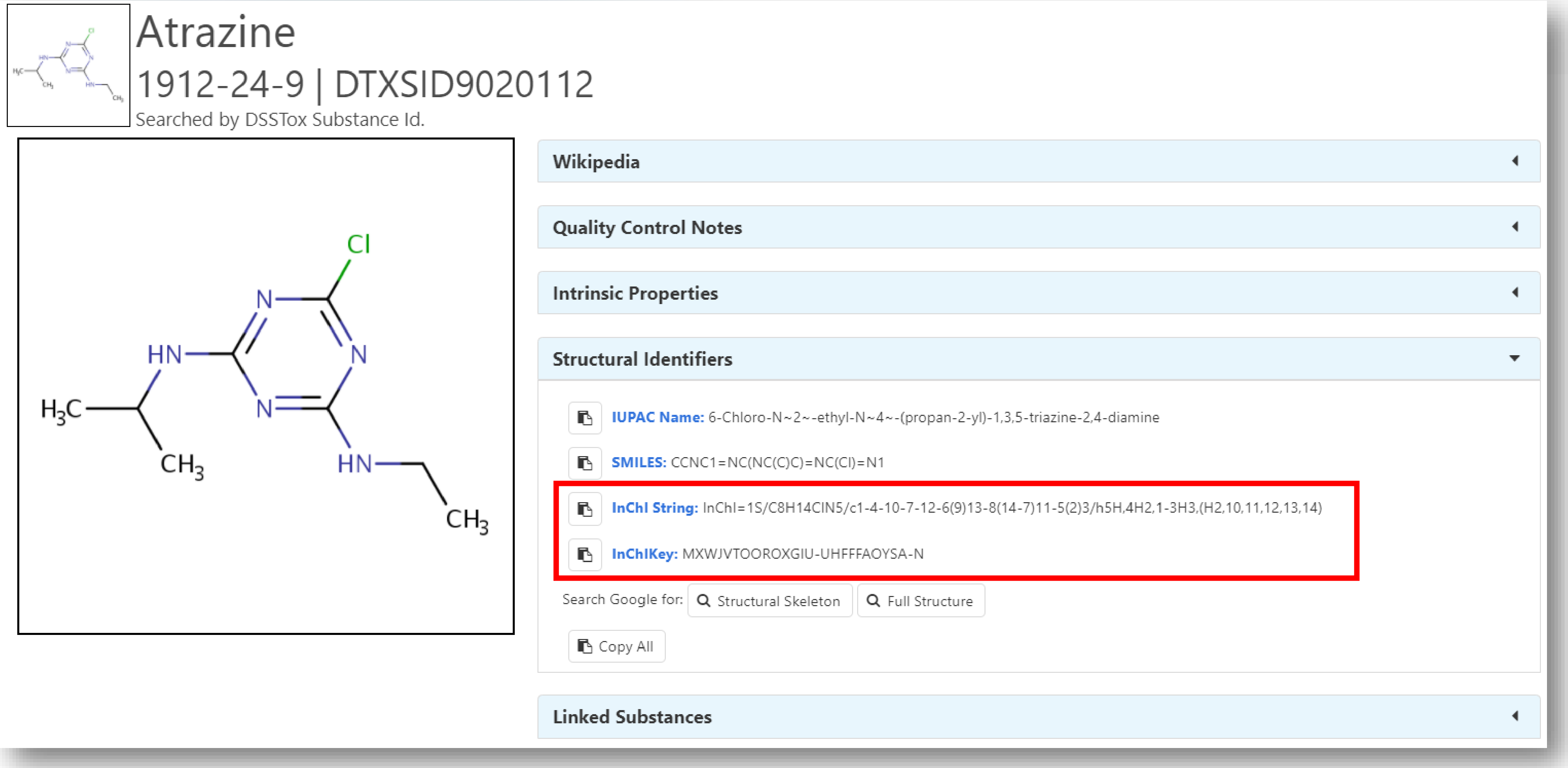


TSCA Chemical Substance Inventory

Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials (UVCB Substance) on the TSCA Inventory

- InChI identifiers are **essential** in our curation process

Structural Identifiers



Atrazine

1912-24-9 | DTXSID9020112

Searched by DSSTox Substance Id.

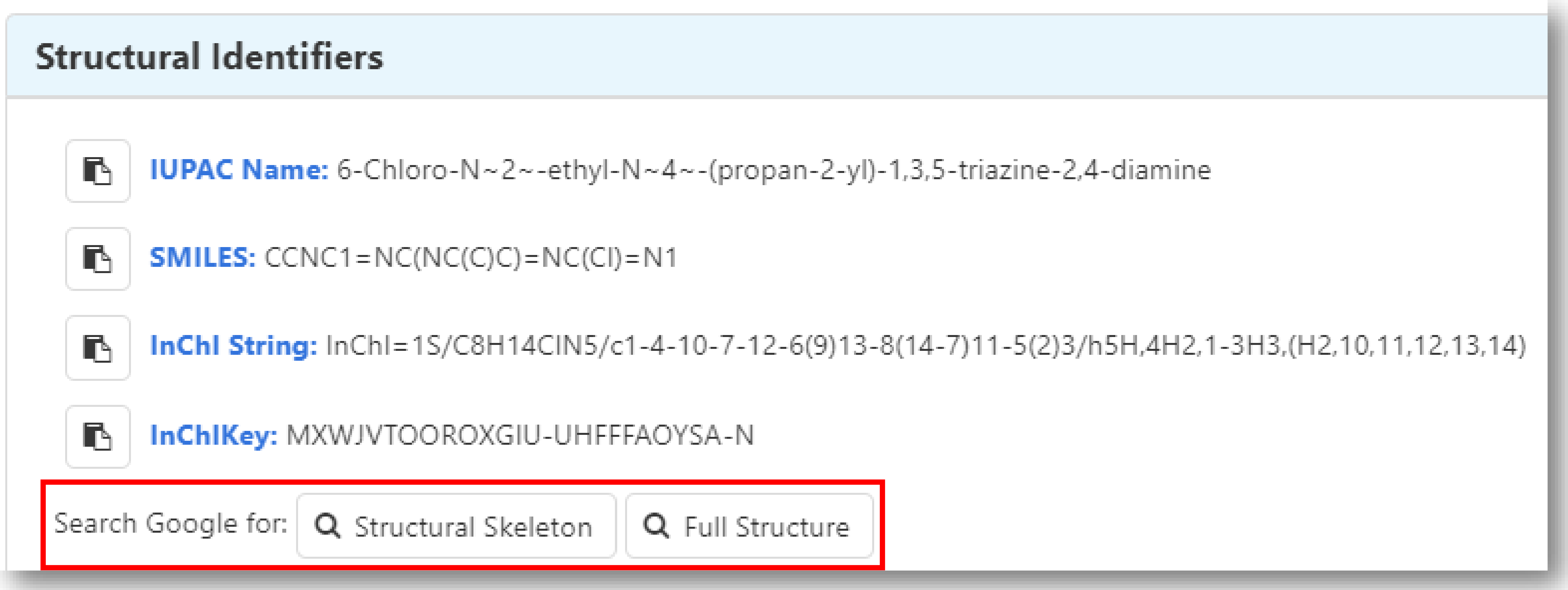
IUPAC Name: 6-Chloro-N~2~-ethyl-N~4~-(propan-2-yl)-1,3,5-triazine-2,4-diamine

SMILES: CCNC1=NC(NC(C)C)=NC(Cl)=N1

InChI String: InChI=1S/C8H14ClN5/c1-4-10-7-12-6(9)13-8(14-7)11-5(2)3/h5H,4H2,1-3H3,(H2,10,11,12,13,14)

InChIKey: MXWJVTOOROXGIU-UHFFFAOYSA-N

External Searches



Structural Identifiers

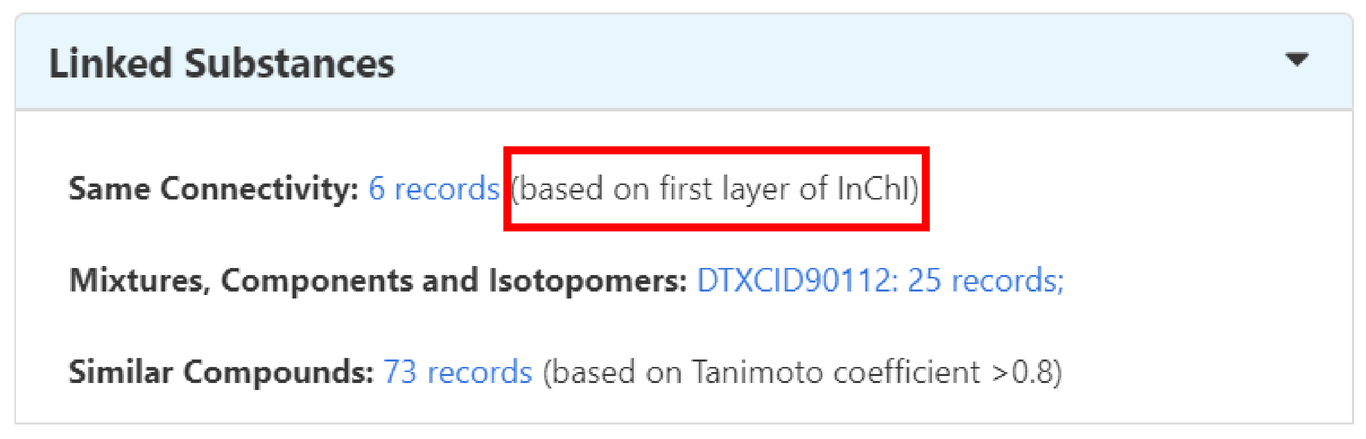
IUPAC Name: 6-Chloro-N~2~-ethyl-N~4~-(propan-2-yl)-1,3,5-triazine-2,4-diamine

SMILES: CCNC1=NC(NC(C)C)=NC(Cl)=N1

InChI String: InChI=1S/C8H14ClN5/c1-4-10-7-12-6(9)13-8(14-7)11-5(2)3/h5H,4H2,1-3H3,(H2,10,11,12,13,14)

InChIKey: MXWJVTOOROXGIU-UHFFFAOYSA-N

Linked Substances



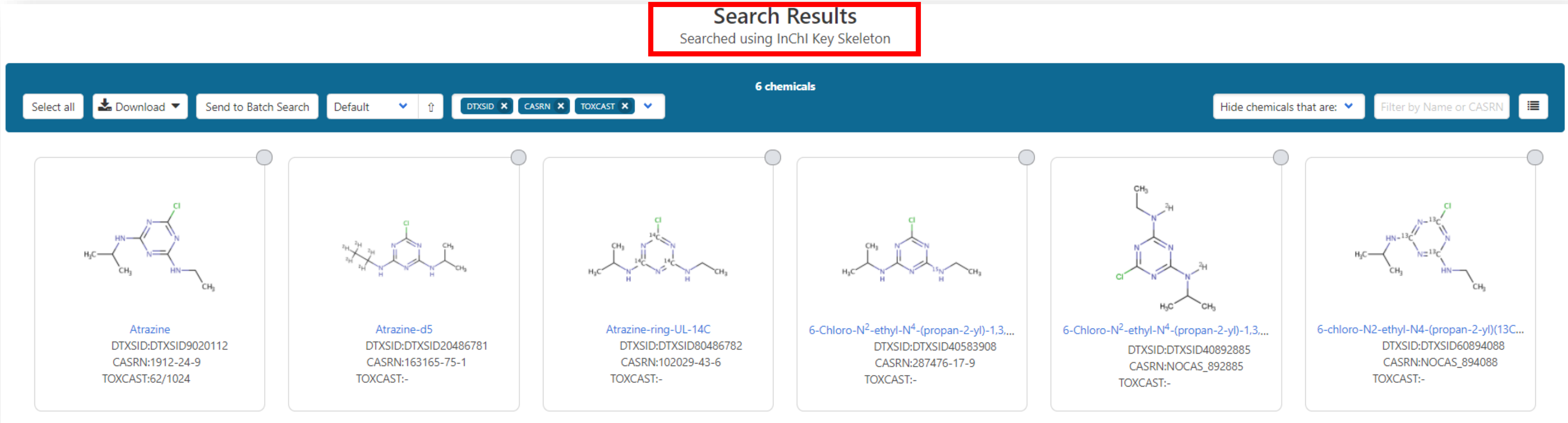
Linked Substances

Same Connectivity: 6 records (based on first layer of InChI)

Mixtures, Components and Isotopomers: DTXCID90112: 25 records;

Similar Compounds: 73 records (based on Tanimoto coefficient >0.8)

https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?input_type=inchikey_skeleton&inputs=MXWJVTOOROXGIU



