Mixtures

informatics for formulations and consumer products

Leah McEwen & Alex M. Clark

Who

- InChI Trust / IUPAC
  - https://www.inchi-trust.org
  - Mixtures InChI notation

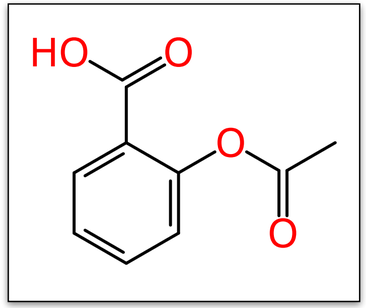
- Collaborative Drug Discovery
  - https://collaborativedrug.com
  - Mixfiles & tools

Introduction

✤ Cheminformatics has 40 years of practice representing abstract molecules

✤ Very successful applications for pharmaceutical drug discovery

Molfile




```
CC(=O)OC1=CC=CC=C1C(=O)O
```

SMILES

```
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2
-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
```

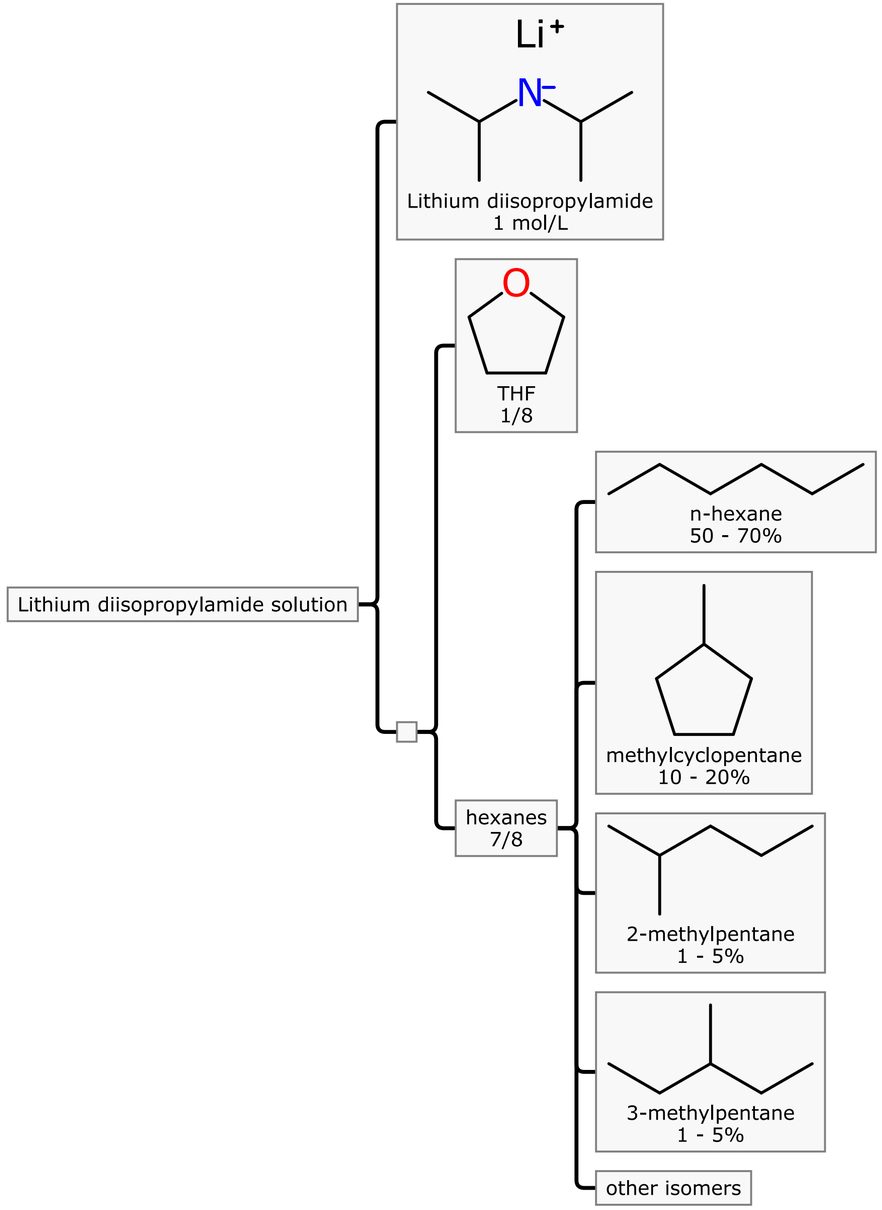
InChI

✤ But the reality of chemicals in the lab is that

- nothing is ever completely pure
- most activities involve explicitly mixing chemicals

✤ Lack of standard way to describe mixtures (for informatics)

Mixfile/MInChI

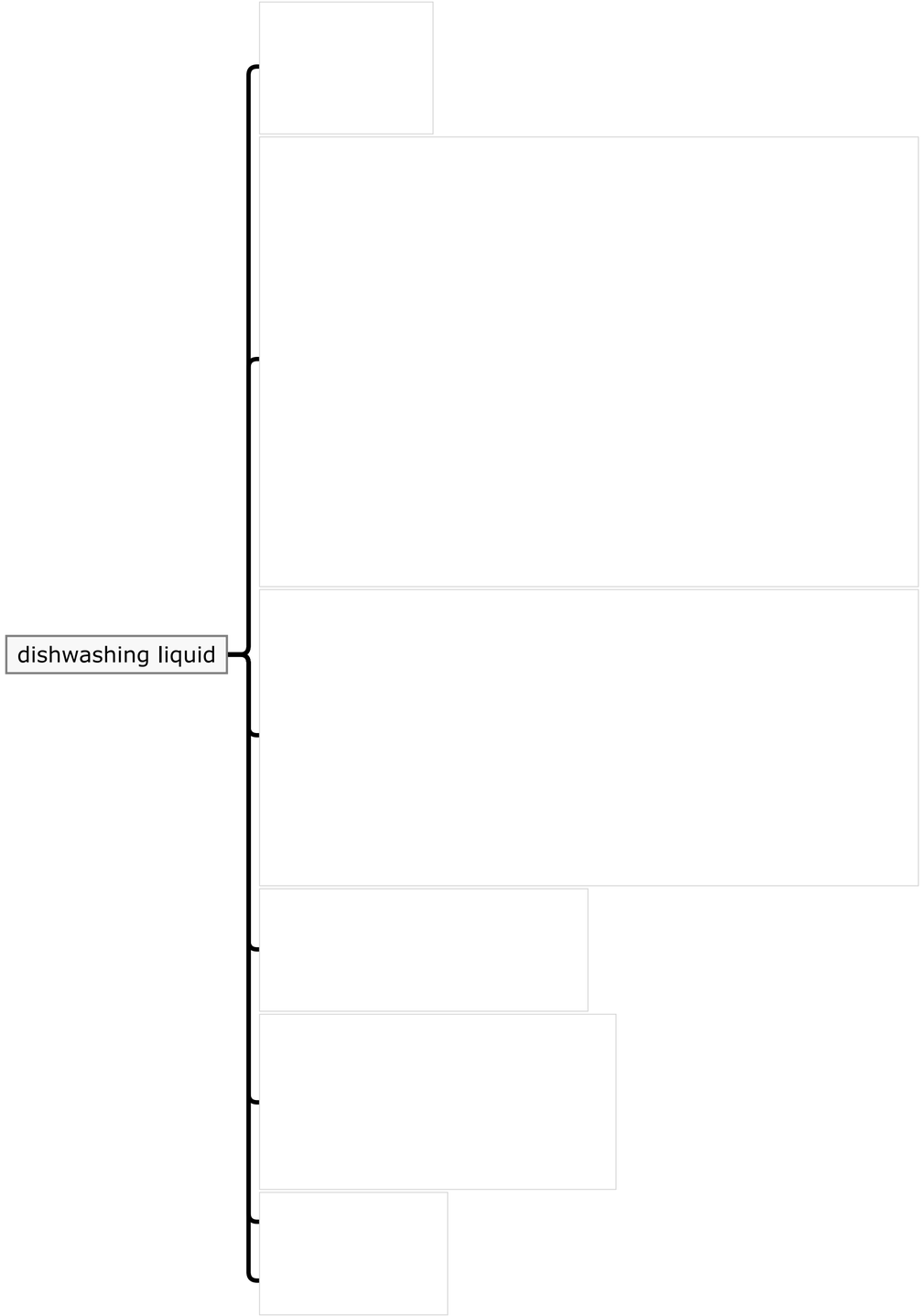


- Format needs to be:
  - hierarchical
  - embed structures when possible
  - include concentration information
  - tolerate uncertainty
- More verbose ELN-friendly form is **Mixfile**
- Concise form with canonical components is **MInChI** (*mixtures* InChI)

```
MInChI=0.00.1S/C4H8O/c1-2-4-5-3-1/h1-4H2&C6H12/
c1-6-4-2-3-5-6/h6H,2-5H2,1H3&C6H14/c1-3-5-6-4-2/
h3-6H2,1-2H3&C6H14/c1-4-5-6(2)3/h6H,4-5H2,1-3H3&C6H14/
c1-4-6(3)5-2/h6H,4-5H2,1-3H3&C6H14N.Li/c1-5(2)7-6(3)4;/
h5-6H,1-4H3;/q-1;+1/n{6&{1&{3&2&4&5}}}/
g{1mr0&{1vp0&{5:7pp1&1:2pp1&1:5pp0&1:5pp0}7vp0}}
```

