United States Environmental Protection Agency

Integration of Markush Structures into EPA's DSSTox Database to Represent and Enumerate UVCB Substances

Christopher Grulke†
Antony Williams
Ann Richard

NCCT, U.S. EPA

American Chemical Society Meeting, Spring 2018

18 March 2018, New Orleans, LA

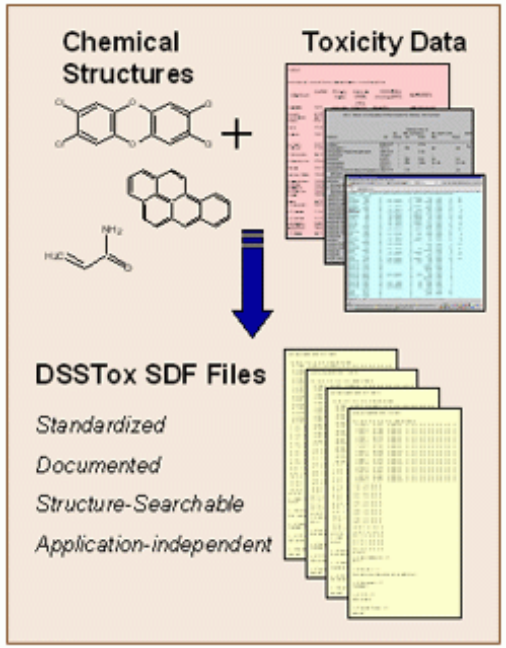
Office of Research and Development

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

DSSTox Ancient History

Goal: Linking chemical structures to data 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
  - etc...



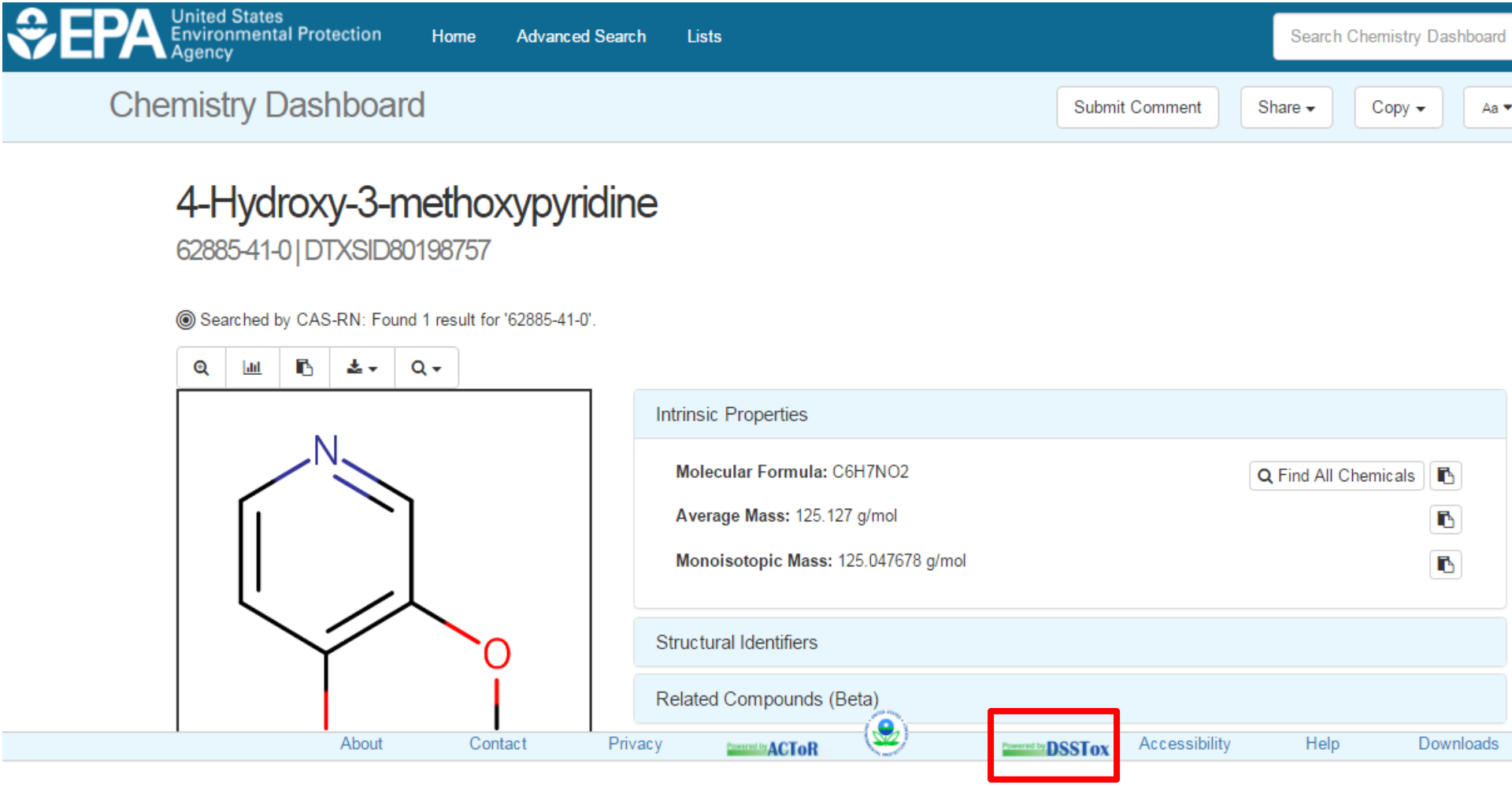
- Managed all chemical registration for ToxCast and Tox21 chemicals
- By 2014, roughly 20K manually curated substance records

DSSTox Current History

- 761K substance records (27.5K manually curated)
- Central database for the Comptox Chemical Dashboard
- More Goals:
  - Become a hub for all chemical data relevant to an environmental scientist
  - Provide batch extraction of chemical data for our user community
  - Offer chemical list based views of our data
  - Provide list specific search capabilities

- Check out: https://comptox.epa.gov/dashboard

Comptox Chemistry Dashboard



Comptox Chemistry Dashboard (Cont.)

