

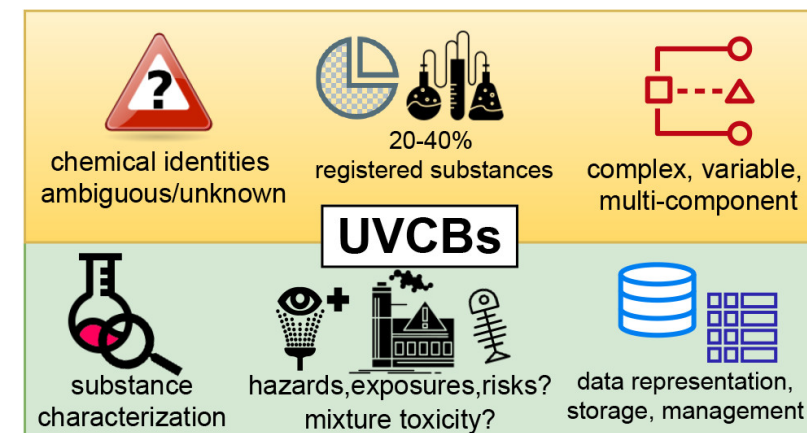
Integrating UVCBs and Related Data into Open Chemical Knowledgebases

Emma L. Schymanski, Anjana Elapavalore
Environmental Cheminformatics, Luxembourg Centre for
Systems Biomedicine, University of Luxembourg
Contact: emma.schymanski@uni.lu Twitter: [@ESchymanski](https://twitter.com/ESchymanski)

Evan E. Bolton, Qingliang Li, Paul A. Thiessen,
Leonid Zaslavsky, Jian Zhang
National Center for Biotechnology Information, National
Library of Medicine, National Institutes of Health

SETAC EUROPE 33RD ANNUAL MEETING

30 APRIL - 4 MAY 2023 | DUBLIN, IRELAND + ONLINE



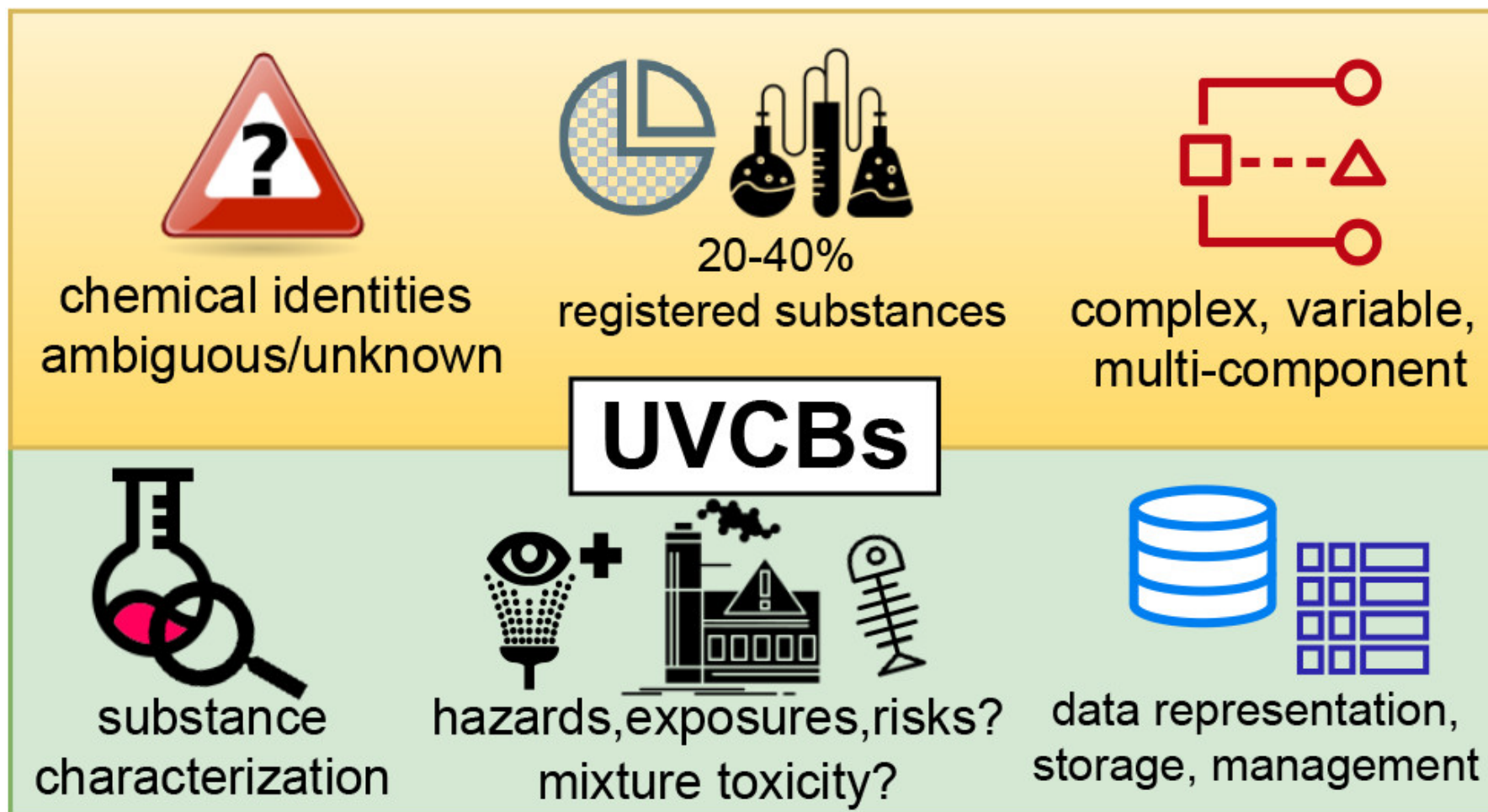
Lai et al. (2022) ES&T. DOI:
[10.1021/acs.est.2c00321](https://doi.org/10.1021/acs.est.2c00321)

Slides available at DOI:
[10.5281/zenodo.7883658](https://doi.org/10.5281/zenodo.7883658)



What are UVCBs?

Unknown or variable composition, complex reaction products, or biological materials



Why are UVCBs Challenging?

COMPOUND SUMMARY

Polyethylene Glycol 300

See also:  [Ethylene Oxide](#) (has monomer).

PubChem CID

Not available because this is not a discrete structure.

Polyethylene glycol 300 (PEG 300) is a water-miscible polyether with an average molecular weight of 300 g/mol. It is a clear viscous liquid at room temperature with non-volatile, stable properties.

Polyethylene glycols are widely used in biochemistry, structural biology, and medicine in addition to pharmaceutical and chemical industries. They serve as solubilizers, excipients, lubricants, and chemical reagents. Low molecular weight glycols are observed to exhibit antibacterial activity as well. PEG 300 is found in eye drops as a lubricant to temporarily relieve redness, but can cause irritation of the eyes.

