

Consensus ranking and fragmentation prediction for identification of unknowns in high resolution mass spectrometry

*Antony Williams¹, Andrew D. McEachran², Tommy Cathey³,
Tom Transue³ and Jon R. Sobus⁴*

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC

2) Oak Ridge Institute of Science and Education (ORISE) Research Participant, RTP, NC

3) GDIT, Research Triangle Park, North Carolina, United State

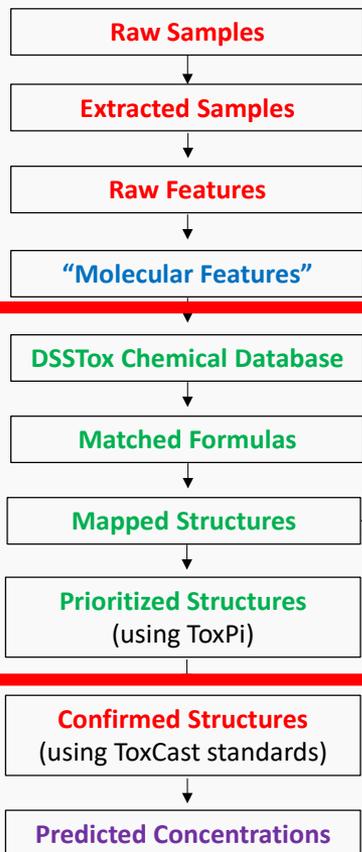
4) National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC

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

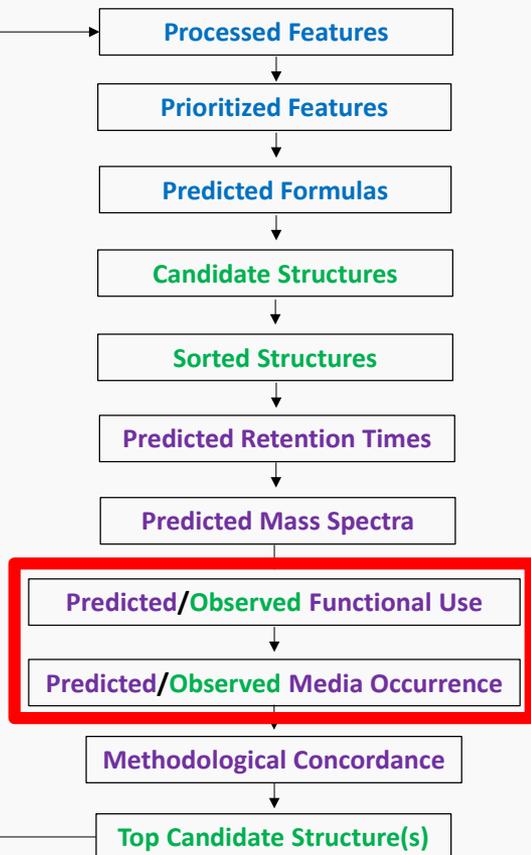
*Spring 2019
ACS Spring Meeting, Orlando*

Suspect Screening and Non-Targeted Analysis Workflows

Suspect Screening



Non-Targeted Analysis



Color Key

Red = Analytical Chemistry

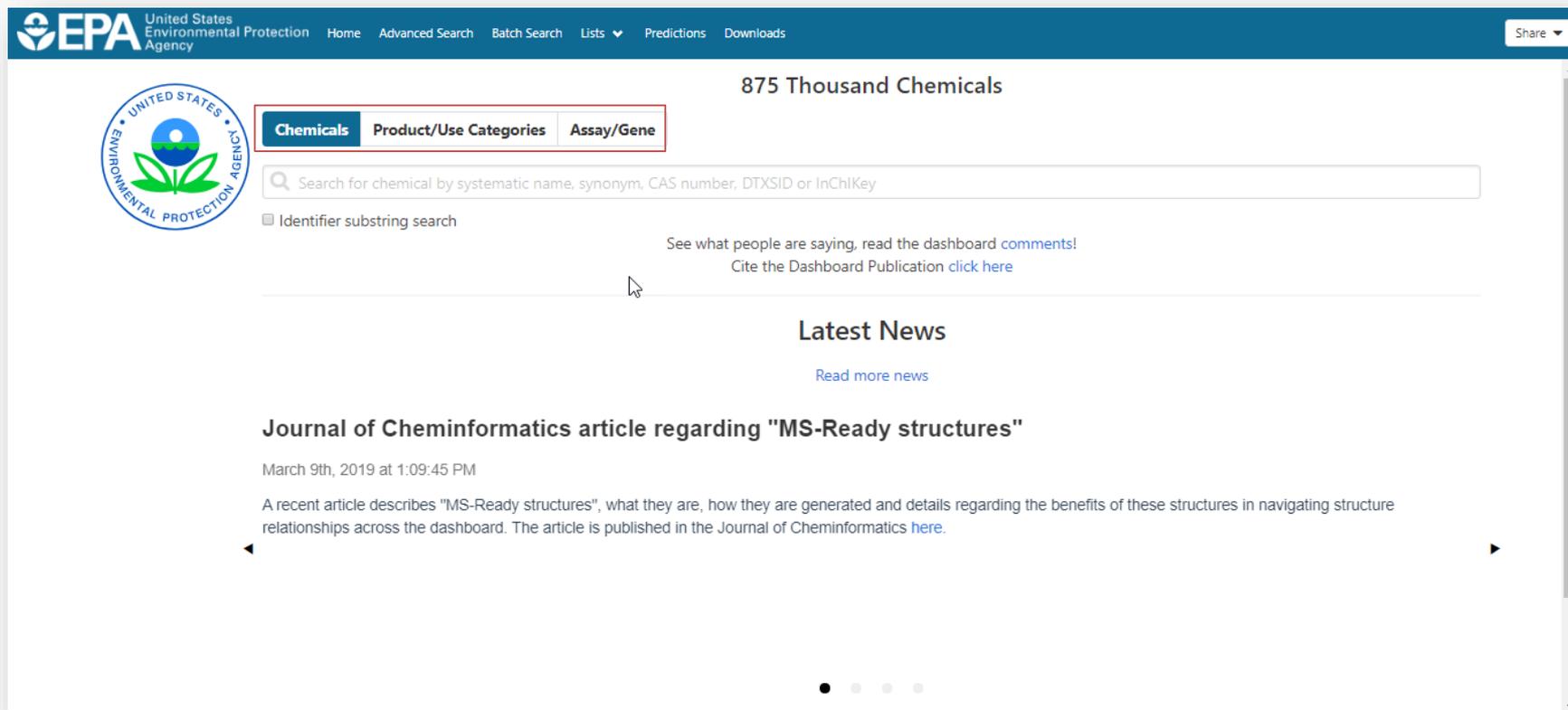
Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

Green = Informatics & Web Services



875k Chemical Substances



The screenshot displays the EPA CompTox Chemicals Dashboard. At the top, the EPA logo and navigation links (Home, Advanced Search, Batch Search, Lists, Predictions, Downloads) are visible. The main heading reads "875 Thousand Chemicals". Below this, there are three tabs: "Chemicals" (selected), "Product/Use Categories", and "Assay/Gene". A search bar is present with the placeholder text "Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey". Below the search bar, there is a checkbox for "Identifier substring search" and links to "See what people are saying, read the dashboard comments!" and "Cite the Dashboard Publication [click here](#)". The "Latest News" section features an article titled "Journal of Cheminformatics article regarding 'MS-Ready structures'" dated March 9th, 2019. The article text states: "A recent article describes 'MS-Ready structures', what they are, how they are generated and details regarding the benefits of these structures in navigating structure relationships across the dashboard. The article is published in the Journal of Cheminformatics [here](#)."

Detailed Chemical Pages

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
 - ▶ ADME
 - ▶ EXPOSURE
 - ▶ BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- ▶ LITERATURE
- LINKS
- COMMENTS

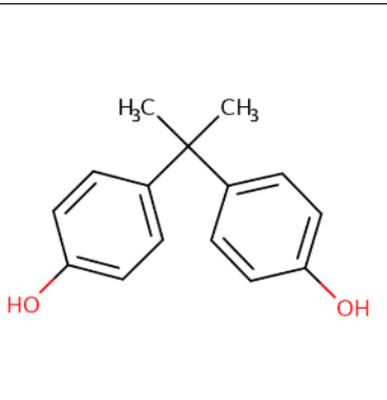
Home Advanced Search Batch Search Lists Predictions Downloads

Copy Share Submit Comment Search all data

Bisphenol A

80-05-7 | DTXSID7020182

Searched by DSSTox Substance Id.



Cc1ccc(O)cc1C(C)c2ccc(O)cc2

Wikipedia

Bisphenol A (BPA) is an organic synthetic compound with the chemical formula $(\text{CH}_3)_2\text{C}(\text{C}_6\text{H}_4\text{OH})_2$ belonging to the group of diphenylmethane derivatives and bisphenols, with two hydroxyphenyl groups. It is a colorless solid that is soluble in organic solvents, but poorly soluble in water (0.344 wt % at 83 °C). BPA is a starting material for the synthesis of plastics, primarily certain polycarbonates

[Read more](#)

Intrinsic Properties

Molecular Formula: $\text{C}_{15}\text{H}_{16}\text{O}_2$ [Mol File](#) [Find All Chemicals](#)

Average Mass: 228.291 g/mol [Isotope Mass Distribution](#)

Monoisotopic Mass: 228.11503 g/mol

Structural Identifiers

Linked Substances

Presence in Lists

Record Information

Quality Control Notes

Access to Chemical Hazard Data

DETAILS

EXECUTIVE SUMMARY

PROPERTIES

ENV. FATE/TRANSPORT

HAZARD

▶ ADME

▶ EXPOSURE

▶ BIOACTIVITY

SIMILAR COMPOUNDS

GENRA (BETA)

RELATED SUBSTANCES

SYNONYMS

▶ LITERATURE

LINKS

COMMENTS

Hazard

Data Type: Toxicity Value

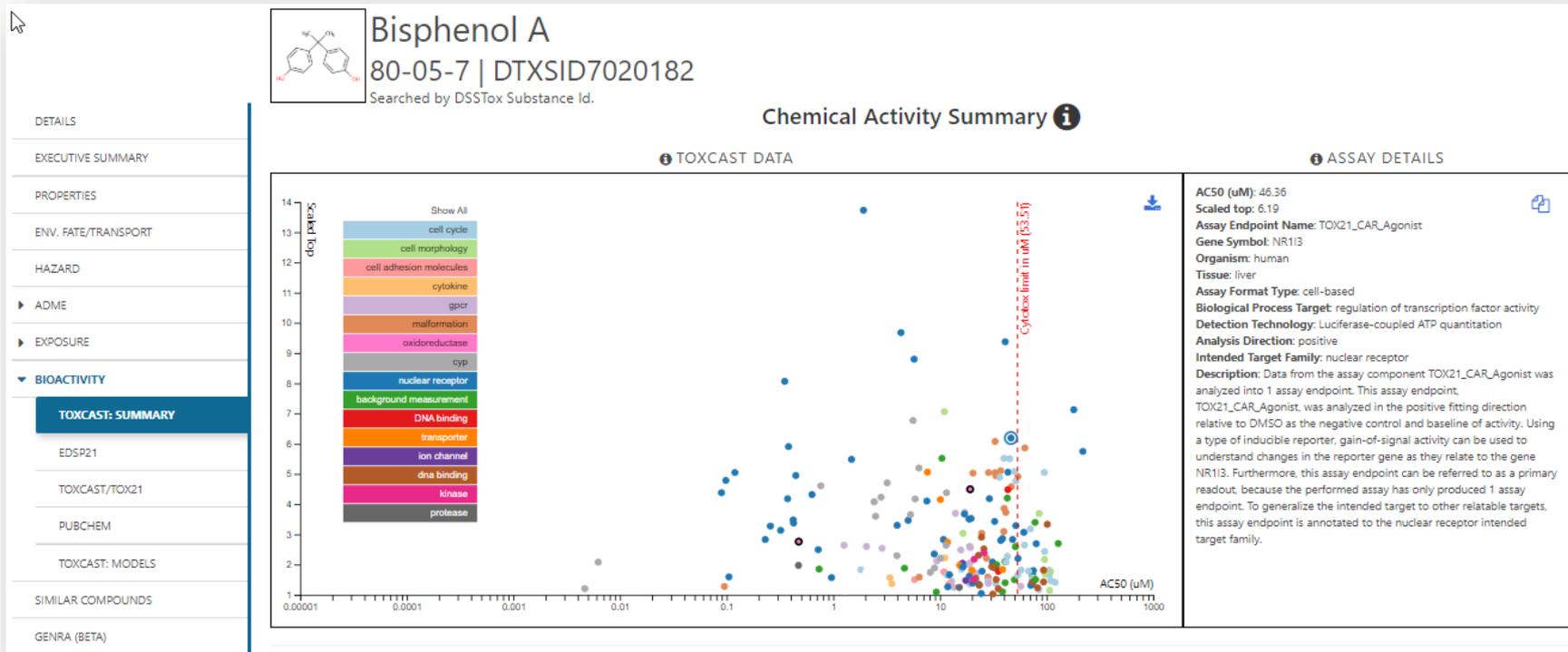
Human
Eco

Download
Columns
Search query

More	Priority	Type	Subtype	Risk assessment class	Value	Units	Study type	Exposure route	Species	Subsource	Source
	7	MEG	Short-term Critical Air	short-term	500	mg/m3	-	inhalation	-	TG 230 Military Exposure Guidelines Table	DOD
	7	MEG	Short-term Marginal Air	short-term	100	mg/m3	-	inhalation	-	TG 230 Military Exposure Guidelines Table	DOD
	7	MEG	Short-term Negligible Air	short-term	15	mg/m3	-	inhalation	-	TG 230 Military Exposure Guidelines Table	DOD
	7	MEG	Soil Negligible Soil	chronic	106000	mg/kg	-	soil	-	TG 230 Military Exposure Guidelines Table	DOD
	7	MEG	Long-Term, SL/d Negligible Water	chronic	7	mg/L	-	oral	-	TG 230 Military Exposure Guidelines Table	DOD
	6	RfD	-	chronic	0.05	mg/kg-day	-	oral	rat	Wignall	Wignall
	5	RfD	-	chronic	0.05	mg/kg-day	-	-	-	MSC Table 5	Pennsylvania DEP ToxValues
	4	RfD	-	chronic	0.05	mg/kg-day	chronic	oral	rat	IRIS	Chiu
	3	RfD	-	chronic	0.6	mg/kg-day	-	oral	rat	EPA/ORNL/OLEM	HEAST
	1	RfD	-	chronic	0.05	mg/kg-day	-	oral	-	EPA NCEA	IRIS

In Vitro Bioassay Screening

ToxCast and Tox21



Sources of Exposure to Chemicals

Bisphenol A

80-05-7 | DTXSID7020182

Searched by DSSTox Substance Id.

