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

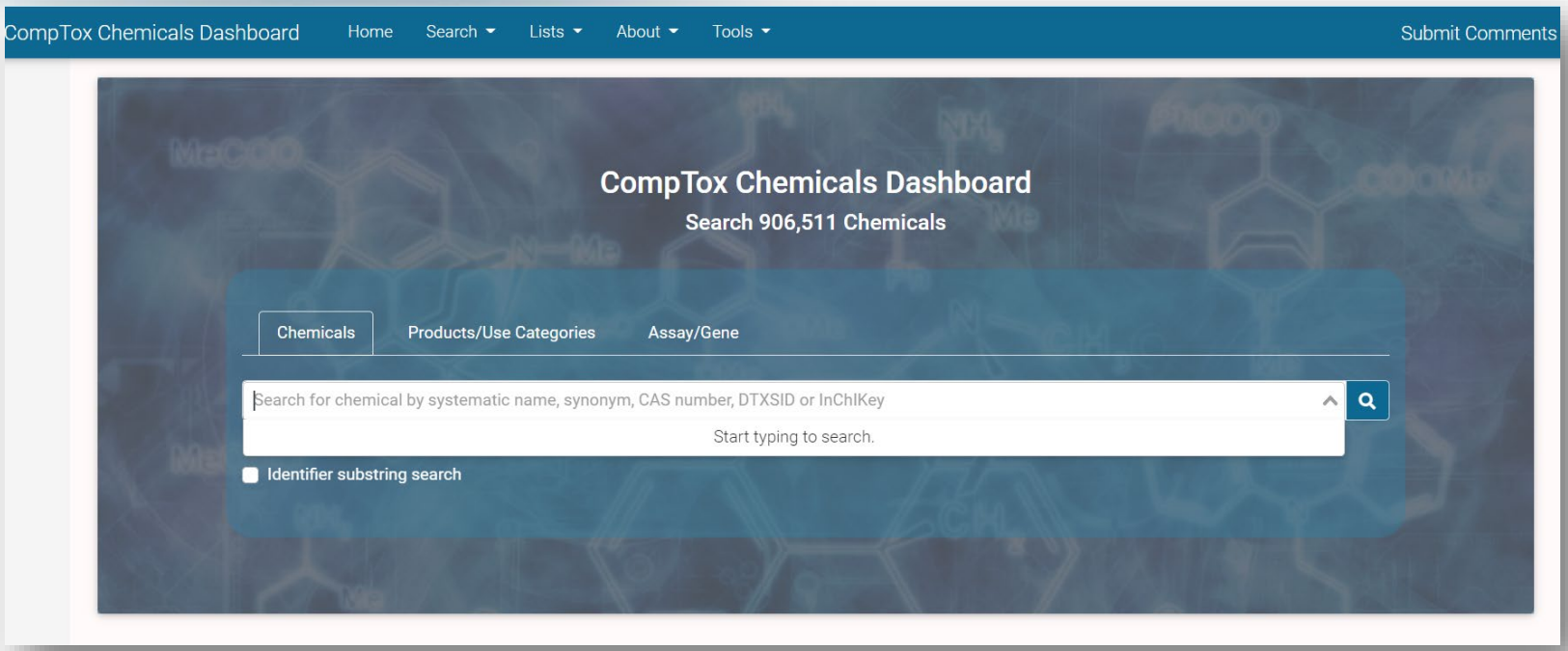
Structure standardization approaches for mass spectrometry data integration

Antony J. Williams[1], *Charles Lowe*[1], *Gabriel Sinclair*[2], *Todd Martin*[1] *and Valery Tkachenko*[3]

1 Center for Computational Toxicology & Exposure, U.S. Environmental Protection Agency,
2 ORAU Student Services Contractor
3 Science Data Experts

August 2022: ACS Chicago

CompTox Chemicals Dashboard



- Structure standardization is essential for the Dashboard
- Standardization underpins many pieces of functionality
- “MS-Ready” structures are the foundation of our support for mass spectrometry

Different flavors of standardization

- We have standardization approaches at the time of chemical registration
- Our ChemReg system is based on ChemAxon tools so we have SMILES standardization and both Jchem InChIs and StdInChIs are generated
- But we also add “QSAR- and MS-Ready” structures based on **OPERA**

OPERA

- Originally developed by Kamel Mansouri while he was postdoc'ing at EPA

OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri, Chris M. Grulke, Richard S. Judson & Antony J. Williams

Journal of Cheminformatics **10**, Article number: 10 (2018) | Cite this article

13k Accesses | **156** Citations | **25** Altmetric | Metrics

Research article | Open Access | Published: 18 September 2019

Open-source QSAR models for pKa prediction using multiple machine learning approaches

Kamel Mansouri, Neal F. Cariello, Alexandru Korotcov, Valery Tkachenko, Chris M. Grulke, Catherine S. Sprankle, David Allen, Warren M. Casey, Nicole C. Kleinstreuer & Antony J. Williams

Journal of Cheminformatics **11**, Article number: 60 (2019) | Cite this article

21k Accesses | **42** Citations | **20** Altmetric | Metrics

OPERA has standardization

https://github.com/kmansouri/OPERA

- OPERA is open source and under maintenance and expansion with new data and models

kmansouri / OPERA Public

Notifications | Fork 31 | Star 54

Code | Issues 6 | Pull requests | Actions | Projects | Security | Insights

master | 1 branch | 34 tags | Go to file | Code

Mansouri v2.8.4 — 6db6c16 on Jun 17 — 165 commits

OPERA_Source_code	v2.8.4	2 months ago
Icon.png	OPERA 1.2 icon	5 years ago
Install_guide.pdf	v2.7-beta1	14 months ago
LICENSE	Initial commit	6 years ago
Logo.png	Added logo and icon	6 years ago
OPERA1.5_Source_code.zip	MATLAB source code for OPERA1.5	4 years ago
OPERA2.0_Source_code.zip	MATLAB source code for OPERA 2.0	4 years ago
OPERA_Data.zip	v2.8.1	5 months ago
OPERA_models_2.8.xlsx	v2.8.1	5 months ago

About

Free and open-source application (command line and GUI) providing QSAR models predictions as well as applicability domain and accuracy assessment for physicochemical properties, environmental fate and toxicological endpoints.
==========>Download the latest compiled version from the "releases" tab and run the executable installer.

Readme
MIT license
54 stars
8 watching

OPERA standardization

- QSAR-Ready standardization is to support QSAR modeling – i.e., remove stereocenters, desalting, ignore multicomponent chemicals

- MS-Ready is a derivative work of QSAR-Ready and has different rules for processing chemicals ESPECIALLY for multi-component chemicals and the handling of certain organometallics

"MS-ready" structures

McEachran *et al. J Cheminform (2018) 10:45*
https://doi.org/10.1186/s13321-018-0299-2

Journal of Cheminformatics

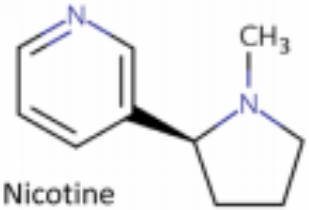
METHODOLOGY

Open Access

"MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

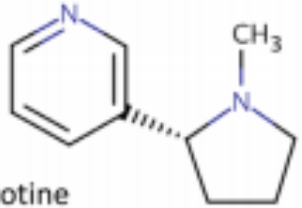
Andrew D. McEachran[1,2*], Kamel Mansouri[1,2,3], Chris Grulke[2], Emma L. Schymanski[4], Christoph Ruttkies[5] and Antony J. Williams[2*]

LEGEND: Name, SMILES
DTXSID | InChIKey 1st Block
CAS | **Monoiso. Mass** | logP | **Sources**
Data on: **Toxicity** | **Exposure** | **Bioassays**



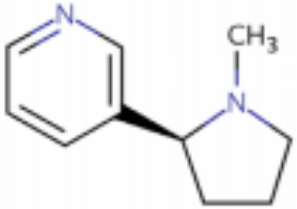
Nicotine

CN1CCC[C@H]1C1=CN=CC=C1
DTXSID1020930| SNICXCGAKADSCV
54-11-5 | **162.1157**| 0.929| **72**
Tox: **yes**| Expo: **yes**| Bioassay: **yes**



D-Nicotine

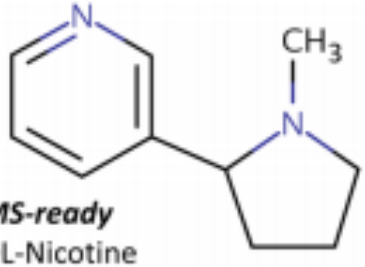
CN1CCC[C@@H]1C1=CN=CC=C1
DTXSID004635| SNICXCGAKADSCV
25162-00-9 | **162.1157**| 0.929| **20**
Tox: **no**| Expo: **yes**| Bioassay: **yes**