► [DrugBank](#)

PubChem Ethylene Oxide (Compound)

5.4 Other Relationships



Benzonate (monomer of)



Polidocanol (monomer of)



Polysorbate 80 (monomer of)



Nonoxynol-9 (monomer of)



Polysorbate 20 (monomer of)

Polyethylene Glycol 3350 (monomer of)



Pentaethylene glycol monododecyl ether (monomer of)

Polyethylene Glycol 400 (monomer of)

Poloxalene (monomer of)



3,6,9,12,15,18,21,24,27,30,33,36-Dodecaoxaoctatetracontan-1-ol (monomer of)

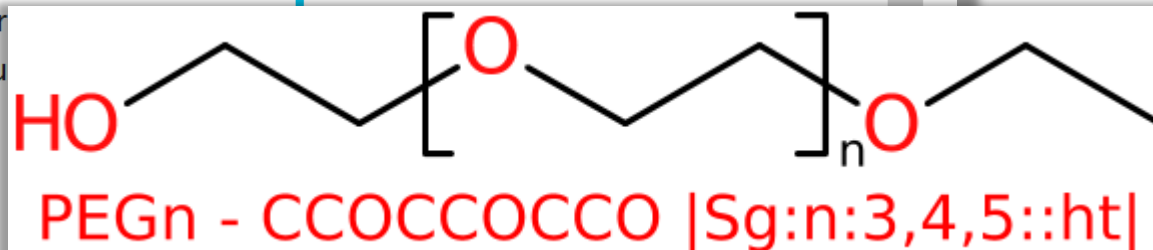
1 Synonyms

2 Names and Identifiers

3 Chemical and Physical Properties

4 Related Records

5 Drug and Medication Information



Expert Knowledge: NORMAN Suspect List Exchange

<https://www.norman-network.com/nds/SLE/>



NORMAN Database System



NORMAN Suspect List Exchange

The NORMAN Suspect List Exchange (NORMAN-SLE) was established to facilitate the exchange of information on suspected substances. The NORMAN-SLE documents all individual collections of suspected substances (see Source column in SusDat). NORMAN-SLE versions are listed in the table below. Comments and contributions are welcome - please email us at nds@norman-network.com. Please refer to our [documentation](#) pages for: [citation instructions](#).

No.	Abbreviation	Description	Link
S0	SUSDAT	Merged NORMAN Suspect List: SusDat	Introduction

Antibiotic Resistance Bacteria/Genes

A database of ARBs/ARGs in environmental matrices

Mohammed Taha *et al.* (2022) DOI: [10.1186/s12302-022-00680-6](https://doi.org/10.1186/s12302-022-00680-6)

RESEARCH

Open Access



The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry

Hiba Mohammed Taha¹, Reza Aalizadeh², Nikiforos Alygizakis^{3,2}, Jean-Philippe Antignac⁴, Hans Peter H. Arp^{5,6}, Richard Bade⁷, Nancy Baker⁸, Lidia Belova⁹, Lubertus Bijlsma¹⁰, Evan E. Bolton¹¹, Werner Brack^{12,13}, Alberto Celma^{10,14}, Wen-Ling Chen¹⁵, Tiejun Cheng¹¹, Parviel Chirsir¹, Ľuboš Čirka^{16,3}, Lisa A. D'Agostino¹⁷, Yannick Djoumbou Feunang¹⁸, Valeria Dulio¹⁹, Stellan Fischer²⁰, Pablo Gago-Ferrero²¹, Aikaterini Galani², Birgit Geueke²², Natalia Głowacka³, Juliane Glüge²³, Ksenia Groh²⁴, Sylvia Grosse²⁵, Peter Haglund²⁶, Pertti J. Hakkinen¹¹, Sarah E. Hale⁵, Felix Hernandez¹⁰, Elisabeth M.-L. Janssen²⁴, Tim Jonkers²⁷, Karin Kiefer²⁴, Michal Kirchner²⁸, Jan Koschorreck²⁹, Martin Krauss¹², Jessy Krier¹, Marja H. Lamoree²⁷, Marion Letzel³⁰, Thomas Letzel³¹, Qingliang Li¹¹, James Little³², Yanna Liu³³, David M. Lunderberg^{34,35}, Jonathan W. Martin¹⁷, Andrew D. McEachran³⁶, John A. McLean³⁷, Christiane Meier²⁹, Jeroen Meijer³⁸, Frank Menger¹⁴, Carla Merino^{39,40}, Jane Muncke²², Matthias Muschket¹², Michael Neumann²⁹, Vanessa Neveu⁴¹, Kelsey Ng^{34,2}, Herbert Oberacher⁴³, Jake O'Brien⁷, Peter Oswald³, Martina Oswaldova³, Jaqueline A. Picache³⁷, Cristina Postigo^{44,14}, Noelia Ramirez^{45,39}, Thorsten Reemtsma¹², Justin Renaud⁴⁶, Pawel Rostkowski⁴⁷, Heinz Rüdel⁴⁸, Reza M. Salek⁴¹, Saer Samanipour⁴⁹, Martin Scherlinger^{23,42}, Ivo Schliebner²⁹, Wolfgang Schulz⁵⁰, Tobias Schulze¹², Manfred Sengl³⁰, Benjamin A. Shoemaker¹¹, Kerry Sims⁵¹, Heinz Singer²⁴, Randolph R. Singh^{1,52}, Mark Sumarah⁴⁶, Paul A. Thiessen¹¹, Kevin V. Thomas⁷, Sonia Torres³⁹, Xenia Trier⁵³, Annemarie P. van Wezel⁵⁴, Roel C. H. Vermeulen³⁸, Jelle J. Vlaanderen³⁸, Peter C. von der Ohe²⁹, Zhanyun Wang⁵⁵, Antony J. Williams⁵⁶, Egon L. Willighagen⁵⁷, David S. Wishart⁵⁸, Jian Zhang¹¹, Nikolaos S. Thomaidis², Juliane Hollender^{23,24}, Jaroslav Slobodnik³ and Emma L. Schymanski¹

NORMAN Suspect List Exchange in PubChem



The NORMAN network enhances the exchange of information on emerging environmental substances, and encourages the validation and harmonisation of common measurement methods and monitoring tools so that the requirements of risk assessors and risk managers can be better met. The NORMAN Suspect List Exchange (NORMAN-SLE) is a central access point to find suspect lists relevant for various environmental monitoring questions, described in [DOI:10.1186/s12302-022-00680-6](https://doi.org/10.1186/s12302-022-00680-6)

Organization	NORMAN Network (c/o UniLu)
Category	Research and Development
URL	https://www.norman-network.com/nds/SLE
License Note	Data: CC-BY 4.0; Code (hosted by ECI, LCSB): Artistic-2.0
License URL	https://creativecommons.org/licenses/by/4.0/
Contact Name	Emma Schymanski
Address	6 avenue du Swing, Belvaux, Luxembourg 4367
Data Source ID	23819
Data in PubChem	117,695 Live Substances 20,670 Annotations 1 Classification
Last Updated	2023/04/12

