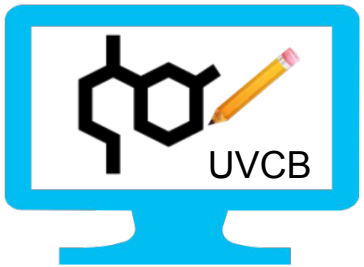
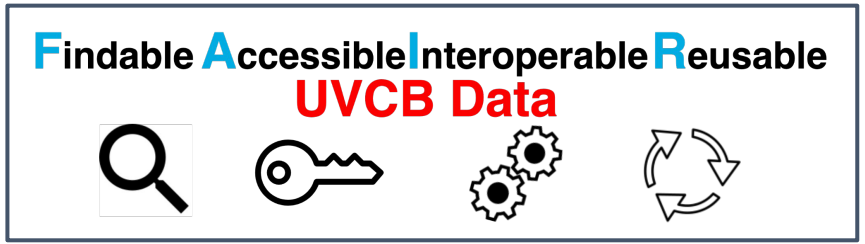
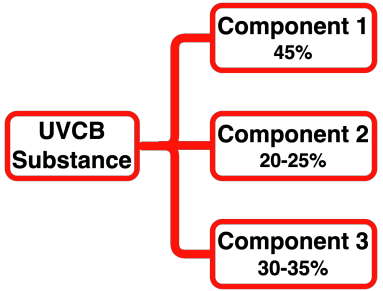
In silico Structure Elucidation & FAIR Information Management

for Improved UVCB Assessment







Adelene Lai[1,2], Nina Sachdev[3], Alex M. Clark[4], Marc Fernandez[5], Leah R. McEwen[6,7], Alexander Okonski[5], Katrina Sullivan[5], and Emma L. Schymanski[1]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg; [2]Friedrich-Schiller-University; [3]Wellesley College; [4]Collaborative Drug Discovery; [5]Environment and Climate Change Canada; [6]Cornell University; [7]International Union of Pure and Applied Chemistry

SETAC Europe, 18 May 2022

Luxembourg National Research Fund

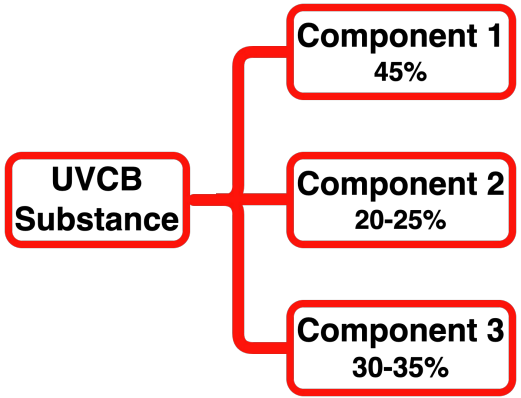
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

uni.lu UNIVERSITÉ DU LUXEMBOURG

LCSB

AMC develops commercial products to deal with mixtures at CCD. No further competing interests.

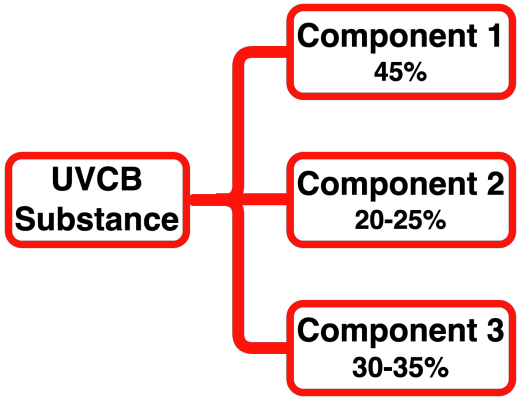
Unknown or
Variable Composition,
Complex Reaction Products or
Biological Materials

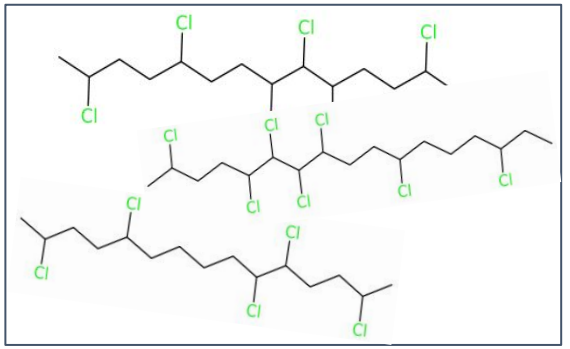


[1]Sauer *et al.* (2020) DOI: 10.1016/j.yrtph.2020.104583 [2]Kutsarova *et al.* (2019) DOI: 10.1002/etc.4358 [3]ECHA (2017) DOI: 10.2823/794394

Unknown or Variable Composition, Complex Reaction Products or Biological Materials

20-40% registered substances[1-3]



