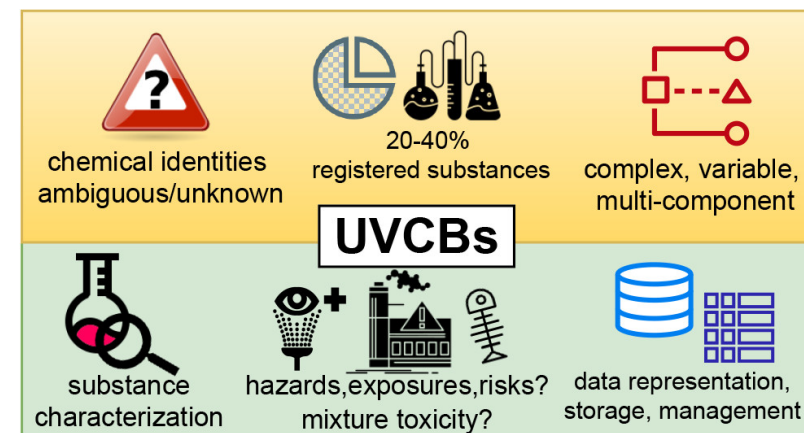


Integrating UVCBs and Related Data into Open Chemical Knowledgebases

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



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

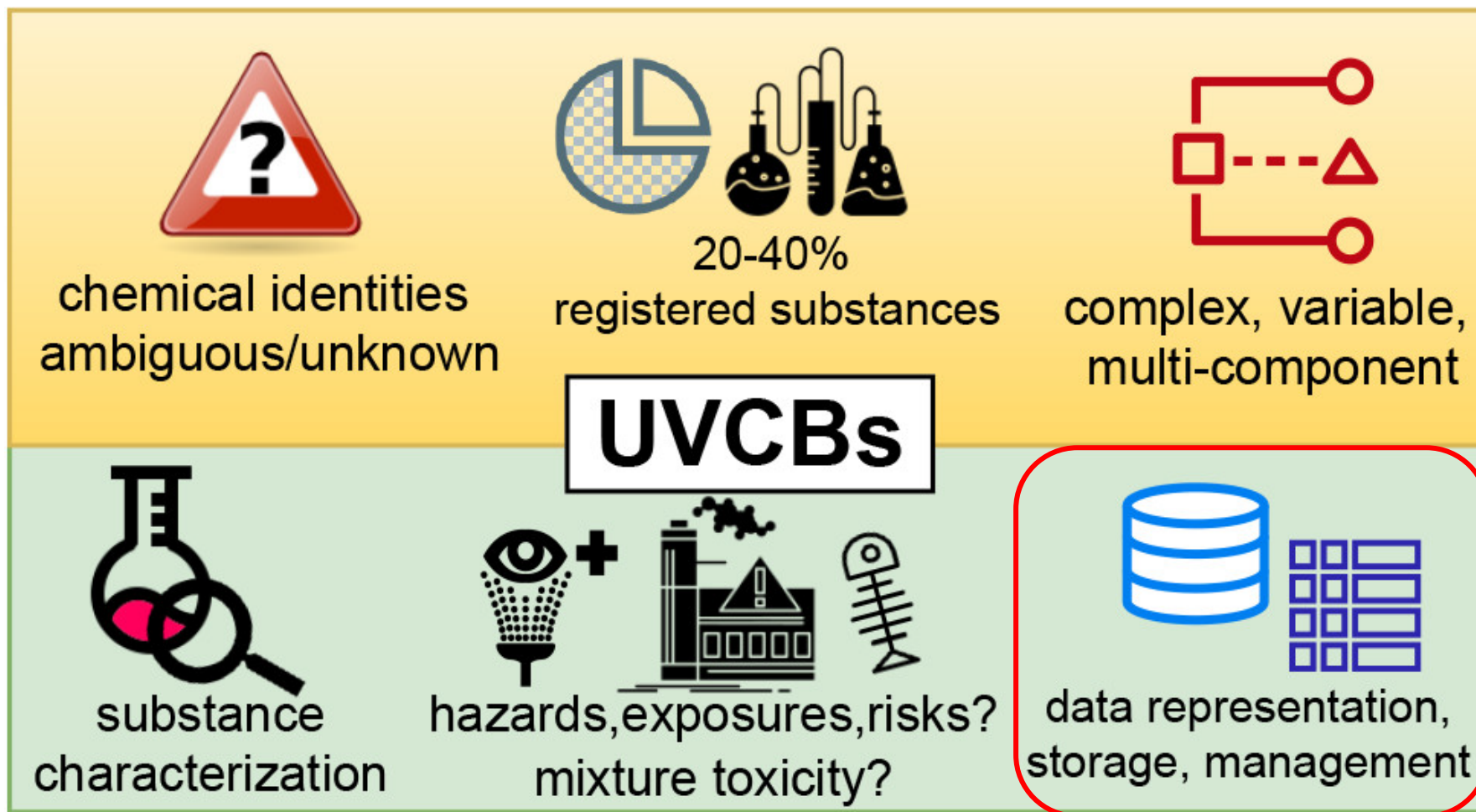
Workshop, 18-19 Sept 2023
(presented remotely)

