February 2021

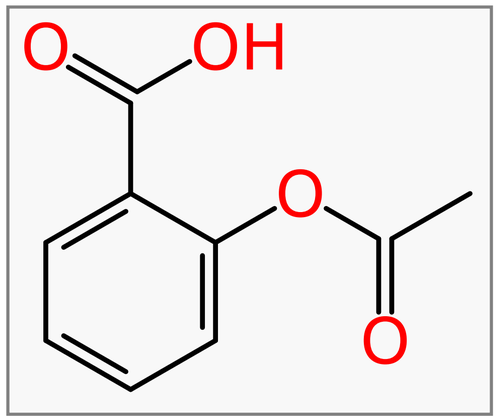
Mixtures QSAR

modelling collections of chemicals

Alex M. Clark

alex@collaborativedrug.com

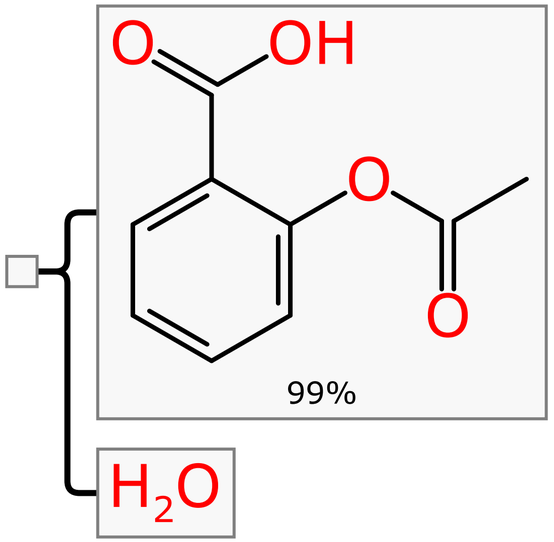
Real world chemistry is mixtures



Molfile

InChI

1980's



Mixfile

MInChI

2020's

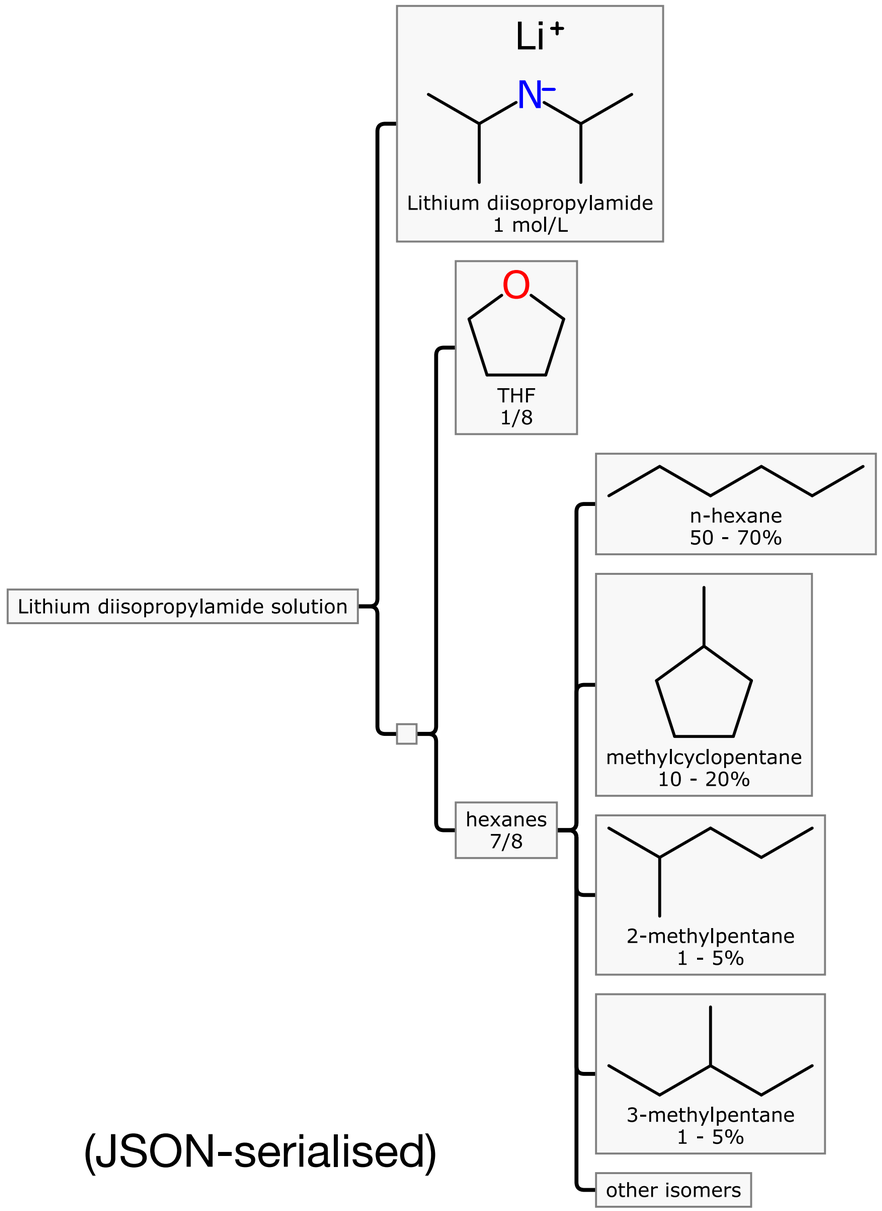
- Most mixtures stored as text or custom spreadsheets
- Value of upgrading to cheminformatics is well established...
- ... with the right datastructure, always just one script away from what you need
- If you can represent it, you can model it

CDD VAULT® Complexity Simplified

InChI TRUST

IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

Mixfile/MInChI



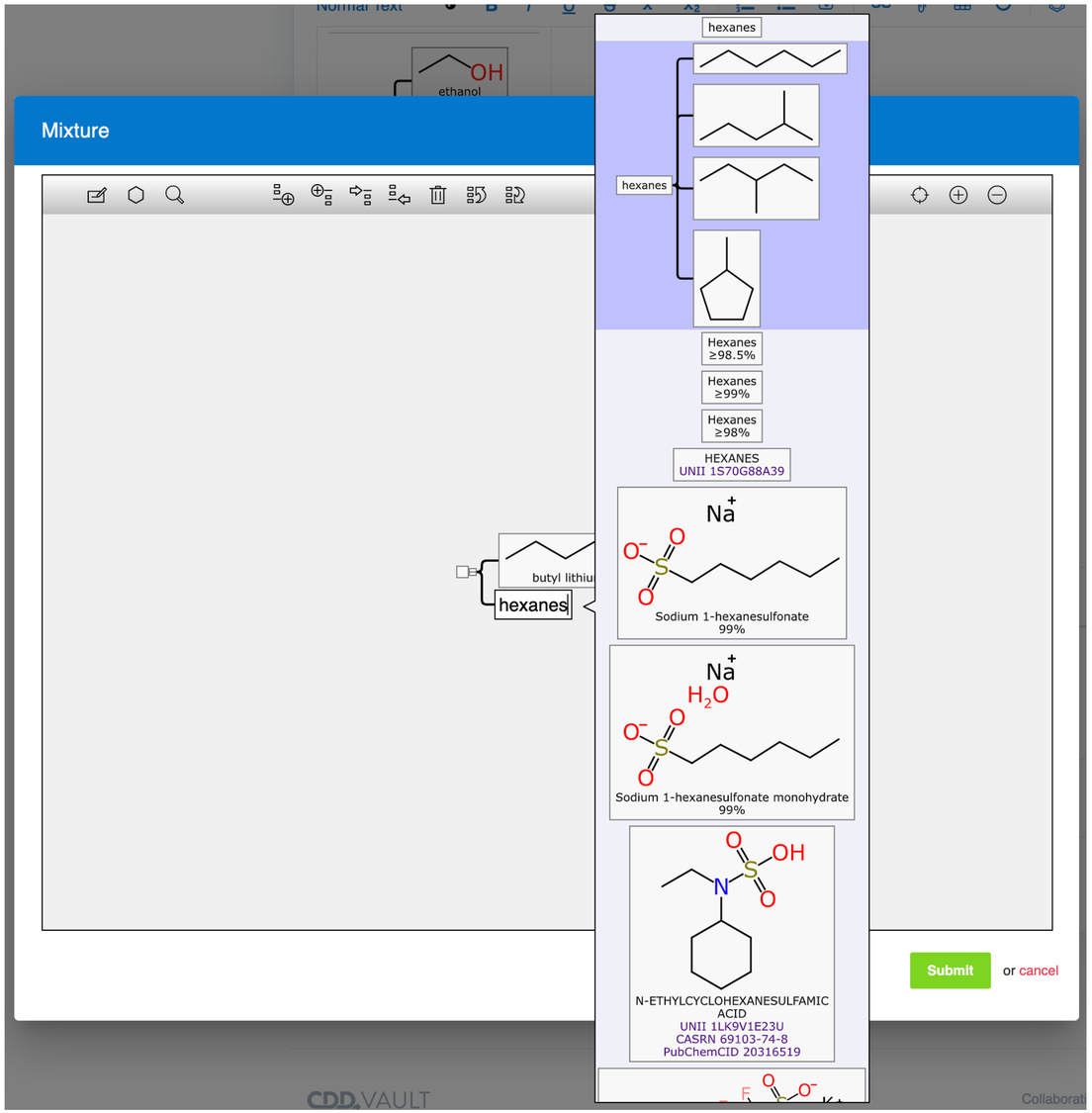
(JSON-serialised)