United States Environmental Protection Agency

Home | Advanced Search | Lists | Search Chem

Chemistry Dashboard

Submit Comment | Share | Co

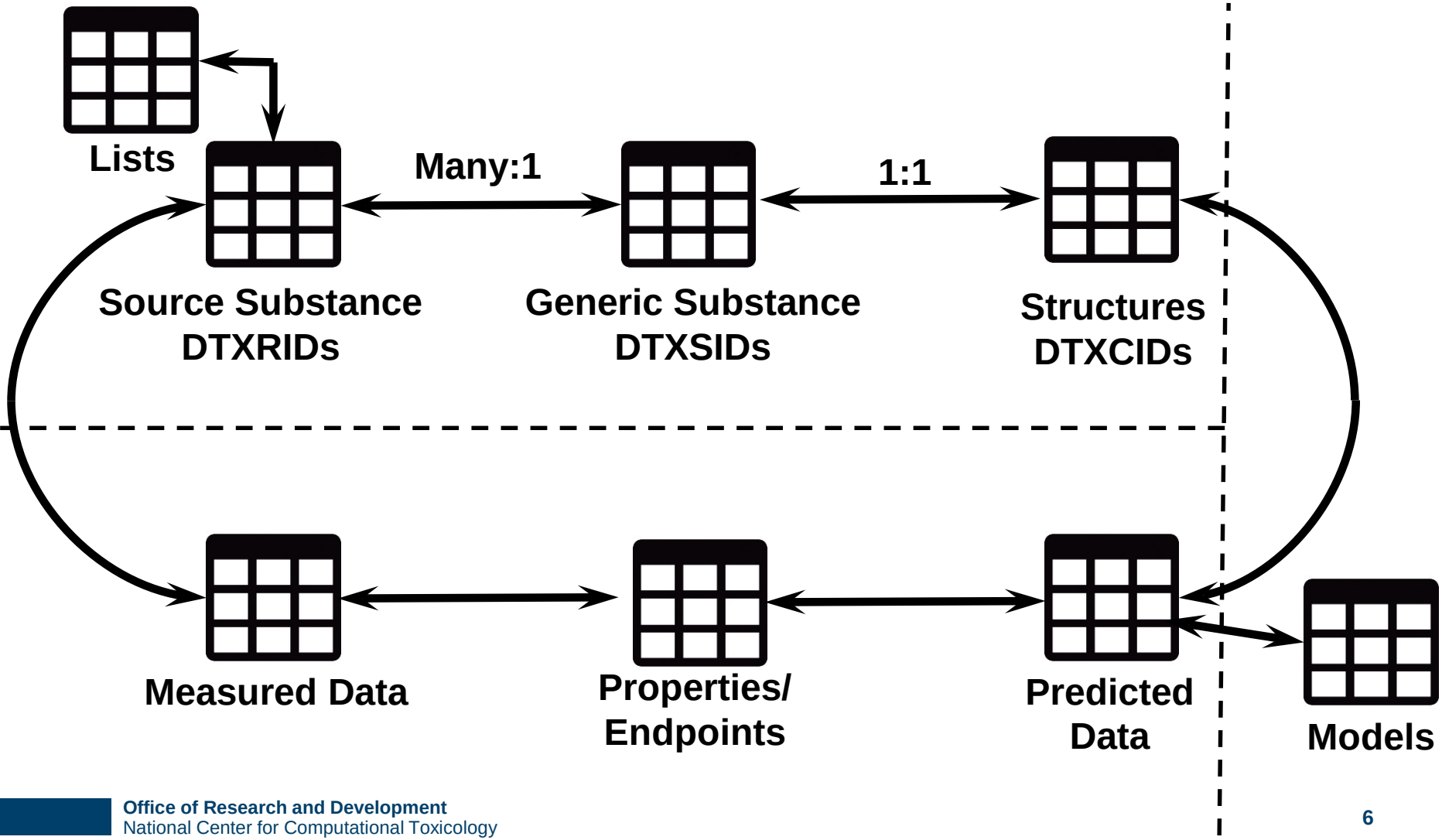
Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

- **Summary**
- LogP: Octanol-Water
- Water Solubility
- Density
- Melting Point
- Boiling Point
- Surface Tension
- Vapor Pressure
- LogKoa: Octanol-Air
- Henry's Law
- Index of Refraction
- Molar Refractivity
- pKa Basic Apparent

Download as: TSV | Excel | SDF

Property	Average		Median			Range	Unit
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
LogP: Octanol-Water	2.61 (1)	2.76 (4)	2.61 to 2.61	2.76	2.61	2.50 to 3.05	-
Water Solubility	1.30e-04 (1)	1.46e-02 (4)	1.30e-04 to 1.30e-04	1.46e-02	1.30e-04	1.50e-04 to 5.71e-02	mol/L
Density	-	1.27 (1)	-	1.27	-	-	g/cm^3
Melting Point	174 (6)	151 (3)	173 to 177	151	173 to 177	114 to 185	°C
Boiling Point	-	312 (3)	-	312	-	284 to 339	°C
Surface Tension	-	53.8 (1)	-	53.8	-	-	dyn/cm
Vapor Pressure	7.21e-11 (1)	4.47e-06 (3)	7.21e-11 to 7.21e-11	4.47e-06	7.21e-11	2.06e-07 to 1.27e-05	mmHg
LogKoa: Octanol-Air	-	8.38 (1)	-	8.38	-	-	-
Henry's Law	-	4.20e-10 (1)	-	4.20e-10	-	-	atm-m3/r
Index of Refraction	-	1.61 (1)	-	1.61	-	-	-
Molar Refractivity	-	58.5 (1)	-	58.5	-	-	cm^3
pKa Basic Apparent	-	2.27 (1)	-	2.27	-	-	-
Molar Volume	-	170 (1)	-	170	-	-	cm^3
Polarizability	-	23.2 (1)	-	23.2	-	-	Å^3

General Data Model



So Where is the Data?

C10-16 Alcohols

67762-41-8 | DTXSID4028331

Searched by Approved Name: Found 1 result for 'C10-16 Alcohols'.

Presence in Lists

Federal | Safer Choice Chemical List

US State

International

Other | TSCAACTIVE

Record Information

Quality Control Notes

Related Substances | Synonyms | Links | Bioassays | Exposure | Hazard | Comments | Chemical Properties | Literature

No Chemical Properties Found.

United States Environmental Protection Agency

UVCB Chemicals

Environmental Topics | Laws & Regulations | About EPA | Search EPA.gov

CONTACT US SHARE

TSCA Chemical Substance Inventory

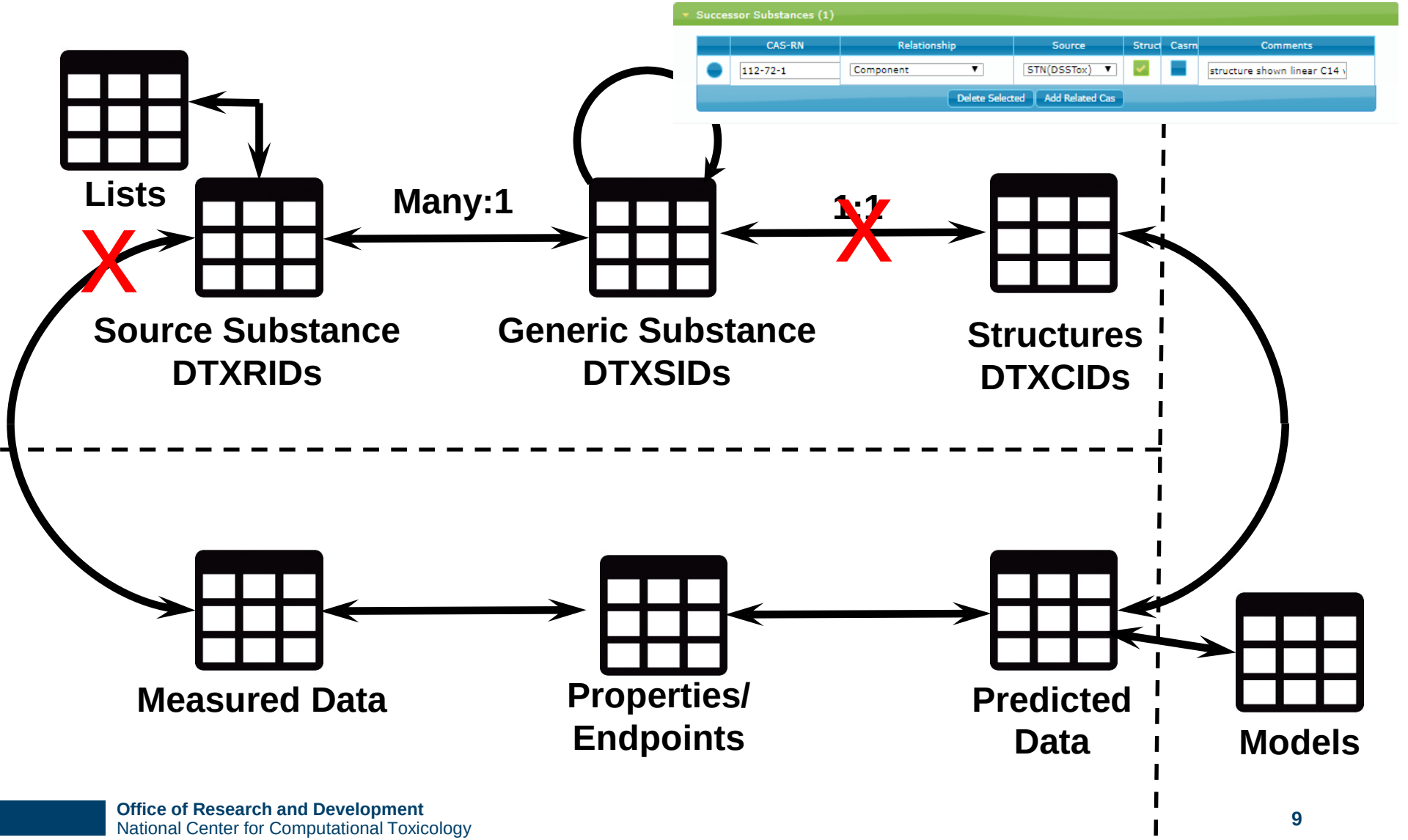
- TSCA Inventory Home
- About the Inventory
- Access the Inventory
- Policy and Guidance