Search Results

Searched using InChI Key Skeleton

6 chemicals

Atrazine
DTXSID:DTXSID9020112
CASRN:1912-24-9
TOXCAST:62/1024

Atrazine-d5
DTXSID:DTXSID20486781
CASRN:163165-75-1
TOXCAST:-

Atrazine-ring-UL-14C
DTXSID:DTXSID80486782
CASRN:102029-43-6
TOXCAST:-

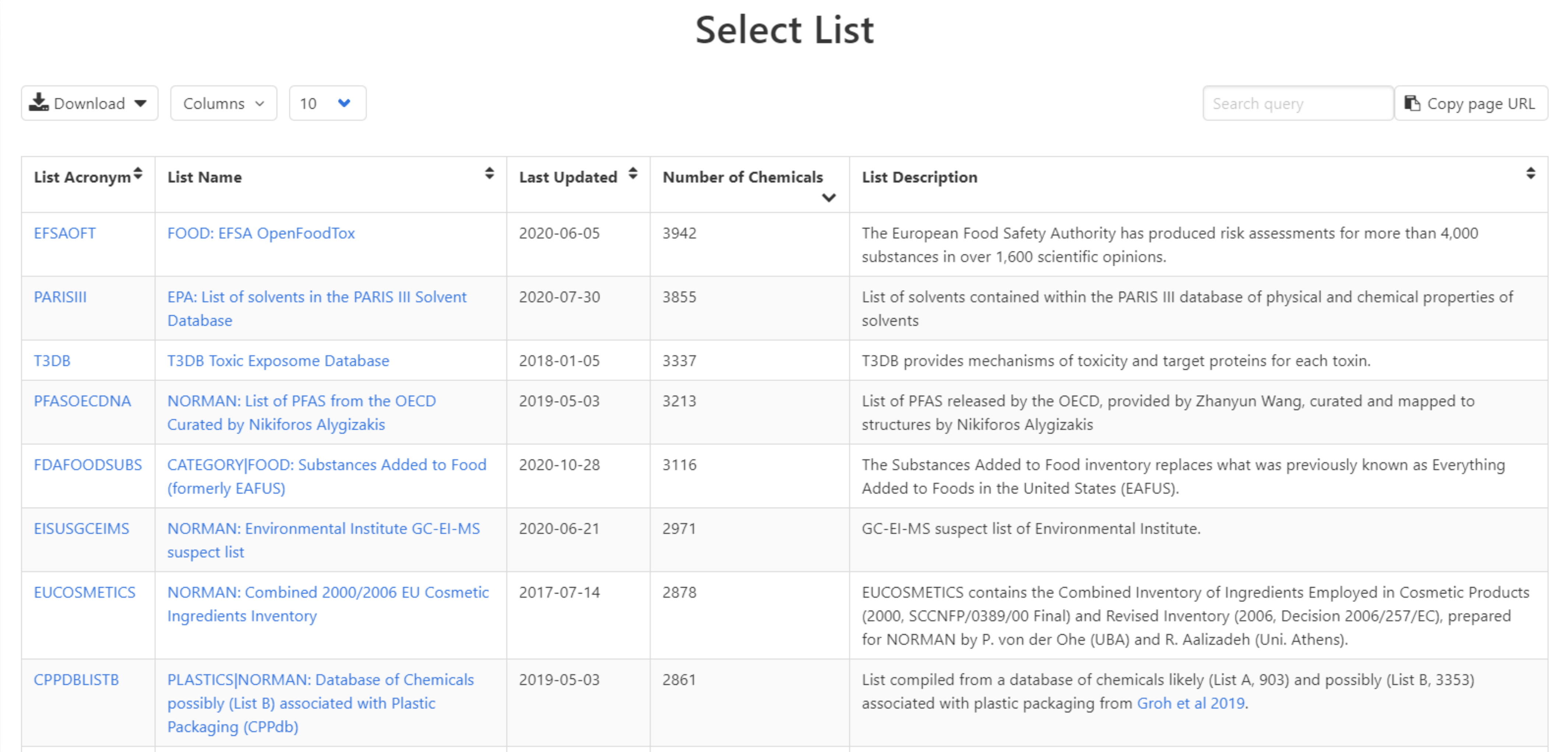
6-Chloro-N^2-ethyl-N^4-(propan-2-yl)-1,3,...
DTXSID:DTXSID40583908
CASRN:287476-17-9
TOXCAST:-

6-Chloro-N^2-ethyl-N^4-(propan-2-yl)-1,3,...
DTXSID:DTXSID40892885
CASRN:NOCAS_892885
TOXCAST:-

6-chloro-N2-ethyl-N4-(propan-2-yl)(13C...
DTXSID:DTXSID60894088
CASRN:NOCAS_894088
TOXCAST:-

Chemical List collections

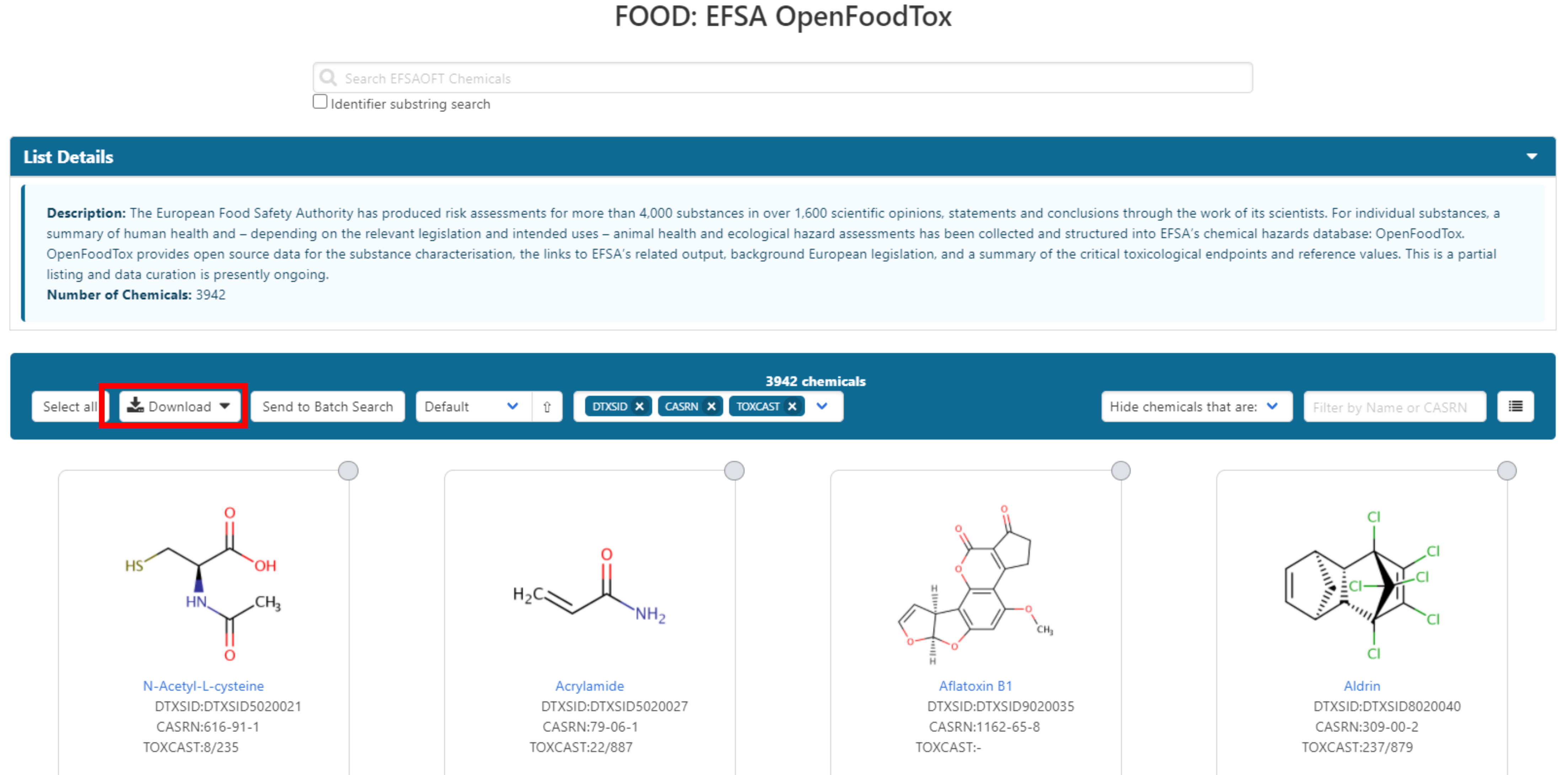
- ~300 lists of chemicals aggregated by content or category



Select List

List Acronym	List Name	Last Updated	Number of Chemicals	List Description
EFSAOFT	FOOD: EFSA OpenFoodTox	2020-06-05	3942	The European Food Safety Authority has produced risk assessments for more than 4,000 substances in over 1,600 scientific opinions.
PARISIII	EPA: List of solvents in the PARIS III Solvent Database	2020-07-30	3855	List of solvents contained within the PARIS III database of physical and chemical properties of solvents
T3DB	T3DB Toxic Exposome Database	2018-01-05	3337	T3DB provides mechanisms of toxicity and target proteins for each toxin.
PFASOECDNA	NORMAN: List of PFAS from the OECD Curated by Nikiforos Alygizakis	2019-05-03	3213	List of PFAS released by the OECD, provided by Zhanyun Wang, curated and mapped to structures by Nikiforos Alygizakis
FDAFOODSUBS	CATEGORY|FOOD: Substances Added to Food (formerly EAFUS)	2020-10-28	3116	The Substances Added to Food inventory replaces what was previously known as Everything Added to Foods in the United States (EAFUS).
EISUSGCEIMS	NORMAN: Environmental Institute GC-EI-MS suspect list	2020-06-21	2971	GC-EI-MS suspect list of Environmental Institute.
EUCOSMETICS	NORMAN: Combined 2000/2006 EU Cosmetic Ingredients Inventory	2017-07-14	2878	EUCOSMETICS contains the Combined Inventory of Ingredients Employed in Cosmetic Products (2000, SCCNFP/0389/00 Final) and Revised Inventory (2006, Decision 2006/257/EC), prepared for NORMAN by P. von der Ohe (UBA) and R. Aalizadeh (Uni. Athens).
CPPDBLISTB	PLASTICS|NORMAN: Database of Chemicals possibly (List B) associated with Plastic Packaging (CPPdb)	2019-05-03	2861	List compiled from a database of chemicals likely (List A, 903) and possibly (List B, 3353) associated with plastic packaging from Groh et al 2019.

Downloadable with InChIs by default



FOOD: EFSA OpenFoodTox

Description: The European Food Safety Authority has produced risk assessments for more than 4,000 substances in over 1,600 scientific opinions, statements and conclusions through the work of its scientists. For individual substances, a summary of human health and – depending on the relevant legislation and intended uses – animal health and ecological hazard assessments has been collected and structured into EFSA's chemical hazards database: OpenFoodTox. OpenFoodTox provides open source data for the substance characterisation, the links to EFSA's related output, background European legislation, and a summary of the critical toxicological endpoints and reference values. This is a partial listing and data curation is presently ongoing.

