SI: Solvmate - A Hybrid Physical/ML Approach to Solvent Recommendation Leveraging a Rank-based Problem Framing

Contents

1 Data summary statistics

Table 1: Top solvent counts by data source.

open notebook	
solvent	count
water	610
water : PEG-400	369
1-octanol	322
methanol	294
ethanol	284
ethanol : water	264
tetrahydrofuran	152
2-propanol	123
acetonitrile	121
1-propanol	117
toluene	112
chloroform	109
1-butanol	102
dimethyl sulfoxide	97
acetone	97
PEG-400	97
ethyl acetate	90
1-pentanol	79
cyclohexane	71
1-hexanol	66

nova	
solvent	count
methanol	319
nitromethane	319
acetone	319
dimethylformamide	319
methyl isobutyl ketone	319
ethyl acetate	319
ethanol	319
toluene	319
acetonitrile	318
methyl t-butylether	318
chlorobenzene	318
chloroform	317
2-propanol	315
2-butanone	315
tetrahydrofuran	315
methylene chloride	315
diethyl ether	315
hexane	314

bayer_internal	
solvent	count
ethanol	95
acetonitrile	57
tetrahydrofuran	56
acetone	56
ethyl acetate	56
2-propanol	55
heptane	53
cyclohexane	53
ditertbutylether	53
toluene	51
water	49
methanol	47
dichloromethane	33

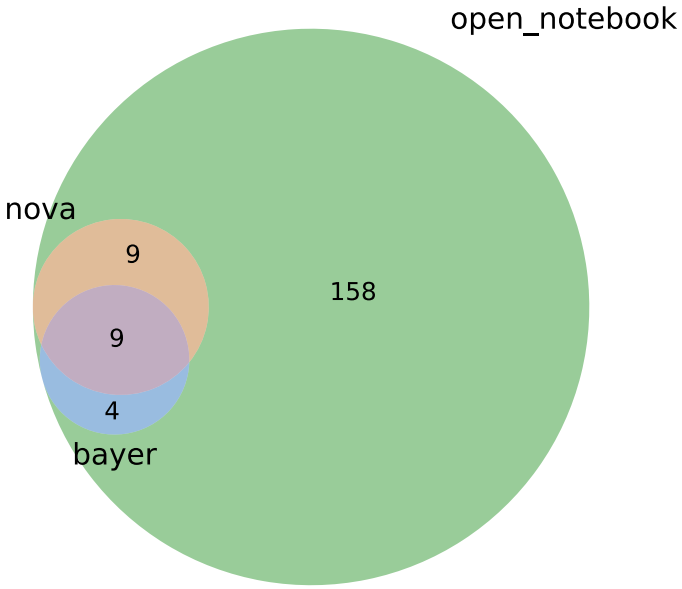


Figure S1: Venn diagram of solvent overlaps between data sources

2 Evaluation on full dataset

Table 2: Results of evaluation on the complete dataset.

feature	relation	rho (rand)	rho (butina)	rho (solvent)	rho (ood(novartis))	rho (ood(bayer))
cddd	absolute	0.628±0.03	0.572±0.03	0.231±0.12	0.397±0.02	0.617±0.01
	relative_concat	0.637±0.04	**0.622**±0.05	0.477±0.32	0.686±0.05	0.59±0.02
	relative_diff	**0.646**±0.02	0.615±0.04	0.42±0.34	0.706±0.03	0.609±0.03
ecfp_bit	absolute	0.468±0.05	0.317±0.09	0.074±0.17	0.177±0.08	0.3±0.06
	relative_concat	0.523±0.03	0.471±0.07	0.279±0.28	0.647±0.05	0.281±0.2
	relative_diff	0.533±0.07	0.527±0.04	0.25±0.27	0.654±0.09	0.284±0.18
ecfp_count	absolute	0.564±0.04	0.461±0.05	0.115±0.19	0.167±0.06	0.52±0.02
	relative_concat	0.583±0.02	0.502±0.09	0.38±0.21	0.701±0.05	0.483±0.03
	relative_diff	0.53±0.07	0.428±0.11	0.351±0.21	0.699±0.05	0.475±0.04
prior	absolute	0.458±0.04	0.446±0.05	-0.112±0.17	0.298±0.02	0.571±0.0
rand	absolute	0.035±0.11	-0.064±0.14	-0.004±0.19	0.012±0.1	0.096±0.16
xtb	absolute	0.604±0.04	0.552±0.05	0.429±0.09	0.389±0.03	0.456±0.13
	relative_concat	0.637±0.02	0.595±0.07	**0.531**±0.24	**0.729**±0.01	0.63±0.02
	relative_diff	0.616±0.04	0.583±0.05	0.466±0.26	0.684±0.04	**0.639**±0.01

3 XTB configuration

We used the xtb executable in version 6.5.1 as provided here: `https://github.com/grimme-lab/xtb/releases` The calls to the xtb executable follow this scheme:

```
xtb input.xyz --opt --alpb {solvent} > output.out 2> err.out
```