HCl

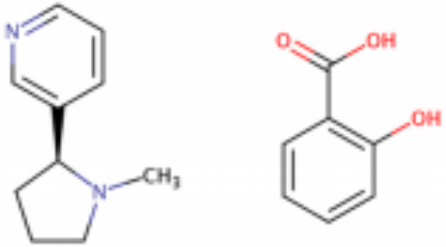
Nicotine hydrochloride

Cl.CN1CCC[C@H]1C1=CN=CC=C1
DTXSID602093| HDJBTCAJIMNXEW
2820-51-1 | **198.0924** | 0.929| **9**
Tox: **no**| Expo: **yes**| Bioassay: **yes**



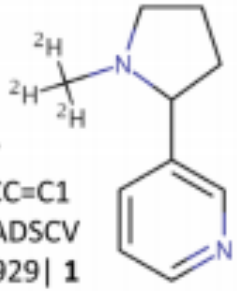
MS-ready
DL-Nicotine

CN1CCCC1C1=CN=CC=C1
DTXSID3048154 | SNICXCGAKADSCV
22083-74-5 | **162.1157**| 0.953| 9
Tox: **yes**| Expo: **no**| Bioassay: **yes**



Benzoic acid, 2-hydroxy-, compd. with
3-[(2S)-1-methyl-2-pyrrolidinyl]pyridine (1:1)

OC(=O)C1=C(O)C=CC=C1.CN1CCC[C@H]1C1=CN=CC=C1
DTXSID5075319| AIBWPBUAKCMKNS
29790-52-1| **300.1474**| 0.929| **6**
Tox: **no**| Expo: **yes**| Bioassay: **no**



DL-Nicotine-d3

[2H]C([2H])([2H])N1CCCC1C1=CN=CC=C1
DTXSID80442666| SNICXCGAKADSCV
69980-24-1| **165.1345**| 0.929| **1**
Tox: **no**| Expo: **no**| Bioassay: **no**

Environmental Science & Technology
Viewpoint
pubs.acs.org/est

Open Science for Identifying "Known Unknown" Chemicals

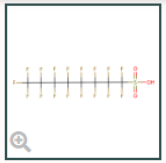
Emma L. Schymanski and Antony J. Williams

Overview of MS-Ready Structures

- All structure-based chemical substances are algorithmically processed to
  - Split multicomponent chemicals into individual structures
  - Desalt and neutralize individual structures
  - Remove stereochemical bonds from all chemicals

- MS-Ready structures are then mapped to original substances to provide a path between chemicals detected by mass spectrometry to original substances

MS-Ready Mappings from Details Page

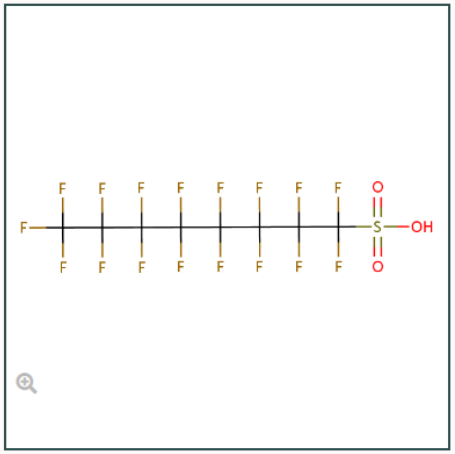


Perfluorooctanesulfonic acid

1763-23-1 | DTXSID3031864

Searched by DTXSID3031864.

Chemical Details



- Wikipedia
- Quality Control Notes
- Intrinsic Properties
- Structural Identifiers
- Linked Substances

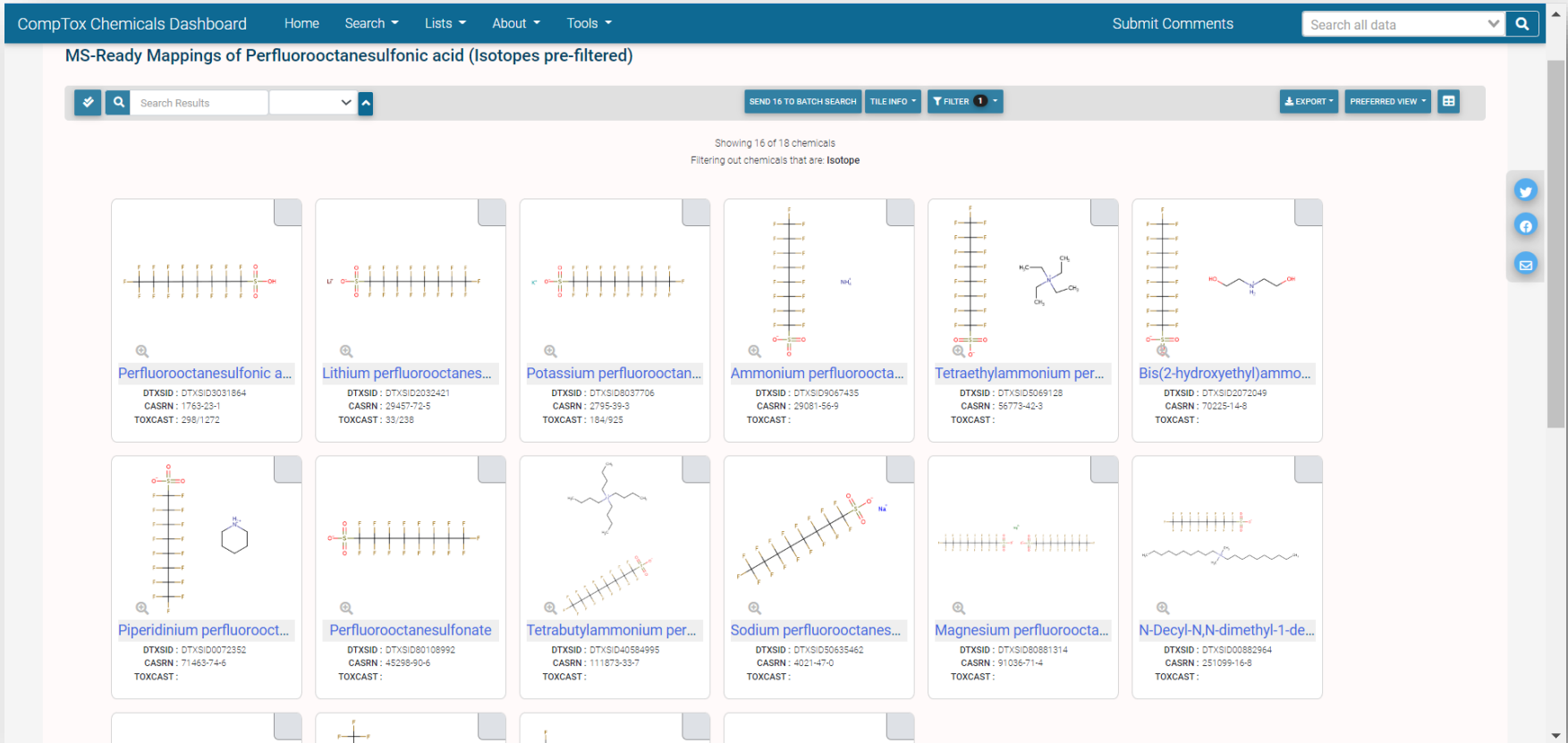
Same Connectivity: 4 records (based on first layer of InChI)

Mixtures, Components and Isotopomers: DTXCID1011864: 18 records;

Similar Compounds: 127 records (based on Tanimoto coefficient >0.8)

