Mapping of chemical identifiers to DSSTox to enable data integration in the US-EPA CompTox Chemicals Dashboard

Christopher Grulke[1]
Inthirany Thillainadarajah[1,2]
Patience Brown[3]
Antony Williams[1]
Ann Richard[1]

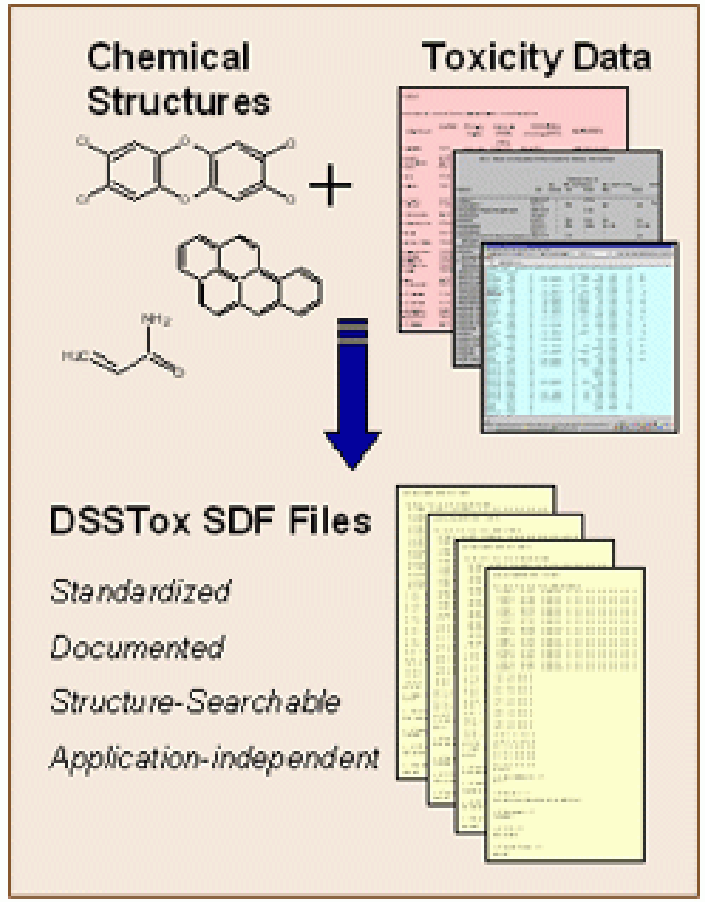
1. National Center of Computational Toxicology, U.S. EPA
2. Senior Environmental Employment Program, U.S. EPA
3. OECD/EHS, Paris France

American Chemical Society Meeting, Spring 2019
3 April 2019, Orlando FL

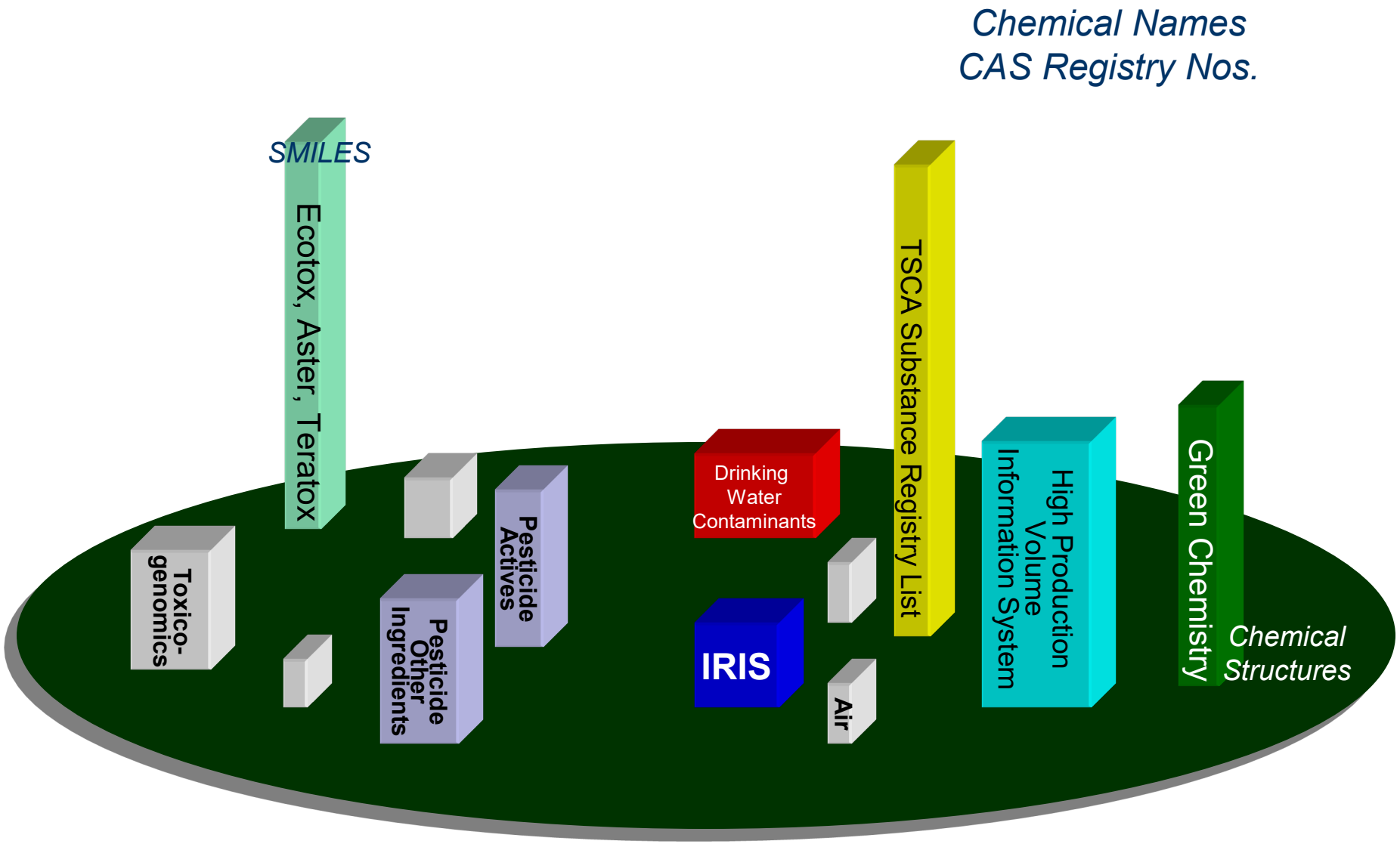
DSSTox Background

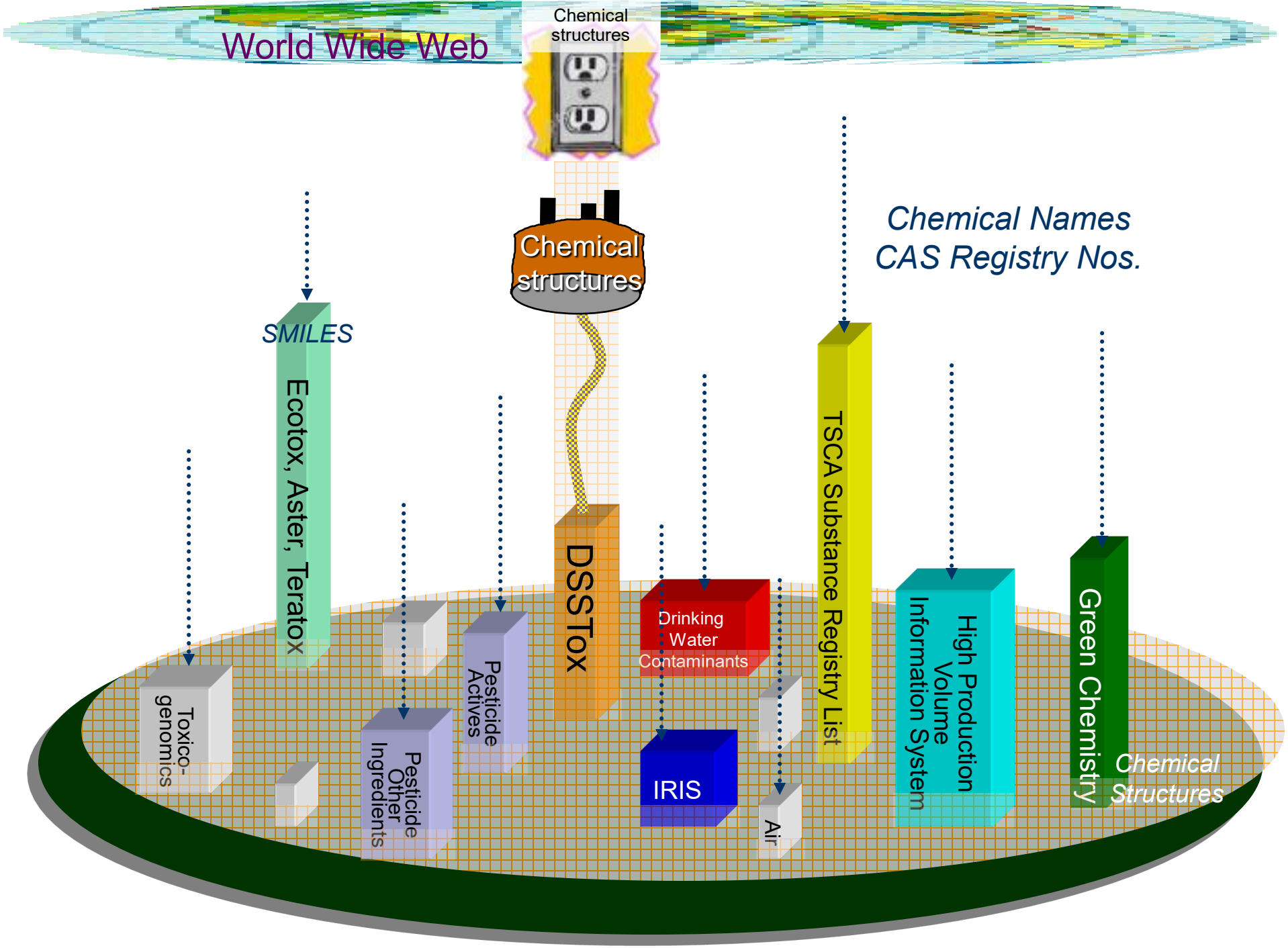
Goal: Linking data to chemical structures enabling SAR

- First release of data files in 2004
- Focused on high impact sets of data
  - Carcinogenic Potency Database
  - Drinking water disinfection by-products
  - EPA's Integrated Risk Information System
  - FDA's Maximum Daily Dose dataset
  - EPA's Fat Head Minnow Toxicity dataset
  - ToxCast and Tox21 chemicals
- Currently contains: 876K records (32K manually curated)
- Check it out: https://comptox.epa.gov/dashboard