Number of Chemicals: 3942

3942 chemicals

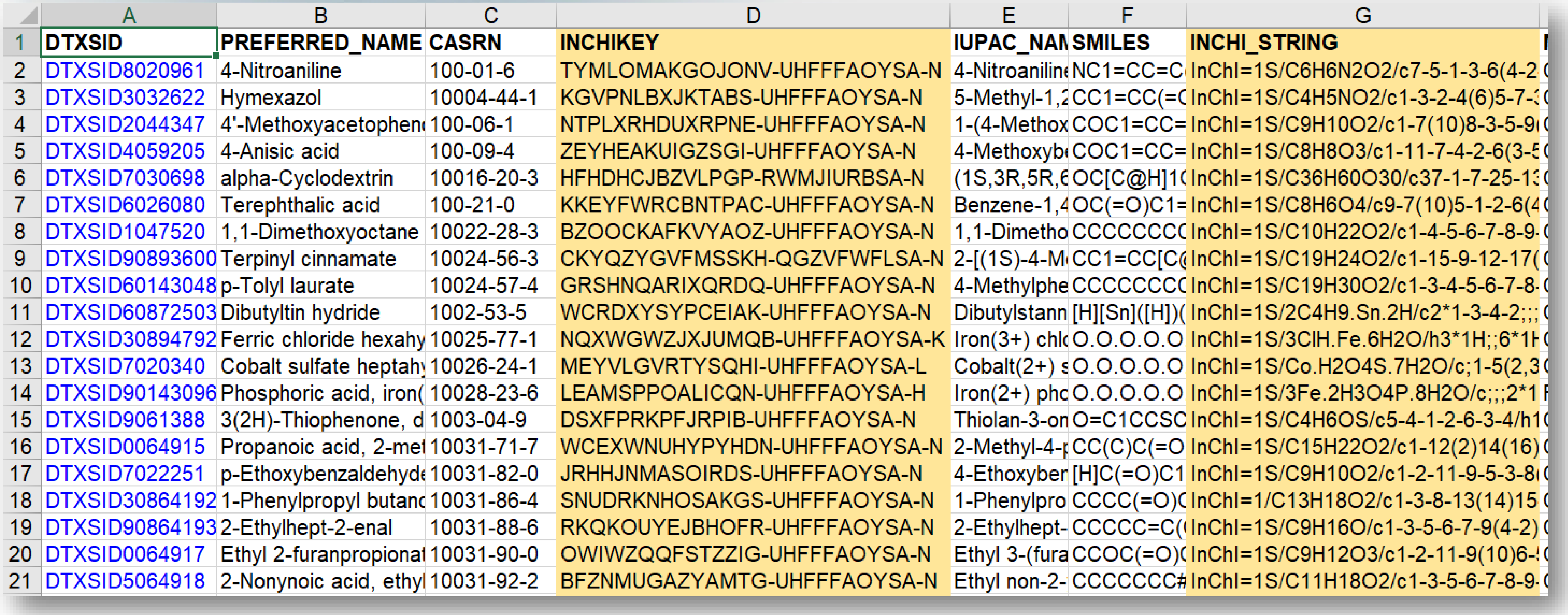
N-Acetyl-L-cysteine
DTXSID:DTXSID5020021
CASRN:616-91-1
TOXCAST:8/235

Acrylamide
DTXSID:DTXSID5020027
CASRN:79-06-1
TOXCAST:22/887

Aflatoxin B1
DTXSID:DTXSID9020035
CASRN:1162-65-8
TOXCAST:-

Aldrin
DTXSID:DTXSID8020040
CASRN:309-00-2
TOXCAST:237/879

Exported: Excel, TSV, SDF

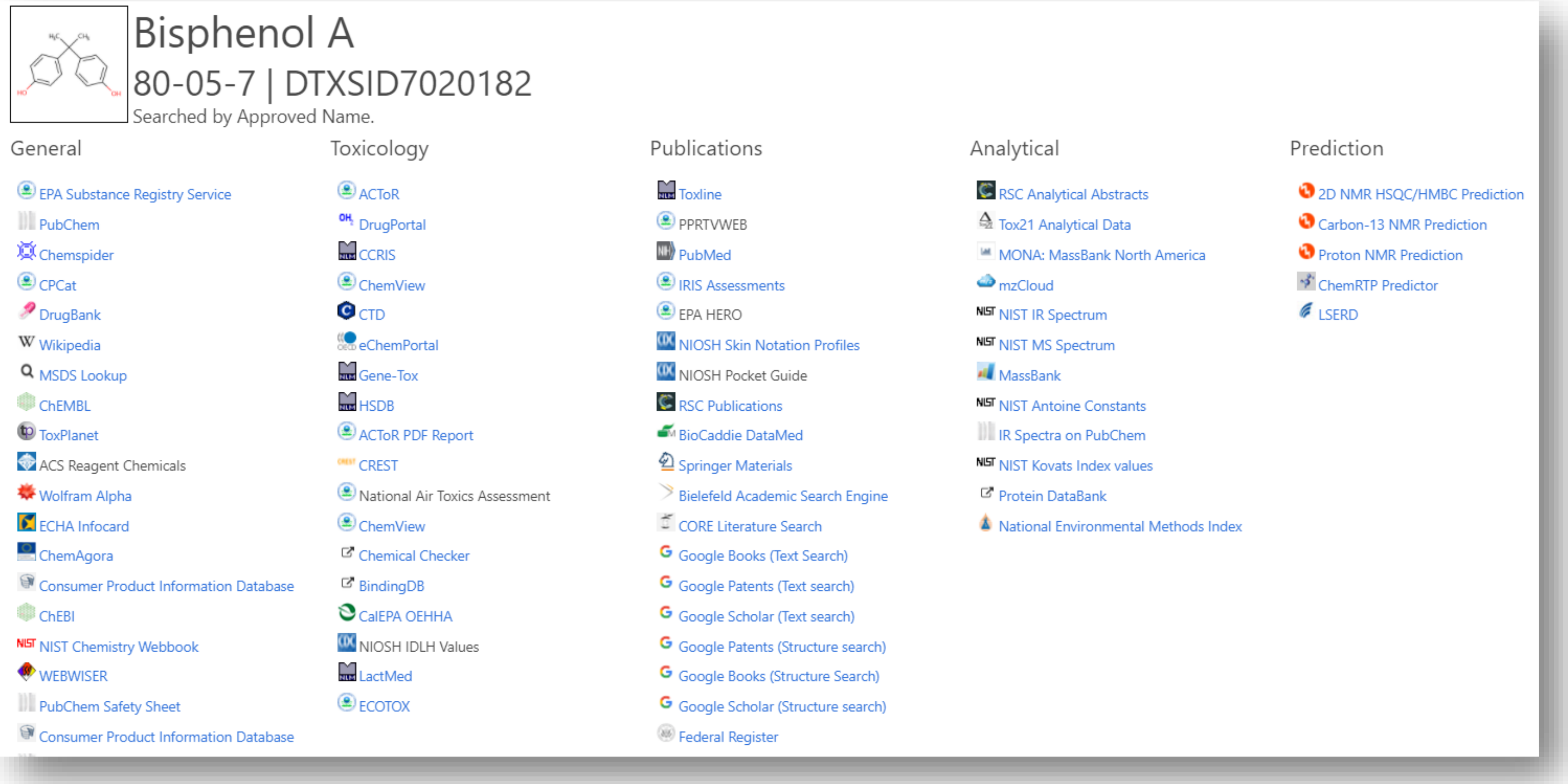


	A	B	C	D	E	F	G
1	**DTXSID**	**PREFERRED_NAME**	**CASRN**	**INCHIKEY**	**IUPAC_NAM**	**SMILES**	**INCHI_STRING**
2	DTXSID8020961	4-Nitroaniline	100-01-6	TYMLOMAKGOJONV-UHFFFAOYSA-N	4-Nitroanilin	NC1=CC=C	InChI=1S/C6H6N2O2/c7-5-1-3-6(4-2
3	DTXSID3032622	Hymexazol	10004-44-1	KGVPNLBXJKTABS-UHFFFAOYSA-N	5-Methyl-1,	CC1=CC(=	InChI=1S/C4H5NO2/c1-3-2-4(6)5-7-
4	DTXSID2044347	4'-Methoxyacetophen	100-06-1	NTPLXRHDUXRPNE-UHFFFAOYSA-N	1-(4-Methox	COC1=CC=	InChI=1S/C9H10O2/c1-7(10)8-3-5-9
5	DTXSID4059205	4-Anisic acid	100-09-4	ZEYHEAKUIGZSGI-UHFFFAOYSA-N	4-Methoxyb	COC1=CC=	InChI=1S/C8H8O3/c1-11-7-4-2-6(3-5
6	DTXSID7030698	alpha-Cyclodextrin	10016-20-3	HFHDHCJBZVLPGP-RWMJIURBSA-N	(1S,3R,5R,	OC[C@H]1	InChI=1S/C36H60O30/c37-1-7-25-1
7	DTXSID6026080	Terephthalic acid	100-21-0	KKEYFWRCBNTPAC-UHFFFAOYSA-N	Benzene-1,4	OC(=O)C1=	InChI=1S/C8H6O4/c9-7(10)5-1-2-6(4
8	DTXSID1047520	1,1-Dimethoxyoctane	10022-28-3	BZOOCKAFKVYAOZ-UHFFFAOYSA-N	1,1-Dimetho	CCCCCCC	InChI=1S/C10H22O2/c1-4-5-6-7-8-9-
9	DTXSID90893600	Terpinyl cinnamate	10024-56-3	CKYQZYGVFMSSKH-QGZVFWFLSA-N	2-[(1S)-4-M	CC1=CC[C	InChI=1S/C19H24O2/c1-15-9-12-17(
10	DTXSID60143048	p-Tolyl laurate	10024-57-4	GRSHNQARIXQRDQ-UHFFFAOYSA-N	4-Methylphe	CCCCCCC	InChI=1S/C19H30O2/c1-3-4-5-6-7-8-
11	DTXSID60872503	Dibutyltin hydride	1002-53-5	WCRDXYSYPCEIAK-UHFFFAOYSA-N	Dibutylstann	[H][Sn]([H])(	InChI=1S/2C4H9.Sn.2H/c2*1-3-4-2;;;
12	DTXSID30894792	Ferric chloride hexahy	10025-77-1	NQXWGWZJXJUMQB-UHFFFAOYSA-K	Iron(3+) chl	O.O.O.O.O	InChI=1S/3ClH.Fe.6H2O/h3*1H;;6*1H
13	DTXSID7020340	Cobalt sulfate heptah	10026-24-1	MEYVLGVRTYSQHI-UHFFFAOYSA-L	Cobalt(2+)	O.O.O.O.O	InChI=1S/Co.H2O4S.7H2O/c;1-5(2,3
14	DTXSID90143096	Phosphoric acid, iron(	10028-23-6	LEAMSPPOALICQN-UHFFFAOYSA-H	Iron(2+) ph	O.O.O.O.O	InChI=1S/3Fe.2H3O4P.8H2O/c;;;2*1
15	DTXSID9061388	3(2H)-Thiophenone, d	1003-04-9	DSXFPRKPFJRPIB-UHFFFAOYSA-N	Thiolan-3-on	O=C1CCSC	InChI=1S/C4H6OS/c5-4-1-2-6-3-4/h1
16	DTXSID0064915	Propanoic acid, 2-me	10031-71-7	WCEXWNUHYPYHDN-UHFFFAOYSA-N	2-Methyl-4-	CC(C)C(=O	InChI=1S/C15H22O2/c1-12(2)14(16)
17	DTXSID7022251	p-Ethoxybenzaldehyd	10031-82-0	JRHHJNMASOIRDS-UHFFFAOYSA-N	4-Ethoxyber	[H]C(=O)C1	InChI=1S/C9H10O2/c1-2-11-9-5-3-8
18	DTXSID30864192	1-Phenylpropyl butan	10031-86-4	SNUDRKNHOSAKGS-UHFFFAOYSA-N	1-Phenylpro	CCCC(=O)	InChI=1/C13H18O2/c1-3-8-13(14)15
19	DTXSID90864193	2-Ethylhept-2-enal	10031-88-6	RKQKOUYEJBHOFR-UHFFFAOYSA-N	2-Ethylhept-	CCCCC=C(	InChI=1S/C9H16O/c1-3-5-6-7-9(4-2)
20	DTXSID0064917	Ethyl 2-furanpropiona	10031-90-0	OWIWZQQFSTZZIG-UHFFFAOYSA-N	Ethyl 3-(fura	CCOC(=O)	InChI=1S/C9H12O3/c1-2-11-9(10)6-
21	DTXSID5064918	2-Nonynoic acid, ethyl	10031-92-2	BFZNMUGAZYAMTG-UHFFFAOYSA-N	Ethyl non-2-	CCCCCCC#	InChI=1S/C11H18O2/c1-3-5-6-7-8-9-

- InChIKeys are commonly used for mapping to our datasets

External Link farm

- Because InChIs have proliferated linking to many useful external sources has been simplified. Linking is based on:
  - Database_ID
  - Name
  - CASRN
  - SMILES
  - **InChIKey**



Bisphenol A
80-05-7 | DTXSID7020182
Searched by Approved Name.