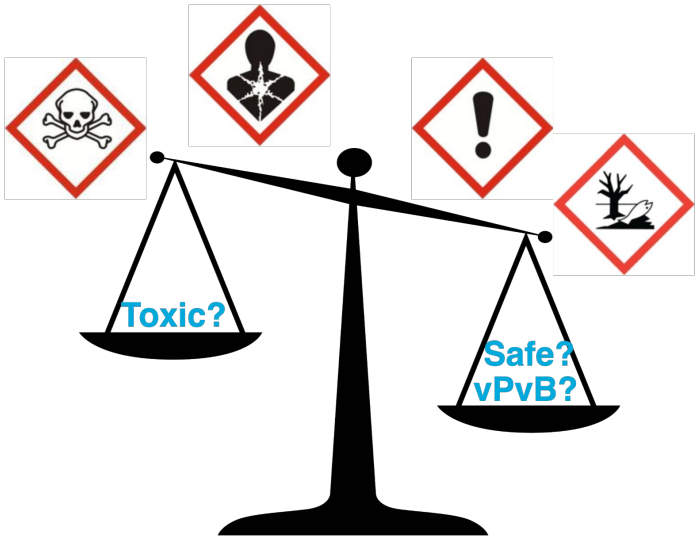


Alkanes, C_{14-17}, chloro (MCCPs)

Cedarwood Oil

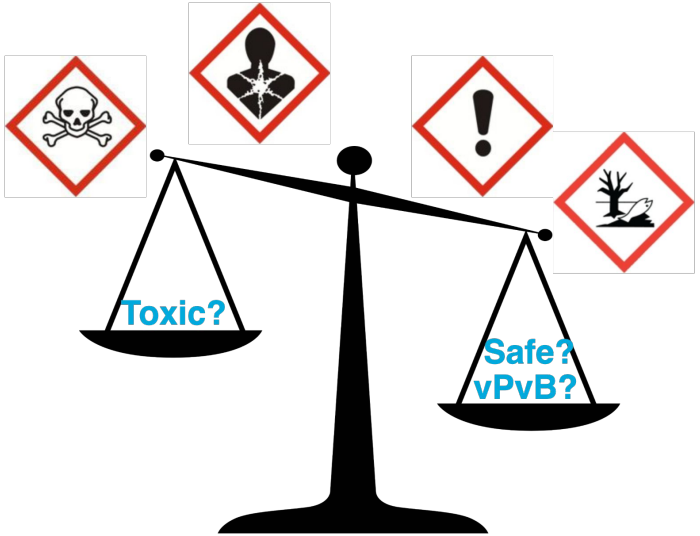
Petroleum Distillates

[1]Sauer *et al.* (2020) DOI: 10.1016/j.yrtph.2020.104583 [2]Kutsarova *et al.* (2019) DOI: 10.1002/etc.4358 [3]ECHA (2017) DOI: 10.2823/794394



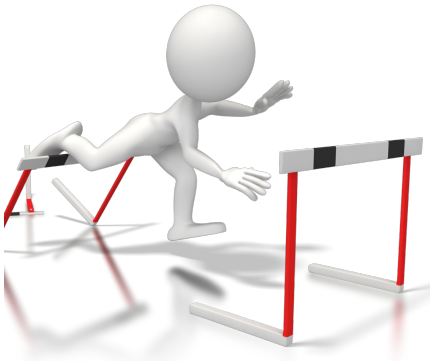
UVCB Regulatory Assessment is challenging[4-5]

[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321 [5]Salvito *et al.* (2020) DOI: 10.1002/etc.4846



UVCB Regulatory Assessment is challenging[4-5]

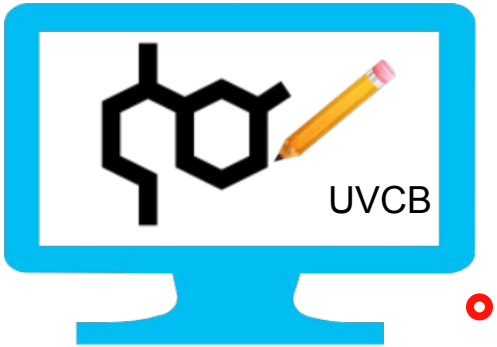
UVCB Data & Cheminformatics



1. Limited Structural Info in public domain

2. UnFAIR Info Management of UVCBs in databases

1. Limited Structural Info:
***In silico* Structure Elucidation**



UVCB representative structures needed for assessment

with
M. Fernandez, A. Okonski, K. Sullivan
(Environment and Climate Change Canada)

[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321 [5]Salvito *et al.* (2020) DOI: 10.1002/etc.4846