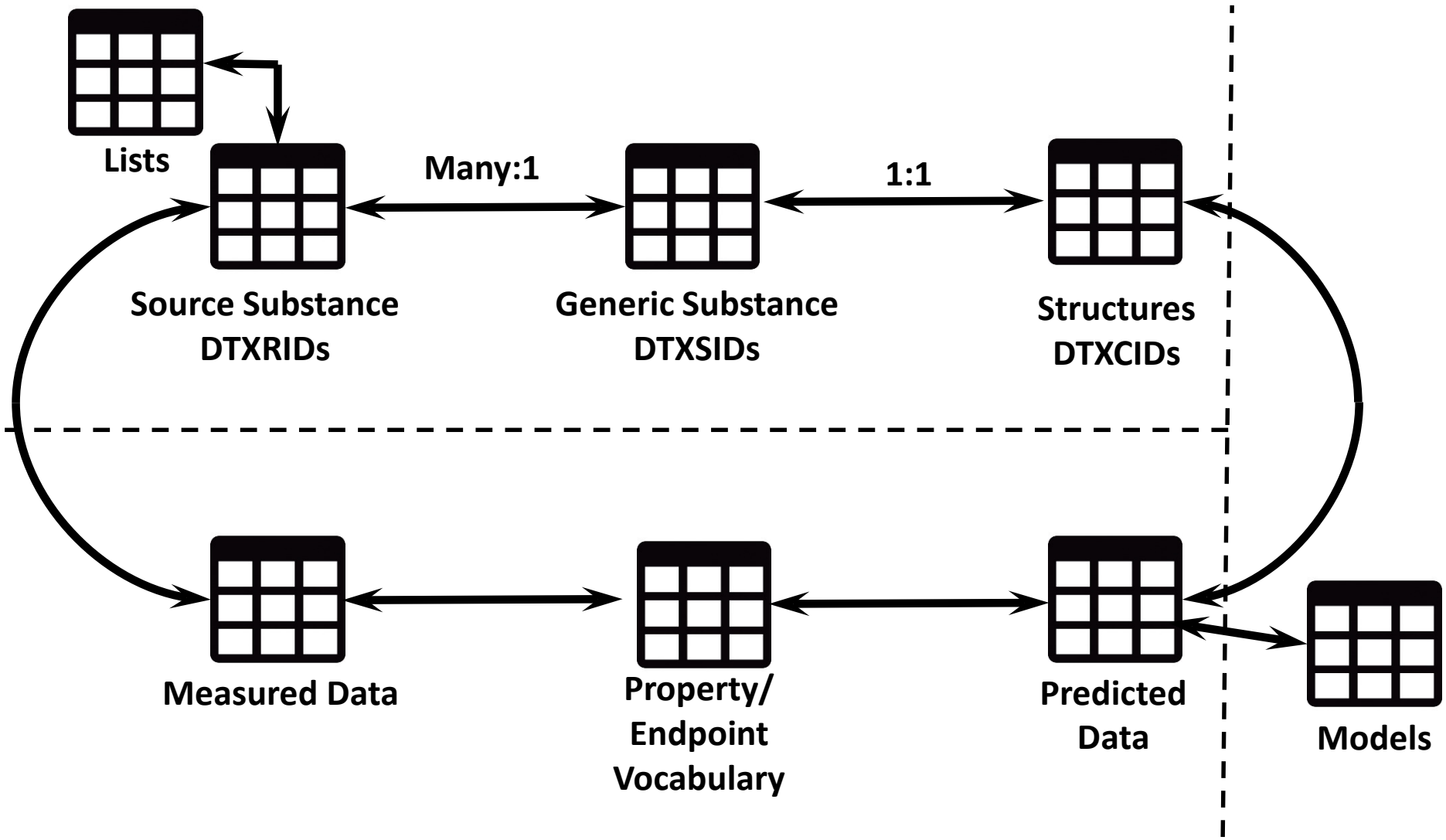


EPA's data islands … circa 2000



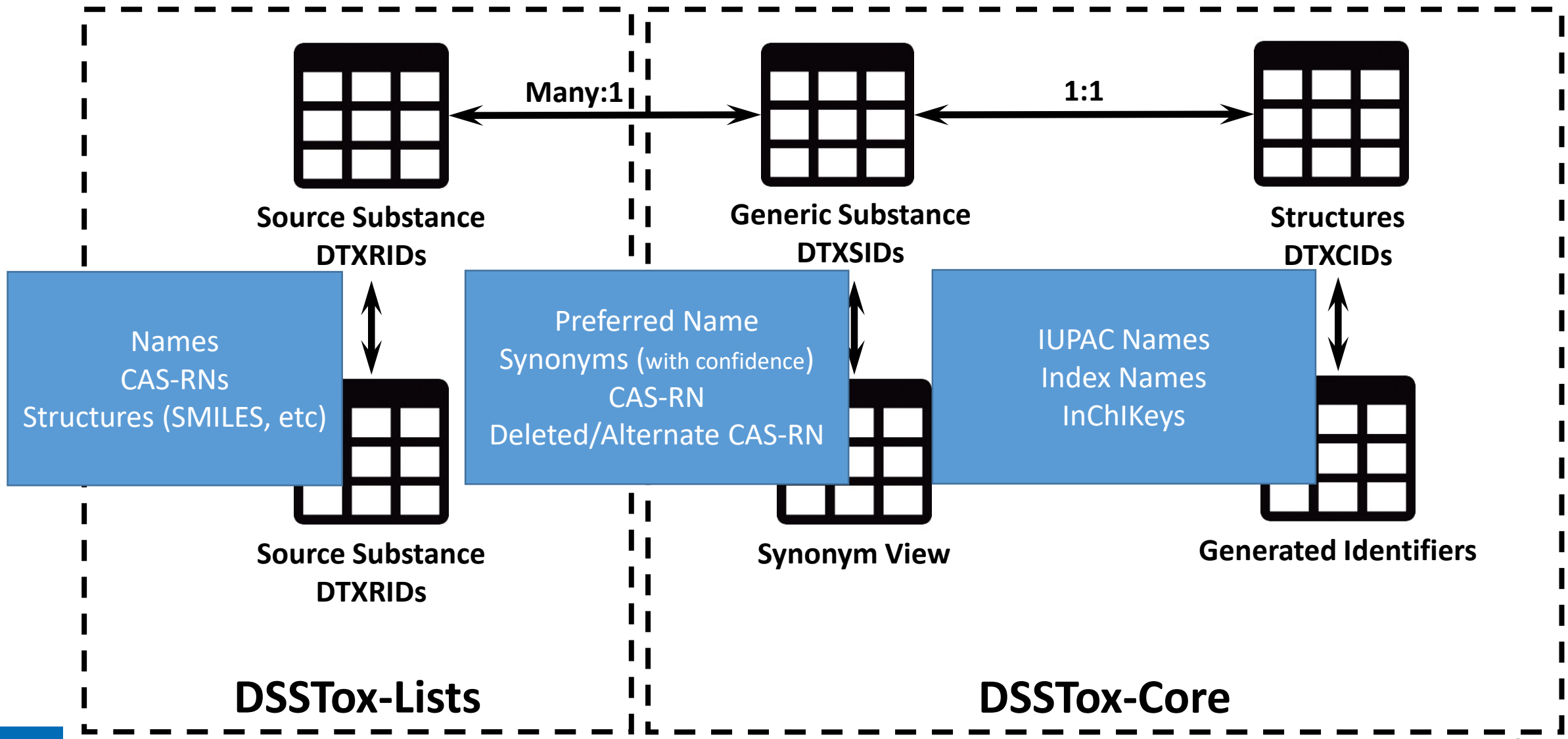

World Wide Web
Chemical structures
Chemical structures
Chemical Names
CAS Registry Nos.
SMILES
Ecotox, Aster, Teratox
DSSTox
Drinking Water Contaminants
TSCA Substance Registry List
High Production Volume Information System
Green Chemistry
Chemical Structures
Toxico-genomics
Pesticide Actives
Pesticide Other Ingredients
IRIS
Air

Data linkage in DSSTox



5

Chemical Identifiers in DSSTox



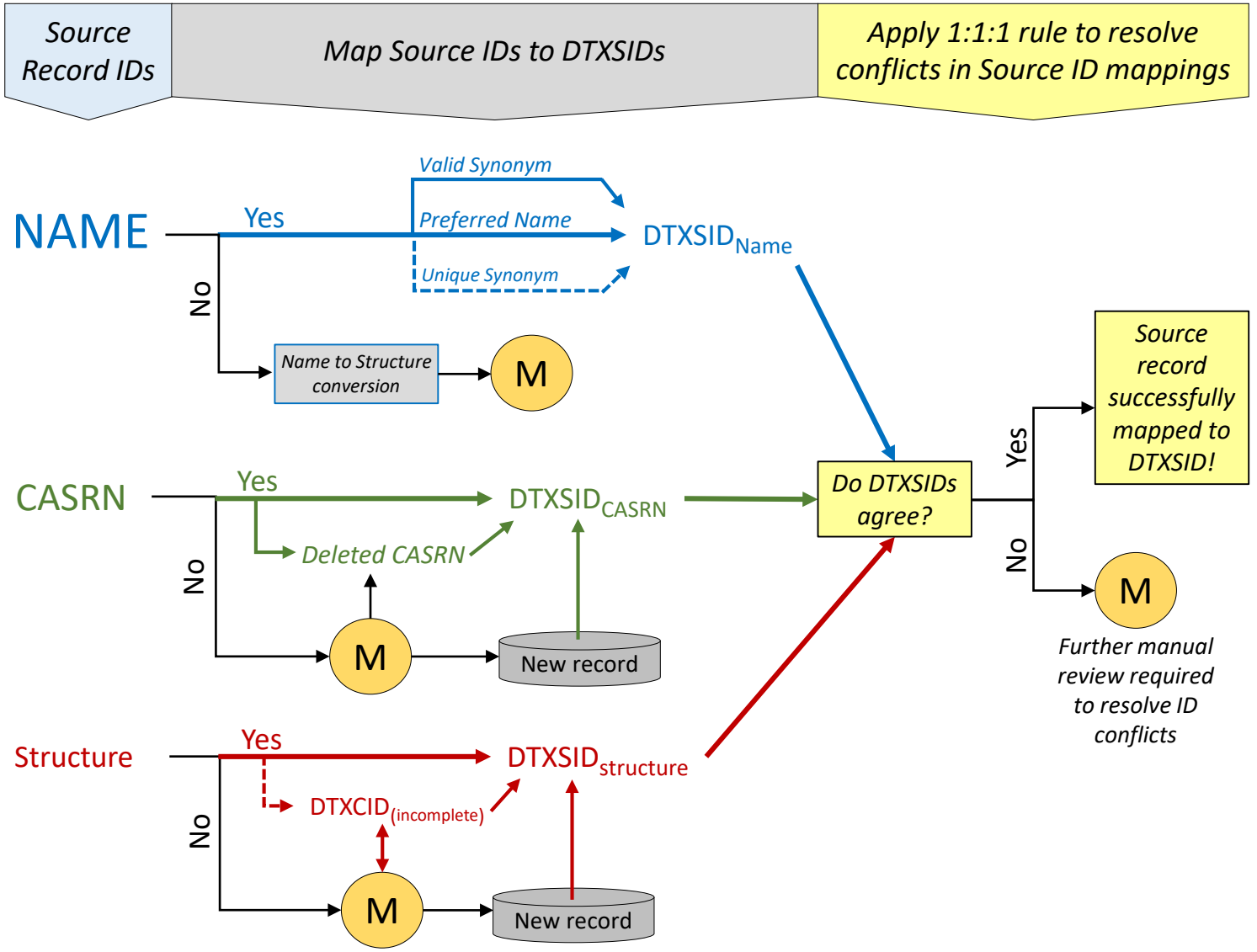


Figure 6. Schematic illustrating general DSSTox curation workflow for processing a set of Source chemical IDs for a single input record in which each Source ID is either mapped to an existing DTXSID, or requires manual curator review (M) and may lead to registration of a new DTXSID record: 1) NAME is most frequently provided with Source lists and, typically, is most closely associated with the original data record; 2) CASRNs are often provided with Source lists, but are most often collated by the Source from public records, so tend to be less reliable; 3) Structure is most often not included with Source lists, but when provided by, e.g., chemical suppliers, tends to be the least reliable identifier in relation to the original Source record. Bold lines indicate curation steps amenable to automated processing; M indicates expert manual review is required.