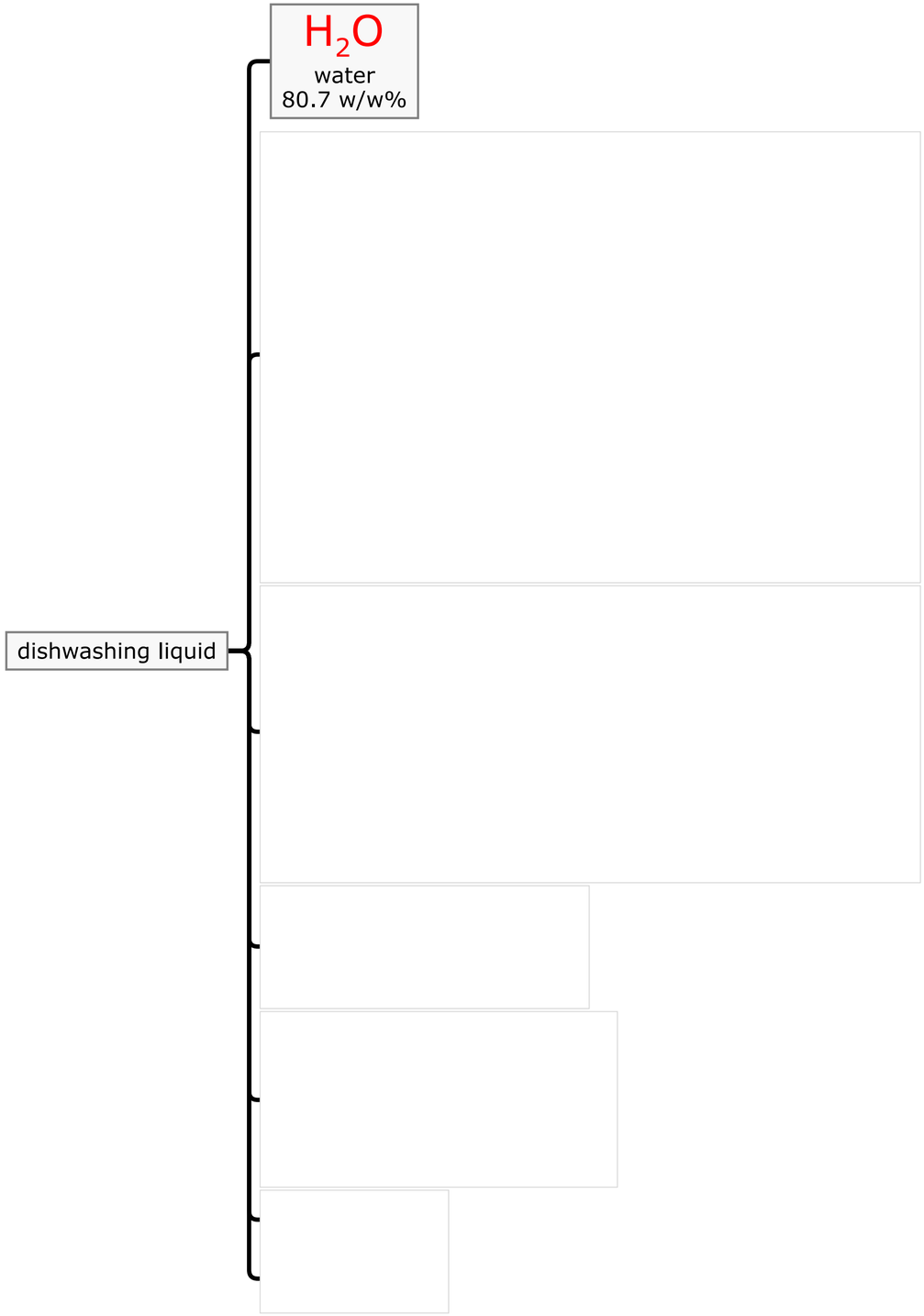
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



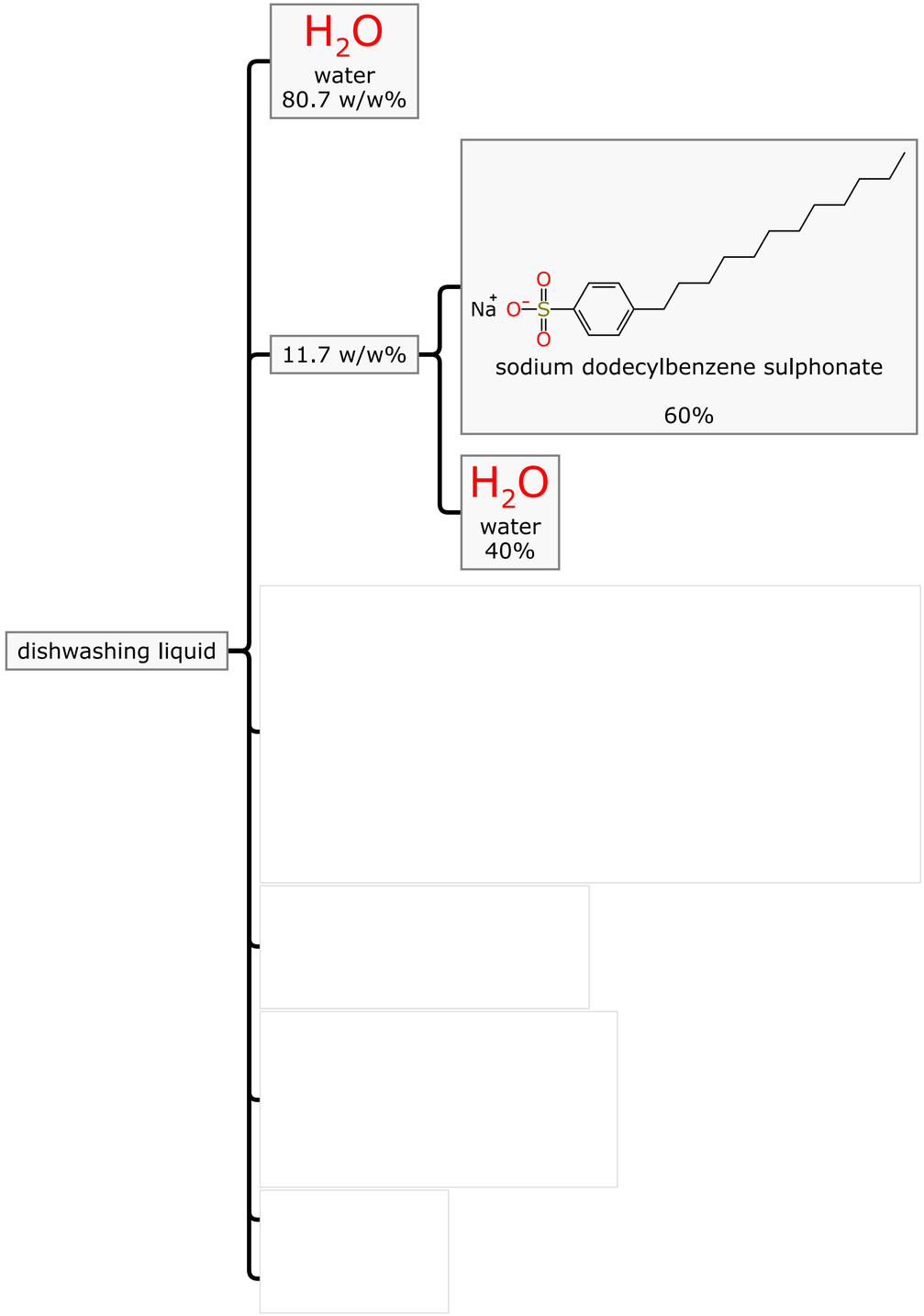
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



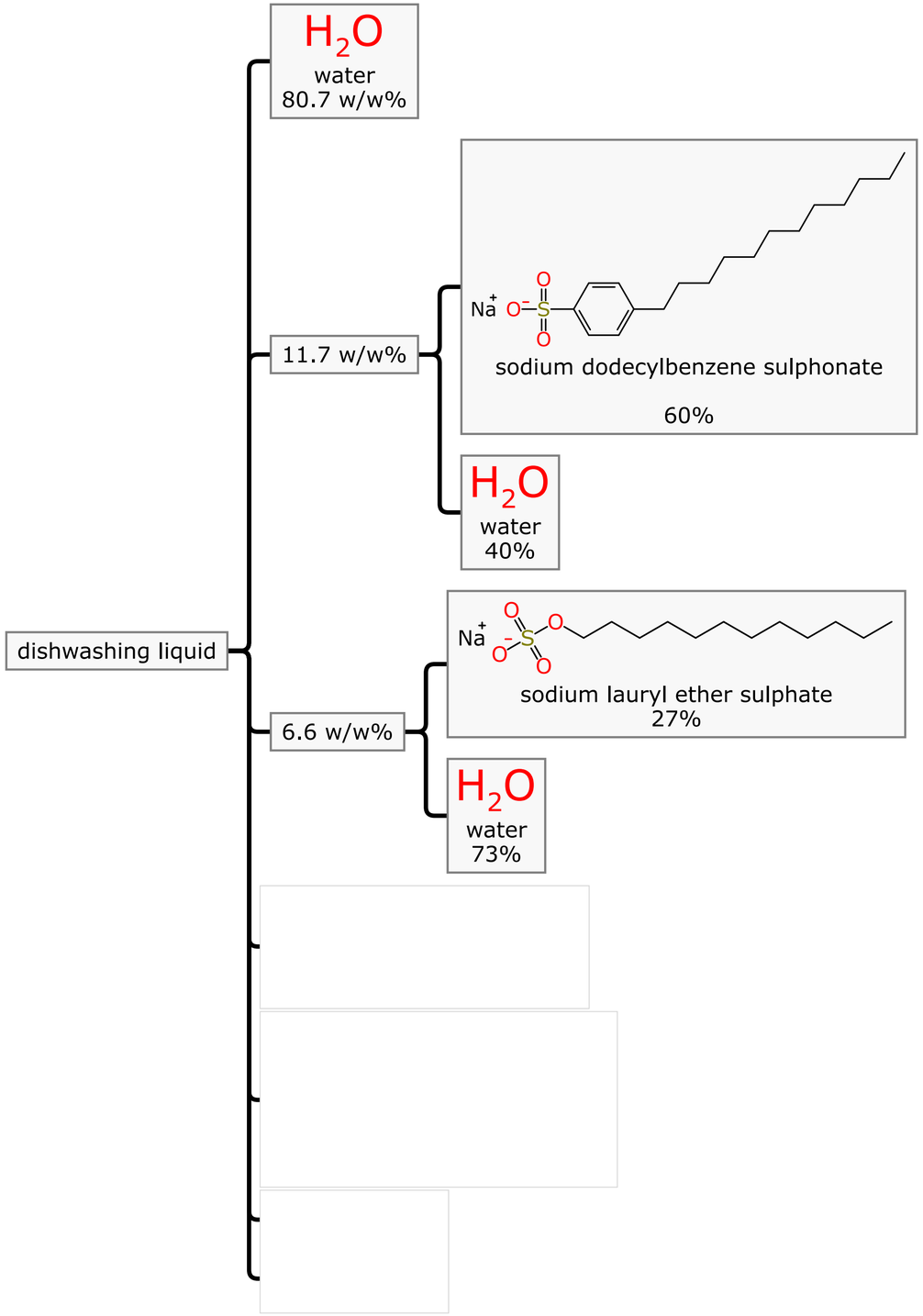
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