- Presence in Lists
- Record Information

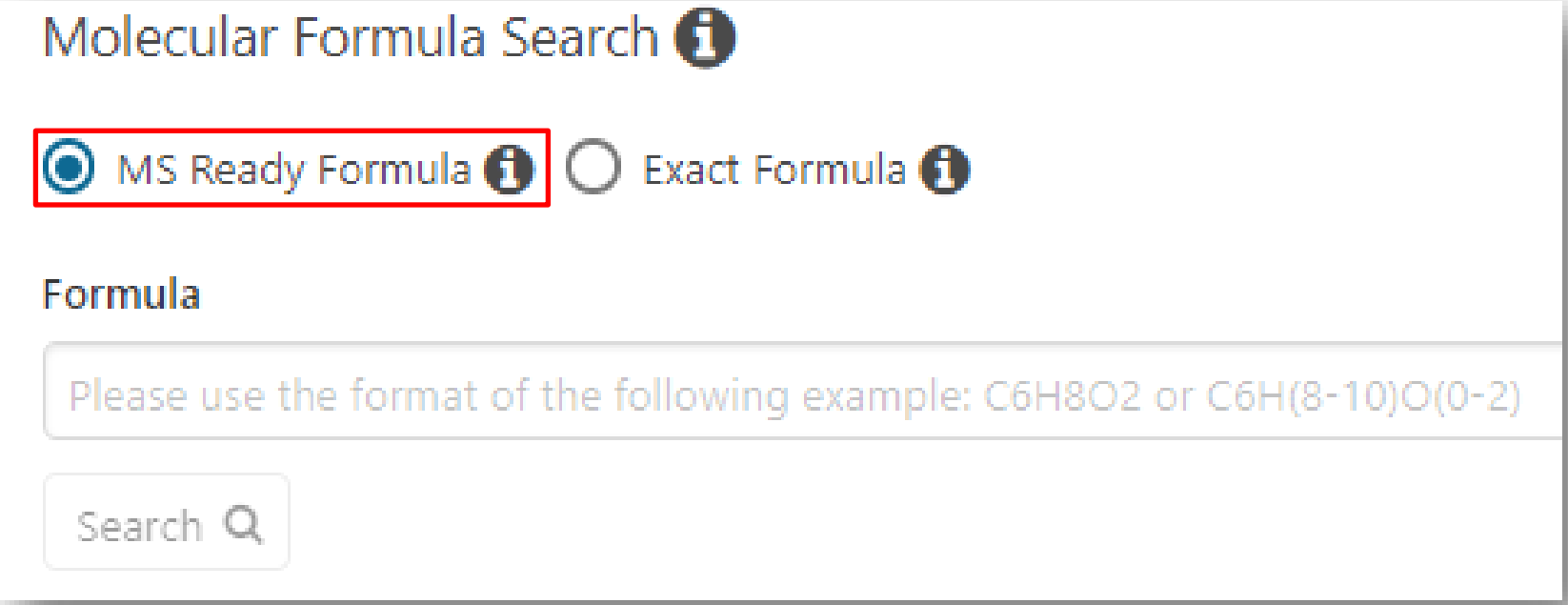
MS-Ready Mappings
Set of substances for “PFOS”



MS-Ready mappings support structure identification

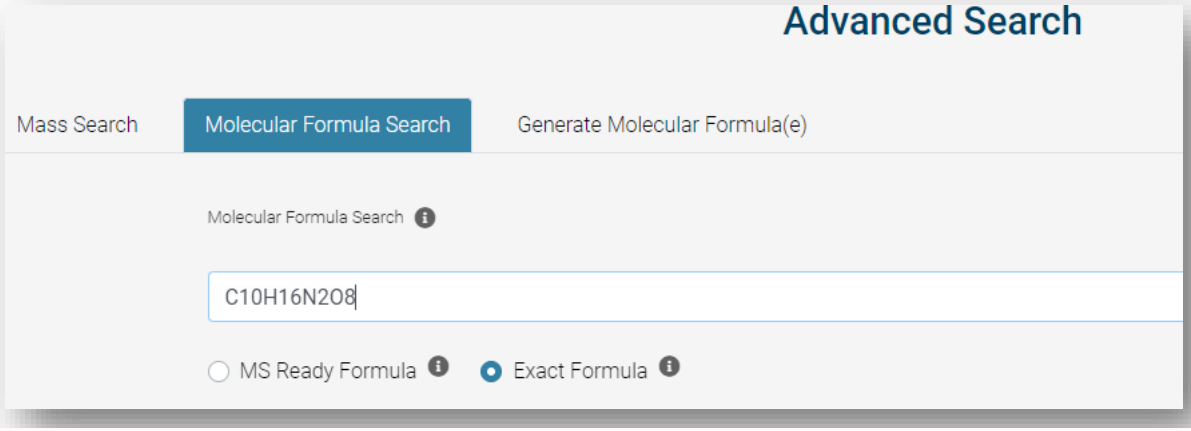
- Mass spectrometry detects single components and is non-stereospecific so all multi-component forms of PFOS are “equivalent” for detection
- Not all forms of PFOS are equivalent in commerce, or availability of data, literature, and standards
- Mass/formula searches find the component.MS-Ready mapping finds everything with the component

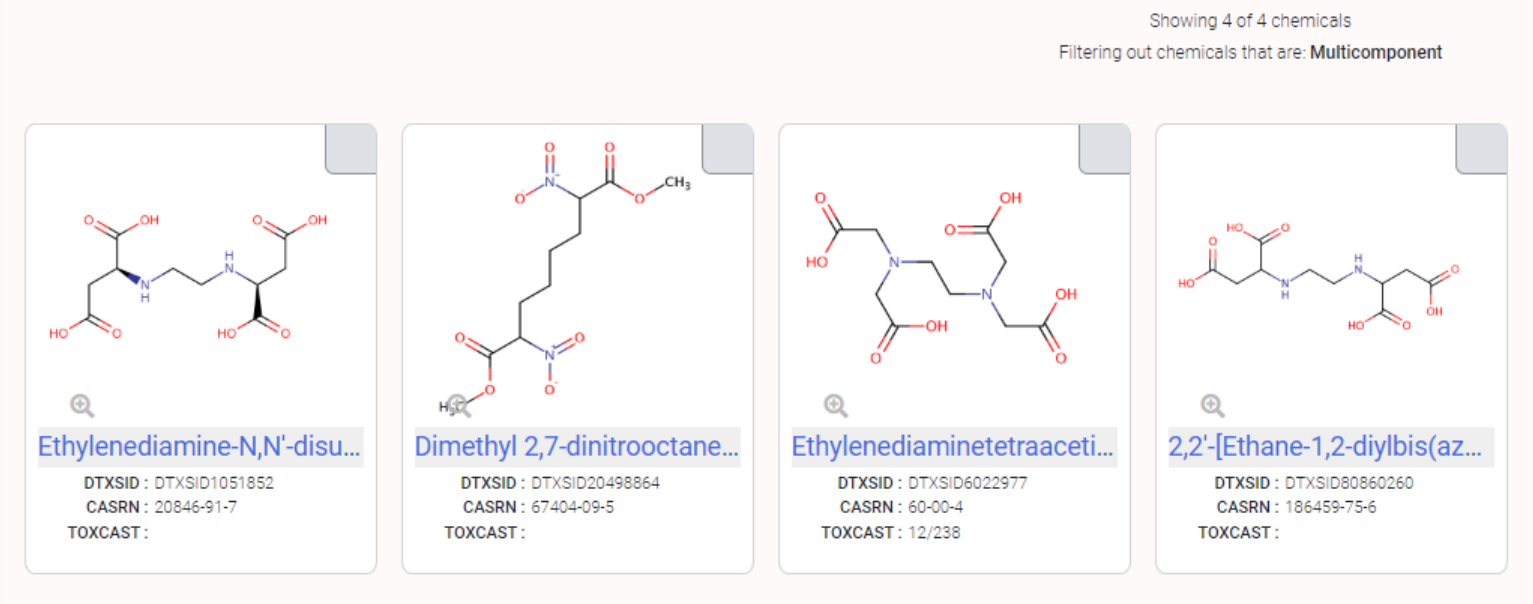
MS-Ready Structures for **Formula Search**