Scoring mapped hits

```
Map<String, Double> prefixes = new HashMap<~>();
//Simple recall
prefixes.put("DTXSID matched",  1.0);
prefixes.put("DTXCID matched",   1.0);
prefixes.put("DTXRID matched",   1.0);
prefixes.put("DTXSID not found",  0.0);
prefixes.put("DTXCID not found",    0.0);
prefixes.put("DTXRID not found",  0.0);
prefixes.put("InChI not found",  0.0);
prefixes.put("GSID matched", 0.9);
prefixes.put("CID matched", 0.6);
prefixes.put("RID matched", 0.6);
prefixes.put("Structure matched", 1.0);
prefixes.put("Structure connectivity matched", 0.3);
prefixes.put("Structure not found", 0.0);
prefixes.put("CAS-RN matched", 0.75);
prefixes.put("Other CAS-RN matched", 0.65);
prefixes.put("CAS-RN not found", 0.0);
prefixes.put("Preferred Name matched", 0.8);
prefixes.put("Valid Synonym matched", 0.72);
prefixes.put("Unique Synonym matched", 0.55);
prefixes.put("Ambiguous Synonym matched", 0.35);
prefixes.put("Mapped Identifier matched", 0.15);
prefixes.put("Name2Structure matched", 0.34);
prefixes.put("Name2Structure connectivity matched", 0.14);
prefixes.put("Identifier not found", 0.0);
prefixes.put("DTXID format error", 0.0);

PREFIXES = Collections.unmodifiableMap(prefixes);
```

```
Map<String, Double> penalty = new HashMap<~>();
//Simple recall
penalty.put("zero-stripped",  0.201);
penalty.put("stereo-corrected",   0.201);
penalty.put("formatted",  0.401);
penalty.put(null,  0.0);

PENALTY = Collections.unmodifiableMap(penalty);
```

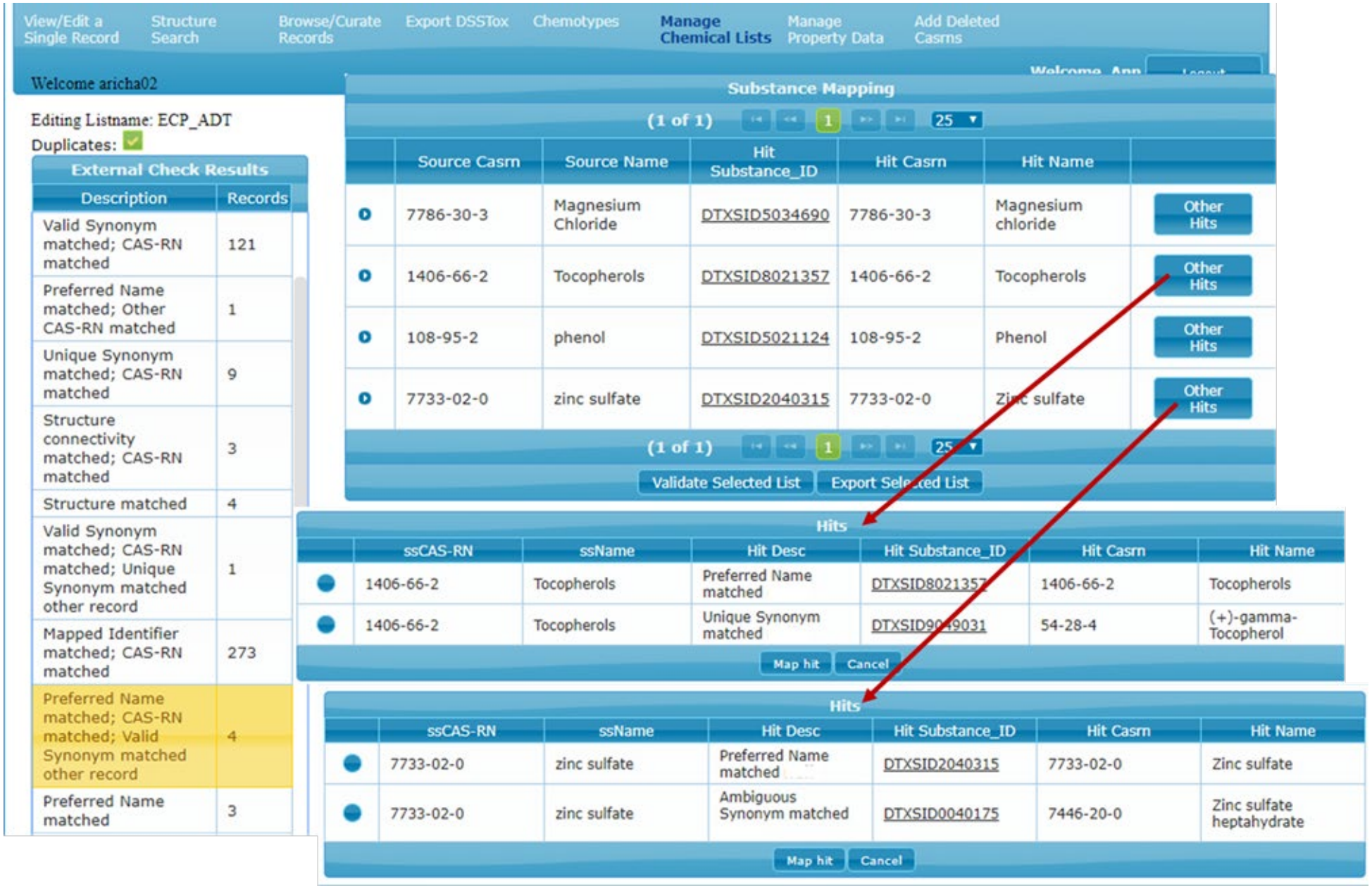
$$MatchScore_{Hit=DTXSID_j}^{Query=SSI_i} = Prefix_{i,j} - Penalty_{i,j}$$

$$HitScore_{Hit=DTXSID_j} = \sum_{i=1}^{n} MatchScore_{Hit=DTXSID_j}^{Query=SSI_i}$$

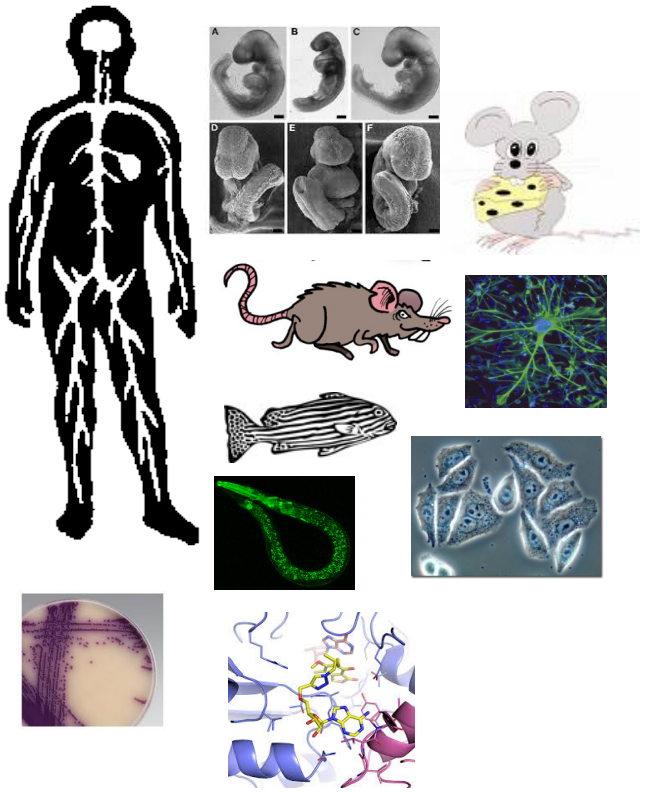
$$MapScore_{Hit=DTXSID_j}$$

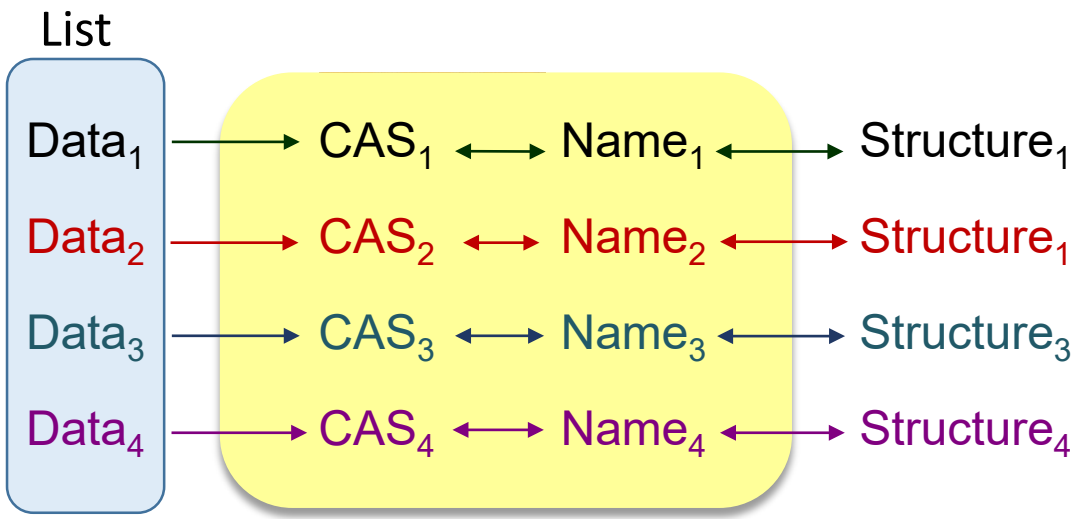
$$= HitScore_{Hit=DTXSID_j} - \sum_{k<>j}^{n} HitScore_{Hit=DTXSID_k}$$