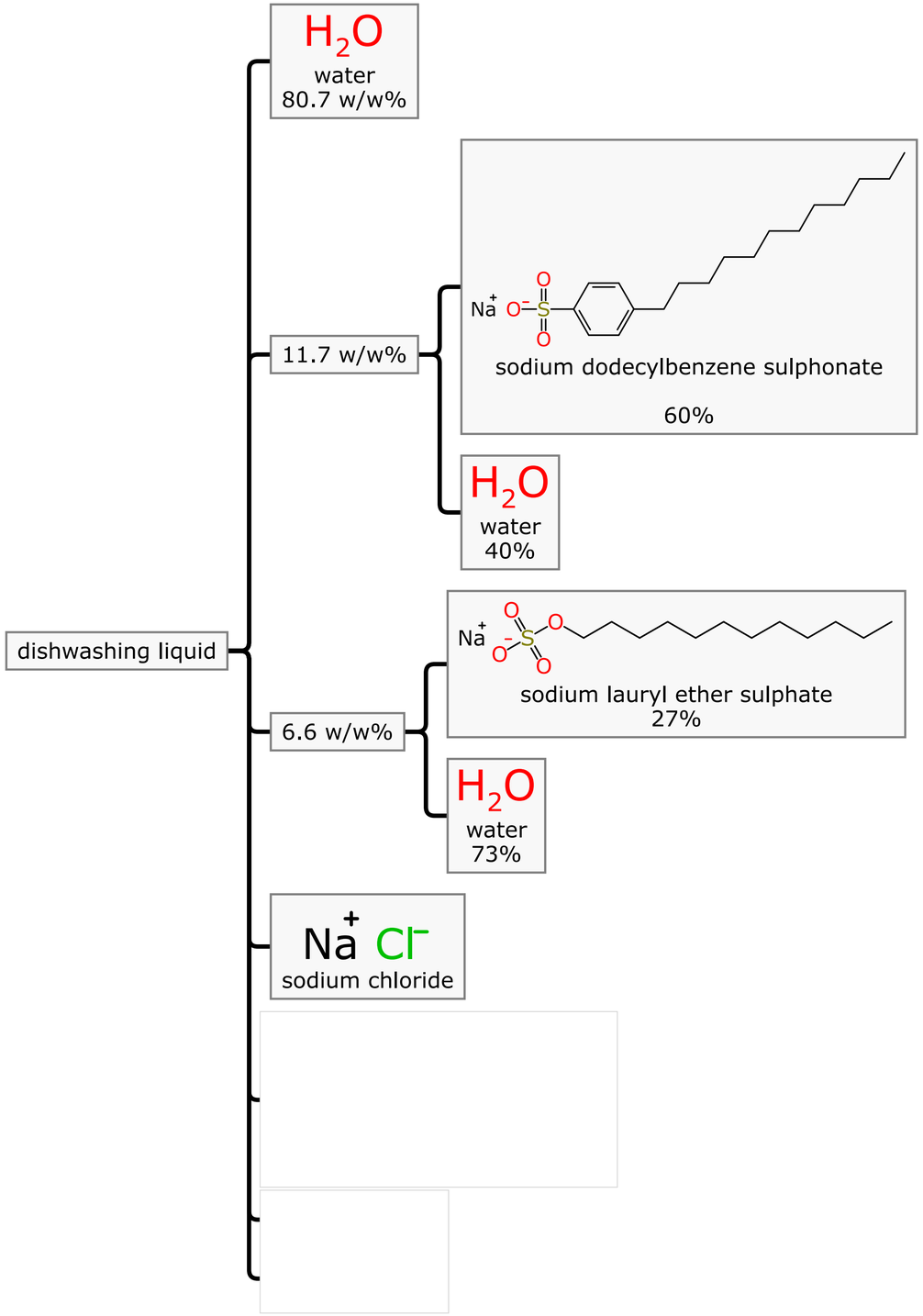
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



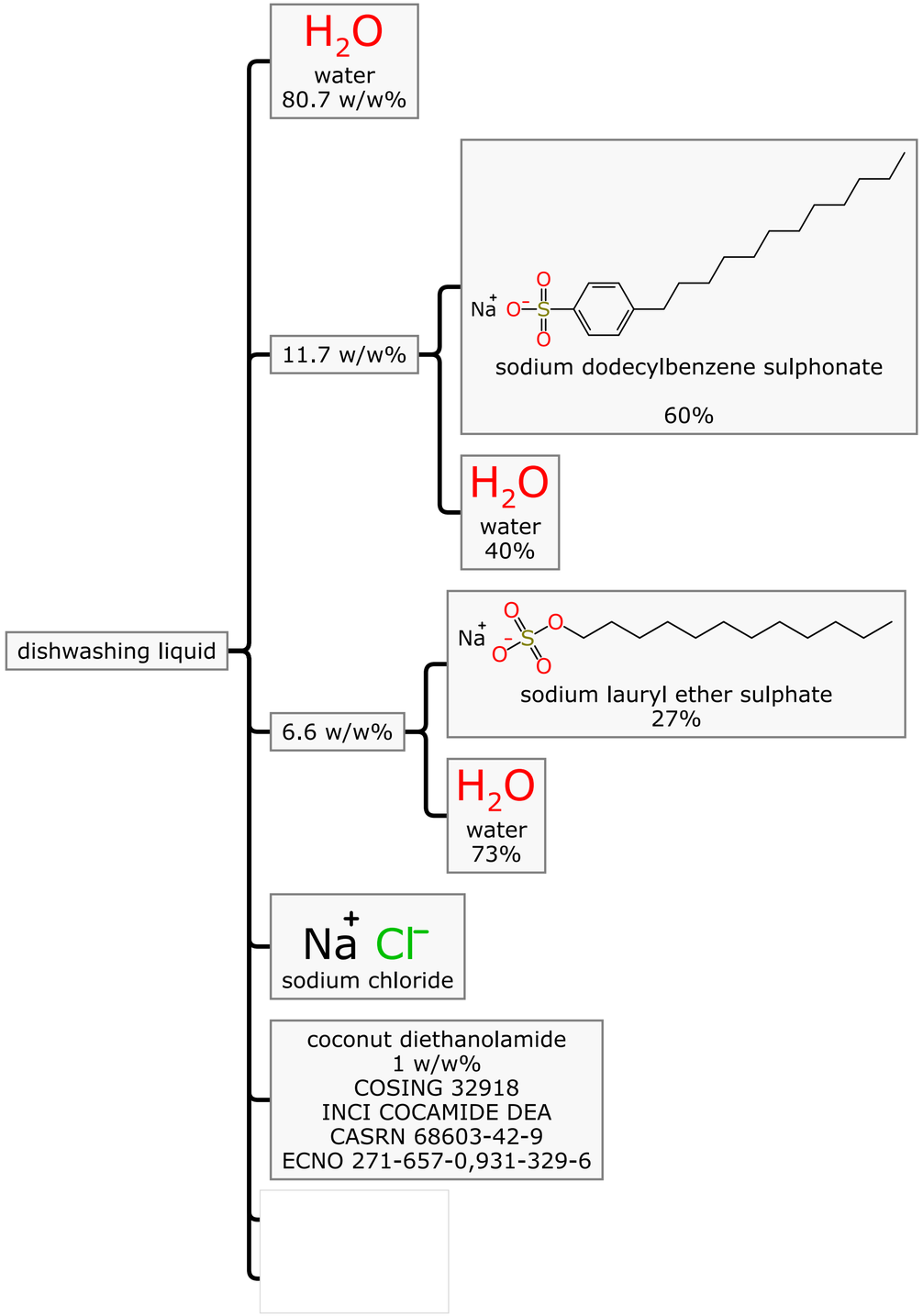
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