NORMAN Suspect List Exchange Classification ? ↗ 114,200

- ▶ S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) ? 3,778
- ▶ S25 | OECDPFAS | List of PFAS from the OECD POPs Inventory
- ▶ S36 | UBAPMT | Potential Persistent, Mobile and Transformable
- ▶ S47 | ECHAPLASTICS | A list from the ECHA POPs Inventory
- ▶ S50 | CCSCOMPEND | The Unified Collapsing and Consolidation of Chemicals
- ▶ S60 | SWISSPEST19 | Swiss Pesticides and Plant Protection Products
- ▶ S61 | UJICCSLIB | Collision Cross Section Library
- ▶ S66 | EAWAGTPS | Parent-Transformation Products
- ▶ S68 | HSDBTPS | Transformation Products
- ▶ S69 | LUXPEST | Pesticide Screening List
- ▶ S72 | NTUPHTW | Pharmaceutically Active Substances
- ▶ S75 | CyanoMetDB | Comprehensive data on cyanometabolites
- ▶ S77 | FCCDB | Food Contact Chemicals Database
- ▶ S79 | UACCSCEC | Collision Cross Section Library
- ▶ S80 | PFASGLUEGE | Overview of PFAS Uses ? 1,251
- S00 | SUSDAT | Merged NORMAN Suspect List: SusDat ? 98,145
- S01 | MASSBANK | NORMAN Compounds in MassBank EU ? 7,123
- S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances ? 11,239
- S03 | NORMANCT15 | NORMAN Collaborative Trial Targets and Suspects ? 625
- S04 | UJIBADE | Target List from UJI used in Bade et al 2015 ? 541

S80 | PFASGLUEGE | Overview of PFAS Uses ? 1,251

- ▶ PFAS Polymer Type ? 489
 - Non-polymer ? 487
 - Polymer ? 1
 - Unclear ? 1
- ▶ PFAS Structural Category ? 1,231
 - ▶ Cyclic PFAS ? 32
 - ▶ Fluoropolymers ? 17
 - ▶ Fluoropolymers - copolymers ? 8
 - ▶ Fluoropolymers - monoconstituents ? 5
 - ▶ Fluoropolymers - terpolymers ? 4

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>

What is PubChem? <https://pubchem.ncbi.nlm.nih.gov/>



Explore Chemistry

Quickly find chemical information from authoritative sources

UVCB

Try covid-19 aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)4/h1-2H3



Use Entrez



Compounds



Substances



BioAssays



Draw Structure



Upload ID List



Browse Data



Periodic Table

115M Compounds

304M Substances

304M Bioactivities

36M Literature

43M Patents

919 Data Sources

[See More Statistics >](#)

[Explore Data Sources >](#)

Recognising UVCBs & Automated Curation

- **Substance name** is the most widely available identifier of UVCBs across all databases¹ => **text-based recognition** (aka “Regex”)

Gather
authoritative
UVCB names



~62,000 “UVCB concepts”

~500,000 synonyms (names)

Recognising UVCBs & Automated Curation

- **Substance name** is the most widely available identifier of UVCBs across all databases¹ => **text-based recognition** (aka “Regex”)

Gather
authoritative
UVCB names



Generate **Regex** to recognise UVCBs
(excerpt from “c” to “d”)

```
elseif ( $s =~ /charcol/i ) { $is_uvcb = 1; $why = "charcol"; }
elseif ( $s =~ /chlorinated/i ) { $is_uvcb = 1; $why = "chlorinated"; }
elseif ( $s =~ /complex/i ) { $is_uvcb = 1; $why = "complex"; }
elseif ( $s =~ /compounds/i ) { $is_uvcb = 1; $why = "compounds"; }
elseif ( $s =~ /compd\./i ) { $is_uvcb = 1; $why = "compd."; }
elseif ( $s =~ /compds\./i ) { $is_uvcb = 1; $why = "compds."; }
elseif ( $s =~ /concentrate/i ) { $is_uvcb = 1; $why = "concentrate"; }
elseif ( $s =~ /condensed/i ) { $is_uvcb = 1; $why = "condensed"; }
elseif ( $s =~ /condensation/i ) { $is_uvcb = 1; $why = "condensation"; }
elseif ( $s =~ /distillate/i ) { $is_uvcb = 1; $why = "distillate"; }
elseif ( $s =~ /dervs/i ) { $is_uvcb = 1; $why = "dervs"; }
elseif ( $s =~ /derivs/i ) { $is_uvcb = 1; $why = "derivs"; }
elseif ( $s =~ /derivatives/i ) { $is_uvcb = 1; $why = "derivatives"; }
elseif ( $s =~ /diluent/i ) { $is_uvcb = 1; $why = "diluent"; }
```

Recognising UVCBs & Automated Curation

- **Substance name** is the most widely available identifier of UVCBs across all databases¹ => **text-based recognition** (aka “Regex”)

Gather
authoritative
UVCB names



Generate
Regex to
recognise UVCBs

```
elseif ( $s =~ /charcol/i ) {  
elseif ( $s =~ /chlorinated/i ) {  
elseif ( $s =~ /complex/i ) {  
elseif ( $s =~ /compounds/i ) {  
elseif ( $s =~ /compd\. /i ) {  
elseif ( $s =~ /compds\. /i ) {  
elseif ( $s =~ /concentrate/i ) {  
elseif ( $s =~ /condensed/i ) {  
elseif ( $s =~ /condensation/i ) {  
elseif ( $s =~ /distillate/i ) {  
elseif ( $s =~ /dervs/i ) { $is  
elseif ( $s =~ /derivs/i ) { $i  
elseif ( $s =~ /derivatives/i ) {  
elseif ( $s =~ /diluent/i ) {
```

Validate on
various datasets



Recognising UVCBs & Automated Curation

- **Substance name** is the most widely available identifier of UVCBs across all databases¹ => **text-based recognition** (aka “Regex”)
 - 113 regular expressions in total after PubChemLite/NORMAN-SLE
 - Only 8 removed following validation on whole PubChem*

Expression	Example - detected expression in bold , plus other matches
“ tree”	Camphor tree whole (see next slides)
“chlorinated”	1-Propene, tetramer, chlorinated
C#-C# Regex	Carboxylic acids , di-, C4-11 (see next slides)
“Amines”	Amines , di- C14-18 -alkylmethyl (see next slides)

*

fruit
mace
resin
rosin
shell
solvent
steroid
tannin

Variable: Connecting Chemicals & UVCBs (e.g. “C#-C# RegEx”)

Expression	Example - detected expression in bold , plus other matches	
C#-C# RegEx	Carboxylic acids , di-, C4-11	(see next slides)
“Amines”	Amines , di- C14-18 -alkylmethyl	(see next slides)

- Homologous series are (relatively) easy to recognise and map up

OngLai: An Algorithm to Classify Homologous Series

Powered by RDKit

License Apache 2.0

Maintained? yes

issues 3 open

contributors 5

DOI 10.5281/zenodo.7035149

release v.1.1.0

pypi package ???

