

Mixtures InChI

A story of how standards drive upstream products

Alex M. Clark & Leah R. McEwen

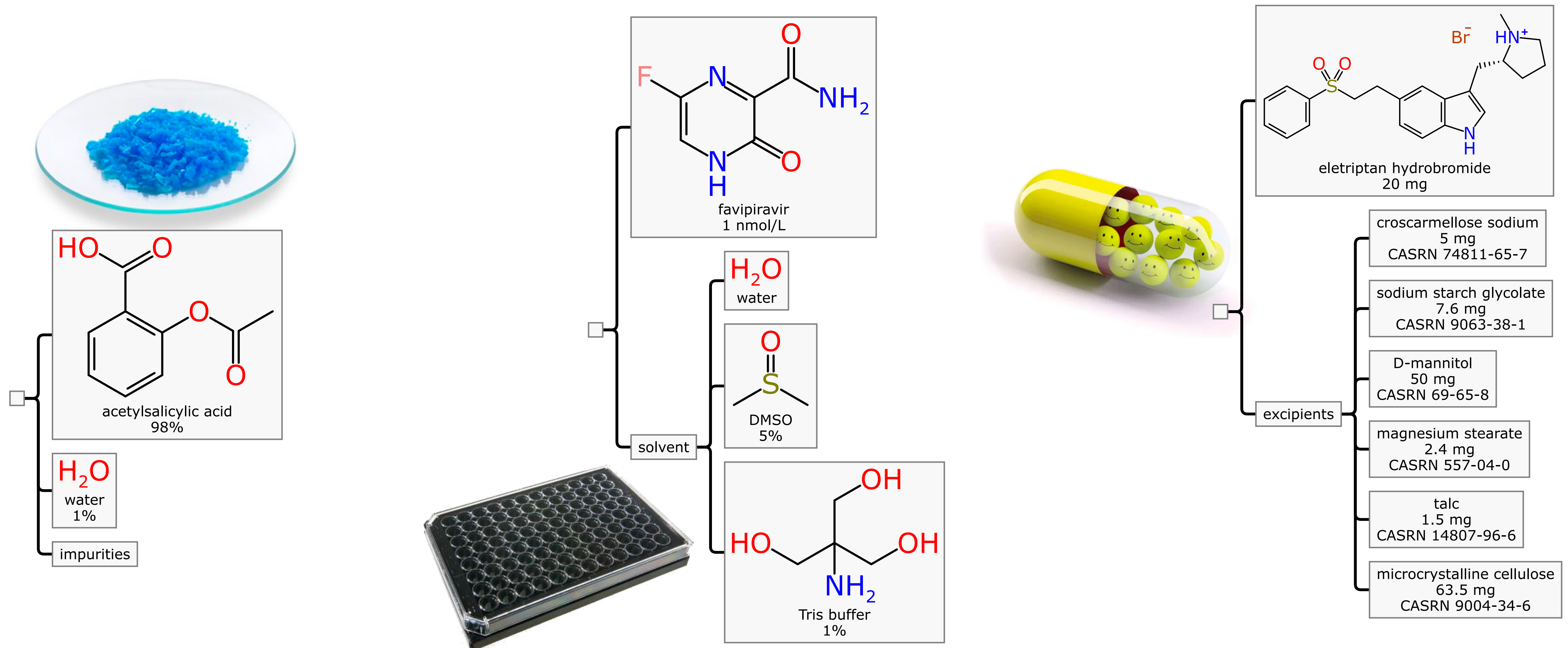
alex@collaborativedrug.com



CDD VAULT[®]
Complexity Simplified

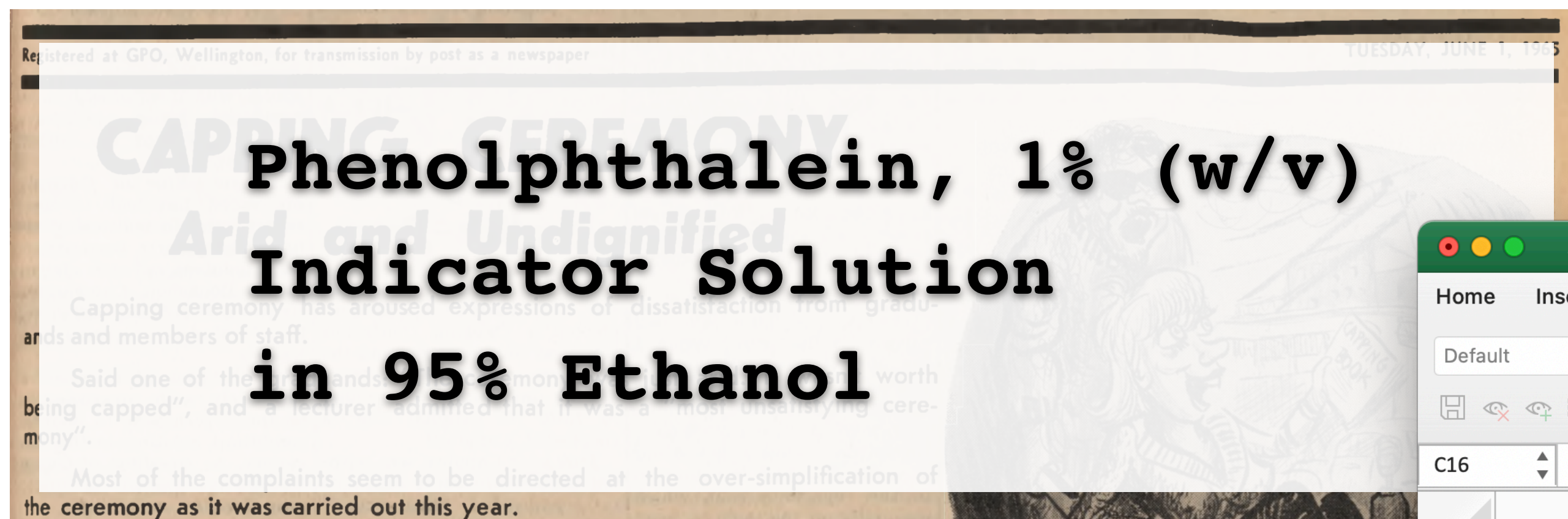
It's always a mixture

❖ The pure molecule approximation has value... but in the real world:



State of mixture informatics

❖ Machine readable molecules ~ 1/2 century ago, but mixtures limited to text



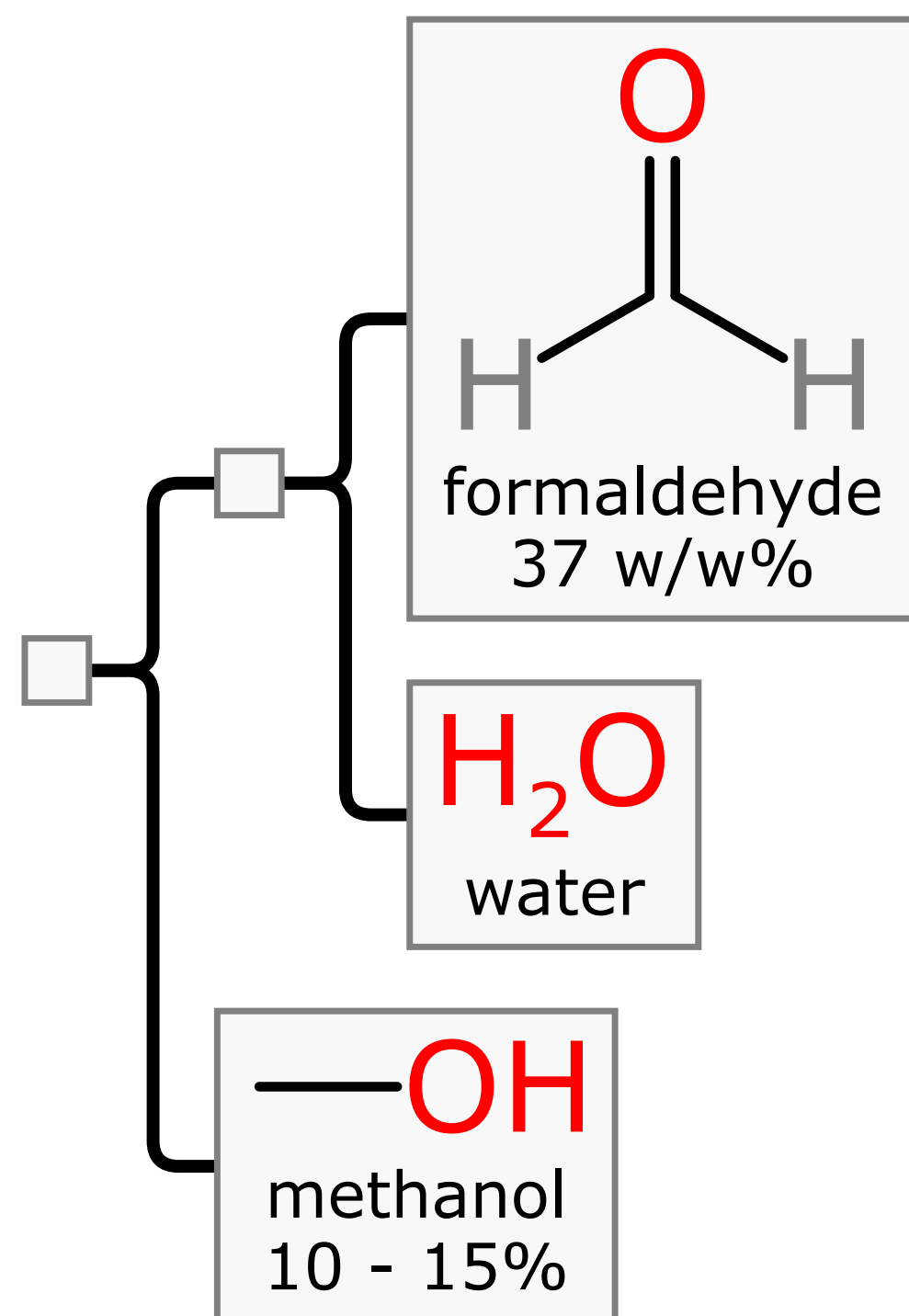
	A	B	C
1	<u>MOLECULE</u>	<u>CONC</u>	<u>SOLVENT</u>
2	copper sulfate	0.01 mol/L	water
3	hydrochloric acid	0.02 M	aqueous
4	perchloric acid	60%	
5	potassium sodium tartrate	30 w/v%	H2O
6	tellurium (IV) bromide	99.9%	n/a
7	Boc2O		dioxane
8	dithiothreitol & alpha-thioglycerol	2:3	
9	acetic acid	5% (v/v)	
10	buffer solution (ph 7.0)		
11			
12			

