DSSTox: The Open Environmental Chemistry Data underlying the CompTox Chemical Dashboard

Chris Grulke†
Antony Williams
Ann Richard

NCCT, U.S. EPA

OpenTox USA 2017

12-13 July 2017, Durham, NC

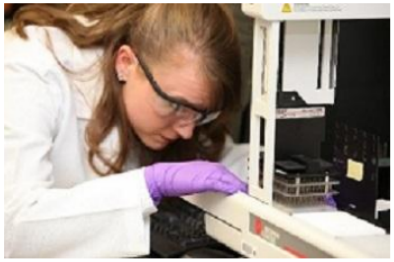
Office of Research and Development

NCCT's Purpose

- NCCT outputs: include a lot of data, models, algorithms and software applications
- We produce Open Data – we want people to interrogate it, learn from it, develop understanding

Toxicity Forecasting

Advancing the Next Generation of Chemical Evaluation



EPA needs rapid and efficient methods to prioritize, screen and evaluate thousands of chemicals. EPA's Toxicity Forecaster (ToxCast) generates data and predictive models on thousands of chemicals of interest to the EPA. ToxCast uses high-throughput screening methods and computational toxicology approaches to rank and prioritize chemicals. In fact, EPA's Endocrine Disruption Screening Program (EDSP) is working to use ToxCast to rank and prioritize chemicals.

- ToxCast has data on over 1,800 chemicals from a broad range of sources including industrial and consumer products, food additives, and potentially "green" chemicals that could be safer alternatives to existing chemicals.
- ToxCast screens chemicals in over 700 high-throughput assays that cover a range of high-

Downloadable Computational Toxicology Data

EPA's computational toxicology research efforts evaluate the potential health effects of thousands of chemicals. The process of evaluating potential health effects involves generating data that investigates the potential harm, or hazard of a chemical, the degree of exposure to chemicals as well as the unique chemical characteristics.

As part of EPA's commitment to share data, all of the computational toxicology data is publicly available for anyone to access and use.

High-throughput Screening Data

EPA researchers use rapid chemical screening (called high-throughput screening assays) to limit the number of laboratory animal tests while quickly and efficiently testing thousands of chemicals for potential health effects.

- ToxCast Data: High-throughput screening data on thousands of chemicals.

Rapid Exposure and Dose Data

EPA researchers develop and use rapid exposure estimates to predict potential exposure for thousands of chemicals.

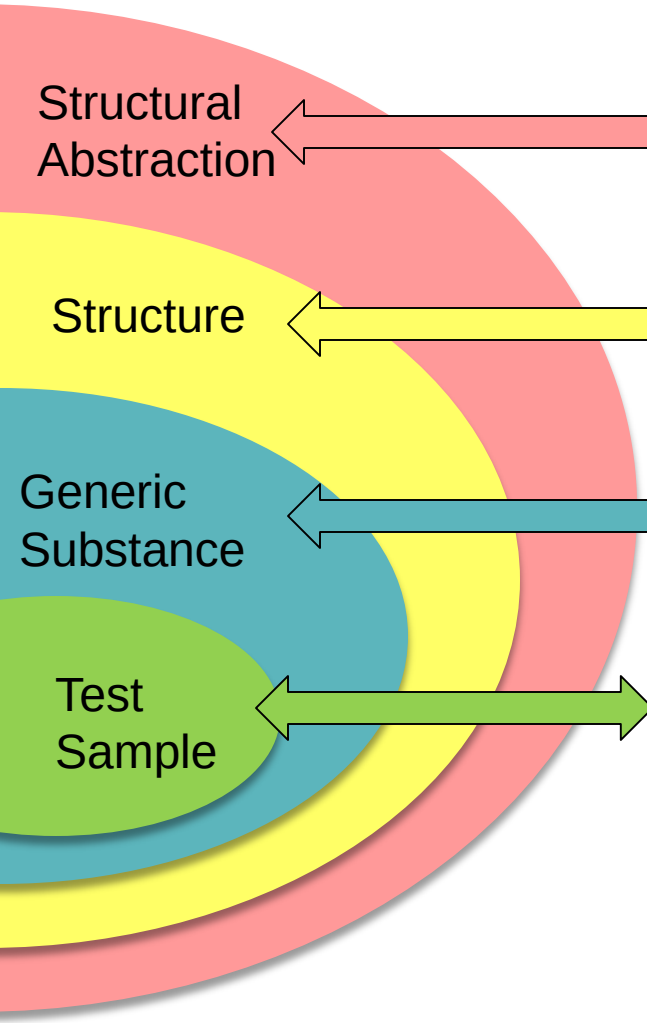
- High-throughput toxicokinetics data: It is important to link the external dose of a chemical to an internal blood or tissue concentration, this process is called toxicokinetics. EPA researchers measure the critical factors that determine the distribution

EPA's Challenge

- Inventory was initially published in 1979
- Second version, containing about 62k chemical substances, was published in 1982
- Continues to grow and now lists ~85k chemicals, about 15k are confidential business information

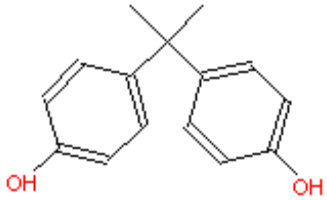
https://www.epa.gov/tsca-inventory/how-access-tsca-inventory

Chemical representation levels supporting data integration



Chemotypes, fingerprints, physchem properties, ...