Methods: *In silico* Structure Elucidation

- Search databases & literature

PubChem ECHA EUROPEAN CHEMICALS AGENCY

CompTox Chemicals Dashboard

pubchem.ncbi.nlm.nih.gov/
echa.europa.eu/
comptox.epa.gov/dashboard/

→ structures of analogous UVCBs

→ composition info:
major & representative components

[6]Mayfield, J. (2021) simolecule.com/cdkdepict/depict.html

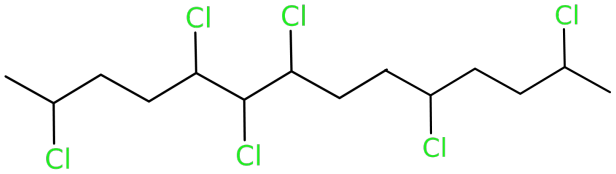
Methods: *In silico* Structure Elucidation

- Search databases & literature
  - → structures of analogous UVCBs
  - → composition info: major & representative components

PubChem ECHA EUROPEAN CHEMICALS AGENCY

CompTox Chemicals Dashboard

pubchem.ncbi.nlm.nih.gov/
echa.europa.eu/
comptox.epa.gov/dashboard/



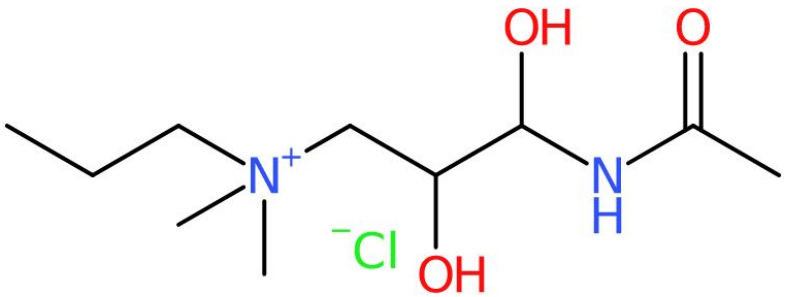
- UVCB name → (sub)structure
  - Draw & combine substructures using CDKDepict[6]
    - Or: ChemDraw, Structure Generation (LMC, Bulgaria)

[6]Mayfield, J. (2021) simolecule.com/cdkdepict/depict.html

Results: *In silico* Structure Elucidation 1

1-Propanaminium, N-(3-aminopropyl)-2,3-dihydroxy-N,N-dimethyl-, N-limnanthes alba seed oil-acyl derivs., chlorides (CASRN 345648-01-3)

Step 1: Search & draw Substructure 1



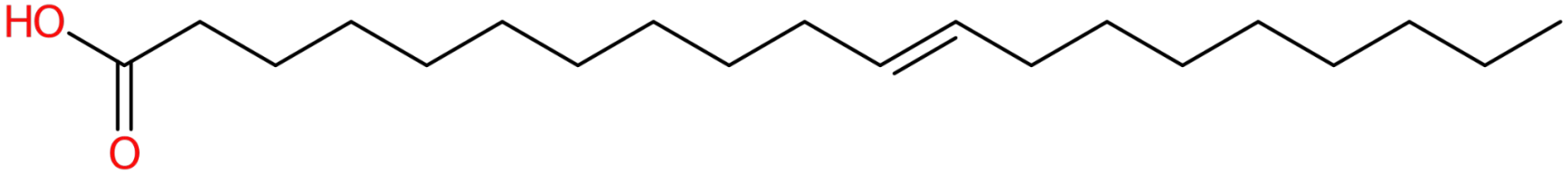
via analogous structures in databases (PubChem)

Results: *In silico* Structure Elucidation 1

1-Propanaminium, N-(3-aminopropyl)-2,3-dihydroxy-N,N-dimethyl-, N-limnanthes alba seed oil-acyl derivs., chlorides (CASRN 345648-01-3)

Step 2: Literature search for major component (Substructure 2)

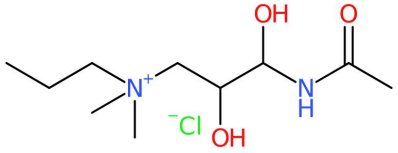
~60% Eicosenoic acid 20:1[7,8]

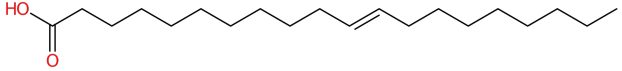


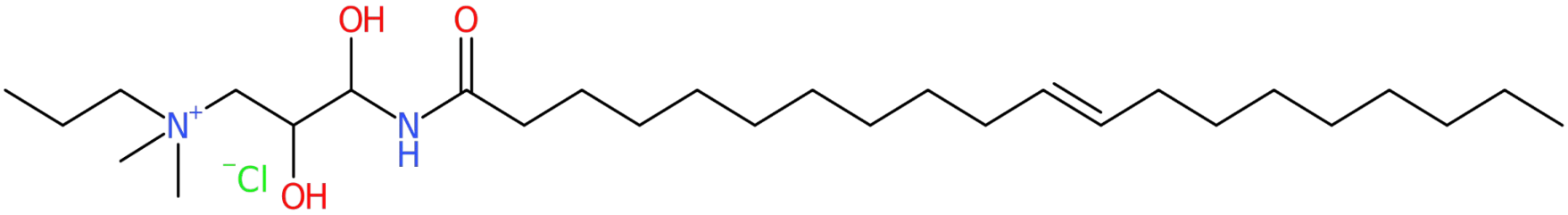
[7]Moser *et al.* (2010) DOI: 10.1039/B923740M [8]UL Prospector (2022) ulprospector.com/en/eu/PersonalCare/Detail/2562/198451/FANCOR-MEADOWFOAM-SEED-OIL