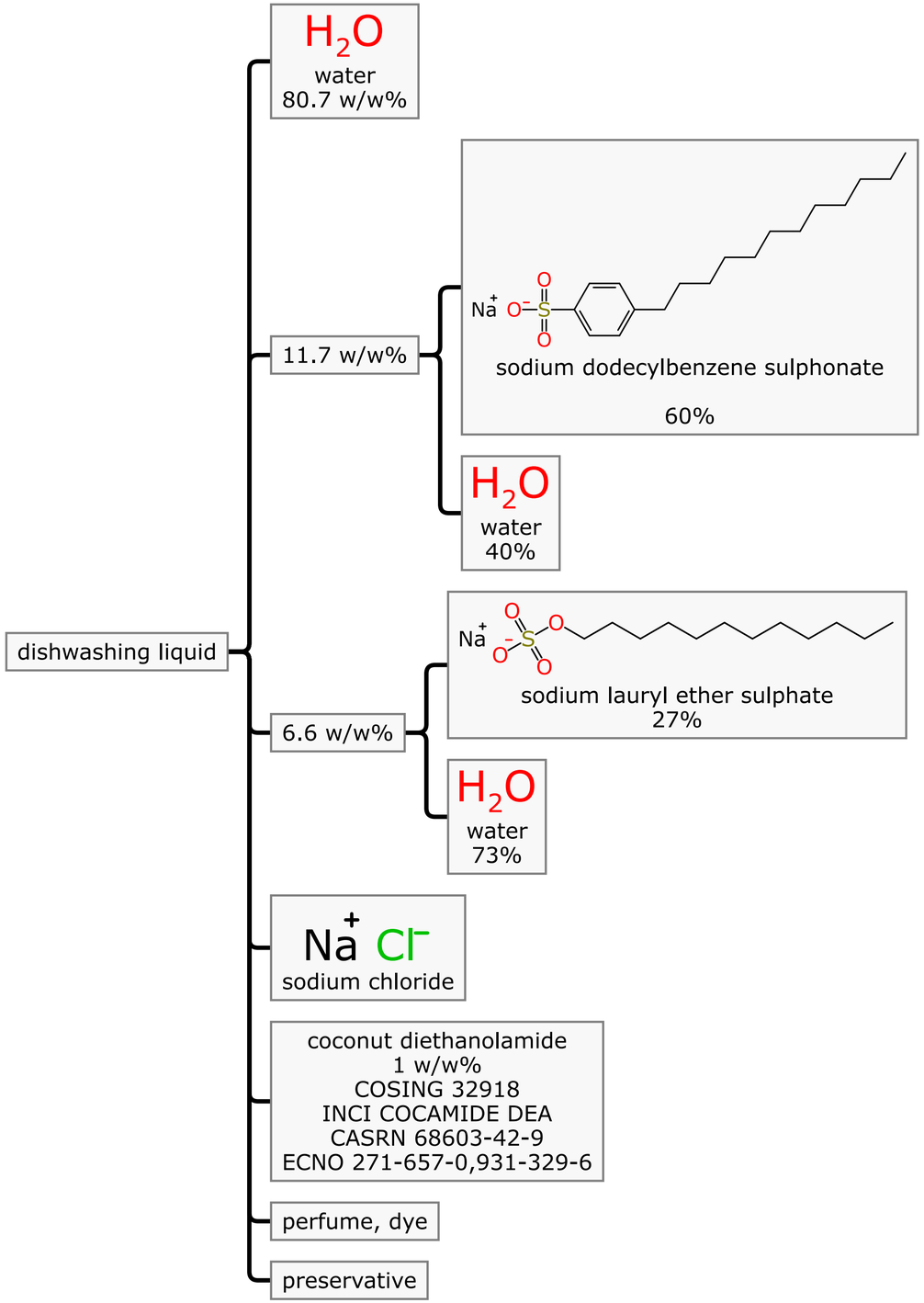
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



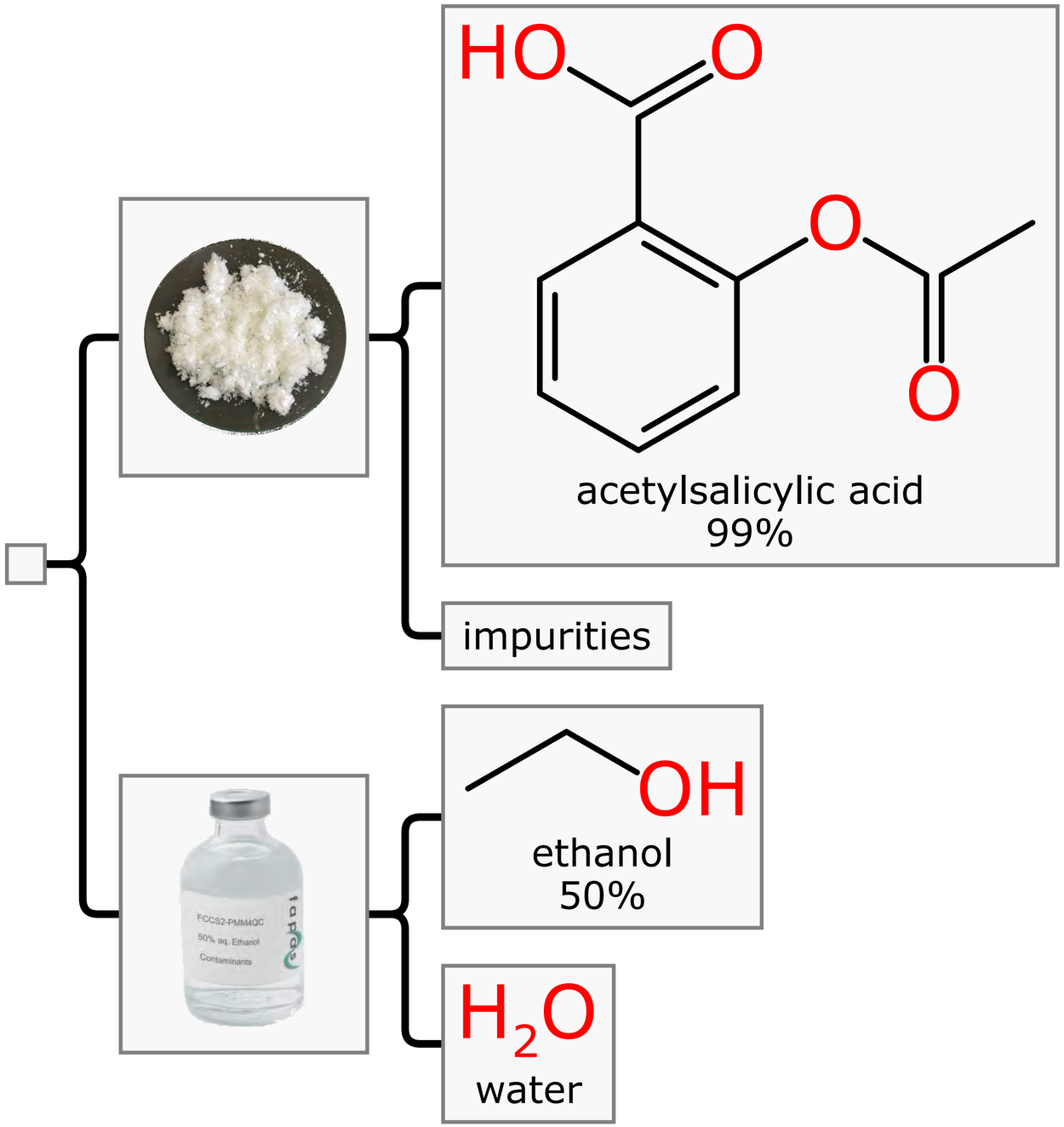
Formulation Example

- Many consumer products are well described from a chemical perspective
- Some components are more easily defined than others
- When structure is not available, can use external identifiers
- Hierarchy encodes information about the design of the product
- Concentrations can be expressed with uncertainties



Design of Mixtures

- Each branch is a *thing*
- Each leaf is a *concept*
- Layout can correspond to how the mixture is formulated



Knowledge Capture

- Capture what we know about the mixture: and nothing more
  - ideally each leaf node has well defined structure & precise concentration
  - the closer we get to this, the more analysis we can do
- Concentrations often unknown, vague, or implied
- Structure(s) can be hard to pin down...
  - ... not always a single, well defined, easy to draw molecule

Structures by External Definition

- Sometimes have to resort to describing a component by method of preparation, means of extraction, measured properties, etc.

- External database identifiers can be useful:

  - **CASRN**: Chemical Abstracts literature extraction

  - **INCI**: International Nomenclature of Cosmetics Ingredients

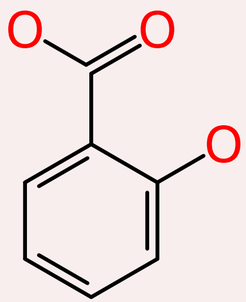
  - **UNII**: Food & Drug Administration database

- Database identifiers are not ideal for machine readability, but they can be used to establish equivalence:

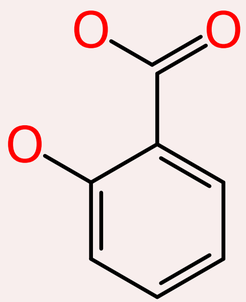
HYDROGENATED COCONUT OIL
COSING 34340
INCI HYDROGENATED COCONUT OIL
CASRN 84836-98-6
ECNO 284-283-8