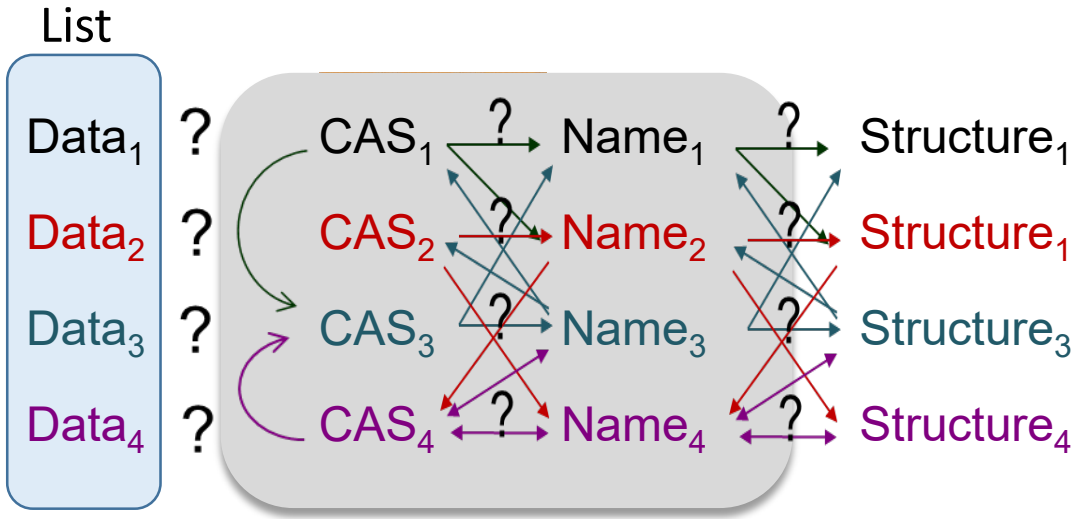
DSSTox Chemical Lists



Errors in Data-Structure Linkages







EPA
United States
Environmental Protection
Agency

Case 1: TSCA Chemical Substance Inventory

Download the non-confidential TSCA Inventory

EPA provides a Microsoft Access version and a generic comma-delimited "CSV" text version of the non-confidential TSCA Inventory for users to download. If you don't have Microsoft Access, use the CSV file. Both files are compressed ".zip" files. The .zip files contain the actual data files. See generic information about zip files.

- A Microsoft Access file (ZIP) (4 MB) (Last created: 03/15/2019)
- A Comma Separated Value text file (ZIP) (2 MB) (Last created: 03/15/2019)

The Microsoft Access file contains two tables:

- TSCAINV_032019
  - Contains non-confidential chemical substance listings on the TSCA Inventory, as identified by Chemical Abstract Service (CAS) Registry Number and Chemical Abstracts (CA) Index Name.
- PMNACC_032019

TSCA Non-Confidential Inventory Input

Release 2/20/2019

Records: ~60K

Active: 32898

Curator Validated	23856
CASRN and Name Match	39
CASRN Match, Name questionable	8298
Conflicted Mappings	105
No Hits	600

id	CASRN	casregno	ChemName	DF	UV	FL	CS
10588	9/4/2851	2851094	Benzoxazole, 2-(1-piperidinyl)-				INACTIVE
10589	2851-94-7	2851947	1H-Imidazole, 1-ethenyl-2-phenyl-				INACTIVE
10590	2851-95-8	2851958	1H-Imidazole, 1-ethenyl-2-methyl-				ACTIVE

TSCAINV_022019.csv

```
1591  7522,1038-66-0,1038660,"[1,1'-Biphenyl]-4,4'-diamine, 2,2',3,3',5,5',6,6'-octafluoro-",,,,ACTIVE
1592  7523,1038-95-5,1038955,"Phosphine, tris(4-methylphenyl)-",,,,ACTIVE
1593  7524,1041-00-5,1041005,"Benzoxazole, 2,2'-(1,2-ethenediyl)bis[5-methyl-",,,,ACTIVE
1594  7525,1042-84-8,1042848,"Ethanone, 2-(1-methylnaphtho[1,2-d]thiazol-2(1H)-ylidene)-1-phenyl-",,,,
1595  7526,1046-56-6,1046566,"1,2,4-Triazine, 5,6-diphenyl-3-(2-pyridinyl)-",,,,ACTIVE
1596  7527,1047-16-1,1047161,"Quino[2,3-b]acridine-7,14-dione, 5,12-dihydro-",,,,ACTIVE
1597  7528,1048-05-1,1048051,"Germane, tetraphenyl-",,,,ACTIVE
1598  7529,1048-08-4,1048084,"Benzene, 1,1',1'',1'''-silanetetrayltetrakis-",,,,ACTIVE
1599  7530,1052-38-6,1052386,"1,3-Benzenediamine, 4,4'-[1,3-phenylenebis(2,1-diazenediyl)]bis-",,,,ACTI
1600  7531,1058-71-5,1058715,"1,2,4-Triazine, 5,6-diphenyl-3-(4-phenyl-2-pyridinyl)-",,,,INACTIVE
1601  7532,1058-92-0,1058920,"2,7-Naphthalenedisulfonic acid, 3-[2-(5-chloro-2-hydroxyphenyl)diazenyl]-
1602  7533,1062-96-0,1062960,"Cholest-5-en-3-ol (3.beta.)-, hexanoate",,,,INACTIVE
1603  10578,4/6/2836,2836046,"1,3-Benzenediamine, N1,N1-dimethyl-",,,,INACTIVE
1604  10579,2836-32-0,2836320,"Acetic acid, 2-hydroxy-, sodium salt (1:1)",,,,ACTIVE
1605  10580,2837-89-0,2837890,"Ethane, 2-chloro-1,1,1,2-tetrafluoro-",,,PMN,ACTIVE
1606  10581,2840-00-8,2840008,"4-Pyridinamine, 3,5-dichloro-2,6-difluoro-",,,PMN; S; 5E,INACTIVE
1607  10582,2840-28-0,2840280,"Benzoic acid, 3-amino-4-chloro-",,,PMN,ACTIVE
1608  10583,2842-44-6,2842446,"Ethanol, 2-[methyl(4-methylphenyl)amino]-",,,PMN,ACTIVE
1609  10584,2845-89-8,2845898,"Benzene, 1-chloro-3-methoxy-",,,,INACTIVE
1610  10585,11/9/2846,2846119,"2,7-Naphthalenedisulfonic acid, 5-ethoxy-4-hydroxy-3-[2-[4-(6-methyl-2-
1611  10586,2847-16-7,2847167,"Decanoic acid, cadmium salt (2:1)",,,,INACTIVE
1612  10587,2847-30-5,2847305,"Pyrazine, 2-methoxy-3-methyl-",,,,ACTIVE
1613  10588,9/4/2851,2851094,"Benzoxazole, 2-(1-piperidinyl)-",,,,INACTIVE
1614  10589,2851-94-7,2851947,"1H-Imidazole, 1-ethenyl-2-phenyl-",,,,INACTIVE
1615  10590,2851-95-8,2851958,"1H-Imidazole, 1-ethenyl-2-methyl-",,,,ACTIVE
1616  10591,2855-13-2,2855132,"Cyclohexanemethanamine, 5-amino-1,3,3-trimethyl-",,,,ACTIVE
```

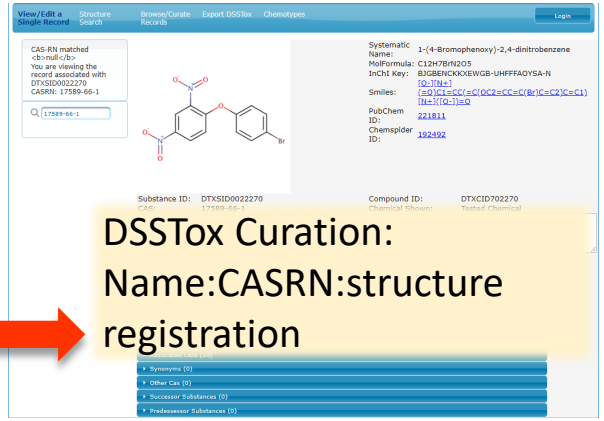
Registering “No Hits” from Trusted CASRN:Name Source

DSSTox
“No Hits”

No match to:
Names
Synonyms
Source names
CASRNs
Structures

Expert visual review

Structurable



DSSTox Curation:
Name:CASRN:structure registration

Keyword filters:
“poly”, “deriv”
“rxn”, “reaction”

