Structure identification by Mass Spectrometry Non-Targeted Analysis using the US EPA’s CompTox Chemistry Dashboard

Antony Williams¹, Andrew D. McEachran², Chris Grulke¹, Seth Newton³, Kristin Isaacs³, Katherine Phillips³, Nancy Baker¹ and Jon R. Sobus³

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC
2) Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC
3) 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

March 2018
ACS Spring Meeting, New Orleans
Two Years of Development for the CompTox Chemistry Dashboard

• The Chemistry Dashboard went online on April 1^{st} 2016
• Two years later, with 10k users a week, it is fulfilling the promise with an underlying architecture for integrating CompTox data
• Used by our mass spectrometry team (in NERL) on a daily basis – suspect screening and non-targeted analysis
Suspect screening and non-targeted analysis

Seth R. Newtona, *, Rebecca L. McMahonb, Jon R. Sob, Andrew D. MeEachranb,c, Mark J. Sryanad

Integrating tools for non-targeted analysis

Suspect Screening Analysis of

Katherine A. Phillips†, Alice Yau†, Kristin A. Gruke†, Ann M. Richard†, Antony J. Williams†,

Open Science for Identifying “Known Unknown” Chemicals

Emma L. Schymanski† and Antony J. Williams†
The CompTox Chemistry Dashboard

• A publicly accessible website delivering access:
  – ~760,000 chemicals with related property data
  – Experimental and predicted physicochemical property data
  – Experimental Human and Ecological hazard data
  – Integration to “biological assay data” for 1000s of chemicals
  – Information regarding consumer products containing chemicals
  – Links to other agency websites and public data resources
  – “Literature” searches for chemicals using public resources
  – “Batch searching” for thousands of chemicals
  – Real time prediction of physchem and toxicity endpoints
Propylene glycol (IUPAC name: propane-1,2-diol) is a synthetic organic compound with the chemical formula C₃H₈O₂. It is a viscous colorless liquid which is nearly odorless but possesses a faintly sweet taste. Chemically it is classed as a diol and is miscible with a broad range of solvents, including water, acetone, and chloroform. It is produced on a large scale and is primarily used in the production of polymers, but also sees use in food...Read more

Intrinsic Properties

Structural Identifiers

Linked Substances

Presence in Lists

Record Information

Quality Control Notes
### Access to Chemical Hazard Data

#### Chemistry Dashboard | EPAHFR

**Download table as:** TSV Excel

**Risk Assessment Class:** Human Eco

<table>
<thead>
<tr>
<th>Priority</th>
<th>Type</th>
<th>Subtype</th>
<th>Risk Assessment Class</th>
<th>Values</th>
<th>Units</th>
<th>Study Type</th>
<th>Exposure Route</th>
<th>Species</th>
<th>Subsource</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Cardiovascular</td>
<td>subchronic</td>
<td>5000.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Endocrine</td>
<td>subchronic</td>
<td>5000.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>LOEL</td>
<td>Hematological</td>
<td>subchronic</td>
<td>2500.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>LOEL</td>
<td>Hepatic</td>
<td>subchronic</td>
<td>2500.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Immune</td>
<td>immunotopic</td>
<td>5000.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Renal</td>
<td>subchronic</td>
<td>5000.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>LOEL</td>
<td>Systemic</td>
<td>subchronic</td>
<td>2500.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rat</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Hematological</td>
<td>subchronic</td>
<td>1500.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rabbit</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
<tr>
<td>+</td>
<td>NOEL</td>
<td>Systemic</td>
<td>subchronic</td>
<td>1500.0</td>
<td>mg/kg-day</td>
<td>subchronic</td>
<td>oral</td>
<td>rabbit</td>
<td>Vaille et al</td>
<td>PPRTV(…)</td>
</tr>
</tbody>
</table>
In Vitro Bioassay Screening
ToxCast and Tox21
Sources of Exposure to Chemicals
Identifiers to Support Searches