Coconut oil (hydrogenated)
INCI HYDROGENATED COCONUT OIL

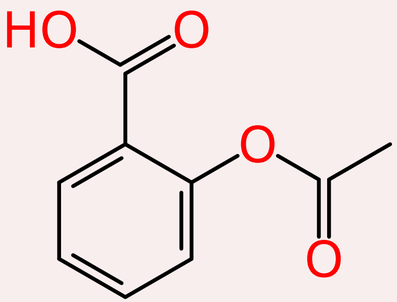
Comparisons with Structures



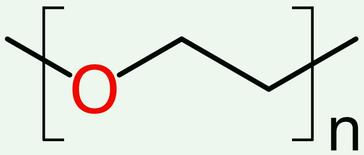
≡



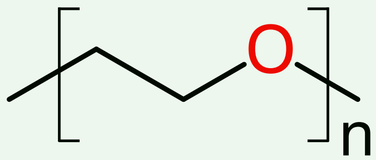
substructure of



`InChI=1S/C7H6O3/c8-6-4-2-1-3-5(6)7(9)10/h1-4,8H,(H,9,10)`



≅

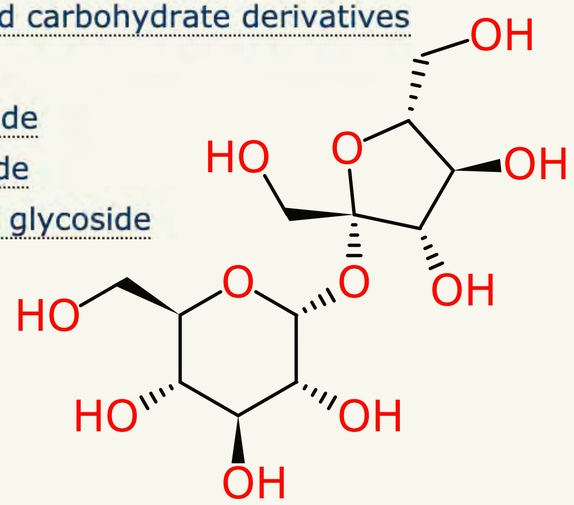


MW 300-500 MW 400-700

$Y_{1.2}Ba_{0.8}CuO_4$ ≅ $YBa_2Cu_3O_{7-\delta}$

- CHEBI:36962 organochalcogen compound
  - CHEBI:36963 organooxygen compound
    - CHEBI:78616 carbohydrates and carbohydrate derivatives
      - CHEBI:16646 carbohydrate
        - CHEBI:50699 oligosaccharide
          - CHEBI:36233 disaccharide
            - CHEBI:24407 glycosyl glycoside

≅



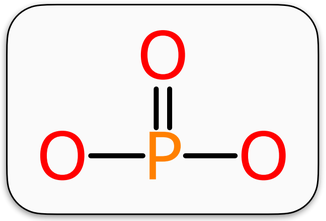
Search Queries

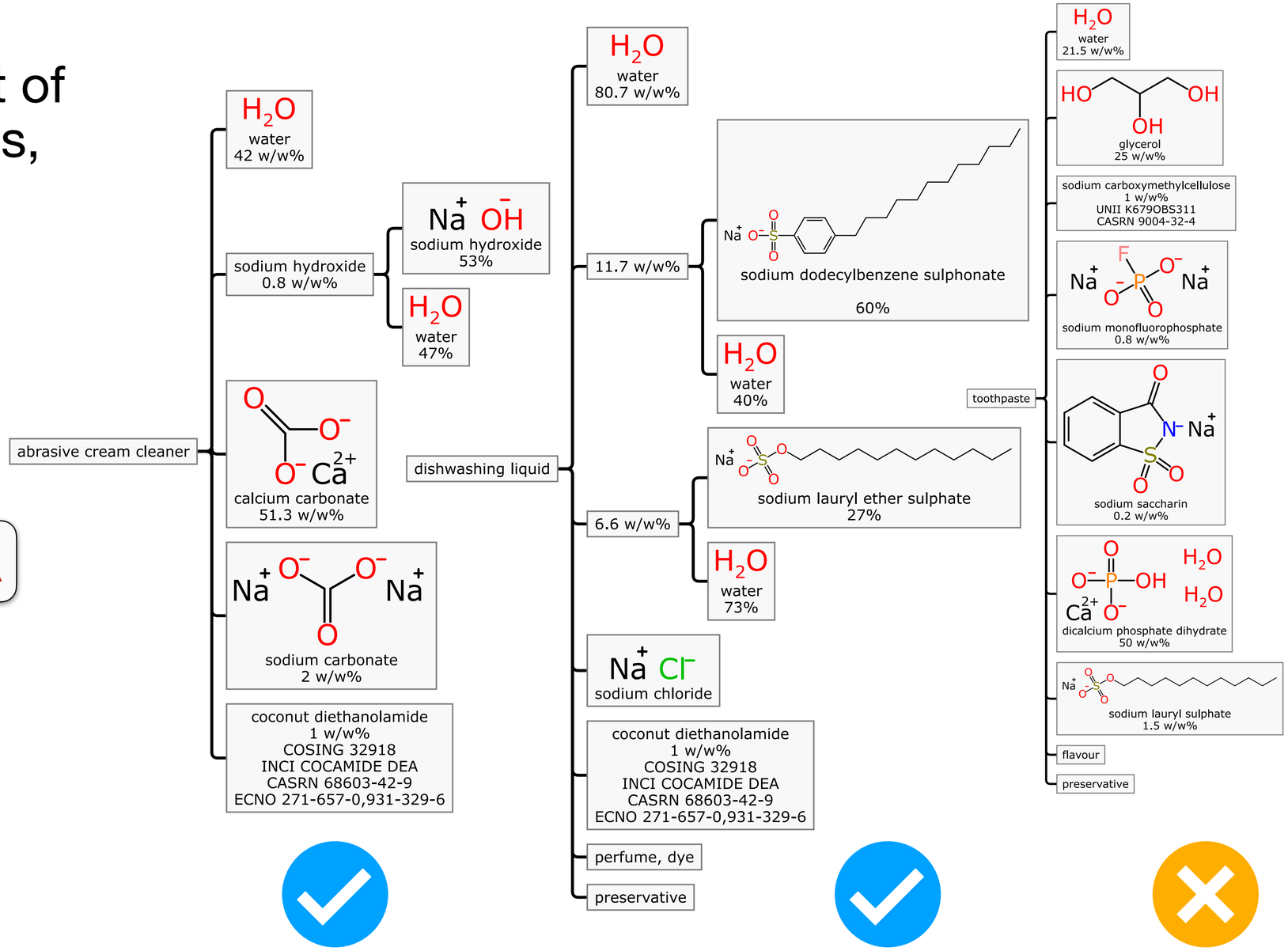
- Looking for a certain subset of external cleaning surfactants, phosphate-free