MS-Ready Mappings

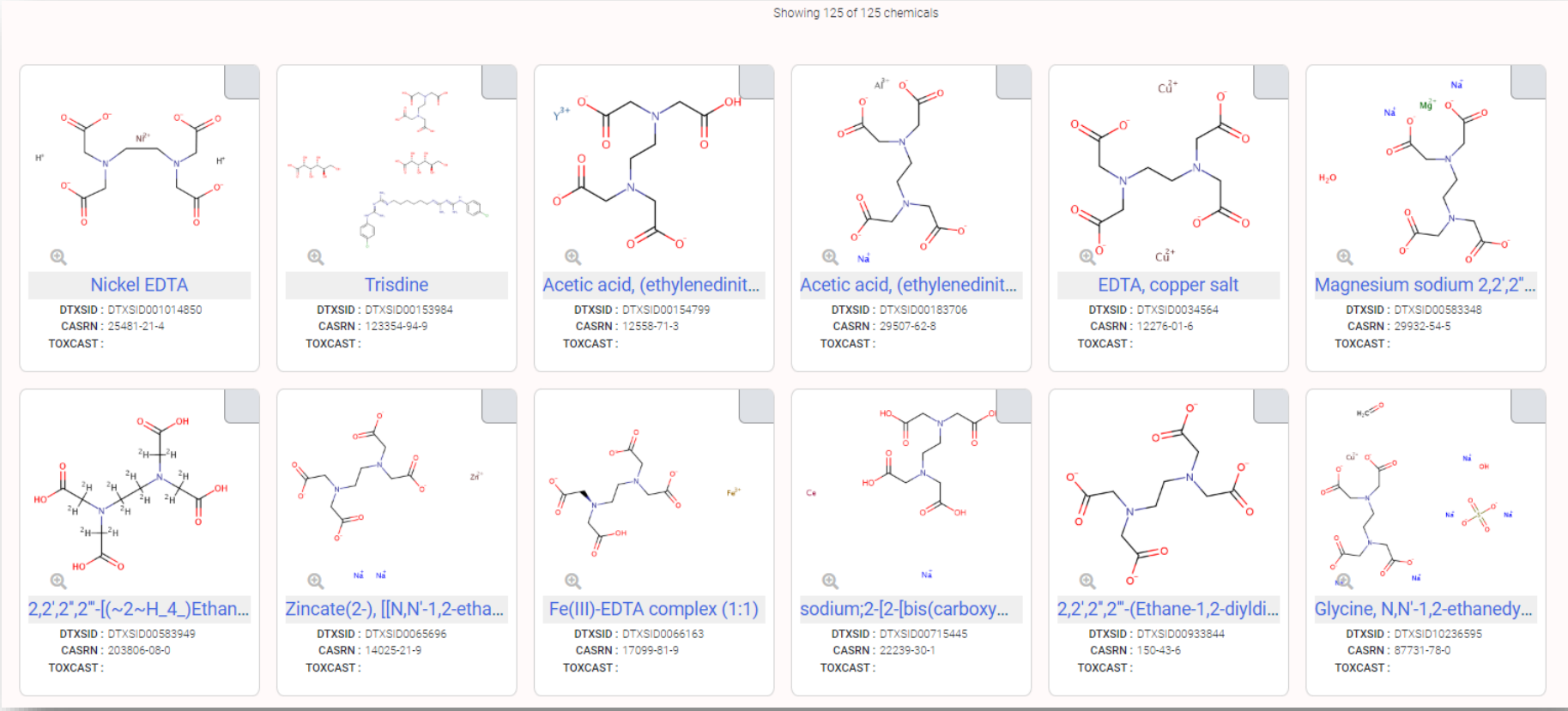
- **EXACT Formula**: C10H16N2O8: 4 Hits





MS-Ready Mappings

- **Same** Input Formula: C10H16N2O8
- **MS Ready Formula** Search: **125** Chemicals



MS-Ready Mappings

- 125 chemicals returned in total
  - 9 of the 125 are **single component** chemicals
  - 4 of the 125 are **isotope-labeled**

- Multiple components, stereo, isotopes and charge all mapped through MS-Ready

Batch Searching mass and formula

Batch Searching

- Singleton searches are useful but we work with **thousands** of masses and formulae!

- Typical questions
  - What is the list of chemicals for the formula $C_xH_yO_z$
  - What is the list of chemicals for a mass +/- error
  - Can I get chemical lists in Excel files? In SDF files?
  - Can I include properties in the download file?

Batch Searching Formula/Mass

Batch Search

1 Select Input Type(s)

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

+/- 5 ppm

This search is based on what we refer to as "Mass Spec (MS) Ready" structures. All chemicals within the database are treated in a manner such that all are desalted, mixtures are separated, and stereochemistry is removed as Mass Spectrometry detects the major components of a salt or mixture and is insensitive to stereochemistry. As an example, a search for the monoisotopic mass of phenol will return phenol, sodium phenolate and calcium phenoxide. See the publication for more details: https://doi.org/10.1186/s13321-018-0299-2.

2 Enter Identifiers to Search

(Please enter one identifier per line and limit the number of identifiers to 10,000 or less)

200.083730
352.146330
318.089209
382.287180
58499
57082
15030
96651

DISPLAY ALL CHEMICALS or CHOOSE EXPORT OPTIONS

Batch Search

1 Select Input Type(s)

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

2 Enter Identifiers to Search

(Please enter one identifier per line and limit the number of identifiers to 10,000 or less)

C14H22N2O3
C18H34N2O6S
C10H12N2O
C14H18N4O3
C12H11N7
C8H9NO2

3 DISPLAY ALL CHEMICALS or CHOOSE EXPORT OPTIONS

Searching batches using MS-Ready Formula (or mass) searching

	A	B	C	D	E	F	G
1	INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS	DATA SOURCES
2	C14H22N2O3	DTXSID2022628	29122-68-7	Atenolol	C14H22N2O3	266.163042576	46
3	C14H22N2O3	DTXSID0021179	6673-35-4	Practolol	C14H22N2O3	266.163042576	32
4	C14H22N2O3	DTXSID4048854	841-73-6	Bucolome	C14H22N2O3	266.163042576	20
5	C14H22N2O3	DTXSID1045407	13171-25-0	Trimetazidine dihydrochloride	C14H24Cl2N2O3	338.116398	19
6	C14H22N2O3	DTXSID0045753	56715-13-0	R-(+)-Atenolol	C14H22N2O3	266.163042576	19
7	C14H22N2O3	DTXSID2048531	5011-34-7	Trimetazidine	C14H22N2O3	266.163042576	14
8	C14H22N2O3	DTXSID10239405	93379-54-5	Esatenolol	C14H22N2O3	266.163042576	12
9	C14H22N2O3	DTXSID50200634	52662-27-8	N-(2-Diethylaminoethyl)-2-(4-hydroxyphenoxy)acetamide	C14H22N2O3	266.163042576	7
10	C14H22N2O3	DTXSID4020111	51706-40-2	dl-Atenolol hydrochloride	C14H23ClN2O3	302.1397203	6
11	C14H22N2O3	DTXSID1068693	51963-82-7	Benzenamine, 2,5-diethoxy-4-(4-morpholinyl)-	C14H22N2O3	266.163042576	5
12	C18H34N2O6S	DTXSID3023215	154-21-2	Lincomycin	C18H34N2O6S	406.213757997	35
13	C18H34N2O6S	DTXSID7047803	859-18-7	Lincomycin hydrochloride	C18H35ClN2O6S	442.1904357	22
14	C18H34N2O6S	DTXSID20849438	1398534-62-7	PUBCHEM_71432748	C18H35ClN2O6S	442.1904357	1
15	C10H12N2O	DTXSID1047576	486-56-6	Cotinine	C10H12N2O	176.094963014	40
16	C10H12N2O	DTXSID8075330	50-67-9	Serotonin	C10H12N2O	176.094963014	22
17	C10H12N2O	DTXSID8044412	2654-57-1	4-Methyl-1-phenylpyrazolidin-3-one	C10H12N2O	176.094963014	18
18	C10H12N2O	DTXSID80165186	153-98-0	Serotonin hydrochloride	C10H13ClN2O	212.0716407	11
19	C10H12N2O	DTXSID2048870	29493-77-4	(4R,5S)-4-methyl-5-phenyl-4,5-dihydro-1,3-oxazol-2-amine	C10H12N2O	176.094963014	10
20	C10H12N2O	DTXSID10196105	443-31-2	6-Hydroxytryptamine	C10H12N2O	176.094963014	9
21	C10H12N2O	DTXSID90185693	31822-84-1	1,4,5,6-Tetrahydro-5-phenoxypyrimidine	C10H12N2O	176.094963014	7
22	C10H12N2O	DTXSID40178777	2403-66-9	2-Benzimidazolepropanol	C10H12N2O	176.094963014	7
23	C10H12N2O	DTXSID80157026	13140-86-8	N-Cyclopropyl-N'-phenylurea	C10H12N2O	176.094963014	6
24	C10H12N2O	DTXSID30205607	570-14-9	4-Hydroxytryptamine	C10H12N2O	176.094963014	6
25	C14H18N4O3	DTXSID5023900	17804-35-2	Benomyl	C14H18N4O3	290.137890456	68
26	C14H18N4O3	DTXSID3023712	738-70-5	Trimethoprim	C14H18N4O3	290.137890456	51
27	C14H18N4O3	DTXSID40209671	60834-30-2	Trimethoprim hydrochloride	C14H19ClN4O3	326.1145682	8
28	C14H18N4O3	DTXSID70204210	55687-49-5	Benzenemethanol, 4-((2,4-diamino-5-pyrimidinyl)methyl)-2,	C14H18N4O3	290.137890456	5
29	C14H18N4O3	DTXSID20152671	120075-57-2	6-Methoxy-4-(3-(N,N-dimethylamino)propylamino)-5,8-quin	C14H18N4O3	290.137890456	4
30	C14H18N4O3	DTXSID30213742	63931-79-3	1H-1,2,4-Benzotriazepine-3-carboxylic acid, 4,5-dihydro-4-	C14H18N4O3	290.137890456	3
31	C14H18N4O3	DTXSID30219608	69449-07-6	2,4-Pyrimidinediamine, 5-((3,4,5-trimethoxyphenyl)methyl)-	C14H20N4O4	308.14845514	3
32	C14H18N4O3	DTXSID20241155	94232-27-6	L-Aspartic acid, compound with 5-((3,4,5-trimethoxyphenyl	C18H25N5O7	423.175398165	3
33	C14H18N4O3	DTXSID80241156	94232-28-7	L-Glutamic acid, compound with 5-((3,4,5-trimethoxypheny	C19H27N5O7	437.191048229	3
34	C14H18N4O3	DTXSID20143781	101204-93-7	1H-Pyrido(2,3-e)-1,4-diazepine-2,3,5-trione, 4-(2-(diethylam	C14H18N4O3	290.137890456	3
35	C12H11N7	DTXSID6021373	396-01-0	Triamterene	C12H11N7	253.107593382	52
36	C12H11N7	DTXSID00204465	5587-93-9	Ampyrimine	C12H11N7	253.107593382	7
37	C12H11N7	DTXSID5064621	7300-26-7	Benzenamine, 4-azido-N-(4-azidophenyl)-	C12H9N7	251.091943318	4
38	C12H11N7	DTXSID00848025	90293-82-6	Sulfuric acid--6-phenylpteridine-2,4,7-triamine (1/1)	C12H13N7O4S	351.074973101	1
39	C12H11N7	DTXSID50575293	92310-83-3	(1E)-N-Phenyl-1,2-bis(1H-1,2,4-triazol-1-yl)ethan-1-imine	C12H11N7	253.107593382	1
40	C8H9NO2	DTXSID2020006	103-90-2	Acetaminophen	C8H9NO2	151.063328534	75
41	C8H9NO2	DTXSID6025567	134-20-3	Methyl 2-aminobenzoate	C8H9NO2	151.063328534	50