<table>
<thead>
<tr>
<th>Chemical Properties</th>
<th>Env. Fate/Transport</th>
<th>Hazard</th>
<th>ADME (Beta)</th>
<th>Exposure</th>
<th>Bioassays</th>
<th>Similar Compounds</th>
<th>Related Substances</th>
<th>Synonyms</th>
</tr>
</thead>
</table>

Found 78 synonyms

- **Legend:** Valid Synonyms, Good Synonyms, Other Synonyms
- **Copy all Synonyms**

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Active CAS RN</th>
<th>Betckeim Registry Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2-Propylene glycol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propane-1,2-diol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,2-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57-55-6</td>
<td><strong>Active CAS RN</strong></td>
<td></td>
</tr>
<tr>
<td>alpha-Propylene glycol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+/-) 1,2-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(RS)-1,2-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dl-Propylene glycol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-01-00-02142</td>
<td><strong>Betckeim Registry Number</strong></td>
<td></td>
</tr>
<tr>
<td>1,2-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+/-)-1,2-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+/-)-Propylene glycol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,2-(RS)-Propanediol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,2-DIHYDROXYPROPANE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,2-PROPANDIOL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Literature Searches and Links

Google Scholar

PubMed Abstract Sifter

PubMed Articles

PubMed Patents

PPRTV

IRIS

Exposure Bioassays Similar Compounds Related Substances Synonyms Literature Links Comments

13 of 13 articles loaded...

(Optionally, edit the query before retrieving)

("1,2-Propanediol" OR "Propanediol") AND (NOEL OR NOEL OR LC50 OR LD50 OR "reference dose" OR "reference concentration" OR "adverse effect level"[lab] OR "cancer slope factor"[lab])

1. Zulkifli; Abidin; Abidin; Amar Nordin; Pravaeza; Sye...
2. Welley; Kirkpatrick; Ocham; Jerome; Langston; Lil...
3. Poet; Ball; Hays
4. Kienhus; Soeteman-Hernandez; Bos; Cremers; Kle...
5. Qiu; Jo; Minowa; Sanada; Nejishima; Matsuishi; ...
6. Blythe; 0; Minowa; Sanada; Minowa; Caravaca; ...
7. Schneeberger; Wang; Schweiger; Messick; Valle; De...
8. Warley; McDonald; Lily; Kirkpatrick; Wallery; Byron;...
9. Carney; Potenger; Johnson; Liberack; Tomes; Dry;...
Cardiac morbidity and mortality associated with occupational exposure to 1, 2 propylene glycol dinitrate.

SA Forman, JC Helmkamp, CM Bone - Journal of occupational ... 1987 - europepmc.org

Abstract Myocardial infarction and angina pectoris are conditions long associated with occupational exposure to nitroglycerin and related explosives. Cardiac sentinel events in selected munitions workers exposed to the related nitrated ester 1, 2 propylene glycol ...

Experimenta exposure to propylene glycol mist in aviation emergency training: acute ocular and respiratory effects

G Wieslander, D Norbäck, T Lindgren - ... and Environmental Medicine, 2001 - oem.bmj.com

... to high concentrations of propylene glycol in workplaces and public places may cause ocular and respiratory irritation, and that sensitive subjects should be protected or avoid extreme or prolonged exposure. Propylene glycol (PG) (1–2 propanediol; CAS nr 57–55–6) is a ...

Evaluation of the neurophysiologic effects of 1, 2-propylene glycol dinitrate by quantitative ataxia and oculomotor function tests

EP Horvath, RA Ikla, J Boyd - ... - American journal of ... , 1981 - Wiley Online Library

... 1,2-Propylene glycol dinitrate (PGDN), a nitrated ester found in the torpedo propellant Otto Fuel 11, has been suspected of causing neurologic and ... of chronic neurotoxicity was found, even among a subgroup of workers (CE-UB) with the longest total duration of exposure ...