DETAILS

EXECUTIVE SUMMARY

PROPERTIES

ENV. FATE/TRANSPORT

HAZARD

ADME

EXPOSURE

PRODUCT & USE CATEGORIES

CHEMICAL WEIGHT FRACTION

CHEMICAL FUNCTIONAL USE

TOXICS RELEASE INVENTORY

MONITORING DATA

EXPOSURE PREDICTIONS

PRODUCTION VOLUME

Download

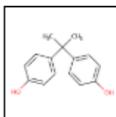
Columns 10

Search query

Product and Use Categories (PUCs) i

Product or Use Categorization	Categorization type	Number of Unique Products
manufacturing, metals	CPCat Cassette	17
adhesive	CPCat Cassette	17
	CPCat Cassette	16
	CPCat Cassette	12
	CPCat Cassette	11
	CPCat Cassette	8
	CPCat Cassette	8
	CPCat Cassette	8
	CPCat Cassette	7
	CPCat Cassette	6

First << < 1 2 3 4 5 6 7 8 9 10 > >> Last



Bisphenol A

80-05-7 | DTXSID7020182

Searched by Approved Name.

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ▶ ADME
- ▶ EXPOSURE
- ▶ BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- ▶ LITERATURE
- LINKS**
- COMMENTS

General

- EPA Substance Registry Service
- Household Products Database
- Chemical Entities of Biological Interest (ChEBI)
- PubChem
- Chempider
- CPCat
- DrugBank
- HMDB
- Wikipedia
- MSDS Lookup
- ChEMBL
- Chemical Vendors
- CalEPA Office of Environmental Health Hazard Assessment
- NIOSH Chemical Safety Cards
- ToxPlanet
- ACS Reagent Chemicals
- Wikidata
- ChemHat: Hazards and Alternatives Toolbox
- Wolfram Alpha
- ScrubChem
- ECHA Brief Profile

Toxicology

- ACToR
- DrugPortal
- CCRIS
- ChemView
- CTD
- eChemPortal
- Gene-Tox
- HSDB
- ToxCast Dashboard 2
- LactMed
- International Toxicity Estimates for Risk
- ATSDR Toxic Substances Portal
- Superfund Chemical Data matrix
- NIOSH IDLH Values
- ACToR PDF Report
- Toxics Release Inventory
- CREST
- National Air Toxics Assessment

Publications

- Toxline
- Environmental Health Perspectives
- NIEHS
- National Toxicology Program
- Google Books
- Google Scholar
- Google Patents
- PPRTVWEB
- PubMed
- IRIS Assessments
- EPA HERO
- NIOSH Skin Notation Profiles
- NIOSH Pocket Guide
- RSC Publications
- BioCaddie DataMed
- Springer Materials
- Federal Register
- Regulations.gov
- Bielefeld Academic Search Engine
- CORE Literature Search

Analytical

- FOR-IDENT
- NEMI: National Environmental Methods Index
- RSC Analytical Abstracts
- Tox21 Analytical Data
- MONA: MassBank North America
- mzCloud
- NIST IR Spectrum
- NIST MS Spectrum

Prediction

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

Analytical

- FOR-IDENT
- NEMI: National Environmental Methods Index
- RSC Analytical Abstracts
- Tox21 Analytical Data
- MONA: MassBank North America**
- mzCloud
- NIST NIST IR Spectrum
- NIST NIST MS Spectrum

MoNA - MassBank of North America | Spectra | Downloads | Upload | Help

Search...

Display Generated Query

10 records/page

Bisphenol A Score: ★★★★★

Cc1ccc(O)c1C(C)(C)c2ccc(O)cc2

instrument	LTQ Orbitrap XL Thermo Sc...
instrument type	LC-ESI-ITFT
ms level	MS2
ionization	ESI
collision energy	30 % (nominal)
retention time	14.0 min
precursor m/z	229.1223
precursor type	[M-H] ⁺
ionization mode	positive
accession	EA016309

Originally submitted to the MassBank High Quality Mass Spectral Database

MassBank | LC-MS | Display Full Record

Mass/Formula Searching and Metadata Ranking

Advanced Searches

Mass Search

Mass Search

\pm Min/Max

Adduct

Neutral 

All Adducts



Choose adduct from dropdown

191.131

Da

\pm

5

Da

ppm

Search 

Advanced Searches

Mass Search

Search Results

Searched by Mass: 191.131 +/- 5.0 ppm.

329 of 329 chemicals visible

Select all

Download

Send to Batch Search

Mass Difference

DTXSID

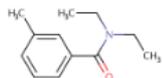
CASRN

TOXCAST

Mass Diff

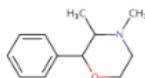
Multicomponent Chemicals

Filter by Name or CASRN



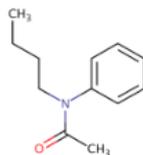
DEET

DTXSID: DTXSID2021995
CASRN: 134-62-3
TOXCAST: 12/768
Mass Diff: 0.000014



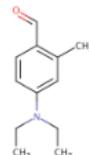
Phendimetrazine

DTXSID: DTXSID1023447
CASRN: 634-03-7
TOXCAST: -
Mass Diff: 0.000014



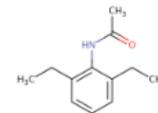
N-Butylacetanilide

DTXSID: DTXSID2042197
CASRN: 91-49-6
TOXCAST: -
Mass Diff: 0.000014



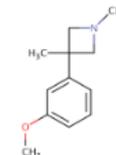
Benzaldehyde, 4-(diethylamino)-2-methyl-

DTXSID: DTXSID90168148
CASRN: 92-14-8
TOXCAST: -
Mass Diff: 0.000014



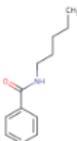
Acetanilide, 2',6'-diethyl-

DTXSID: DTXSID90168148
CASRN: 16665-89-7
TOXCAST: -
Mass Diff: 0.000014



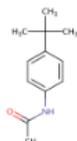
Azetidine, 1,3-dimethyl-3-(m-methoxyphenyl)-

DTXSID: DTXSID40173560
CASRN: 19832-26-9
TOXCAST: -
Mass Diff: 0.000014



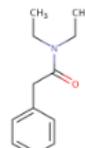
Benzamide, N-pentyl-

DTXSID: DTXSID20174196
CASRN: 20308-43-4
TOXCAST: -
Mass Diff: 0.000014



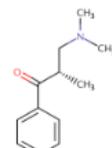
p-t-Butylacetanilide

DTXSID: DTXSID00174238
CASRN: 20330-45-4
TOXCAST: -
Mass Diff: 0.000014



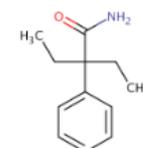
N,N-Diethylphenylacetamide

DTXSID: DTXSID00179048
CASRN: 2431-96-1
TOXCAST: -
Mass Diff: 0.000014



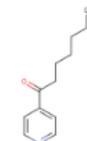
3-(Dimethylamino)-2-methylpropiofen...

DTXSID: DTXSID60180796
CASRN: 26171-50-6
TOXCAST: -
Mass Diff: 0.000014



Butyramide, 2-ethyl-2-phenyl-

DTXSID: DTXSID60184653
CASRN: 30568-39-9
TOXCAST: -
Mass Diff: 0.000014



1-Heptanone, 1-(4-pyridyl)-

DTXSID: DTXSID40186594
CASRN: 32941-30-3
TOXCAST: -
Mass Diff: 0.000014

MS-Ready Structures for Formula Search

Molecular Formula Search

MS Ready Formula  Exact Formula 

Formula

Please use the format of the following example: C₆H₈O₂ or C₆H(8-10)O(0-2)

Search 

“MS-Ready Structures”

<https://doi.org/10.1186/s13321-018-0299-2>

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

Journal of Cheminformatics

METHODOLOGY

Open Access

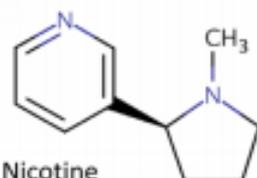
“MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies



Andrew D. McEachran^{1,2*}, Kamel Mansouri^{1,2,3}, Chris Grulke², Emma L. Schymanski⁴, Christoph Ruttkies⁵ and Antony J. Williams^{2*}

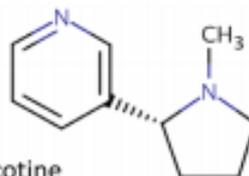
Open Science for Identifying “Known Unknown” Chemicals

Emma L. Schymanski*[†] and Antony J. Williams*[‡]



Nicotine

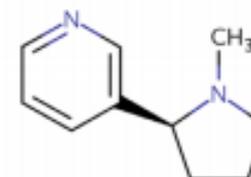
CN1CCC[C@H]1C1=CN=CC=C1
DTXSID1020930 | SNICXCGAKADSCV
54-11-5 | **162.1157** | 0.929 | **72**
Tox: **yes** | Expo: **yes** | Bioassay: **yes**



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**

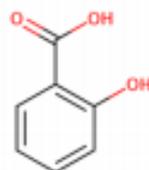
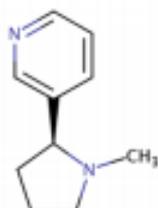
LEGEND: Name, SMILES
DTXSID | InChIKey 1st Block
CAS | Monoiso. Mass | logP | Sources
Data on: Toxicity | Exposure | Bioassays



HCl

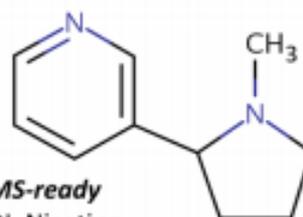
Nicotine hydrochloride

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



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

OC(=O)C1=CC(O)=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**



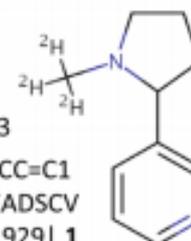
MS-ready

DL-Nicotine

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

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**



MS-Ready Mappings

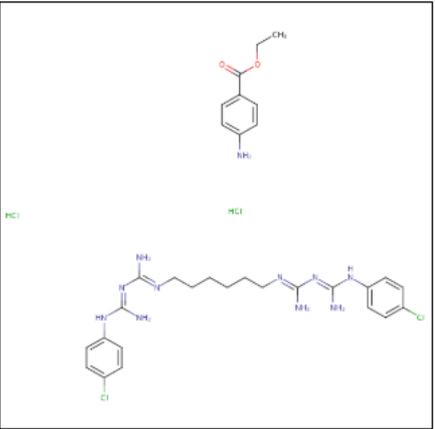
 United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads

Copy Share Submit Comment Search all data

Progaron

108532-15-6 | DTXSID20148579
Searched by DSSTox Substance Id.