Building a New Standardizer

Standardization rules are imperfect

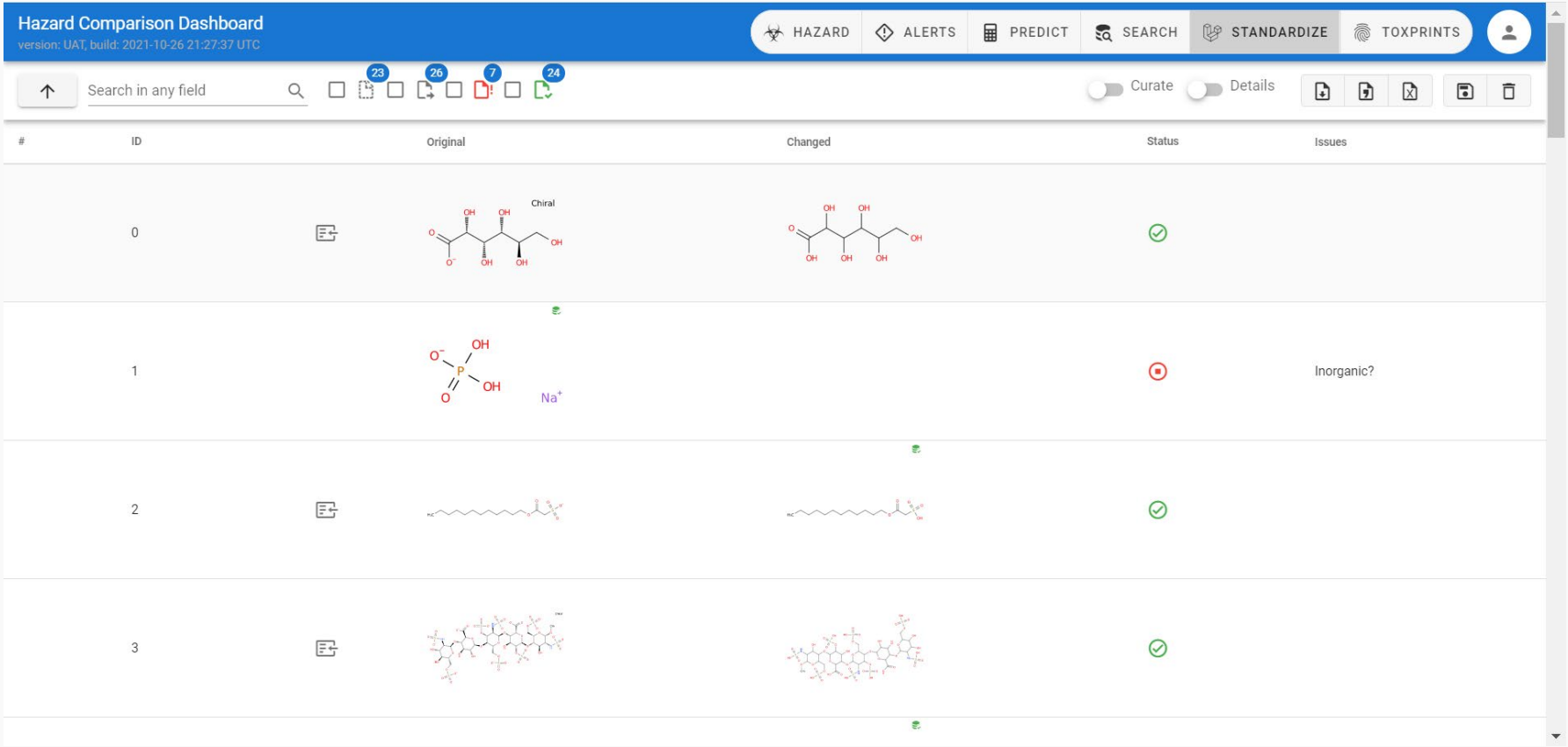
- We have found multiple examples of where standardization rules need to be optimized and tweaked – not just for MS but also for chemical registration
  - Tautomer handling
  - Salt handling – especially C-Metal bonds
  - Multi-component systems – removal of solvents

Proof-of-concept standardizer

- The EPA Cheminformatics Modules are proof-of-concept tools to test approaches
- Five modules at present
  - Hazard comparison module
  - Structure/substructure/similarity searching
  - Batch QSAR prediction for property and toxicity
  - Application of “ToxPrint chemotypes”
  - Structure Alerts

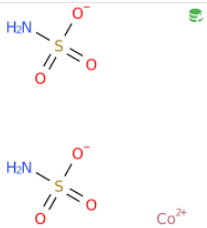
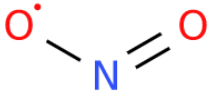
The Standardizer PoC Module

https://www.epa.gov/chemical-research/cheminformatics



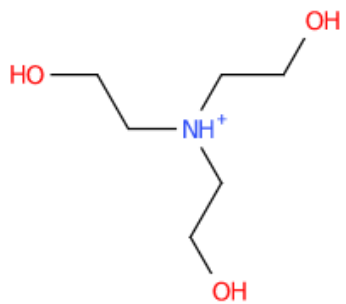
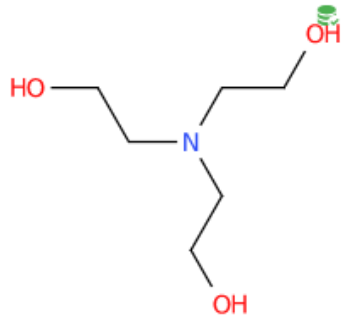
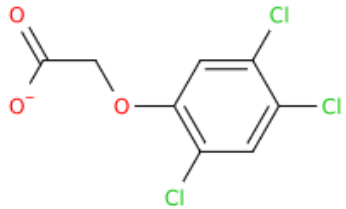
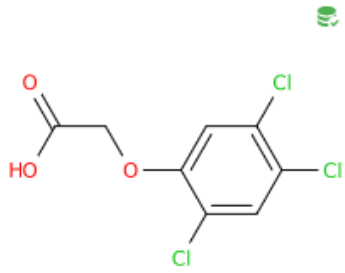
Standardizer

- Inorganic detection PLUS curation interface

Original	Changed	Status	Issues	Curation	Actions
NaH			Inorganic?		
			Inorganic?		
			Inorganic?		