- 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) notation

```
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}}
```

Journal of Cheminformatics (2019)
10.1186/s13321-019-0357-4

Content creation



- Draw with editor

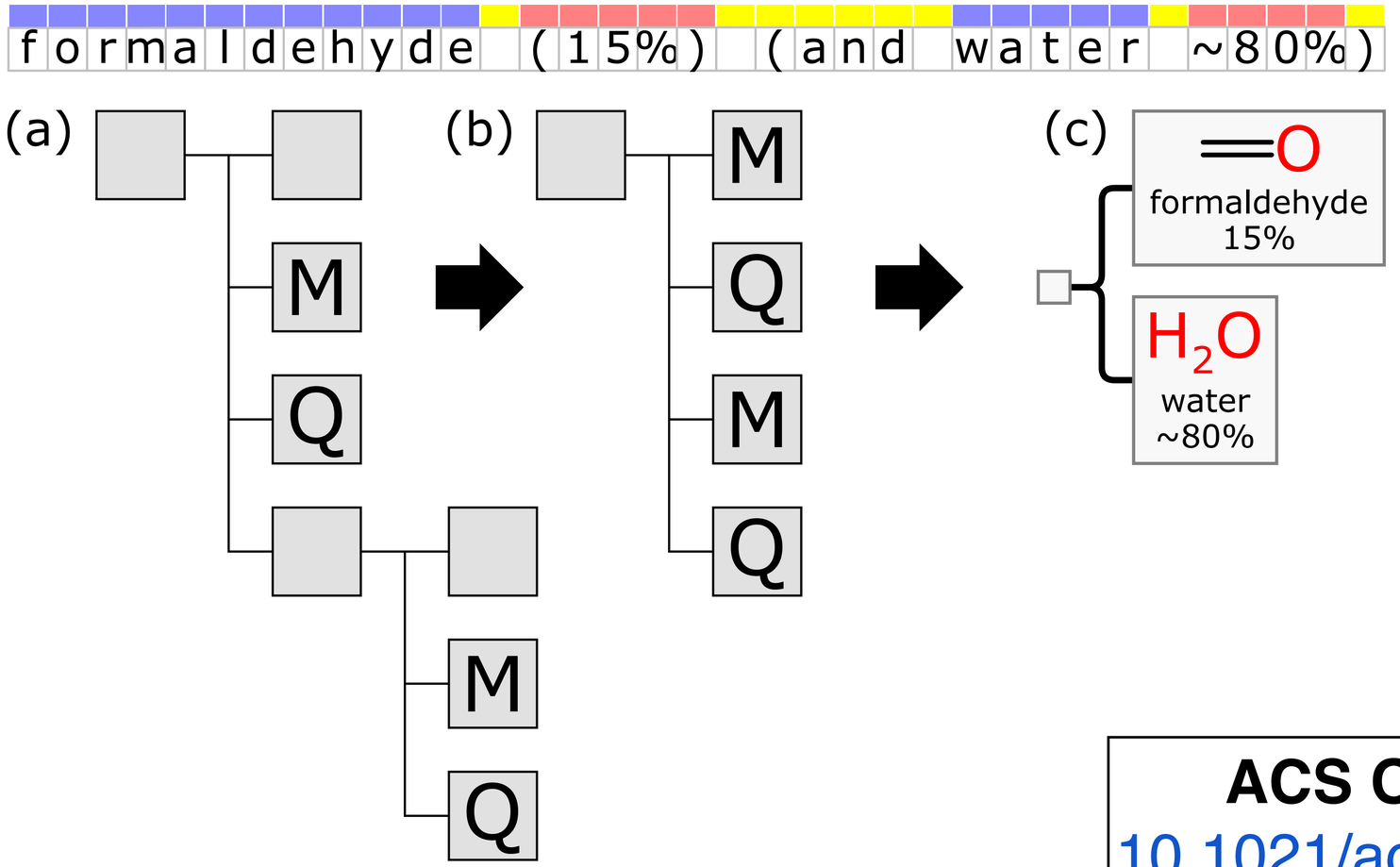
  github.com/cdd/mixtures

- ... or import existing data

Fine chemicals as text

✤ Catalogs & inventories favour short descriptions of mixtures

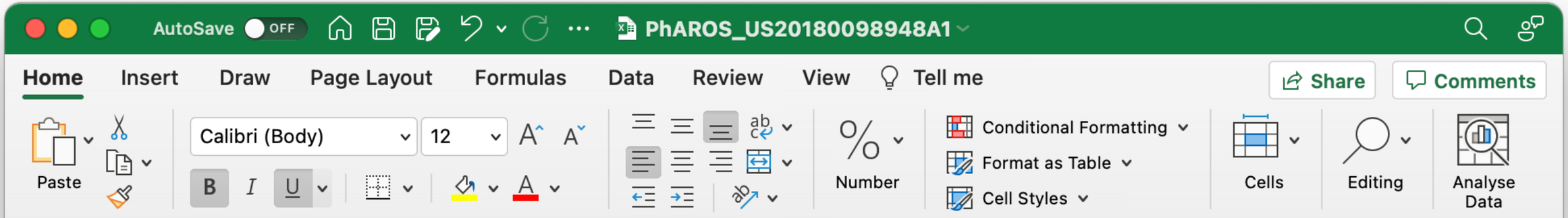
✤ Machine learning for text-to-mixture works well:



ACS Omega (2021)
10.1021/acsomega.1c03311

Mixtures as spreadsheets

✤ Each component defined by a column (or several)...



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Code**	**Major**	**Conc (uM)**	**MPP-Cyto%**	**Minor1**	**Minor2**	**Minor3**	**Minor4**	**Minor5**	**Minor6**	**Minor7**	**Minor8**
2	ENT1a	cannabidiol	10	62	cannabichromene	cannabigerol	cannabidivarin	limonene	linalool	nerolidol	alpha-pinene	cis-phytol
3	ENT2a	cannabidiol	10	16	limonene	linalool	nerolidol	alpha-pinene	cis-phytol			
4	ENT3a	cannabidiol	10	31	cannabichromene	cannabigerol	cannabidivarin					
5	ENT4a	cannabidiol	10	16	cannabichromene	limonene	linalool	nerolidol	alpha-pinene	cis-phytol		
6	ENT5a	cannabidiol	10	14	cannabigerol	limonene	linalool	nerolidol	alpha-pinene	cis-phytol		
7	ENT6a	cannabidiol	10	51	cannabidivarin	limonene	linalool	nerolidol	alpha-pinene	cis-phytol		
8	ENT7a	cannabidiol	10	20	cannabichromene	cannabigerol						
9	ENT8a	cannabidiol	10	44	cannabichromene	cannabidivarin						
10	ENT9a	cannabidiol	10	36	cannabigerol	cannabidivarin	limonene	linalool	nerolidol	alpha-pinene	cis-phytol	
11	ENT10a	cannabidiol	10	64	cannabichromene	cannabigerol	cannabidivarin	limonene	linalool			
12	ENT11a	cannabidiol	10	33	cannabichromene	cannabigerol	cannabidivarin	nerolidol				
13	ENT12a	cannabidiol	10	27	cannabichromene	cannabigerol	cannabidivarin	alpha-pinene	cis-phytol			
14	ENT1b	cannabinol	10	48	cannabichromene	cannabigerol	cannabidivarin	limonene	linalool	nerolidol	alpha-pinene	cis-phytol
15	ENT2b	cannabinol	10	6	limonene	linalool	nerolidol	alpha-pinene	cis-phytol			
16	ENT3b	cannabinol	10	41	cannabichromene	cannabigerol	cannabidivarin					
17	ENT4b	cannabinol	10	9	cannabichromene	limonene	linalool	nerolidol	alpha-pinene	cis-phytol		