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

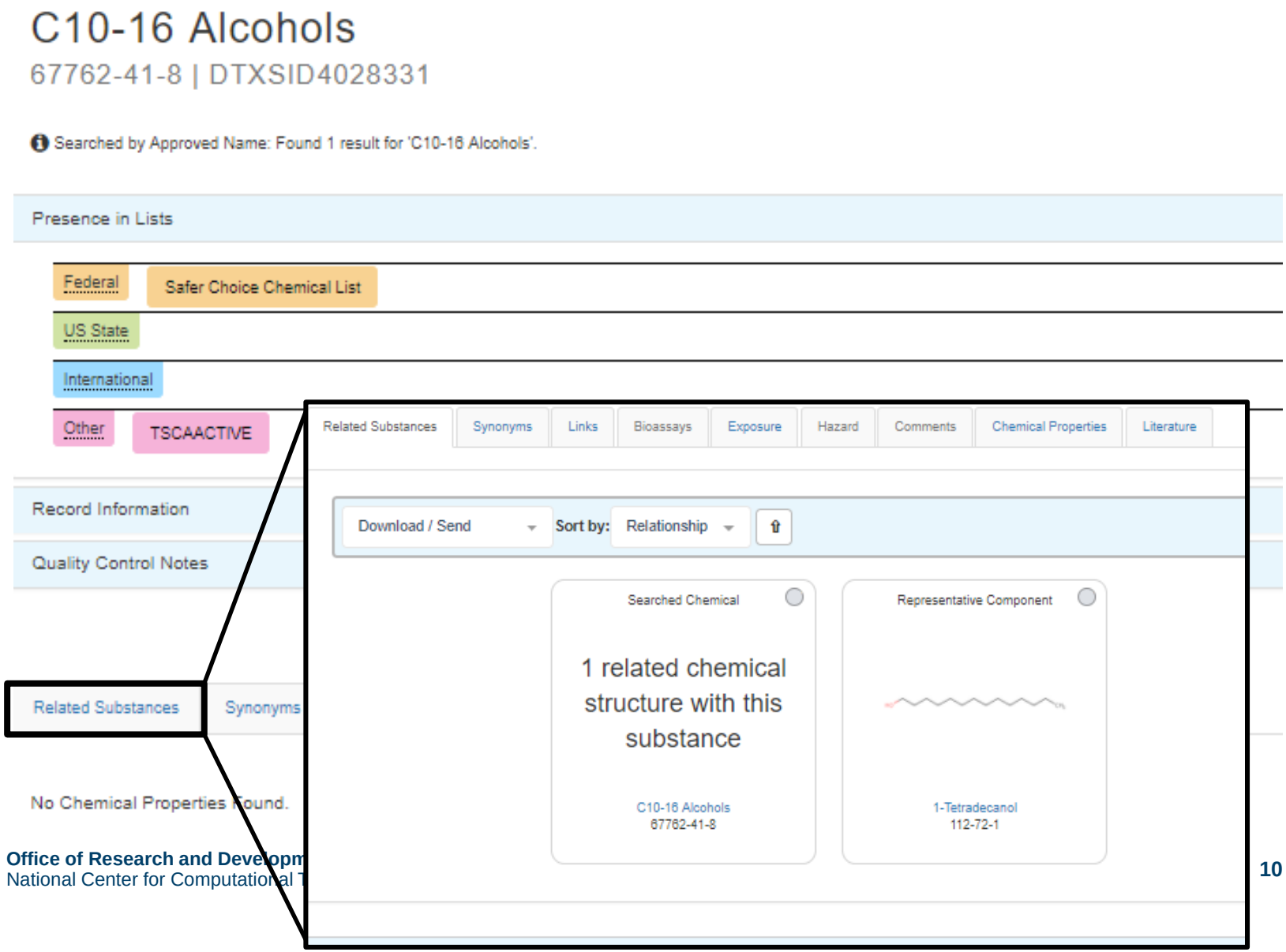
This paper is a compendium of information related to the broad class of chemical substances referred to as UVCBs for the Toxic Substances Control Act (TSCA) Chemical Substance Inventory. These chemical substances cannot be represented by unique structures and molecular formulas.

- UVCB chemical examples
  - Surfactants with undefined composition
  - Petroleum Distillates
  - Gelatins, hydrozylates
  - Formaldehyde, reaction products with diethanolamine
  - Fatty acids, linseed-oil, compds. with triethylamine

Data linkage breakdown



So Where is the Data?



Chemical Families (e.g. PCBs)

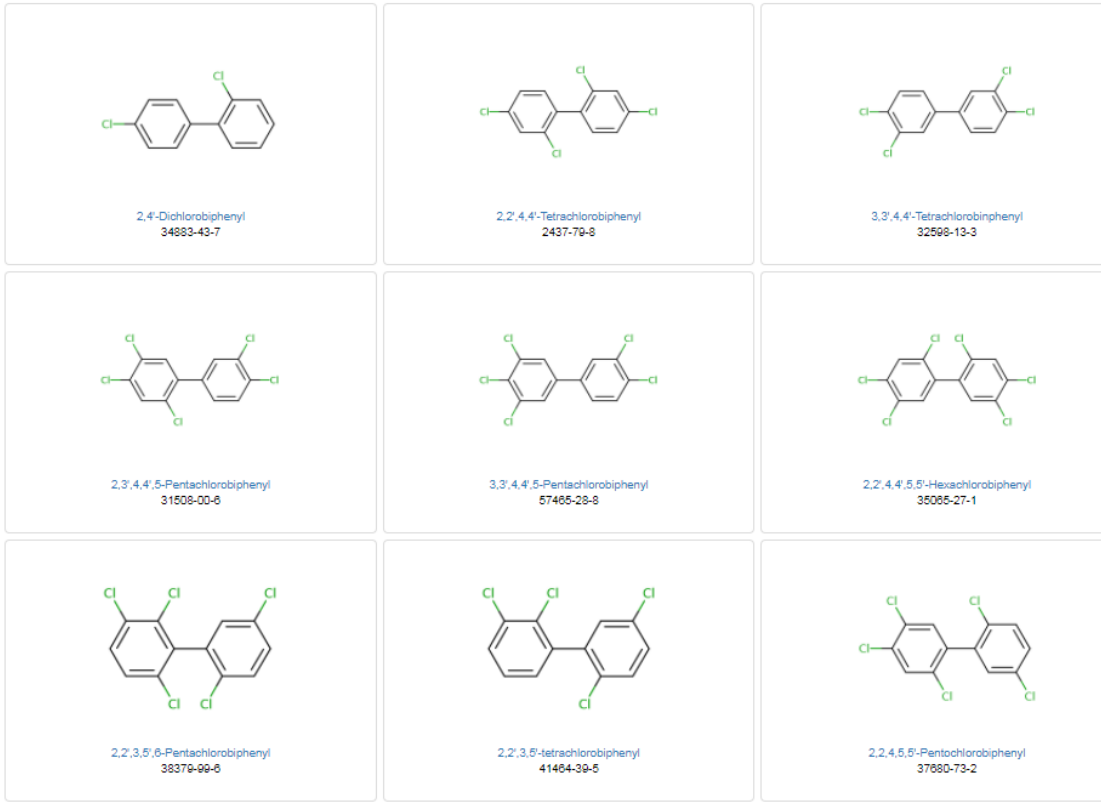
Successor Substances (209)