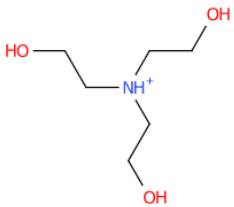
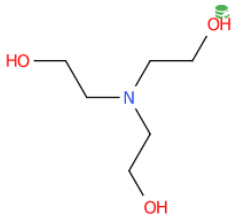
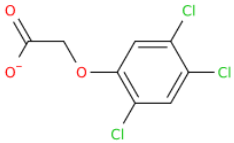
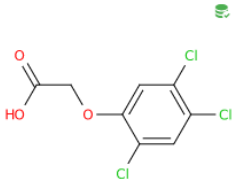
Standardizer – MS and QSAR-Ready

- Neutralization

Original	Changed	Status
		✓
		✓

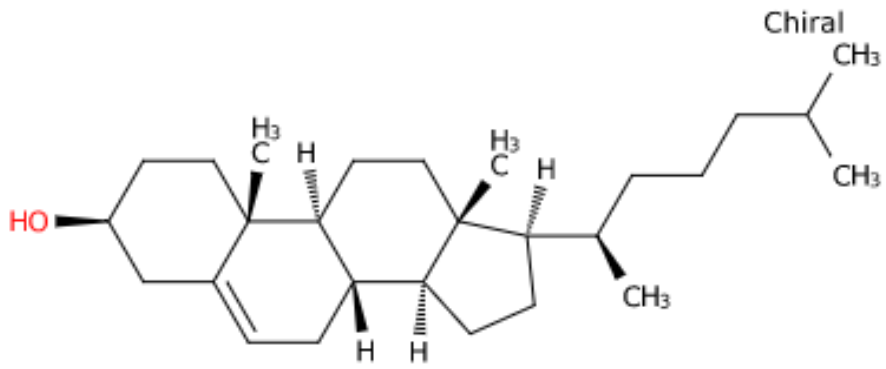
Standardizer – MS and QSAR-Ready

- Neutralization

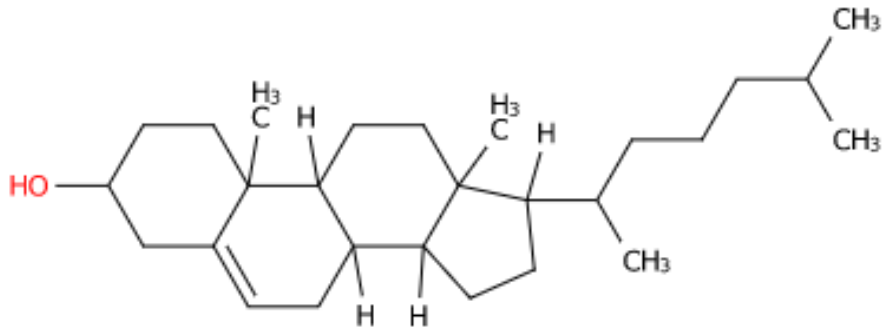
Original	Changed	Status
 InChI=1/C6H15NO3/c8-4-1-7(2-5-9)3-6-10/h8-10H,1-6H2/p+1/fC6H16NO3/h7H/q+1 GSEJCLTVZPLZKY-ADELPXGTNA-O [NH+](CCO)(CCO)CCO OCC[NH+](CCO)CCO	 InChI=1/C6H15NO3/c8-4-1-7(2-5-9)3-6-10/h8-10H,1-6H2 GSEJCLTVZPLZKY-UHFFFAOYNA-N N(CCO)(CCO)CCO OCCN(CCO)CCO	✓
 InChI=1/C8H5Cl3O3/c9-4-1-6(11)7(2-5(4)10)14-3-8(12)13/h1-2H,3H2,(H,12,13)/p-1/fC8H4Cl3O3/q-1 SMYMJHWAQXWPDB-RYONWTJXNA-M C1(OCC([O-])=O)=C(Cl)C=C(Cl)C(Cl)=C1 [O-]C(=O)COC1C=C(Cl)C(Cl)=CC=1Cl	 InChI=1/C8H5Cl3O3/c9-4-1-6(11)7(2-5(4)10)14-3-8(12)13/h1-2H,3H2,(H,12,13)/f/h12H SMYMJHWAQXWPDB-XWKXFZRBNA-N OC(=O)COC1C(Cl)=CC(Cl)=C(Cl)C=1 OC(=O)COC1=CC(Cl)=C(Cl)C=C1Cl	✓

Stereo Removal

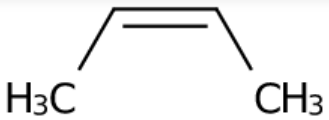
Original



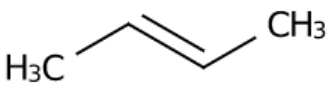
Reset Stereocenters



Stereo Removal

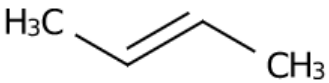


InChI=1/C4H8/c1-3-4-2/h3-4H,1-2H3/b4-3-
IAQRGUVFOMOMEM-ARJAWSKDNA-N
C/C=C\C
C/C=C\C

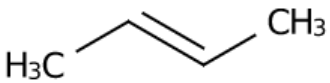


InChI=1/C4H8/c1-3-4-2/h3-4H,1-2H3
IAQRGUVFOMOMEM-UHFFFAOYNA-N
CC=CC
CC=CC

- Cross-bond depiction required improvement



InChI=1/C4H8/c1-3-4-2/h3-4H,1-2H3/b4-3+
IAQRGUVFOMOMEM-ONEGZZNKNA-N
C/C=C/C
C/C=C/C

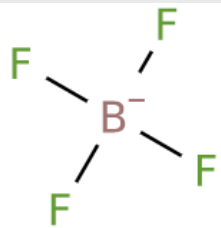
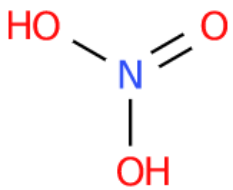
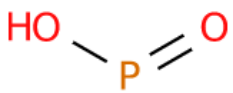


InChI=1/C4H8/c1-3-4-2/h3-4H,1-2H3
IAQRGUVFOMOMEM-UHFFFAOYNA-N
CC=CC
CC=CC

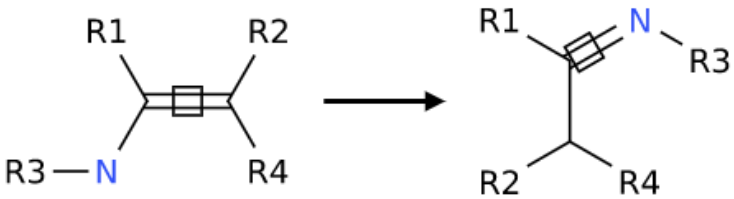
Rules can be defined

Type	Name	Frozen	Title	Description	Records #
LibraryGroup	mesomers		Mesomeric transformations		13
LibraryGroup	tautomers		Tautomers standardization		7
LibraryGroup	neutralize		Neutralize and de-radicalize		59
LibraryGroup	break-salts		Break salts		11
LibraryGroup	ms-ready-exclusions		MS-Ready exclusions		32
LibraryGroup	qsar-ready-exclusions		QSAR-Ready exclusions		189
LibraryGroup	standard-checks		Standard checks		11
Workflow	ms-ready		MS-Ready		13
Workflow	qsar-ready		QSAR-Ready		14
LibraryGroup	markush-checks		Markush checks		7

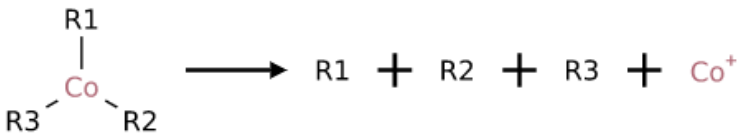
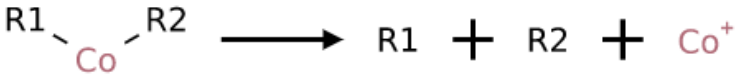
MS-Ready Exclusions

Check	SMILES		WARNING	SKIP	Tetrafluoroboranuide
Check	SMILES		WARNING	SKIP	Bromide
Check	SMILES		WARNING	SKIP	Nitrate Uncharged
Check	SMILES		WARNING	SKIP	Hypophosphite Uncharged

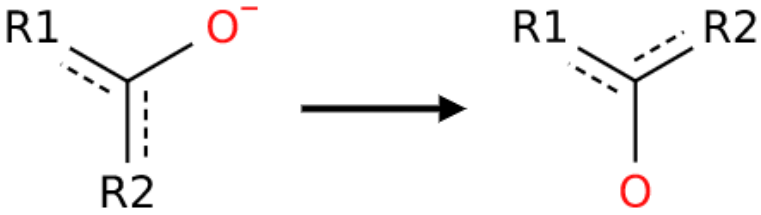
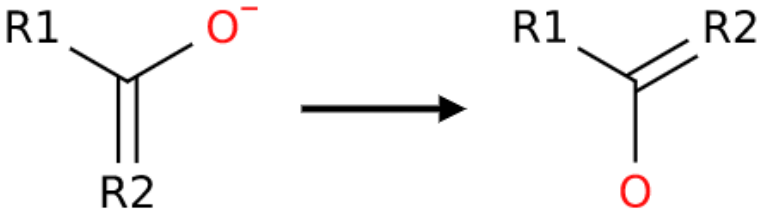
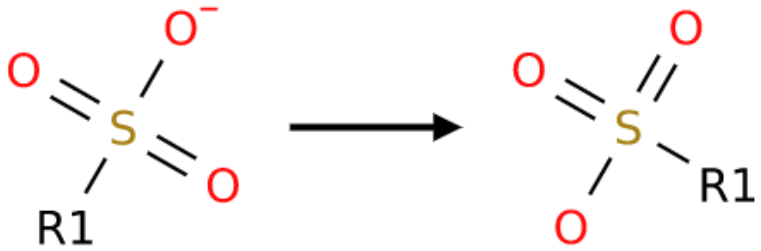
Tautomers