Slides available at DOI:
[10.5281/zenodo.8302440](https://doi.org/10.5281/zenodo.8302440)



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.

Description

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 properties. PEG 300 is found in eye drops as a lubricant to temporarily relieve irritation of the eyes.

▶ DrugBank

PubChem Ethylene Oxide (Compound)

5.4 Other Relationships

PEG-4 diethylhexanoate (monomer of)	Polyethylene Glycol 600 (monomer of)	Cetyl peg/ppg-10
PEG-8 stearate (monomer of)	Polyethylene Glycol 7000 (monomer of)	Demplatin Pegrag
PEG/PPG-20/6 dimethicone (monomer of)	Polyethylene Glycol 800 (monomer of)	Disteareth-100 isc
Pegamotecan (monomer of)	Polyethylene Glycol 8000 (monomer of)	Firtecan Pegol (m
Poloxamer 188 (monomer of)	Polyethylene oxide 2000000 (monomer of)	Glycereth-12 (mo
Poloxamer 407 (monomer of)	Beheneth-10 (monomer of)	Glycereth-18 ethy
Polyethylene Glycol 1000 (monomer of)	Bis-PEG-18 methyl ether dimethyl silane (monomer of)	Glycereth-26 (mo
Polyethylene Glycol 1450 (monomer of)	Bis-peg/ppg-14/14 dimethicone (monomer of)	Glycereth-7 trime
Polyethylene Glycol 4000 (monomer of)	Ceteth-10 phosphate (monomer of)	Hydroxyethyl Cell
Polyethylene Glycol 4500 (monomer of)	Ceteth-20 (monomer of)	Hydroxyethyl cellu

▶ PubChem

2 Names and Identifiers

3 Chemical and Physical Properties

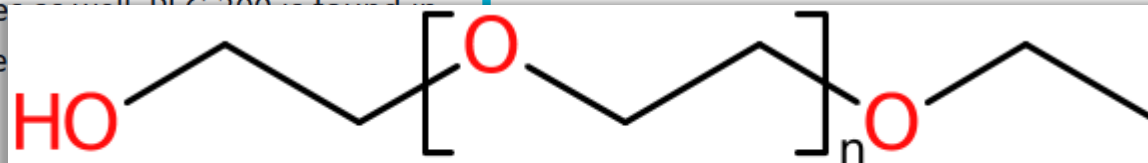
4 Related Records

5 Drug and Medication Information

6 Pharmacology and Biochemistry

7 Use and Manufacturing

8 Toxicity



PEGn - CCOC(O)CCO |Sg:n:3,4,5::ht|

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 in 2011 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 regularly updated. 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 Glowacka³, 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^{3,42}, 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 Rüdell⁴⁸, 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



SEARCH All Databases

Searching for individual substance or group(s) of substances

Note: Click on a link below to go to an individual database home page



Substance Database

A merged list of NORMAN substances; Central Database to access various lists of substances for suspect screening and prioritisation



Suspect List Exchange

Central Database to access various lists of substances for suspect screening and prioritisation



Antibiotic Resistance Bacteria/Genes

A database of ARBs/ARGs in environmental matrices

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

Organization	NORMAN Network (c/o UniLu)
Category	Research and Development
URL	https://www.norman-network.com/normansle/
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	118,487 Live Substances 22,314 Annotations 1 Classification
Last Updated	2023/08/23