❖ Bespoke formats exist

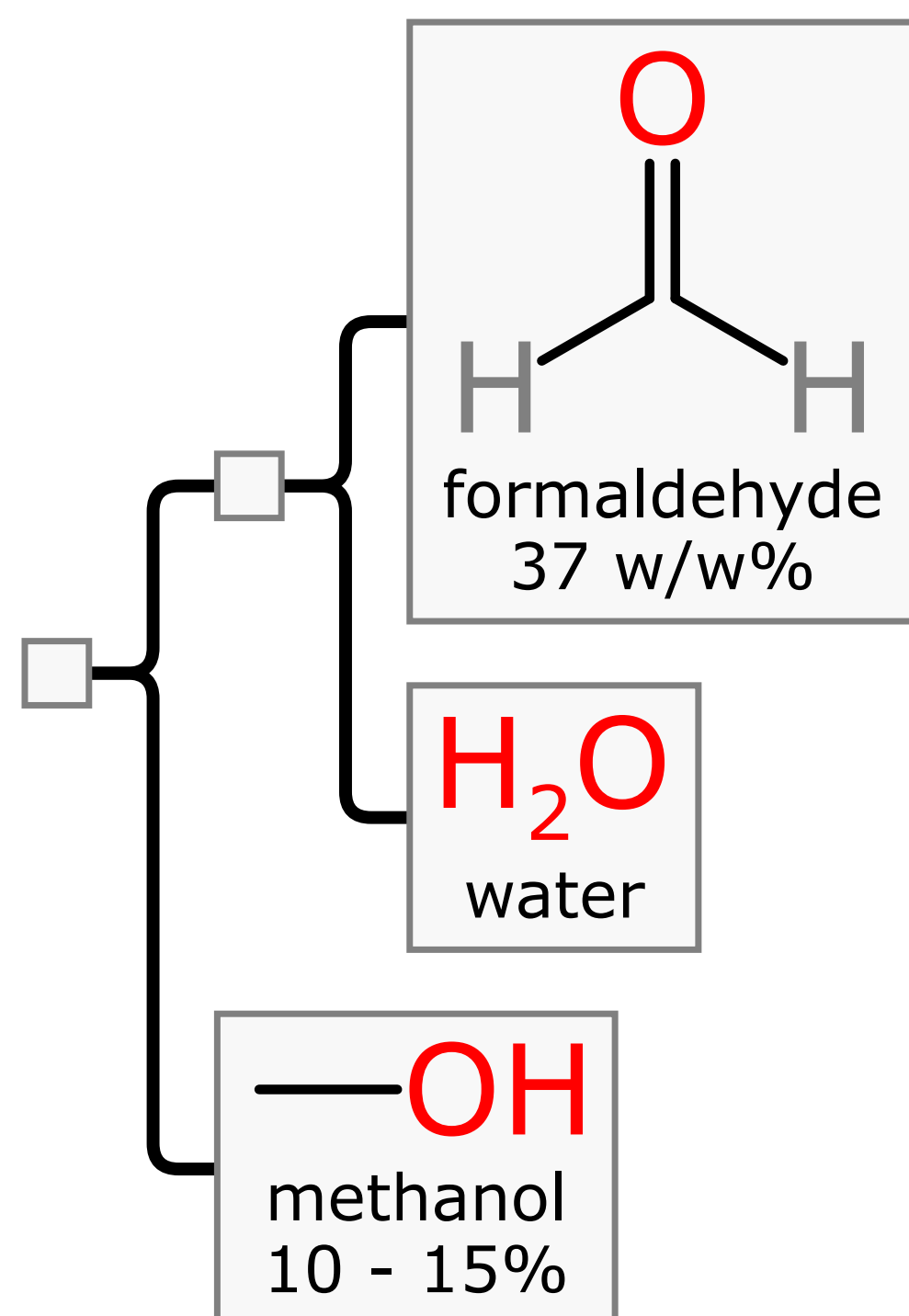
❖ Generally in silos:

- ▶ low machine readability
- ▶ low interoperability

Mixtures InChI



Mixtures InChI



CH₂O/c1-2/h1H2

1

37wF-2

H₂O/h1H2

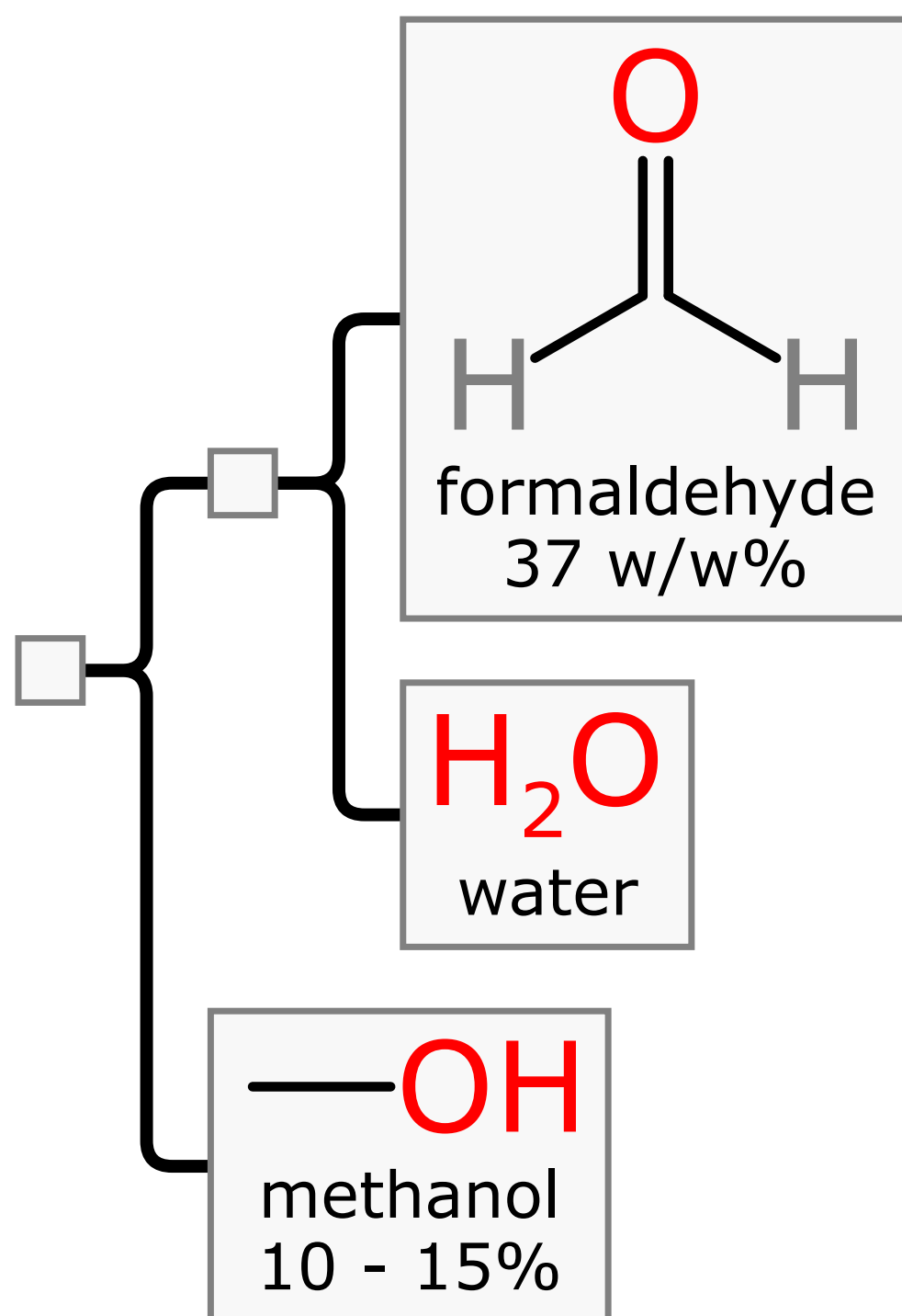
3

CH₄O/c1-2/h2H,1H3

2

10:15pp0

Mixtures InChI



CH₂O/c1-2/h1H2

1

37wF-2

H₂O/h1H2

3

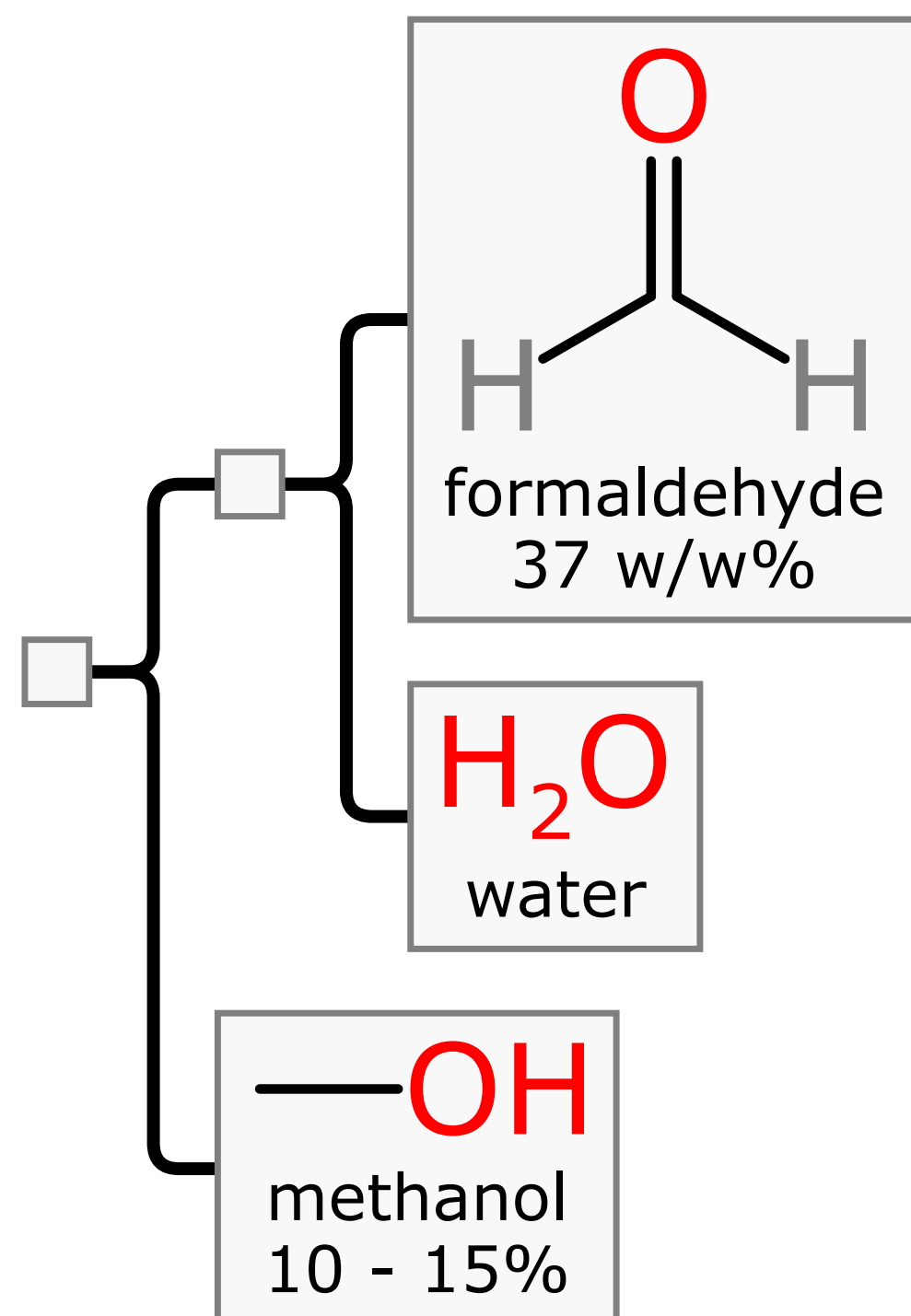
CH₄O/c1-2/h2H,1H3

2

10:15pp0

MInChI=0.00.1S/ & & & /n { { & } & } /g { { & } & }

Mixtures InChI



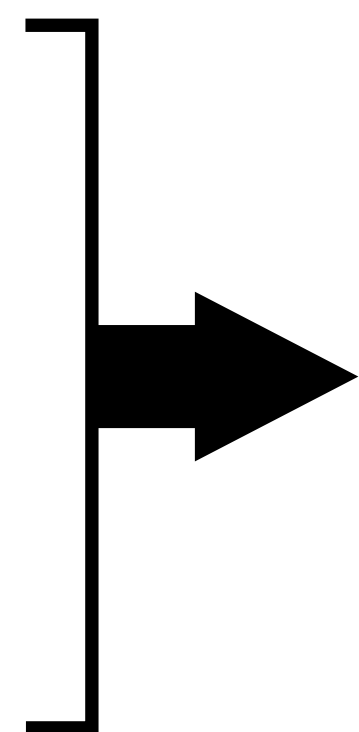
```
MInChI=0.00.1S/CH2O/c1-2/h1H2 & CH4O/c1-2/h2H,1H3 &  
H2O/h1H2/n{{1 & 3} & 2}/g{{37wf-2 & } & 10:15pp0}
```