Propylene Glycol Dinitrate

C on Emergency, CEGL for Selected... - 2009 - ncbi.nlm.nih.gov

# External Links to Data and Services

The table below lists various external links related to data and services. The table is organized by categories such as General, Toxicology, Publications, Analytical, and Prediction.

### General
- EPA Substance Registry
- Household Products Data
- PubChem
- CPCat
- DrugBank
- Wikipedia
- MSDS Lookup
- ChEMBL
- Chemical Vendors
- NIOSH Chemical Safety
- ToxPlanet
- ACS Reagent Chemicals
- Wikidata
- ChemHat: Hazards and A...
- Wolfram Alpha

### Toxicology
- ACToR
- DrugPortal
- CCRIS
- ChemView
- CTD
- eChemPortal
- Gene-Tox
- HSDB
- ToxCast Dashboard 2
- LactMed
- International Toxicity Esti...
- ATSDR Toxic Substances...
- ACToR PDF Report
- CREST

### Publications
- Toxline
- Environmental Health Per...
- NIEHS
- National Toxicology Progr...
- Google Books
- Google Scholar
- Google Patents
- PPRTVWEB
- PubMed
- IRIS Assessments
- EPA HERO
- RSC Publications
- BioCaddie DataMed
- Springer Materials
- Federal Register

### Analytical
- RSC Analytical Abstracts
- Tox21 Analytical Data
- MONA: MassBank North ...
- NIST NIST IR Spectrum
- NIST NIST MS Spectrum

### Prediction
- 2D NMR HSQC/HCMB Predict
- Carbon-13 NMR Prediction
- Proton NMR Prediction
- ChemRTP Predictor
- LSERD
MassBank of America (MoNA) is a metadata-centric, auto-curating repository designed for efficient storage and querying of mass spectral records. It intends to serve as a framework for a centralized, collaborative database of metabolite mass spectra, metadata and associated compounds. MoNA currently contains over 200,000 mass spectral records from experimental and in-silico libraries as well as from user contributions.
Dashboard for Structure ID

• Structure Identification using the dashboard
  – Formula/mass-based searching – 1 chemical at a time
Advanced Searches

**Advanced Search**

**Mass Search**

- **Mass**
- **Min/Max**
- **Error**
- **Da ppm**

**Molecular Formula Search**

- **Molecular Formula**
- **MS Ready Formula**
- **Exact Formula**

**Generate Molecular Formula(e)**

- **Mass**
- **Error**
- **Da ppm**

**Default Options:**
- C[1-50]
- H[0-100]
- O[0-20]
- N[0-20]
- P[0-20]
- S[0-10]

**Include Halogens:**
- F[0-20]
- Cl[0-20]
- Br[0-20]
- I[0-20]
Advanced Searches
Mass Based Search

Mass Search

191.131 Da ± 5 Da ppm

Search
Advanced Searches

Chemistry Dashboard

Search Results
Searched by Mass: '191.1381 +/- 5 ppm'.
298 of 298 chemicals visible

DEET
134-62-3

Phendimetrazine
834-03-7

N-Butylisocyanatide
91-40-8

Benzyldihydridodecylamine-4-deuterylamine-4... 92-14-8

Acetanilide, 2,6-dimethyl-18665-89-7

Acetanilide, 1,3-dimethyl-3-(m-met... 10032-26-9

Benzamide, N-propyl- 20300-43-4

p-Butylacetanilide
20330-45-4

N,N-Dimethylphenylacetamide
2401-96-1

3-(Dimethylaminio)-2-methylpropylamine-122617-50-6

Butylxymide, 2-ethyl-2-phenyl- 30560-39-0

1-Heptane, 1-(4-pyridyl)- 32941-30-3

18
Formula Searches

Molecular Formula Search

C12H17NO

- MS Ready Formula
- Exact Formula

Search
Dashboard for Structure ID

• Structure Identification using the dashboard
  – Formula/mass-based searching – 1 chemical at a time
  – Distilling structures into “MS-Ready form”
Specific Data-Mappings
“MS-Ready Structures”