Change	SMIRKS	
Change	SMIRKS	
Change	SMIRKS	

Break Salts

Change	SMIRKS	A—K ⟶ A + K^+
Change	SMIRKS	
Change	SMIRKS	

Neutralize and Deradicalize

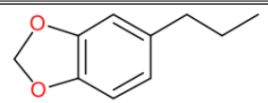
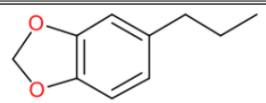
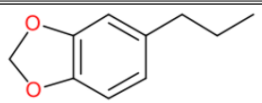
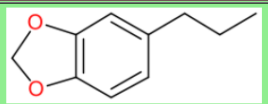
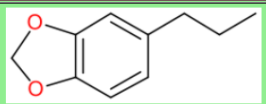
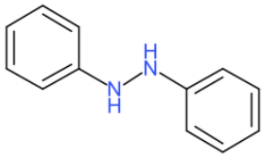
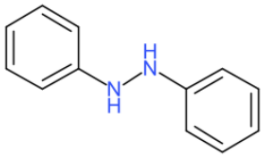
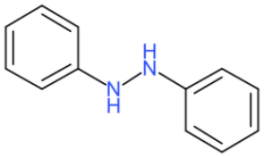
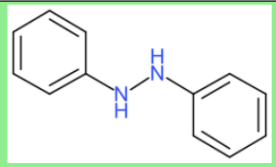
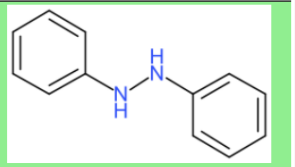
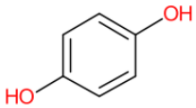
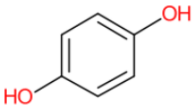
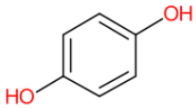
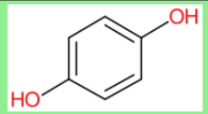
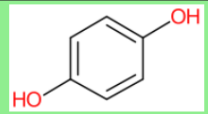
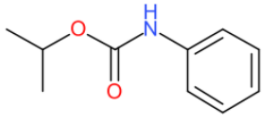
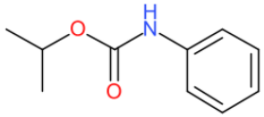
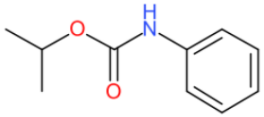
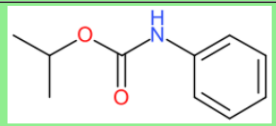
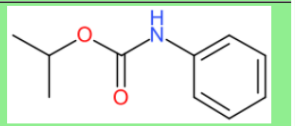
Change	SMIRKS	
Change	SMIRKS	
Change	SMIRKS	

Markush structure handling

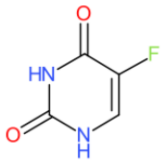
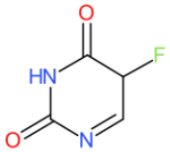
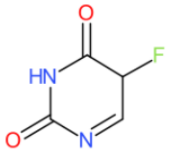
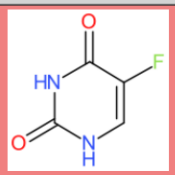
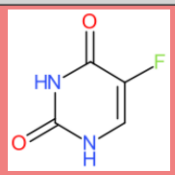
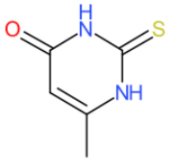
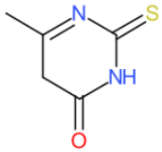
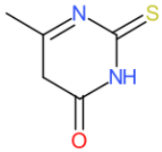
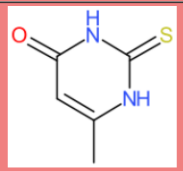
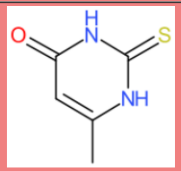
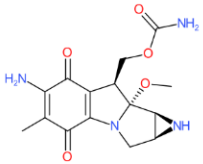
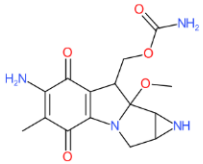
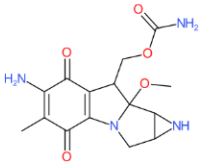
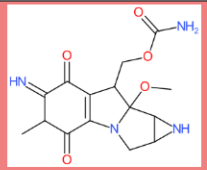
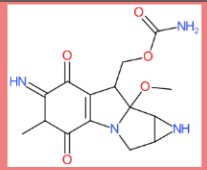
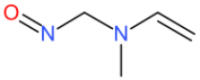
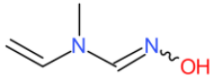
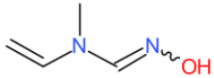
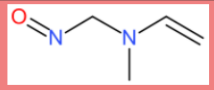
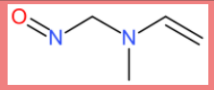
Class	Type	Value	Severity	Action	Title
Check	METHOD	hasQuery	ERROR	SKIP	Has Query?
Check	METHOD	hasPseudoatom	ERROR	SKIP	Pseudo-atom?
Check	METHOD	hasGenericSGroups	ERROR	SKIP	Generic SGroups?
Check	METHOD	hasSuperatom	ERROR	SKIP	Super-atom?
Check	METHOD	hasRepeatingUnits	ERROR	SKIP	Repeating Units?
Check	METHOD	hasMultipleGroups	ERROR	SKIP	Multiple Groups?
Check	METHOD	hasRGroups	ERROR	SKIP	R-Groups?

Work in Progress

- Validation study comparing OPERA with new standardizer for both QSAR and MS-Ready

DTXSID	ORIGINAL	DSSTOX_QSAR_READY	DSSTOX_MS_READY	SCI_DATA_EXPERTS_QSAR_READY	SCI_DATA_EXPERTS_MS_READY
DTXSID7020475	 CCCC1=CC2=C(OCO2)C=C1	 CCCC1=CC2=C(OCO2)C=C1	 CCCC1=CC2=C(OCO2)C=C1	 CCCC1=CC2OCOC=2C=C1	 CCCC1=CC2OCOC=2C=C1
DTXSID7020710	 N(NC1=CC=CC=C1)C1=CC=CC=C1	 N(NC1=CC=CC=C1)C1=CC=CC=C1	 N(NC1=CC=CC=C1)C1=CC=CC=C1	 C1=CC=CC=C1NNC1C=CC=CC=1	 C1=CC=CC=C1NNC1C=CC=CC=1
DTXSID7020716	 OC1=CC=C(O)C=C1	 OC1=CC=C(O)C=C1	 OC1=CC=C(O)C=C1	 OC1C=CC(O)=CC=1	 OC1C=CC(O)=CC=1
DTXSID7020766	 CC(C)OC(=O)NC1=CC=CC=C1	 CC(C)OC(=O)NC1=CC=CC=C1	 CC(C)OC(=O)NC1=CC=CC=C1	 CC(C)OC(=O)NC1C=CC=CC=1	 CC(C)OC(=O)NC1C=CC=CC=1

DIFFERENCES are of interest

DTXSID	ORIGINAL	DSSTOX_QSAR_READY	DSSTOX_MS_READY	SCI_DATA_EXPERTS_QSAR_READY	SCI_DATA_EXPERTS_MS_READY
DTXSID2020634	 FC1=CNC(=O)NC1=O	 FC1C=NC(=O)NC1=O	 FC1C=NC(=O)NC1=O	 O=C1NC=C(F)C(=O)N1	 O=C1NC=C(F)C(=O)N1
DTXSID2020890	 CC1=CC(=O)NC(=S)N1	 CC1=NC(=S)NC(=O)C1	 CC1=NC(=S)NC(=O)C1	 CC1=CC(=O)NC(=S)N1	 CC1=CC(=O)NC(=S)N1
DTXSID2020898	 CO[C@]12[C@H]3N[C@H]3CN1C1=C([C@H]2COC(N)=O)C(=O)C(N)=C(C)C1=O	 COC12C3NC3CN1C1=C(C2COC(N)=O)C(=O)C(N)=C(C)C1=O	 COC12C3NC3CN1C1=C(C2COC(N)=O)C(=O)C(N)=C(C)C1=O	 CC1C(=N)C(=O)C2C(COC(N)=O)C3(OC)C4NC4CN3C=2C1=O	 CC1C(=N)C(=O)C2C(COC(N)=O)C3(OC)C4NC4CN3C=2C1=O
DTXSID3042211	 CN(CN=O)C=C	 CN(C=C)C=NO	 CN(C=C)C=NO	 CN(CN=O)C=C	 CN(CN=O)C=C