4 Feature importance

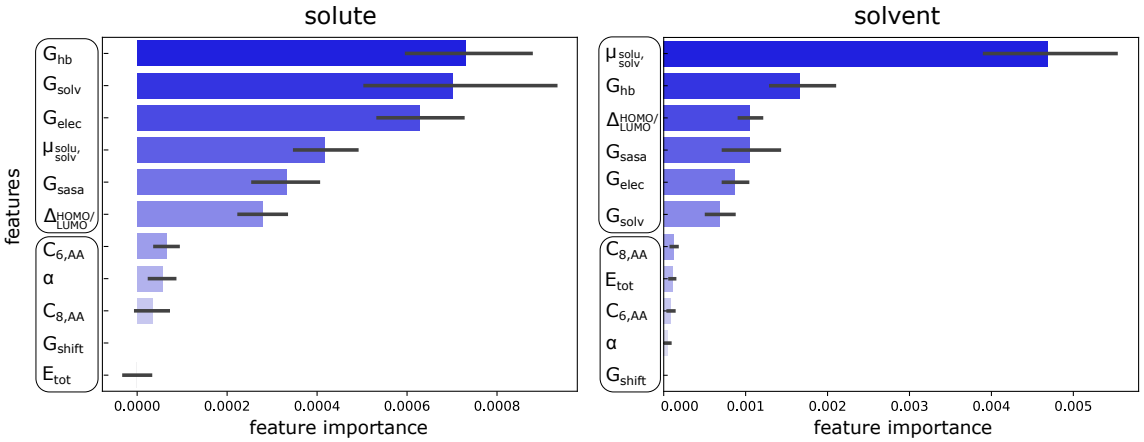


Figure S2: Feature importance analysis for solute features (*left*) and solvent features (*right*).

Table 3: Overview of the features selected from the xTB output.

variable	description	unit
G_{solv}	energy of solvation	kcal/mol
G_{elec}	electronic energy	kcal/mol
G_{sasa}	part. energy (solvent accessible surface area)	kcal/mol
G_{hb}	part. energy (hydrogen bonding)	kcal/mol
G_{shift}	shift energy	kcal/mol
E_{tot}	total energy	kcal/mol
$\Delta_{\text{HOMO/LUMO}}$	HOMO-LUMO gap	eV
$\mu_{solu,solv}$	dipole moment	Db
$C_{6,\text{AA}}$	C_6 dispersion coefficient	au $\cdot$ bohr^6
$C_{8,\text{AA}}$	C_8 dispersion coefficient	au $\cdot$ bohr^8
α	$\alpha(0)$ dispersion coefficient	au