One brick at a time...

♣ Have:	<input checked="" type="checkbox"/> use cases		talking to people: demand is real
	<input checked="" type="checkbox"/> standard		MInChI specification, IUPAC endorsed
	<input type="checkbox"/> tools		... custom or limited purpose
	<input type="checkbox"/> data		... not machine readable
	<input type="checkbox"/> community		... not informatics oriented

♣ What to do? Chicken vs. egg problems...

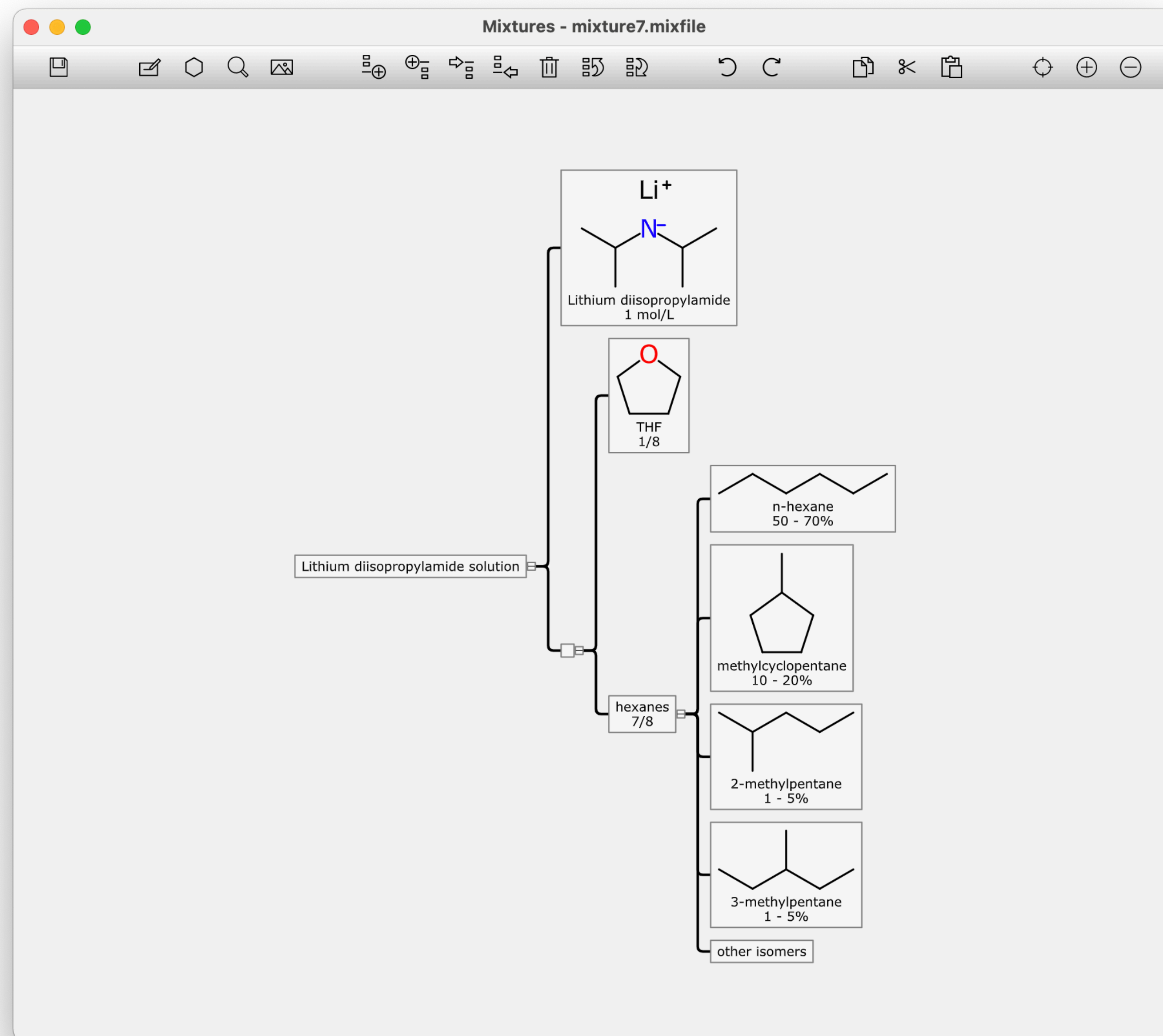
- ▶ no **community** without **data**
- ▶ no **data** without **tools**
- ▶ no **tools** without **community**