Not structurable

Auto-load:
Name:CASRN registration

Reviewing/Validating Mapped Records

ACToR-DSSTox Chemical Registration

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

Welcome cgrulke

Editing Listname:
TSCA_ACTIVE_NCTI
Duplicates: ✓

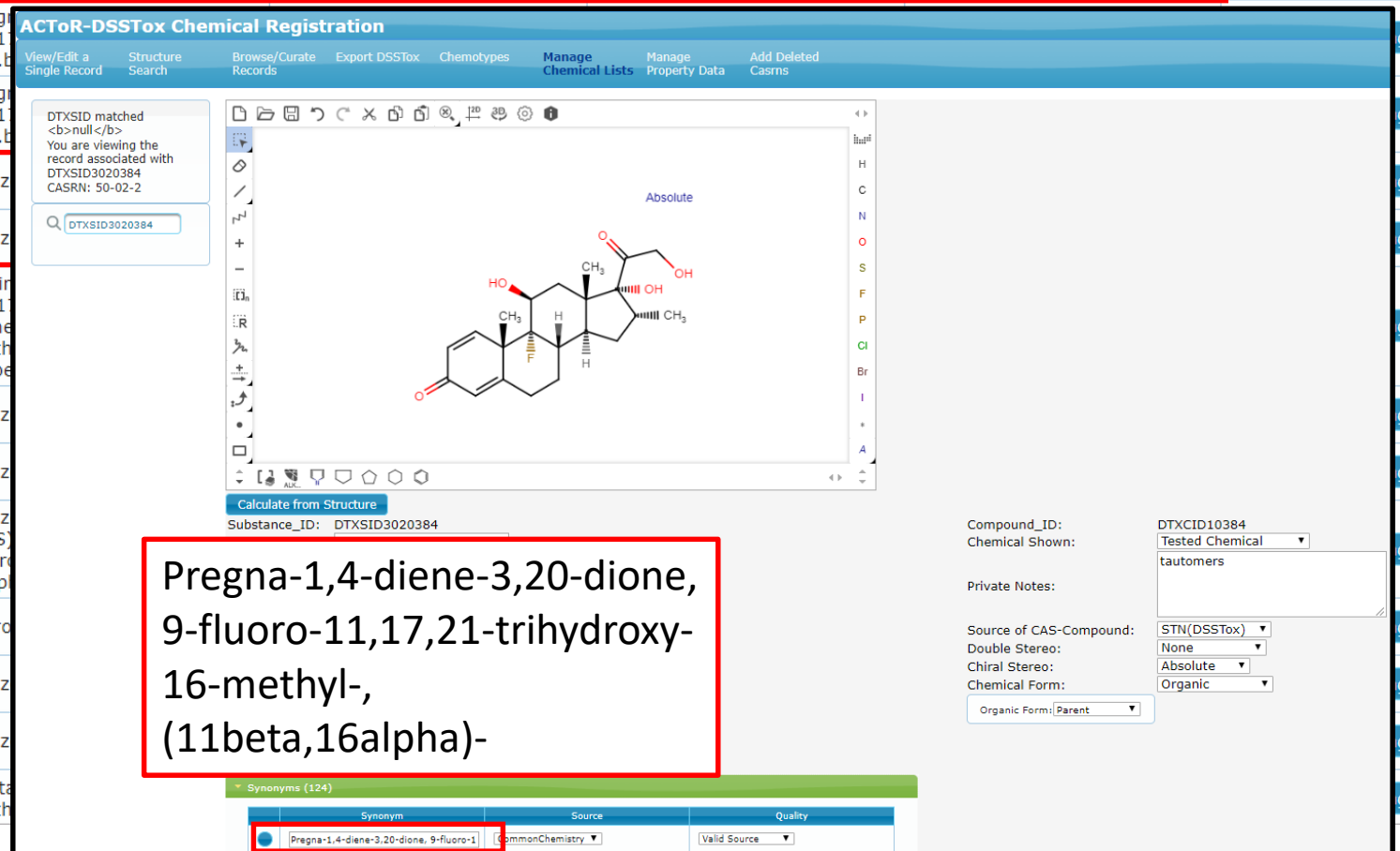
External Check Results

Description	Records
CAS-RN matched **casregno** Name2Structure matched **ChemName**	736
CAS-RN matched **casregno** Preferred Name matched **ChemName** Unique Synonym matched other record: **ChemName**	19
Other CAS-RN matched **casregno** Name2Structure matched **ChemName**	2
CAS-RN matched **casregno** Valid Synonym matched **ChemName** Unique Synonym matched other record: **ChemName**	3
CAS-RN matched **casregno** Mapped Identifier matched **ChemName**	4279
CAS-RN matched **casregno** Preferred Name matched **ChemName** Ambiguous	3

Substance Mapping

(1 of 172) 1 2 3 4 5 6 7 8 9 10 25

Source Casrn	Source Name	Hit Substance_ID	Hit Casrn	Hit Name	
50022	Pregna-1,4-diene-3,20-dione, 9-fluoro-11,17,21-trihydroxy-16-methyl-, (11.beta.,16.alpha.)-	DTXSID3020384	50-02-2	Dexamethasone	Validate Mapping
50237	Pregn 11,1 (11.b				g
50248	Pregn 11,1 (11.b				g
50306	Benz				g
50453	Benz				g
50555	Yohin 11,1 trime meth (3.be				g
50793	Benz				g
50840	Benz				g
50986	Benz [(1S) hydr (.alp				g
51354	L-Pro				g
51365	Benz				g
51445	Benz				g
51661	Aceta meth				g



Pregna-1,4-diene-3,20-dione, 9-fluoro-11,17,21-trihydroxy-16-methyl-, (11beta,16alpha)-

EPA
United States
Environmental Protection
Agency

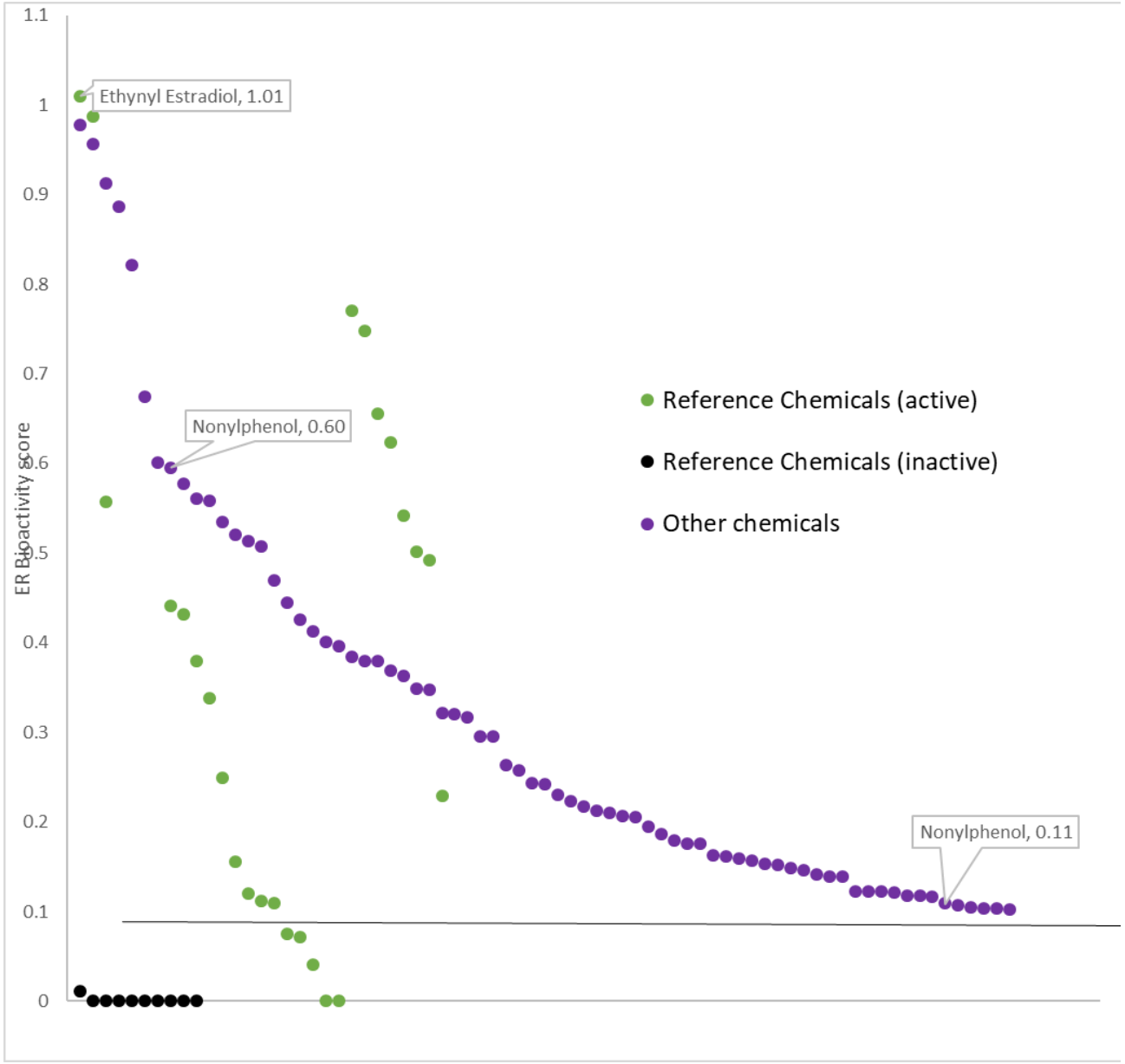
Case 2: Endocrine Disruption Reference Data

Endocrine Disruption Reference Set Background

- Interest in identifying reference chemicals for:
  - Validating assays
    - Large set of structurally diverse chemicals across a range of potencies define applicability domain and dynamic response of assays
  - Linking endocrine mechanisms of action with adverse in vivo responses
    - Explicit or implicit regulatory requirement for identifying “endocrine disruptors”
  - Evaluating the performance of non-animal alternative methods
    - Are the same chemicals active in vitro as in vivo
- Reference chemicals were identified from:
  - Systematic literature reviews (e.g. Kleinstreuer et al. 2016, 2017; Browne et al. 2018)
  - Validation of existing methods (e.g. OECD validation reports; ICCVAM validation reports)
  - Scoping documents (e.g. OECD No. 207)

Pre-Chemical Curation Problems

- Many different chemical names for the same substance
- Many different substances for the same chemical name
  - may have very different bioactivities (e.g. “nonylphenol”)



Endocrine Disruption Reference Input 1

Received: 6/21/2017

Records: ~250

Mapping:

~140 Good Mapping

~110 Questionable/No Mapping

CASRN	Chemical	Pathway	ER	AR	thyroid
	N,N,N',N'-tetramethylthiourea	T			TPO inhibition
	11-ketotestosterone				
57-91-0	17a-Estradiol	EA	agonist	agonist (negative)	
80657-17-6	17a-trenbolone	A		AR agonist	
50-28-2	17beta-Estradiol	EAT	ER agonist	AR antagonist	effect TR transactivation
10161-33-8	17b-Trenbolone	A			
57-63-6	17-ethynyl estradiol	E	ER agonist; ER antagonist (negative)		
434-22-0	19-Nortestosterone	Androgenic			
72-55-9	2,2-bis(4-chlorophenyl)-1,1-dichloroethylene (p,p'-DDE)	EA	ER agonist	AR antagonist	
778-22-3	**2,2-diphenylpropane**	Estrogenic			
93-76-5	2,4,5-tricholorphenoxyacetic acid				

DSSTox 2014

DSSTox_GSID	DSSTox_CID	TS_ChemName	TS_ChemName_Synonyms
20001	1	A-alpha-C	
39224	2	Acetaldehyde	Acetaldehyde (Ethanal)
39225	3	Acetaldehyde methylformylhydrazone	
20004	4	Acetaldehyde oxime	Acetaldoxime
20005	5	Acetamide	
20006	6	Acetaminophen	4-hydroxyacetanilide; 4-Acetamidophenol; APAP; Paracetamol

Endocrine Disruption Reference Set Examples

Other specific questions:

26037-38-3 Nonoxynol-9

The CASRN is for an unspecific mixture of Nonoxynols. Nonoxynol-9 has a defined structure and is linked to the casrn 14409-72-4

Perfluorooctane (PFOS)

Perfluorooctane and PFOS are different

68359-37-5 b-cyfluthrin

If b is short for beta, it is a specific set of stereoisomers (with its own casrn) whereas the casrn provided is for completely unspecified stereochemistry

52315-07-8 b-cypermethrin

If b is short for beta, it is a specific set of stereoisomers (with its own casrn) whereas the casrn provided is for completely unspecified stereochemistry

From list:

102676-31-3 Fadrozole

102676-47-1 Fadrozole

Explanation: casrn 102676-31-3 is actually the HCl salt of fadrazole.

From list:

6153-64-6 Oxytetracycline

Explanation: casrn 6153-64-6 is a dihydrate form of Oxytetracycline

Endocrine Disruption Reference Set Examples (Cont…)

You have the name "6-OH-BDE-47" with CASRN "1017894-49-3". We believe this CAS-RN to be better represented by the name "6'-OH-BDE-85" (if you want to use this naming convention). Please see here: https://comptox.epa.gov/dashboard/DTXSID30873880. If you like the name as is, I would consider this casrn: 79755-43-4 as seen here: https://comptox.epa.gov/dashboard/DTXSID60229856.

You have the name "6-OH-BDE-85" with CASRN "1219628-90-6". We believe this CAS-RN to be "5-OH-BDE-85" (if you want to use this naming convention). Please see here: https://comptox.epa.gov/dashboard/DTXSID10873904. If you like your name as is, I would consider this casrn: 35162-01-7 as seen here: https://comptox.epa.gov/dashboard/DTXSID60474586.