General
- EPA Substance Registry Service
- PubChem
- Chemspider
- CPCat
- DrugBank
- Wikipedia
- MSDS Lookup
- ChEMBL
- ToxPlanet
- ACS Reagent Chemicals
- Wolfram Alpha
- ECHA Infocard
- ChemAgora
- Consumer Product Information Database
- ChEBI
- NIST Chemistry Webbook
- WEBWISER
- PubChem Safety Sheet
- Consumer Product Information Database

Toxicology
- ACToR
- DrugPortal
- CCRIS
- ChemView
- CTD
- eChemPortal
- Gene-Tox
- HSDB
- ACToR PDF Report
- CREST
- National Air Toxics Assessment
- ChemView
- Chemical Checker
- BindingDB
- CalEPA OEHHA
- NIOSH IDLH Values
- LactMed
- ECOTOX

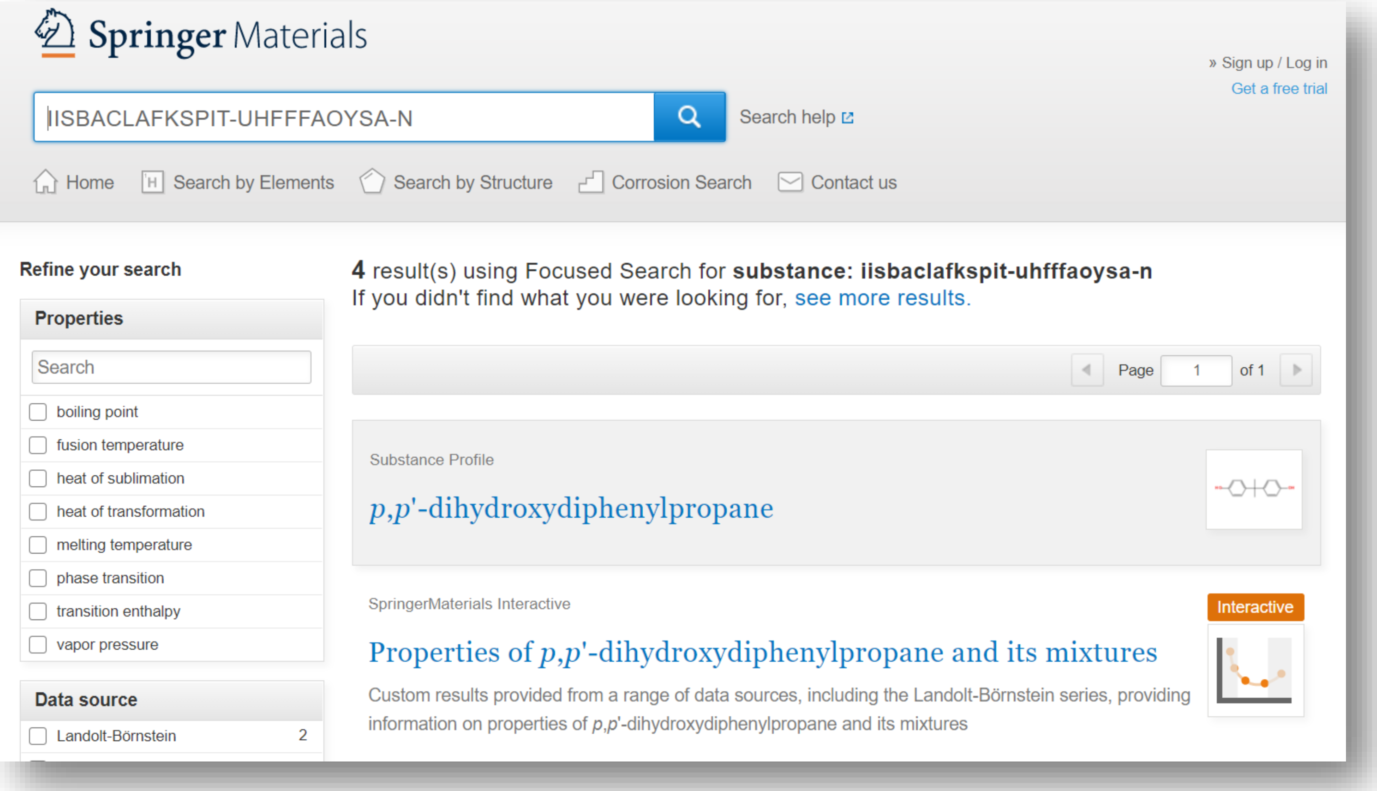
Publications
- Toxline
- PPRTVWEB
- PubMed
- IRIS Assessments
- EPA HERO
- NIOSH Skin Notation Profiles
- NIOSH Pocket Guide
- RSC Publications
- BioCaddie DataMed
- Springer Materials
- Bielefeld Academic Search Engine
- CORE Literature Search
- Google Books (Text Search)
- Google Patents (Text search)
- Google Scholar (Text search)
- Google Patents (Structure search)
- Google Books (Structure Search)
- Google Scholar (Structure search)
- Federal Register

Analytical
- RSC Analytical Abstracts
- Tox21 Analytical Data
- MONA: MassBank North America
- mzCloud
- NIST IR Spectrum
- NIST MS Spectrum
- MassBank
- NIST Antoine Constants
- IR Spectra on PubChem
- NIST Kovats Index values
- Protein DataBank
- National Environmental Methods Index

Prediction
- 2D NMR HSQC/HMBC Prediction
- Carbon-13 NMR Prediction
- Proton NMR Prediction
- ChemRTP Predictor
- LSERD

Example linkages of value

- MassBank (Europe)
- MoNA (MassBank of North America)
- Protein Databank
- ChemRTP Predictor
- Springer Materials
- Google Patents
- Google Books
- Google Scholar
- …and growing



Springer Materials

IISBACLAFKSPIT-UHFFFAOYSA-N

Refine your search

Properties

- ☐ boiling point
- ☐ fusion temperature
- ☐ heat of sublimation
- ☐ heat of transformation
- ☐ melting temperature
- ☐ phase transition
- ☐ transition enthalpy
- ☐ vapor pressure

Data source

- ☐ Landolt-Börnstein 2

4 result(s) using Focused Search for **substance: iisbaclafkspit-uhfffaoysa-n**
If you didn't find what you were looking for, see more results.

Substance Profile

p,p'-dihydroxydiphenylpropane

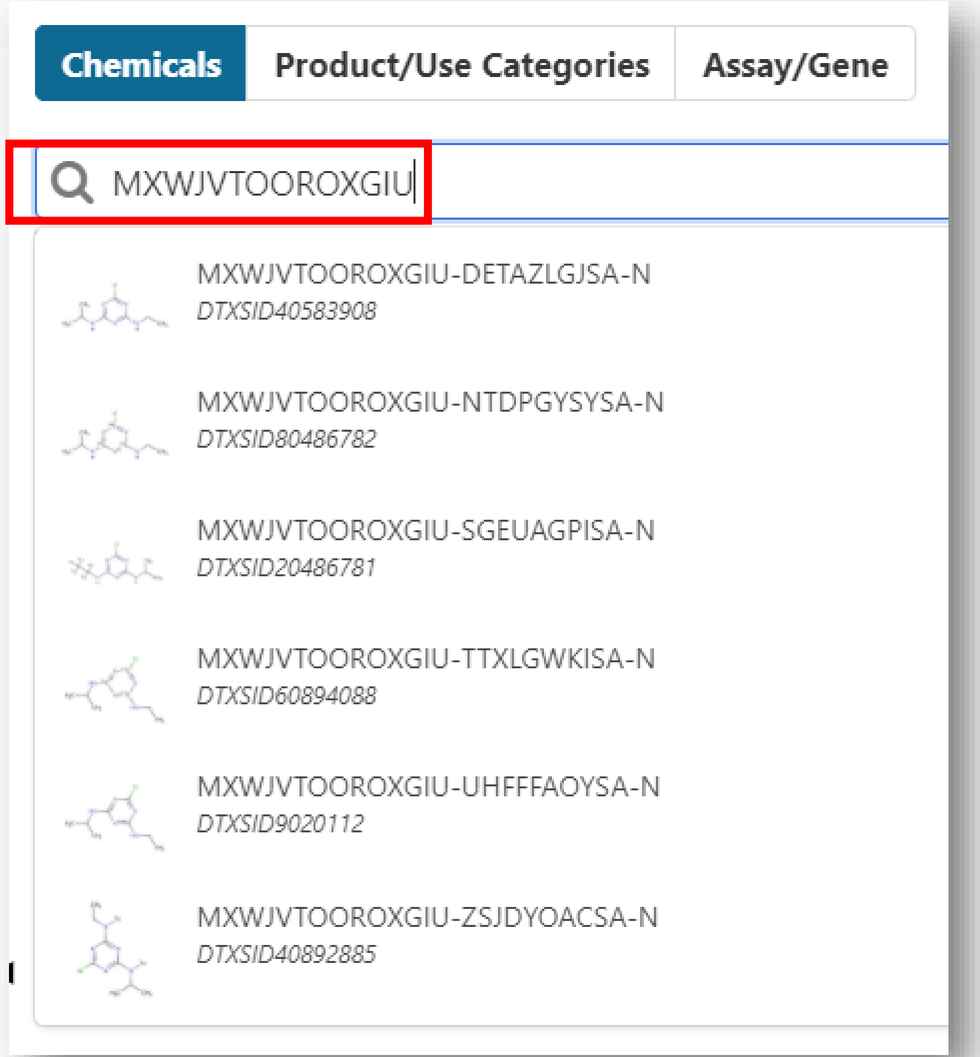
SpringerMaterials Interactive

Properties of *p,p'*-dihydroxydiphenylpropane and its mixtures

Custom results provided from a range of data sources, including the Landolt-Börnstein series, providing information on properties of *p,p'*-dihydroxydiphenylpropane and its mixtures

Single Chemical Searches

- Of course we have the basic searches…
- URL landing pages for integration
  - Full InChIkey: https://comptox.epa.gov/dashboard/dsstoxdb/results?search=MXWJVTOOROXGIU-DETAZLGJSA-N
  - Partial InChIkey: https://comptox.epa.gov/dashboard/dsstoxdb/results?search=MXWJVTOOROXGIU



MXWJVTOOROXGIU

MXWJVTOOROXGIU-DETAZLGJSA-N
DTXSID40583908

MXWJVTOOROXGIU-NTDPGYSYSA-N
DTXSID80486782

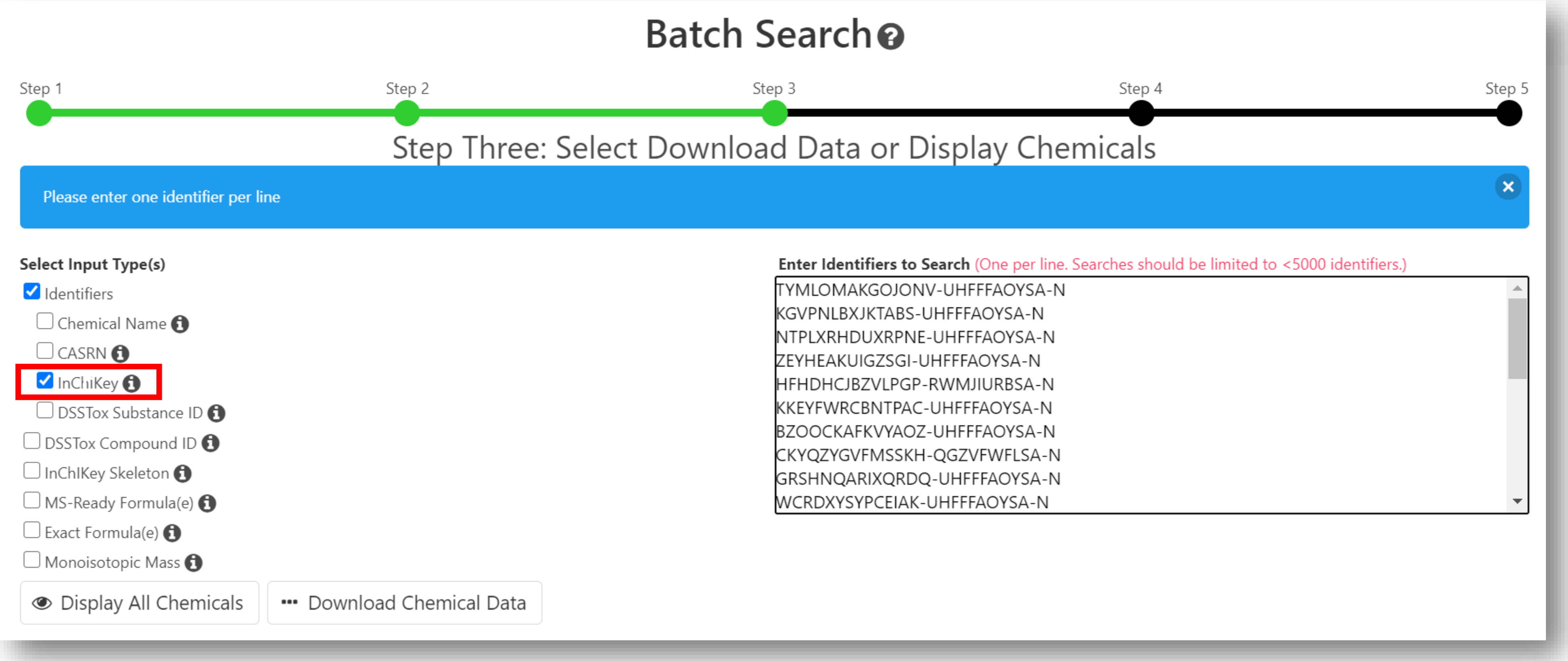
MXWJVTOOROXGIU-SGEUAGPISA-N
DTXSID20486781

MXWJVTOOROXGIU-TTXLGWKISA-N
DTXSID60894088

MXWJVTOOROXGIU-UHFFFAOYSA-N
DTXSID9020112

MXWJVTOOROXGIU-ZSJDYOACSA-N
DTXSID40892885

Batch Search – Full InChIKey search



Batch Search

Step Three: Select Download Data or Display Chemicals

Please enter one identifier per line

Select Input Type(s)