- ▼ NORMAN Suspect List Exchange Classification ? ↗ 115,694
 - ▶ S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) ? 3,935
 - ▶ S25 | OECDPFAS | List of PFAS from the
 - ▶ S36 | UBAPMT | Potential Persistent, Mo
 - ▶ S47 | ECHAPLASTICS | A list from the Pl
 - ▶ S50 | CCSCOMPEND | The Unified Collis
 - ▶ S60 | SWISSPEST19 | Swiss Pesticides a
 - ▶ S61 | UJICCSLIB | Collision Cross Sectio
 - ▶ S66 | EAWAGTPS | Parent-Transformatio
 - ▶ S68 | HSDBTPS | Transformation Product
 - ▶ S69 | LUXPEST | Pesticide Screening Lis
 - ▶ S72 | NTUPHTW | Pharmaceutically Activ
 - ▶ S75 | CyanoMetDB | Comprehensive data
 - ▶ S77 | FCCDB | Food Contact Chemicals L
 - ▶ S79 | UACCSCEC | Collision Cross Sectio
 - ▶ S80 | PFASGLUEGE | Overview of PFAS Uses ? 1,251
 - S00 | SUSDAT | Merged NORMAN Suspect List: SusDat ? ↗ 98,982
 - S01 | MASSBANK | NORMAN Compounds in MassBank EU ? ↗ 7,175
 - S02 | STOFFIDENT | HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances ? ↗ 11,256
 - S03 | NORMANCT15 | NORMAN Collaborative Trial Targets and Suspects ? ↗ 625
 - S04 | UJIBADE | Target List from UJI used in Bade et al 2015 ? ↗ 542
- ▼ 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

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

116M Compounds

308M Substances

292M Bioactivities

36M Literature

933 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

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"; }
```

Validate on
various datasets



Recognising UVCBs & Automated Curation

- 113 regular expressions after PubChemLite and NORMAN-SLE validation

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)

- Only 8 removed following validation on whole PubChem*

*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 ...

The screenshot shows the GitHub interface for the repository 'elixir-europe / biohackathon-projects-2022'. The main content area displays a file tree for 'Project 26 note...'. The tree includes folders like 'PubChemLite', 'Biochemical Tr...', 'UVCBs: NORMA...', 'Outlining SL...', 'UVCB Names', 'NORMAN-SLE...', 'Bioschemas JS...', 'RDF Subset Pu...', 'CCS / Lipids', 'Grouping Hom...', 'Federated quer...', 'Feasibility of m...', and 'Please add mo...'. A sidebar on the left shows the repository's main branch and a list of files, including 'DeniseSI22 Script from Eva...', 'dataAnnotation_UVCBs', and 'regex_UVCBs'.

- Emma and Evan did a lot of number crunching and the output sent by Jeff. An excerpt:

```
Synonyms found in 1 concepts: 142172
Synonyms found in 2 concepts: 13248
Synonyms found in 3 concepts: 1668
Synonyms found in 4 concepts: 313
Synonyms found in 5 concepts: 76
Synonyms found in 6 concepts: 20
Synonyms found in 7 concepts: 1
Synonyms found in 11 concepts: 1
Synonyms found in 22 concepts: 2
Synonyms found in 23 concepts: 2
```

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

The screenshot shows a GitLab repository page titled 'UVCBs / Concepts'. The page content includes a note: 'Note that the contents of this folder are "work in progress".' Below this, there is a section titled 'Concepts and CIDs' with the following text: 'The concept_cids folder contains mapping files (txt ending) containing the CIDs that mentioned in the file name (concept_XXXXX...). A csv file mapping between concept name, concept number and matching OngLai series'. There is also a section titled 'PubChemLite EXPOSOMICS' with a logo featuring a pineapple and the name 'Ong Lai'. The text continues: 'These CIDs were obtained by running the OnLai algorithm on PubChemLite EXPOSOMICS annotation content. The OngLai outputs were theory that mapping CIDs were likely component hits). The matches were reviewed manually, and concepts were selected that are already on the dev site.' At the bottom, it states: 'The code associated with this work is currently under development in a private GitLab repository: <https://gitlab.lcsb.uni.lu/anjana.elapavalore/pubchem-concepts>. This work is an outcome of collaborative efforts of Evan Bolton, Anjana Elapavalore and Emma Schymanski, as part of Project 26 at BioHackathon Europe 2022.'



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

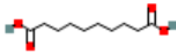
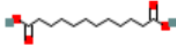
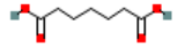
4 Related Records

UVCB Carboxylic acids, di-, C4-11

8 items

Search

SORT BY UVCB Component - A to Z

UVCB Component	Type	Component Structure	Source	Reference
Decanedioic acid	Component		DOI:10.5281/zenodo.7414795	DOI:10.5281/zenodo.7883657
Dodecanedioic acid	Component		DOI:10.5281/zenodo.7414795	DOI:10.5281/zenodo.7883657
Heptanedioic acid	Component		DOI:10.5281/zenodo.7414795	DOI:10.5281/zenodo.7883657

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 Amines, di-C14-18-alkylmethyl

4 items

UVCB Component	Type	Component Structure	Source	Reference
Di(hexadecyl)methylamine	Component		DOI:10.5281/zenodo.7883657	DOI:10.5281/zenodo.7414795
Di(octadecyl)methylamine	Component		DOI:10.5281/zenodo.7883657	DOI:10.5281/zenodo.7414795

Amines, di-C14-18-alkylmethyl

Related Records

UVCB Amines, di-C14-18-alkylmethyl

4 items

UVCB Component	Type	Component Structure	Component CID	Reference	Source	Comment
Di(hexadecyl)methylamine	Component		105560	DOI:10.5281/zenodo.7883657	DOI:10.5281/zenodo.7414795	
Di(octadecyl)methylamine	Component		77709	DOI:10.5281/zenodo.7883657	DOI:10.5281/zenodo.7414795	Longest chain length with annotation content

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/S1516-8913200000300011 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)