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

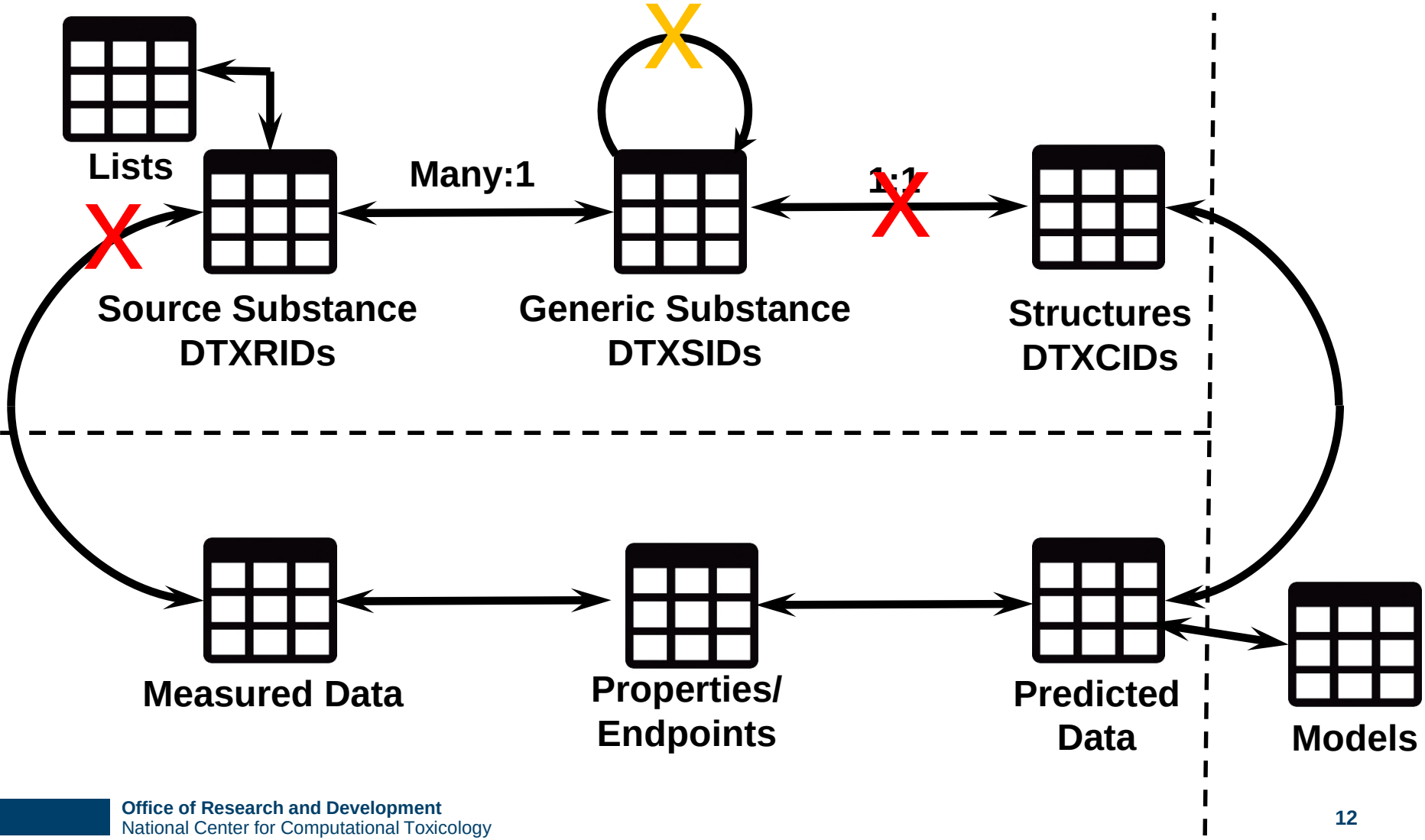
Related Chemicals

Found 209 chemicals

Download as: TSV Excel SDF



Data linkage breakdown



United States Environmental Protection Agency

ChemAxon Markush Technology

ChemAxon Support Tickets | Events | Jobs | Contact Us

Search ChemAxon...

Products Download Support Services About us Web Store tony27587@gmail.com

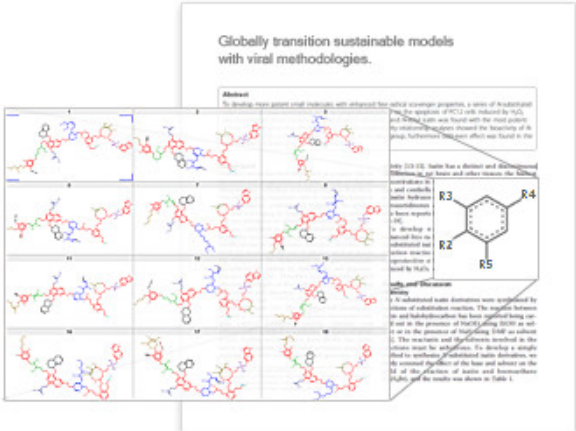
Home > Products > Markush Technology

Markush Technology

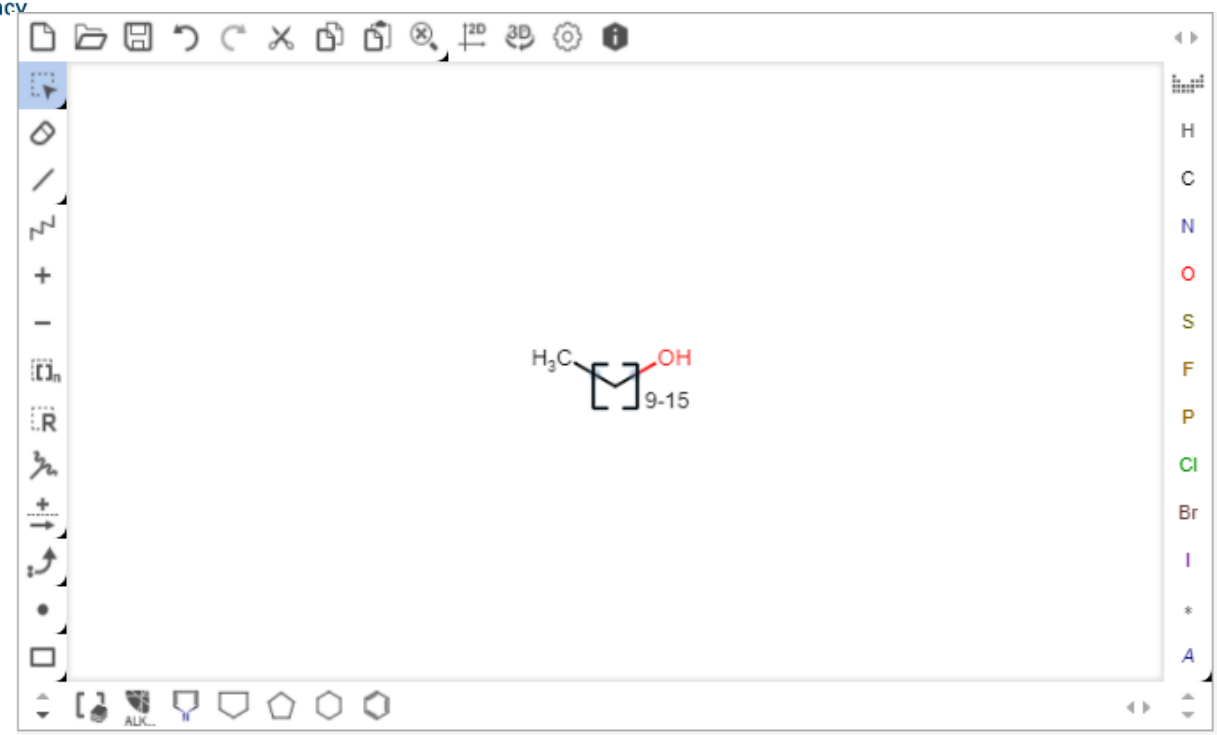
Download in JChem Download the Editor

TOOLKIT FOR THE ANALYSIS OF VIRTUAL COMBINATORIAL LIBRARY AND PATENT MARKUSH STRUCTURES

Markush structures, or generic structures, are widely used in combinatorial libraries and chemical patents to define large chemical spaces. ChemAxon provides the most advanced Markush technology, including rapid structure search in Markush space, enumeration, overlap analyses, and automatic Markush composition. All Markush analysis features are available as an add-on for ChemAxon's JChem technology (JChem Base, JChem Cartridge and Instant JChem).