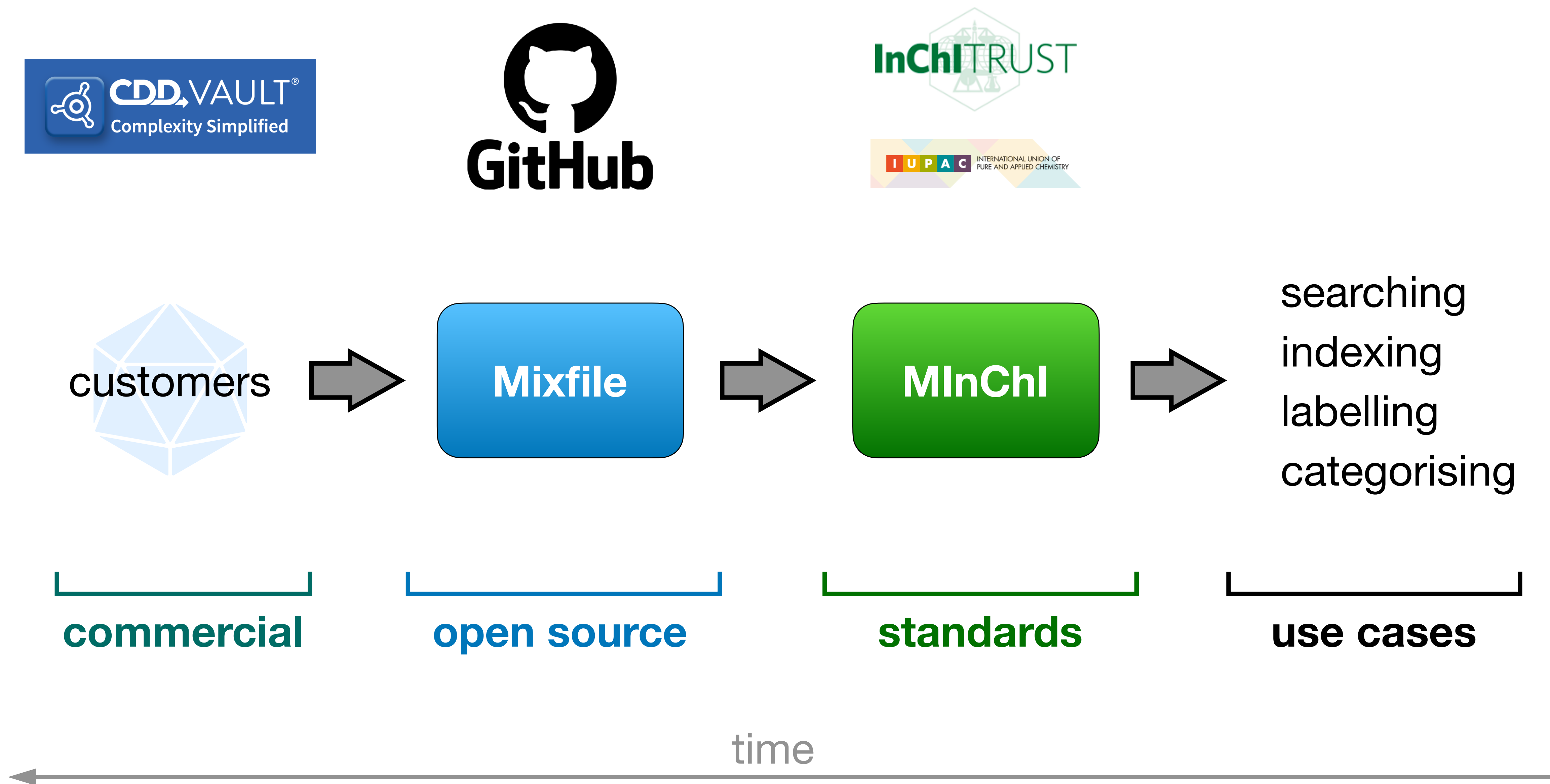
have to build simultaneously

Tools

- ❖ 2018: NIH SBIR grant awarded to *Collaborative Drug Discovery*
- ❖ First step: open source mixture editor and software libraries
- ❖ Coded in TypeScript, cross-compiled to JavaScript, for:
 - ▶ **web**
 - ▶ **desktop** (via Electron)
 - ▶ **server** (via NodeJS)
- ❖ Operates on "**Mixfile**" which is ELN-like, JSON-based, mixture analog of Molfile