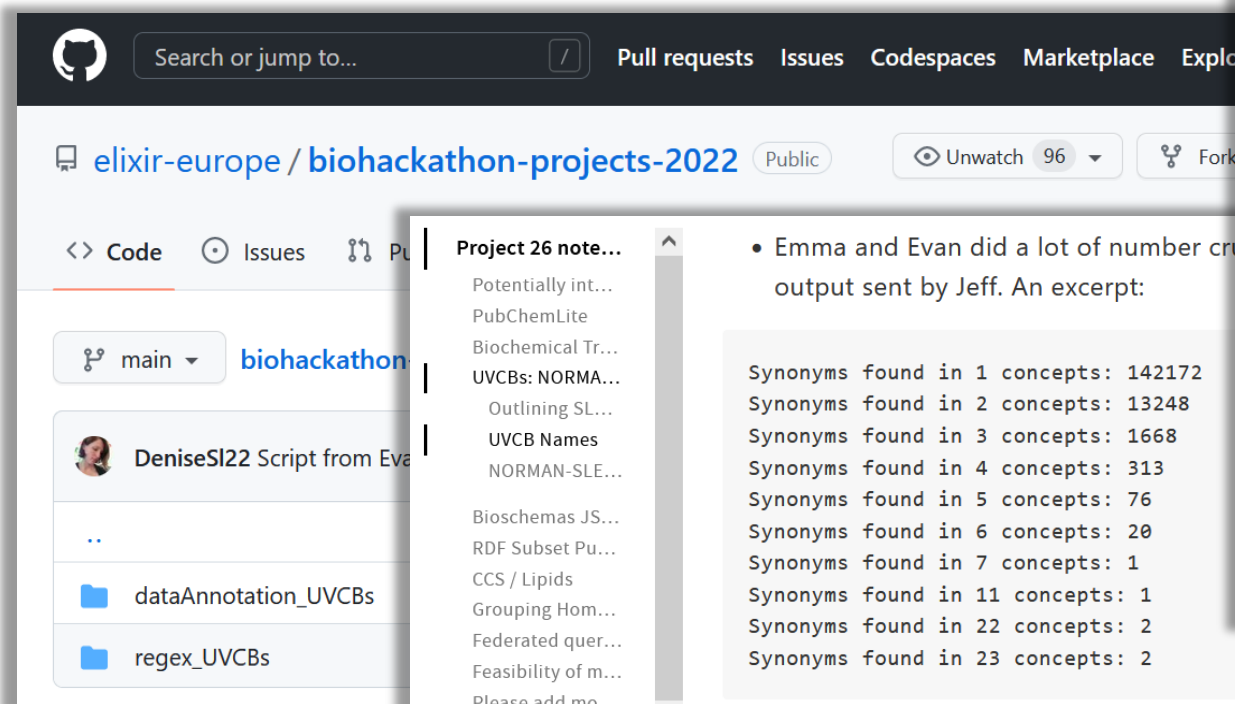


Adelene Lai
adelenelai

- Many concepts had ≥ 1 “representative” already associated with them

Variable: Connecting Chemicals & UVCBs (e.g. “C#-C# RegEx”)

- Quite a bit of (bio)hacking later ...



- Evan and Emma came up with a set of RegEx / name detections to catch “typical” UVCBs (i.e. records in PubChem that should not have CIDs associated with them)

```
if ( $s =~ /C\d+\-C\d+/ ) { $is_uvcb = 1; }  
elsif ( $s =~ /C\d+\-\d+/ ) { $is_uvcb = 1; }  
elsif ( $s =~ /C\>\d+/ ) { $is_uvcb = 1; }
```



Variable: Connecting Chemicals & UVCBs (e.g. “C#-C# RegEx”)

PubChem Carboxylic acids, di-, C4-11 (Compound)

5.2 U.S. Production

Aggregated Product Volume

2019: <1,000,000 lb

2018: <1,000,000 lb

2017: <1,000,000 lb

2016: 1,000,000 lb - <20,000,000 lb

<https://www.epa.gov/chemical-data-reporting>

► EPA Chemicals under the TSCA

5.3 General Manufacturing Information

Industry Processing Sectors

All Other Basic Organic Chemical Manufacturing

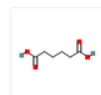
► EPA Chemicals under the TSCA

EPA TSCA Commercial Activity Status

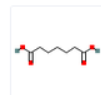
Carboxylic acids, di-, C4-11: ACTIVE

4 Related Records

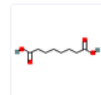
UVCB Components



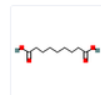
CID 196



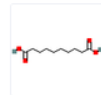
CID 385



CID 10457



CID 2266



CID 5192

Cite

Download

CONTENTS

Title and Summary

1 Synonyms

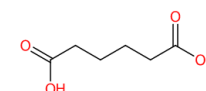
2 Names and Identifiers

3 Chemical and Physical Properties

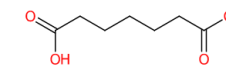
4 Related Records

5 Use and Manufacturing

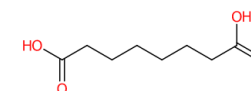
6 Information Sources



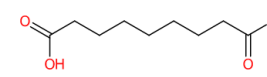
hexanedioic acid CID:196



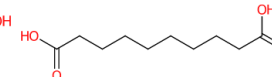
heptanedioic acid CID:385



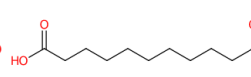
octanedioic acid CID:10457



nonanedioic acid CID:2266



decanedioic acid CID:5192



undecanedioic acid CID:15816

Variable: Connecting Chemicals & UVCBs (e.g. “C#-C# RegEx”)

PubChem Amines, di-C14-18-alkylmethyl (Compound)

5.1 Uses

5.1.1 Industry Uses

Intermediates
Viscosity adjustors

<https://www.epa.gov/chemical-data-reporting>

► EPA Chemicals under the TSCA

5.2 U.S. Production

Aggregated Product Volume

2019: 1,000,000 lb - <20,000,000 lb
2018: 1,000,000 lb - <20,000,000 lb
2017: 1,000,000 lb - <20,000,000 lb
2016: 1,000,000 lb - <20,000,000 lb

<https://www.epa.gov/chemical-data-reporting>

► EPA Chemicals under the TSCA

4 Related Records

UVCB Components

-  CID 110447
-  CID 110257
-  CID 105560
-  CID 77709

CONTENTS

- Title and Summary
- 1 Synonyms
- 2 Names and Identifiers
- 3 Chemical and Physical Properties
- 4 Related Records
- 5 Use and Manufacturing
- 6 Information Sources

Di(tetradecyl)methylamine CID: 110447 Di(pentadecyl)methylamine CID: 110257

Di(hexadecyl)methylamine CID: 105560 Di(octadecyl)methylamine CID: 77709

Biological: Associating Biological Information & UVCBs

Expression

“ tree”

Example - detected expression in **bold**, plus other matches

Camphor **tree** whole

PubChem COVID-19 Vaccine, mRNA (Compound)

PubChem CID Not available because this is not a discrete structure.

1 Synonyms

COVID-19 Vaccine, mRNA

► PubChem

2 Drug and Medication Information

2.1 FDA Purple Book

Application Number	125742
Proprietary Name	Comirnaty
Proper Name	COVID-19 Vaccine, mRNA
Status	Rx

PubChem

camphor tree whole

Substances
(37)

Proteins
(2)

Taxonomy
(1)

Pathways
(2)

Literature
(130)

Searching chemical names and synonyms in the substance records submitted by PubChem's contributors. [Read](#)

37 results

Filters

SORT BY

Relevance



Camphor; **D-CAMPHOR**; **DL-Camphor**; **Camphora**; ... **Camphora**
Tree; ...

Substance SID: 123094738 Compound CID: 159055

Data Source: ChEMBL External ID: ChEMBL504760

Data Source Category: Curation Efforts; Research and Development

Deposit Date: 2011-06-16 Last Modified Date: 2023-03-02

Structure
not available

Camphorwood; **Laurus** **Camphora**; **Camphor** **Laurel**; **Camphora**
Tree; **Camphor** **Whole**; ...