Update Record



Calculate from Structure

Substance_ID: DTXSID4028331

CAS: 67762-41-8

Name: C10-16 Alcohols

Substance Type: Mixture/Formulation

QC Level: DSSTox_High

Data Source: STN(DSSTox)

QC Notes: SDA (Soap and Detergent Association) Reporting Number: 15-060-00. SDA Substance Name: C10-C16 alkyl alcohol

Compound_ID:

Chemical Shown: Markush Enumerable

Private Notes:

Source of CAS-Compound: Public

Double Stereo: None

Chiral Stereo: None

Chemical Form: Organic

Organic Form: Parent

Look Away: Work In Progress

```
service
hibernate.cfg.xml
hibernate-pg.cfg.xml
log4j.properties
.gitignore
aromatic.txt
ChEBI_complete.sdf
ChEBI_complete_3star.sdf
chemregscripts.iml
DocumentCategoryNOSlist1.txt
DocumentCategoryNOSlist2.txt
DocumentCategoryNOSlist3.txt
DocumentCategoryNOSlist4.txt
DocumentCategoryNOSlist5.txt
DSSToxTOX21S_with ACDLABSPhysChem.sdf
DSSToxTOXCSTwith ACDLABSPhysChem.sdf
mixtures2.sdf
new_tox21_records.csv
tox21_qsur_predictions_in_domain.csv
TSCA_missing.xlsx
Updated_STAR_Flags_WS.sdf
External Libraries
< 9.0 > C:\Program Files\Java\jdk-9.0.4
com.epam.indigo:indigo-inchi:1.2.3.r0
```

```
44          for(Compound compound : compounds){
45              Calendar cal = Calendar.getInstance();
46              List<String> inchikeys = JChemMarkush.EnumerateInchi
47              System.out.println(compound.getDsstoxCompoundId() +
48              for(String key:inchikeys){
49                  Compound newComp = compoundDAO.findByInchiKey(ke
50                  if(newComp!=null){
51                      System.out.println("\t"+key+"\t" +newComp.ge
52                      CompoundRelationship cr = new CompoundRelati
53                      cr.setRelationship("Markush Child");
54                      cr.setSource("ChemRegScripts/MarkushEnumerat
55                      cr.setFkCompoundrelationshipTypeId(31);
56                      cr.setPredecessorCompound(compound);
57                      cr.setSuccessorCompound(newComp);
58                      cr.setCreatedBy("cgrulke");
59                      cr.setUpdatedBy("cgrulke");
60                      sf.getCurrentSession().save(cr);
61                  }else{
62                      System.out.println("\t"+key);
63                  }
64              }
65          }
66
67          tx.commit();
68  //        tx.rollback();
69      }
70  }
71
```

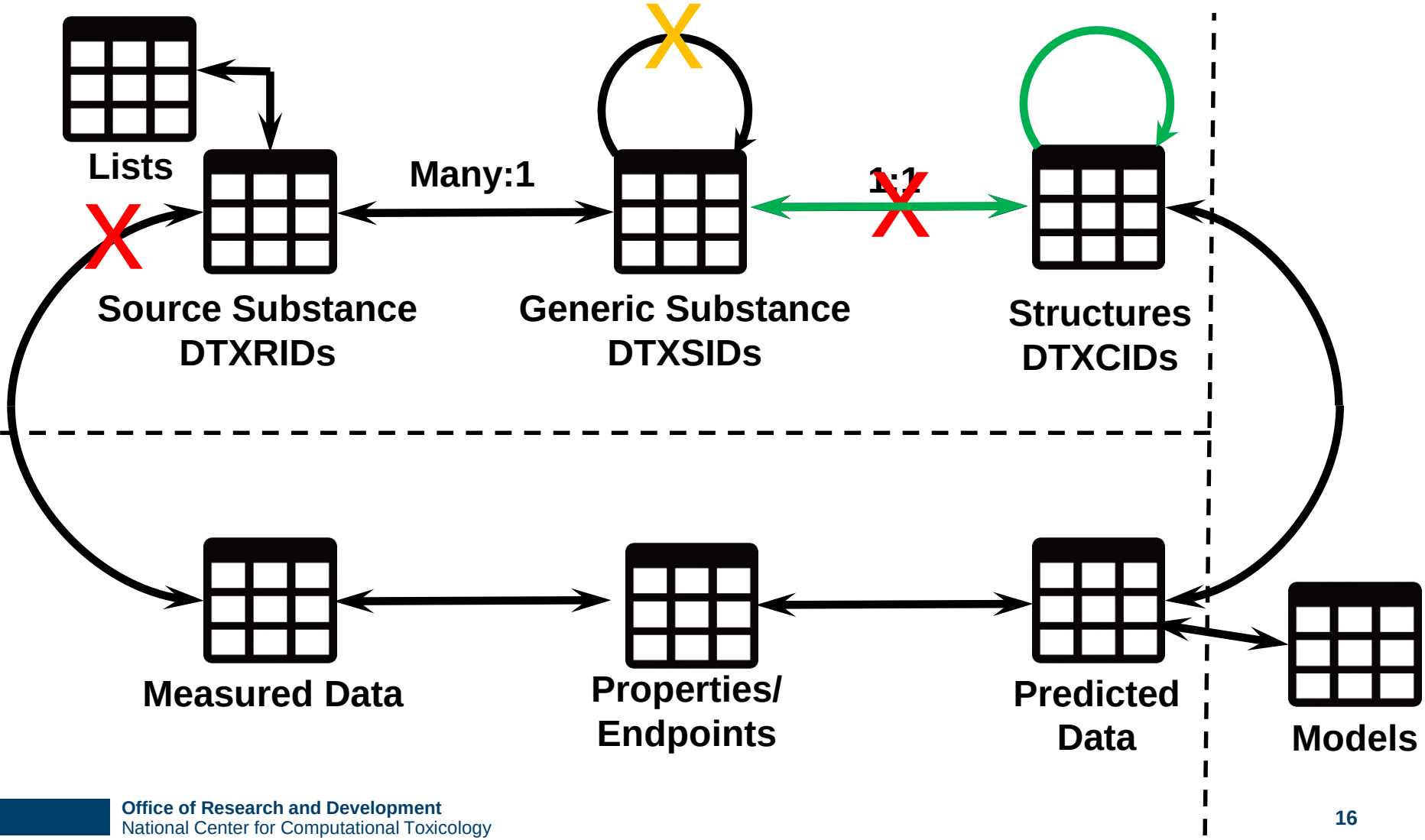
MarkushEnumeration › main()