Upstream/Downstream



ELN integration

- ❖ Mixture creation is also part of a commercial product
- ❖ Scientists use the ELN already...
- ❖ ... machine readable data is a side effect of normal use
- ❖ First class citizens:
 - ▶ molecules
 - ▶ reactions
 - ▶ mixtures

collaborativedrug.com

CDD.VAULT · Mixture Vault Help · Log out

Explore Data **ELN** Import Data Reports Settings Full-Access User

How to make dishwashing liquid

ID: 973478671

Project: Mixture Registration

Normal Text 🔍 **B** / U **S** x^2 x_2 ☰ ☰ ☑ 🔗 📅 🕒 🏠 📄 🛡 🍽 Saved

water (80.7 w/w%), sodium dodecylbenzene sulphonate (60%), water (40%) (11.7 w/w%), sodium lauryl ether sulphate (27%), water (73%) (6.6 w/w%), sodium chloride, coconut diethanolamide (1 w/w%), perfume, dye, preservative

```
graph LR; DL[dishwashing liquid] --- H2O1[H2O water 80.7 w/w%]; DL --- 11.7[11.7 w/w%]; DL --- 6.6[6.6 w/w%]; DL --- NaCl[Na+ Cl- sodium chloride]; DL --- CD[coconut diethanolamide 1 w/w%]; DL --- PD[perfume, dye]; DL --- PR[preservative]; 11.7 --- H2O2[H2O water 40%]; 11.7 --- SDS[sodium dodecylbenzene sulphonate 60%]; 6.6 --- H2O3[H2O water 73%]; 6.6 --- SLES[sodium lauryl ether sulphate 27%];
```

Data

- ❖ Bootstrap from text sources
- ❖ Proprietary deep learning algorithm
- ❖ ~30K mixtures marked up, public release

Name:

Nitric acid, >90% solution in water

Confidence in result: *high*

Auto-generated name: *Nitric acid (>90%) in water*

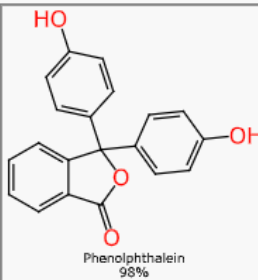
- ❖ Substantial body of exemplars, and **upstream test data** for MInChI generation
- ❖ Can rapidly markup inventories and vendor catalogs
- ❖ Integrated into software-as-a-service products

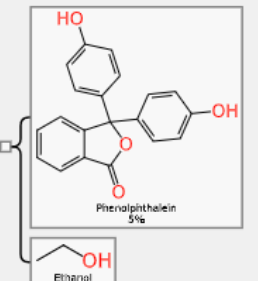
Support resources

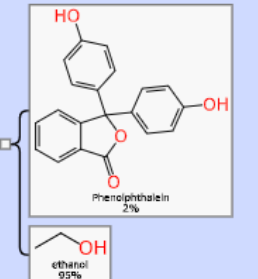
- Looking up known content speeds up data creation...

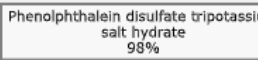
Mixture

phenolph Edit Parse

 Phenolphthalein, 98%


 Phenolphthalein solution 5% in Ethanol

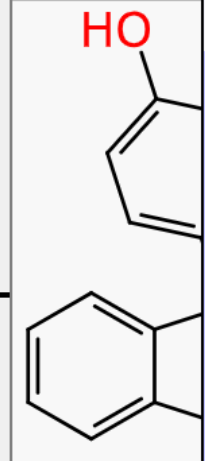
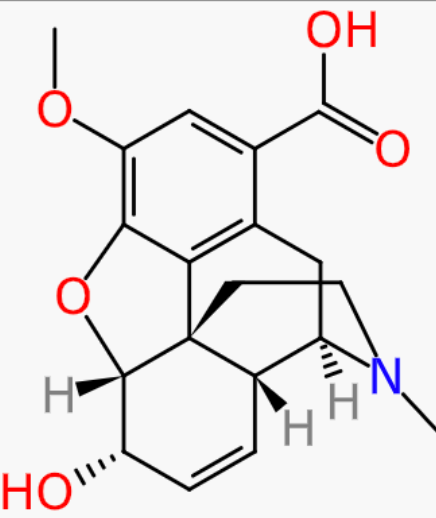
 Phenolphthalein, 2% soln. in 95% ethanol