CDD Vault

- Mixture composer
- Destination for each column
- Build mixture hierarchy from rows
- Name-to-structure, lookup databases, links to ID codes

CDD VAULT · McKerrow Vault

Help · Log out

Explore Data | ELN | Import Data | Reports | Settings | 4 | Full-Access User

Step 1: Choose Data File → Step 2: Map Fields → Step 3: Commit Data

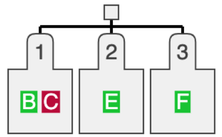
File: PhAROS_US20180098948A1.xlsx

Project: Mixtures · **Owner:** Full-Access User

Compose mixtures from columns Edit

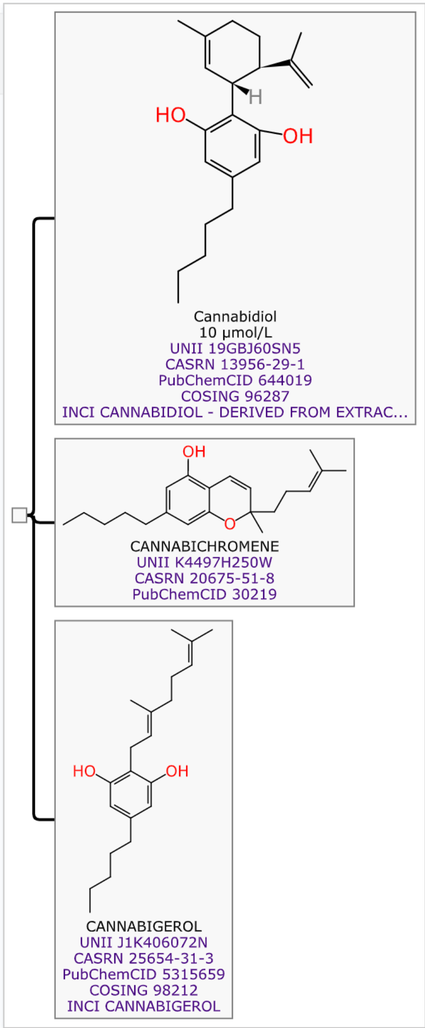
- ○ Not part of mixture
- ○ Chemical structure [Molfile]
- ○ Chemical structure [SMILES]
- ◉ Molecule lookup by name
- ○ CDD Vault identifier
- ○ Component name (text)
- ○ Quantity
- ○ Units

Component: 1 2 3 / 4 5 6 / 7 8 9 ?



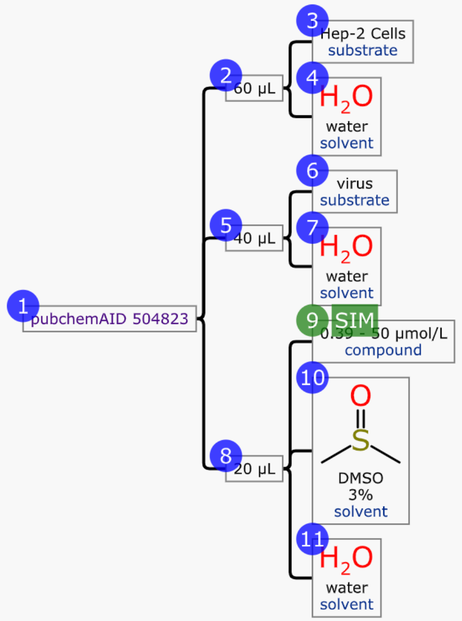
B ⇒ 1 molecule name	C ⇒ 1 quantity	D	E ⇒ 2 molecule name	F ⇒ 3 molecule name	G
Major	**Conc (uM)**	**MPP-Cyto%**	**Minor1**	**Minor2**	**Minor3**
cannabidiol	10	62	cannabichromene	cannabigerol	cannabidivari
cannabidiol	10	16	limonene	linalool	nerolidol
cannabidiol	10	31	cannabichromene	cannabigerol	cannabidivari
cannabidiol	10	16	cannabichromene	limonene	linalool
cannabidiol	10	14	cannabigerol	limonene	linalool
cannabidiol	10	51	cannabidivarin	limonene	linalool
cannabidiol	10	20	cannabichromene	cannabigerol	
cannabidiol	10	44	cannabichromene	cannabidivarin	
cannabidiol	10	36	cannabigerol	cannabidivarin	limonene
cannabidiol	10	64	cannabichromene	cannabigerol	cannabidivari

and 56 additional rows



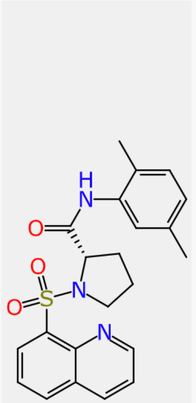
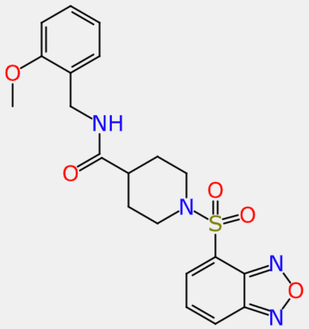
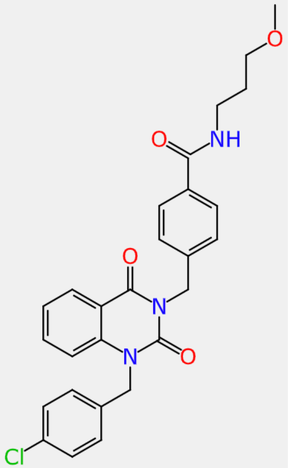
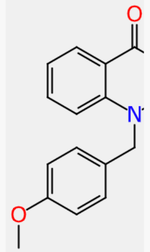
High throughput screening

Mixture Export - pubchem504823.mixfile



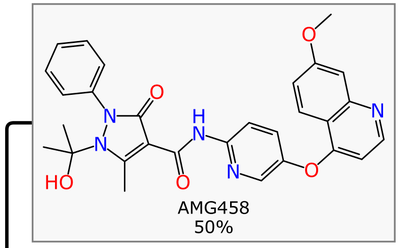
Edit Template

Activity | Protocol | Results

		1	2	3	
Molecule	structure 9				
PUBCHEM_RESULT_TAG	ignore	1	2	3	4
PUBCHEM_SID	identifier 9 Edit pubchemSID	104169543	104169547	118045990	118045991
PUBCHEM_CID	identifier 9 Edit pubchemCID	49842897	6619281	15989137	51051584
PUBCHEM_ACTIVITY_OUTCOME	ignore	Active	Active	Active	Active
PUBCHEM_ACTIVITY_SCORE	ignore	100	100	100	100
PUBCHEM_ACTIVITY_URL	ignore				
PUBCHEM_ASSAYDATA_COMMENT	ignore				
IC50 Modifier	ignore			<	
IC50	metadata 9 Edit IC50 micromoles per litre (umol/L)	3.2	2.9	1.2	2.4
% CPE Inhibition @ 150 uM	ignore			70.12	6.22

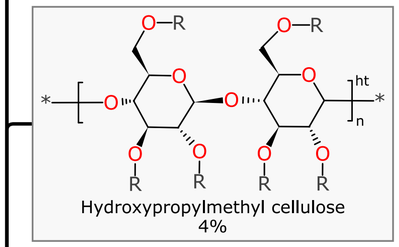
Drug tablet formulation

- Represent *active ingredient* by structure & concentration
- Other materials (*excipients*):
  - polymers
  - organic molecules
  - salts
  - amorphous materials

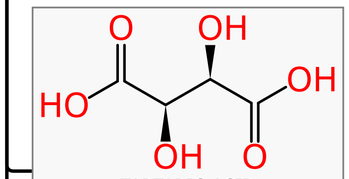


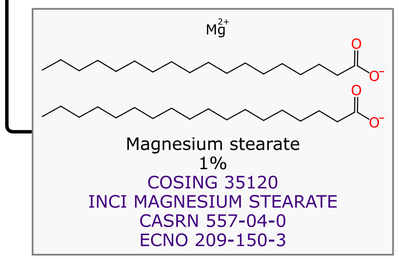
active ingredient

MICROCRYSTALLINE CELLULOSE
16%
UNII OP1R32D61U
CASRN 9004-34-6
ECNO 232-674-9
COSING 78508
INCI MICROCRYSTALLINE CELLULOSE