has H_2O >40%

has **INCI**: COCAMIDE DEA

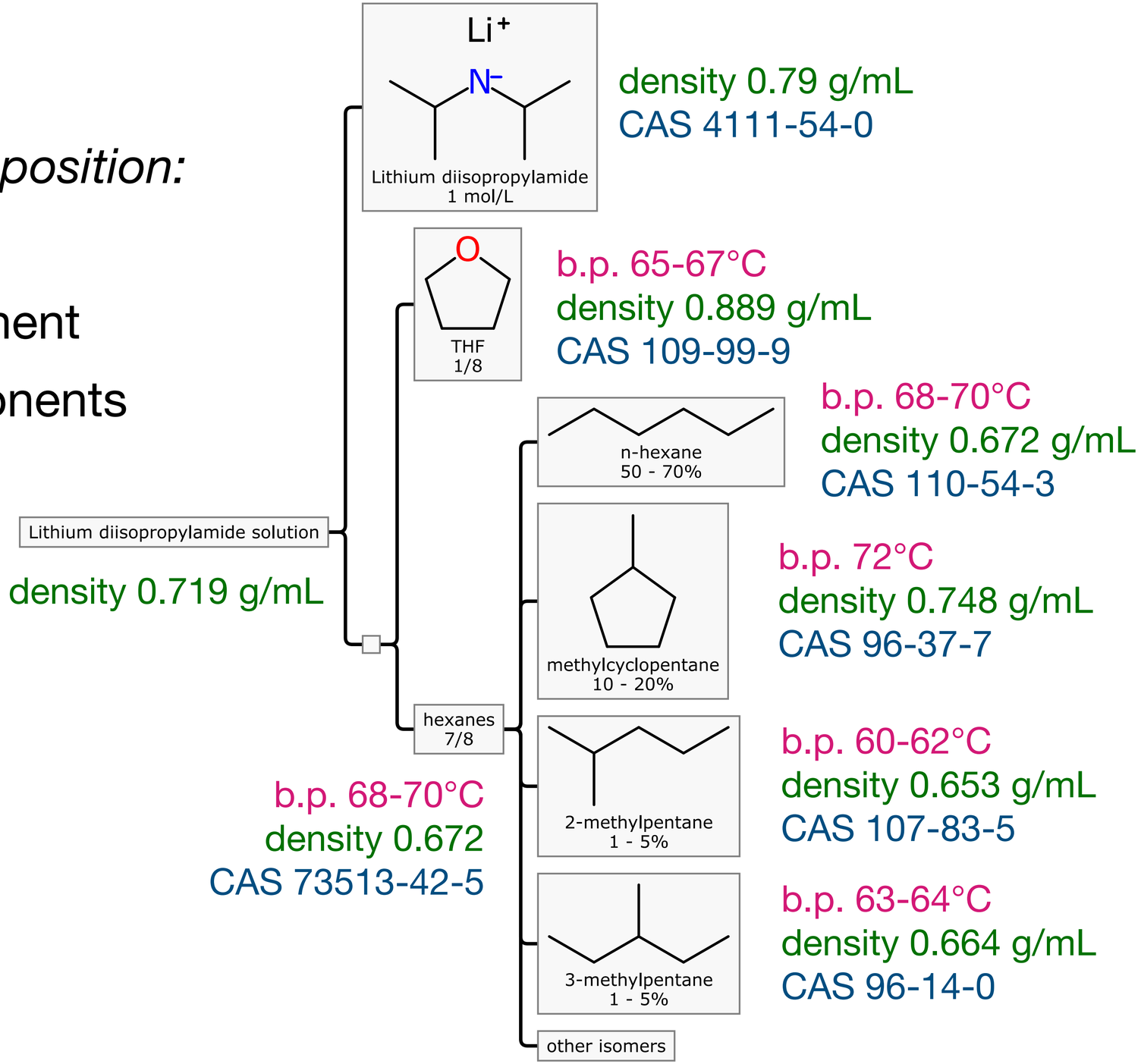
has not substructure



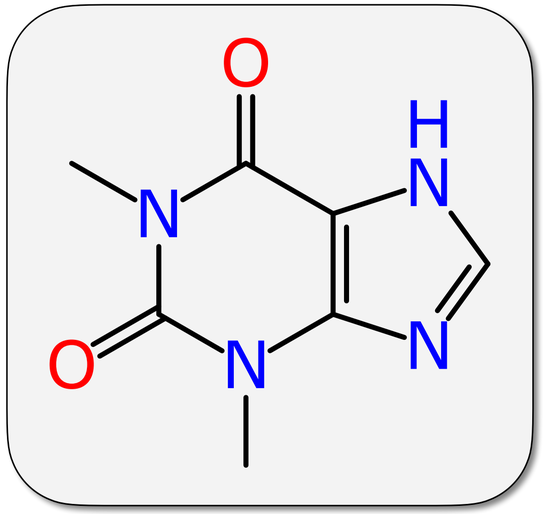


Properties

- Metadata is attached to a *position:*
  - root = the whole thing
  - leaf = individual component
  - branch = several components



Informatics Example



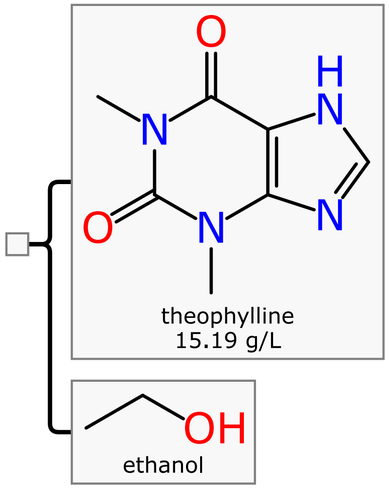
theophylline

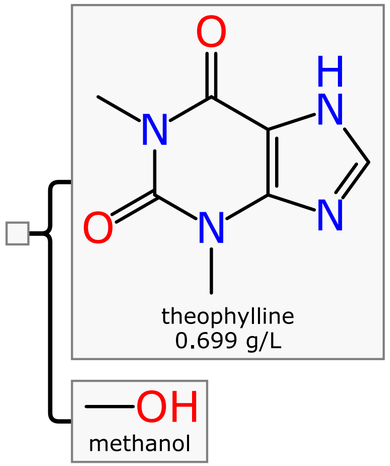
nasal anti-inflammatory

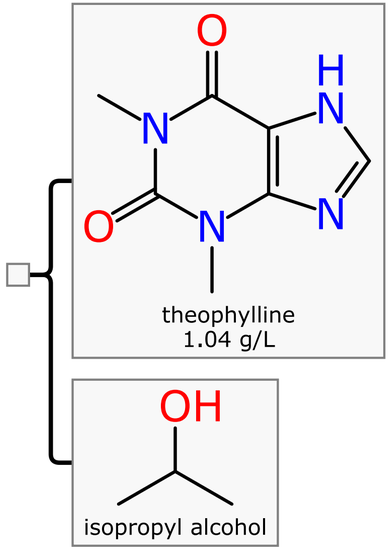
- Solubility of theophylline
- Often delivered in liquid form with mixed solvents: optimising proportion of drug is important
- Consider a scenario where:
  - all data was provided in *Mixtures InChI* form
  - these data exist in openly available repositories
- Query:
  - check that *theophylline* is present and has concentration
  - check that other ingredients are *solvents*
- Consider 4 papers with relevant solubility, published over 20 years...

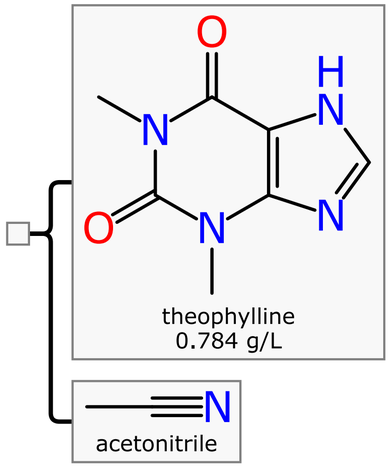
Paper #1

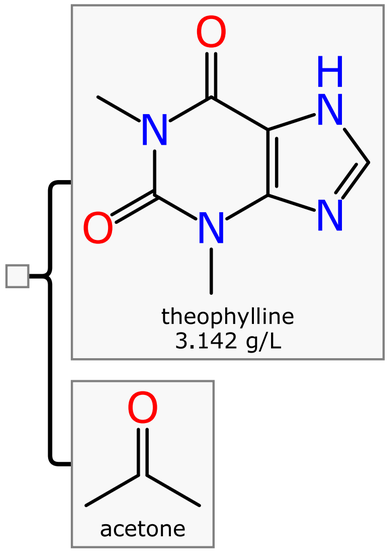
✤ Valizadeh et al, *Adv. Pharm. Bull.* (2011), DOI 10.5681/apb.2011.003

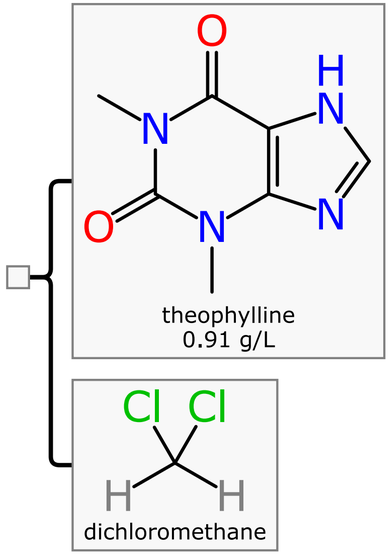






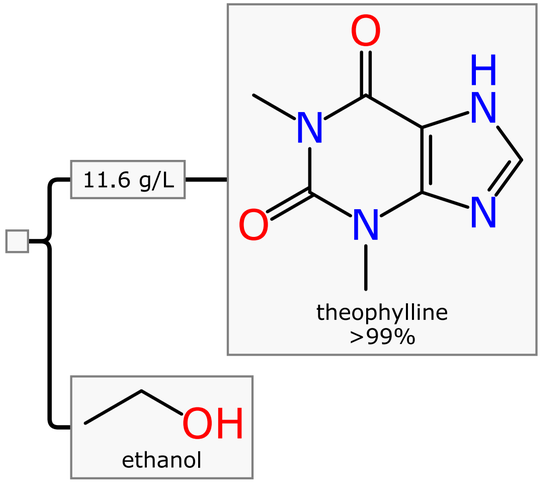


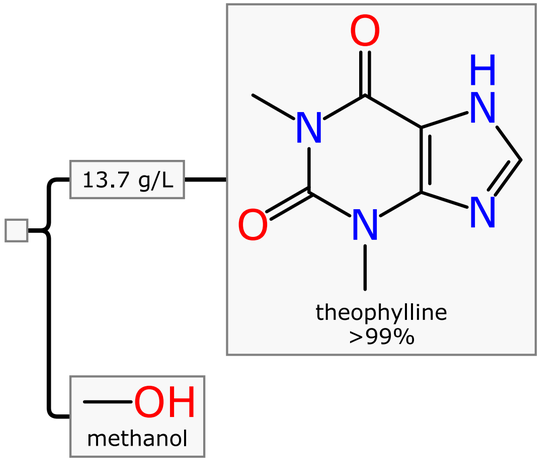


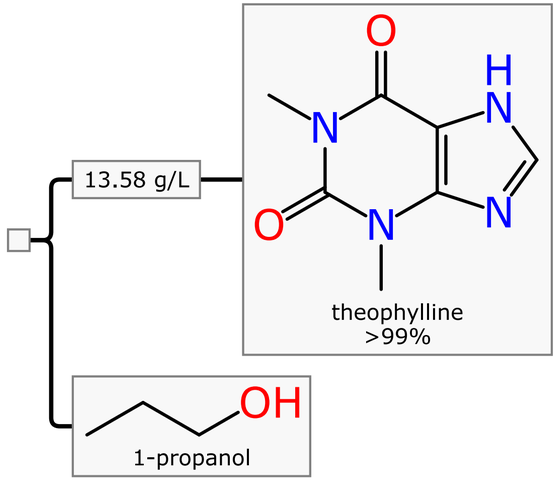


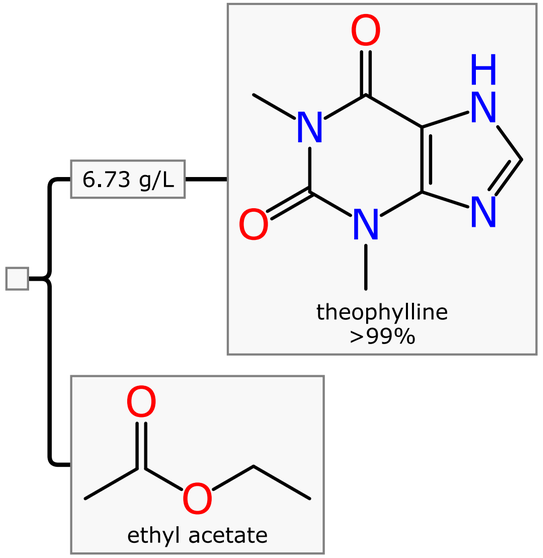
Paper #2

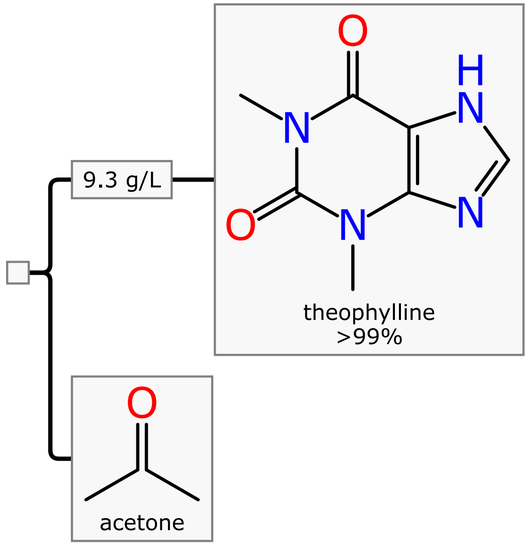
✤ Yan et al, *J. Chem. Eng. Data* (2017), DOI 10.1021/acs.jced.7b00065