 Phenolphthalein disulfate tripotassium salt hydrate, 98%

- INCI and UNII collections available to quick search

Mixture



coconut

COCONUT
UNII 3RT3536DHY

COCONUT OIL
UNII Q9L0O73W7L
CASRN 8001-31-8
ECNO 232-282-8

COCONUT CRAB
UNII 35NE41P38E

COCONUT ACID
UNII 40U37V505D
CASRN 61788-47-4
PubChemCID 57180994
ECNO 262-978-7

COCONUT ACID
COSING 75280
INCI COCONUT ACID
CASRN 61788-47-4
ECNO 262-978-7

COCONUT SHELL
UNII 704218UHJ4

COCONUT ALCOHOL
COSING 75281
INCI COCONUT ALCOHOL
CASRN 68425-37-6
ECNO 270-351-4

COCONUT ALKANES
COSING 86954
INCI COCONUT ALKANES

COCONUT FLOWER SUGAR
COSING 94994
INCI COCONUT FLOWER SUGAR

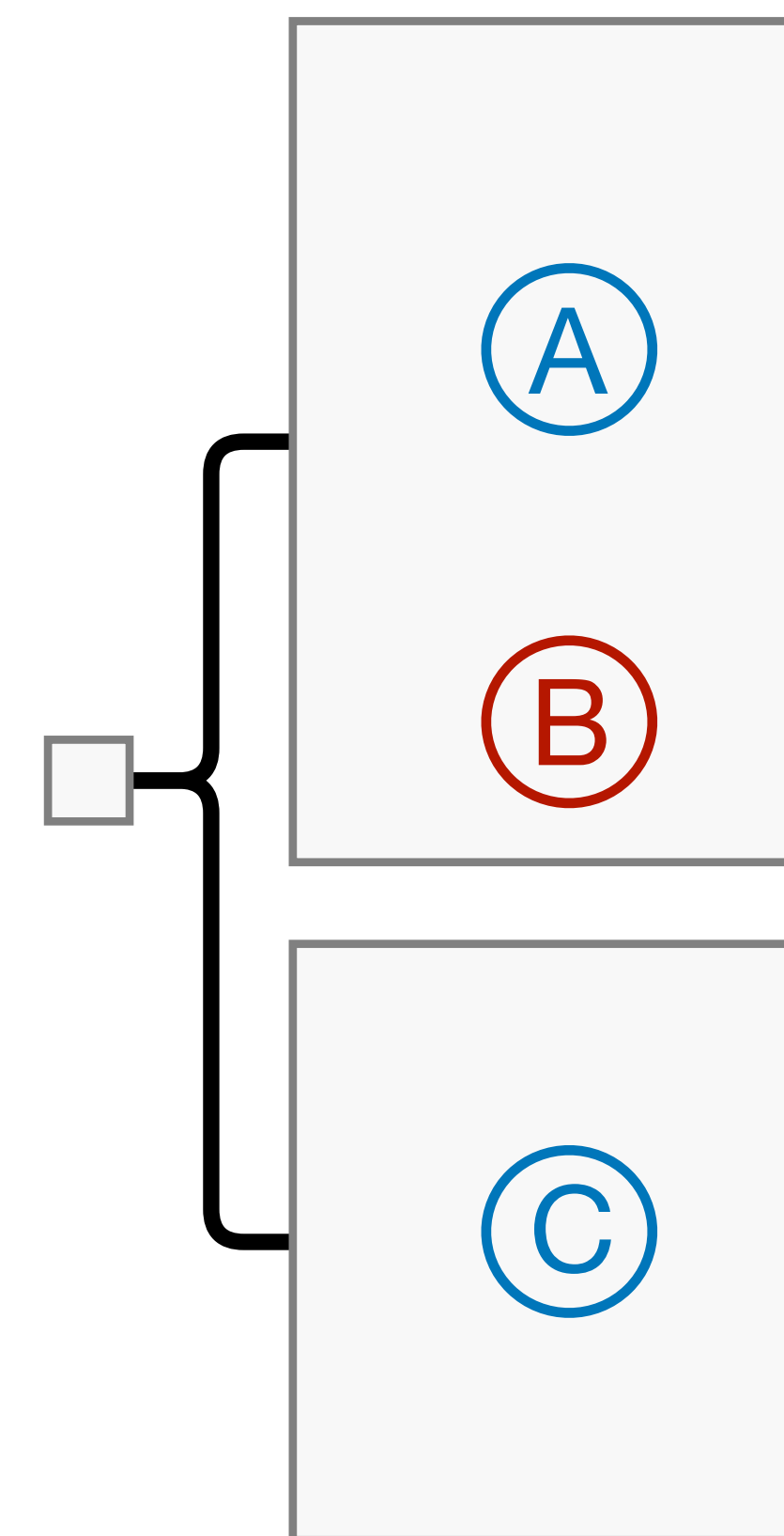
COCONUT OIL MIPA AMIDES
COSING 91913
INCI COCONUT OIL MIPA AMIDES
CASRN 68333-82-4

Submit or cancel

Demidigital

❖ Partially marked up data can be upgraded by document-wide options...

	MATERIAL	QUANTITY	MATERIAL
	A	B	C
1	<u>MOLECULE</u>	<u>CONC</u>	<u>SOLVENT</u>
2	copper sulfate	0.01 mol/L	water
3	hydrochloric acid	0.02 M	aqueous
4	perchloric acid	60%	
5	potassium sodium tartrate	30 w/v%	H2O
6	tellurium (IV) bromide	99.9%	n/a
7	Boc2O		dioxane
8	dithiothreitol & alpha-thioglycerol	2:3	
9	acetic acid	5% (v/v)	
10	buffer solution (ph 7.0)		



❖ Currently in design phase

Community

- ❖ Creating technology is easy, getting everyone to use it is hard...
- ❖ Requires concurrent strategies



- ❖ Endorsement by respected standards organisations is a good start
- ❖ InChI derivatives have enthusiastic champions (that's us!)

Engagement

- ❖ **Code it up:** using MInChI notation is easy enough
- ❖ **Got use cases?** Let us know
- ❖ **Spread the word:** data resources need to be digitised

Further viewing

❖ Peer reviewed literature:

Research article | [Open Access](#) | Published: 23 May 2019

Capturing mixture composition: an open machine-readable format for representing mixed substances

[Alex M. Clark](#) ✉, [Leah R. McEwen](#), [Peter Gedeck](#) & [Barry A. Bunin](#)

[Journal of Cheminformatics](#) 11, Article number: 33 (2019) | [Cite this article](#)

❖ Webinars:

2019: www.youtube.com/watch?v=PcAJ4HoRnFU

Capturing mixtures — bringing informatics to the world of practical chemistry

2020: www.youtube.com/watch?v=aSQEVKKnrWw (starts at 4:13:00)

Mixtures: informatics for formulations and consumer products

2021: www.youtube.com/watch?v=0ILc0owuEzQ (starts at 1:05:00)

Mixtures as first class citizens in the realm of informatics

Further work

- ❖ Finalising **MInChI 1.0** specification, reference implementation, validation
- ❖ MInChI needs to extend to less well defined chemical entities
 - variable structures
 - polymers
 - biologics
 - nanomaterials
 - reaction products
- ❖ Properties and metadata: ontology based / IUPAC Gold Book
- ❖ Implementation at scale: registration systems

MInChI Open Meeting: 20 April 11am-1pm (EDT)

Questions?

❖ Contact:

- ▶ Leah R. McEwen irm1@cornell.edu (Cornell University, IUPAC/InChI Trust)
- ▶ Alex M. Clark alex@collaborativedrug.com (Collaborative Drug Discovery)

❖ Thanks to the MInChI team