Chemical analogs, Read-across, SAR modeling



SMILES InChI

Structure searching & modeling

Chemical Name CASRN

Public toxicity datasets

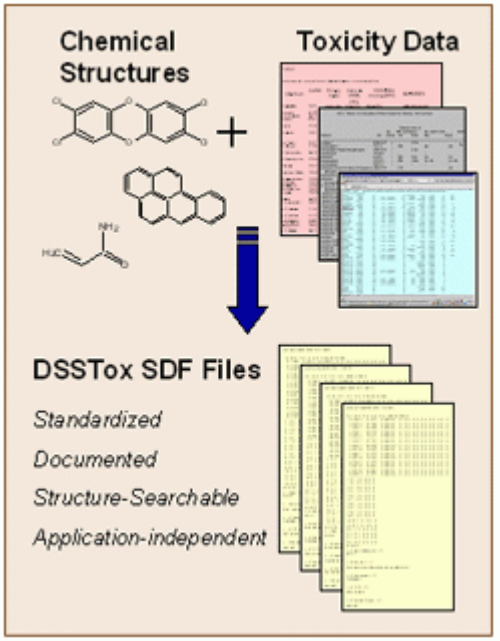
Supplier, Lot/Batch, physical description

Experimental Endpoint Data

DSSTox 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...

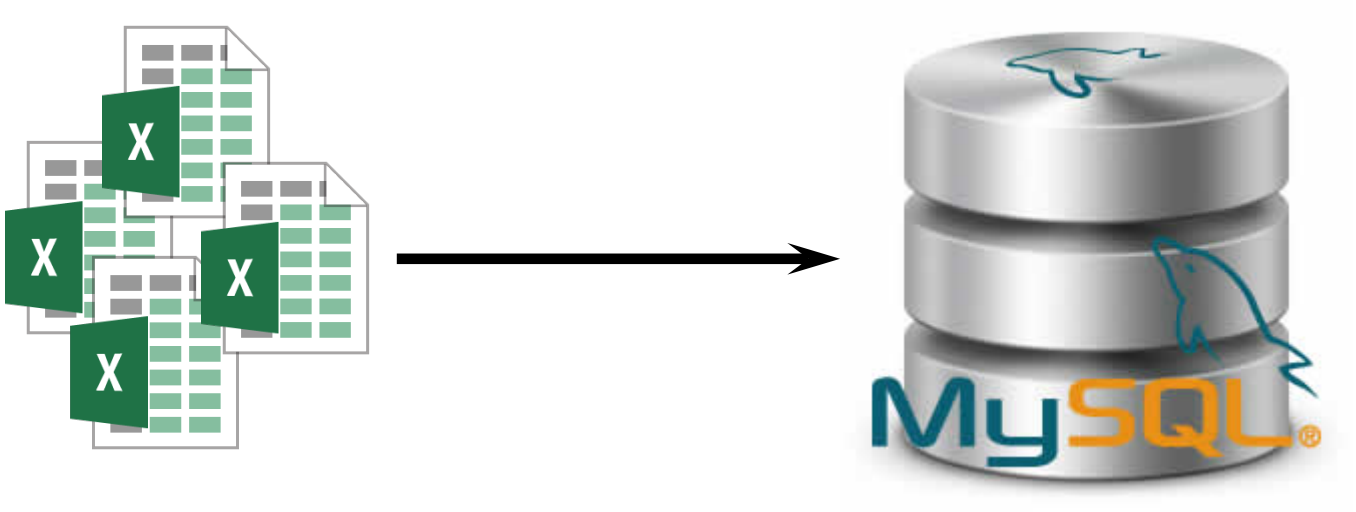


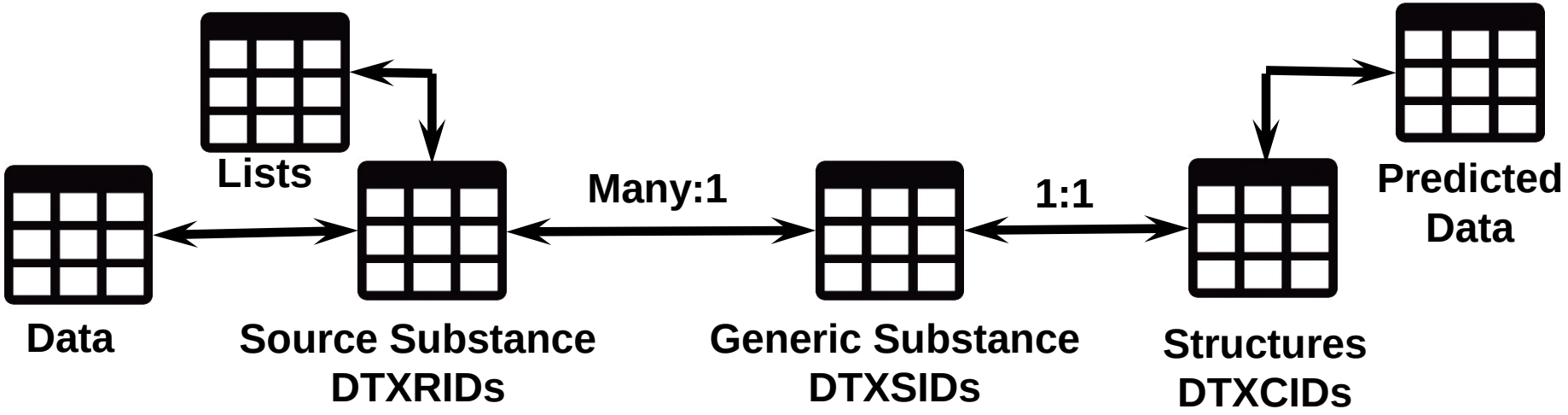
- Managed all chemical registration for ToxCast and Tox21 chemicals
- By 2014, roughly 20K manually curated substance records