CROSPOVIDONE
10%
UNII 2S7830E561
CASRN 9003-39-8





excipients

J Pharm Sci (2009)
10.1002/jps.21422

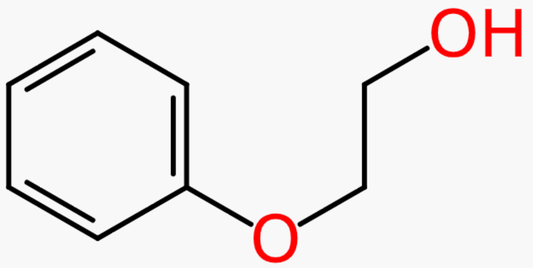
Cosmetics

Eucerin

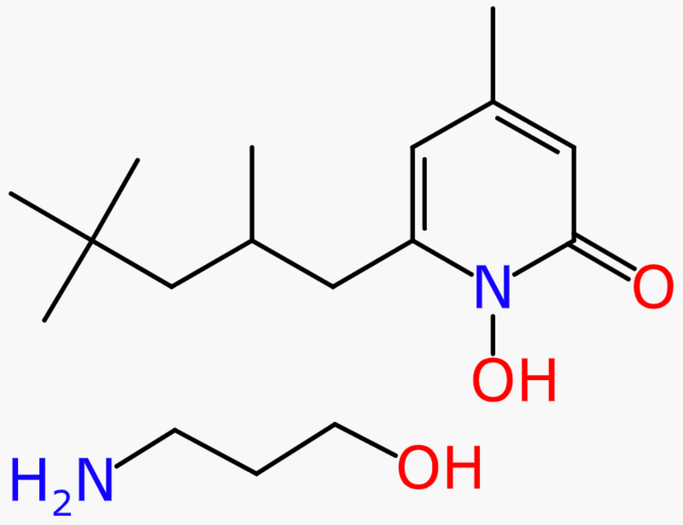
PETROLATUM
UNII 4T6H12BN9U
CASRN 8009-03-8
ECNO 232-373-2

MINERAL OIL
UNII T5L8T28FGP
CASRN 8042-47-5
ECNO 232-455-8

LANOLIN ALCOHOL
COSING 78357
INCI LANOLIN ALCOHOL
CASRN 8027-33-6
ECNO 232-430-1



2-Phenoxyethanol
COSING 36522
INCI PHENOXYETHANOL
CASRN 122-99-6
ECNO 204-589-7

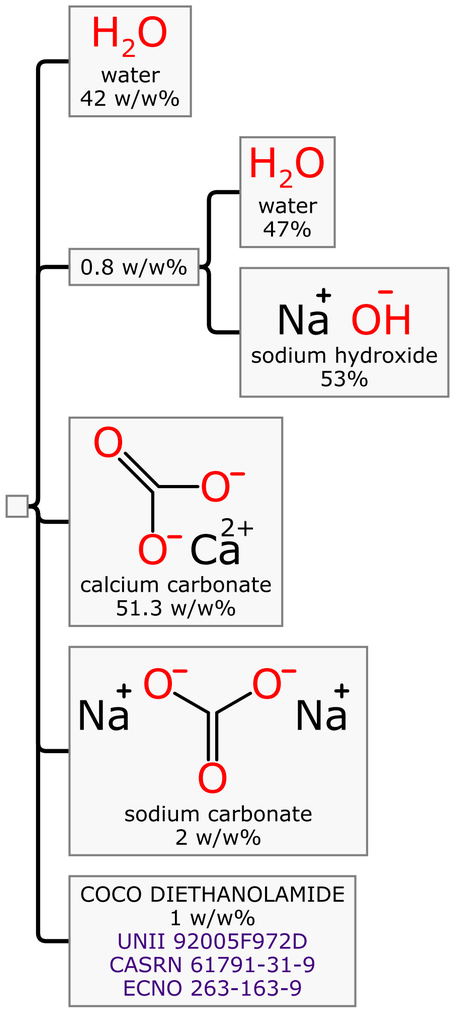


Piroctone olamine
COSING 36604
INCI PIROCTONE OLAMINE
CASRN 68890-66-4
ECNO 272-574-2

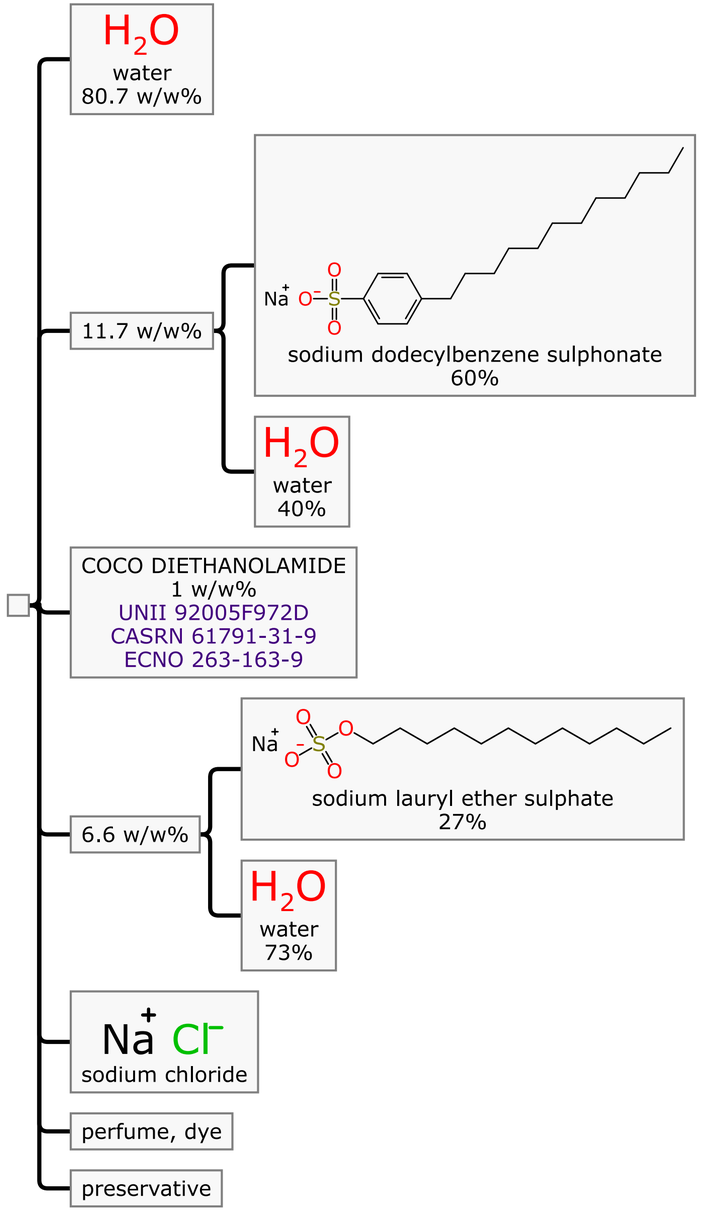
H_2O
WATER
UNII 059QF0KO0R
CASRN 7732-18-5
PubChemCID 962
ECNO 231-791-2