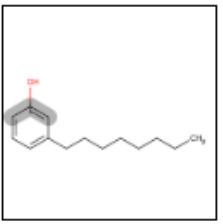
From list:
531-95-3 Equol
94105-90-5 Equol

Explanation: casrn 531-95-3 indicates Equol which is a name associated with defined stereo. Casrn 94105-90-5 means (R,S)-Equol with undefined stereo.

From list:
25154-52-3 4-nonylphenol, branched
84852-15-3 4-nonylphenol, branched

Explanation: casrn 25154-52-3 is n-nonylphenol (linear chain with unknown location of the chain and hydroxyl group on the benzene ring). Casrn 84852-15-3 is indeed 4-nonylphenol, branched (undefined branching in the nonyl alkyl chain).

Endocrine Disruption Reference Set Examples (Nonylphenols)



n-N 4-Nonylphenol, branched

2515 84852-15-3 | DTXSID5029055

Searched Searched by CAS-RN.

Presence in Lists

Record Information

Citation: U.S. Environmental Protection Agency. Chemistry Dashboard. https://comptox.epa.gov/dashboard/DTXSID5029055 (accessed Apr 2nd, 2019),

Data Quality:

Level 1: Expert curated, highest confidence in accuracy and consistency of unique chemical iden

Level 2: Expert curated, nique chemical identifiers using multiple sources

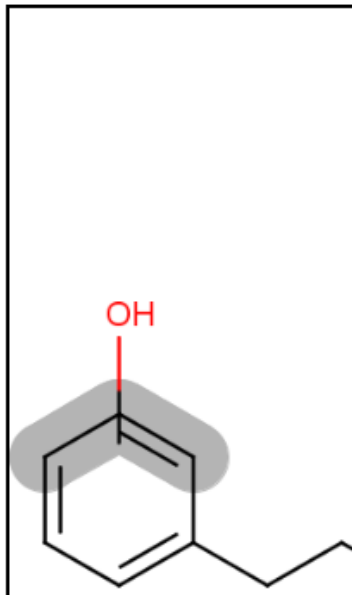
Level 3: Programmatically curated from high quality EPA source, unique chemical identifiers have

Level 4: Programmatically curated from ChemID, unique chemical identifiers have no conflicts in

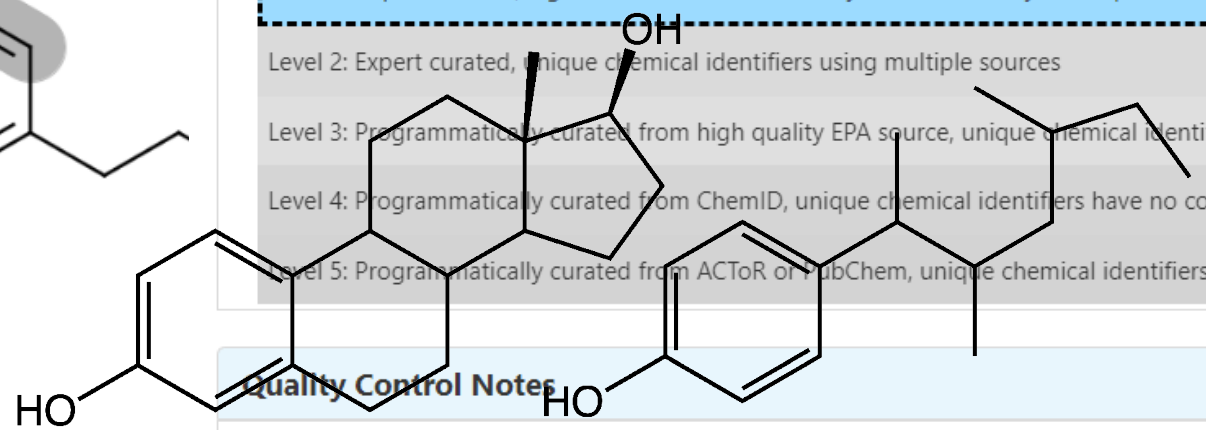
Level 5: Programmatically curated from ACToR or PubChem, unique chemical identifiers with low

Quality Control Notes

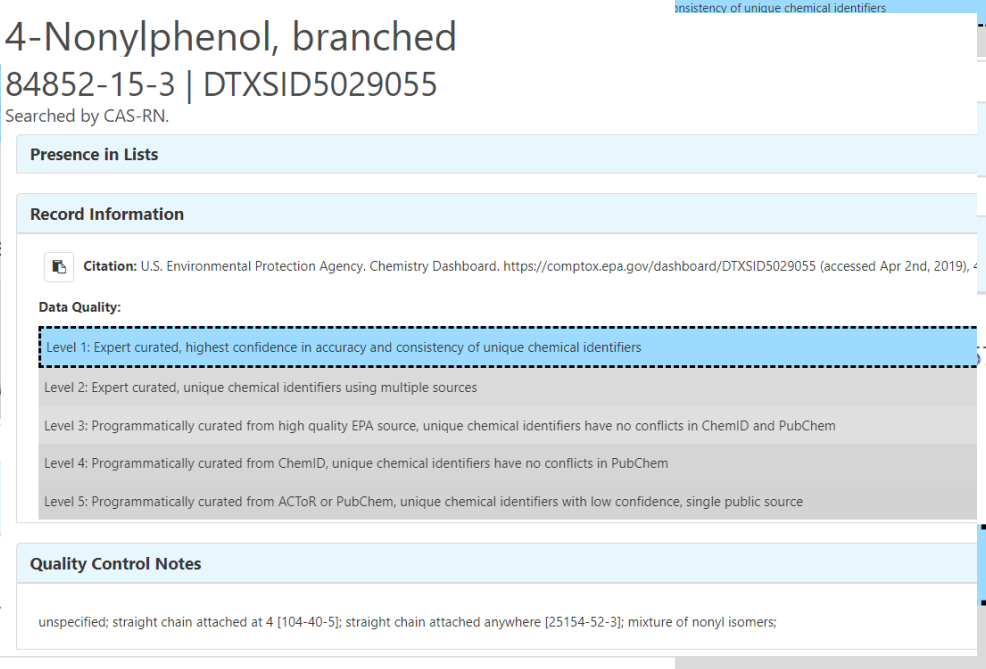
unspecified; straight chain attached at 4 [104-40-5]; straight chain attached anywhere [25154-52-



Estradiol

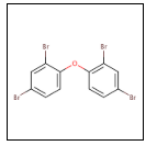


Branched 4-nonylphenol



Endocrine Disruption Reference Set Examples (6-OH-BDE-47)

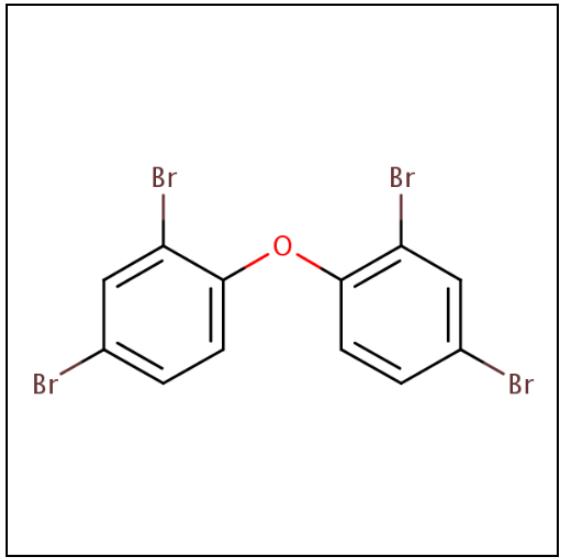
You have the name "6-OH-BDE-47" with CASRN "1017894-49-3". We believe this CAS-RN to be better represented by the name "6'-OH-BDE-85" (if you want to use this naming convention). Please see here: https://comptox.epa.gov/dashboard/DTXSID30873880. If you like the name as is, I would consider this casrn: 79755-43-4 as seen here: https://comptox.epa.gov/dashboard/DTXSID60229856.



2,2',4,4'-Tetrabromodiphenyl ether

5436-43-1 | DTXSID3030056

Searched by DSSTox Substance Id.



Intrinsic Properties

Molecular Formula: $C_{12}H_6Br_4O$ | Mol File | Find All Chemicals

Average Mass: 485.795 g/mol | Isotope Mass Distribution

Monoisotopic Mass: 481.715217 g/mol

Structural Identifiers

Linked Substances

Presence in Lists

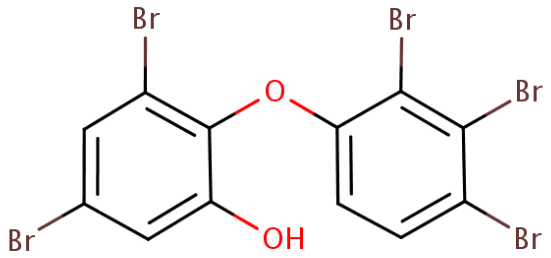
Record Information

Citation: U.S. Environmental Protection Agency. Chemistry Dashboard. https://comptox.epa.gov/d

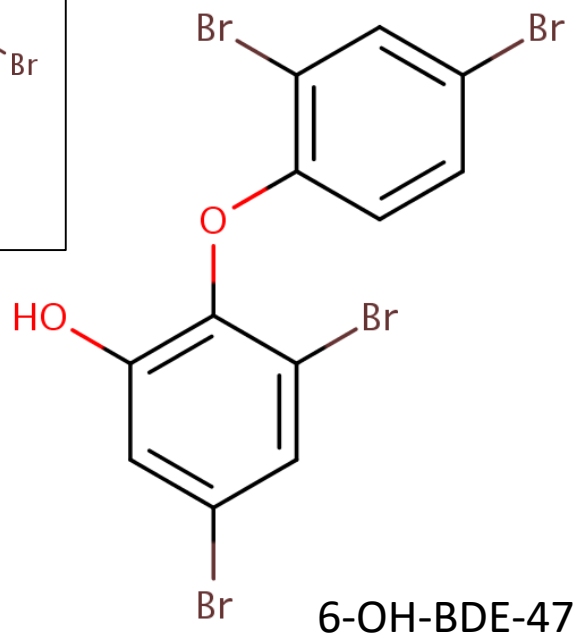
(accessed Apr 2nd, 2019), 2,2',4,4'-Tetrabromodiphenyl ether

Data Quality:

Level 1: Expert curated, highest confidence in accuracy and consistency of unique chemical identifiers



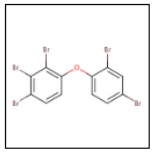
1017894-49-3



6-OH-BDE-47

Endocrine Disruption Reference Set Examples (6-OH-BDE-85)

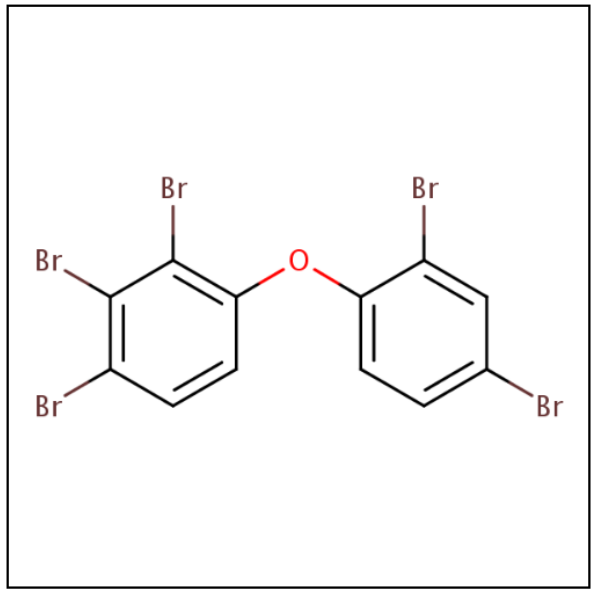
You have the name "6-OH-BDE-85" with CASRN "1219628-90-6". We believe this CAS-RN to be "5-OH-BDE-85" (if you want to use this naming convention). Please see here: https://comptox.epa.gov/dashboard/DTXSID10873904. If you like your name as is, I would consider this casrn: 35162-01-7 as seen here: https://comptox.epa.gov/dashboard/DTXSID60474586.