DETAILS

- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS

Intrinsic Properties

Molecular Formula: $C_{31}H_{43}Cl_4N_1O_2$ Mol File Find All Chemicals

Average Mass: 743.56 g/mol Isotope Mass Distribution

Structural Identifiers

Linked Substances

Same Connectivity: 1 record (nChl)

Mixtures, Component: 0 records

MS-Ready Mappings: DTXCID301804: 12 records; DTXCID0013314: 11 records;

Similar Compounds: 0 records

Quality Control Notes

- EXACT Formula:** C₁₀H₁₆N₂O₈: 3 Hits

MS Ready Formula  Exact Formula 

Formula

C₁₀H₁₆N₂O₈

Select all

Download

Send to Batch Search

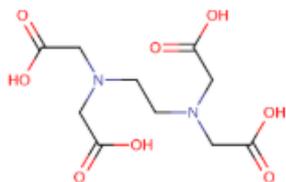
Default

DTXSID

PubChem

CPDAT

3 of 3 chemi

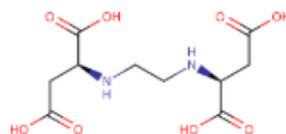


Ethylenediaminetetraacetic acid

DTXSID: DTXSID6022977

PubChem: 158

CPDAT: 387

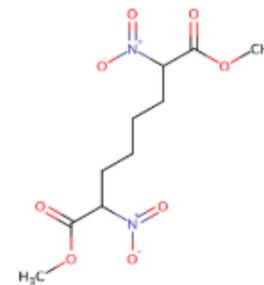


N,N'-Ethylenedi-L-aspartic acid

DTXSID: DTXSID1051852

PubChem: 25

CPDAT: 8



Dimethyl 2,7-dinitrooctanedioate

DTXSID: DTXSID20498864

PubChem: 5

CPDAT: 0

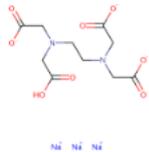
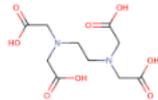
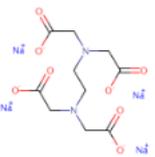
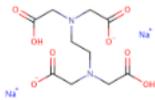
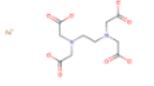
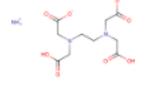
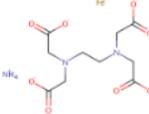
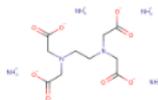
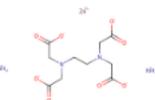
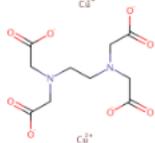
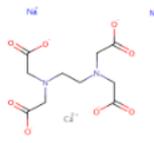
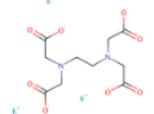
MS-Ready Mappings

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

125 chemicals

Select all Download Send to Batch Search Default DTXSID PubChem CPDAT

Hide chemicals that are: Filter by Name or CASRN

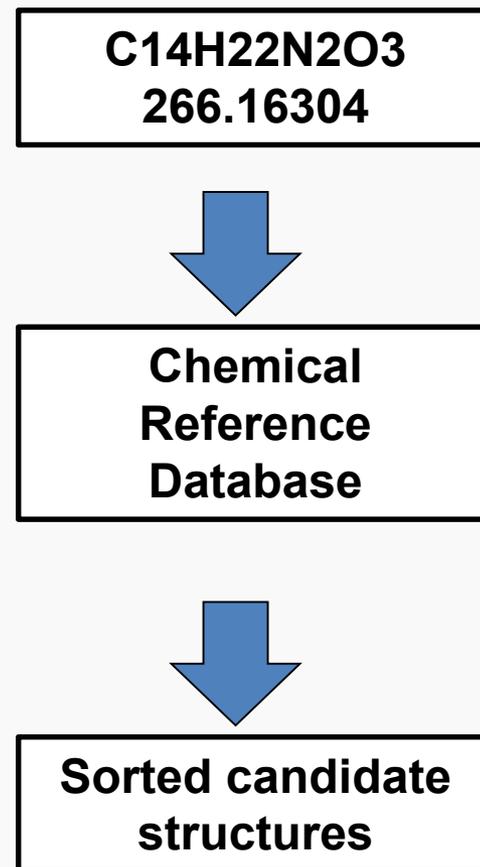
 <p>Trisodium ethylenediaminetetraacetate DTXSID: DTXSID7020556 PubChem: 33 CPDAT: 82</p>	 <p>Ethylenediaminetetraacetic acid DTXSID: DTXSID6022977 PubChem: 158 CPDAT: 387</p>	 <p>Ethylenediaminetetraacetic acid tetrasod... DTXSID: DTXSID3026350 PubChem: 57 CPDAT: 1227</p>	 <p>Ethylenediaminetetraacetic acid, disodiu... DTXSID: DTXSID9027073 PubChem: 56 CPDAT: 1359</p>	 <p>Ethylenediaminetetraacetic acid ferric so... DTXSID: DTXSID5027774 PubChem: 53 CPDAT: 62</p>	 <p>Diammonium dihydrogen ethylenediami... DTXSID: DTXSID9027813 PubChem: 12 CPDAT: 17</p>
 <p>Ferrate(1-), [[N,N'-1,2-ethanediy]bis[N-]](... DTXSID: DTXSID9027815 PubChem: 24 CPDAT: 20</p>	 <p>Tetraammonium ethylenediaminetetraac... DTXSID: DTXSID8027820 PubChem: 11 CPDAT: 12</p>	 <p>Zincate(2-), [[N,N'-1,2-ethanediy]bis[N-]](... DTXSID: DTXSID8028343 PubChem: 5 CPDAT: 10</p>	 <p>EDTA, copper salt DTXSID: DTXSID0034564 PubChem: 8 CPDAT: 10</p>	 <p>Calcium disodium ethylenediaminetetra... DTXSID: DTXSID2036409 PubChem: 42 CPDAT: 29</p>	 <p>Tetrapotassium ethylenediaminetetra... DTXSID: DTXSID3036442 PubChem: 25 CPDAT: 36</p>

- 125 chemicals returned in total
 - 8 of the 125 are **single component** chemicals
 - 3 of the 8 are **isotope-labeled**
 - 3 are **neutral compounds** and 2 are **charged**

Candidate ranking using public resources

Data Source Ranking of “known unknowns”

- Mass and/or formula is for an unknown chemical but contained within a reference database
- Most likely candidate chemicals have the most associated data sources, most associated lit. articles or both



Is a bigger database better?



© American Society for Mass Spectrometry, 2011

J. Am. Soc. Mass Spectrom. (2012) 23:179–185
DOI: 10.1007/s13361-011-0265-y

RESEARCH ARTICLE

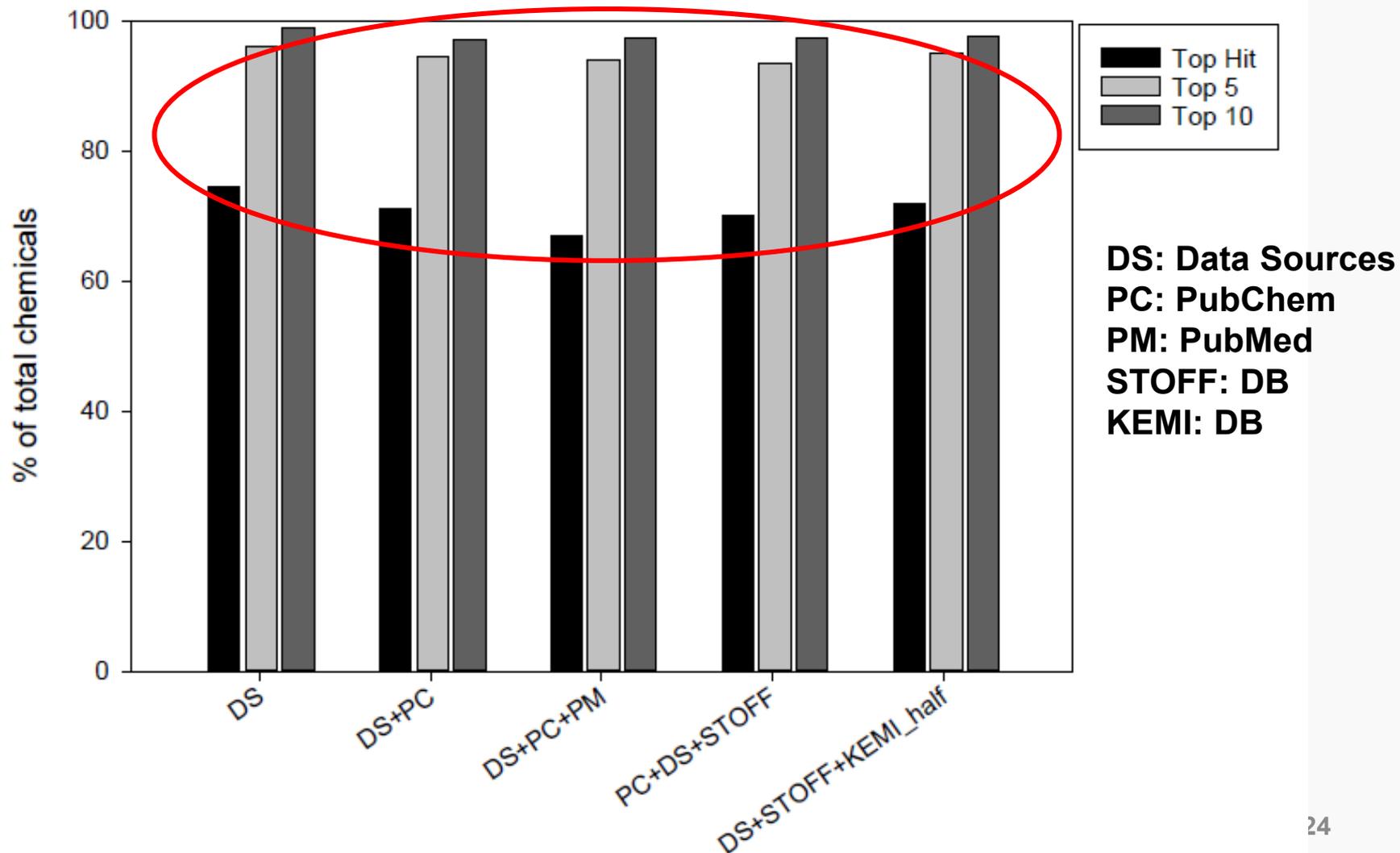
Identification of “Known Unknowns” Utilizing Accurate Mass Data and ChemSpider

- ChemSpider was 26 million chemicals then
- Much BIGGER today
- Is bigger better??

71 Million
chemical structures

- Use available metadata to rank candidates
 - Associated data sources
 - Associated lists in the underlying database
 - Associated data sources in PubChem
 - Specific types (e.g. water, surfactants, pesticides etc.)
 - Number of associated literature articles (Pubmed)
 - **Chemicals in the environment** – the number of products/categories containing the chemical is a very important source of data

Identification ranks for 1783 chemicals using multiple data streams



^{vs}
Anal Bioanal Chem (2017) 49:1729–1735
DOI 10.1007/s00216-016-0139-z



RAPID COMMUNICATION

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

Andrew D. McEachran¹ · Jon R. Sobus² · Antony J. Williams³

- Dashboard content *was* 720k chemicals
- Only 3% of ChemSpider size
- What was the comparison in performance?

SAME dataset for comparison

Compound class	Number in class	Average rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical drug	72	1.4	55	9	6	2	
Industrial chemicals	42	5.5	28	6	3		5
Personal care products	8	6.1	3	1			4
Steroid hormones	7	1.0	7				
Perfluorochemicals	6	1.2	5	1			
Pesticides	12	2.3	6	2	3		1
Veterinary drugs	3	1.3	2	1			
Dyes	2	1.0	2				
Food product/natural compounds	4	3.8	2			1	1
Illicit drugs	2	2.0	1		1		
Misc. molecules	3 ^a	1.3	2	1			

How did performance compare?

	Mass-based searching		Formula-based searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average rank position	1.3	2.2 ^a	1.2	1.4
Percent in #1 position	85%	70%	88%	80%

^a Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5

**For the same 162 chemicals,
Dashboard outperforms
ChemSpider**

How did performance compare?

Summary statistics and rank-ordered position in the CompTox Chemistry Dashboard and ChemSpider of the 89 compound subset from the Little et al. [7] study

		Average rank	Number in each position rank-ordered				
		(±SD)	#1	#2	#3	#4	#5+
Mass-based	Dashboard	1.2 ± 0.7	77 ^a	5	3	3	
	ChemSpider	2.2 ± 6.1 ^b	68	8	7	1	5
Formula-based	Dashboard	1.1 ± 0.4	78 ^a	8	2		
	ChemSpider	1.3 ± 1.0	77	8	2	1	2

^aOne chemical (tephrosin) not present in the Dashboard

Data Quality is important

- Data quality in free web-based databases!



The screenshot shows two overlapping journal article pages from Elsevier's 'Drug Discovery Today'. The top page is for Volume 17, Issues 13-14, July 2012, pages 685-701. The bottom page is for Volume 16, Issues 17-18, September 2011, pages 747-750. The bottom page features an editorial by Alex M Clark, Antony J Williams, and Sean Ekins, titled 'Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data'. The article is published in the Journal of Cheminformatics, 2015, 7:9. The URL is <https://doi.org/10.1186/s13321-015-0057-7>. The article was received on 24 November 2014, accepted on 23 February 2015, and published on 22 March 2015.