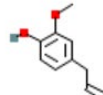
3.2 Natural Products

246 items View More Rows & Details

Structure	Compound CID	Compound
	2537	Camphor
	2758	Eucalyptol

3.1 Metabolites

160 items View More Rows & Details

Structure	Compound CID	Compound	Evidence IDs
	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 biological or biological products (camphor tree) via the embedded hyperlink. This document outlines how to retrieve further information about chemical components related to PubChem. Associate chemicals of interest to the biological section, but note that not all of these subjects are necessarily of interest. 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)

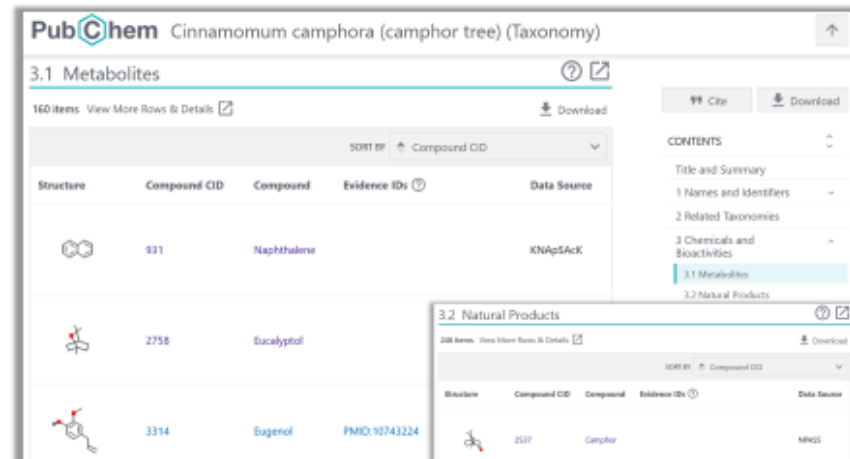


Figure 1: Associated chemicals of interest for *Cinnamomum camphora* on PubChem

Challenges / Perspectives – Example of PCBs

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

PubChem Search: PCBs

Compounds (1) Substances (29) BioAssays (38) **Literature (14,176)**

Searching PubMed abstracts and metadata. [Read More...](#)

14,176 results Filters SORT BY Relevance Download

Dioxins and PCBs in feed and food--review from European perspective

PMID: [24804623](#) Publication Type: Review
Publication Date: 2014-09-01
Journal: The Science of the total environment
Author(s): Rainer Malisch, Alexander Kotz

Abstract: During the 1990s, a number of adverse contamination incidents the general public on food safety. This led to the evaluation of safety mea food. Important aspects regarding dioxins and PCBs in the food chain are understanding of the present situation through its chronological develop exposure of humans to dioxins and PCBs results from dietary intake, with predominant source. Therefore, animal feed contributes considerably to t The detection of the 'real' source of a contamination event in the food ch requires specific knowledge on production processes and changes of pat demonstrated by complex investigations performed in three studies on t

“NORMANUVCBs” being born at the NORMAN GA, Dec. 2022

[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)

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



- Adding 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		
UVCB_Name	Type	Name	Synonym	PubChem_SID	SMILES	InChIKey	CAS_RN	DTXSID	Compositio	Source
PCBs	Component	2,3,3'-Trichlo	2,3,3'-Trichloro-	38034	C1C=CC(=C)SXLTKVKNQVZGI	38444-84	DTXSID8073502			Randolph Singh
PCBs	Component	2,2',4,5',6-Pe	2,2',4,5',6-Penta	63086	C1C=CC=C(C)PQHZWWBJPCNN	60145-21	DTXSID5074189			Randolph Singh
DL-PCBs	Component	2,3,3',4,4',5,5'	2,3,3',4,4',5,5'-H	38306	C1C=CC(=C)XUAWBXYHDDRR	39635-31	DTXSID4074144			Randolph Singh
PCBs	Component	2,2',3,4',6'-Pe	2,2',3,4',6'-Penta	43238	C1C=CC(C)=GOFFZTAPOOICF	60233-25	DTXSID9074193			Randolph Singh
PCBs	Component	3,3',4'-Trichlo	3,3',4'-Trichloro-	37806	C1C=CC(=C)JHBVPKZLIBDTJR	137680-69	DTXSID60865879			Randolph Singh
PCBs	Component	2,3',5,5'-Tetra	2,3',5,5'-Tetrach	38878	C1C=CC(=C)WBTMFEPLVQOV	41464-42	DTXSID8074152			Randolph Singh
PCBs	Component	2,2',3,4,6-Pen	2,2',3,4,6-Pentac	41362	C1C=CC(C)=QGDKRLQRLFUJPF	55215-17	DTXSID6074178			Randolph Singh