Substance SID: 472387794

Data Source: FDA Global Substance Registration System (GSRS)

External ID: 0B27814T7X

Biological: Associating Biological Information & UVCBs

Expression

Example - detected expression in **bold**, plus other **matches**

“ tree”

Camphor **tree** whole

PubChem Cinnamomum camphora (camphor tree) (Taxonomy)

3.2 Natural Products

246 items View More Rows & Details

Download

SORT BY Compound CID

Structure	Compound CID	Compound	Evidence IDs ?	Data Source
	2537	Camphor		NPASS
	2537	Camphor	DOI:10.1590/S1516-89132000000300011 DOI:10.1080/10412905.1998.9700962 DOI:10.1093/JAT/9.1.24 DOI:10.1080/10412905.1993.9698258	LOTUS
	2758	Eucalyptol	DOI:10.1007/S11418-006-0039-1	LOTUS

PubChem

camphor tree whole

Substances
(37)

Proteins
(2)

**Taxonomy
(1)**

Pathways
(2)

Literature
(130)

Searching chemical names and synonyms in the substance records submitted by PubChem's contributors. Read

37 results

Filters

SORT BY

Relevance



Camphor; **D-CAMPHOR**; **DL-Camphor**; **Camphora**; ...; **Camphora Tree**; ...

Substance SID: 123094738 Compound CID: 159055

Data Source: ChEMBL External ID: ChEMBL504760

Data Source Category: Curation Efforts; Research and Development

Deposit Date: 2011-06-16 Last Modified Date: 2023-03-02

Structure
not available

Camphorwood; **Laurus Camphora**; **Camphor Laurel**; **Camphora Tree**; **Camphor Whole**; ...

Substance SID: 472387794

Data Source: FDA Global Substance Registration System (GSRS)

External ID: 0B27814T7X

Biological: Extracting & Using Related Chemical Information

Expression	Example - detected expression in bold
"tree"	Camphor tree whole (see

PubChem Cinnamomum camphora (camphor tree) (Taxonomy)

3.2 Natural Products

246 items View More Rows & Details

Download

3.1 Metabolites

160 items View More Rows & Details

Structure Compound CID Compound



2537

Camphor



2758

Eucalyptol



931

Naphthalene



3314

Eugenol

PMID:10743224

Retrieving *Camphor tree* Chemicals from PubChem

Emma SCHYMANSKI

15/04/2023

Background

Camphor is an example of an ambiguous name that can refer to a discrete chemical entity (camphor) or more abstract entities (e.g. camphor tree, camphor oil). It is a good example of a "biological or biological product" (see Figure 1) via the embedded hyperlink. This document outlines how to retrieve further information about chemical components related to

PubChem. Associate chemicals of interest to be retrieved from the section, but note that not all of these substances are necessarily of interest. For Cinnamomum camphora, the associated *Natural Products* and *Natural Products* appear to be of most interest (see Figure 1).

Automated extraction & formatting for workflows (e.g. RMarkdown)

PubChem Cinnamomum camphora (camphor tree) (Taxonomy)

3.1 Metabolites

160 items View More Rows & Details

Download

Sort BY: Compound CID

Structure	Compound CID	Compound	Evidence IDs	Data Source
<chem>c1ccc2ccccc2c1</chem>	931	Naphthalene		KNAPSAcK
<chem>CC1=CC(=C(C=C1)C)C</chem>	2758	Eucalyptol		
<chem>CC1=CC(=C(C=C1)C)C</chem>	3314	Eugenol	PMID:10743224	

3.2 Natural Products

246 items View More Rows & Details

Download

Sort BY: Compound CID

Structure	Compound CID	Compound	Evidence IDs	Data Source
<chem>C1CCC2C(C1)CCC2</chem>	2537	Camphor		KNAPSAcK

Figure 1: Associated chemicals of interest for *Cinnamomum camphora* on PubChem

Biological: Extracting & Using Related Chemical Information

Environmental Cheminformatics > PubChem Docs > Repository



added camphor tree example ...
Emma Schymanski authored 1 minute ago

main ▾

pubchem-docs / taxonomy / Camphor_tree / Camphor_tree_metabolites.Rmd

Find file

Blame

History



Camphor_tree_metabolites.Rmd

```
title: "Retrieving _Camphor tree  
author: "Emma SCHYMANSKI"  
date: "15/04/2023"  
output: pdf_document  
urlcolor: blue
```

```
knitr::opts_chunk$set(echo = TR
```

Gathering Chemical Information

Next, get all information we need via *webchem*. The following code uses the parent CIDs (recommended). (the parent mapping), uncomment out the last line ('CIDs <- all_CIDs') below before continuing. The parent CIDs are best for mass spectral applications. The other CIDs may be charged, which may be better for some metabolites.

```
library(webchem)  
library(data.table)  
CIDs <- parent_CIDs  
#CIDs <- all_CIDs
```

Then run the following. This will return CID, Name, Monoisotopic Mass, Molecular Formula, XlogP, InChI, SMILES, Count and Literature Count for all CIDs:

```
selected_properties <- c("Title", "MonoisotopicMass", "MolecularFormula", "XlogP",  
"InChI", "IsomericSMILES", "InChIKey", "IUPACName",  
"PatentCount", "LiteratureCount")
```

```
# retrieve info with webchem  
CID_info_all <- as.data.table(webchem::pc_prop(CIDs, selected_properties))
```

```
# output  
write.csv(CID_info_all, "Camphor_tree_all_Struct_Info.csv", row.names = F)
```

Retrieving *Camphor tree* Chemicals from PubChem

Emma SCHYMANSKI

15/04/2023

Background

Camphor is an example of an ambiguous name that can refer to a discrete chemical entity (*camphor*) or more abstract entities (*Camphor tree*, *Camphor tree*, *Camphor tree*). It is a good example of a chemical entity that can be referred to by multiple names. This document outlines how to retrieve further information about chemical components via the *webchem* package. For *Cinnamomum camphora*, the associated *metabolites* and *Natural Products* appear to be of most interest (see Figure 1).