2,2',3,4,4'-Pentabromodiphenyl ether

182346-21-0 | DTXSID4052685

Searched by DSSTox Substance Id.



Intrinsic Properties

Molecular Formula: $C_{12}H_5Br_5O$ Mol File Find All Chemicals

Average Mass: 564.691 g/mol Isotope Mass Distribution

Monoisotopic Mass: 559.62573 g/mol

Structural Identifiers

Linked Substances

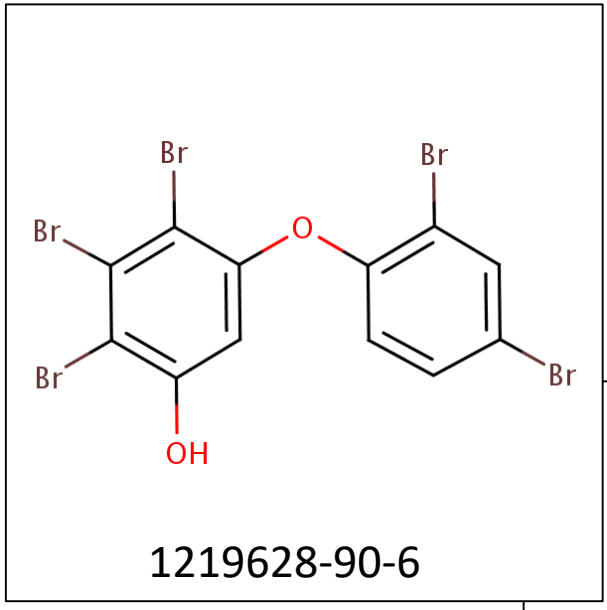
Presence in Lists

Record Information

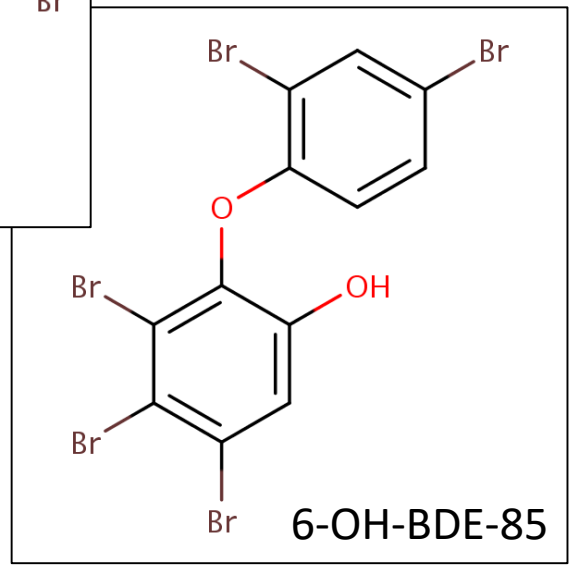
Citation: U.S. Environmental Protection Agency. Chemistry Dashboard. https://comptox.epa.gov/ (accessed Apr 2nd, 2019), 2,2',3,4,4'-Pentabromodiphenyl ether

Data Quality:

Level 1: Expert curated, highest confidence in accuracy and consistency of unique chemical identifiers

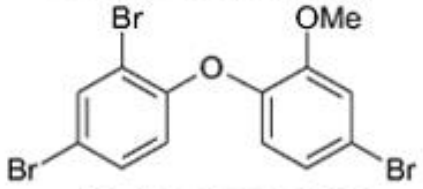
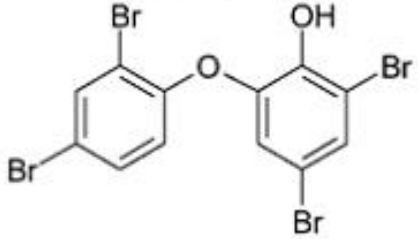
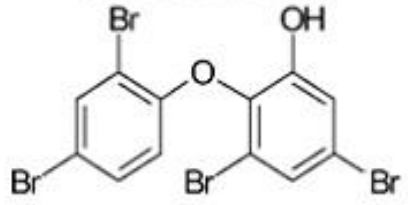


1219628-90-6



6-OH-BDE-85

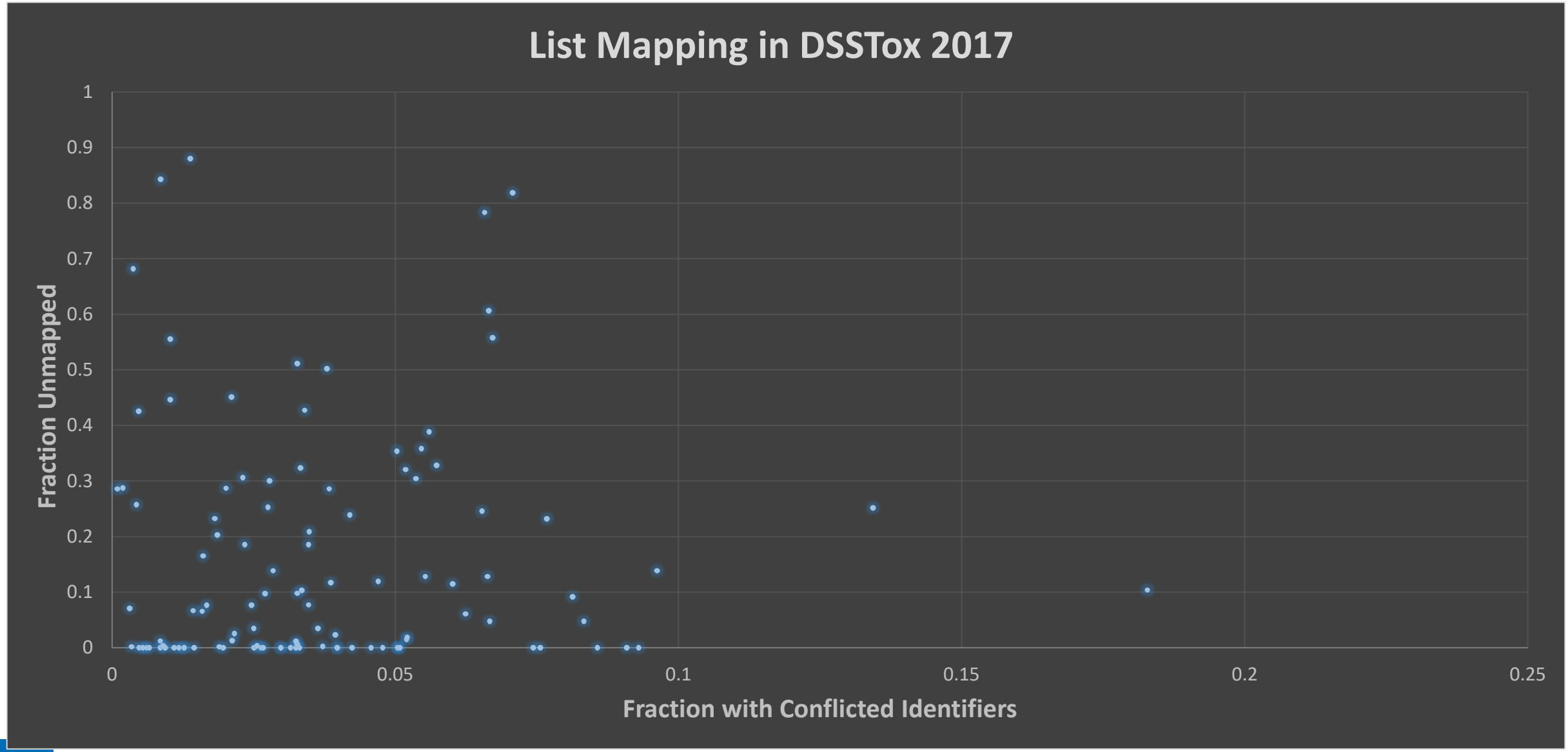
Endocrine Disruption Reference Set Examples (Metabolized BDE)

NAME/STRUCTURE FROM PAPER; Canton et al. 2007	NAME FROM PINTO ET AL.	NAME FROM PINTO ET AL. (ASSIGNED BY OSCP)
2'-MeO-BDE28 	2'-MeO-BDE-68	96920-28-4
2'-OH-BDE 68 	2'-OH-BDE-68	79755-43-4
6-OH-BDE47 	6-OH-BDE-47	1017894-49-3
Could not be located	5-Cl-6-MeO-BDE-47	497106-81-7
Could not be located	5-Cl-6-OH-BDE-47	497069-18-8
Could not be located	6-MeO-BDE-47	102739-99-1
Could not be located	6-OH-BDE-85	1219628-90-6