A) Molecular Ion

\[
\begin{align*}
&\text{m/z} = 256.1702 \\
&\text{monoisotopic mass} = 255.1623 \\
&\text{C}_{17}\text{H}_{21}\text{NO} \\
&\text{DTXCID802949}
\end{align*}
\]

B) MS-Ready Form

\[
\begin{align*}
\text{Diphenhydramine} \\
\text{C}_{17}\text{H}_{21}\text{NO} | 255.1623 \\
\text{DTXSID4022949}
\end{align*}
\]

\[
\begin{align*}
\text{Diphenhydramine citrate} \\
\text{C}_{23}\text{H}_{29}\text{NO}_3 | 447.1893 \\
\text{DTXSID80237211}
\end{align*}
\]

\[
\begin{align*}
\text{Diphenhydramine hydrochloride} \\
\text{C}_{17}\text{H}_{21}\text{CINO} | 291.1390 \\
\text{DTXSID4020537}
\end{align*}
\]

\[
\begin{align*}
\text{Diphenhydramine salicylate} \\
\text{C}_{24}\text{H}_{27}\text{NO}_4 | 393.1940 \\
\text{DTXSID10225883}
\end{align*}
\]
Diphenhydramine
15 Total MS-Ready Mappings
“MS Ready” Formula Search C12H17NO
354 Chemicals
Dashboard for Structure ID

- Structure Identification using the dashboard
  - Formula/mass-based searching – 1 chemical at a time
  - Distilling structures into “MS-Ready form”
  - Ranking based on metadata
## Table 1

Searching ChemSpider by Elemental Composition then Sorting by Number of Associated References

<table>
<thead>
<tr>
<th>Class of compounds</th>
<th>Number compounds in class</th>
<th>Position of compound sorted in descending order by number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drugs</td>
<td>45</td>
<td>#3 1 1</td>
</tr>
<tr>
<td>Pesticides</td>
<td>8</td>
<td>#4 1</td>
</tr>
<tr>
<td>Toxins</td>
<td>2</td>
<td>#1 2</td>
</tr>
<tr>
<td>Polymer antioxidants</td>
<td>15</td>
<td>#5</td>
</tr>
<tr>
<td>Polymer UV stabilizers</td>
<td>10</td>
<td>#2 1 1</td>
</tr>
</tbody>
</table>
### Identifying known unknowns using the US EPA’s CompTox Chemistry Dashboard

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

<table>
<thead>
<tr>
<th></th>
<th>Mass-based searching</th>
<th></th>
<th>Formula-based searching</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dashboard</td>
<td>ChemSpider</td>
<td>Dashboard</td>
<td>ChemSpider</td>
</tr>
<tr>
<td>Average rank position</td>
<td>1.3</td>
<td>2.2a</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Percent in #1 position</td>
<td>85%</td>
<td>70%</td>
<td>88%</td>
<td>80%</td>
</tr>
</tbody>
</table>

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
Additional Metadata Ranking

- US EPA CompTox Chemistry Dashboard Data Sources
- “CPDat” Consumer Product Database
- PubChem Data Source Count
- PubMed Reference Count
Additional Metadata Ranking
C12H17NO: 354 Chemicals