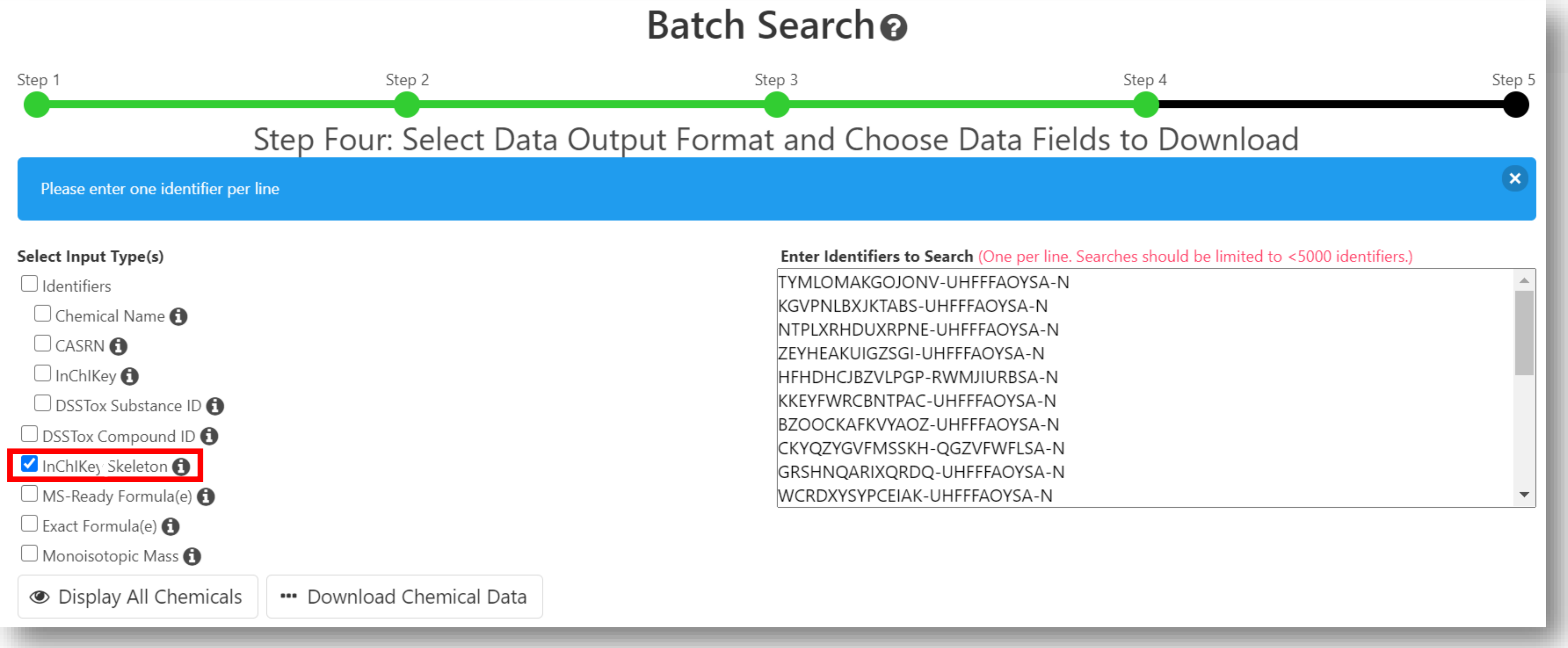
☑ Identifiers
  ☐ Chemical Name
  ☐ CASRN
  ☑ InChIKey
  ☐ DSSTox Substance ID
☐ DSSTox Compound ID
☐ InChIKey Skeleton
☐ MS-Ready Formula(e)
☐ Exact Formula(e)
☐ Monoisotopic Mass

Enter Identifiers to Search (One per line. Searches should be limited to <5000 identifiers.)

TYMLOMAKGOJONV-UHFFFAOYSA-N
KGVPNLBXJKTABS-UHFFFAOYSA-N
NTPLXRHDUXRPNE-UHFFFAOYSA-N
ZEYHEAKUIGZSGI-UHFFFAOYSA-N
HFHDHCJBZVLPGP-RWMJIURBSA-N
KKEYFWRCBNTPAC-UHFFFAOYSA-N
BZOOCKAFKVYAOZ-UHFFFAOYSA-N
CKYQZYGVFMSSKH-QGZVFWFLSA-N
GRSHNQARIXQRDQ-UHFFFAOYSA-N
WCRDXYSYPCEIAK-UHFFFAOYSA-N

…to download property and tox data

Batch Search – InChIKey skeleton



Step Four: Select Data Output Format and Choose Data Fields to Download

Please enter one identifier per line

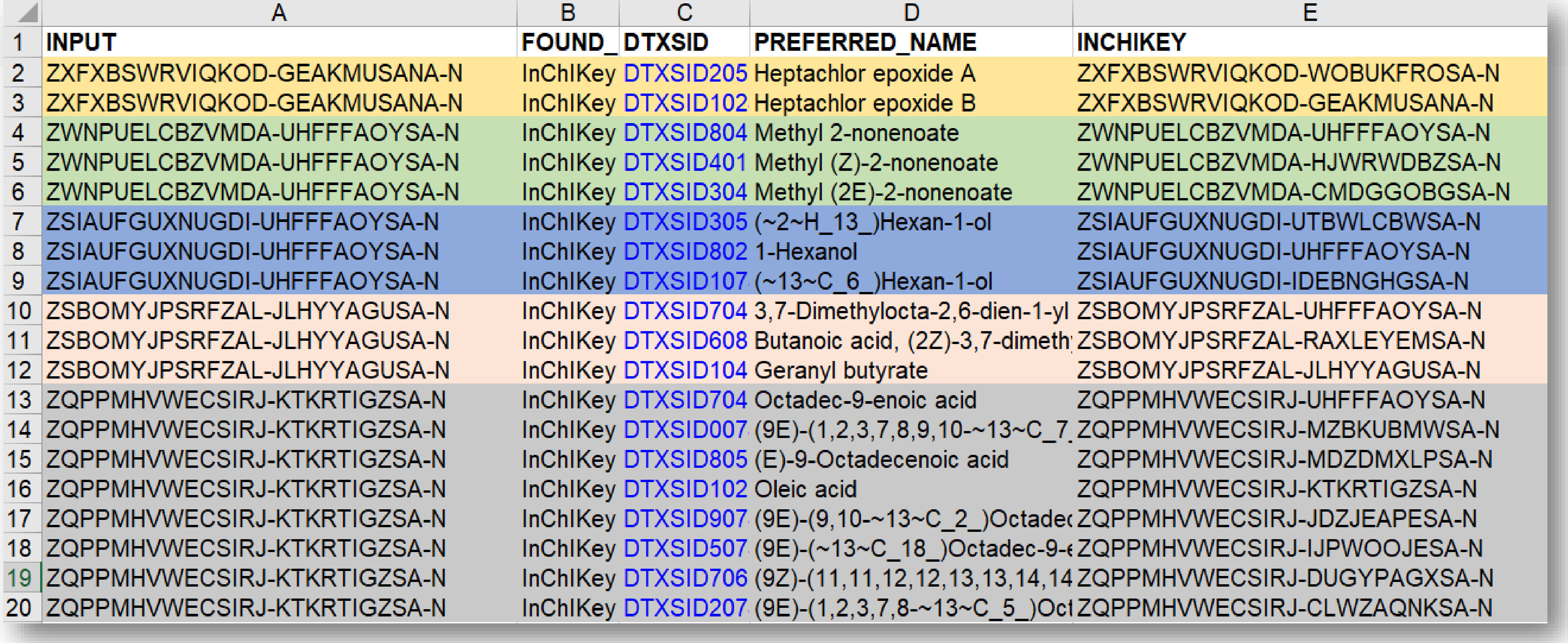
Select Input Type(s)

- ☐ Identifiers
  - ☐ Chemical Name
  - ☐ CASRN
  - ☐ InChIKey
  - ☐ DSSTox Substance ID
- ☐ DSSTox Compound ID
- ☑ InChIKey Skeleton
- ☐ MS-Ready Formula(e)
- ☐ Exact Formula(e)
- ☐ Monoisotopic Mass

Enter Identifiers to Search (One per line. Searches should be limited to <5000 identifiers.)

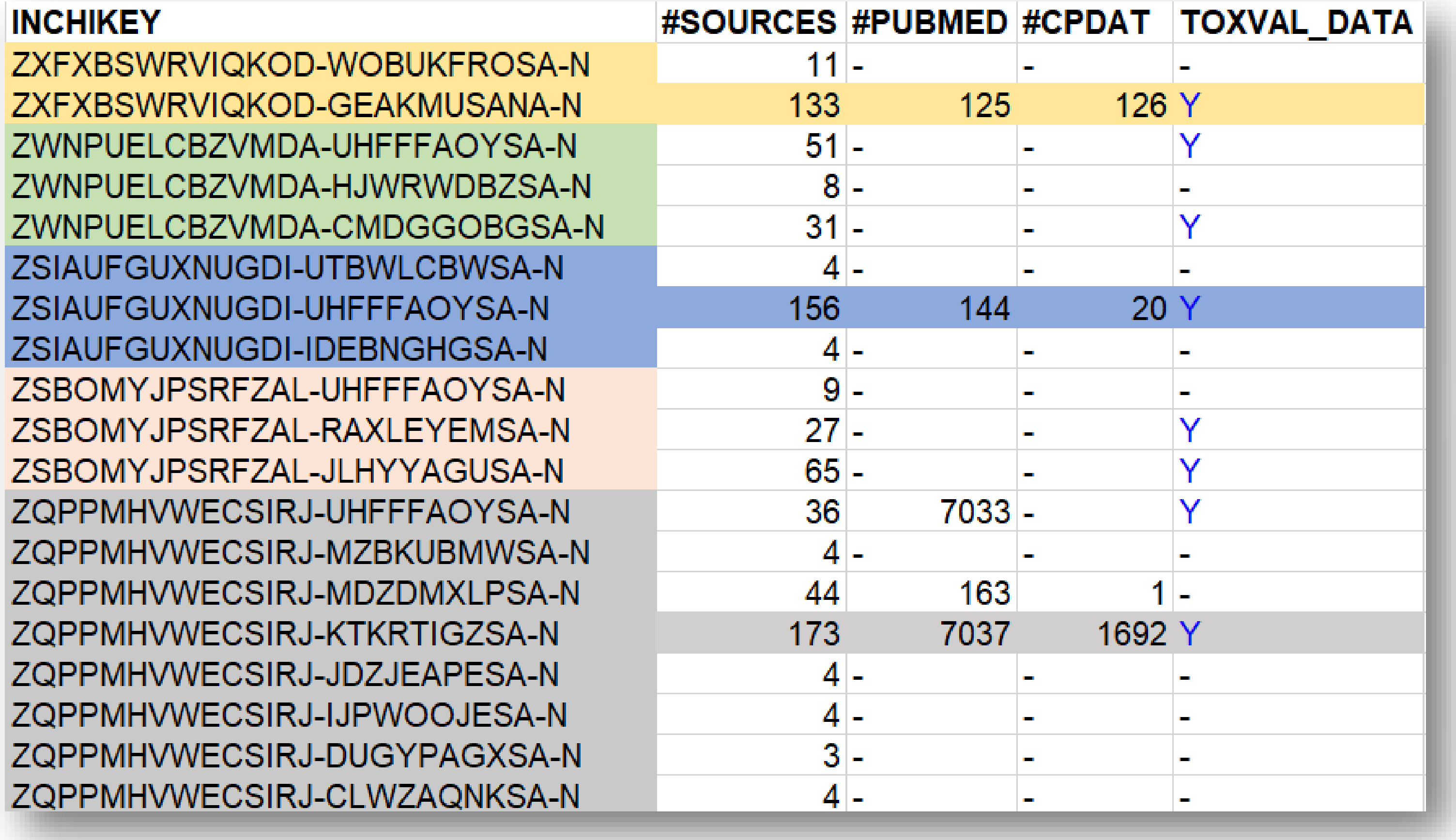
```
TYMLOMAKGOJONV-UHFFFAOYSA-N
KGVPNLBXJKTABS-UHFFFAOYSA-N
NTPLXRHDUXRPNE-UHFFFAOYSA-N
ZEYHEAKUIGZSGI-UHFFFAOYSA-N
HFHDHCJBZVLPGP-RWMJIURBSA-N
KKEYFWRCBNTPAC-UHFFFAOYSA-N
BZOOCKAFKVYAOZ-UHFFFAOYSA-N
CKYQZYGVFMSSKH-QGZVFWFLSA-N
GRSHNQARIXQRDQ-UHFFFAOYSA-N
WCRDXYSYPCEIAK-UHFFFAOYSA-N
```