Automated extraction & formatting for workflows (e.g. RMarkdown)

PubChem Cinnamomum camphora (camphor tree) (Taxonomy)

3.1 Metabolites

160 items View More Rows & Details

Structure	Compound CID	Compound	Evidence IDs	Data Source
	931	Naphthalene		KNAPSAcK
	2758	Eucalyptol		
	3314	Eugenol	PMID:10743224	

3.2 Natural Products

248 items View More Rows & Details

Structure	Compound CID	Compound	Evidence IDs	Data Source
	2507	Camphor		HMDB

Figure 1: Associated chemicals of interest for *Cinnamomum camphora* on PubChem

Biological: Using Chemical Information in MetFrag



MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Can use as a custom-made suspect list or database

Database Settings

Database:

Upload File:

Neutral Mass: Search ppm:

Formula:

Identifiers:

Parent Ion:

Candidate Filter & Score Settings

☒ Exact Spectral Similarity (MoNA)

☐

☒ LiteratureCount

☒ PatentCount

☐ XLogP

2 of 3 item(s) selected

...with scoring terms
for environmental
applications

Biological: Using Chemical Information in MetFrag

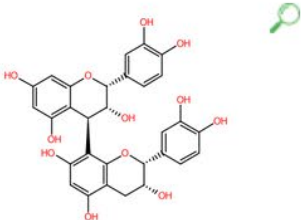
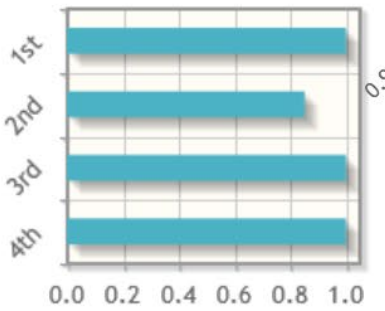
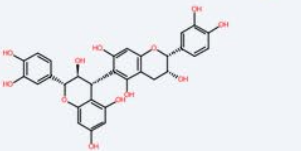


Weights		
MetFrag (1st)	<input type="range"/>	100 %
ExactSpectralSimilarity (2nd)	<input type="range"/>	100 %
LiteratureCount (3rd)	<input type="range"/>	100 %
PatentCount (4th)	<input type="range"/>	100 %

Scores View

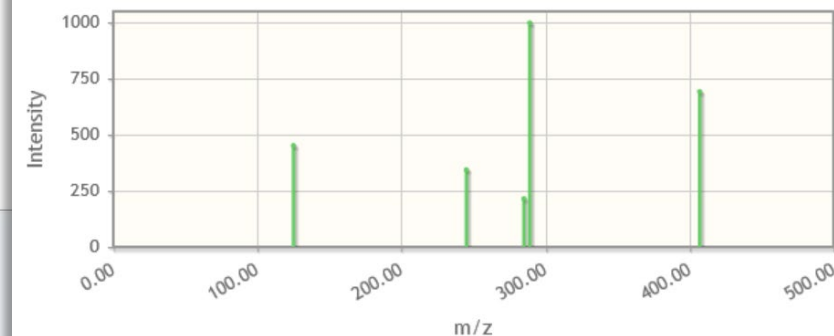
Candidate Name: Procyanidin B2
Candidate Identifier: 122738|53

	Name	Normalized Value	Raw Value
1	MetFrag	1.0	170.1718
2	ExactSpectralSimilarity	0.8513	0.8513
3	LiteratureCount	1.0	753.0
4	PatentCount	1.0	2336.0

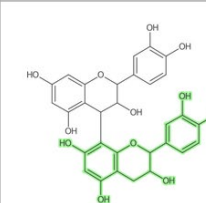
#	Molecule	Identifier	Mass	Formula	Normalized Scores
1	 Procyanidin B2	122738 11250133 InChIKeyBlock1 = XFZJEEAOWLFHDH	578.14243	C ₃₀ H ₂₆ O ₁₂	
2	 Procyanidin B8	474541 InChIKeyBlock1 = GMISZFQPFDAPI	578.14243	C ₃₀ H ₂₆ O ₁₂	

Fragments View

Select area to zoom in. Double click to return.
Click on apex of explained peak to select fragment.



Fragments



Fragment 4

Peak m/z: 289.0718
Fragment Mass: 289.0718 Da
Fragment Formula: [C₁₅H₁₃O₆]⁻

Challenges / Perspectives – Example of PCBs

- Authoritative names don't match with “our” community names

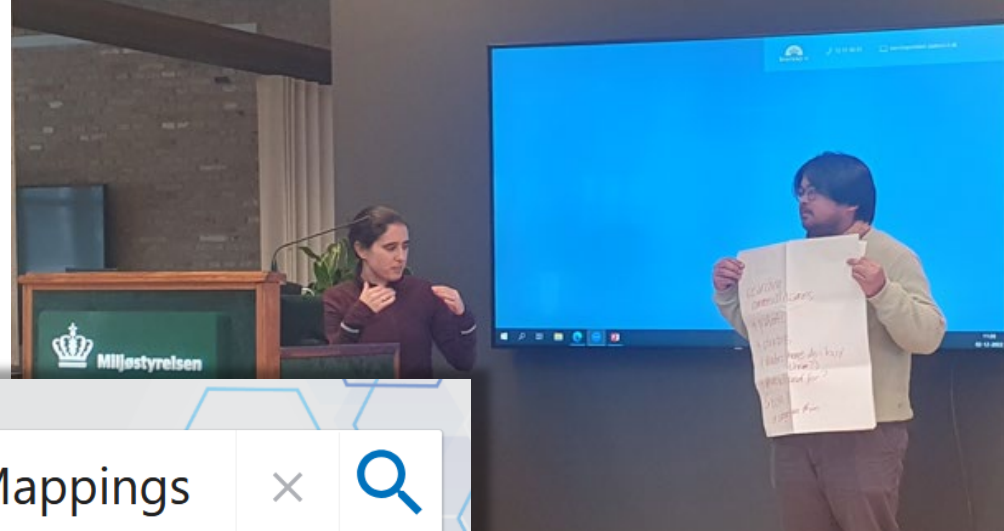
The image shows a screenshot of the PubChem website. The main search bar at the top left contains the text 'PCBs', which is highlighted with a red box. Below the search bar, there are tabs for 'Compounds (1)', 'Substances (29)', 'BioAssays (38)', and 'Literature (14,176)'. The 'Literature' tab is selected and highlighted with a red box. Below the tabs, there is a search bar with the text 'EPA DSSTox: [PCBCHEMICALS]: Polychlorinated biphenyl (PCB) collection'. Below this, there is a section for 'Compounds' with 208 results. The first result is '3,3'-DICHLOBIPHENYL; 2050-67-1; 1,1'-Biphenyl, 3,3'-dichloro-; Clophen A30; Clophen A 30; ...'. The chemical structure of 3,3'-Dichlorobiphenyl is shown. Below the structure, there is a list of identifiers: Compound CID: 16307, MF: C₁₂H₈Cl₂, MW: 223.09g/mol, IUPAC Name: 1-chloro-3-(3-chlorophenyl)benzene, Isomeric SMILES: C1=CC(=CC(=C1)Cl)C2=CC=CC=C2Cl, InChIKey: KTXUOWUHFLBZPW-UHFFFAOYSA-N, InChI: InChI=1S/C12H8Cl2/c13-11-5-1-3-9(7-11)10-4-2-6-12(14)8-10/h1-8H, Create Date: 2005-03-27. To the right of the compound details, there is a section for 'ACTIONS ON RESULTS WITH ID TYPE: Compounds' with buttons for 'Push to Entrez', 'Save for Later', and 'Linked Data Sets'. Below the compound details, there is a section for 'Substances' with buttons for 'Save for Later' and 'Linked Data Sets'. The abstract of the first literature result is visible: 'Dioxins and PCBs in feed and food--review from European perspective'. The abstract text is: 'During the 1990s, a number of adverse contamination incidents focussed the attention of the media and the general public on food safety. This led to the evaluation of safety measures with regard to dioxin intake from food. Important aspects regarding dioxins and PCBs in the food chain are reviewed here, allowing a contextual understanding of the present situation through its chronological developments. About 90-98% of the average exposure of humans to dioxins and PCBs results from dietary intake, with food of animal origin being the predominant source. Therefore, animal feed contributes considerably to the presence of these compounds in food. The detection of the 'real' source of a contamination event in the food chain is a complex scientific problem and requires specific knowledge on production processes and changes of patterns during bioaccumulation. This is demonstrated by complex investigations performed in three studies on two continents to identify the source (e.g.