<table>
<thead>
<tr>
<th>Structure</th>
<th>DTX SID</th>
<th>Preferred Name</th>
<th>CAS RN</th>
<th>QC Level</th>
<th>CPDat Count</th>
<th>Number of Sources</th>
<th>PubChem Data Source</th>
<th>PubMed Data Source</th>
<th>Monoisotopic Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DTXSID2021995</td>
<td>DEET</td>
<td>134-60-3</td>
<td>Level 1</td>
<td>111</td>
<td>104</td>
<td>155</td>
<td>753</td>
<td>191.131014</td>
</tr>
<tr>
<td></td>
<td>DTXSID1023447</td>
<td>Phenothiazine</td>
<td>634-00-7</td>
<td>Level 2</td>
<td>12</td>
<td>27</td>
<td>35</td>
<td>50</td>
<td>191.131014</td>
</tr>
<tr>
<td></td>
<td>DTXSID2042197</td>
<td>N-Butylacrylamide</td>
<td>91-49-6</td>
<td>Level 2</td>
<td>1</td>
<td>28</td>
<td>50</td>
<td>1</td>
<td>191.131014</td>
</tr>
<tr>
<td></td>
<td>DTXSID0179048</td>
<td>N,N-Diethylphenylacetamide</td>
<td>2431-06-1</td>
<td>Level 4</td>
<td>0</td>
<td>18</td>
<td>52</td>
<td>34</td>
<td>191.131014</td>
</tr>
<tr>
<td></td>
<td>DTXSID0965298</td>
<td>N,N-Diethyl-2-methylbenzamide</td>
<td>2728-04-3</td>
<td>Level 1</td>
<td>0</td>
<td>11</td>
<td>48</td>
<td>0</td>
<td>191.131014</td>
</tr>
</tbody>
</table>
### Additional Metadata Ranking

**C12H17NO: 354 Chemicals**

<table>
<thead>
<tr>
<th>Structure</th>
<th>DTXSID</th>
<th>Preferred Name</th>
<th>CPDat Count</th>
<th>Number of Sources</th>
<th>PubChem Data Source</th>
<th>PubMed Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DTXSID2021995</td>
<td>DEET</td>
<td>111</td>
<td>104</td>
<td>155</td>
<td>753</td>
</tr>
<tr>
<td></td>
<td>DTXSID1023447</td>
<td>Phendimetrazine</td>
<td>12</td>
<td>27</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>DTXSID2042197</td>
<td>N-Butylacetanilide</td>
<td>1</td>
<td>26</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DTXSID00179048</td>
<td>N,N-Diethylphenylacetamide</td>
<td>0</td>
<td>18</td>
<td>52</td>
<td>34</td>
</tr>
</tbody>
</table>
Top Ranked Chemical

DEET
134-62-3 | DTXSID2021995

Wikipedia
N,N-Diethyl-meta-toluamide, also called DEET () or diethyltoluamide, is the most common active ingredient in insect repellents. It is a slightly yellow oil intended to be applied to the skin or to clothing and provides protection against mosquitoes, ticks, fleas, chiggers, liceis, and many biting insects. Read more

Executive Summary

https://comptos.epa.gov/dashboard/downloads
Additional data streams in development

- US EPA CompTox Chemistry Dashboard Data Sources
- “CPDat” Consumer Product Database
- PubChem Data Source Count
- PubMed Reference Count
- Retention Time Prediction \[ SC_{TOTAL} = SC_{DS} + SC_{PM} + SC_{RT} + SC_{MO} + \cdots \]
- Predicted Environmental Media Occurrence
- Presence in Lists
### Select List