InChIKey Skeleton Searches to support Non-Targeted Analysis

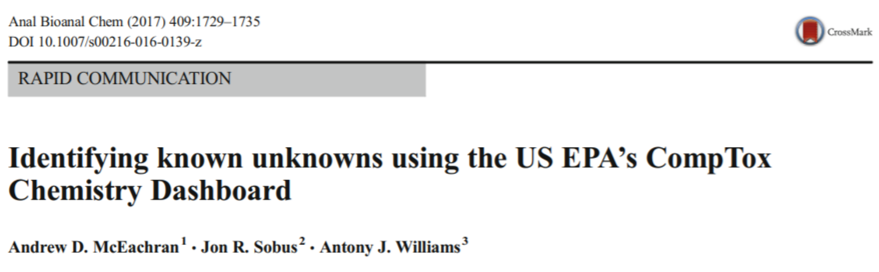


	A	B	C	D	E
1	**INPUT**	**FOUND_**	**DTXSID**	**PREFERRED_NAME**	**INCHIKEY**
2	ZXFXBSWRVIQKOD-GEAKMUSANA-N	InChIKey	DTXSID205	Heptachlor epoxide A	ZXFXBSWRVIQKOD-WOBUKFROSA-N
3	ZXFXBSWRVIQKOD-GEAKMUSANA-N	InChIKey	DTXSID102	Heptachlor epoxide B	ZXFXBSWRVIQKOD-GEAKMUSANA-N
4	ZWNPUELCBZVMDA-UHFFFAOYSA-N	InChIKey	DTXSID804	Methyl 2-nonenoate	ZWNPUELCBZVMDA-UHFFFAOYSA-N
5	ZWNPUELCBZVMDA-UHFFFAOYSA-N	InChIKey	DTXSID401	Methyl (Z)-2-nonenoate	ZWNPUELCBZVMDA-HJWRWDBZSA-N
6	ZWNPUELCBZVMDA-UHFFFAOYSA-N	InChIKey	DTXSID304	Methyl (2E)-2-nonenoate	ZWNPUELCBZVMDA-CMDGGOBGSA-N
7	ZSIAUFGUXNUGDI-UHFFFAOYSA-N	InChIKey	DTXSID305	(~2~H_13_)Hexan-1-ol	ZSIAUFGUXNUGDI-UTBWLCBWSA-N
8	ZSIAUFGUXNUGDI-UHFFFAOYSA-N	InChIKey	DTXSID802	1-Hexanol	ZSIAUFGUXNUGDI-UHFFFAOYSA-N
9	ZSIAUFGUXNUGDI-UHFFFAOYSA-N	InChIKey	DTXSID107	(~13~C_6_)Hexan-1-ol	ZSIAUFGUXNUGDI-IDEBNGHGSA-N
10	ZSBOMYJPSRFZAL-JLHYYAGUSA-N	InChIKey	DTXSID704	3,7-Dimethylocta-2,6-dien-1-yl	ZSBOMYJPSRFZAL-UHFFFAOYSA-N
11	ZSBOMYJPSRFZAL-JLHYYAGUSA-N	InChIKey	DTXSID608	Butanoic acid, (2Z)-3,7-dimeth	ZSBOMYJPSRFZAL-RAXLEYEMSA-N
12	ZSBOMYJPSRFZAL-JLHYYAGUSA-N	InChIKey	DTXSID104	Geranyl butyrate	ZSBOMYJPSRFZAL-JLHYYAGUSA-N
13	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID704	Octadec-9-enoic acid	ZQPPMHVWECSIRJ-UHFFFAOYSA-N
14	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID007	(9E)-(1,2,3,7,8,9,10-~13~C_7_	ZQPPMHVWECSIRJ-MZBKUBMWSA-N
15	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID805	(E)-9-Octadecenoic acid	ZQPPMHVWECSIRJ-MDZDMXLPSA-N
16	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID102	Oleic acid	ZQPPMHVWECSIRJ-KTKRTIGZSA-N
17	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID907	(9E)-(9,10-~13~C_2_)Octade	ZQPPMHVWECSIRJ-JDZJEAPESA-N
18	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID507	(9E)-(~13~C_18_)Octadec-9-	ZQPPMHVWECSIRJ-IJPWOOJESA-N
19	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID706	(9Z)-(11,11,12,12,13,13,14,14	ZQPPMHVWECSIRJ-DUGYPAGXSA-N
20	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	InChIKey	DTXSID207	(9E)-(1,2,3,7,8-~13~C_5_)Oc	ZQPPMHVWECSIRJ-CLWZAQNKSA-N

Then use Metadata to Rank Candidates



INCHIKEY	#SOURCES	#PUBMED	#CPDAT	TOXVAL_DATA
ZXFXBSWRVIQKOD-WOBUKFROSA-N	11	-	-	-
ZXFXBSWRVIQKOD-GEAKMUSANA-N	133	125	126	Y
ZWNPUELCBZVMDA-UHFFFAOYSA-N	51	-	-	Y
ZWNPUELCBZVMDA-HJWRWDBZSA-N	8	-	-	-
ZWNPUELCBZVMDA-CMDGGOBGSA-N	31	-	-	Y
ZSIAUFGUXNUGDI-UTBWLCBWSA-N	4	-	-	-
ZSIAUFGUXNUGDI-UHFFFAOYSA-N	156	144	20	Y
ZSIAUFGUXNUGDI-IDEBNGHGSA-N	4	-	-	-
ZSBOMYJPSRFZAL-UHFFFAOYSA-N	9	-	-	-
ZSBOMYJPSRFZAL-RAXLEYEMSA-N	27	-	-	Y
ZSBOMYJPSRFZAL-JLHYYAGUSA-N	65	-	-	Y
ZQPPMHVWECSIRJ-UHFFFAOYSA-N	36	7033	-	Y
ZQPPMHVWECSIRJ-MZBKUBMWSA-N	4	-	-	-
ZQPPMHVWECSIRJ-MDZDMXLPSA-N	44	163	1	-
ZQPPMHVWECSIRJ-KTKRTIGZSA-N	173	7037	1692	Y
ZQPPMHVWECSIRJ-JDZJEAPESA-N	4	-	-	-
ZQPPMHVWECSIRJ-IJPWOOJESA-N	4	-	-	-
ZQPPMHVWECSIRJ-DUGYPAGXSA-N	3	-	-	-
ZQPPMHVWECSIRJ-CLWZAQNKSA-N	4	-	-	-

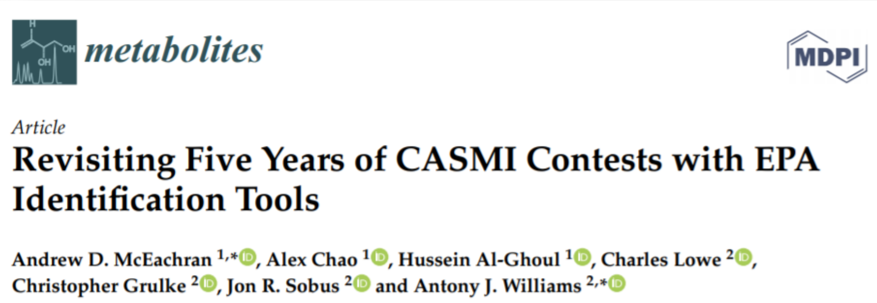


Anal Bioanal Chem (2017) 409:1729–1735
DOI 10.1007/s00216-016-0139-z

RAPID COMMUNICATION

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

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



metabolites

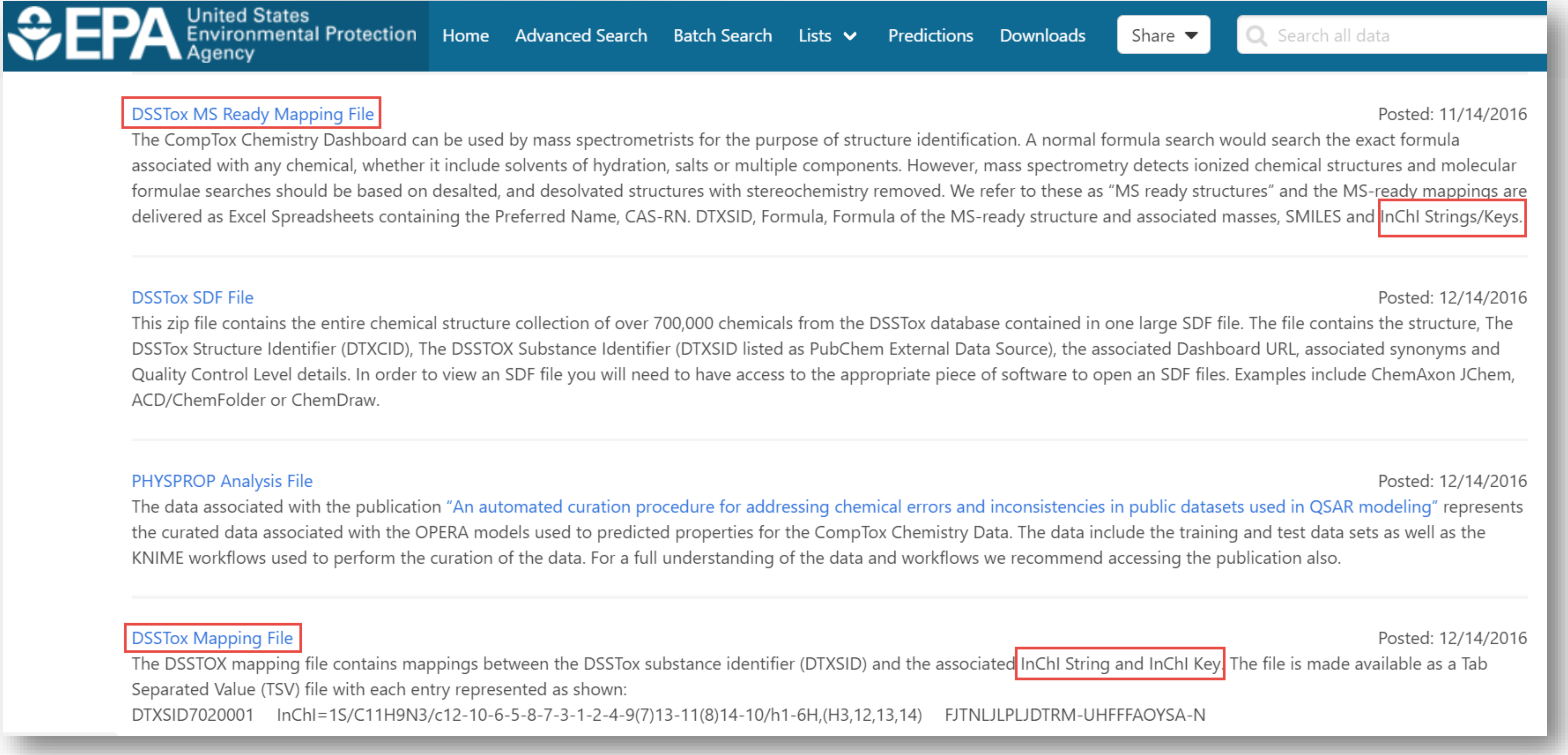
Article

Revisiting Five Years of CASMI Contests with EPA Identification Tools

Andrew D. McEachran [1,*], Alex Chao [1], Hussein Al-Ghoul [1], Charles Lowe [2], Christopher Grulke [2], Jon R. Sobus [2] and Antony J. Williams [2,*]

Download Files

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



DSSTox MS Ready Mapping File

Posted: 11/14/2016

The CompTox Chemistry Dashboard can be used by mass spectrometrists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formulae searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as “MS ready structures” and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN. DTXSID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys.

DSSTox SDF File

Posted: 12/14/2016

This zip file contains the entire chemical structure collection of over 700,000 chemicals from the DSSTox database contained in one large SDF file. The file contains the structure, The DSSTox Structure Identifier (DTXCID), The DSSTOX Substance Identifier (DTXSID listed as PubChem External Data Source), the associated Dashboard URL, associated synonyms and Quality Control Level details. In order to view an SDF file you will need to have access to the appropriate piece of software to open an SDF files. Examples include ChemAxon JChem, ACD/ChemFolder or ChemDraw.

PHYSPROP Analysis File

Posted: 12/14/2016

The data associated with the publication “An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling” represents the curated data associated with the OPERA models used to predicted properties for the CompTox Chemistry Data. The data include the training and test data sets as well as the KNIME workflows used to perform the curation of the data. For a full understanding of the data and workflows we recommend accessing the publication also.