<https://pubchem.ncbi.nlm.nih.gov/#query=PCBs&tab=pubmed>

[https://pubchem.ncbi.nlm.nih.gov/classification/#hid=105&search=Polychlorinated%20biphenyl%20\(PCB\)%20collection&view=tree](https://pubchem.ncbi.nlm.nih.gov/classification/#hid=105&search=Polychlorinated%20biphenyl%20(PCB)%20collection&view=tree)

Challenges / Perspectives

- Community efforts: NORMAN-UVCBs



SEARCH FOR

S103 | NORMANUVCB | NORMAN Dataset of Curated UVCB Mappings

Treating this as a previously computed list of identifiers.

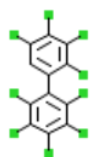
Compounds

220 results

Filters

SORT BY

Complexity



40186-72-9; 2,2',3,3',4,4',5,5',6-NONACHLOROBIPHENYL;
Nonachlorobiphenyl; NONACHLORO-1,1'-BIPHENYL; 53742-07-7

Compound CID: 38411

MF: $C_{12}HCl_9$ MW: 464.2g/mol Complexity (sort by): 351

IUPAC Name: 1,2,3,4,5-pentachloro-6-(2,3,4,5-tetrachlorophenyl)benzene

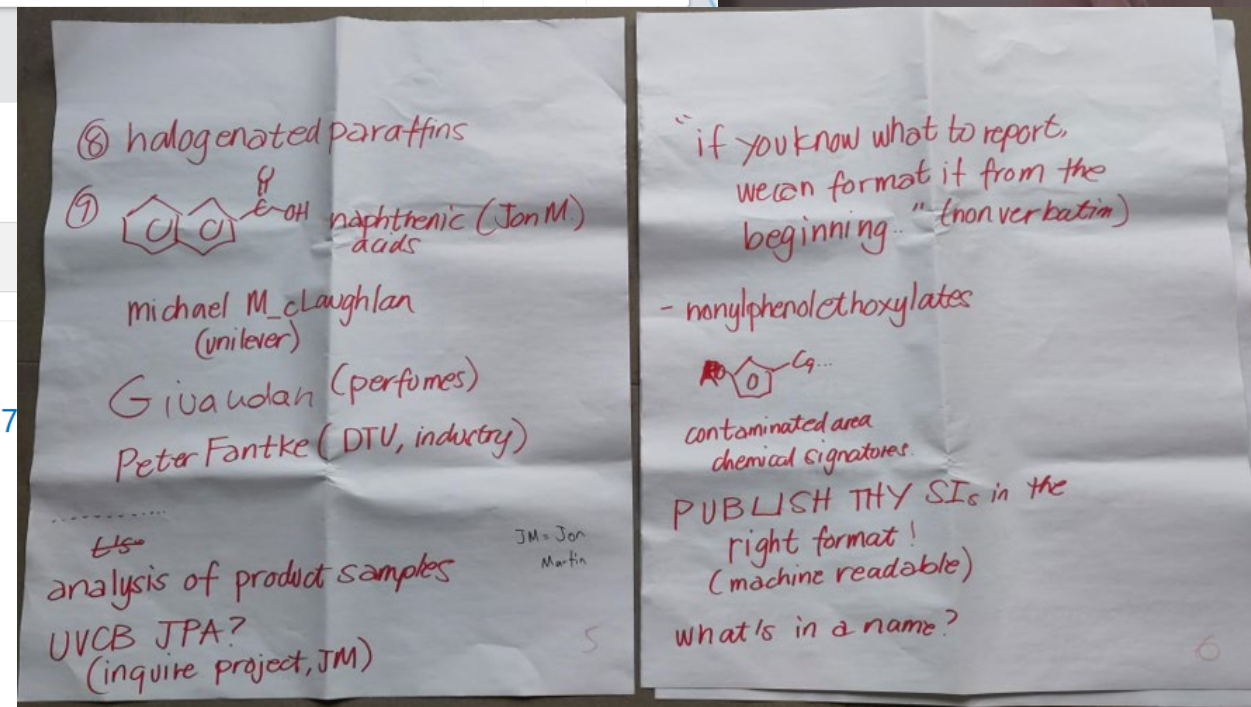
Isomeric SMILES: C1=C(C(=C(C(=C1Cl)Cl)Cl)Cl)C2=C(C(=C(C(=C2Cl)Cl)Cl)Cl)Cl

InChIKey: JFIMDKGRGPNPRQ-UHFFFAOYSA-N

InChI: InChI=1S/C12HCl9

/c13-3-1-2(5(14)9(18)6(3)15)4-7(16)10(19)12(21)11(20)8(4)17/h1H

Create Date: 2005-03-27



NORMAN-UVCBs & UVCB Template (a little complex!)