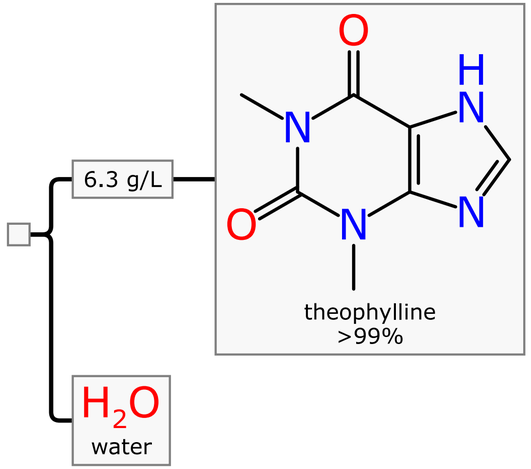






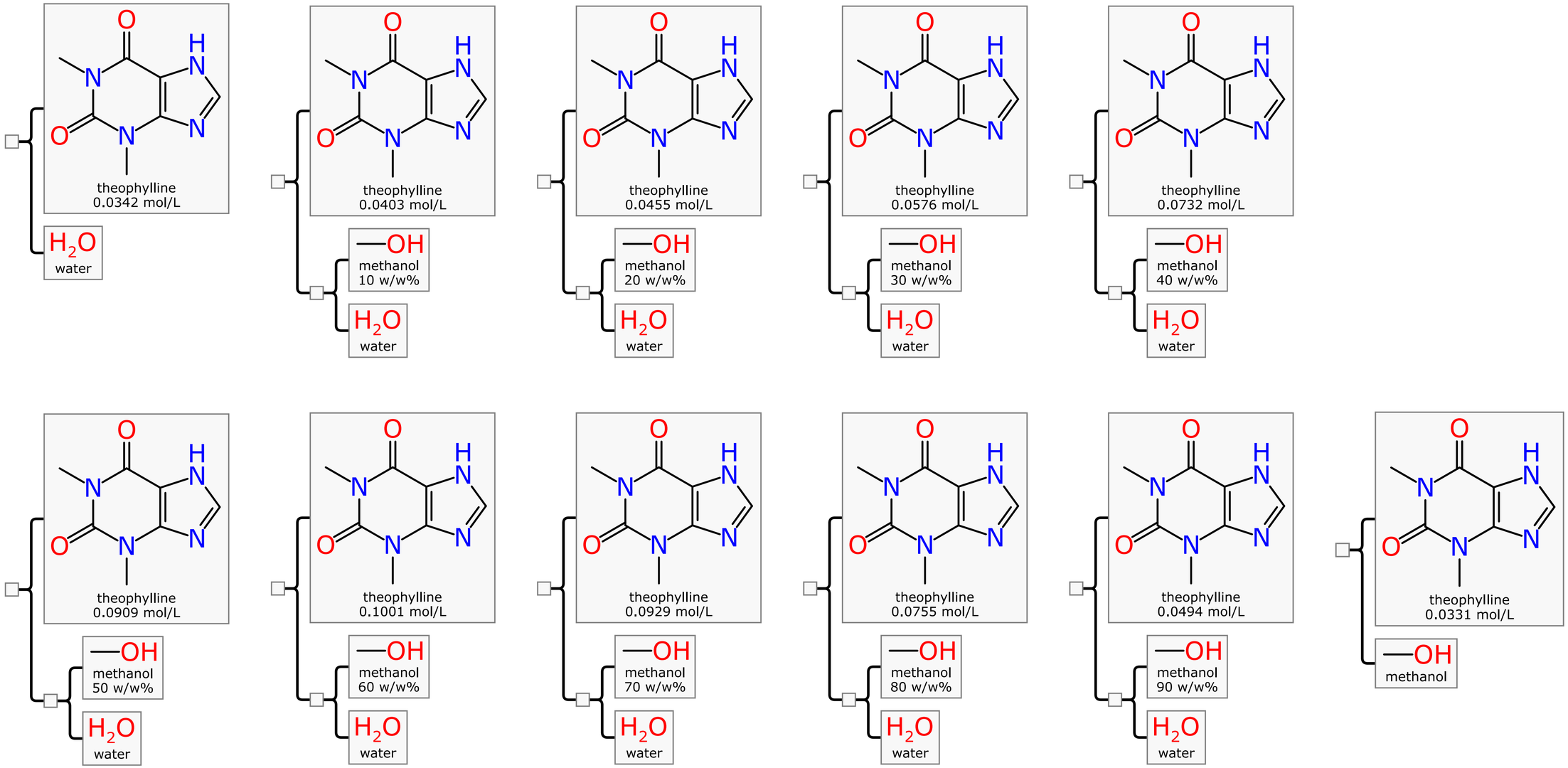






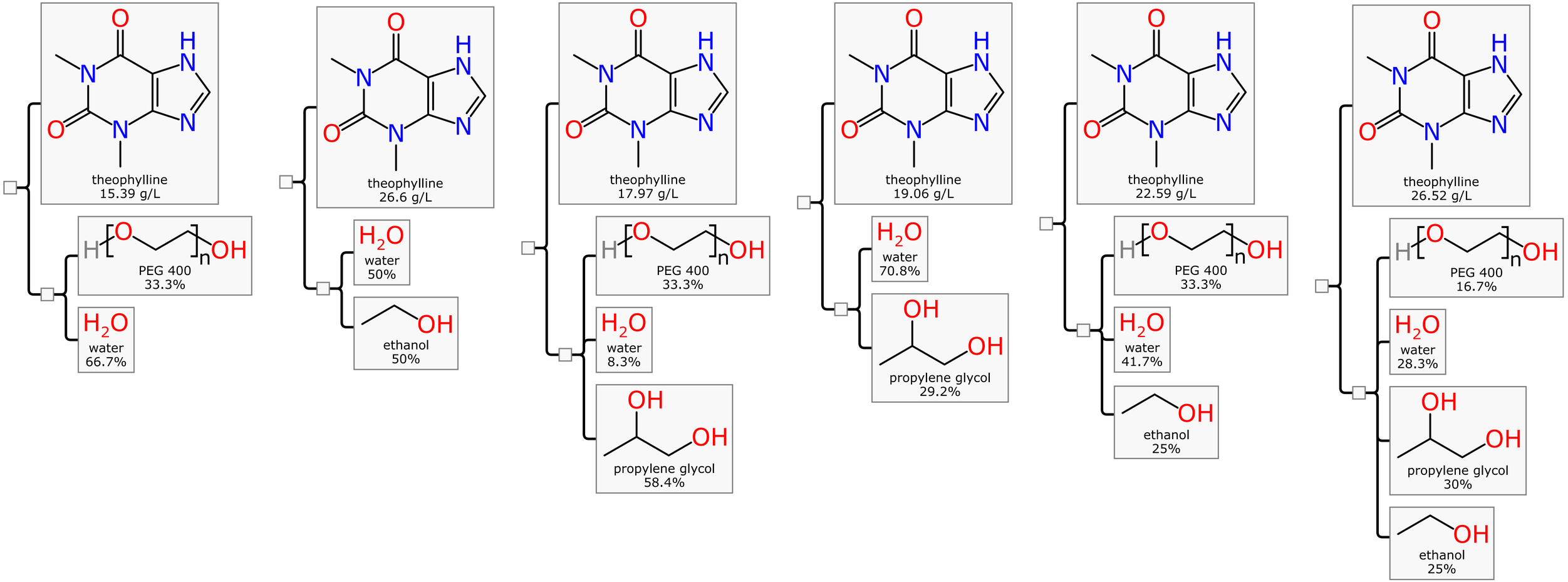
Paper #3

✤ Martínez et al, *J. Solution Chem.* (2017), DOI 10.1007/s10953-017-0666-z



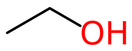
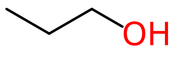
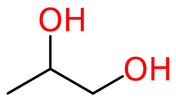
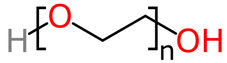
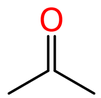
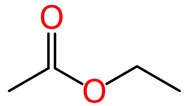
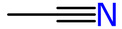
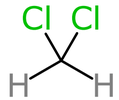
Paper #4