<table>
<thead>
<tr>
<th>List Name</th>
<th>Number of Chemicals</th>
<th>List Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drinking Water Suspects, KWR Water, Netherlands</td>
<td>136</td>
<td>KWRSJERPS is a list of prioritized suspects relevant for human health in drinking water from KWR Water in Nieuwegein, The Netherlands. The methods are detailed in Sjerps et al 2016, DOI: 10.1016/j.watres.2016.02.034</td>
</tr>
<tr>
<td>EPA Consumer Products Suspect Screening Results</td>
<td>1705</td>
<td>This is a compiled list of the suspects reported in the supporting information of Phillips et al 2018, DOI: 10.1021/acs.est.7b04781 - Suspect Screening Analysis of Chemicals in Consumer Products with GCxGC-TOF/MS.</td>
</tr>
<tr>
<td>EPA Integrated Risk Information System (IRIS)</td>
<td>510</td>
<td>EPA's IRIS Program identifies and characterizes the health hazards of chemicals found in the environment. Each IRIS assessment can cover a chemical, a group of related chemicals, or a complex mixture.</td>
</tr>
<tr>
<td>EPAHFR - EPA Chemicals associated with hydraulic fracturing</td>
<td>1640</td>
<td>EPAHFR lists chemicals associated with hydraulic fracturing from 2005-2013, as reported in EPA's Hydraulic Fracturing Drinking Water Assessment Final Report (Dec 2016)</td>
</tr>
<tr>
<td>STOFF-IDENT Database of Water-Relevant Substances</td>
<td>8885</td>
<td>STOFF-IDENT is a database of water relevant substances collated from various sources within the STOFF-IDENT and FOR-IDENT projects, hosted by IFU, HSWT and TUM. The database at <a href="https://www.itu.bayern.de/stoffident/#home">https://www.itu.bayern.de/stoffident/#home</a> has additional functionality.</td>
</tr>
<tr>
<td>Superfund Chemical Data Matrix</td>
<td>220</td>
<td>The Superfund Chemical Data Matrix (SCDM) generates a list of the corresponding Hazard Ranking System (HRS) factor values, benchmarks, and data elements for a particular chemical.</td>
</tr>
<tr>
<td>Surfactant List Screened in Swiss Water (2014)</td>
<td>122</td>
<td>EAWAGSURF is a list of surfactants screened in Swiss wastewater effluents as part of a 2014 study. Structures/mixtures are being progressively curated and linked (Schymanski/Williams). Further details in Schymanski et al 2014, DOI: 10.1021/es4044374</td>
</tr>
</tbody>
</table>
Dashboard for Structure ID

• Structure Identification using the dashboard
  – Formula/mass-based searching – 1 chemical at a time
  – Distilling structures into “MS-Ready form”
  – Ranking based on metadata
  – Batch searching of formulae and masses
Batch Search

Step One: Select Input

Select Input Type(s)
- Chemical Name
- CASRN
- InChIKey
- Skeleton
- DSSTox Substance ID
- MS-Ready Formula
- Exact Formula(e)
- Monoisotopic Mass

This search is based on what we refer to as 'Mass Spec Ready Formulae'. All chemicals within the database are treated in a manner that all are desalted and stereochemistry is removed as Mass Spectrometry detects the major components of a salt and is insensitive to stereochemistry. As an example, a search for the formula associated with phenol will return phenol, sodium phenolate and calcium phenoxide.
Batch Search

Select Input Type(s)
- Chemical Name
- CASRN
- InChIKey
- DSSTox Substance ID
- MS-Ready Formula(s)
- Exact Formula(s)
- Monoisotopic Mass

Enter Identifiers to Search (searches should be limited to <1000 identifiers)
- C6H12O3
- C7H7N3
- C8H11NO
- C7H5NO5
- C9H15NO
- C11H12O
- C9H8O3
- C6H12O5
- C9H15NO2

Metadata
- Curation Level Details
- Data Sources
- Assay Hit Count
- Include links to ACToR reports - SLOWI (BETA)
- NHANES/Predicted Exposure
- Include ToxVal Data Availability
- Number of PubMed Articles
- Abstract Sifter Input File (Beta)
- MetFrag Input File (Beta)
- IRIS
- PPRTV
- PubChem Data Sources
- ToxPrint fingerprints