DSSTox's Strengths and Weakness (circa 2014)

- Strengths
  - Data as accurate as humanly possible
  - Structure searching within subsets of data
  - 1:1:1 relationships between names, CASRNs, and structures
  - 3 table data model: list records, generic substances, compound structures
- Weaknesses
  - Too many spreadsheets
  - Too small
  - Too manual

New Storage Architecture





DSSTox's Strengths and Weakness (circa 2014)

- Strengths
  - Data as accurate as humanly possible
  - Structure searching within subsets of data
  - 1:1:1 relationships between names, CASRNs, and structures
  - 3 table data model: list records, generic substances, compound structures
- Weaknesses
  - Too many spreadsheets
  - Too small
  - Too manual

Example Data Load (EPASRS)

	EPA SRS											
Total Records	76944											
Internal Structure Check	All inchikeys agree			There is an agreement			There is only 1 inchikey			There is disagreement	No inchikey	
	16527			2815			11180			9702	36620	
Other Source Checks	agree	disagree	no structure	agree	disagree	no structure	agree	disagree	no structure		no smiles/name same	other
	15326	816	385	1923	719	173	5858	3598	1724		17378	19242

SRS Internal Disagreement
= 9702/76944 = 12.5%

SRS External Disagreement
=7415/30522 = 24.3%

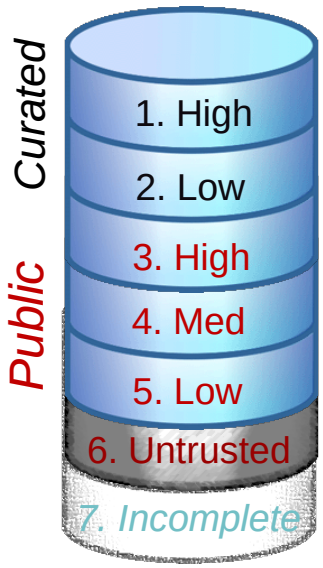
Example Data Load (EPASRS)

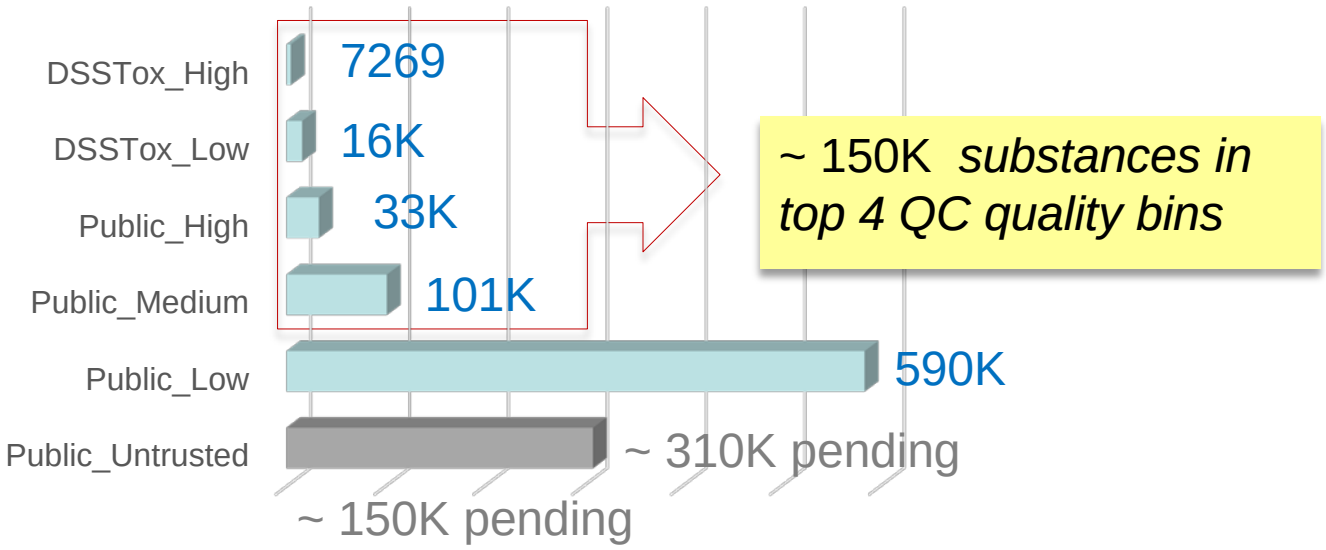
Passing other source checks	With Structure						No Structure			
	23107						17378			
DSSTox Checks	agree	disagree	no structure data	inchi match different casrn	name mismatch	no record	same name (or synonym)	different name	name maps to different casrn	no record
	5703	304	46	144	8	16902	44	1109	72	16153