Drug Discovery Today
Volume 17, Issues 13–14, July 2012, Pages 685-701

Drug Discovery Today
Volume 16, Issues 17–18, September 2011, Pages 747-750

Editorial

Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data

Alex M Clark, Antony J Williams and Sean Ekins

Journal of Cheminformatics 2015 7:9
<https://doi.org/10.1186/s13321-015-0057-7> | © Clark et al.; licensee Springer. 2015

Received: 24 November 2014 | Accepted: 23 February 2015 | Published: 22 March 2015

Will the correct Microcystin LR Stand Up? ChemSpider Skeleton Search

Matches any text strings used to describe a molecule.

ZYZCGGRZINLQBL



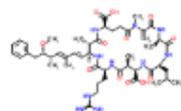
Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID ?

FILTER ▾

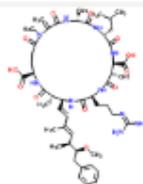
Search Hits Limit: 100 ▾

Found 9 results

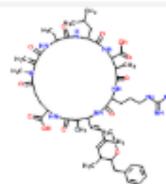
Search term: ZYZCGGRZINLQBL (Found by InChIKey (skeleton match))



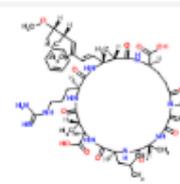
[Cyanoginosin](#)



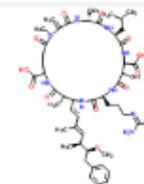
[MCYST-LR](#)



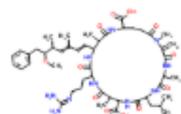
[15-\(3-Carbamimidamidopropyl\)-8-isobutyl-18-\(\(1E,3E\)-6-methoxy-3,5-dimethyl-7-phenyl-1,3-heptadien-1-](#)



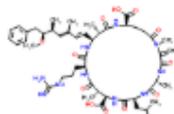
[\(5R,8S,11R,12S,15S,18R,19S,22R\)-15-\(3-Carbamimidamidopropyl\)-8-isobutyl-18-\(\(1E,3E,5S,6S\)-6-methoxy-3,5-](#)



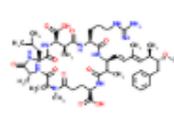
[\(5R,8S,15S\)-15-\(3-\(\(Diaminomethylene\)amino\)propyl\)-8-isobutyl-18-\(\(1E,3E,5S,6S\)-6-methoxy-3,5-](#)



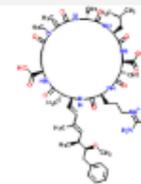
[15-\(3-\(\(Diaminomethylene\)amino\)propyl\)-8-isobutyl-18-\(6-methoxy-3,5-dimethyl-7-](#)



[\(5R,8S,11R,12S,15S,18S,19S,22R\)-15-\(3-Carbamimidamidopropyl\)-8-isobutyl-18-\(\(1E,5S,6S\)-6-methoxy-3,5-dimethyl-7-](#)

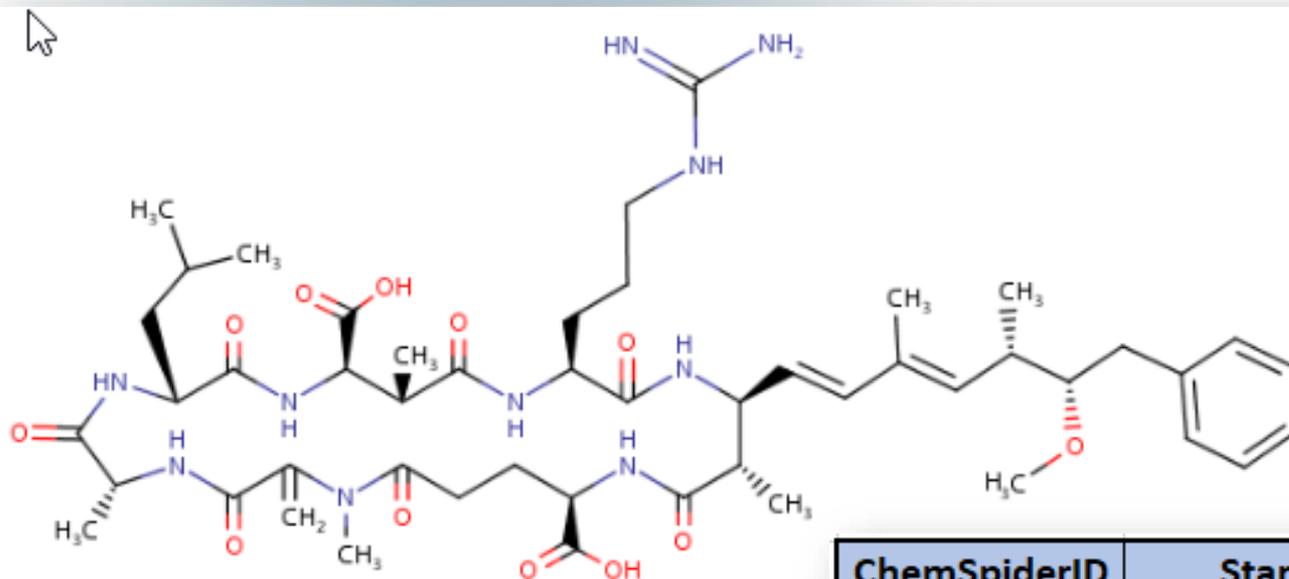


[\(5R,8R,11R,12S,15S,18S,22R\)-15-\(3-Carbamimidamidopropyl\)-8-isobutyl-18-\(\(1E,3E,5R,6R\)-6-methoxy-3,5-dimethyl-7-](#)



[Diamino-N-\(3-\(5R,8S,11R,12S,15S,18S,19S,22R\)-11,2'-dicarboxy-8-isobutyl-18-\(\(1E,3E,5S,6S\)-6-methoxy-3,5-dimethyl-7-](#)

Comparing ChemSpider Structures



ChemSpiderID	Standard InChIKey Stereolayer
WIKIPEDIA	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
CompTox	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
4941647	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
393078	t28-,29-,30-,31+,34-,35-,36+, 37- ,38-,40+
57618348	t28-,29-,30-,31+,34-,35-,36+, 37- ,38-,40+
29342071	t28-,29-,30-,31+, 34+ ,35-,36+, 37- ,38-,40+
7987594	t28-, 29? , 30? ,31+, 34? ,35-, 36? , 37- ,38-, 40?
22900854	t28-, 29? , 30+ , 31- , 34+ , 35+ , 36- , 37- ,38-, 40-
19692240	NONE
2831283	NONE

Comparing ChemSpider Structures

ChemSpiderID	InChIKey	# Stereocenters	# Different
WIKIPEDIA	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
CompTox	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
4941647	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
393078	ZYZCGGRZINLQBL-GWRQVWKTSA-N	10/10	1
57618348	ZYZCGGRZINLQBL-UPPCHHEJSA-N	10/10	1
29342071	ZYZCGGRZINLQBL-IIJTUTQBSA-N	10/10	2
7987594	ZYZCGGRZINLQBL-BESLYTPASA-N	5/10	6
22900854	ZYZCGGRZINLQBL-QAXSDTKVSA-N	9/10	8
19692240	ZYZCGGRZINLQBL-ORZJCNCZSA-N	0/10	10
2831283	ZYZCGGRZINLQBL-UHFFFAOYSA-N	0/10	10

Other Searches



PubChem [About](#)

SEARCH FOR

ZYZCGGRZINLQBL

Treating this query as a text search.

**Compounds
(17)**

Show **All** entries

CMR. Query InChI...	src_id	Source	src_compound_id
...matches...	1	ChEMBL	CHEMBL444092
...matches...	4	Guide to Pharmacology	4735
...matches...	6	KEGG Ligand	C05371
...matches...	7	ChEBI	6925
...matches...	9	ZINC	ZINC000169715525
...matches...	9	ZINC	ZINC000255288110
...matches...	9	ZINC	ZINC000255288111
...matches...	9	ZINC	ZINC000255288112
...matches...	9	ZINC	ZINC000255288113
...matches...	9	ZINC	ZINC000255288114
...matches...	9	ZINC	ZINC000255288115
...matches...	9	ZINC	ZINC000583653042
...matches...	9	ZINC	ZINC000669680403
...matches...	10	eMolecules	26754757
...matches...	10	eMolecules	31239828
...matches...	11	IBM Patent System	DA3C2F25F29692734272194ED0E2C009
...matches...	14	FDA SRS	EQ8332842Y

Processing thousands of chemical signatures as mass and formula

- 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



Step Five: Choose Data Fields to Download

Please enter one identifier per line 

Select Input Type(s)

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

+/- ppm

Enter Identifiers to Search (searches should be limited to <5000 identifiers)

```
41.0265
56.02621
53.0265
58.0418|
93.0578
113.9639
151.8754
69.9377
77.9872
```

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>.

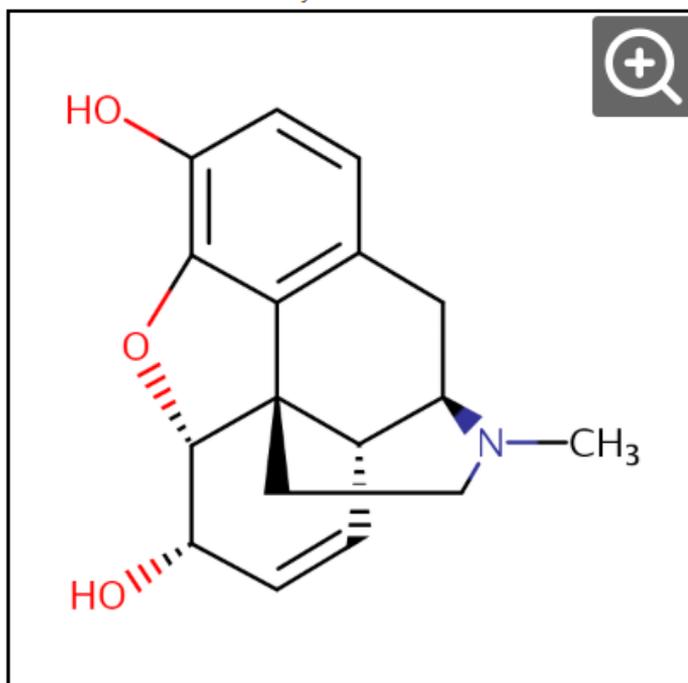
 Display All Chem

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	DTXSID6026567	134-20-3	Methyl 2-aminobenzoate	C8H9NO2	151.063328534	60

Related Searches to Support Mass Spectrometry

Find me “related structures” Formula-Based Search



Wikipedia

Morphine is a pain medication of the opiate family which is found naturally in a number of plants and animals. It acts directly on the central nervous system (CNS) to decrease the feeling of pain. It can be taken for both acute pain and chronic pain. It is frequently used for pain from myocardial infarction and during labor. It can be given by mouth, by injection into a muscle, by injection under the skin, intravenously, injection into the space around the

...

[Read more](#)

Intrinsic Properties

 **Molecular Formula:** C17H19NO3  Mol File  Find All Chemicals

 **Average Mass:** 285.343 g/mol  Isotope Mass Distribution

 **Monoisotopic Mass:** 285.136493 g/mol

Structural Identifiers

Select Chemicals of Interest

Searched by Exact Molecular Formula: C17H19NO5

6 of 120 chemicals selected

Search

Default



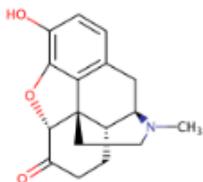
DTXSID

PubChem

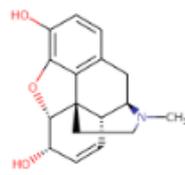
CPDAT

Hide chemicals that are:

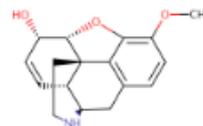
- Unselected
- Isotopes
- Multicomponent Chemicals
- No Structures



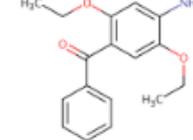
Hydromorphone
DTXSID: DTXSID8023133
PubChem: 39
CPDAT: 21



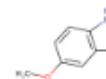
Morphine
DTXSID: DTXSID9023336
PubChem: 52
CPDAT: 37



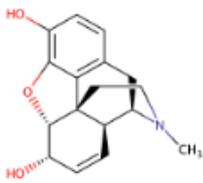
Norcodeine
DTXSID: DTXSID8046327
PubChem: 20
CPDAT: 2



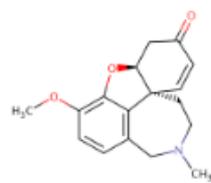
Methanone, (4-amino-2,5-diethoxyphenyl)phenyl
DTXSID: DTXSID3071696
PubChem: 16
CPDAT: 0



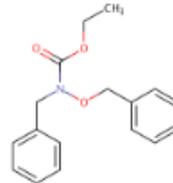
Carbazone
DTXSID: DTXSID...
PubChem: 1
CPDAT: 0