Run MarkushEnumeration

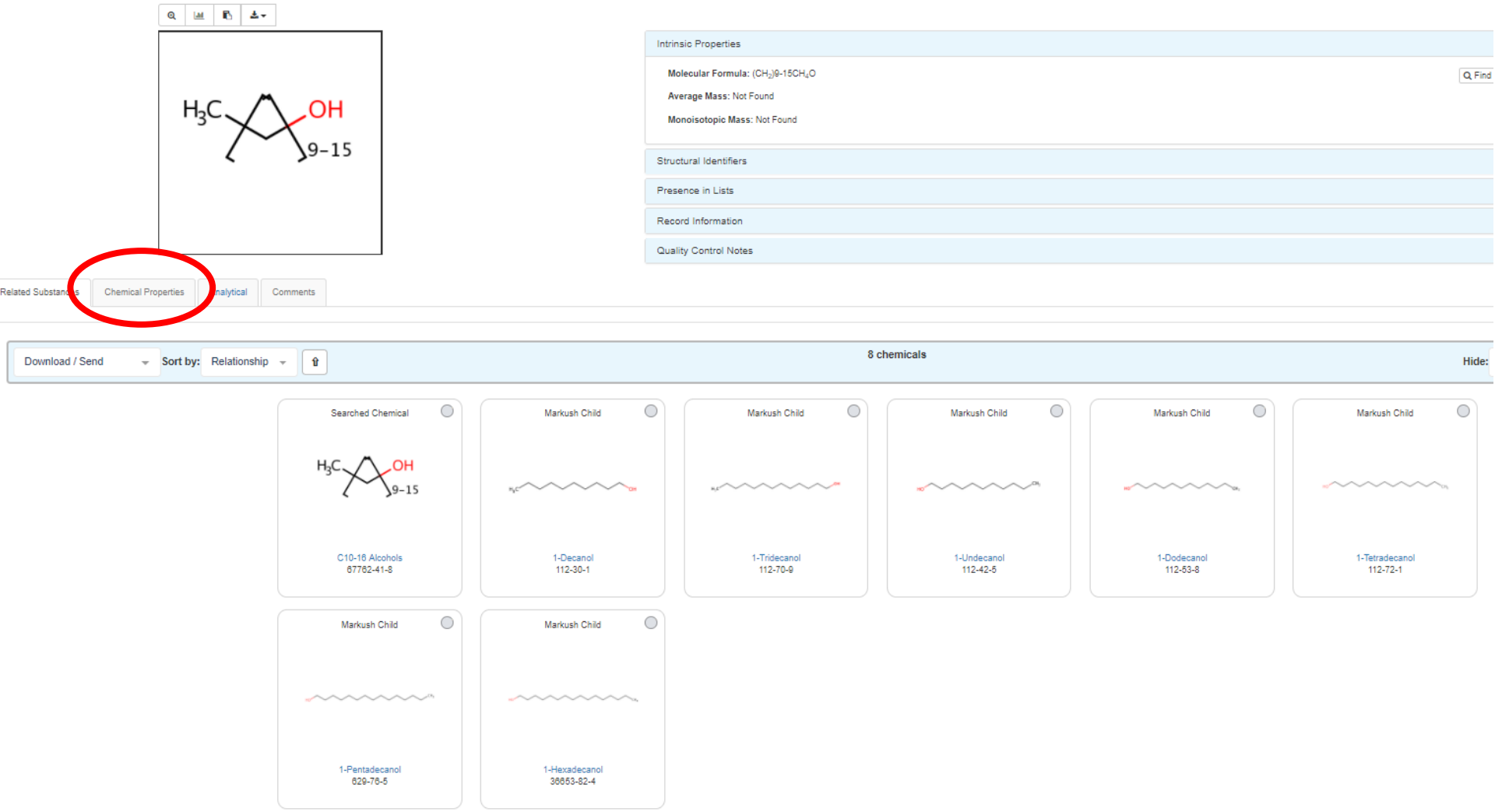
```
        \-[NAMED_PARAM] ParameterNode: '?' {name=inchiKey, expectedType=org.hibernate.type.StringType@6ed16657}

DEBUG [main] (ErrorCounter.java:113) - throwQueryException() : no errors
DEBUG [main] (QueryTranslatorImpl.java:263) - HQL: from database.Compound where jchemInchikey = :inchiKey
DEBUG [main] (QueryTranslatorImpl.java:264) - SQL: select compound0_.id as id1_20_, compound0_.acd_index_name as acd_inde2_20_, com
DEBUG [main] (ErrorCounter.java:113) - throwQueryException() : no errors
    MWKFXSUHUHTGQN-UHFFFAOYSA-N DTXCID501946
    KJIOQYGWTQBHNH-UHFFFAOYSA-N DTXCID706915
    LQZZUXJYWNFBMV-UHFFFAOYSA-N DTXCID906918
    XFRVVPUIAFSTFO-UHFFFAOYSA-N DTXCID901947
    HLZKNKRTKFSKGZ-UHFFFAOYSA-N DTXCID406926
    REIUXOLGHVXAEO-UHFFFAOYSA-N DTXCID207270
    BXWNKGSJHAJOGX-UHFFFAOYSA-N DTXCID607991

Process finished with exit code 0
```


Lists
Many:1
1:1
Source Substance
DTXRIDs
Generic Substance
DTXSIDs
Structures
DTXCIDs
Measured Data
Properties/
Endpoints
Predicted
Data
Models

New Relationships in Comptox Chemistry Dashboard



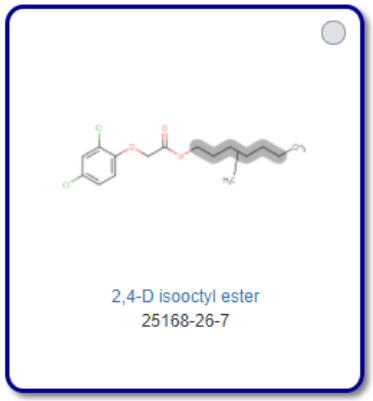
Collecting Property Distributions for Markush ~~Children~~

- Melting Point
- Boiling Point
- Surface Tension
- **Vapor Pressure**
- LogKoa: Octanol-Air
- Henry's Law
- Index of Refraction
- Molar Refractivity
- Molar Volume
- Polarizability

Experimental		
Source	**Result**	
PhysPropNCCT	2.52e-05 mmHg	
Predicted		
Source	**Result**	**Calcu**
NICEATM	5.23e-05 mmHg	Not Available
ACD/Labs	1.24e-04 mmHg	Not Available
TEST	1.24e-05 mmHg	TEST Report
OPERA	1.33e-05 mmHg	OPERA Model Report