5 Parity plots

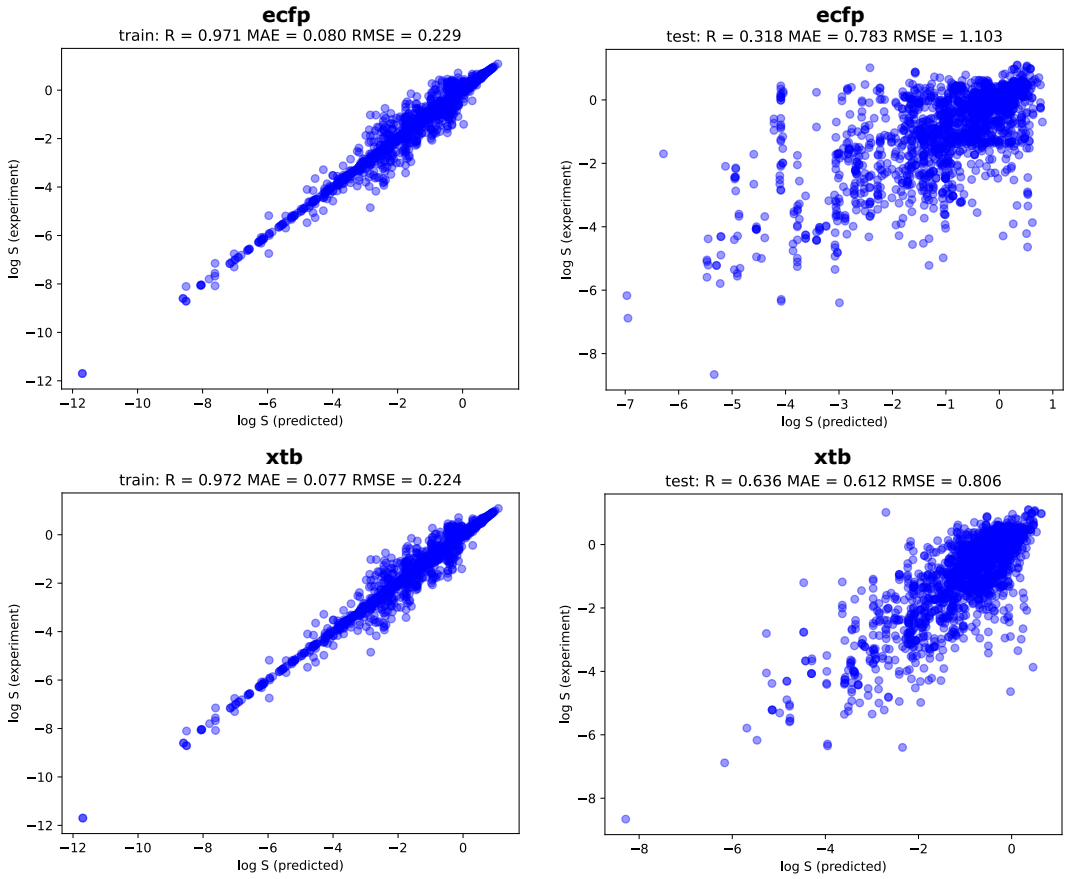


Figure S3: Parity plots of absolute solubility ET regression models for random train/test (60%:40%) split. *top left*: ecfp-model on training set. *top right*: ecfp-model on test set. *bottom left*: xtb-model on training set. *bottom right*: xtb-model on test set.

6 Error analysis by solvent

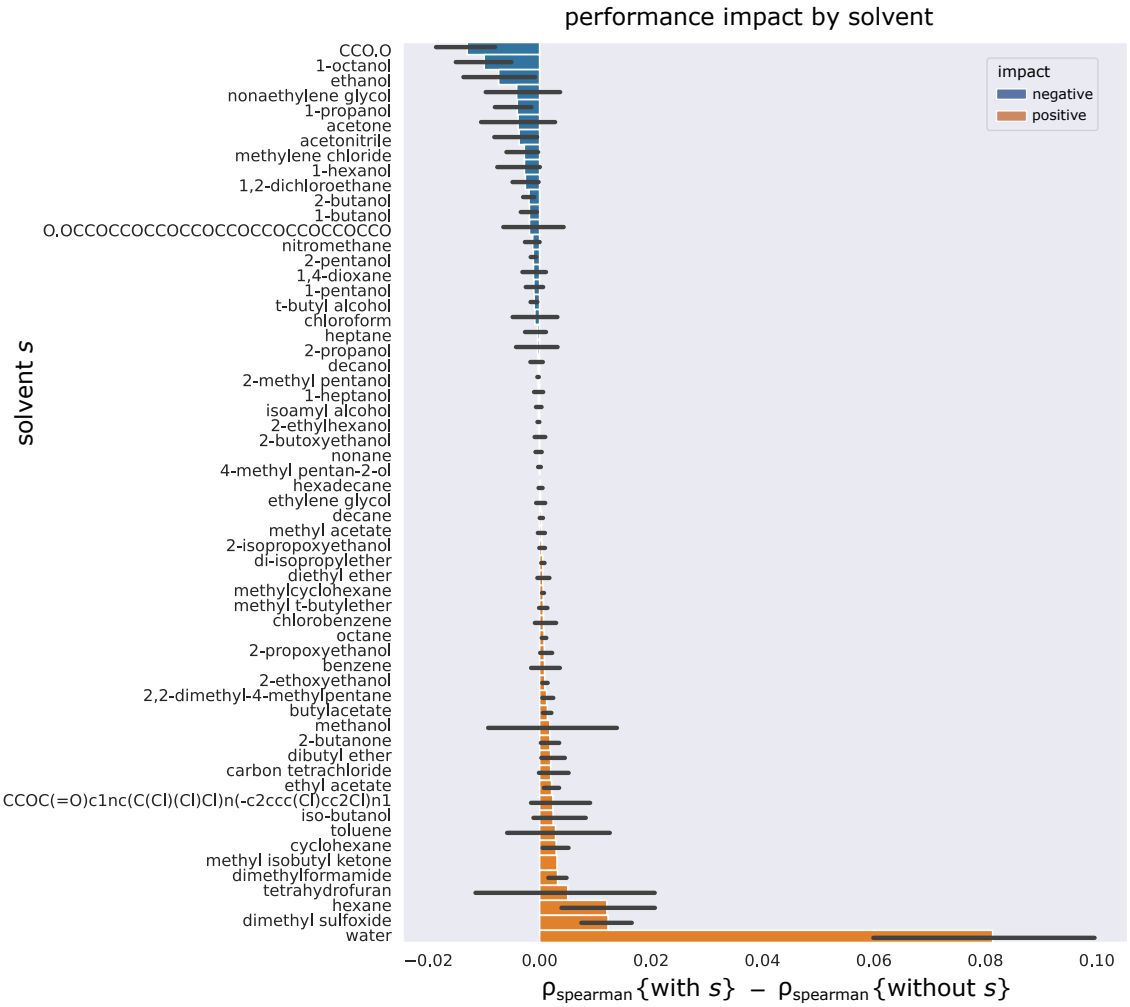


Figure S4: Performance impact by solvent sorted from most negative/hard-to-rank sol- vent at the top (ethanol-water mixture, CCO.O), to most positive/easy-to-rank at the bottom (water, DMSO, hexane). IUPAC names are reported where possible, else the SMILES is reported.