DSSTox / SRS Disagreement
=502/6205 = 8.1%

WARNING: Always consider the accuracy/consistency of chemical data

DSSTox current content





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

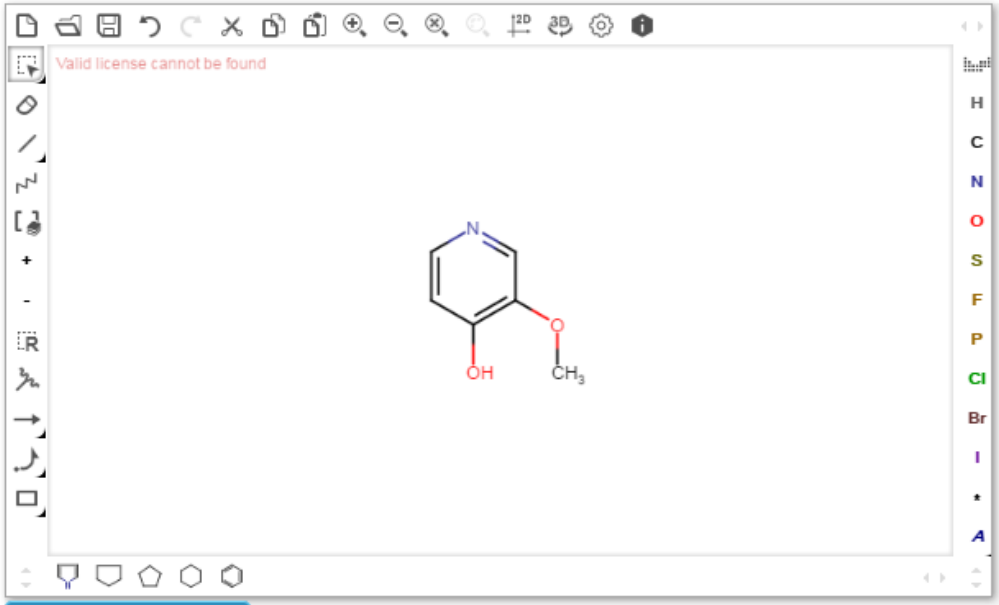
Keeping Manual Curation Alive

United States Environmental Protection Agency

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

Preferred Name matched null
You are viewing the record associated with
DTXSID80198757
CASRN: 62885-41-0