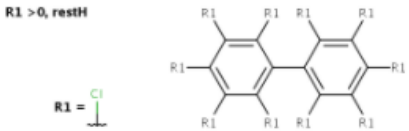
Markush in DSSTox

Download / Send | Sort by: DTXSID | 126 chemicals | Hide: | Select all

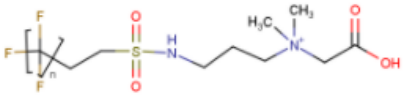


Span of Predicted Properties for UVCBs

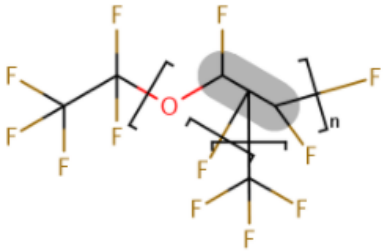
Problems: Queryable, but Not Enumerable



Polychlorinated biphenyls
1336-36-3



Fluorotelomer Sulfonamido Betaines
NOCAS_892972



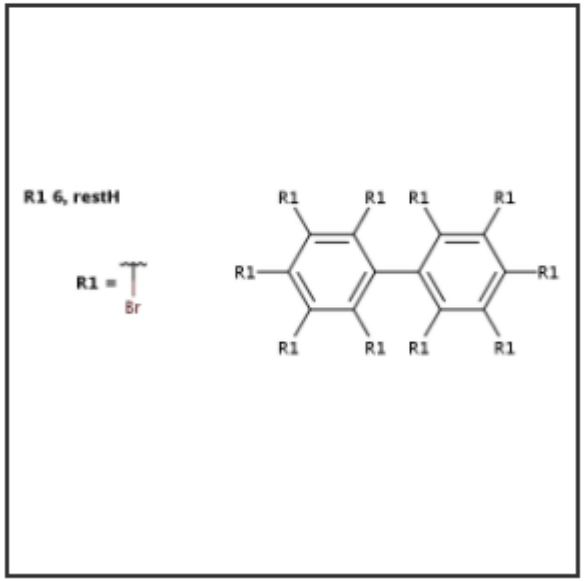
Krytox 143
60164-51-4

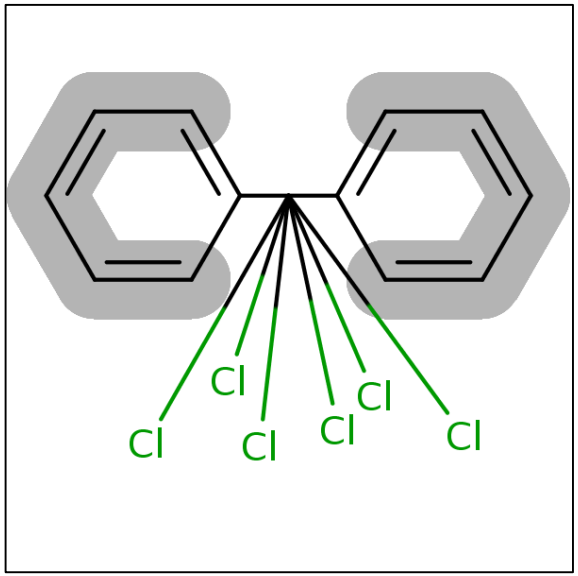
Problem: Human Readable or Computable

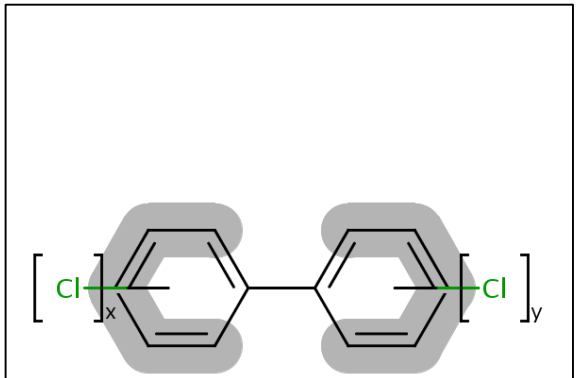
Hexabromobiphenyl

36355-01-8 | DTXSID3025382

Searched by Approved Name: Found 1 result for 'Hexabromobiphenyl'.

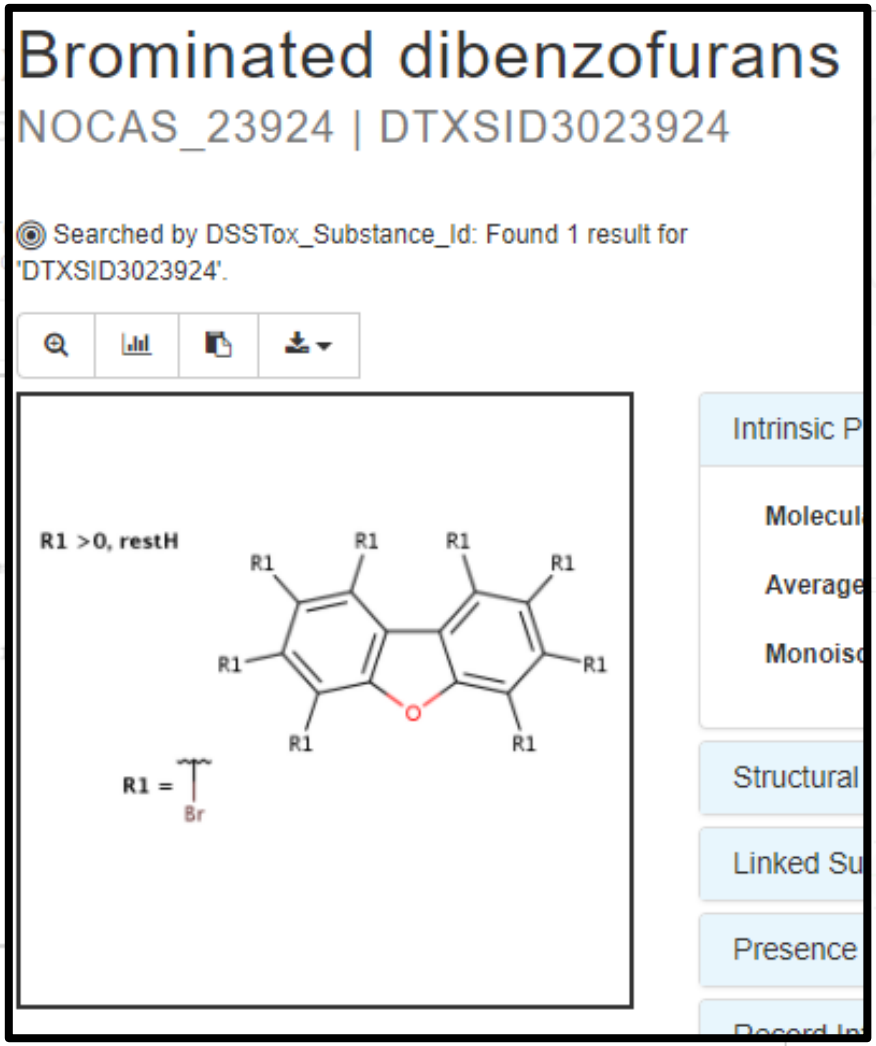






x+y=6

Problem: Markush Uniqueness



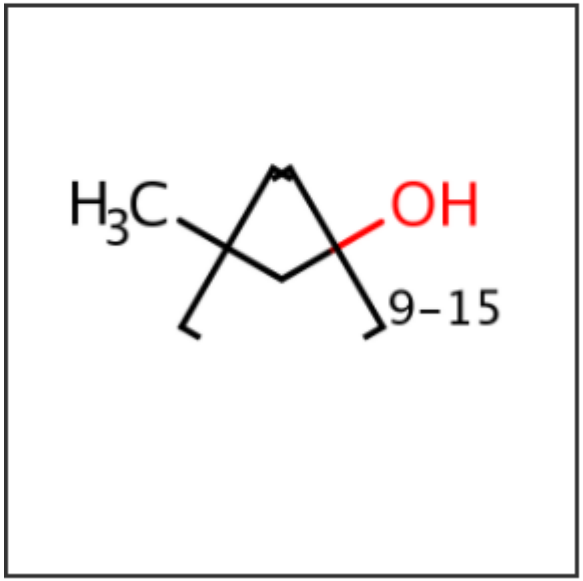
- InChI-Keys for Markush?
- Canconical Extend Smiles?
- Enumerate and compare children?

Problem: Do We Know the Substance?

C10-16 Alcohols

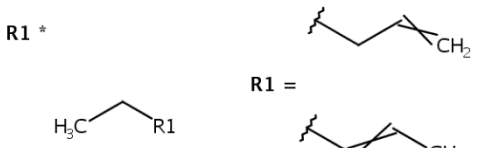
67762-41-8 | DTXSID4028331

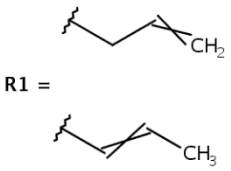
Searched by Approved Name: Found 1 result for 'C10-16 Alcohols'.



Problem: Drawing Limitations

- Variable position bonds (e.g, Pentene)