7 Runtime analysis

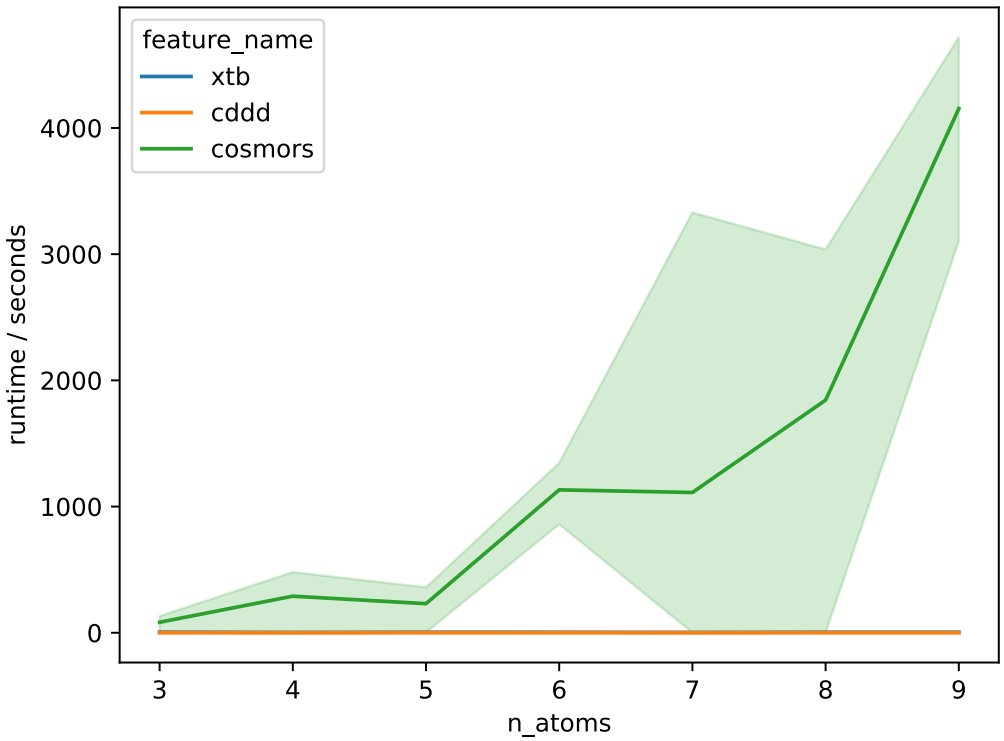


Figure S5: Runtime of the CDDD, XTB, and COSMO-RS featurizers against number of atoms for randomly drawn SMILES containing the elements carbon, nitrogen and oxygen. Extrapolated to typical drug-like molecules, CDDD calulations take seconds, the XTB calculations take minutes, while COSMO+COSMO-RS calculations require runtimes of several days to weeks for a single solute compound.

8 Linear ranking: Comparing algorithms for to-seq step

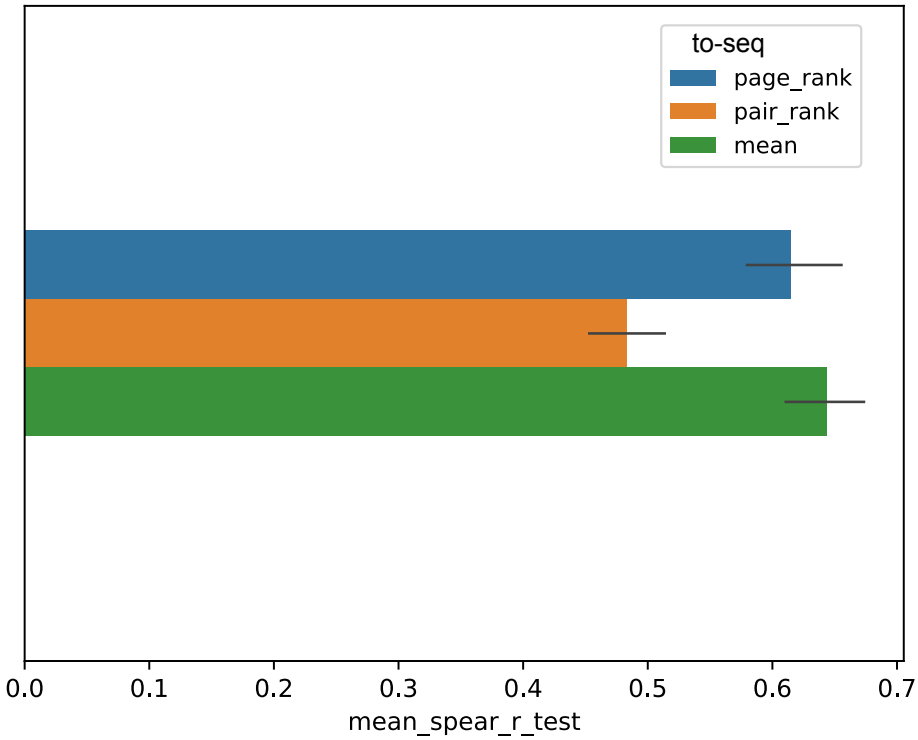


Figure S6: Both pair rank and page rank algorithms were not found to outperform the simpler procedure of mean aggregation. The superiority of the mean procedure over pair rank is statistically significant (T-test: $t \approx 7$, p-value 0.00016), while page rank and mean perform comparably (T-test: $t \approx 1$, p-value 0.32). Therefore, it was decided to adhere to the simpler and slightly better performant mean aggregation procedure.