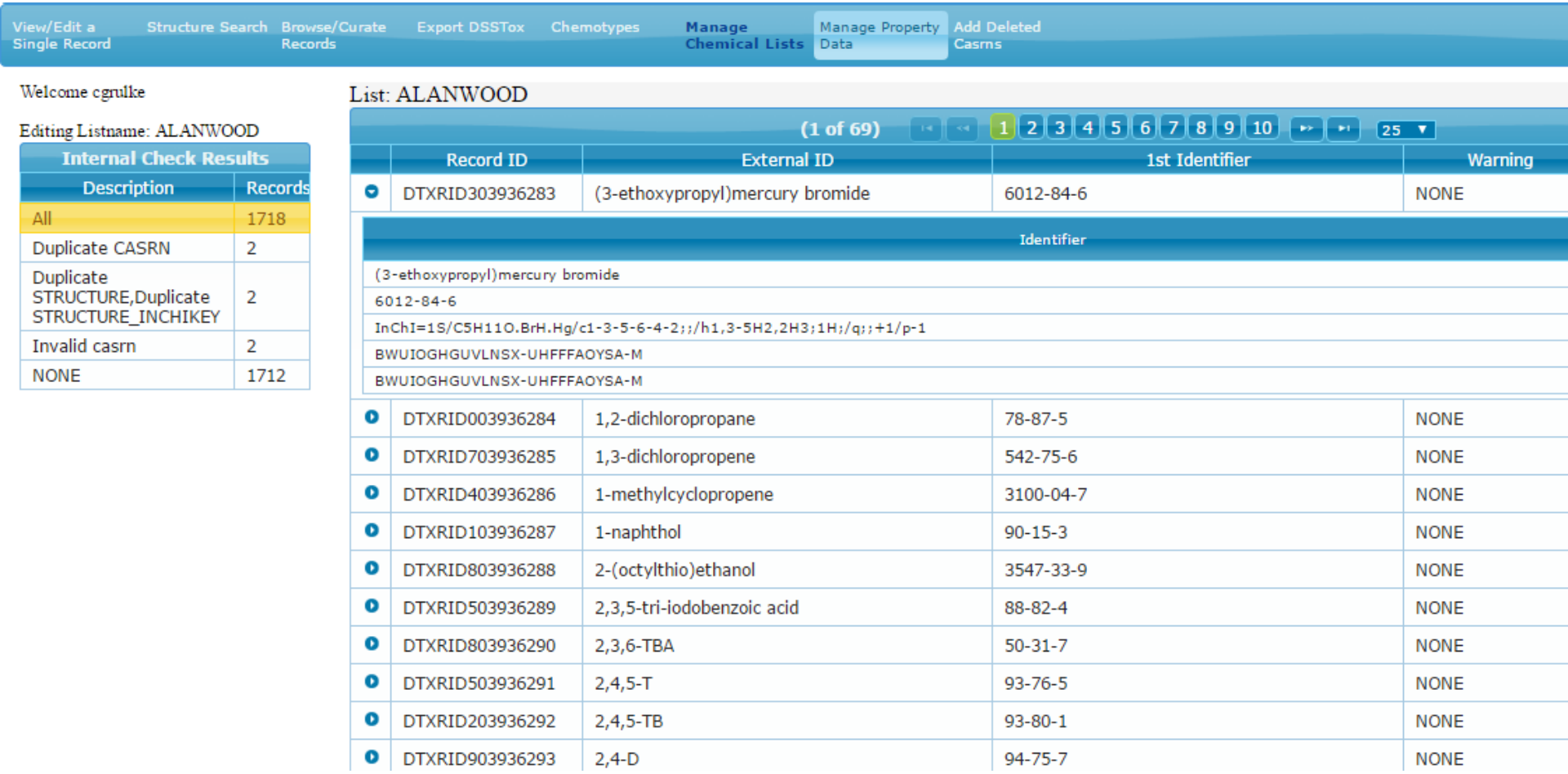
4-Hydroxy-3-methox



Calculate from Structure

Substance_ID:	DTXSID80198757	Compound_ID:	DTXCID40121248
CAS:	62885-41-0	Chemical Shown:	Tested Chemical
Name:	4-Hydroxy-3-methoxypyridine	Private Notes:	
Substance Type:	Single Compound		
QC Level:	DSSTox_High		
Data Source:	STN(DSSTox)	Source of CAS-Compound:	STN(DSSTox)
QC Notes:	CAS [50700-60-2] assigned by DSSTox to pyridin-one tautomer form, which resolves to hydroxy form thru InChI	Double Stereo:	None
		Chiral Stereo:	None
		Chemical Form:	Organic
		Organic Form:	Parent

Internal List Conflicts: ChemReg List Curation



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

Welcome cgrulke

Editing Listname: ALANWOOD

Internal Check Results

Description	Records
All	1718
Duplicate CASRN	2
Duplicate STRUCTURE,Duplicate STRUCTURE_INCHIKEY	2
Invalid casrn	2
NONE	1712