CERESIN
UNII Q1LS2UJO3A
CASRN 8001-75-0
ECNO 232-290-1

Consumer Products

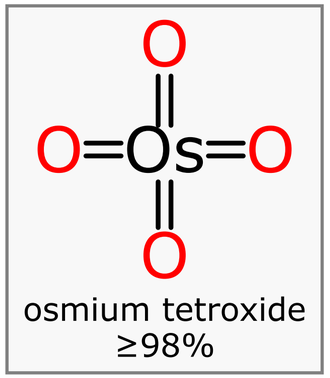


abrasive hand cleaner

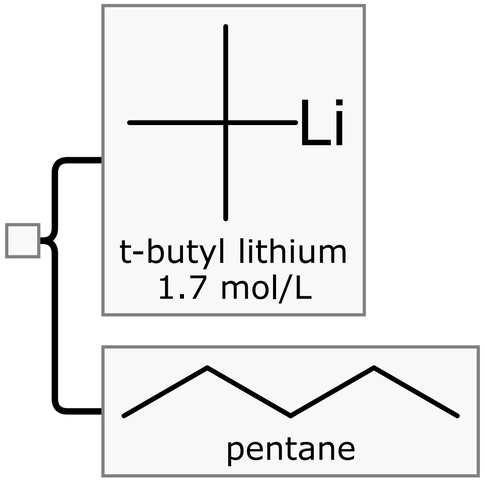


dishwashing liquid

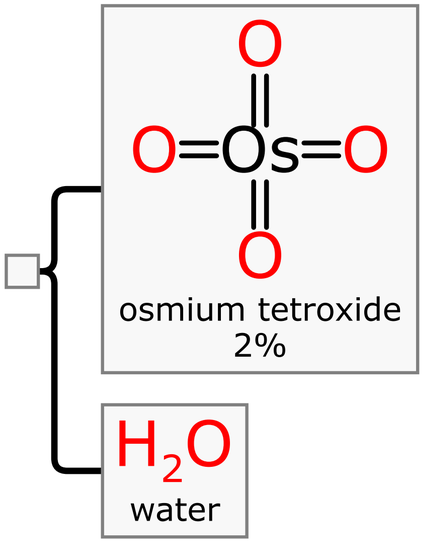
Hazards



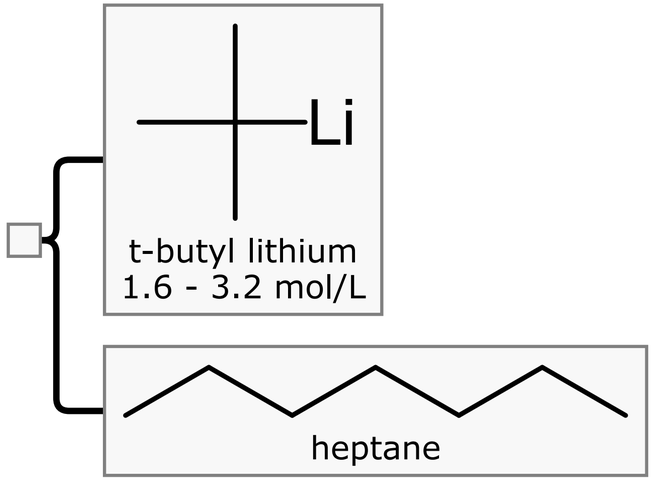
volatile
extreme danger
sealed ampule



storage 2-8 °C
flash point -49 °C



less
dangerous

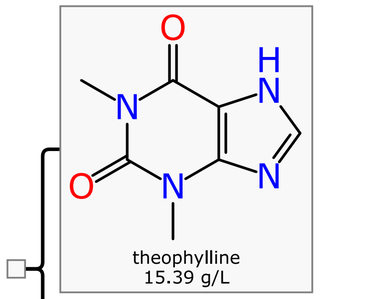


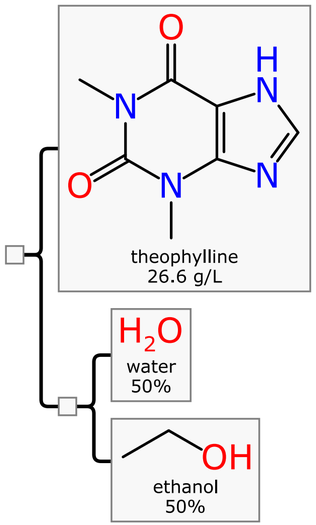
storage 15 °C
flash point -4 °C

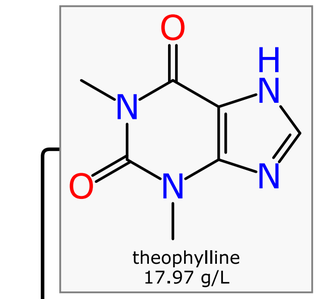
Mixtures QSAR

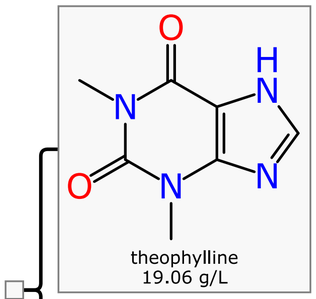
- Using chemical structure to model properties has a long history
- Modelling mixtures adds additional layers:
  - data capture has to be solved first
  - adapt cheminformatics for *multiple* structures
  - factor in concentrations
  - use other descriptors when structures absent
- Several demo examples...

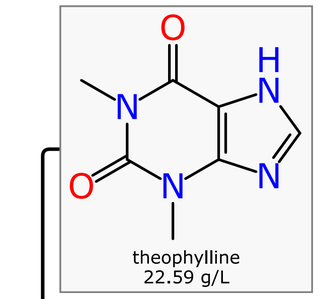
Example 1: Theophilline

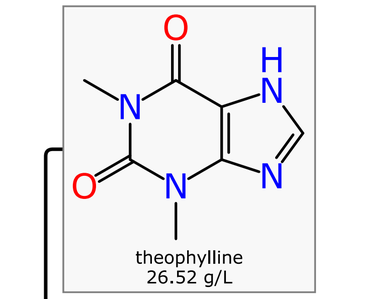


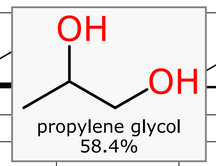










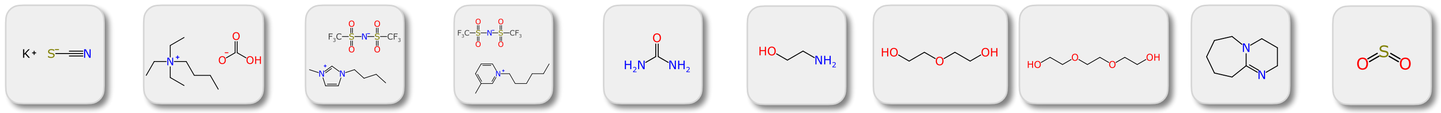


Solubility	H_2O	—OH	OH	OH	OH	OH OH						
0.699		1										
15.19			1									
1.04					1							
3.142								1				
0.784										1		
0.91											1	
6.3	1											
13.7		1										
11.6			1									
13.58				1								
6.73									1			
9.3								1				
8.20	0.8	0.2										
16.38	0.5	0.5										
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)											

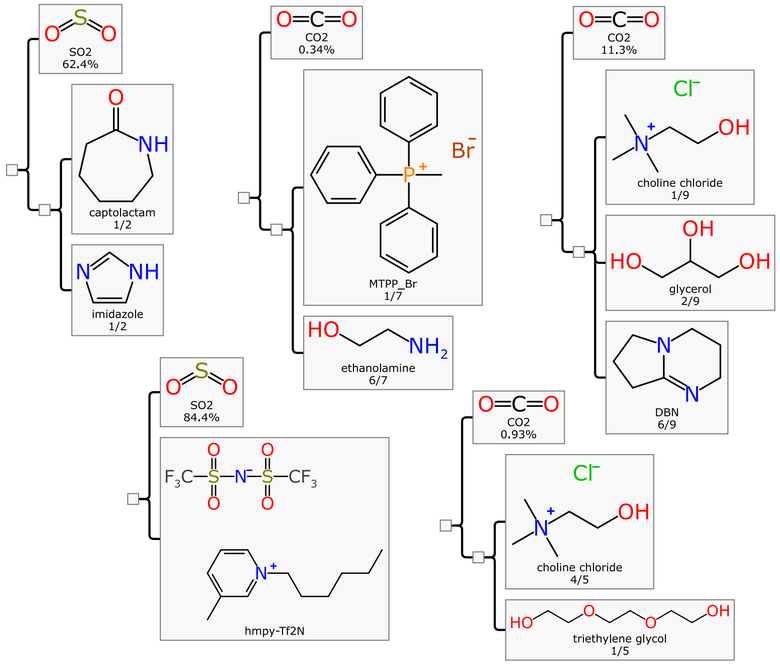
- https://bit.ly/3Kqb5VA
- Solubility is limiting factor
- Sourced from multiple papers

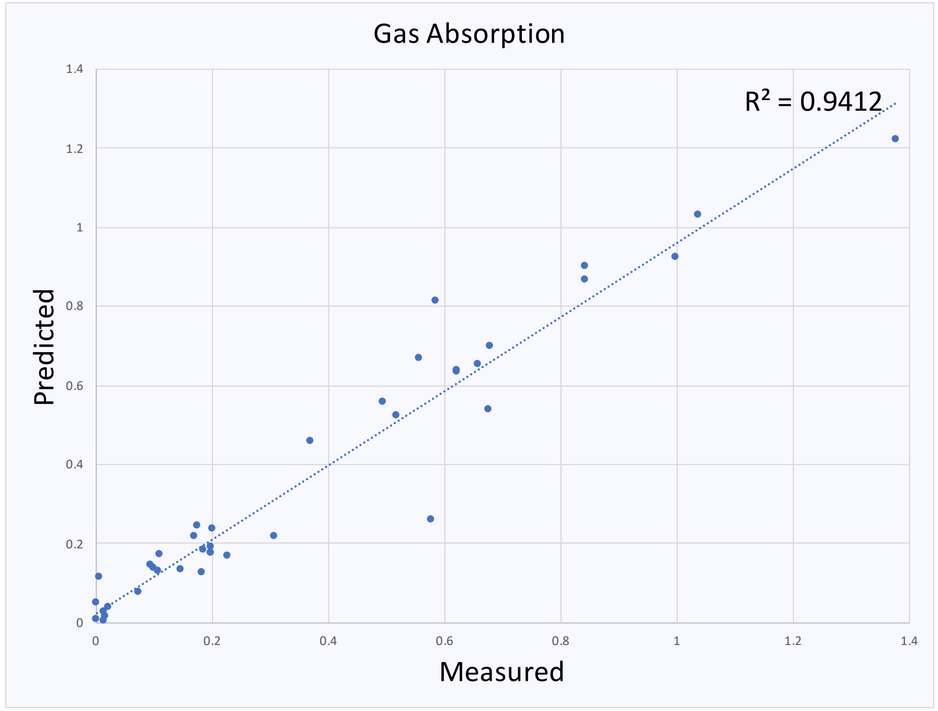
Example 2: Eutectic solubility

- https://bit.ly/3IWFKJV
- Have solubility of CO_2 and SO_2 in various solvent combinations