9 Examples

In the following section, the solvent ranking is shown (screenshots of the web application) for ten neutral molecules and two salts.
In each case, the top shows the solute and solvents selected as input. For convenience, two solvent sets have been defined: The ess (extended-solvent-screening) dataset consisting of 15 representative solvents, and the default dataset consisting of a comprehensive set of 59 organic solvents.
The ranking output is shown as the bottom panel, and consists of a boxplot across five ensemble predictions. This way, the uncertainty of the ranking can be assessed. Lower pos_index values correspond to a higher solubility. The vertical axis has been arranged so that the best solvent (lowest pos_index) is shown at the top.

#1 — 1,1-Dicyclohexylpropane

#2 — hydroxylation of 1,1-Dicyclohexylpropane

#3 — chlorination of 1,1-Dicyclohexylpropane

#4 — Aspirin

#5 — Daptomycin

#6 — Rivaroxaban

#7 — Copanlisib

#8 — adamantane

#9 — cyclodextrin A

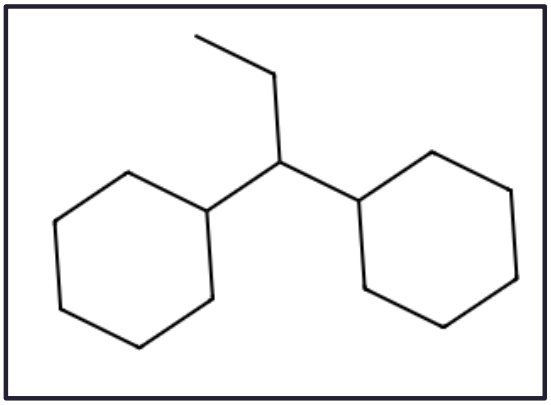
#10 — Abexinostat

#11 — tetramethylammonium chloride

#12 — sodium hexafluorophosphate

#1

solute:

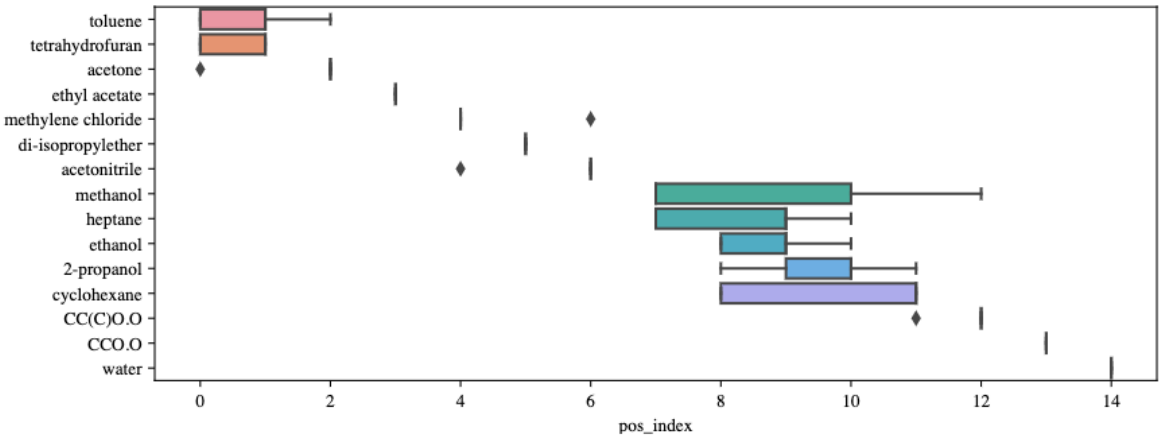


solvents:

ess default +

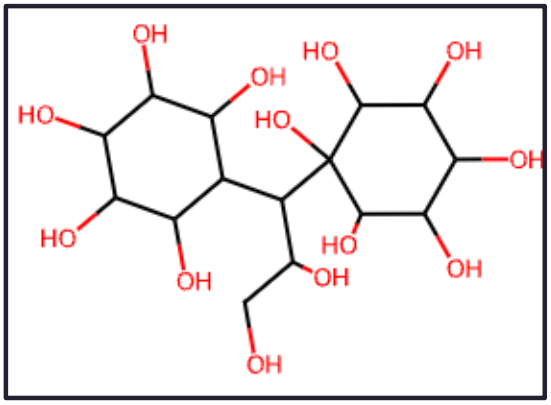
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#2

solute:

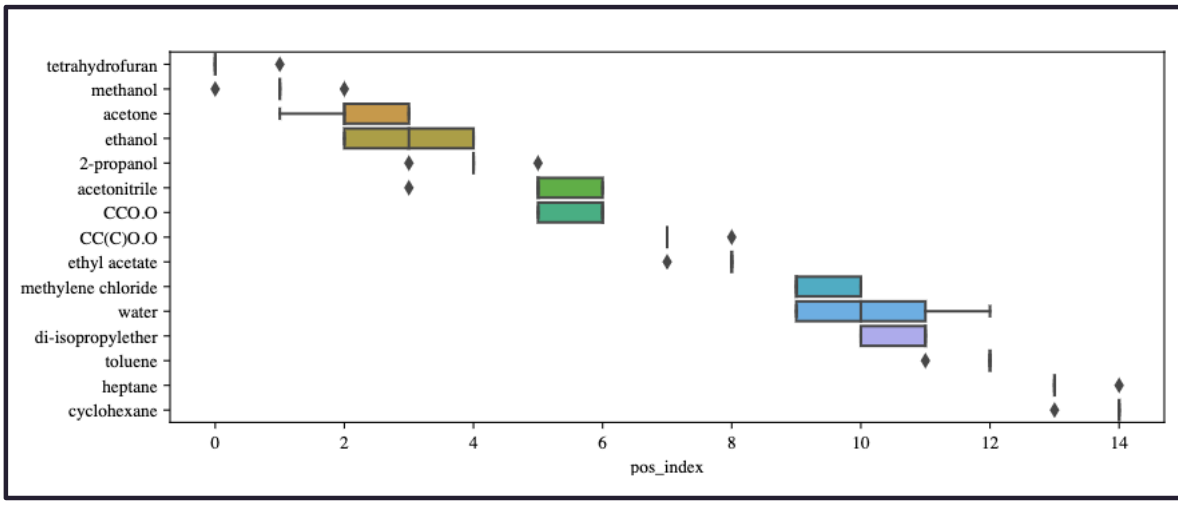


solvents:

ess default +

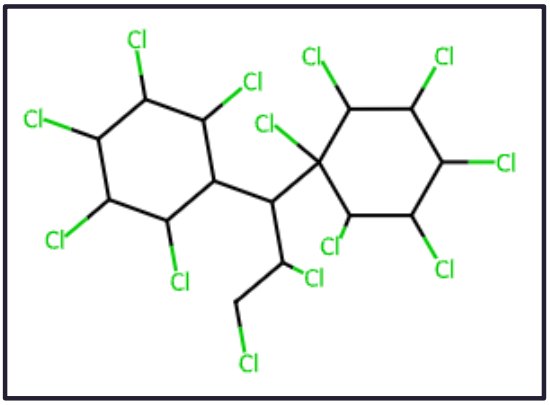
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#3

solute:

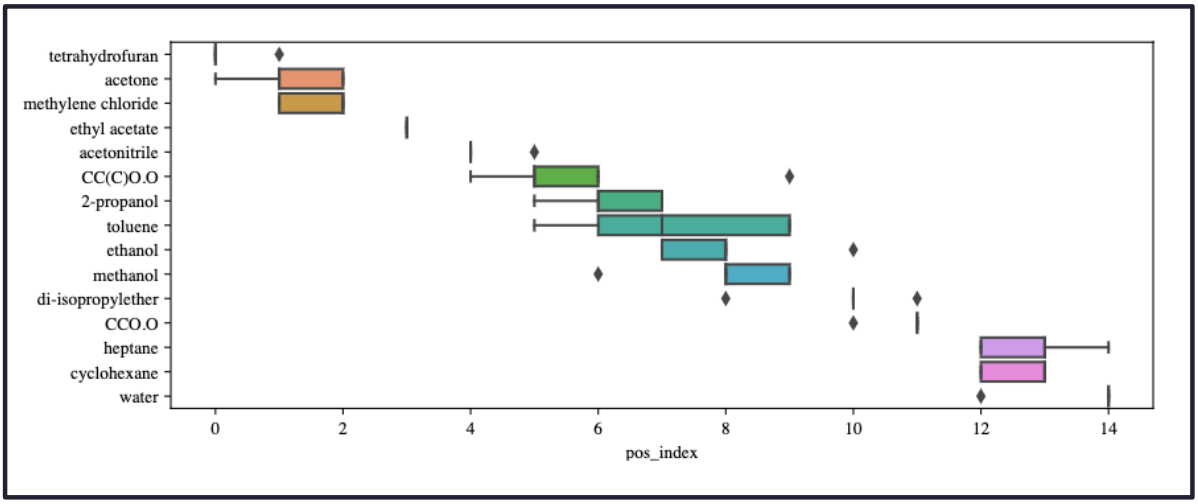


solvents:

ess default +

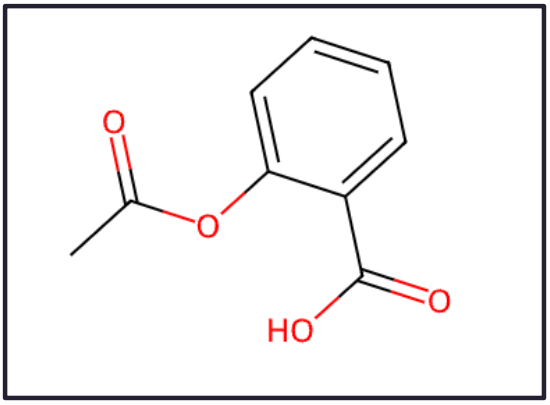
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#4*