- NIOSH IDLH Values
- NIOSH International Chemical Safety Cards
- NIOSH Pocket Guide to Chemical Hazards
- NIOSH Skin Notation Profiles
- NORMAN Collaborative Trial 2015 Targets and Suspects
- Norman Network PFAS (KEMI Report)
- NORMAN Network Priority List
- NormaNEWS: Norman Early Warning System
- PFAS list provided by X.Trier et al
- Pharmaceutical List with EU, Swiss and US Consumption Data
- Provisional Peer Reviewed Toxicity Values
- Stockholm Convention on Organic Pollutants
- STOFF-IDENT Database of Water-Relevant Substances
- Superfund Chemical Data Matrix
- Surfactant List Screened in Swiss Waste (2014)
<table>
<thead>
<tr>
<th>INPUT</th>
<th>FOUND_BY</th>
<th>DTXCID</th>
<th>IN</th>
<th>DATA_SQL</th>
<th>TOXVAL</th>
<th>TOXCAST</th>
<th>TOXCAST</th>
<th>NUMBER</th>
<th>PUBCHEM</th>
<th>STO</th>
</tr>
</thead>
<tbody>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID701</td>
<td>Y</td>
<td>0.36</td>
<td>2/562</td>
<td>24</td>
<td>83</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID003</td>
<td>Y</td>
<td>0.36</td>
<td>1/276</td>
<td>376</td>
<td>80</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID061</td>
<td>Y</td>
<td>4.42</td>
<td>5/113</td>
<td>6</td>
<td>77</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID105</td>
<td>Y</td>
<td>0.0</td>
<td>0/163</td>
<td>3</td>
<td>94</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID901</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>14</td>
<td>110</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID402</td>
<td>Y</td>
<td>0.0</td>
<td>0/113</td>
<td>-</td>
<td>53</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID202</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>36</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID021</td>
<td>-</td>
<td>2.54</td>
<td>7/276</td>
<td>-</td>
<td>54</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID109</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID202</td>
<td>Y</td>
<td>0.0</td>
<td>0/113</td>
<td>-</td>
<td>47</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID303</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>89</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID302</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>25</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID407</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12</td>
<td>62</td>
<td>-</td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID704</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>C6H12O3</td>
<td>MS</td>
<td>Ready</td>
<td>DTXCID704</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>49</td>
<td>-</td>
</tr>
</tbody>
</table>
Batch Search Integration to MetFrag

http://c-ruttkies.github.io/MetFrag/projects/metfragweb/

1. Candidate database
2a. Neutral mass of the precursor molecule
3. Relative mass deviation for candidate selection
4. Neutral molecular formula of the precursor (if available)
5. Database specific candidate identifiers
2b. Calculate neutral precursor mass from charged ion mass
MetFrag Input File

- Curation Level Details
- Data Sources
- Assay Hit Count
- Include links to ACToR reports - SLOW! (BETA)
- NHANES/Predicted Exposure
- Include ToxVal Data Availability
- Number of PubMed Articles
- Abstract Sifter Input File (Beta)
- MetFrag Input File (Beta)
- IRIS
- PPRTV
- PubChem Data Sources
Batch Search Integration to MetFrag
http://c-ruttkies.github.io/MetFrag/projects/metfragweb/
Future Work: Combined Substructure/Formula Searching
Future Work: Searching Against Predicted Spectra

Spectra Prediction

Predicts the spectra for a given input molecule. Spectra are computed for low (10V), medium (20V) and high (40V) collision energy levels and are represented by a list of mass intensity pairs, each corresponding to a peak in the spectra.

InChI strings need to start with "InChI=" and are not expected to have any charge - an additional H+ will be added. Maximum compound size is 200 atoms. Load an InChI example, SMILES example, or another SMILES example.

If you wish to run multiple jobs, input larger query molecules, or customize the computation parameters, you can freely download the source code here: http://sourceforge.net/projects/CFM-ID.
Future Work: Searching Against Predicted Spectra

- CFM-ID predicted spectra generated for 700,000 chemicals
  - Positive ion, Negative ion, Electron Impact
  - Three energies
Future Work

Scoring scheme into results

\[
SC_{TOTAL} = SC_{DS} + SC_{PM} + SC_{RT} + SC_{MO} + \ldots
\]
Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- High quality curated data and rich metadata facilitates mass spec analysis
- “MS-Ready” processed data enables structure identification
Acknowledgments

• The CompTox Chemistry Dashboard team
• NERL colleagues:
  – Jon Sobus, Elin Ulrich, Mark Strynar, Seth Newton (NTA Analysis)
  – Katherine Phillips, Kathie Dionisio, Kristin Isaacs (Consumer Products Database)
• Emma Schymanski – Luxembourg Center for Systems Biomedicine (MS-ready/NTA)
Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821