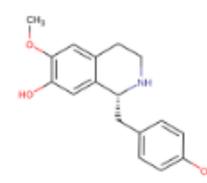
Morphinan-3,6-diol, 7,8-didehydro-4,5-epoxy
DTXSID: DTXSID40167308
PubChem: 8
CPDAT: 0



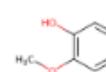
Narwedine, (+/-)-
DTXSID: DTXSID60168190
PubChem: 43
CPDAT: 0



Carbamic acid, (phenylmethoxy)(phenyl)methyl
DTXSID: DTXSID50171931
PubChem: 11
CPDAT: 0

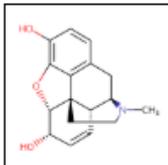


7-Isoquinolinol, 1,2,3,4-tetrahydro-1-((4-hydroxyphenyl)methyl)
DTXSID: DTXSID70176367
PubChem: 21
CPDAT: 0



Cherylin
DTXSID: DTXSID...
PubChem: 1
CPDAT: 0

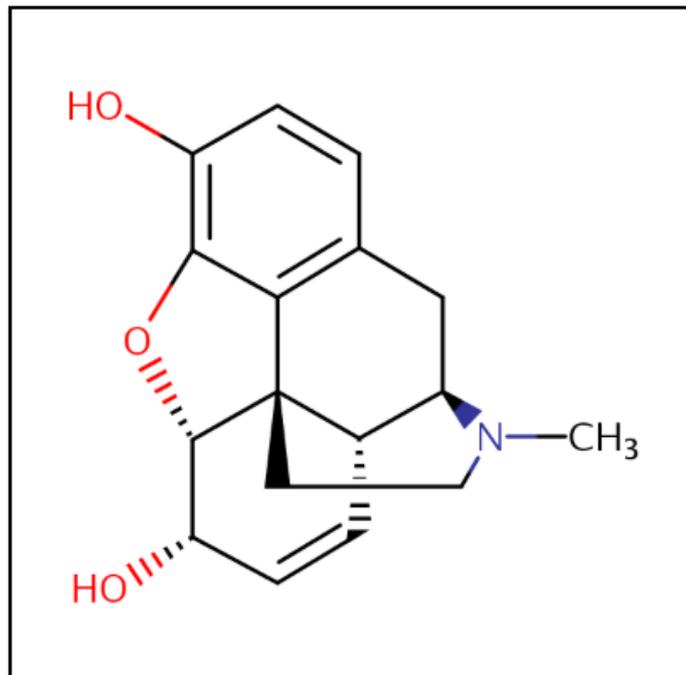
Find me “related structures” Based on Structure Similarity



Morphine

57-27-2 | DTXSID9023336

Searched by Approved Name.



Wikipedia

Intrinsic Properties

Structural Identifiers

Linked Substances

Same Connectivity: [3 records](#) (based on first layer of InChI)

Mixtures, Components and Neutralized Forms: [13 records](#) (based on QSAR ready mappings and with the compound as a component of a mixture)

MS-Ready Mappings: [DTXCID60196731: 21 records](#);

Similar Compounds: [178 records](#) (based on Tanimoto coefficient >0.8)

Presence in Lists

Find me “related structures” Based on Structure Similarity

Searched with a similarity threshold of 0.8

178 chemicals

Select all

Download

Send to Batch Search

Similarity



DTXSID

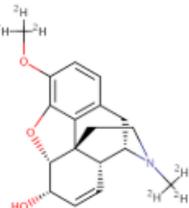
CASRN

TOXCAST

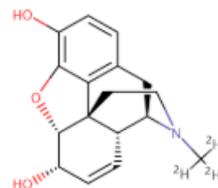
Similarity

Hide chemicals that are:

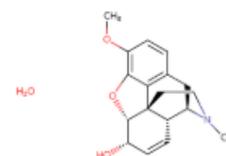
Filter by Name or CASRN



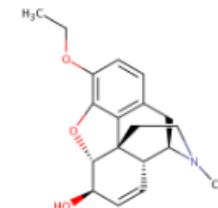
(5alpha,6alpha,9alpha)-17-(²H₂)Methyl-...
DTXSID: DTXSID30747375
CASRN: 1007844-34-9
TOXCAST: -
Similarity: 1.00



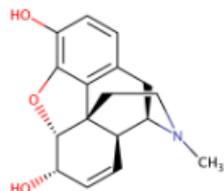
Morphinan-3,6-alpha-diol, 7,8-didehydr...
DTXSID: DTXSID00217656
CASRN: 67293-88-3
TOXCAST: -
Similarity: 1.00



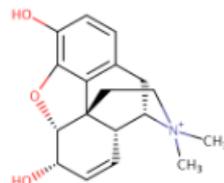
Codeine hydrate
DTXSID: DTXSID20209391
CASRN: 6059-47-8
TOXCAST: -
Similarity: 1.00



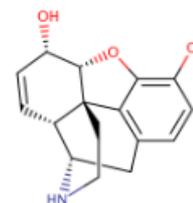
Ethyl-alpha-isomorphine
DTXSID: DTXSID30197073
CASRN: 47252-01-7
TOXCAST: -
Similarity: 1.00



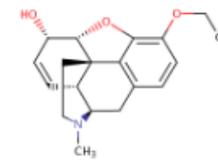
Morphinan-3,6-diol, 7,8-didehydro-4,5-e...



N-Methylmorphine



Normorphine

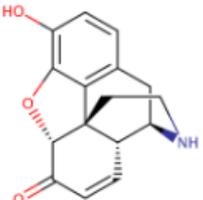
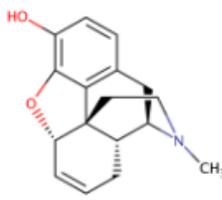
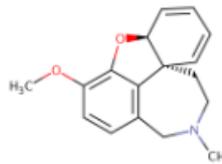
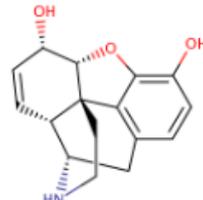
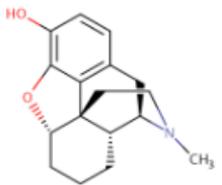
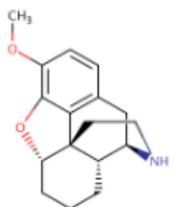
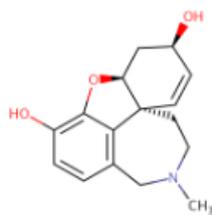
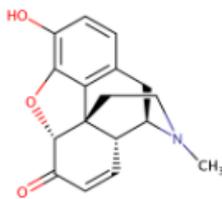


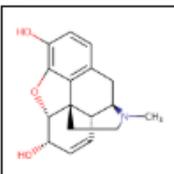
Ethylmorphine

Find me “related structures” Structure Similarity – sort on mass

178 chemicals

Select all Download Send to Batch Search **Mass** CASRN DTXSID Mass Hide chemicals that are: Filter by Name or CASRN

 <p>(5alpha)-7,8-Didehydro-4,5-epoxy-3-hy... CASRN: 6872-48-6 DTXSID: DTXSID00218845 Mass: 269.105193</p>	 <p>Desoxymorphine C CASRN: 63732-65-0 DTXSID: DTXSID20213170 Mass: 269.141579</p>	 <p>R-116937 CASRN: 664995-65-7 DTXSID: DTXSID60216701 Mass: 269.141579</p>	 <p>Normorphine CASRN: 466-97-7 DTXSID: DTXSID1049019 Mass: 271.120843</p>
 <p>Desomorphine CASRN: 427-00-9 DTXSID: DTXSID10195390</p>	 <p>Morphinan-14-ol, 4,5-alpha-epoxy-3-me... CASRN: 55256-27-4 DTXSID: DTXSID90203765</p>	 <p>O-Desmethylgalantamine CASRN: 60755-80-8 DTXSID: DTXSID40209575</p>	 <p>Morphinone CASRN: 467-02-7 DTXSID: DTXSID50196907</p>



Morphine

57-27-2 | DTXSID9023336

Searched by Approved Name.

Abstract Sifter

1) Select PubMed starting point query then 2) click on Retrieve. 

Select a Query Term	Retrieve Articles
Select a Query Term	
Hazard	
Fate and Transport	
Metabolism/PK/PD	
Chemical Properties	
Exposure	
Mixtures	
Male Reproduction	
Androgen Disruption	
Female Reproduction	
GeneTox	
Cancer	
Clinical Trials	
Embryo and embryonic development	
Child (infant through adolescent)	
Dust and Exposure	
Food and Exposure	
Water and Exposure	
Algae	
Disaster / Emergency	

Optionally, edit the query before retrieving.

"57-27-2" OR "Morphine"

Literature Searching

Child (Infant through adolescent)

Dust and Exposure

Food and Exposure

Water and Exposure

Algae

Disaster / Emergency



^{WS} Optionally, edit the query before retrieving.

```
("57-27-2" OR "Morphine") AND ((water OR groundwater OR drinking water) AND Environmental Exposure)
```



Literature Searching

37 of 37 articles loaded...



To find articles quickly, enter terms to sift abstracts.

<input type="checkbox"/>	wastewater	Spectrometry ↓	EPA	Total	PMID	Year	Title	Authors	Journal	Rev
<input type="checkbox"/>	4	2	0	6	29274731	2017	Simultaneous analysis of opioid analgesics and thei...	Krizman-Matasic; Kostanjevecki; Ahel; Terzic	Journal of chromatography. A	
<input type="checkbox"/>	0	1	0	1	25768972	2015	Evaluating external contamination of polybrominate...	Poon; Aleksa; Carnevale; Kapur; Goodyer; Koren	Therapeutic drug monitoring	
<input type="checkbox"/>	0	1	0	1	22544551	2012	Spatial distribution of illicit drugs in surface waters o...	Vazquez-Roig; Andreu; Blasco; Morillas; Picó	Environmental science and pollution research inter...	
<input type="checkbox"/>	1	1	0	2	20801487	2010	Analysis of illicit and illicit drugs in waste, surface an...	Berset; Brenneisen; Mathieu	Chemosphere	
<input type="checkbox"/>	1	1	0	2	17935751	2007	Illicit drugs, a novel group of environmental contami...	Zuccato; Castiglioni; Bagnati; Chiabrando; Grassi; ...	Water research	
<input type="checkbox"/>	2	1	1	4	17607391	2007	Using environmental analytical data to estimate lev...	Bones; Thomas; Paull	Journal of environmental monitoring : JEM	
<input type="checkbox"/>	3	1	2	6	17180984	2006	Simultaneous determination of psychoactive drugs ...	Hummel; Löffler; Fink; Ternes	Environmental science & technology	
<input type="checkbox"/>	6	0	0	6	30583189	2018	Assessment of drugs of abuse in a wastewater trea...	Kumar; Tschärke; O'Brien; Mueller; Wilkins; Padhye	The Science of the total environment	
<input type="checkbox"/>	0	0	3	3	30488421	2018	Effect of enriched environment during adolescence ...	Mohammadian; Najafi; Miladi-Gorji	Developmental psychobiology	
<input type="checkbox"/>	3	0	0	3	29574368	2018	Estimation of the consumption of illicit drugs during ...	Foppe; Hammond-Weinberger; Subedi	The Science of the total environment	
<input type="checkbox"/>	1	0	0	1	28787791	2017	Evaluation of in-sewer transformation of selected illi...	Gao; Banks; Li; Jiang; Lai; Mueller; Thai	The Science of the total environment	
<input type="checkbox"/>	9	0	0	9	28472697	2017	Occurrence and fate of illicit drugs and pharmaceuti...	Causanilles; Ruepert; Ibáñez; Emke; Hernández; d...	The Science of the total environment	
<input type="checkbox"/>	0	0	0	0	28010888	2016	Dose-dependent effects of morphine on lipopolysac...	Mottaz; Schönenberger; Fischer; Eggen; Schirmer; ...	Environmental pollution (Barking, Essex : 1987)	
<input type="checkbox"/>	0	0	0	0	27746311	2016	Effects of voluntary exercise on the viability, prolifer...	Haydari; Safari; Zarbakhsh; Bandegi; Miladi-Gorji	Neuroscience letters	
<input type="checkbox"/>	0	0	0	0	27261879	2016	Genotoxic effects induced by the exposure to an en...	Parolini; Magni; Castiglioni; Binelli	Ecotoxicology and environmental safety	
<input type="checkbox"/>	3	0	0	3	27179320	2016	Temporal trends in drug use in Adelaide, South Aus...	Tschärke; Chen; Gerber; White	The Science of the total environment	