Results: *In silico* Structure Elucidation 1

Step 3: Combine Substructures





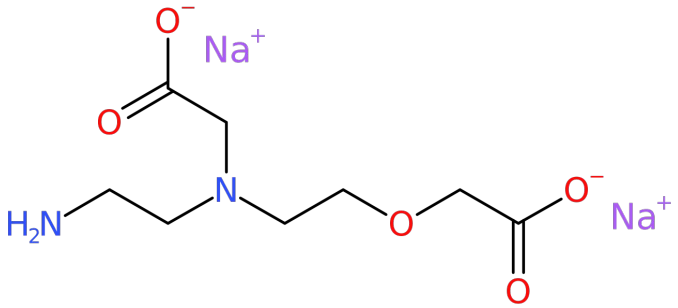


1-Propanaminium, N-(3-aminopropyl)-2,3-dihydroxy-N,N-dimethyl-, N-limnanthes alba seed oil-acyl derivs., chlorides (CASRN 345648-01-3)

Results: *In silico* Structure Elucidation 2

Glycine, N-(2-aminoethyl)-N-[2-(carboxymethoxy)ethyl]-,
N-wheat germ-oil acyl derivs., disodium salts (CASRN 646996-25-0)

Step 1: Seach & draw Substructure 1



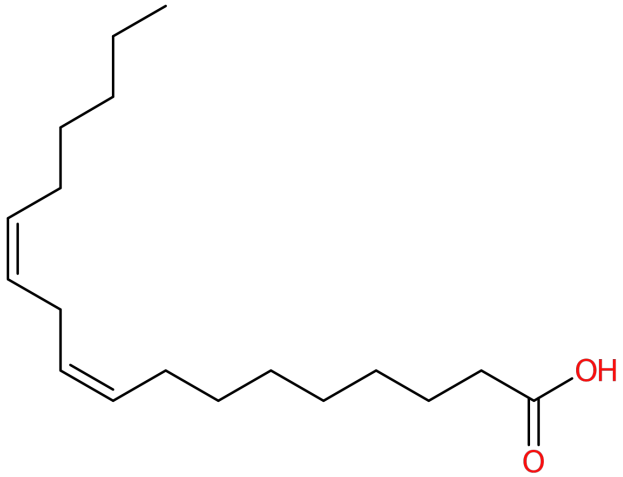
via synonym match in databases (PubChem)

Results: *In silico* Structure Elucidation 2

Glycine, N-(2-aminoethyl)-N-[2-(carboxymethoxy)ethyl]-,
N-wheat germ-oil acyl derivs., disodium salts (CASRN 646996-25-0)

Step 2: Literature search for major component (Substructure 2)

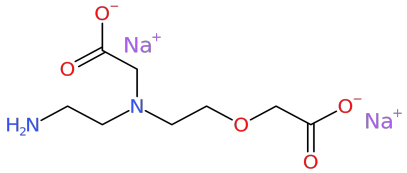
~55% Linoleic acid 18:2[9,10]

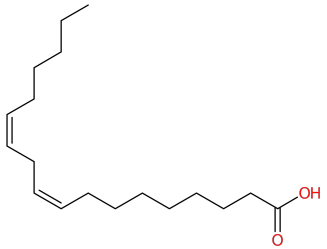


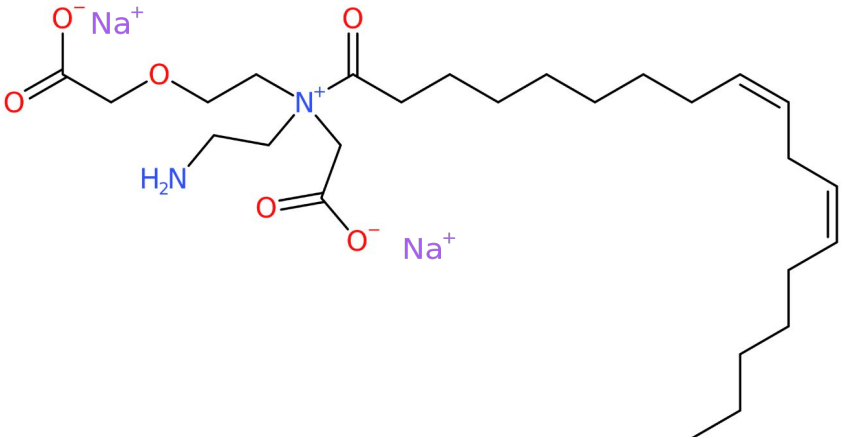
[9]Kumar and Krishna (2015) DOI: 10.1007/s13197-013-1119-3 [10]Ghafoor *et al.* (2017) DOI: 10.1002/ejlt.201600160