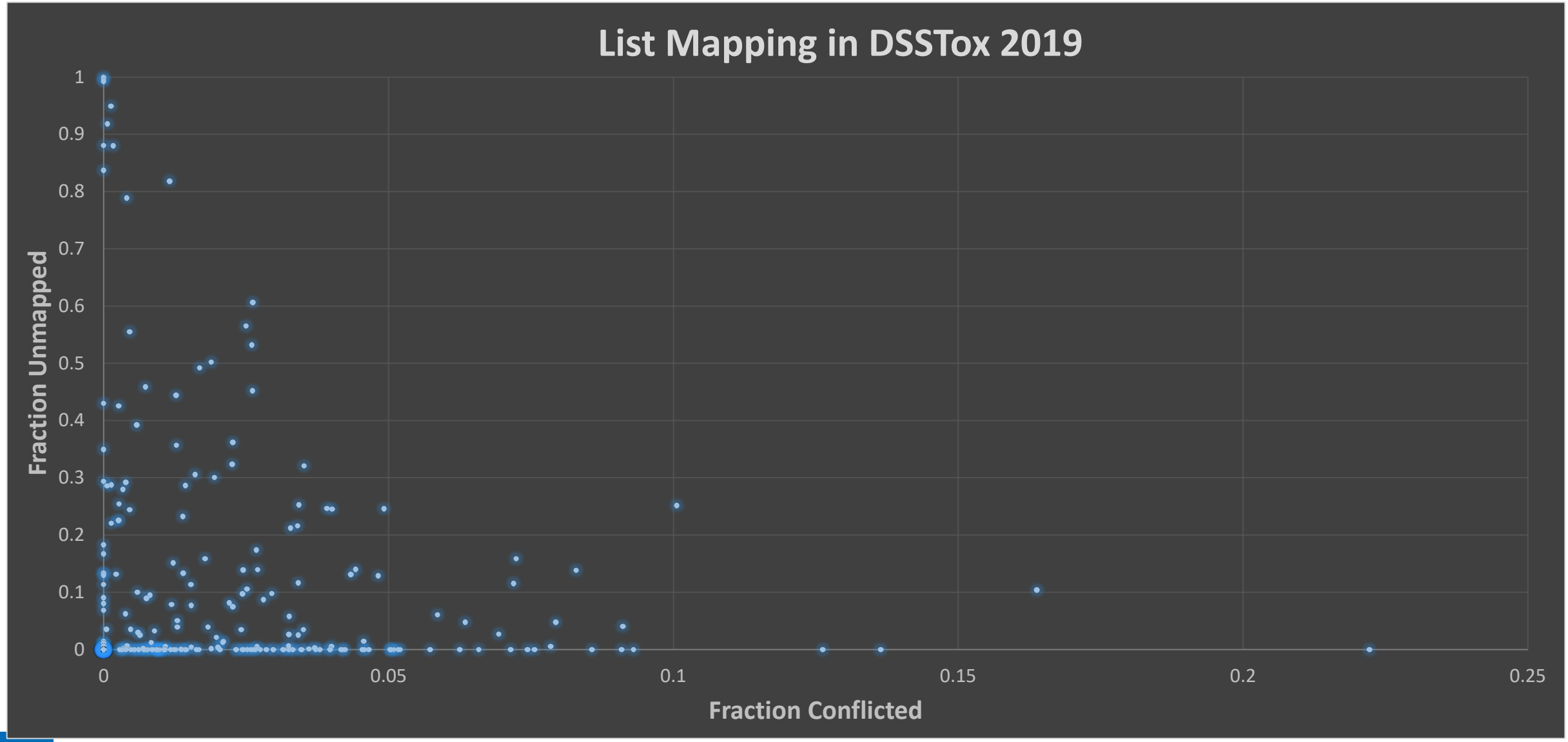
Decision:

- Canton et al synthesized metabolites
  - Trust drawn structures
    - Ignore names
    - Ignore CAS-RNs

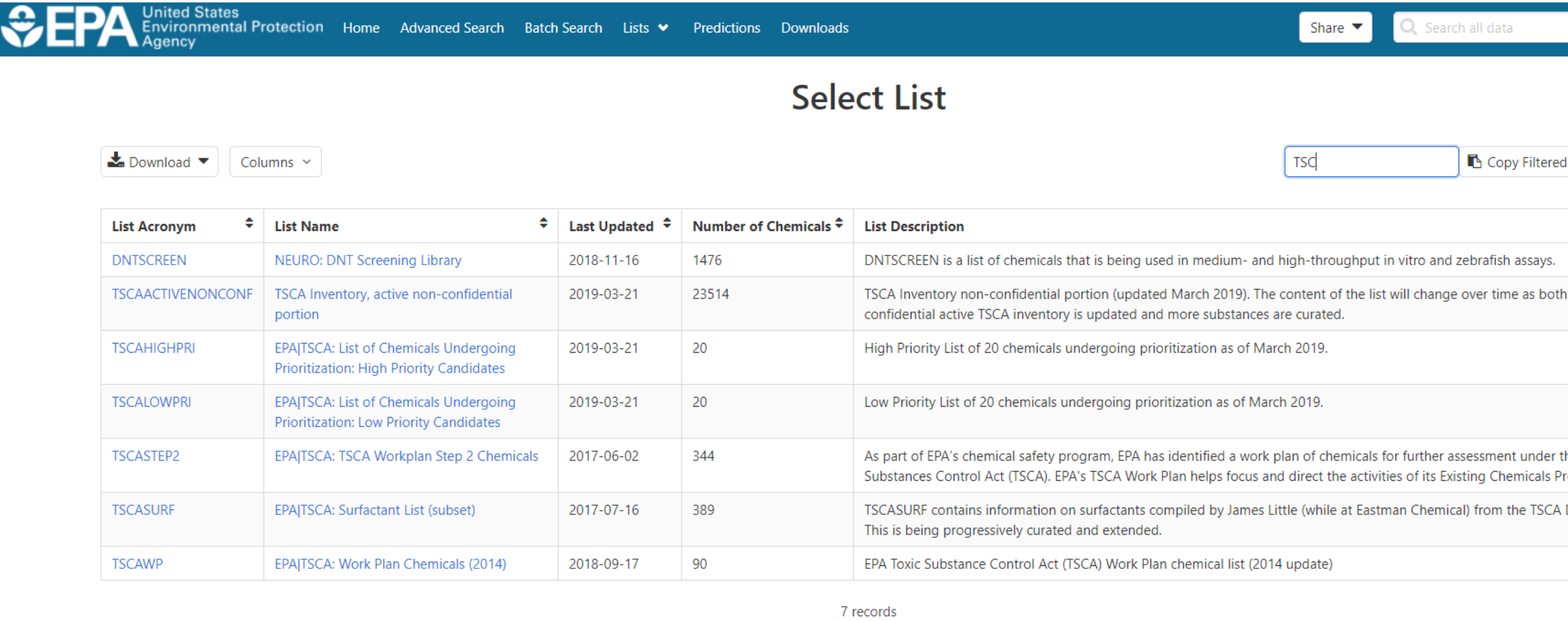
Why Curation is Essential



Why Curation is Essential



Lists in the Dashboard



List Acronym	List Name	Last Updated	Number of Chemicals	List Description
DNTSCREEN	NEURO: DNT Screening Library	2018-11-16	1476	DNTSCREEN is a list of chemicals that is being used in medium- and high-throughput in vitro and zebrafish assays.
TSCAACTIVENONCONF	TSCA Inventory, active non-confidential portion	2019-03-21	23514	TSCA Inventory non-confidential portion (updated March 2019). The content of the list will change over time as both confidential active TSCA inventory is updated and more substances are curated.
TSCAHIGHPRI	EPA|TSCA: List of Chemicals Undergoing Prioritization: High Priority Candidates	2019-03-21	20	High Priority List of 20 chemicals undergoing prioritization as of March 2019.
TSCALOWPRI	EPA|TSCA: List of Chemicals Undergoing Prioritization: Low Priority Candidates	2019-03-21	20	Low Priority List of 20 chemicals undergoing prioritization as of March 2019.
TSCASTEP2	EPA|TSCA: TSCA Workplan Step 2 Chemicals	2017-06-02	344	As part of EPA's chemical safety program, EPA has identified a work plan of chemicals for further assessment under th Substances Control Act (TSCA). EPA's TSCA Work Plan helps focus and direct the activities of its Existing Chemicals Pr
TSCASURF	EPA|TSCA: Surfactant List (subset)	2017-07-16	389	TSCASURF contains information on surfactants compiled by James Little (while at Eastman Chemical) from the TSCA I This is being progressively curated and extended.
TSCAWP	EPA|TSCA: Work Plan Chemicals (2014)	2018-09-17	90	EPA Toxic Substance Control Act (TSCA) Work Plan chemical list (2014 update)

7 records

Conclusions

- Input data often contains confusion in chemical identifiers
- Automated resolution only provides a best guess
- Accurate mapping between data and substances/structures requires thorough investigation and decision making

29

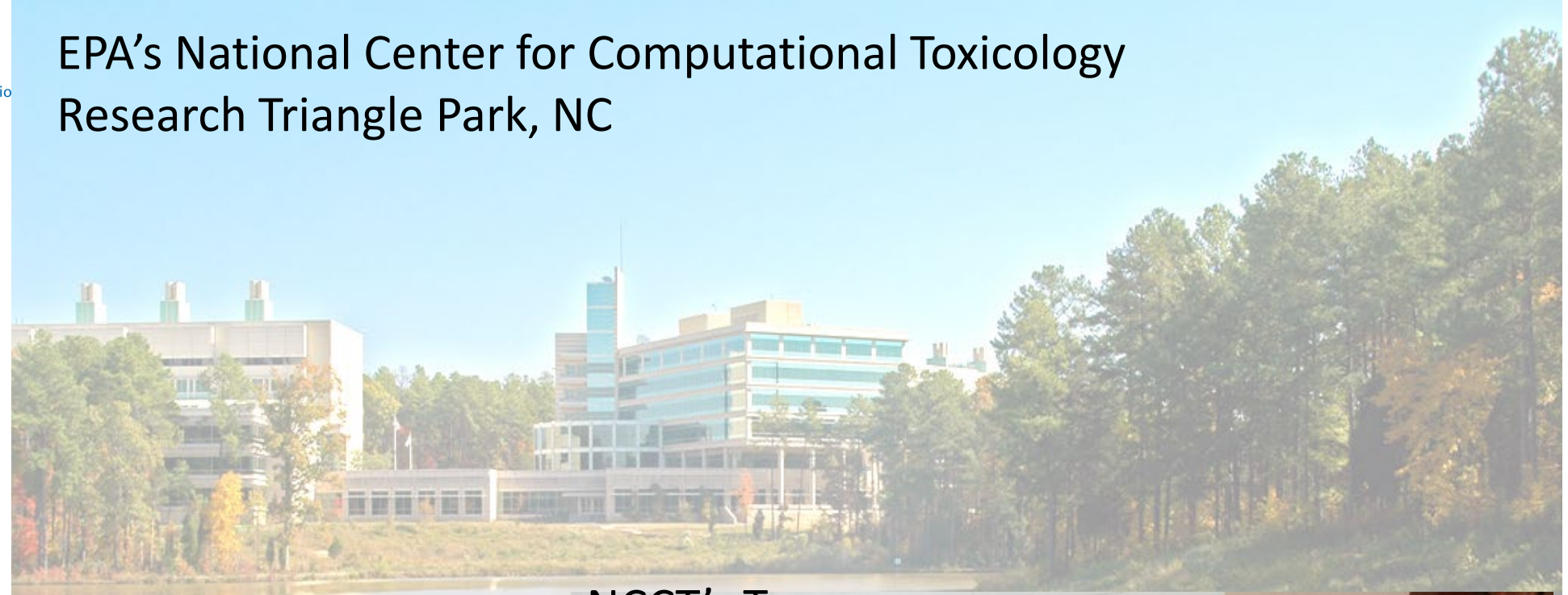
Acknowledgements

EPA NCCT IT
Jeff Edwards
Jeremy Dunne

EPA NCCT Curation
Inthirany Thillainadarajah
Sakuntala Sivasupramaniam

United States Environmental Protection Agency

EPA's National Center for Computational Toxicology
Research Triangle Park, NC



NCCT's Team

EPA United States Environmental Protection Agency

Questions?