✤ Campisi et al, *J. Pharm. Biomed*. (1998), DOI 10.1016/S0731-7085(98)00175-7



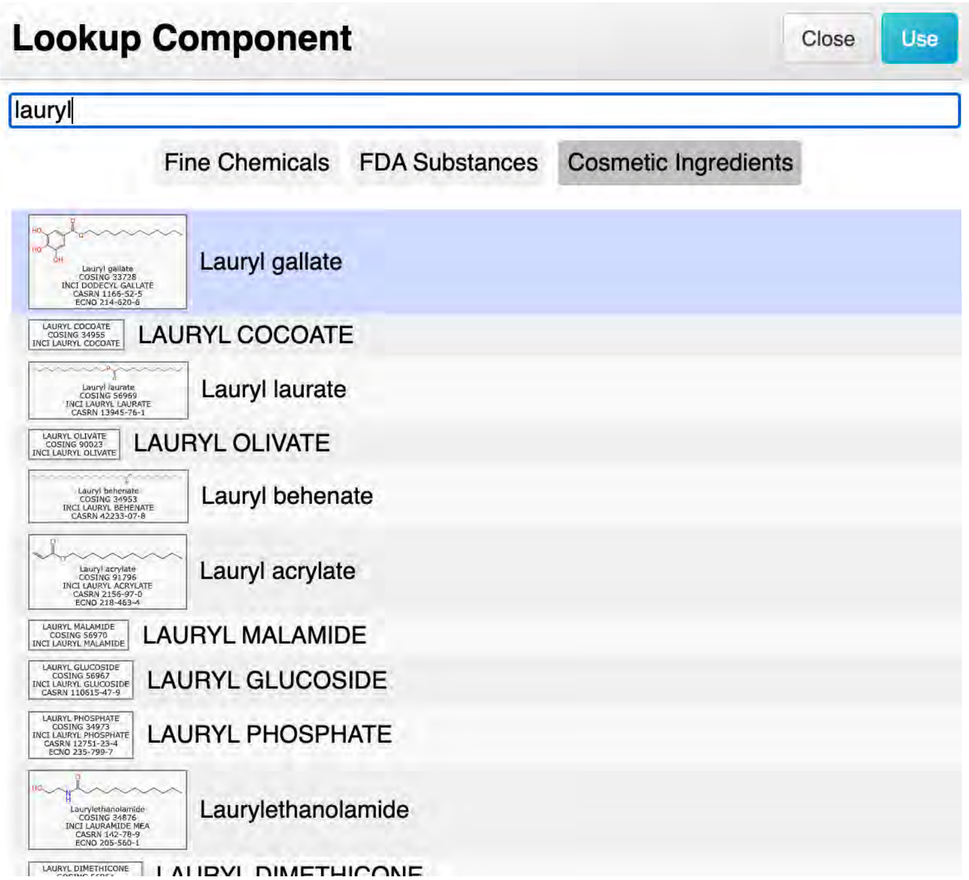
+ 14 more measurements

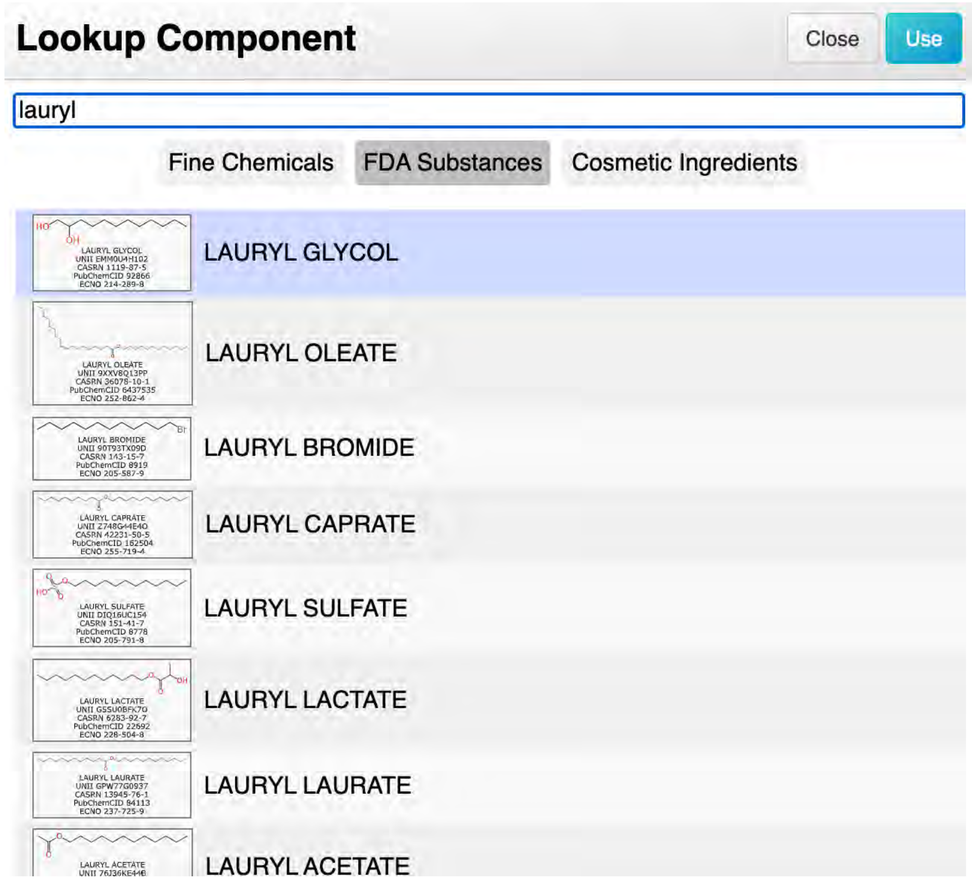
All Together for QSAR

Solubility	H_2O	—OH									
0.699		1									
15.19			1								
1.04					1						
3.142								1			
0.784										1	
0.91											1
6.3	1										
13.7		1									
11.6			1								
13.58				1							
6.73									1		
9.3								1			
8.20	0.8	0.2									
16.38	0.5	0.5									
13.60	0.2	0.8									
	(+8 more similar)										
15.39	0.333						0.667				
26.6	0.5		0.5								
17.97	0.083					0.584	0.333				
19.06	0.708					0.292					
22.59	0.417		0.25				0.333				
26.52	0.283		0.25			0.3	0.167				
	(+14 more similar)										

Databases

✤ Gather public content like INCI and UNII:



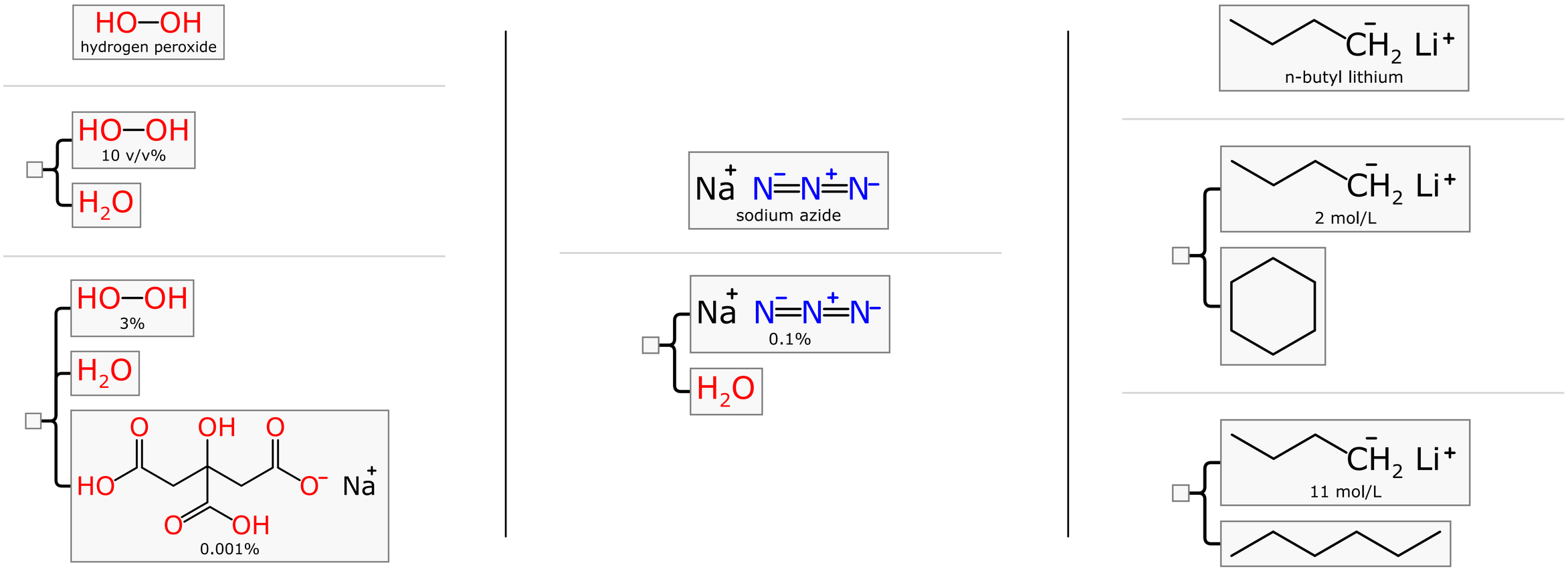


✤ Chemical Abstracts in mixture form would be nice

✤ Reagents and materials from vendors: could search, copy, paste from site

Hazards

- Automated lookup or estimation of safety, toxicity, etc. needs structures *and* mixture context... each of these is not like the other



- Machine readability and open access are both major hurdles

Longevity

- Any ELN, private registration system or public database:
  - capture data in *machine readable form*
  - if a machine can understand it, so can a human
  - if standards are followed, data will always be interpretable
  - data can be shared as much or as little as needed
- Sophisticated queries and analysis become possible
- Institutional knowledge does not evaporate
- An open ecosystem means that tools will evolve
  - tools can be free or proprietary, general purpose or specific

Questions?

✤ Contact:

- Leah McEwen **lrm1@cornell.edu** (Cornell University, IUPAC/InChI Trust)
- Alex M. Clark **alex@collaborativedrug.com** (Collaborative Drug Discovery)