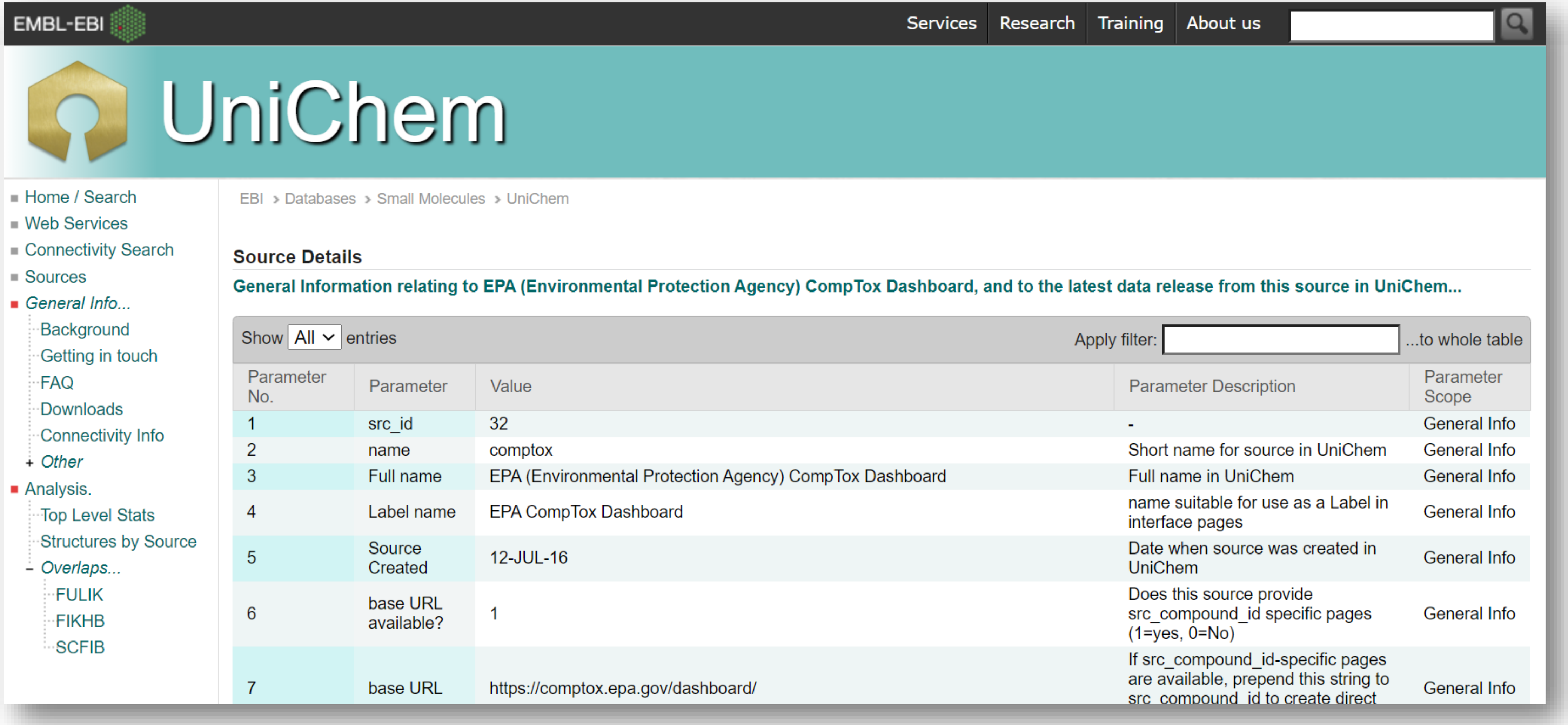
DSSTox Mapping File

Posted: 12/14/2016

The DSSTOX mapping file contains mappings between the DSSTox substance identifier (DTXSID) and the associated InChI String and InChI Key. The file is made available as a Tab Separated Value (TSV) file with each entry represented as shown:

```
DTXSID7020001     InChI=1S/C11H9N3/c12-10-6-5-8-7-3-1-2-4-9(7)13-11(8)14-10/h1-6H,(H3,12,13,14)     FJTNLJLPLJDTRM-UHFFFAOYSA-N
```

Openly Sharing Data



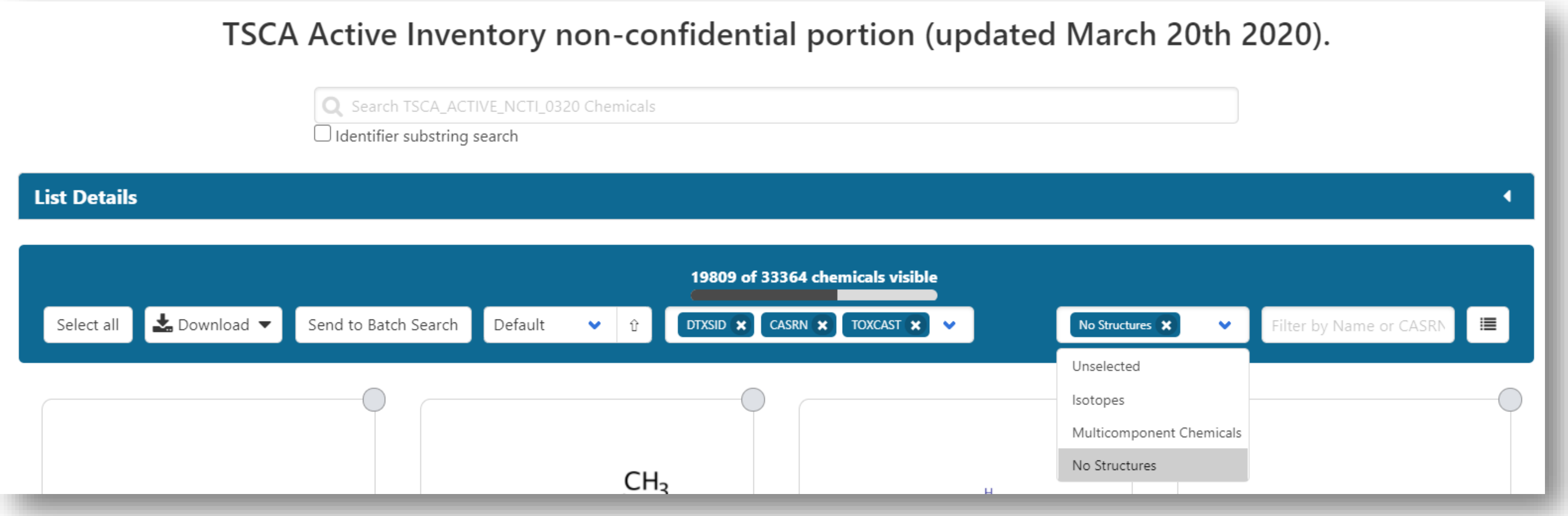
Source Details

General Information relating to EPA (Environmental Protection Agency) CompTox Dashboard, and to the latest data release from this source in UniChem...

Parameter No.	Parameter	Value	Parameter Description	Parameter Scope
1	src_id	32	-	General Info
2	name	comptox	Short name for source in UniChem	General Info
3	Full name	EPA (Environmental Protection Agency) CompTox Dashboard	Full name in UniChem	General Info
4	Label name	EPA CompTox Dashboard	name suitable for use as a Label in interface pages	General Info
5	Source Created	12-JUL-16	Date when source was created in UniChem	General Info
6	base URL available?	1	Does this source provide src_compound_id specific pages (1=yes, 0=No)	General Info
7	base URL	https://comptox.epa.gov/dashboard/	If src_compound_id-specific pages are available, prepend this string to src_compound_id to create direct	General Info

TSCA Inventory – 40% no structures

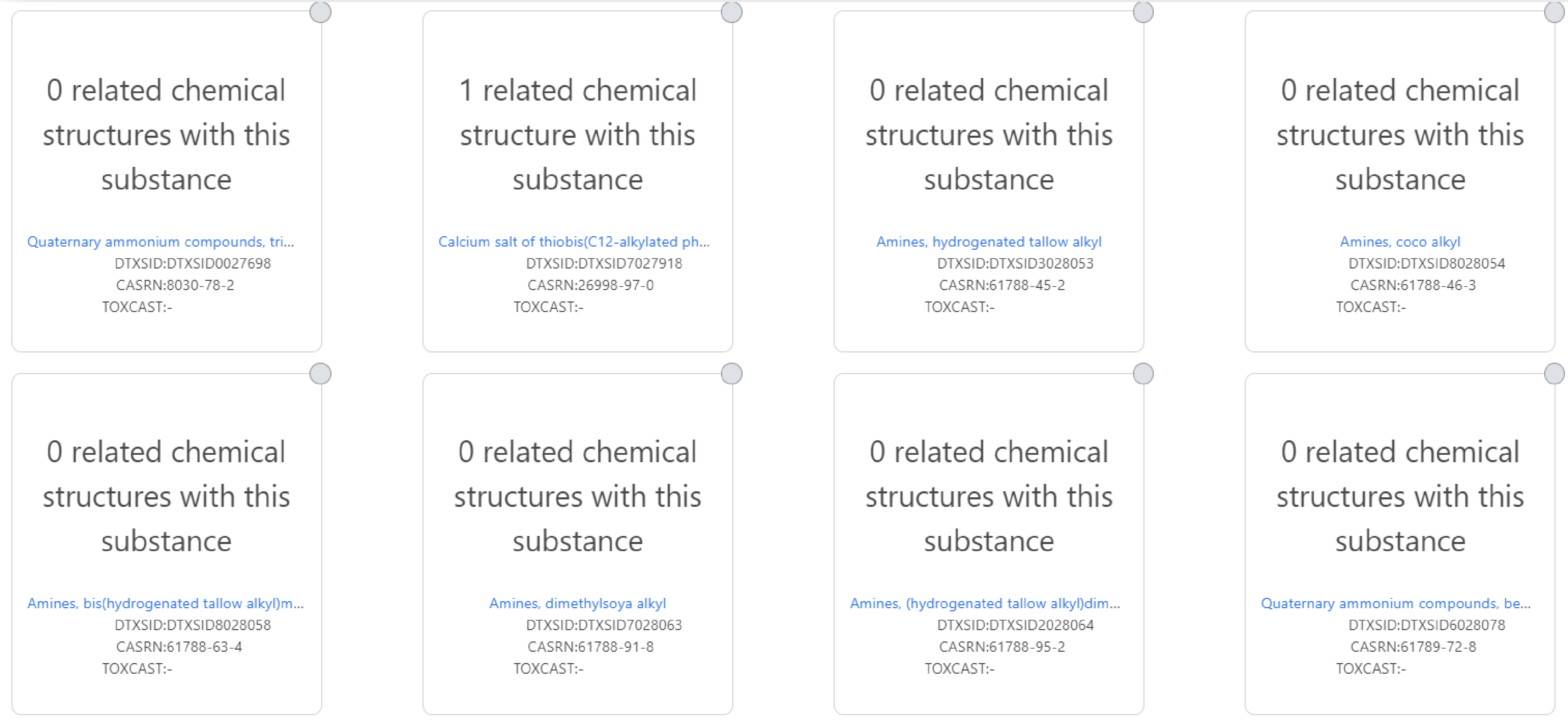
- Many of the substances we deal with at EPA are not structures
- The TSCA active inventory is ~40% non-structurable



TSCA Active Inventory non-confidential portion (updated March 20th 2020).