solute: Aspirin

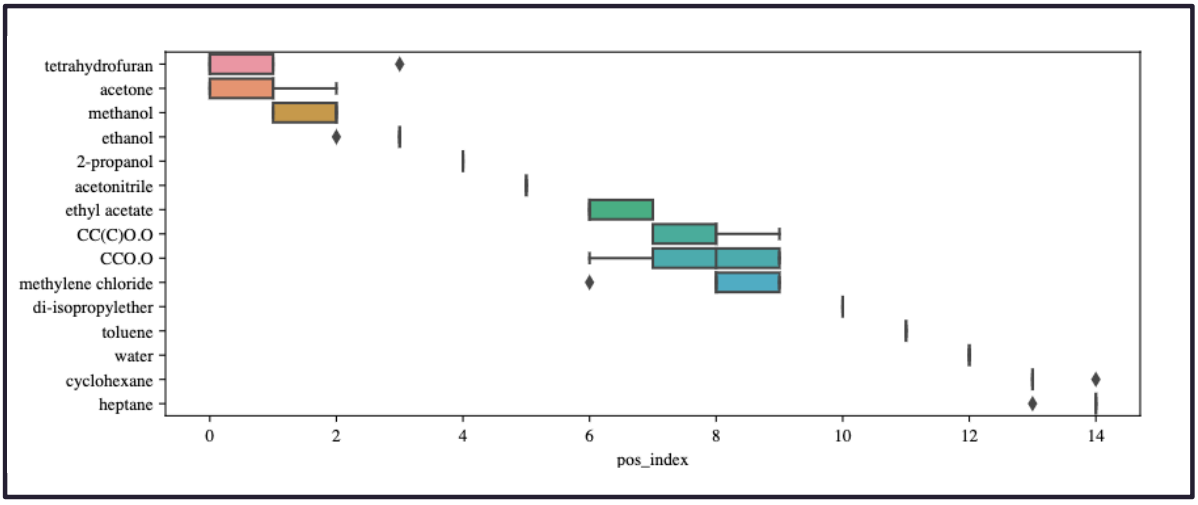


solvents:

ess default +

n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

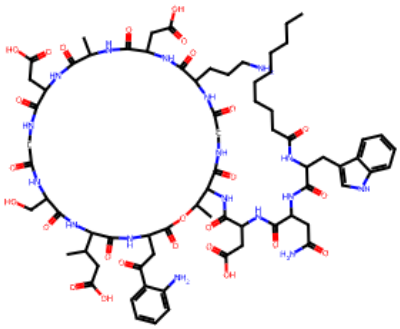
ranking:



*contained in training set

#5

solute: Daptomycin

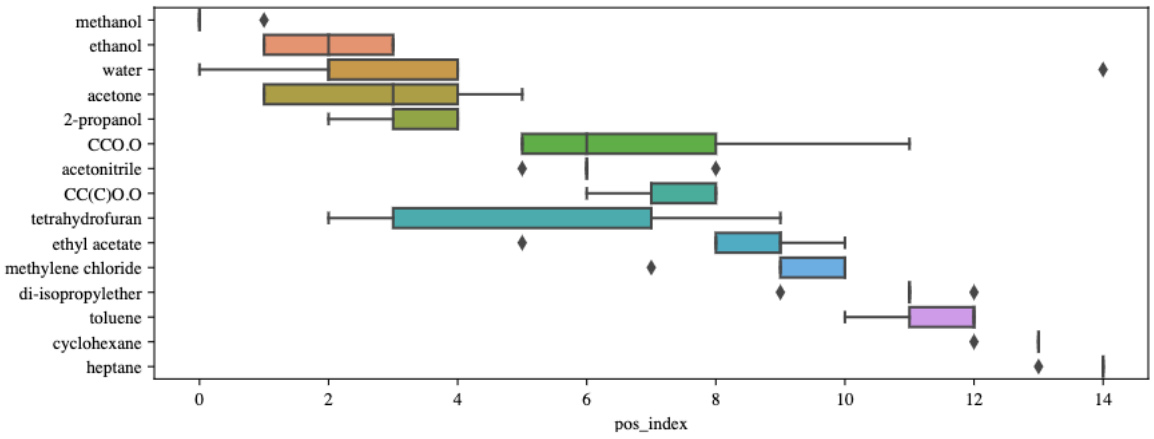


solvents:

ess default +

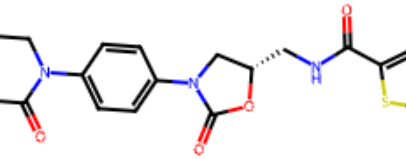
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#6