List: ALANWOOD

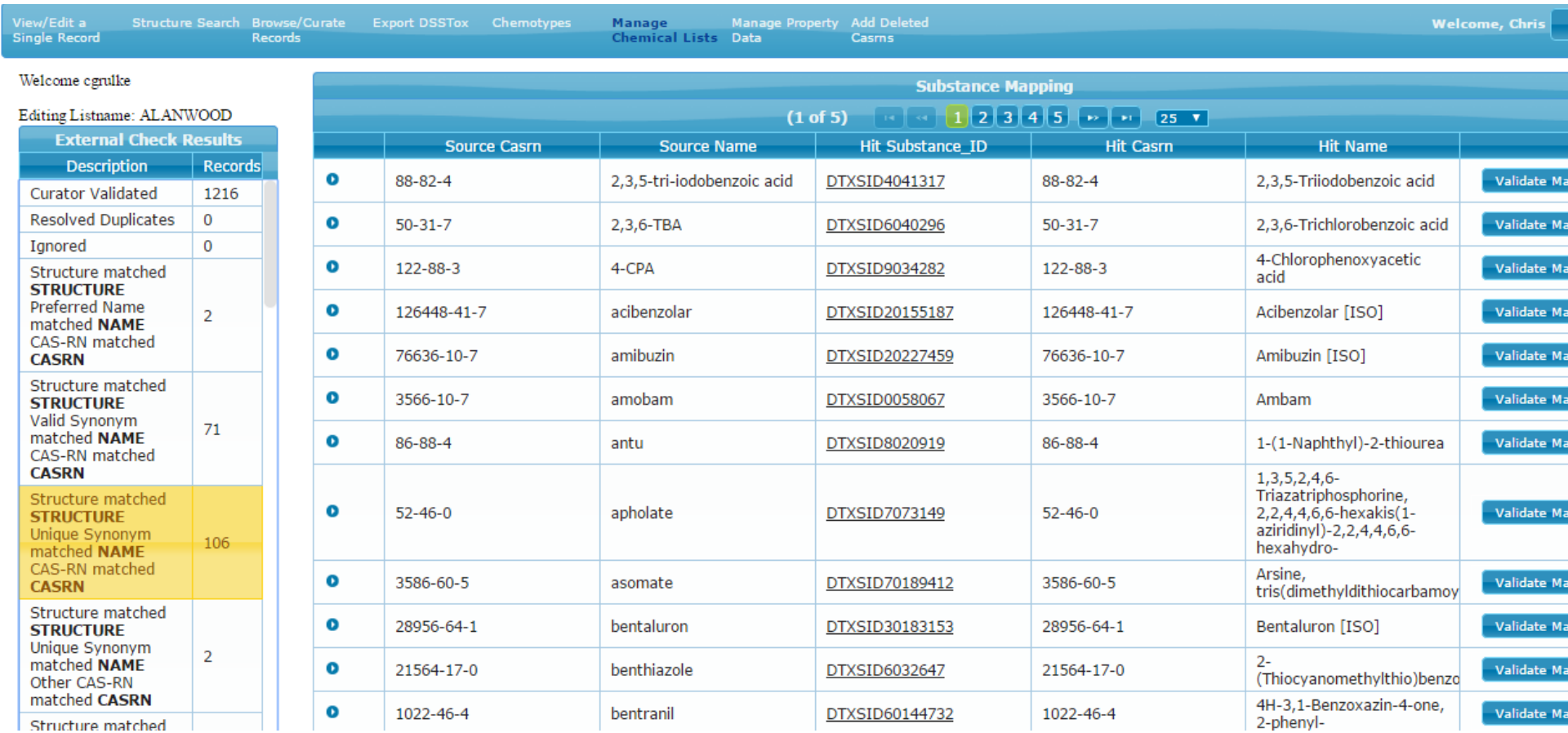
(1 of 69) 1 2 3 4 5 6 7 8 9 10 25

Record ID	External ID	1st Identifier	Warning
DTXRID303936283	(3-ethoxypropyl)mercury bromide	6012-84-6	NONE

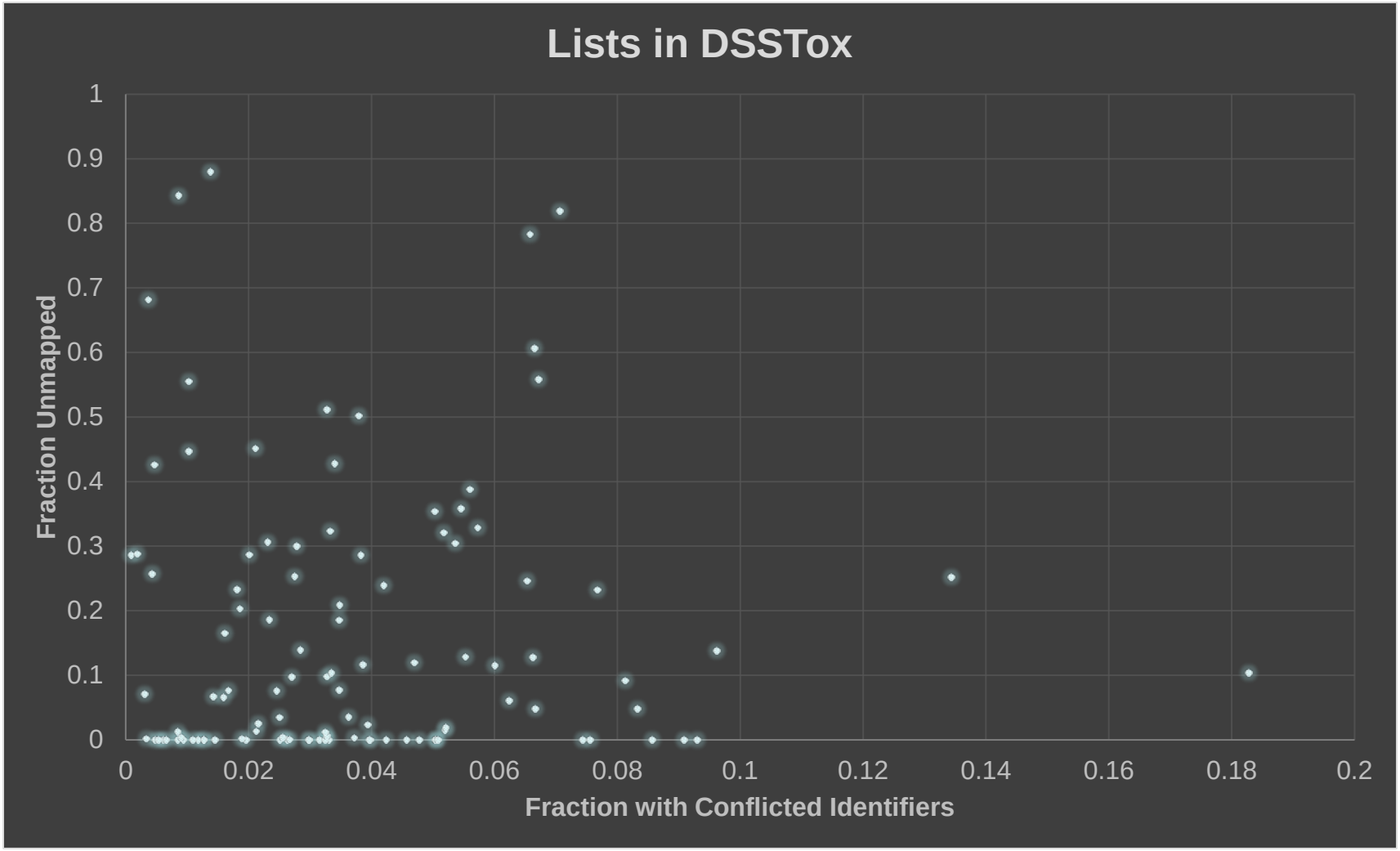
Identifier
(3-ethoxypropyl)mercury bromide
6012-84-6
InChI=1S/C5H11O.BrH.Hg/c1-3-5-6-4-2;;/h1,3-5H2,2H3;1H;/q;;+1/p-1
BWUIOGHGUVLNSX-UHFFFAOYSA-M
BWUIOGHGUVLNSX-UHFFFAOYSA-M