19809 of 33364 chemicals visible

TSCA-related substances



0 related chemical structures with this substance

Quaternary ammonium compounds, tri...
DTXSID:DTXSID0027698
CASRN:8030-78-2
TOXCAST:-

1 related chemical structure with this substance

Calcium salt of thiobis(C12-alkylated ph...
DTXSID:DTXSID7027918
CASRN:26998-97-0
TOXCAST:-

0 related chemical structures with this substance

Amines, hydrogenated tallow alkyl
DTXSID:DTXSID3028053
CASRN:61788-45-2
TOXCAST:-

0 related chemical structures with this substance

Amines, coco alkyl
DTXSID:DTXSID8028054
CASRN:61788-46-3
TOXCAST:-

0 related chemical structures with this substance

Amines, bis(hydrogenated tallow alkyl)m...
DTXSID:DTXSID8028058
CASRN:61788-63-4
TOXCAST:-

0 related chemical structures with this substance

Amines, dimethylsoya alkyl
DTXSID:DTXSID7028063
CASRN:61788-91-8
TOXCAST:-

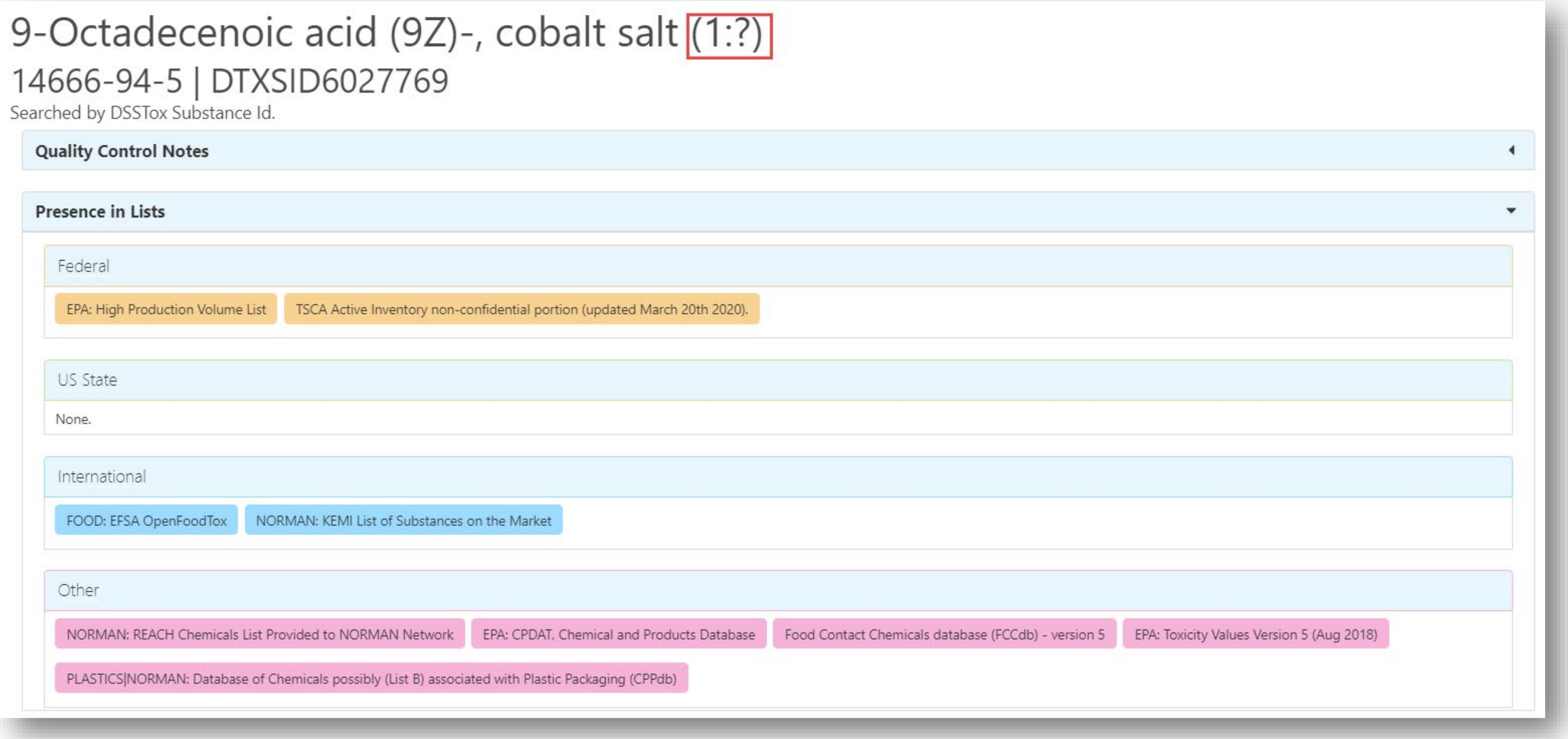
0 related chemical structures with this substance

Amines, (hydrogenated tallow alkyl)dim...
DTXSID:DTXSID2028064
CASRN:61788-95-2
TOXCAST:-

0 related chemical structures with this substance

Quaternary ammonium compounds, be...
DTXSID:DTXSID6028078
CASRN:61789-72-8
TOXCAST:-

Ambiguous stoichiometry
Thousands of chemicals



9-Octadecenoic acid (9Z)-, cobalt salt (1:?)

14666-94-5 | DTXSID6027769

Searched by DSSTox Substance Id.

Federal

EPA: High Production Volume List

TSCA Active Inventory non-confidential portion (updated March 20th 2020).

US State

None.

International

FOOD: EFSA OpenFoodTox

NORMAN: KEMI List of Substances on the Market

Other

NORMAN: REACH Chemicals List Provided to NORMAN Network

EPA: CPDAT, Chemical and Products Database

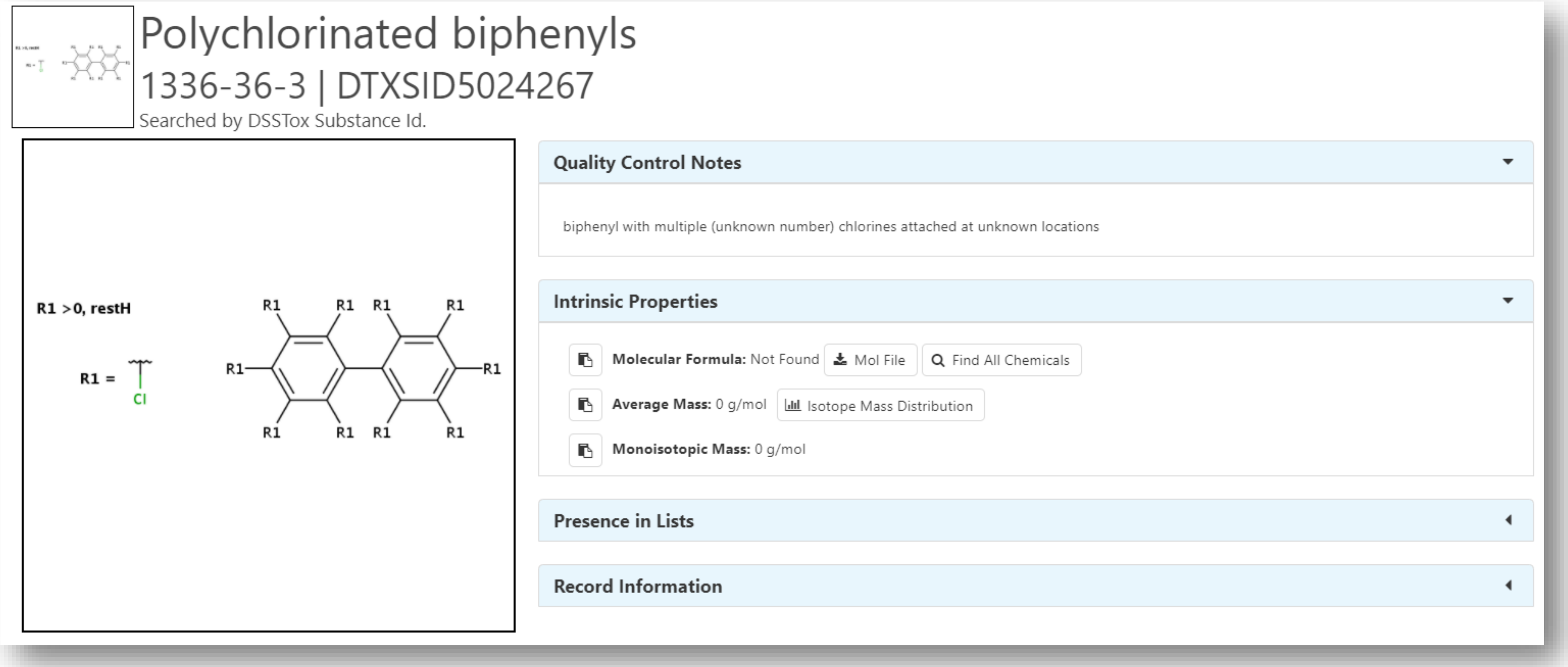
Food Contact Chemicals database (FCCdb) - version 5

EPA: Toxicity Values Version 5 (Aug 2018)

PLASTICS|NORMAN: Database of Chemicals possibly (List B) associated with Plastic Packaging (CPPdb)

Complex chemistry – Markush

- We are ready to test and provide feedback



Polychlorinated biphenyls

1336-36-3 | DTXSID5024267

Searched by DSSTox Substance Id.

Quality Control Notes

biphenyl with multiple (unknown number) chlorines attached at unknown locations

Intrinsic Properties

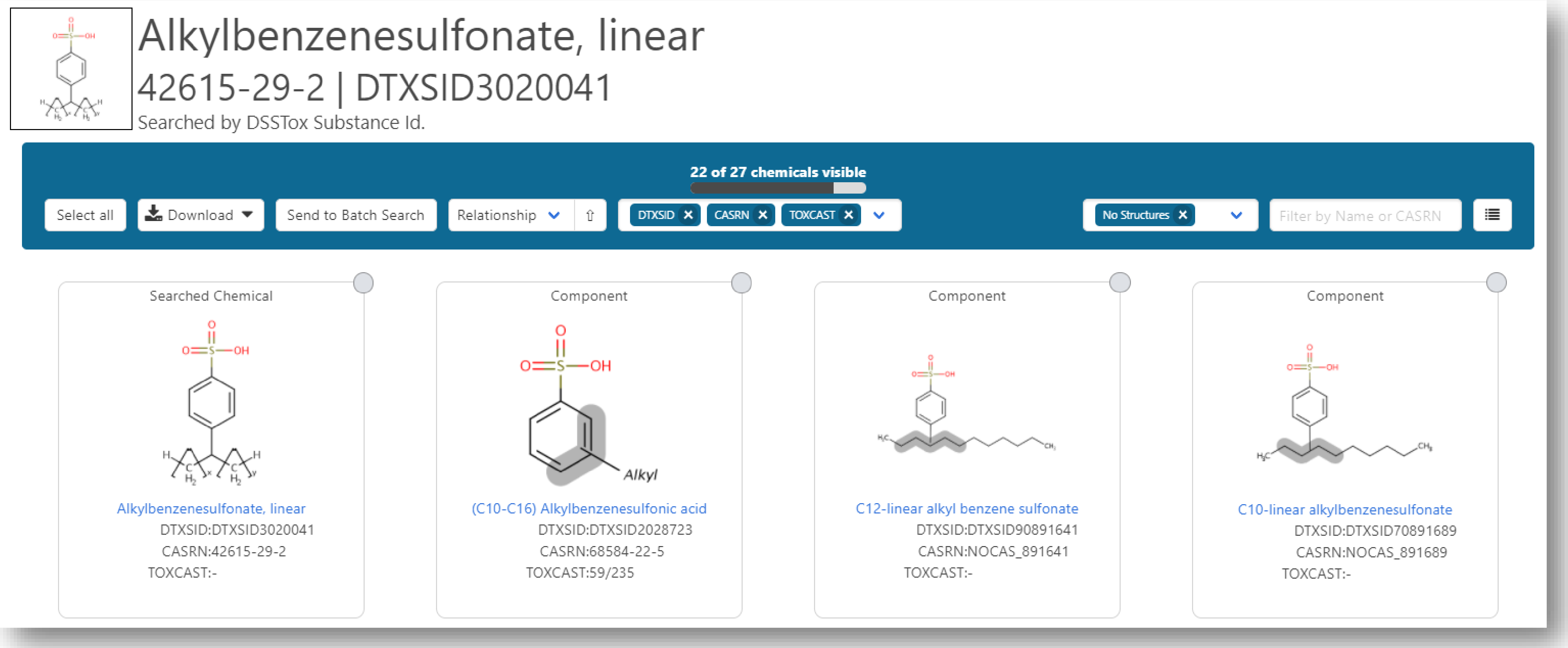
Molecular Formula: Not Found

Average Mass: 0 g/mol

Monoisotopic Mass: 0 g/mol

Complex Chemistry - Markush

- We have appropriate structural representations – but no InChIs



Alkylbenzenesulfonate, linear
42615-29-2 | DTXSID3020041
Searched by DSSTox Substance Id.

Searched Chemical
Alkylbenzenesulfonate, linear
DTXSID:DTXSID3020041
CASRN:42615-29-2
TOXCAST:-

Component
(C10-C16) Alkylbenzenesulfonic acid
DTXSID:DTXSID2028723
CASRN:68584-22-5
TOXCAST:59/235

Component
C12-linear alkyl benzene sulfonate
DTXSID:DTXSID90891641
CASRN:NOCAS_891641
TOXCAST:-

Component
C10-linear alkylbenzenesulfonate
DTXSID:DTXSID70891689
CASRN:NOCAS_891689
TOXCAST:-

Our contribution to teaching

- The majority of our users accessing the dashboard over the past few years had not heard of InChIs
- We take the opportunity to educate our user community about the utility of InChIs
- We encourage InChIs in Supplementary Information Files
- We are sharing InChIs from our system with other EPA internal systems to their advantage

Conclusion

- Thank you, thank you, thank you…
- InChI identifiers are very useful to our efforts – we depend on it…
  - …in our registration and curation processes
  - …for searching single chemicals and batches of chemicals
  - …for mappings within the application
  - …for linking to third-party websites
  - …for registration into resolver databases
  - …to add to our data exports for database mapping
  - …as default information in many of our download files
- We are anxiously awaiting support for our complex chemistry challenges – polymers, organometallics, mixtures, Markush

Acknowledgments

- Contact: Williams.Antony@epa.gov
- Feedback and follow-up is welcomed! Your questions help
- The dashboard is based on the efforts of many more team members than us
- Many collaborators provide data

EPA's Center for Computational Toxicology and Exposure