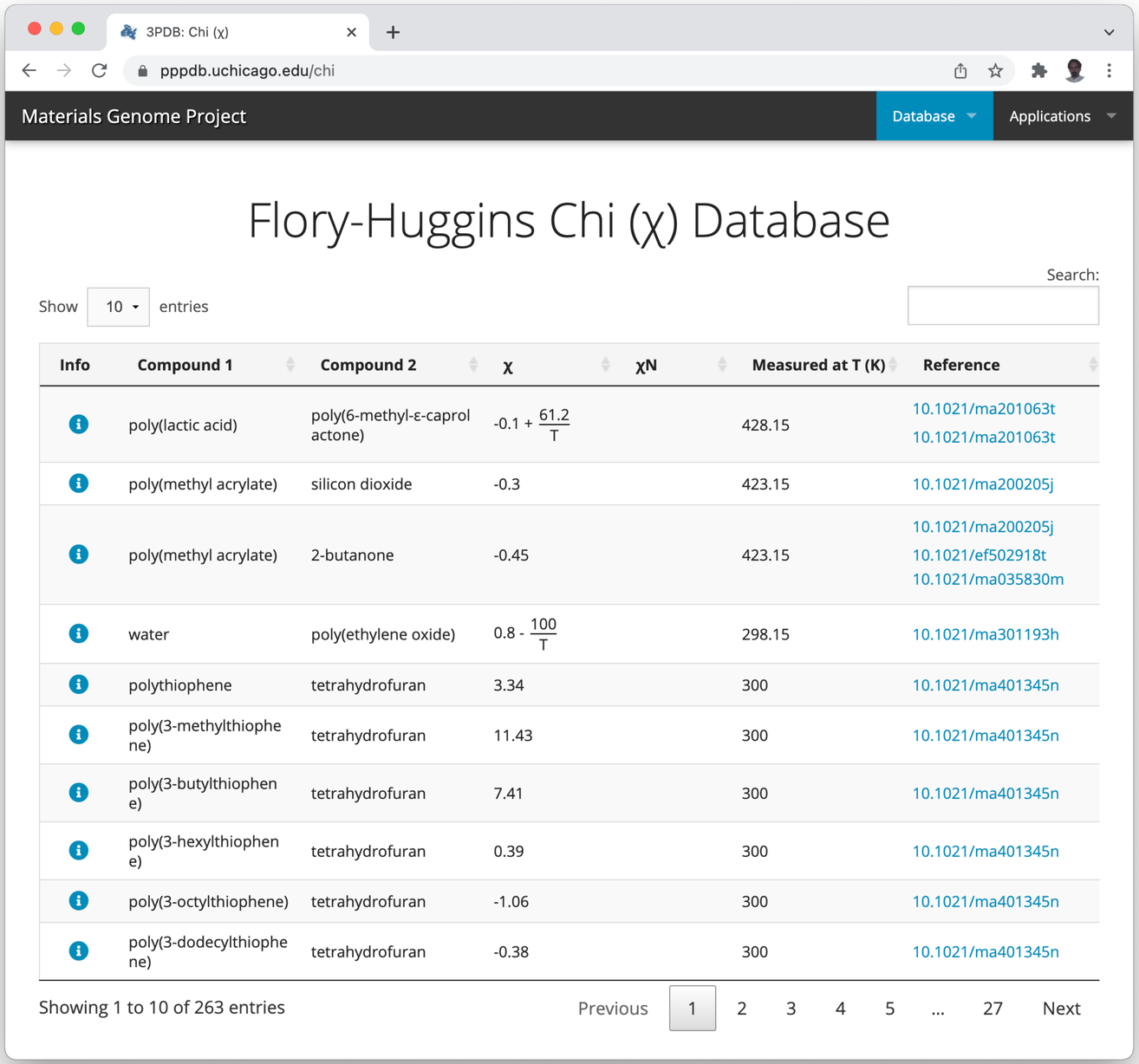
0	0.01852	0.01389	0.01258	0	0.04762	0.04167	0.03922	0.5455	0
0	0.08	0.01905	0.01667	0.6667	0.07843	0.07018	0.06667	0	0
0	0.05556	0.01961	0.01709	0	0.2963	0.2424	0.2222	0	0
0	0.06122	0.02317	0.02041	0.06593	0.8571	0.2449	0.2286	0	0
0	0.06122	0.02317	0.02041	0.06593	0.8571	0.2449	0.2286	0.004082	0
0	0.06548	0.02317	0.02041	0.06593	0.8571	0.2449	0.2286	0	0
0	0.1429	0.6471	0.1509	0	0.0303	0.02857	0.02778	0.02174	0
0	0.07317	0.4054	0.09615	0	0.03448	0.03226	0.03125	0.02381	0
0	0.1333	0.03902	0.03478	0	0.1882	0.1818	0.2	0	0
0	0.1333	0.03902	0.03478	0	0.1882	0.1684	0.16	0.006452	0
0	0.1333	0.03902	0.03478	0	0.1882	0.2	0.1818	0	0
0	0.08333	0.02439	0.02174	0	0.1176	0.1053	0.1	0	1
0	0.25	0.08163	0.07407	0.1154	0.1481	0.1379	0.1333	0	1
0.25	0.01974	0.0163	0.03	0.03409	0	0	0	0.1406	1
0.25	0.1154	0.04167	0.03659	0.4167	0.05	0	0	0	1
0.1875	0.1154	0.04167	0.03659	0.4167	0.05	0	0	0	1
0.1875	0.01974	0.0163	0.03	0.03409	0	0	0	0.1406	1
0.3	0.072	0.01714	0.015	0.6	0.04615	0	0	0	1
0	0.03226	0.0125	0.01111	0	0.1333	0.4545	0.5	0.5	0
0	0.07778	0.5	0.1633	0.01429	0.01351	0.01282	0.0125	0.5	0
0	1	0.1556	0.14	0.12	0.07143	0.06667	0.06452	0	0
0	0.05556	0.01754	0.0155	0	0.1429	0.125	0.1176	0	0
0	0.2292	0.08108	0.07143	0.02273	0.025	0.02273	0.02174	0.09375	1
0	0.07639	0.02703	0.02381	0.007576	0.008333	0.007576	0.007246	0.03125	1
0	0.07692	0.02778	0.02439	0.2778	0.03333	0	0	0.09375	1
0	0.01316	0.01316	0.02	0.02273	0	0	0	0.09375	1
0	0.1489	0.8788	0.3673	0.02703	0.02564	0.02439	0.02381	0.01923	1
0	0.14	0.3265	1	0.025	0.02381	0.02273	0.02222	0.01818	1
0	0.05926	0.02222	0.01951	0	0.05714	0.05	0.04706	0	0
0	0.08654	0.04286	0.0375	0	0.2727	0.2308	0.2143	0	0
0	0.07759	0.03947	0.03488	0	0.15	0.1324	0.125	0	0
0	0.1429	0.03333	0.02963	0.1333	0.07843	0.07018	0.06667	0	0
0	0.072	0.01951	0.01739	0.6	0.09412	0.08421	0.08	0	0
0	0.08	0.01905	0.01667	0.6667	0.07843	0.07018	0.06667	0	0
0	0.08571	0.02041	0.01786	0.7143	0.06723	0.06015	0.05714	0	0
0	0.08333	0.02439	0.02174	0.25	0.1176	0.1053	0.1	0	0
0	0.08333	0.02439	0.02174	0.3	0.1176	0.1053	0.1	0	0
0	0.08333	0.02439	0.02174	0.35	0.1176	0.1053	0.1	0	0
0	0.08333	0.02439	0.02174	0.25	0.075	0.1053	0.1	0	0
0	0.08333	0.02439	0.02174	0.3	0.05	0.1053	0.1	0	0
0	0.08333	0.02439	0.02174	0.35	0.025	0.1053	0.1	0	0