Results: *In silico* Structure Elucidation 2

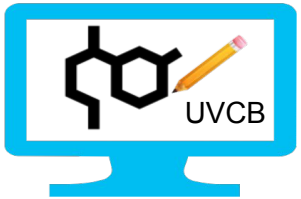
Step 3: Combine Substructures







Glycine,
N-(2-aminoethyl)-N-[2-(carboxymethoxy)ethyl]-,
N-wheat germ-oil acyl derivs., disodium salts
(CASRN 646996-25-0)



Outlook: *In silico* Structure Elucidation

Representative chemical structures proposed for UVCBs

- Only UVCBs w/chemically interpretable names
- Analytical validation ideal[11]

Representative structures needed for assessment

[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321 [11]Chibwe *et al.* (2019) DOI: 10.1021/acs.est.9b03760 [12]Rajan *et al.* (2021) DOI: 10.1186/s13321-021-00512-4
[13]McKay *et al.* (2021) DOI: 10.26434/chemrxiv-2021-gt5lb [14]Clark *et al.* (2021) DOI: 10.1021/acsomega.1c03311

Outlook: *In silico* Structure Elucidation

Representative chemical structures proposed for UVCBs

- Only UVCBs w/chemically interpretable names

Representative structures needed for assessment

- Analytical validation ideal[11]
- Further automated methods
  - within databases[4]
  - open cheminformatics/ML tools:

STOUT[12] (IUPAC name to SMILES)

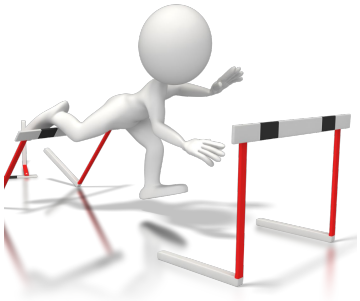
SURGE[13](Open Structure Generator)

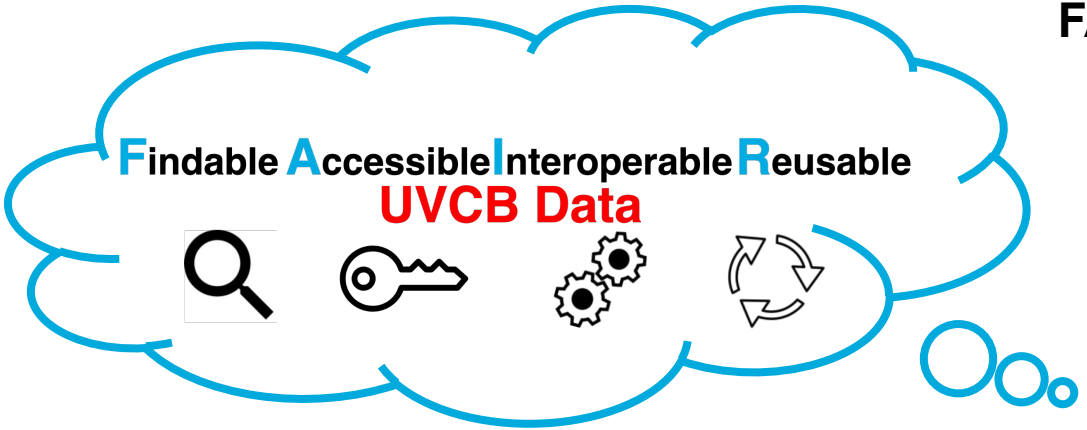
Mixture Description Parser[14]

[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321 [11]Chibwe *et al.* (2019) DOI: 10.1021/acs.est.9b03760 [12]Rajan *et al.* (2021) DOI: 10.1186/s13321-021-00512-4
[13]McKay *et al.* (2021) DOI: 10.26434/chemrxiv-2021-gt5lb [14]Clark *et al.* (2021) DOI: 10.1021/acsomega.1c03311

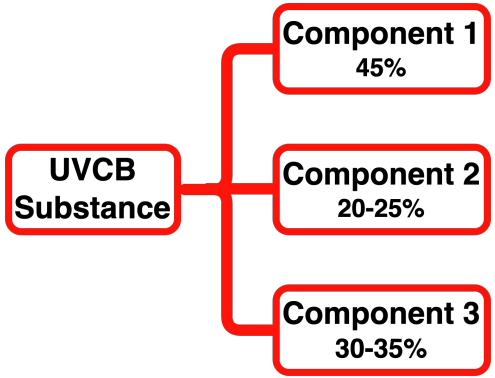
Challenging to find & use UVCB data

- limited searchability - Name? CASRN? EC No.?[4,5]
- non-universal across databases
- closed *e.g.*, G SMILES[15]
- not machine readable (*e.g.*, images)