The benefits of building chemical lists

Chemical Lists

Home Advanced Search Batch Search Lists ▾ Predictions Downloads

Lists of Chemicals

List of Assays

Download ▾

Columns ▾

mass

Copy Filtered Lists URL

List Acronym ▾	List Name ▾	Last Updated ▾	Number of Chemicals ▾	List Description ▾
HDXEXCH	MASSPECDB: Hydrogen Deuterium Exchange Standard Set - Under HDX Conditions	2018-11-07	592	Observed species (deuterated and undeuterated) from the HDXNOEX list under hydrogen deuterium exchange conditions (Ruttkies, Schymanski et al. in prep.)
HDXNOEX	MASSPECDB: Hydrogen Deuterium Exchange Standard Set - No Exchange	2018-11-07	765	Environmental standard set used to investigate hydrogen deuterium exchange in small molecule high resolution mass spectrometry (Ruttkies, Schymanski et al. in prep.)
MASSBANKEUSP	MASSPECDB: MassBank.EU Collection: Special Cases	2017-07-16	263	The MassBank.EU list contains curated chemicals (Schymanski/Williams) associated with the literature/tentative/unknown/SI spectra available on MassBank.EU that are not available as part of the full MassBank collection of reference standard spectra.
MASSBANKREF	MASSPECDB: MassBank Reference Spectra Collection	2017-07-13	1267	This MassBank list contains chemicals associated with the full MassBank collection of reference standard spectra available on MassBank.EU, MassBank.JP and MassBank of North America as well as the Open Data collection, curated by Williams/Schymanski.
MYCOTOXINS	MASSPECDB: Mycotoxins from MassBank.EU	2017-08-02	88	This is a set of mycotoxins, initiated by the contribution of spectra of 90 mycotoxins to MassBank.EU by Justin Renaud and colleagues from Agriculture and Agri-Food Canada, Government of Canada

WATER|EPA; Chemicals associated with hydraulic fracturing

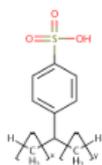
 Identifier substring search

List Details

Description: Chemicals used in hydraulic fracturing fluids and/or identified in produced water from 2005-2013, corresponding to chemicals listed in Appendix H of EPA's Hydraulic Fracking Drinking Water Assessment Final Report (Dec 2016). Citation: U.S. EPA, Hydraulic Fracturing for Oil and Gas: Impacts from the Hydraulic Fracturing Water Cycle on Drinking Water Resources in the United States (Final Report). U.S. Environmental Protection Agency, Washington, D.C. EPA/600/R-16/236F, 2016. <https://www.epa.gov/hfstudy>

*Note that Appendix H chemical listings in Tables H-2 and H-4 were mapped to current DSSTox content, which has undergone additional curation since the publication of the original EPA HF Report (Dec 2016). In the few cases where a Chemical Name and CASRN from the original report map to distinct substances (as of Jan 2018), both were included in the current EPAHF chemical listing for completeness; additionally, 34 previously unmapped chemicals in Table H-5 are now registered in DSSTox (all but 2 assigned CASRN) and, thus, have been added to the current EPAHF listing.

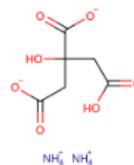
Number of Chemicals: 1640



Alkylbenzenesulfonate, linear
DTXSID: DTXSID3020041
PubChem: 0
CPDAT: 83



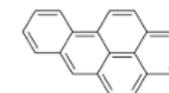
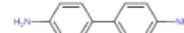
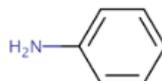
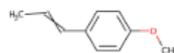
Ammonium chloride
DTXSID: DTXSID0020078
PubChem: 82
CPDAT: 260



Diammonium citrate
DTXSID: DTXSID5020079
PubChem: 19
CPDAT: 18



Ammonium hydroxide
DTXSID: DTXSID4020080
PubChem: 83
CPDAT: 857



PFAS lists of Chemicals

Select List

Download ▾

Columns ▾

PFAS

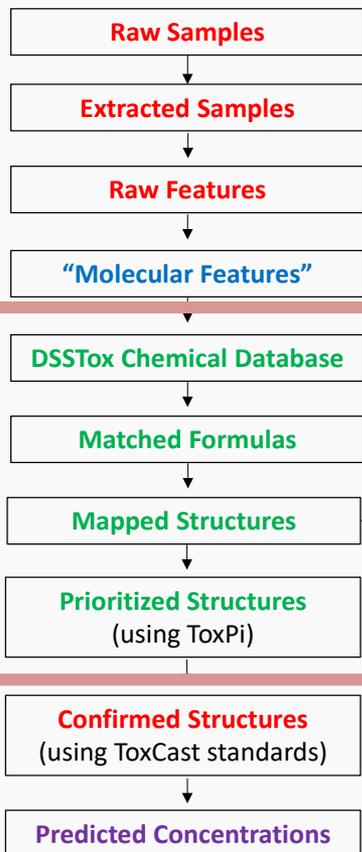
Copy Filtered Lists URL

List Acronym	List Name	Last Updated	Number of Chemicals	List Description
EPAPFAS75S1	PFAS EPA: List of 75 Test Samples (Set 1)	2018-06-29	74	PFAS list corresponds to 75 samples (Set 1) submitted for initial testing screens conducted by EPA researchers in collaboration with researchers at the National Toxicology Program.
EPAPFAS75S2	PFAS EPA: List of 75 Test Samples (Set 2)	2019-02-21	75	PFAS list corresponds to a second set of 75 samples (Set 2) submitted for testing screens conducted by EPA researchers in collaboration with researchers at the National Toxicology Program.
EPAPFASCAT	PFAS EPA Structure-based Categories	2018-06-29	64	List of registered DSSTox "category substances" representing PFAS categories created using ChemAxon's Markush structure-based query representations.
EPAPFASINSOL	PFAS EPA: Chemical Inventory Insoluble in DMSO	2018-06-29	43	PFAS chemicals included in EPA's expanded ToxCast chemical inventory found to be insoluble in DMSO above 5mM.
EPAPFASINV	PFAS EPA: ToxCast Chemical Inventory	2018-06-29	430	PFAS chemicals included in EPA's expanded ToxCast chemical inventory and available for testing.
EPAPFASRL	PFAS EPA: Cross-Agency Research List	2017-11-16	199	EPAPFASRL is a manually curated listing of mainly straight-chain and branched PFAS (Per- & Poly-fluorinated alkyl substances) compiled from various internal, literature and public sources by EPA researchers and program office representatives.
PFASKEMI	PFAS: List from the Swedish Chemicals Agency (KEMI) Report	2017-02-09	2416	Perfluorinated substances from a Swedish Chemicals Agency (KEMI) Report on the occurrence and use of highly fluorinated substances.
PFASMASTER	PFAS Master List of PFAS Substances	2018-07-26	5061	PFASMASTER is a consolidated list of PFAS substances spanning and bounded by the below lists of current interest to researchers and regulators worldwide.
PFASOECD	PFAS: Listed in OECD Global Database	2018-05-16	4729	OECD released a New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances, (PFASs) listing more than 4700 new PFAS
PFASTRIER	PFAS Community-Compiled List (Trier et al., 2015)	2017-07-16	597	PFASTRIER community-compiled public listing of PFAS (Trier et al, 2015)

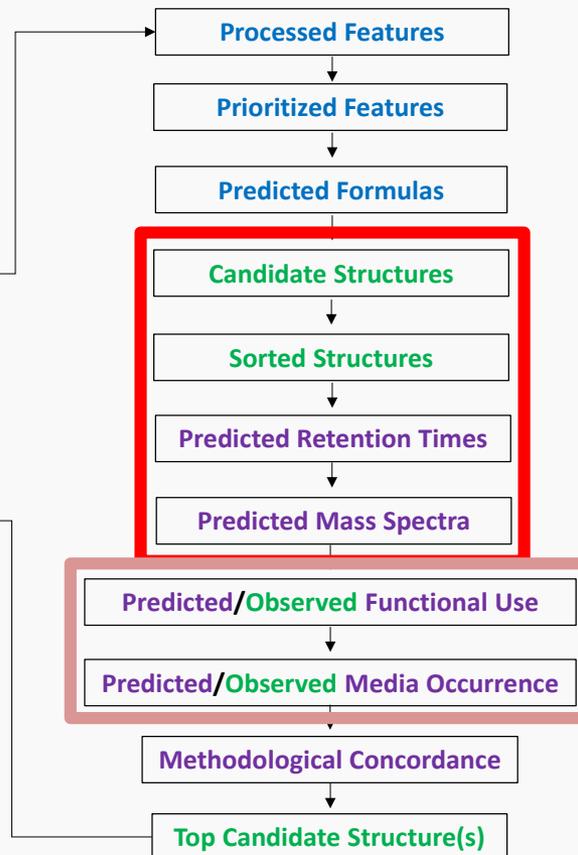
Ongoing Research in Progress

Suspect Screening and Non-Targeted Analysis Workflow

Suspect Screening



Non-Targeted Analysis



Color Key

Red = Analytical Chemistry

Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

Green = Informatics & Web Services



- Predicted Spectra for candidate ranking
 - Viewing and Downloading pre-predicted spectra
 - Search spectra against the database

Predicted Mass Spectra

<http://cfmid.wishartlab.com/>



[Metabolomics](#)

February 2015, Volume 11, [Issue 1](#), pp 98-110 | [Cite as](#)

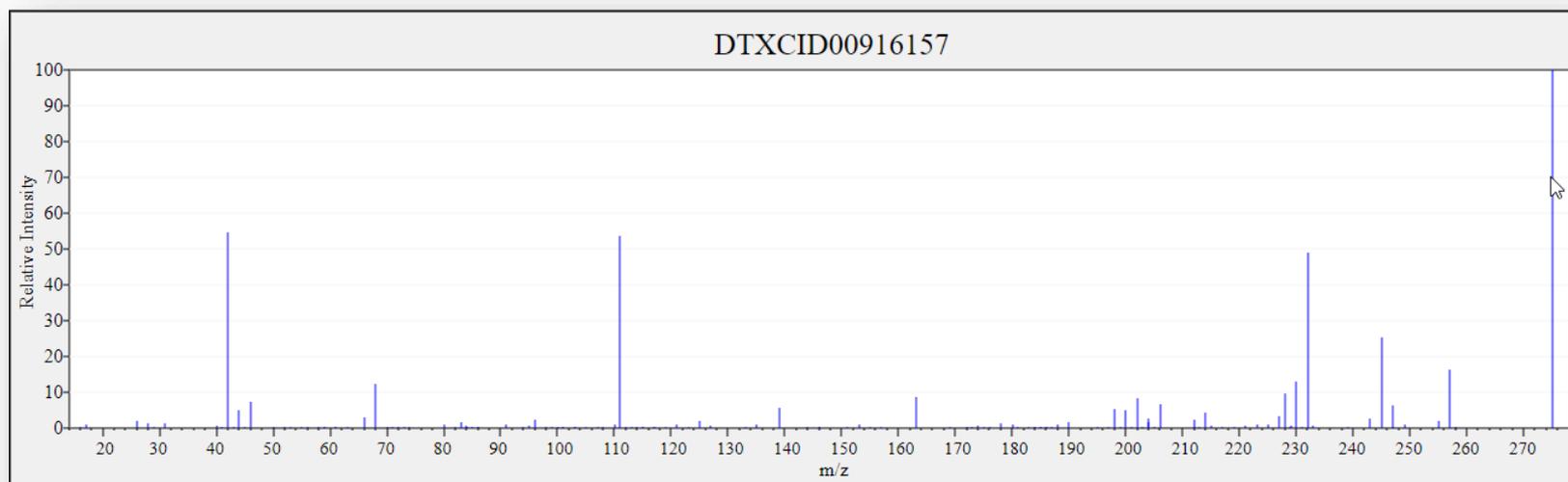
Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification

[Authors](#)

[Authors and affiliations](#)

Felicity Allen , Russ Greiner, David Wishart

- MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >800,000 structures, to be accessible via Dashboard



Search Expt. vs. Predicted Spectra

Non Target Analysis Prototype

Mass Search

Min/Max

Da Da ppm

Molecular Formula Search

Mass or Formula must be entered before searching spectrum

Ionization Type

ESI+ ▼
ESI+
ESI-
EI

Spectra Input

Single Energy Multiple

304.1332052 11.6199475
198.0913404 7.308439699
123.0440559 6.538348292
198.0756904 5.269463115
216.1019051 4.700461978
200.1080005 4.800144384

Peak Match Window: Da ppm

Search Expt. vs. Predicted Spectra

Spectra Input

Single Ener

304.1332052 11.61
198.0913404 7.30
123.0440559 6.53
196.0756904 5.28
216.1019051 4.70
200.4080005 4.80

Peak Match

Search

TSV CSV Excel

Chemical Structure ID

DTXCID101048191

DTXCID101181567

DTXCID50879086

DTXCID60686349

DTXCID00830900

DTXCID10971176

DTXCID60301242

DTXCID40703048

DTXCID60349982

DTXCID10316649

Showing 1 to 10 of 38 entries

Chemical Structure ID

Score (10eV)