- Help add more UVCB data with **FAIR** templates & **open** license

A	B	C	D	E	F	G	H	I	J
UVCB_Name	UVCB_Synonym	CAS_RN	EC_Num	PubChem_SID	Source	Source_ID	Reference_Description	Reference_ID	Related_UVCB
PCBs	polychlorinated biphenyls				Randolph Singh (IFREMER, France), Emma Schymanski (UCSB)	ORCID:0000-0003-4500-3400 ORCID: 0000-0001-6868-8145	J Haedrich et al (2021) Environ Sci Eur 33(1):33	PMID:33828936	
DL-PCBs									
PCBs	Component	2,3,3'-Trichloro	2,3,3'-Trichloro-	38034	C1C1=CC=CC(=C)SXHLTVKPNQVZGI	38444-84	DTXSID8073502	Randolph Singh	
PCBs	Component	2,2',4,5',6-Penta	2,2',4,5',6-Penta-	63086	C1C1=CC=C(C1PQHZWWBJPCNN	60145-21	DTXSID5074189	Randolph Singh	
PCBs	Component	2,3,3',4,4',5,5'-Hexa	2,3,3',4,4',5,5'-Hexa-	38306	C1C1=CC(=CC1XUAWBXBYHDDR	39635-31	DTXSID4074144	Randolph Singh	
PCBs	Component	2,2',3,4',6'-Penta	2,2',3,4',6'-Penta-	43238	C1C1=CC(C1)=GOFFZTAPOOICF	60233-25	DTXSID9074193	Randolph Singh	
PCBs	Component	3,3',4-Trichloro	3,3',4-Trichloro-	37806	C1C1=CC=CC(=C)JHBVPKZLIBDTJR	137680-69	DTXSID60865879	Randolph Singh	
PCBs	Component	2,3',5,5'-Tetra	2,3',5,5'-Tetra-	38878	C1C1=CC(=C(C1WBTMFEPLVQOV	41464-42	DTXSID8074152	Randolph Singh	
PCBs	Component	2,2',3,4,6-Penta	2,2',3,4,6-Penta-	41362	C1C1=CC(C1)=QGDRLQRLFUJPF	55215-17	DTXSID6074178	Randolph Singh	
PCBs	Component	2,3,3',4,5'-Penta	2,3,3',4,5'-Penta-	91704	C1C1=CC(=CC1MPCDNZSLJWJDN	70362-41	DTXSID1074206	Randolph Singh	
PCBs	Component	3,3',4,5'-Tetra	3,3',4,5'-Tetra-	63104	C1C1=CC(=CC1QLCTXEMDCZGPC	41464-48	DTXSID3074155	Randolph Singh	
PCBs	Component	2,3,4,4',5-Penta	2,3,4,4',5-Penta-	53036	C1C1=CC=C(C1SXZSFWHOSHAKN	74472-37	DTXSID9074226	Randolph Singh	
PCBs	Component	3,4'-Dichloro	3,4'-Dichloro-	118101	C1C1=CC=C(C1CJDNEKOMKXLSB	2974-90-5	DTXSID10863067	Randolph Singh	
PCBs	Component	2,2',3,5,6,6'-Hexa	2,2',3,5,6,6'-Hexa-	63070	C1C1=CC(C1)=CLODVDBWNVQL	68194-09	DTXSID70867526	Randolph Singh	
PCBs	Component	2,2',3,3',4,4',5,5'-Hexa	2,2',3,3',4,4',5,5'-Hexa-	40485	C1C1=C(C1C)C(C1JAHJITLJSDRCG	152663-78	DTXSID1074171	Randolph Singh	
PCBs	Component	2,3,4',6-Tetra	2,3,4',6-Tetra-	63110	C1C1=CC=C(C1FXRXQYZALWWC	52663-58	DTXSID3074159	Randolph Singh	
PCBs	Component	2,3,3',4,4',5-Hexa	2,3,3',4,4',5-Hexa-	38019	C1C1=CC(=C(C1LCXMEXLGMKFLQ	38380-08	DTXSID0052706	Randolph Singh	
PCBs	Component	2,2',5,5'-Tetra	2,2',5,5'-Tetra-	37248	C1C1=CC(=C(C1HCWZEPKLWVAE	35693-99	DTXSID3038305	Randolph Singh	
PCBs	Component	2,4-Dichloro	2,4-Dichloro-	1136399	C1C1=CC(C1)=WEJZHJJPXXML	33284-50	DTXSID8040301	Randolph Singh	
PCBs	Component	3,4',5-Trichloro	3,4',5-Trichloro-	38038	C1C1=CC=C(C1SYSBNFJJSJLZMM	38444-88	DTXSID40865913	Randolph Singh	

Letter to the Editor | [Open Access](#) | [Published: 07 July 2021](#)

FAIR chemical structures in the Journal of Cheminformatics

[Emma L. Schymanski](#) & [Evan E. Bolton](#)

Journal of Cheminformatics **13**, Article number: 50 (2021) | [Cite this article](#)

3046 Accesses | **10** Citations | **27** Altmetric | [Metrics](#)

JOURNAL ARTICLE

FAIRifying the exposome journal: Templates for chemical structures and transformations

[Emma L Schymanski](#), [Evan E Bolton](#)

Exposome, Volume 2, Issue 1, 2022, osab006, <https://doi.org/10.1093/exposome/osab006>

Published: 24 December 2021 **Article history**

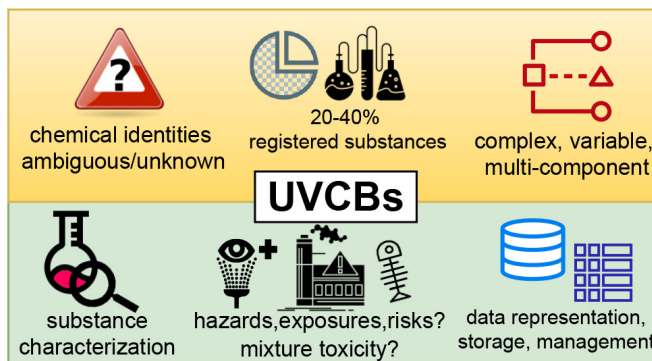
Schymanski, Bolton & Singh (2022) NORMANUVCB (S103.0.1.1) Zenodo. DOI: [10.5281/zenodo.7584082](https://doi.org/10.5281/zenodo.7584082)

Schymanski & Bolton (2022) FAIR-ifying the Exposome Journal: DOI: [10.1093/exposome/osab006](https://doi.org/10.1093/exposome/osab006)

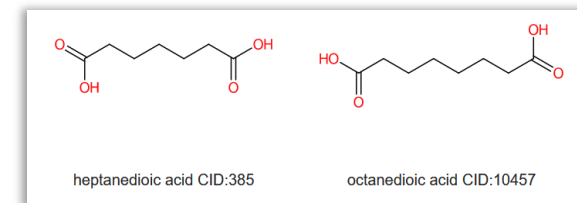
Schymanski & Bolton (2021) FAIR Chemical Structures. *J. Cheminform.* DOI: [10.1186/s13321-021-00520-4](https://doi.org/10.1186/s13321-021-00520-4)

Take home messages

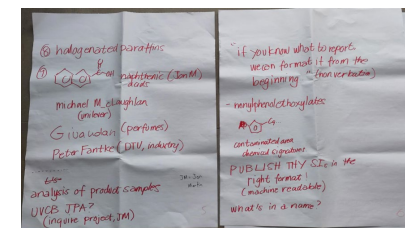
- **UVCBs** are challenging
- Text-based curation to **recognise** UVCBs
- Algorithms can support cross-linking **Variable** cases
- **Biological** cases leverage other disciplines
- Expert use cases will help develop most **relevant** features
 - Please reach out to us if you have suggestions or data to add!
Emma: emma.schymanski@uni.lu or [@ESchymanski](https://twitter.com/ESchymanski)
PubChem: pubchem-help@ncbi.nlm.nih.gov or [@pubchem](https://twitter.com/pubchem)



```
elseif ( $s =~ /charcol/i ) {  
elseif ( $s =~ /chlorinated/i  
elseif ( $s =~ /complex/i ) {  
elseif ( $s =~ /compounds/i )  
elseif ( $s =~ /compd\. /i ) {  
elseif ( $s =~ /comps\. /i )
```



PubChem Cinnamomum camphora (camphor tree) (Taxonomy)
3.2 Natural Products



Acknowledgements!

Today's slides:

DOI: [10.5281/zenodo.7883658](https://doi.org/10.5281/zenodo.7883658)

Email: emma.schymanski@uni.lu

Twitter: [@ESchymanski](https://twitter.com/ESchymanski)

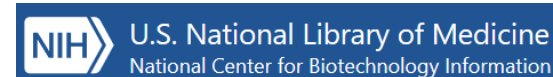


PubChem

[@EvanBolton](https://twitter.com/EvanBolton)
[@pubchem](https://twitter.com/pubchem)



Evan Bolton, Jian (Jeff) Zhang, Paul Thiessen,
Leonid Zaslavsky, Qingliang (Leon) Li,
plus Asta, Ben, Siqian + the whole team
pubchem-help@ncbi.nlm.nih.gov



Adelene Lai
(OngLai)

Anjana
Elapavalore
OngLai+PCL
Mapping



This work was
supported in part by
the National Center
for National Library
of Institutes of Health



H2020:
101036756