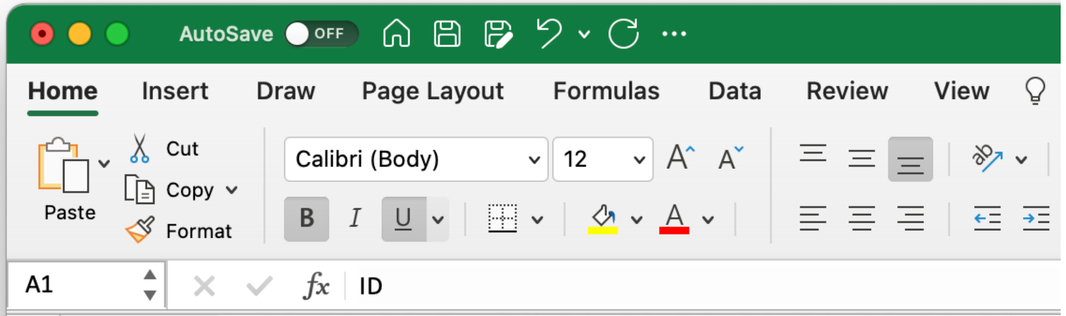
Polymers: Chi

- χ is an measure of entropy of mixing
- Values for polymer + [polymer *or* solvent *or* other]
- 2-component mixtures
- Data taken from Materials Genome Project...



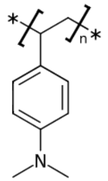
Spreadsheet

- Content arrives in tabular form
- Expanding into full mixtures - 2 steps:
  1. register each named structure
  2. use Vault to compose into mixtures

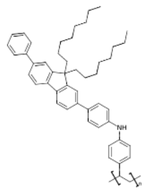


	A	B	C	D
1	ID	Component1	Component2	Chi
2	PolyMix001	poly(lactic acid)	poly(6-methyl-ε-caprolact	0.0429
3	PolyMix002	poly(methyl acrylate)	silicon dioxide	-0.3000
4	PolyMix003	poly(methyl acrylate)	2-butanone	-0.4500
5	PolyMix004	water	poly(ethylene oxide)	0.4646
6	PolyMix005	polythiophene	tetrahydrofuran	3.3400
7	PolyMix006	poly(3-methylthiophene)	tetrahydrofuran	11.4300
8	PolyMix007	poly(3-butylthiophene)	tetrahydrofuran	7.4100
9	PolyMix008	poly(3-hexylthiophene)	tetrahydrofuran	0.3900
10	PolyMix009	poly(3-octylthiophene)	tetrahydrofuran	-1.0600
11	PolyMix010	poly(3-dodecylthiophene)	tetrahydrofuran	-0.3800
12	PolyMix011	polystyrene	poly(4-hydroxystyrene)	0.6800
13	PolyMix012	poly(ethylene oxide)	poly(methyl acrylate)	-0.0400
14	PolyMix013	polybutadiene	tetrahydrofuran	0.3990
15	PolyMix014	polybutadiene	tetrahydrofuran	0.4030
16	PolyMix015	polybutadiene	tetrahydrofuran	0.4000
17	PolyMix016	polybutadiene	tetrahydrofuran	0.4000
18	PolyMix017	polybutadiene	tetrahydrofuran	0.4110
19	PolyMix018	polybutadiene	tetrahydrofuran	0.4200
20	PolyMix019	polyisoprene	cyclohexane	0.3940
21	PolyMix020	polyisoprene	cyclohexane	0.3730
22	PolyMix021	polyisoprene	cyclohexane	0.3950
23	PolyMix022	polyisoprene	cyclohexane	0.4370
24	PolyMix023	polyisoprene	cyclohexane	0.4380
25	PolyMix024	polyisoprene	cyclohexane	0.4400
26	PolyMix025	poly(4-tert-butylstyrene)	poly(methyl methacrylate	0.0966

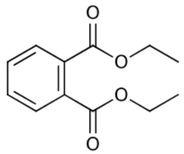
Register structures



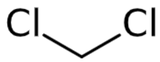
poly[N-(4-vinylbenzyl)-N,N-diethylamine]
Synonyms: *(no synonyms)*
Protocols: 0



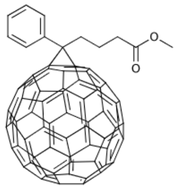
poly[(9,9-dioctylfluorenyl-2,7-diyl)-co-5,5-(4′,7′-di-2-thienyl-2′,1′,3′-benzothiadiazole]
Synonyms: *(no synonyms)*
Protocols: 0



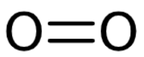
diethyl phthalate
Synonyms: *(no synonyms)*
Protocols: 0



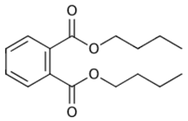
dichloromethane
Synonyms: *(no synonyms)*
Protocols: 0



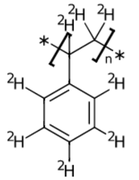
phenyl-C61-butyric acid methyl ester
Synonyms: [6,6]-phenyl-C61 butyric acid methyl ester
Protocols: 0



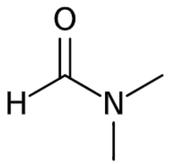
oxygen
Synonyms: *(no synonyms)*
Protocols: 0



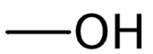
dibutyl phthalate
Synonyms: *(no synonyms)*
Protocols: 0



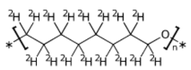
deuterated polystyrene
Synonyms: *(no synonyms)*
Protocols: 0



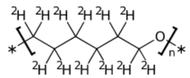
n,n-dimethylformide
Synonyms: *(no synonyms)*
Protocols: 0



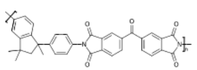
methanol
Synonyms: *(no synonyms)*
Protocols: 0



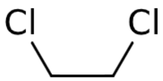
deuterated poly(octylene oxide)
Synonyms: *(no synonyms)*
Protocols: 0



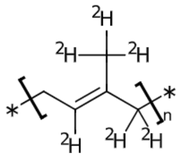
deuterated poly(hexene oxide)
Synonyms: *(no synonyms)*
Protocols: 0



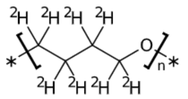
matrimid
Synonyms: *(no synonyms)*
Protocols: 0



ethylene dichloride
Synonyms: 1,2-dichloroethane
Protocols: 0



deuterated poly(ethylene-alt-propylene)
Synonyms: *(no synonyms)*
Protocols: 0



deuterated poly(butylene oxide)
Synonyms: *(no synonyms)*
Protocols: 0

Markup into mixtures

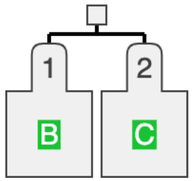
✤ CDD Vault importer: *Compose Mixtures*

- ○ Not part of mixture
- ○ Chemical structure [Molfile]
- ○ Chemical structure [SMILES]
- ○ Molecule lookup by name
- ◉ CDD Vault identifier
- ○ Component name (text)
- ○ Quantity
- ○ Units

Component

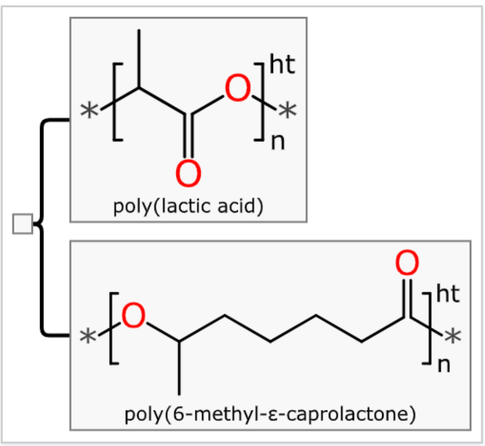
1	2	3
4	5	6
7	8	9

?