DTXRID003936284	1,2-dichloropropane	78-87-5	NONE
DTXRID703936285	1,3-dichloropropene	542-75-6	NONE
DTXRID403936286	1-methylcyclopropene	3100-04-7	NONE
DTXRID103936287	1-naphthol	90-15-3	NONE
DTXRID803936288	2-(octylthio)ethanol	3547-33-9	NONE
DTXRID503936289	2,3,5-tri-iodobenzoic acid	88-82-4	NONE
DTXRID803936290	2,3,6-TBA	50-31-7	NONE
DTXRID503936291	2,4,5-T	93-76-5	NONE
DTXRID203936292	2,4,5-TB	93-80-1	NONE
DTXRID903936293	2,4-D	94-75-7	NONE

ChemReg List Curation (cont.)



Why Curation is Essential



Chemical Families (e.g. PCBs)

United States Environmental Protection Agency

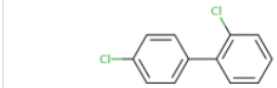
Successor Substances (209)

	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

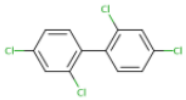
Related Chemicals

Found 209 chemicals

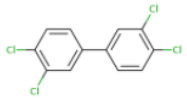
Download as: TSV Excel SDF



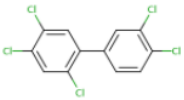
2,4'-Dichlorobiphenyl
34883-43-7



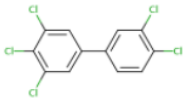
2,2',4,4'-Tetrachlorobiphenyl
2437-79-8



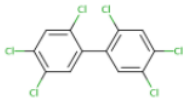
3,3',4,4'-Tetrachlorobinphenyl
32598-13-3



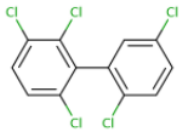
2,3',4,4',5-Pentachlorobiphenyl
31508-00-6



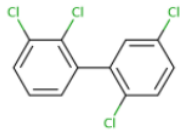
3,3',4,4',5-Pentachlorobiphenyl
57465-28-8



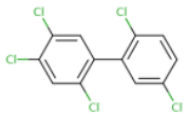
2,2',4,4',5,5'-Hexachlorobiphenyl
35065-27-1



2,2',3,5',6-Pentachlorobiphenyl
38379-99-6



2,2',3,5'-tetrachlorobiphenyl
41464-39-5



2,2,4,5,5'-Pentochlorobiphenyl
37680-73-2

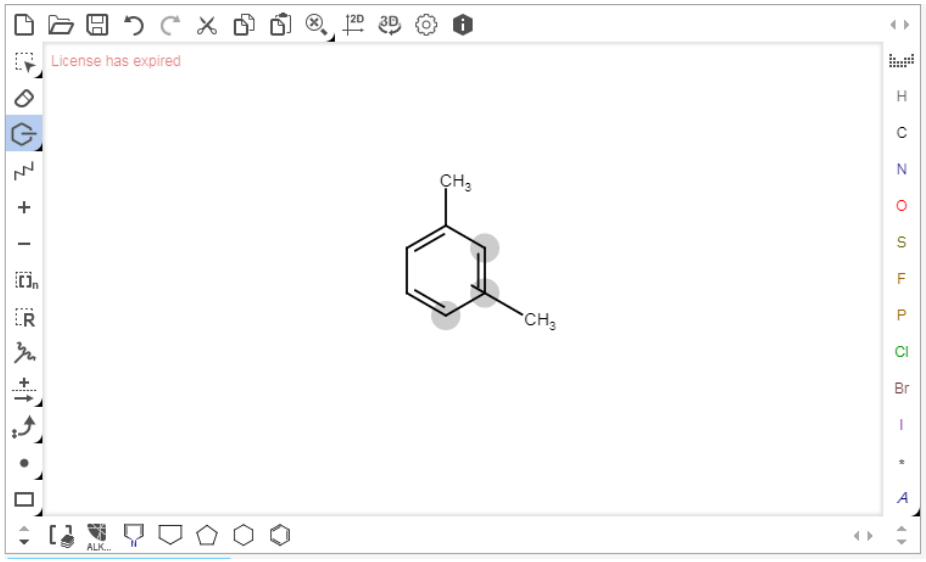
Next Steps: How to Work with UCVBs

Xylene

Successor Substances (3)