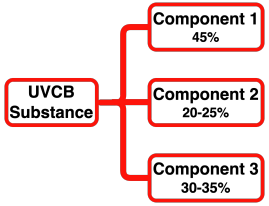


2. UnFAIR Info Management:
FAIR-ify UVCB Data with MInChI



with
N. Sachdev, A. M. Clark, L. R. McEwen, Z. Wang

[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321 [5]Salvito *et al.* (2020) DOI: 10.1002/etc.4846



Methods: FAIR-ify UVCB Data with MInChI

- **Mixture InChI**[16]
  - Based on InChI

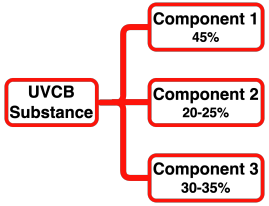
Glutamic Acid

InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/t3-/m0/s1

main

stereochemistry

[16]Clark *et al.* (2019) DOI: 10.1186/s13321-019-0357-4 [17]Sachdev, N. (2020) https://ninasachdev.github.io/UVCB-MInChI/

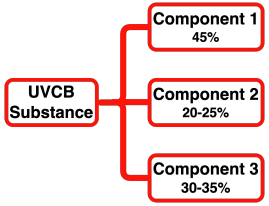


Methods: FAIR-ify UVCB Data with MInChI

- **Mixture InChI**[16]
  - Based on InChI
  - Data format for multiple components
  - Open, machine readable, searchable

Glutamic Acid
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/t3-/m0/s1
main
stereochemistry

compound
quantity
hierarchy



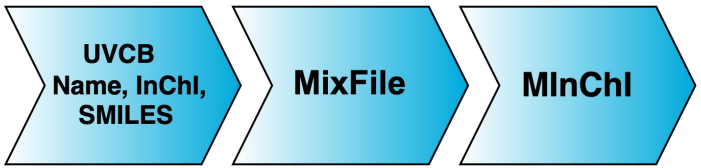
Methods: FAIR-ify UVCB Data with MInChI

- **Mixture InChI**[16]
  - Based on InChI
  - Data format for multiple components
  - Open, machine readable, searchable

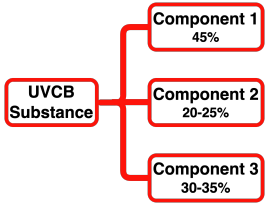
Glutamic Acid
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/t3-/m0/s1
main
stereochemistry

compound
quantity
hierarchy

- Workflow[17] to encode UVCBs as MInChI



[16]Clark *et al.* (2019) DOI: 10.1186/s13321-019-0357-4 [17]Sachdev, N. (2020) https://ninasachdev.github.io/UVCB-MInChI/



Results: FAIR-ify UVCB Data with MInChI 1

Alcohols, C9-11 (CASRN 66455-17-2)

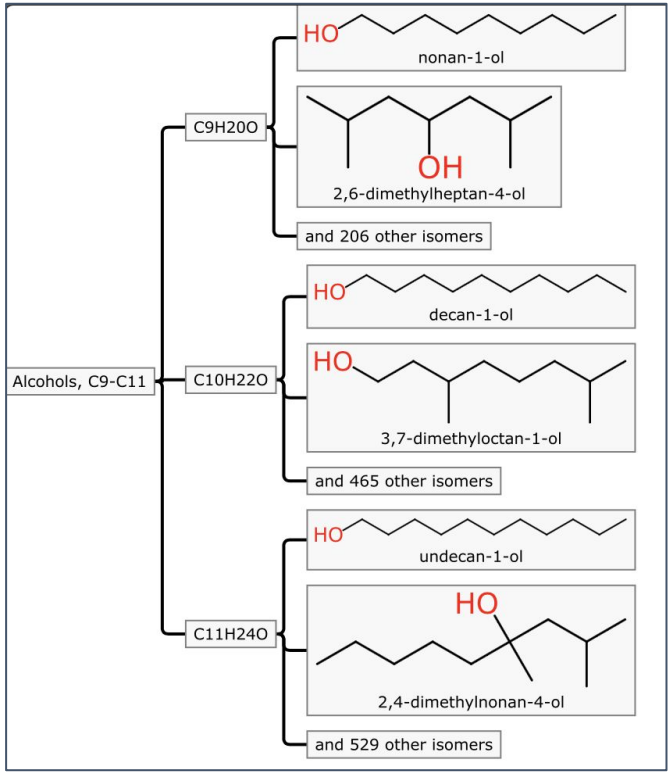
MInChI=0.99.1S/&C10H22O/c1-2-3-4-5-6-7-8-9-10-11/h11H,2-10H2,1H3&C10H22O/c1-9(2)5-4-6-10(3)7-8-11/h9-11H,4-8H2,1-3H3&C11H24O/c1-2-3-4-5-6-7-8-9-10-11-12/h12H,2-11H2,1H3&C11H24O/c1-5-6-7-8-11(4,12)9-10(2)3/h10,12H,5-9H2,1-4H3&C9H20O/c1-2-3-4-5-6-7-8-9-10/h10H,2-9H2,1H3&C9H20O/c1-7(2)5-9(10)6-8(3)4/h7-10H,5-6H2,1-4H3/n{6&7&1}&{2&3&1}&{4&5&1}/g{&&}&{&&}&{&&}