PCBs	Component	2,3,3',4,5'-Per	2,3,3',4,5'-Penta	91704	C1C=CC(=C)MPCDNZSLWJDN	70362-41	DTXSID1074206			Randolph Singh
PCBs	Component	3,3',4,5'-Tetra	3,3',4,5'-Tetrach	63104	C1C=CC(=C)QLCTXEMDCZGPC	41464-48	DTXSID3074155			Randolph Singh
PCBs	Component	2,3,4,4',5-Pen	2,3,4,4',5-Pentac	53036	C1C=CC=C(C)SXZFWHOSHAKN	74472-37	DTXSID9074226			Randolph Singh
PCBs	Component	3,4'-Dichloro	3,4'-Dichloro-	118101	C1C=CC(C)CJIDNEKOMKXLSBI	2974-90	DTXSID10863067			Randolph Singh
PCBs	Component	2,2',3,5,6,6'-H	2,2',3,5,6,6'-Hex	63070	C1C=CC(C)=CLODVBWVNVQL	68194-09	DTXSID70867526			Randolph Singh
PCBs	Component	2,2',3,3',4,4',5,2,2',3,3',4,4',5,6-		40485	C1C=C(C)C(C)JAHJITLJFSDRCG	152663-78	DTXSID1074171			Randolph Singh
PCBs	Component	2,3,4',6-Tetra	2,3,4',6-Tetrachl	63110	C1C=CC=C(C)FXRXQYZALWWC	52663-58	DTXSID3074159			Randolph Singh
PCBs	Component	2,3,3',4,4',5-H	2,3,3',4,4',5-Hex	38019	C1C=CC(=C)CLXCMEXLGMKFLQ	38380-08	DTXSID0052706			Randolph Singh
PCBs	Component	2,2',5,5'-Tetra	2,2',5,5'-Tetrach	37248	C1C=CC(=C)CHCWZEPKLWVAE	35693-99	DTXSID3038305			Randolph Singh
PCBs	Component	2,4-Dichloro	2,4-Dichloro-	1136399	C1C=CC(C)=WEJZHJXPXXML	33284-50	DTXSID8040301			Randolph Singh
PCBs	Component	3,4',5-Trichlo	3,4',5-Trichloro-	38038	C1C=CC=C(C)SYSBNFJJSJLZMM	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)

NORMAN-UVCBs & Polychlorinated Biphenyls

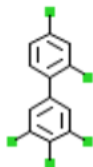
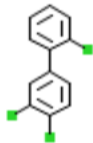
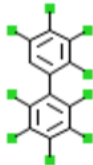
UVCB Polychlorinated biphenyls

209 items

Download

Search

SORT BY UVCB Component - A to Z

UVCB Component	Type	Component Structure	Component CID	Reference	Source
2',3,4,4',5-Pentachlorobiphenyl	Component		47650	https://comptox.epa.gov/dashboard/chemical-lists/PCBCHEMICALS	DOI:10.5281/zenodo.7414795
2',3,4-Trichlorobiphenyl	Component		38036	https://comptox.epa.gov/dashboard/chemical-lists/PCBCHEMICALS	DOI:10.5281/zenodo.7414795
2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl	Component		38411	https://comptox.epa.gov/dashboard/chemical-lists/PCBCHEMICALS	DOI:10.5281/zenodo.7414795

<< First < Previous Page 1 of 21 Next > Last >>

NORMAN-UVCBs & Polychlorinated Biphenyls



UVCB Polychlorinated biphenyls

209 items

Download

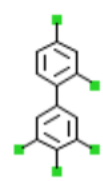
Search

SORT BY UVCB Component - A to Z

UVCB Component	Type	Component Structure	Component CID	Reference	Source
----------------	------	---------------------	---------------	-----------	--------

2',3,4,4',5-Pentachlorobiphenyl

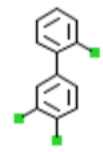
Component



47650

2',3,4-Trichlorobiphenyl

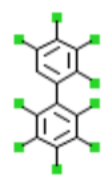
Component



38036

2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl

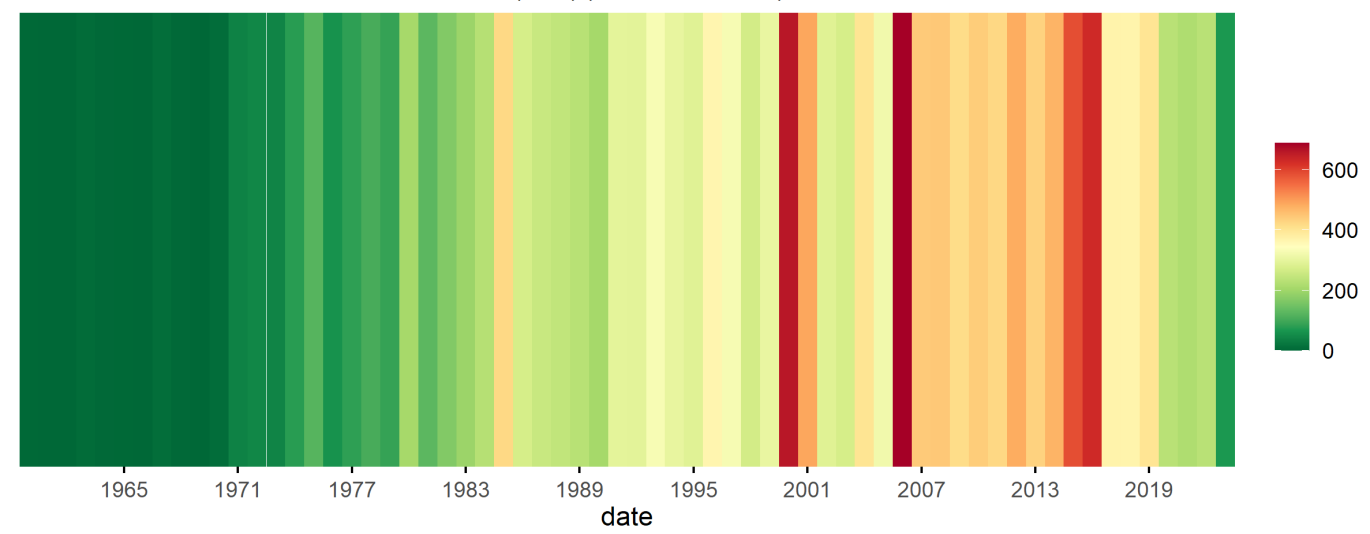
Component



38411

Summarized Chemical Stripes - All PCBs

Literature data compiled from 209 CIDs
First reference of all selected CIDs: 1938 (CID(s): 16323,17348)



<< First < Previous Page 1 of 21 Next > Last >>

Take home messages

- Text-based curation 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 )
```