	CAS-RN	Relationship	Source	St
●	108-38-3	is a Representative Isomer of this	STN(DSSTox)	✓
●	95-47-6	is a Representative Component o	STN(DSSTox)	✓
●	106-42-3	is a Representative Component o	STN(DSSTox)	✓

Delete Selected | Add Related Cas

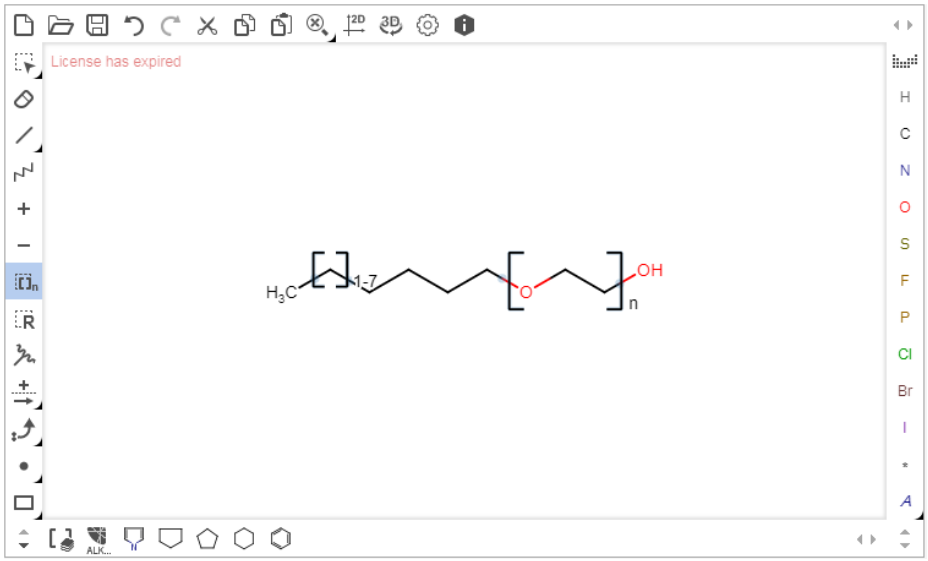


Alcohols, C6-12, ethoxylated

Successor Substances (1)

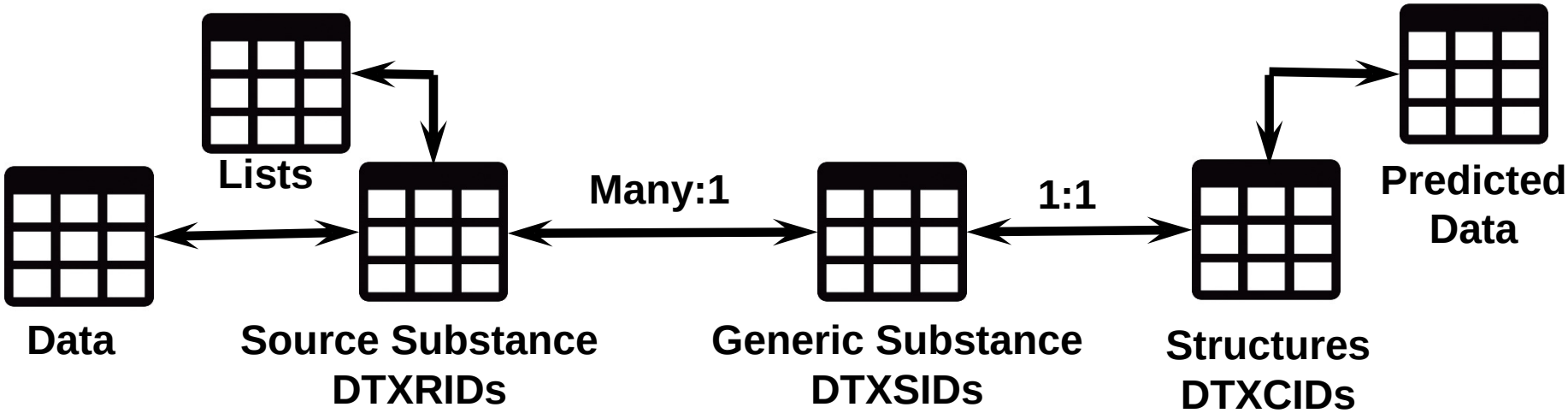
	CAS-RN	Relationship	Source	Struct	Casrn	Cor
●	18913-04-7	is a Representative Component o	STN(DSSTox)	✓	■	C9, monoetho

Delete Selected | Add Related Cas



Conclusion

- 3-Layer data model



- Be careful with public chemical data
- Please take and use our data!

Acknowledgements

EPA NCCT
Imran Shah
Chris Grulke
Jeff Edwards
Ann Richard
Jordan Foster
Jennifer Smith
Richard Judson
Grace Patlewicz
John Wambaugh
Michelle Krzyzanowski