	A	B ⇒ 1 Vault identifier	C ⇒ 2 Vault identifier	D	E	
	ID	Component1	Component2	Chi	Chi-Raw	Chi-N
1	PolyMix001	poly(lactic acid)	poly(6-methyl-ε-caprolactone)	0.04294	-0.1 + 61.2T	
2	PolyMix002	poly(methyl acrylate)	silicon dioxide	-0.30000	-0.3	
3	PolyMix003	poly(methyl acrylate)	2-butanone	-0.45000	-0.45	
4	PolyMix004	water	poly(ethylene oxide)	0.46460	0.8 - 100T	
5	PolyMix005	polythiophene	tetrahydrofuran	3.34000	3.34	
6	PolyMix006	poly(3-methylthiophene)	tetrahydrofuran	11.43000	11.43	
7	PolyMix007	poly(3-butylthiophene)	tetrahydrofuran	7.41000	7.41	
8	PolyMix008	poly(3-hexylthiophene)	tetrahydrofuran	0.39000	0.39	
9	PolyMix009	poly(3-octylthiophene)	tetrahydrofuran	-1.06000	-1.06	
10	PolyMix010	poly(3-dodecylthiophene)	tetrahydrofuran	-0.38000	-0.38	

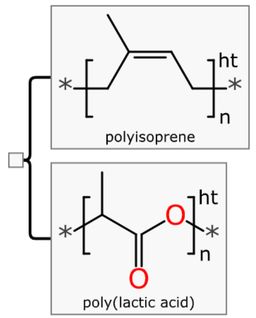
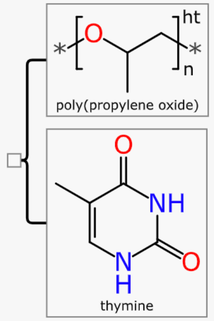
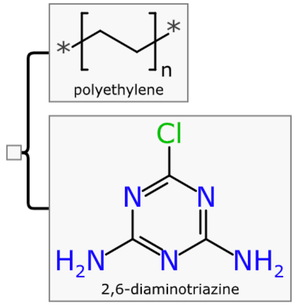
and 237 additional rows



Gather data

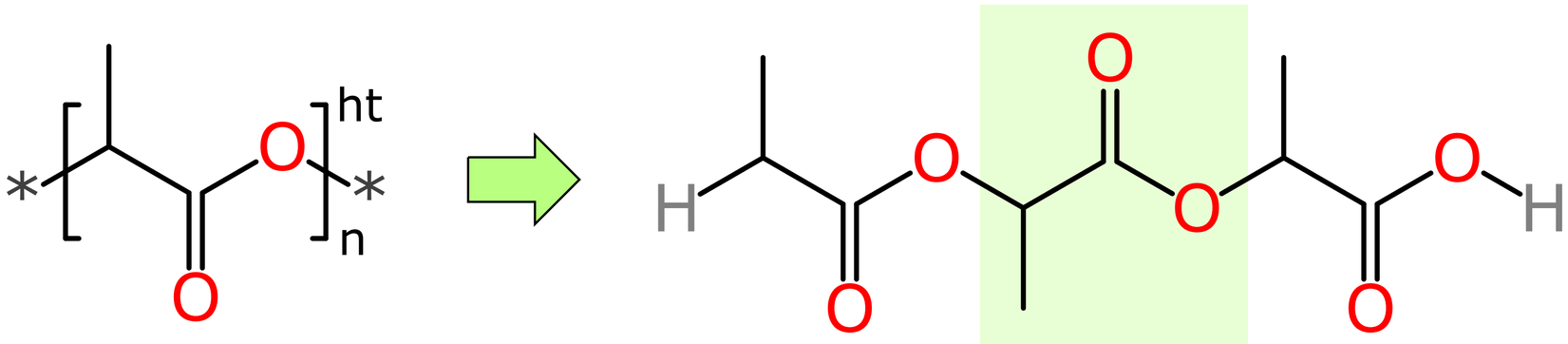
- Select all chemical objects with a *Flory-Huggins Chi* property
- 172 unique datapoints
- Data gathering is generic: applies to any mixture-formatted content with this property

172 Selected: Plot · Export · Add to collection · Build model · Flag outliers | Customize your report · Save this search

Select… all · none	Molecule	Flory-Huggins Chi
☑	 PolyMix245 McKerrow Vault	0.799 0.689 0.419
☑	 PolyMix244 McKerrow Vault	10.6
☑	 PolyMix243 McKerrow Vault	3.00
		8.50

Model

✤ For polymer components, do some simple processing:

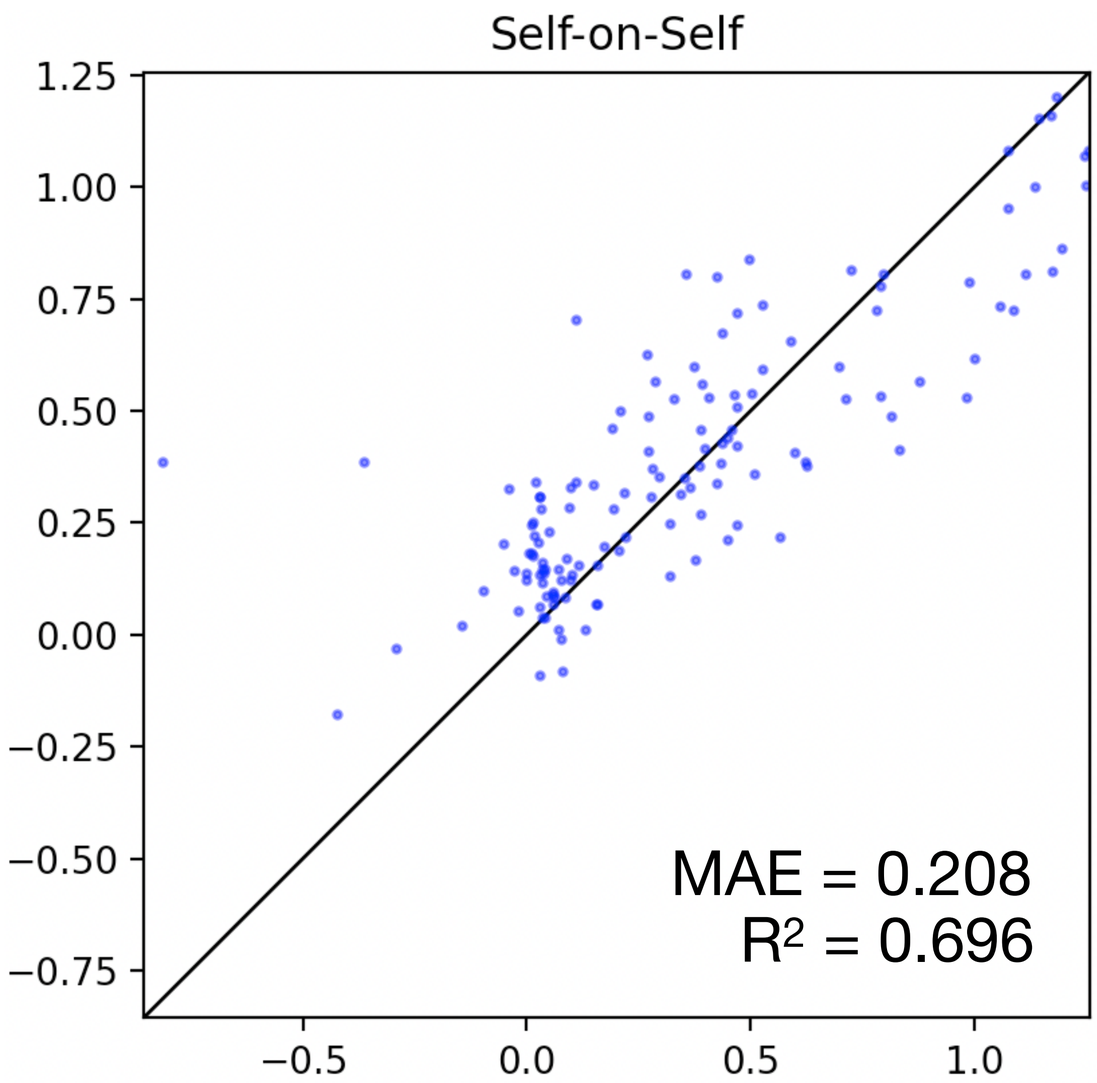


✤ Generate ECFP4 fingerprints / fold into 2048 bit vector

✤ Add vectors for component 1 & component 2: values are [0, 1, 2]

✤ Model **descriptors** ➪ **chi** using LightGBM (tree-based gradient boost)

Results



- Can propose:
  - arbitrary polymer structure
  - arbitrary other molecule
  - predict entropy of mixing (χ)
- Dataset is small and diverse...
- ... cross validation not great, need more data
- Database of mixtures & χ-values?
- Bring into training set

The Future

- Digitise mixtures, like for molecules - *self describing*
- Put them in repositories
  - with measured properties
  - open databases (like *PubChem*)
  - or just a huge data lake
- Then focus on
  - queries, analysis, model building - the fun part!
  - describing components that are not simple molecules

Acknowledgements

- Leah McEwen
- InChI Trust & IUPAC
- NIH Grant 1R43TR002528-01
- The CDD Vault team

alex@collaborativedrug.com

(PS: we're hiring)