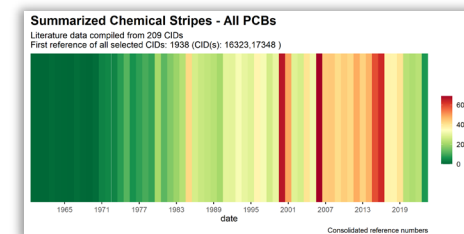
- Algorithms can support cross-linking **Variable** cases



- **Biological** cases leverage other disciplines

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

- Adding in **NORMAN** Network UVCB data case by case



- We would be happy to work with HESI on the **Tier 0 Database!**

- Emma: emma.schymanski@uni.lu or [@ESchymanski](https://twitter.com/ESchymanski)
PubChem: pubchem-help@ncbi.nlm.nih.gov or [@pubchem](https://twitter.com/pubchem)

Acknowledgements!

Today's slides:

DOI: [10.5281/zenodo.8302440](https://doi.org/10.5281/zenodo.8302440)

Email: emma.schymanski@uni.lu

pubchem-help@ncbi.nlm.nih.gov

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



PubChem

NIH U.S. National Library of Medicine
National Center for Biotechnology Information



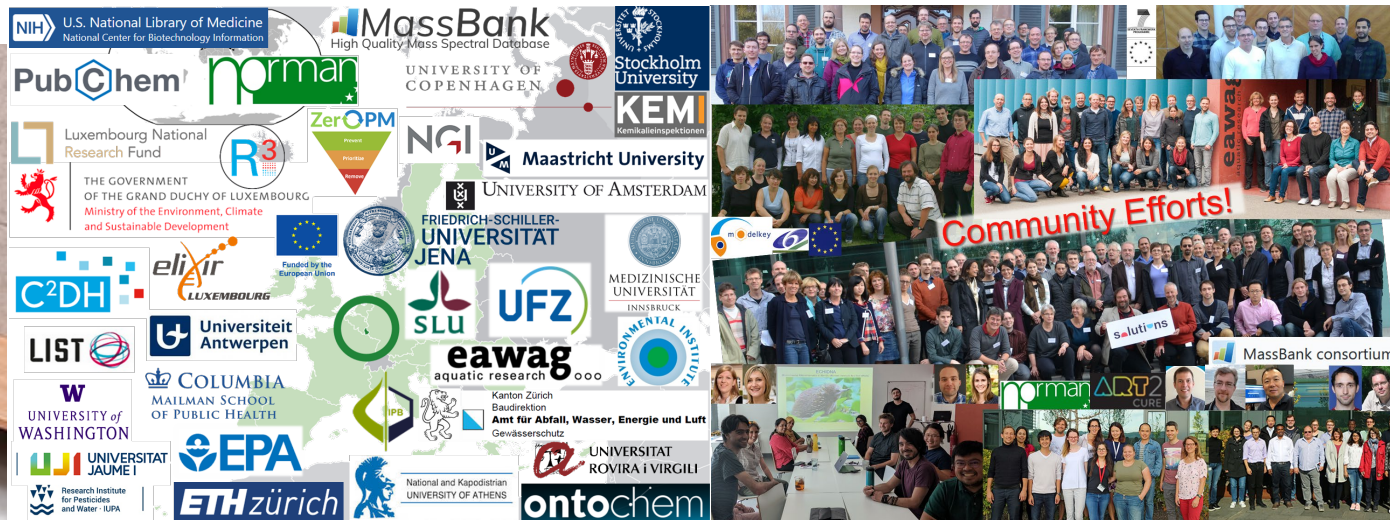
Evan Bolton, Jian (Jeff) Zhang, Paul Thiessen,
Leonid Zaslavsky, Qingliang (Leon) Li,
plus Tiejun, Asta, Ben, Siqian + the whole team

This work was supported in part by the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM), National Institutes of Health (NIH)



Adelene Lai
(OngLai)

Anjana
Elapavalore
OngLai+PCL
Mapping



Funded by the
European Union



H2020:
101036756



Luxembourg
National
Research Fund