DTXCID101048191

0.22

DTXCID101181567

0.19

DTXCID50879086

0.17

DTXCID60686349

0.14

DTXCID00830900

0.13

DTXCID10971176

0.12

DTXCID60301242

0.12

DTXCID40703048

0.11

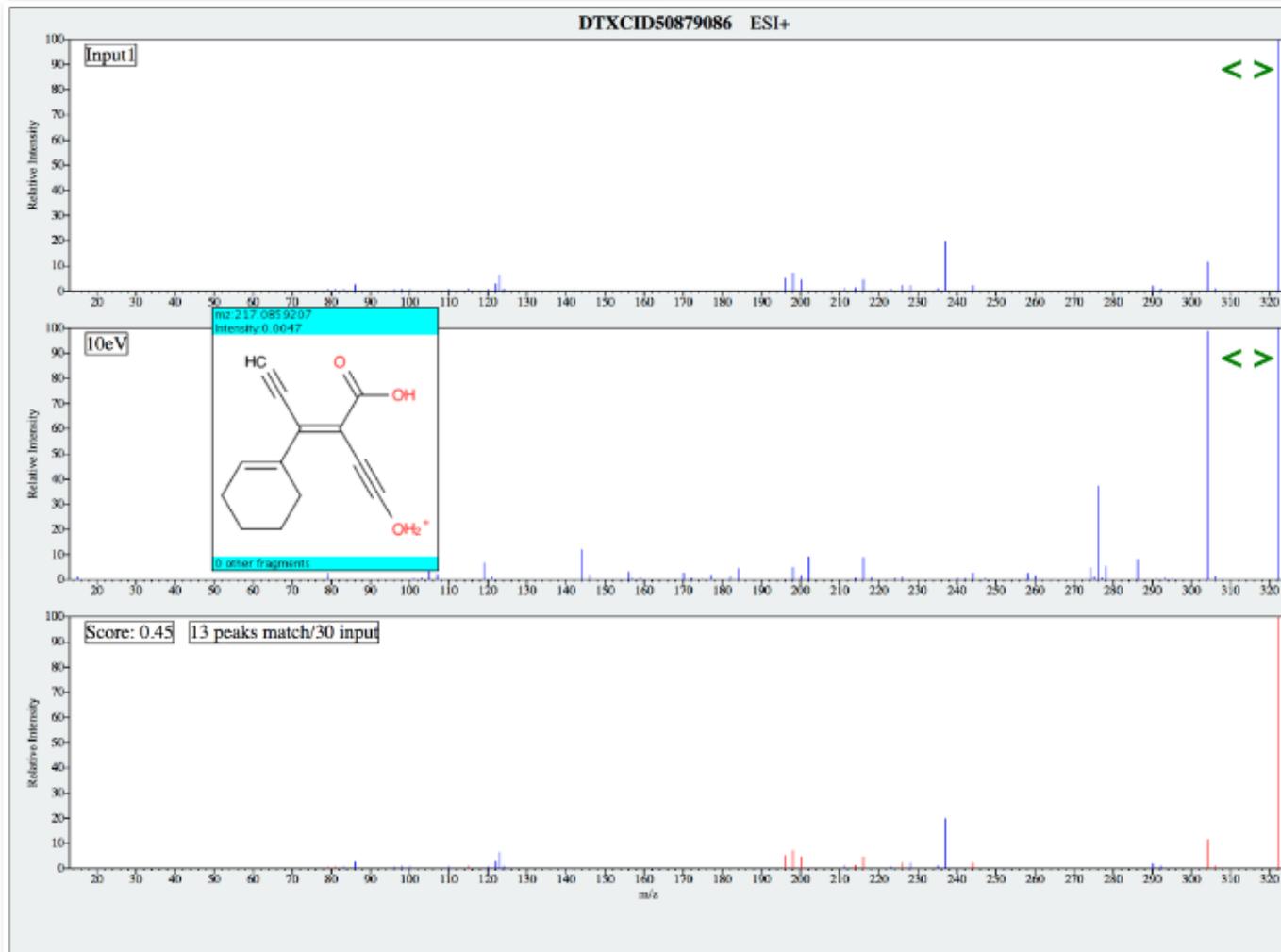
DTXCID60349982

0.11

DTXCID10316649

0.09

Spectral Viewer Comparison



Peak Window: 2.0

ppm Update Score

Input	10eV	20eV	40eV
1	0.45	0.16	0.029
2	0.45	0.16	0.029
3	0.45	0.16	0.029



- Critical Assessment of Small Molecule Identification
 - Training data= 312 peak lists (from 285 substances)
 - 234 MS/MS in positive mode
 - 58 in negative mode
 - Challenge Data= 208 peak lists (from 188 substances)
 - 127 in positive mode
 - 81 in negative mode
- Precursor ion search window= 15 ppm
- Fragment ion match threshold= 0.02 Da
- Candidates limited to Dashboard results within precursor ion search window – **early work on 765K ONLY**



CASMI 2016 Contest Challenge Set (n=208)

CFM-ID only

	# Identified	% of Total
#1 Hits	89	43%
Top 5	154	74%
Top 10	174	84%
Top 20	190	91%

CFM-ID +DSSTox Data Sources

	# Identified	% of Total
#1 Hits	154	74%
Top 5	195	94%
Top 10	198	95%
Top 20	202	97%

- Predicted Spectra for candidate ranking
 - Viewing and Downloading pre-predicted spectra
 - Search spectra against the database
- **Retention Time Index Prediction**

25



Talanta

Volume 182, 15 May 2018, Pages 371-379



A comparison of three liquid chromatography (LC) retention time prediction models

Andrew D. McEachran ^{a, b}  , Kamel Mansouri ^{a, b, 1}  , Seth R. Newton ^c , Brandiese E.J. Beverly ^{a, c, 2}  , Jon R. Sobus ^c , Antony J. Williams ^b  

 [Show more](#)

<https://doi.org/10.1016/j.talanta.2018.01.022>

[Get rights and content](#)



Journal of Hazardous Materials

Volume 363, 5 February 2019, Pages 277-285



Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants

Reza Aalizadeh, Maria-Christina Nika, Nikolaos S. Thomaidis  

 [Show more](#)

<https://doi.org/10.1016/j.jhazmat.2018.09.047>

[Get rights and content](#)

- Predicted Spectra for candidate ranking
 - Viewing and Downloading pre-predicted spectra
 - Search spectra against the database
- Retention Time Index Prediction
- **Structure/substructure/similarity search**

Prototype Development

AADashboard

atrazine

Search

Select properties to predict

H
C
N
O

T.E.S.T. 18

OPERA

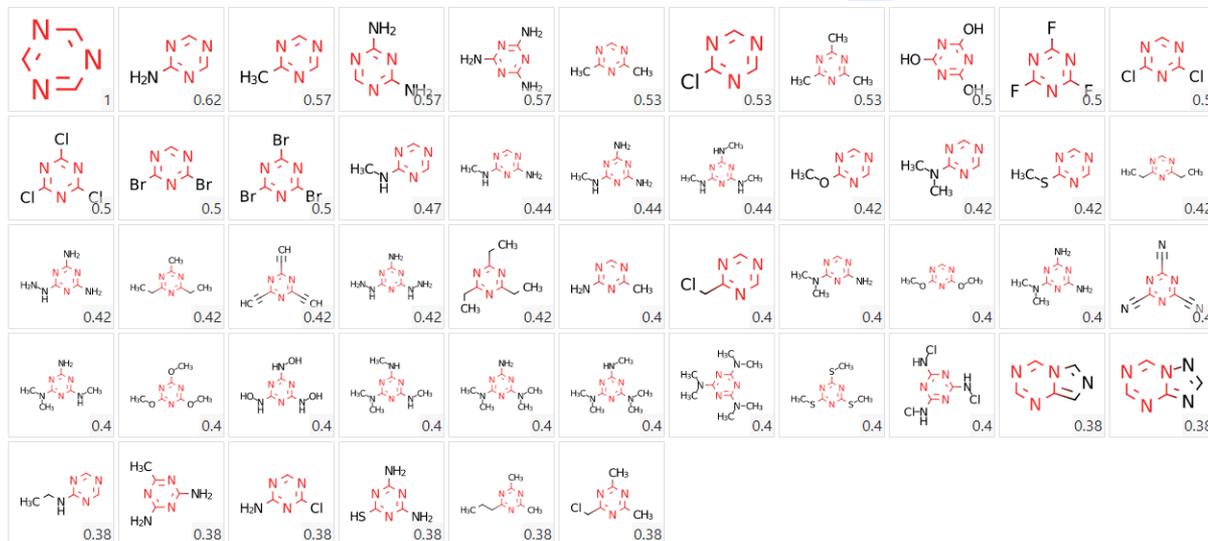
Search

- Exact
- Substructure

Search result 2540

Show Isotopically Labeled Charged Salts or Mixtures

Sort Similarity



Search result 2540

Show Isotopically Labeled

Prototype Development

atrazine Search

100%

Select properties to predict

H T.E.S.T. 18 OPERA Search

C

N

O

S

P

F

Cl

Br

Exact

Substructure

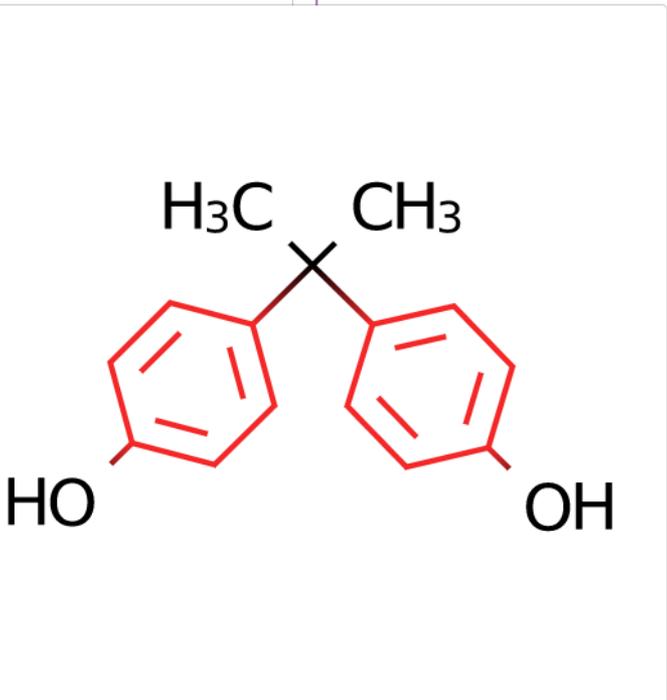
Similarity

Molecular Formula

Molecular Weight

Input formula (e.g. C6 H6):

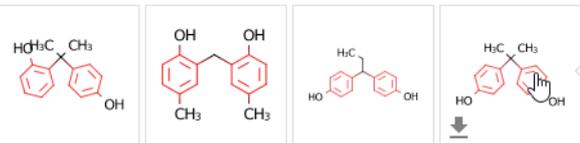
C15H16O2



Search

Search result 5

Show Isotopically Labeled Cl

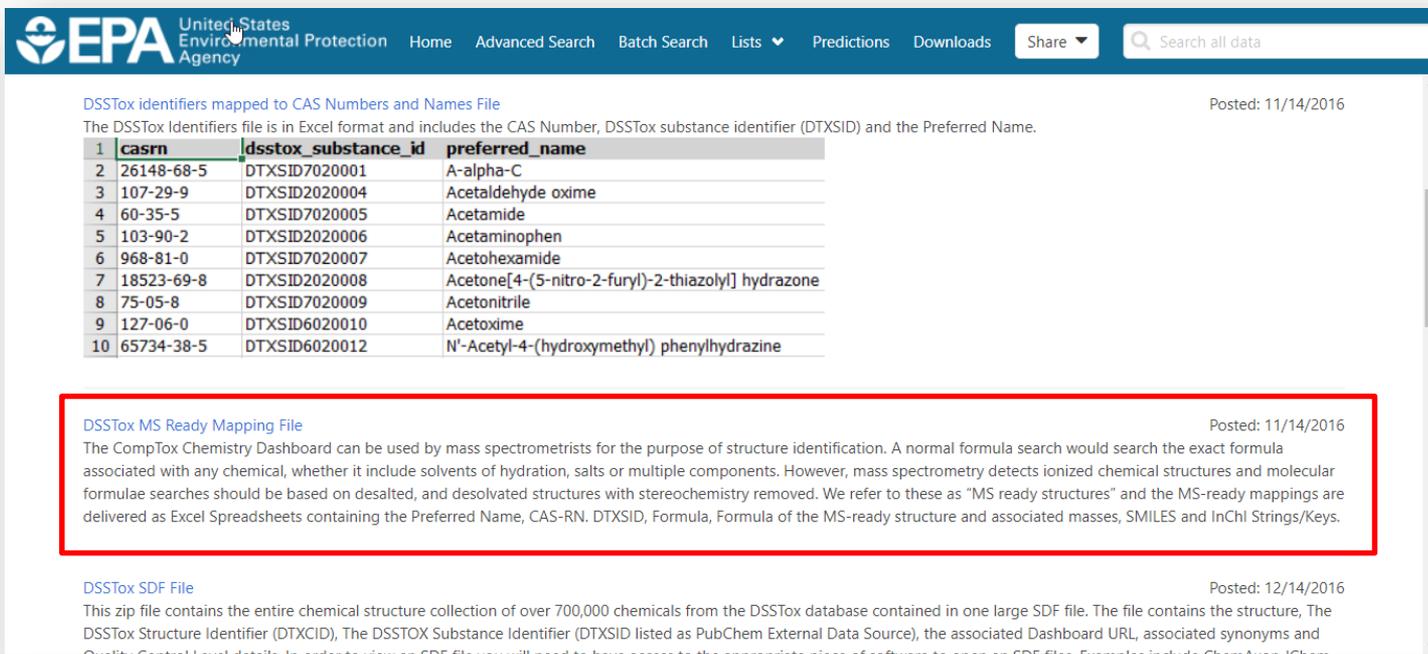


Elements per page 50

1

- Predicted Spectra for candidate ranking
 - Viewing and Downloading pre-predicted spectra
 - Search spectra against the database
- Retention Time Index Prediction
- Structure/substructure/similarity search
- **Access to API and web services for programmatic access**

- Groups waiting on our API and web services
- Mass Spec companies instrument integration
- Release will be in iterations but for now our data are available



The screenshot shows the EPA website interface. At the top, there is a navigation bar with the EPA logo, "United States Environmental Protection Agency", and links for Home, Advanced Search, Batch Search, Lists, Predictions, Downloads, and a Share button. A search bar is also present. Below the navigation bar, there are three main sections:

- DSSTox identifiers mapped to CAS Numbers and Names File** (Posted: 11/14/2016)
The DSSTox Identifiers file is in Excel format and includes the CAS Number, DSSTox substance identifier (DTXSID) and the Preferred Name.