header

structure

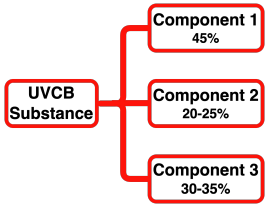
index

concentration



ninasachdev.github.io/UVCB-MInChI/

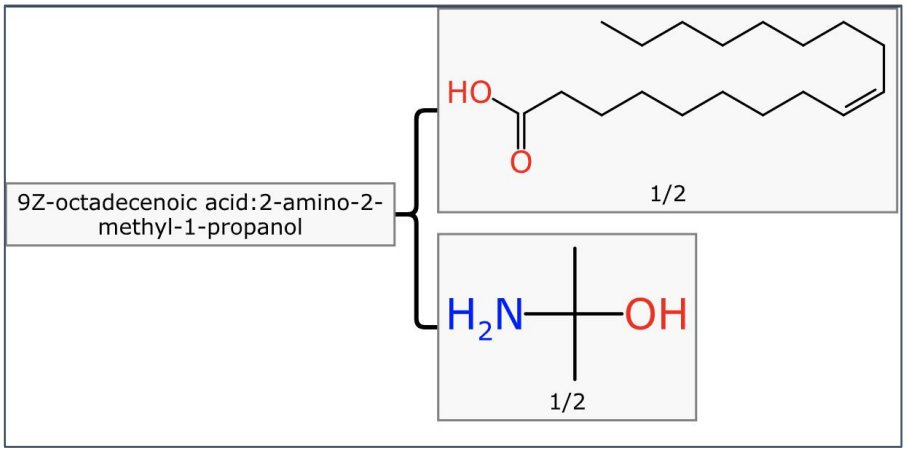
github.com/cdd/mixtures



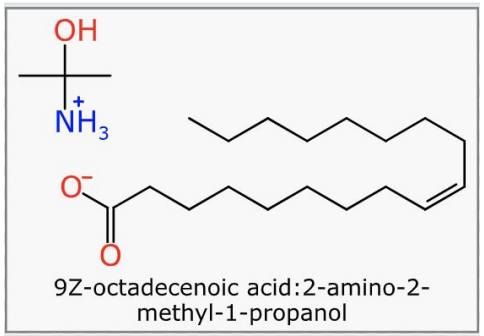
Results: FAIR-ify UVCB Data with MInChI 2

9-Octadecenoic acid (9Z)-, compd. With 2-amino-2-methyl-1-propanol (1:1) (CASRN 68140-41-0)

MInChI=0.99.1S/C18H34O2/c1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18(19)20/h9-10H,2-8,11-17H2,1H3,(H,19,20)/b10-9-&C3H9NO/c1-3(2,4)5/h5H,4H2,1-2H3/n{1&2}/g{1rt0&1rt0}

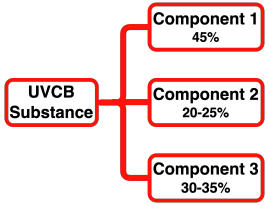


alternatively, as single component:



FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
uni.lu UNIVERSITÉ DU LUXEMBOURG
LCSB

ninasachdev.github.io/UVCB-MInChI/
github.com/cdd/mixtures



Outlook: FAIR-ify UVCB Data with MInChI

MInChI - a FAIR, Open, machine-readable data format for UVCBs

- 68 UVCBs → 68 MInChIs[17]
- Limited by UVCB data gaps

compound

quantity

hierarchy

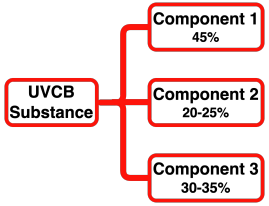
Findable **A**ccessible **I**nteroperable **R**eusable
UVCB Data