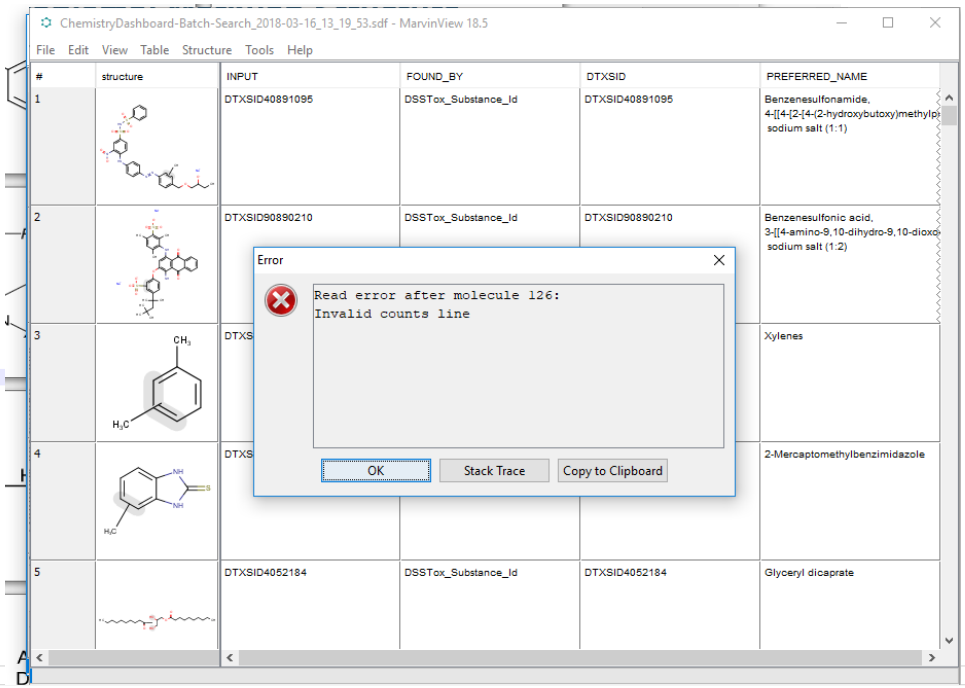




Problems: Markush Structure Transfer

- Mol File

```
4  V30 6 1 6 11
4  V30 7 1 6 12
4  V30 8 1 8 9
4  V30 9 2 8 10
4  V30 10 1 4 3
4  V30 11 1 3 13
4  V30 12 1 5 14
4  V30 13 2 14 16
4  V30 14 2 14 17
4  V30 15 1 14 15
4  V30 16 1 15 18
4  V30 17 1 18 19
4  V30 18 1 19 20
4  V30 19 1 6 20
4  V30 END BOND
4  V30 BEGIN SGROUP
4  V30 1 SRU 0 ATOMS=(3 1 2 3) XBONDS=(2 10 11) BRKXYZ=(9 -10.3158 -0.8296 0 -
4  V30 -10.3987 1.0166 0 0 0 0) BRKXYZ=(9 -8.7797 0.6021 0 -9.7037 -0.9984 0 -
4  V30 0 0 0) CONNECT=HT LABEL=n
4  V30 END SGROUP
4  V30 END CTAB
4  END
```



- CXSMILES

DTXSID4024858	m,p-Cresol mixture	C*.OC1=CC=CC=C1 |c:4,6,t:2,lp:2:2,m:1:6.7|
DTXSID3024364	Cresol	C*.OC1=CC=CC=C1 |c:4,6,t:2,lp:2:2,m:1:6.7.8|
DTXSID5024396	Acetin	CC(*)=O.OCC(O)CO |lp:3:2,4:2,7:2,9:2,m:2:4.7|
DTXSID7024611	Bis(t-butyl dioxyisopropyl)benzene	CC(C)(C)OOC(C)(C)*.CC(C)(C)OOC(C)(C)C1=CC=CC=C1 |c:20,22,t:18,lp:4:2,5:2,14:2,15:2,m:9:21.22|
DTXSID6024701	tert-Butylphenyl diphenyl phosphate	CC(C)(C)*.O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=CC=C1 |c:9,11,17,19,25,27,t:7,15,23,lp:5:2,7:2,14:2,21:2,m:4:11.12.13|
DTXSID6024703	Butyltin tris(isooctylmercaptoacetate)	C*.C*.C*.CCCCCCCOC(=O)CS[Sn](CCCC)(SCC(=O)OCCCCCCC)SCC(=O)OCCCCCCC |lp:13:2,15:2,17:2,23:2,26:2,27:2,35:2,38:2,39:2,m:1:7.8.9.10.11.12,3:40.41.42.43.44.45,5:28.29.30.31.32.33|
DTXSID3024861	Cresyl diphenyl phosphate	C*.O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)OC1=CC=CC=C1 |c:6,8,14,16,22,24,t:4,12,20,lp:2:2,4:2,11:2,18:2,m:1:6.7.8|
DTXSID4025082	Diisodecyl phthalate	C*.C*.CCCCCCCCCOC(=O)C1=C(C=CC=C1)C(=O)OCCCCCCCCC |c:16,18,t:14,lp:13:2,15:2,23:2,24:2,m:1:5.6.7.8.9.10.11.12,3:25.26.27.28.29.30.31.32|
DTXSID7025130	Dimethylnaphthalene	C*.C*.C1=CC2=C(C=C1)C=CC=C2 |c:2,6,9,11,m:1:4.5,3:4.8.9.10.11.12.13|
DTXSID3025172	Dioctyltin-S,S'-bis(isooctylmercaptoacetate)	C*.C*.CCCCCCCC[Sn](CCCCCCCC)(SCC(=O)OCCCCCCC)SCC(=O)OCCCCCCC |lp:21:2,24:2,25:2,33:2,36:2,37:2,m:1:38.39.40.41.42.43,3:26.27.28.29.30.31|
DTXSID7025188	C.I. Direct Black 80	[Na+].[Na+].[Na+].NC1=CC=C2C=C(C(N=NC3=CC=C4C=C(C(N=NC5=CC=C(C=C5)N=NC5=CC=C(N)C6=C5C=CC=C6)=C(O)C4=C3)S([O-])(=O)=O)=C(O)C2=C1)S([O-])(=O)=O.[O-]S(*)(=O)=O |c:5,14,21,23,33,36,38,44,54,t:1,3,10,12,
DTXSID2025216	Divinylbenzene	*C=C.C=CC1=CC=CC=C1 |c:6,8,t:4,m:0:6.8|
DTXSID7025219	Sodium dodecylbenzenesulfonate	[Na+].CCCCCCCCCCCC*.[O-]S(=O)(=O)C1=CC=CC=C1 |c:18,20,t:16,lp:14:3,16:2,17:2,m:13:21.22.23|
DTXSID3025382	Hexabromobiphenyl	Cl*.Cl*.C1=CC=C(C=C1)C1=CC=CC=C1 |c:2,4,6,11,13,t:9,lp:0:3,2:3,m:1:11.12.13.14.15,3:4.5.6.8.9,Sg:n:0:x:ht,Sg:n:2:y:ht|
DTXSID9025295	2-Ethylhexenal	CC[*] |$;;_R0$,RG:_R0={"C(CCC=C)C=O |$_AP1;;;;;;;$,lp:7:2|},{CCC=CC(*)C=O |$;;;;;_AP1;;$,lp:7:2|},{CCCCC(*)=C=O |$;;;;;_AP1;;$,lp:7:2|},{CC=CCC(*)C=O |$;;;;;_AP1;;$,lp:7:2|},{CCCC=C(*)C=O |$;;;;;_AP1;;$,lp:7:2|}|
DTXSID6024129	Hexabromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;6|}
DTXSID3024233	Nonabromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;9|}
DTXSID8024236	Octabromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;8|}
DTXSID2024246	Pentabromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;5|}
DTXSID8024319	Tetrabromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;4|}
DTXSID50881109	Tribromodiphenyl ether	[*]C1=C([*])C([*])=C(OC2=C([*])C([*])=C([*])C([*])=C2[*])C([*])=C1[*] |$_R1;;;_R1;;;_R1;;;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1;;;_R1$,c:1,8,16,21,t:5,12,lp:7:2,RG:_R1={Br* |$;_AP1$,lp:0:2|},LOG={_R1:;H;3|}
DTXSID6027923	Dodecylbenzenesulfonic acid	CCCCCCCCCCCC*.OS(=O)(=O)C1=CC=CC=C1 |c:18,20,t:16,lp:13:2,15:2,16:2,m:12:20.21.22|

Conclusions

- Use of Markush for UVCB enables linking
- A lot of UVCB cannot be depicted with Markush
- Integration of Markush Technology brings its own problems
  - Best depiction vs current enumeration capabilities
  - Determining uniqueness
  - Actually understanding the chemical substance
  - Overcoming limitations through willpower
  - Sharing Markush

Acknowledgements

EPA NCCT IT
Jeff Edwards
Jeremy Dunne

EPA NCCT Curation
David McKee
Inthirany Thillainadarajah
Sakuntala Sivasupramaniam

EPA NERL Curation
Brian Meyer

United States Environmental Protection Agency

Questions?

United States
Environmental Protection
Agency

United States
Environmental Protection
Agency

Conclusion

- The CompTox Chemistry Dashboard is the public interface into DSSTox chemical data (and everything that we can associate)
- Semi-automated curation is effective for expanding our data, while allowing appropriate correction via manual interaction
- We take care of our data
- We let you know how well we take care of our data
- We want your feedback to help us take care of our data!