Develop standard Tautomer Sets

Tautomer Standardization in Chemical Databases: Deriving Business Rules from Quantum Chemistry

Christopher M. Baker*, Nathan J. Kidley, Konstantinos Papachristos, Matthew Hotson, Rob Carson, David Gravestock, Martin Pouliot, Jim Harrison, and Alan Dowling

Cite this: *J. Chem. Inf. Model.* 2020, 60, 8, 3781–3791
Publication Date: July 9, 2020
https://doi.org/10.1021/acs.jcim.0c00232

Article Views	Altmetric	Citations
924	7	2

LEARN ABOUT THESE METRICS

Share Add to Export RIS

- Tautomeric standardization will benefit our mappings within the registration system
- Specifically useful for our project to merge and assemble public domain data

MS-Ready Data feed our *in silico* predictions

EPA United States Environmental Protection Agency

scientific **data**

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > scientific data > data descriptors > article

Data Descriptor | Open Access | Published: 02 August 2019

Linking *in silico* MS/MS spectra with chemistry data to improve identification of unknowns

Andrew D. McEachran, Ilya Balabin, Tommy Cathey, Thomas R. Transue, Hussein Al-Ghoul, Chris Grulke, Jon R. Sobus & Antony J. Williams

Scientific Data **6**, Article number: 141 (2019) | Cite this article

An integration hub for MS data

- In order to support our non-targeted analysis we are presently building:
  - The NTA WebApp for data processing and preparation before using our cheminformatics platforms
  - Processing 1.2M substances into MS-Ready structures to generate CFM-ID calculations
  - Assembling public domain data and mapping to our database – MassBank.eu and MassBank.us, SpectraBase
  - Building a cheminformatically enabled methods database

Building a Methods Database

- Simple Vision: I want to find the best method(s) associated with a chemical (class)
- The Approach:
  - Aggregate MS method documents
  - Extract chemistry (mostly CASRN and Names)
  - Map CASRN and Names to structures
  - Search a database by names and synonyms, CASRNs, InChIKeys and ultimately structure
  - “I cannot find my chemical in any method” – CHEMINFORMATICS can help….

Where are there methods?

- 900 method documents

Related Topics: Pesticide Analytical Methods

CONTACT US

Environmental Chemistry Methods (ECM) Index – 0–9

0-9 | A | B | C | D | E | F | G | H | I | K | L | M | N | O | P | Q | R | S | T | V | Z

Analyte(s) by Pesticide	ECM MRID	Matrix	Method Date
1,2,4-triazole	49762553	Water	2/19/13
1,3-dichloropropene & 1,2-dichloropropane	44536511	Soil	3/27/98
1,3-dichloropropene & 1,2-dichloropropane	44536511	Water	3/27/98
1,3-dichloropropene Degradate 3-chloroallyl Alcohol	44536505	Water	12/12/97

Other Methods…

Approved CWA Test Methods: Organic Compounds

Methods approved under Clean Water Act section 304(h) and published at 40 CFR Part 136 EXIT.

August 2017:

- Method 608.3 *(replaces Method 608)*
- Method 624.1 *(replaces Method 624)*
- Method 625.1 *(replaces Method 625)*
- Validation of SPE Products and Associated Procedures with Method 625.1

Note: The drinking water method 525.1 is approved for Clean Water Act use, but it is not approved for Safe Drinking Water Act (SDWA) compliance monitoring.

- More information on approved SDWA methods

- 420.1: Phenolics (Spectrophotometric, Manual 4AAP With Distillation) (pdf) (89.78 KB, 1978)
- 420.4: Determination of Total Recoverable Phenolics by Semi-Automated Colorimetry; Rev. 1.0 (pdf) (191.55 KB, August 1993)
- 525.1: Determination of Organic Compounds in Drinking Water by Liquid-Solid Extraction and Capillary Column Gas Chromatography/Mass Spectrometry; Rev 2.2 (pdf) (770.69 KB, May 1991)
- 525.2: Determination of Organic Compounds in Drinking Water by Liquid-Solid Extraction

Regulatory Methods

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency

Office of Water

www.epa.gov May 1991

Method 525.1, Revision 2.2: Determination of Organic Compounds in Drinking Water by Liquid-Solid Extraction and Capillary Column Gas Chromatography/Mass Spectrometry

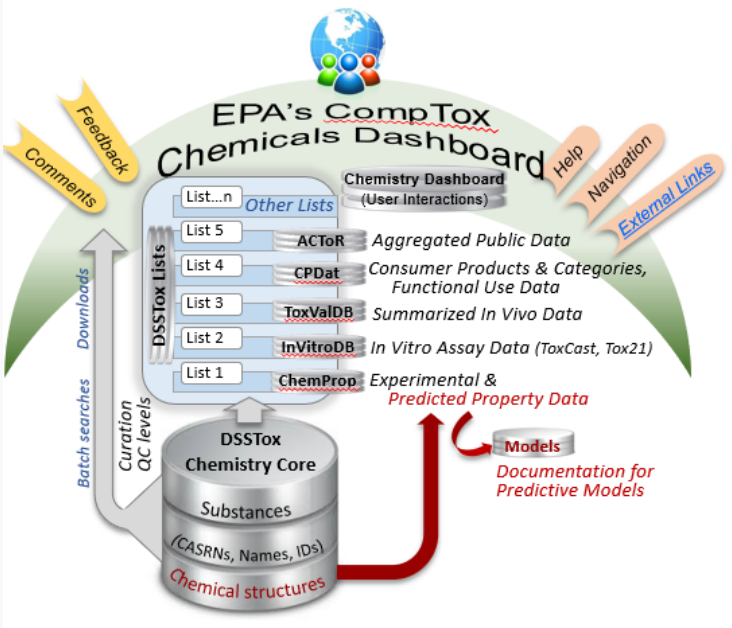
Compound	*MW*[a]	*CAS No.*
Acenaphthylene	152	208-96-8
Alachlor	269	15972-60-8
Aldrin	362	309-00-2
Anthracene	178	120-12-7
Atrazine	215	1912-24-9
Benz[*a*]anthracene	228	56-55-3
Benzo[*b*]fluoranthene	252	205-99-2
Benzo[*k*]fluoranthene	252	207-08-9
Benzo[*a*]pyrene	252	50-32-8
Benzo[*g,h,i*]perylene	276	191-24-2
Butylbenzyl phthalate	312	85-68-7
Chlordane Components		
α-Chlordane	406	5103-71-9
γ-Chlordane	406	5103-74-2
trans-Nonachlor	440	39765-80-5
2-Chlorobiphenyl	188	2051-60-7
Chrysene	228	218-01-9
Dibenz[*a,h*]anthracene	278	53-70-3
Di-*n*-butyl phthalate	278	84-72-2
2,3-dichlorobiphenyl	222	16605-91-7
Diethyl phthalate	222	84-66-2
Bis(2-ethylhexyl) adipate	222	103-23-1
Bis(2-ethylhexy) phthalate	390	117-81-7
Dimethyl phthalate	194	131-11-3
Endrin	378	72-20-8
Fluorene	166	86-73-7
Heptachlor	370	76-44-8
Heptachlor epoxide	386	1024-57-3
2,2′3,3′,4,4′,6-Heptachlorobiphenyl	392	52663-71-5

Structure Standardization is KEY

- To pull data together across public MS spectral databases, and build an integrated MS methods database, requires structure standardization and lots of CURATION
- We will unveil the results of this effort at the ACS Spring meeting in 2023
- Watch this space

Conclusion

- Dashboard access to data for ~906k chemicals with MS-Ready data facilitating structure identification
- Related metadata facilitates candidate ranking
- We need to continue to tweak and modify MS-Ready through new rules
- PoC standardization module allows us to have multiple rule sets and adjust to research performance



- MS-ready rules and generation service will all be publicly available for people to re-use

Acknowledgements

- Kamel Mansouri (he started all of this! We are indebted)
- Chris Grulke (6 years of developing the chemical registration schema)

Contact

Antony Williams
CCTE, US EPA Office of Research and Development,
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821

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

Journal of Cheminformatics

DATABASE

Open Access

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]

https://doi.org/10.1186/s13321-017-0247-6