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

UNIVERSITÉ DU LUXEMBOURG

LCSB

[17] Sachdev, N. (2020) https://ninasachdev.github.io/UVCB-MInChI/

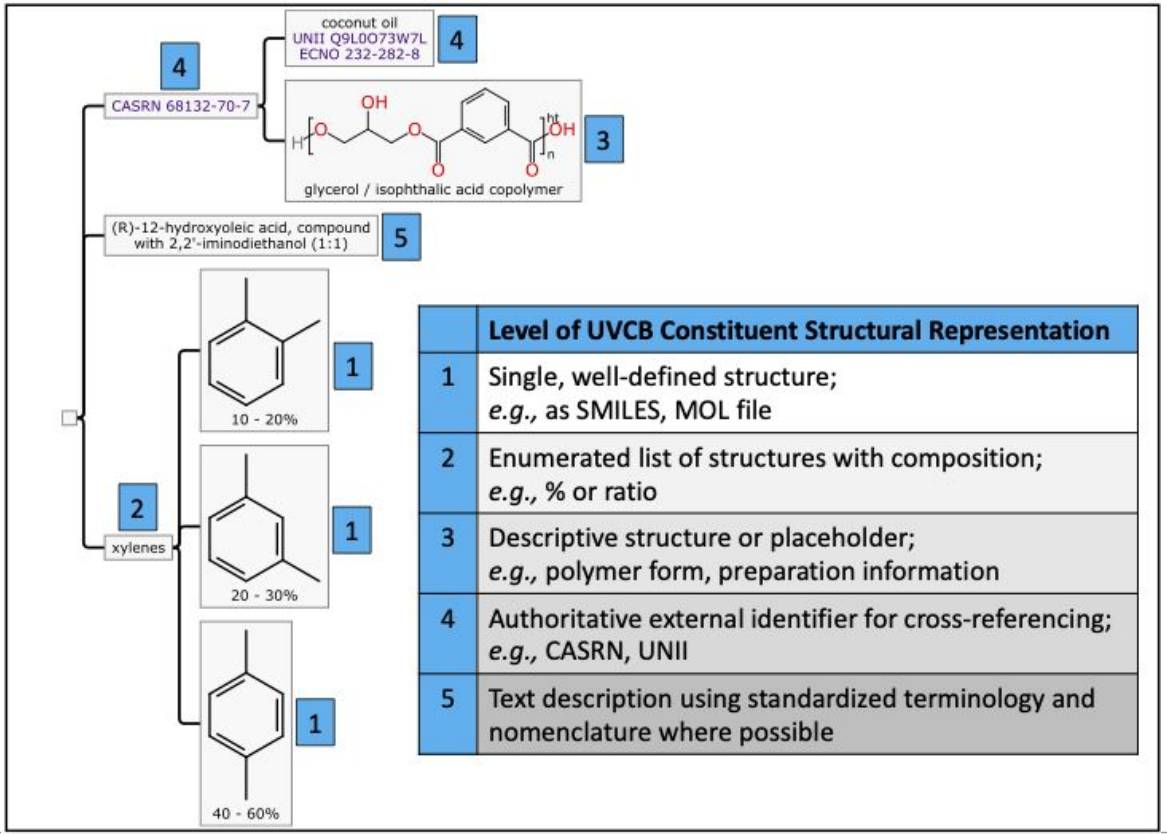


Outlook: FAIR-ify UVCB Data

Further Work:

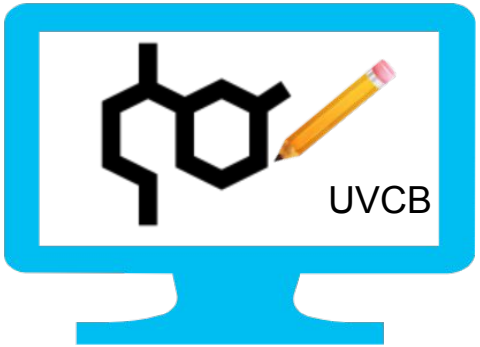
Systematising UVCB data[4]

→ scheme



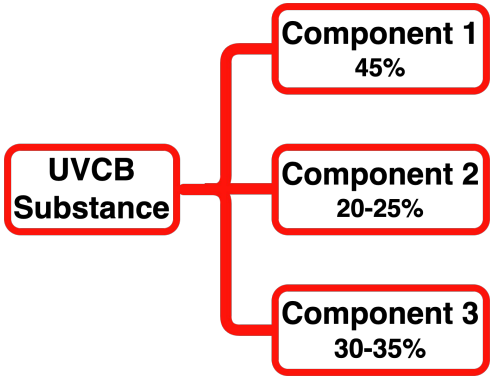
[4]Lai *et al.* (2022) DOI: 10.1021/acs.est.2c00321

1. Limited Structural Info:
***In silico* Structure Elucidation**



with
M. Fernandez, A. Okonski, K. Sullivan
(Environment and Climate Change Canada)

2. UnFAIR Info Management:
FAIR-ify UVCB Data with MInChI



with
N. Sachdev, A. M. Clark, L. R. McEwen, Z. Wang

Acknowledgements

- N. Sachdev, A. Clark, M. Fernandez, L. McEwen, A. Okonski, K. Sullivan
- Zhanyun Wang
- Emma Schymanski, Environmental Cheminformatics Group, LCSB, Luxembourg
- Christoph Steinbeck & Cheminformatics & Computational Metabolomics Group, FSU-Jena, Germany
- Funding:
  - AL & ELS - Luxembourg National Research Fund A18/BM/12341006
  - AMC & LRM - National Institutes of Health 2R44TR002328-02

Slides available: 10.5281/zenodo.6469349

Contact: adelene.lai[at]uni[dot]lu, emma.schymanski[at]uni[dot]lu