1	casrn	dsstox_substance_id	preferred_name
2	26148-68-5	DTXSID7020001	A-alpha-C
3	107-29-9	DTXSID2020004	Acetaldehyde oxime
4	60-35-5	DTXSID7020005	Acetamide
5	103-90-2	DTXSID2020006	Acetaminophen
6	968-81-0	DTXSID7020007	Acetohexamide
7	18523-69-8	DTXSID2020008	Acetone[4-(5-nitro-2-furyl)-2-thiazolyl] hydrazone
8	75-05-8	DTXSID7020009	Acetonitrile
9	127-06-0	DTXSID6020010	Acetoxime
10	65734-38-5	DTXSID6020012	N'-Acetyl-4-(hydroxymethyl) phenylhydrazine
- DSSTox MS Ready Mapping File** (Posted: 11/14/2016)
The CompTox Chemistry Dashboard can be used by mass spectrometrists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formulae searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as "MS ready structures" and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN, DTXSID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys.
- DSSTox SDF File** (Posted: 12/14/2016)
This zip file contains the entire chemical structure collection of over 700,000 chemicals from the DSSTox database contained in one large SDF file. The file contains the structure, The DSSTox Structure Identifier (DTXCID), The DSSTOX Substance Identifier (DTXSID listed as PubChem External Data Source), the associated Dashboard URL, associated synonyms and Quality Control Level details. In order to view an SDF file you will need to have access to the appropriate piece of software to open an SDF file. Examples include ChemAxon JChem

Benefiting the community with Open Data

NORMAN Suspect List Exchange

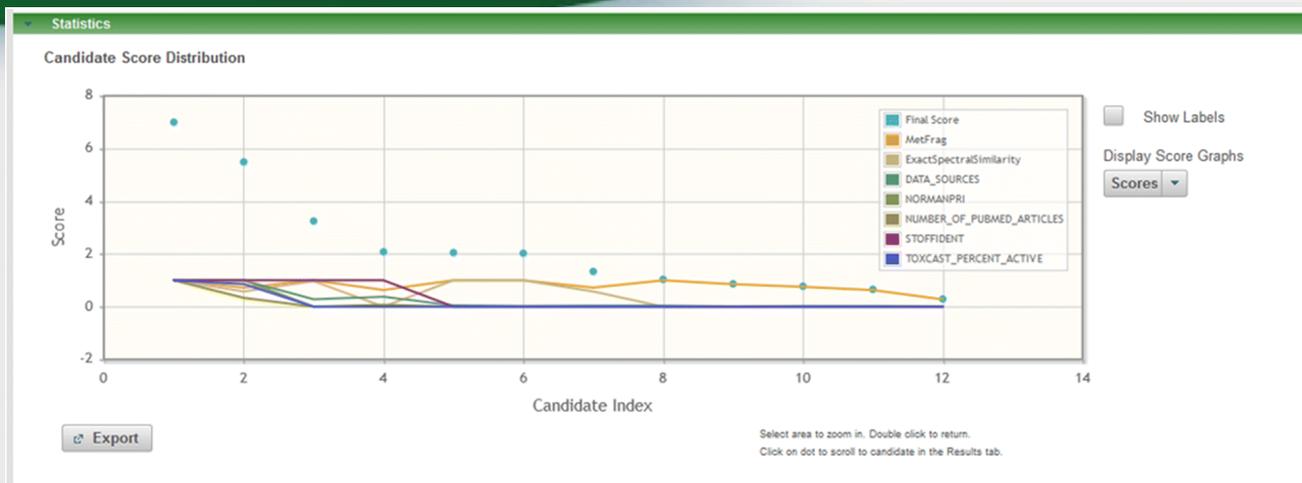
<https://www.norman-network.com/?q=node/236>

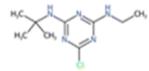
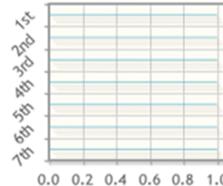
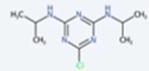
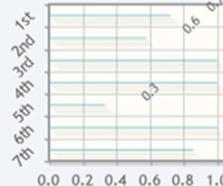
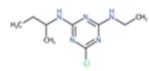
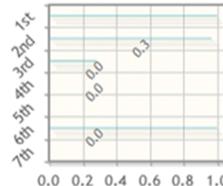


Wastewater Suspect List based on Swedish Product Data	Wastewater Suspect List Original File with Mapped DTXSIDs (12/02/2019)	KEMIWWSUS InChIKeys (12/02/2019)	A prioritized list of 1,123 substances relevant for wastewater based on Swedish product registry data, including scores. Provided by Stellan Fischer, KEMI.
Algal toxins list from CompTox	ALGALTOX XLSX, CSV (14/02/2019) CompTox ALGALTOX List	ALGALTOX InChIKeys (14/02/2019)	List of algal toxins (generated during blooms) from the CompTox Chemicals Dashboard.
CCL 4 Chemical Candidate List	CCL4 XLSX, CSV (14/02/2019) CompTox CCL4 List	CCL4 InChIKeys (14/02/2019)	Contaminants that are not (yet) regulated in the USA but are known or anticipated to occur in public water systems; from CompTox.
Hydrogen Deuterium Exchange (HDX) Standard Set	HDXNOEX XLSX, CSV (14/02/2019) CompTox HDXNOEX List CompTox HDXEXCH List	HDXNOEX InChIKeys (14/02/2019)	Environmental standard set used to investigate hydrogen deuterium exchange in small molecule HRMS (Ruttkies et al. submitted). HDXEXCH list also contains observed deuterated species.
Neurotoxicants Collection from Public Resources	NEUROTOXINS XLSX, CSV (14/02/2019) CompTox NEUROTOXINS List	NEUROTOXINS InChIKeys (14/02/2019)	A list of neurotoxicants compiled from public resources, details on CompTox and Schymanski <i>et al.</i> (submitted).
Statins Collection from Public Resources	STATINS XLSX, CSV (14/02/2019) CompTox STATINS List	STATINS InChIKeys (14/02/2019)	A list of statins (lipid-lowering medications) compiled from public resources, details on CompTox.
Synthetic Cannabinoids and Psychoactive Compounds	SYNTHCANNAB XLSX, CSV (14/02/2019) CompTox SYNTHCANNAB List	SYNTHCANNAB InChIKeys (14/02/2019)	A list of synthetic cannabinoids and psychoactive compounds assembled from public resources, from CompTox.

Integration to MetFrag in place

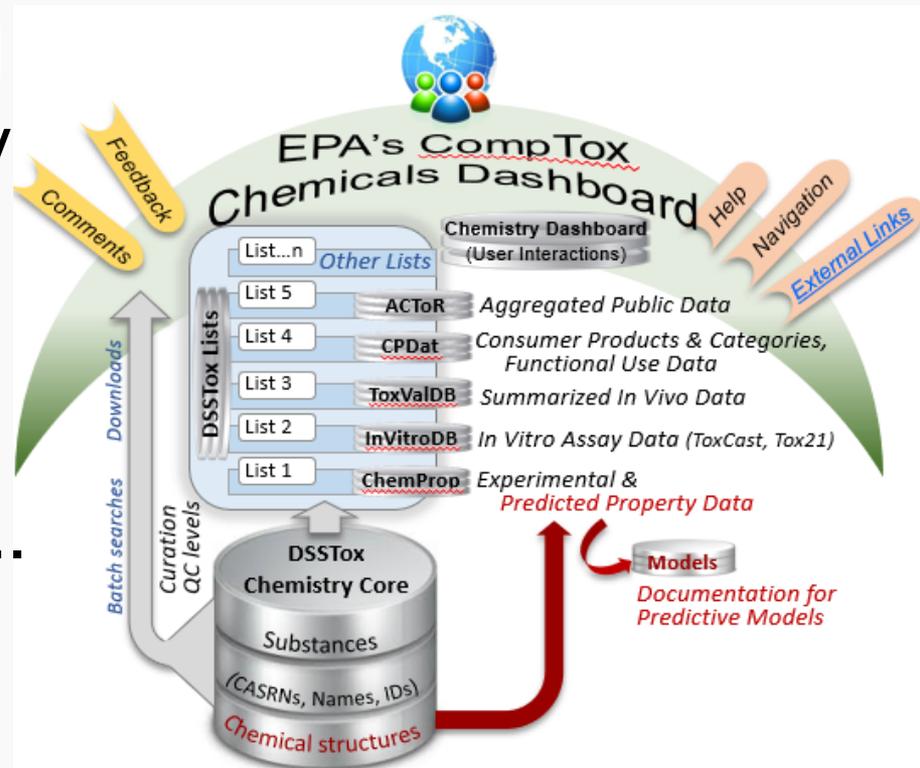
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0299-2>



#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1	 Terbutylazine	DTXSID4027608 InChIKeyBlock1 = FZXISNSWEXTPMF	229.10948	C ₉ H ₁₆ ClN ₅		7.0	Peaks: 10 / 14 Fragments Scores Download
2	 Propazine	DTXSID3021196 InChIKeyBlock1 = WJNRPIHGGKWCK	229.10948	C ₉ H ₁₆ ClN ₅		5.4894	Peaks: 7 / 14 Fragments Scores Download
3	 Sebuthylazine	DTXSID7058171 InChIKeyBlock1 = BZRUVKZGXNSXMB	229.10948	C ₉ H ₁₆ ClN ₅		3.2476	Peaks: 10 / 14 Fragments Scores Download

Conclusion

- Dashboard access to data for ~875,000 chemicals
- MS-Ready data facilitates structure identification
- Related metadata facilitates candidate ranking
- Relationship mappings and chemical lists of great utility
- Dashboard and contents are one part of the solution
- We are committed to open API development with time..



Journal of Exposure Science & Environmental Epidemiology (2018) 28:411–426

<https://doi.org/10.1038/s41370-017-0012-y>

REVIEW ARTICLE



Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA

Jon R. Sobus¹ · John F. Wambaugh² · Kristin K. Isaacs¹ · Antony J. Williams² · Andrew D. McEachran³ · Ann M. Richard² · Christopher M. Grulke² · Elin M. Ulrich¹ · Julia E. Rager^{3,4} · Mark J. Strynar¹ · Seth R. Newton¹



ENVIRONMENTAL
Science & Technology

Article

Cite This: *Environ. Sci. Technol.* 2018, 52, 3125–3135

pubs.acs.org/est

Suspect Screening Analysis of Chemicals in Consumer Products

Katherine A. Phillips,[†] Alice Yau,[‡] Kristin A. Favela,[‡] Kristin K. Isaacs,[†] Andrew McEachran,^{§,||} Christopher Grulke,^{||} Ann M. Richard,^{||} Antony J. Williams,^{||} Jon R. Sobus,[†] Russell S. Thomas,^{||} and John F. Wambaugh^{*,||}

- IT Development team – especially Jeff Edwards and Jeremy Dunne
- Chris Grulke for the ChemReg system
- NERL colleagues – Jon Sobus, Elin Ulrich, Mark Strynar, Seth Newton, Alex Chao
- Reza Aalizadeh & Nikolaos Thomaidis, Athens
- Emma Schymanski, LCSB, Luxembourg
- The NORMAN Network and all contributors

Antony Williams

US EPA Office of Research and Development
National Center for Computational Toxicology

EMAIL: Williams.Antony@epa.gov

ORCID: <https://orcid.org/0000-0002-2668-4821>