solute: Rivaroxaban

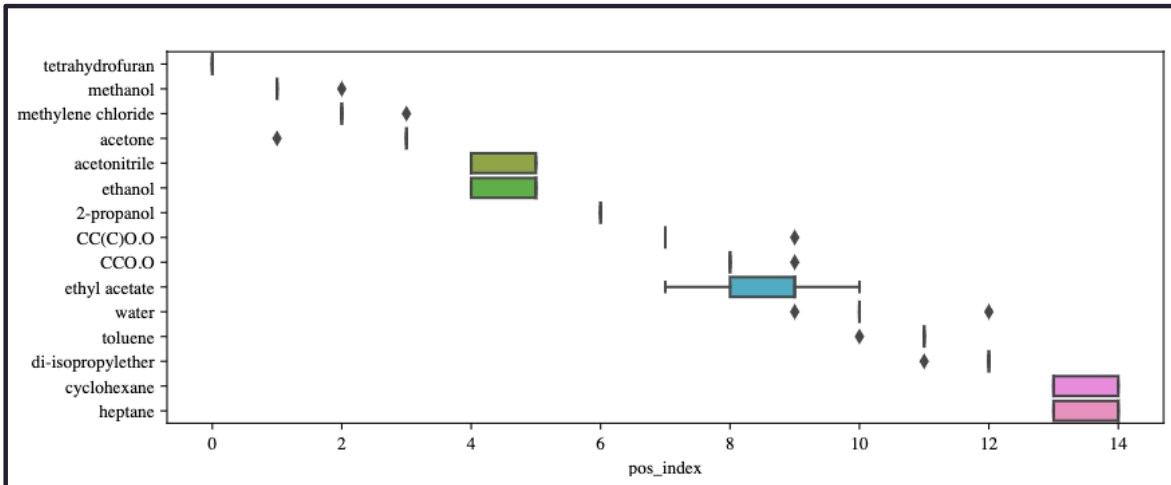


solvents:

ess default +

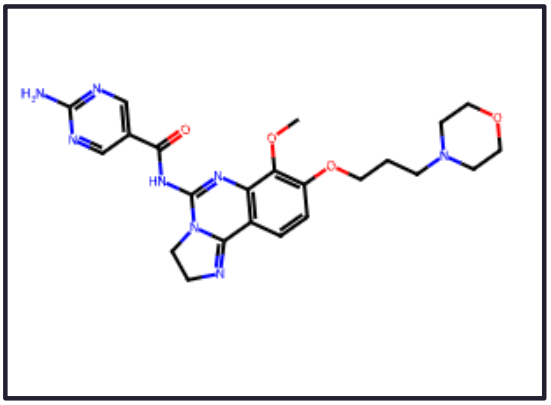
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#7

solute: Copanlisib

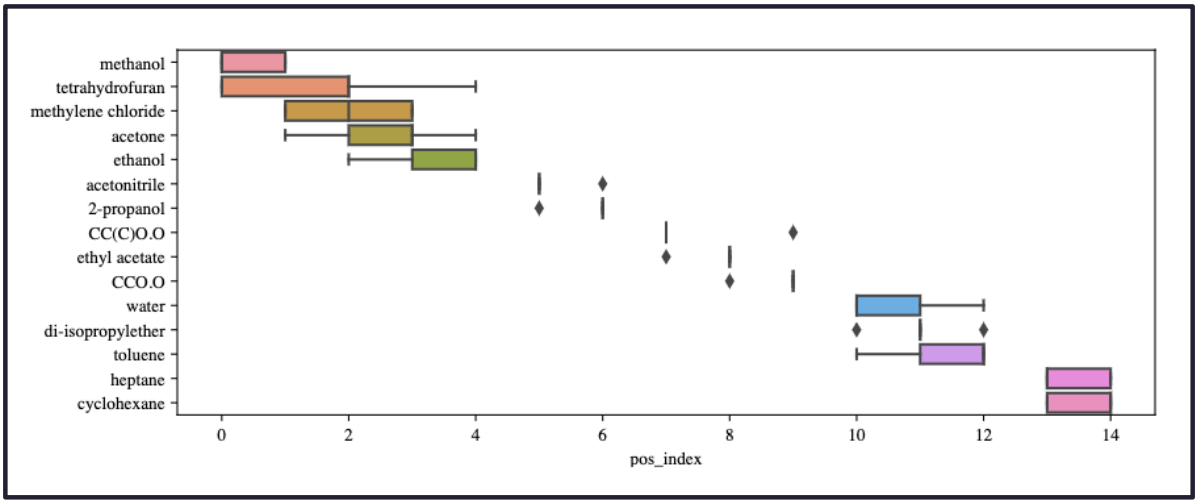


solvents:

ess default +

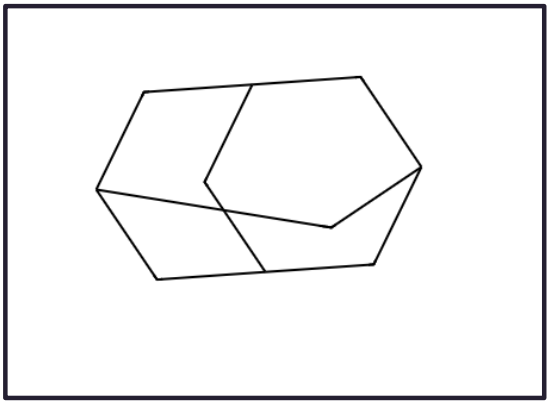
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#8

solute: adamantane

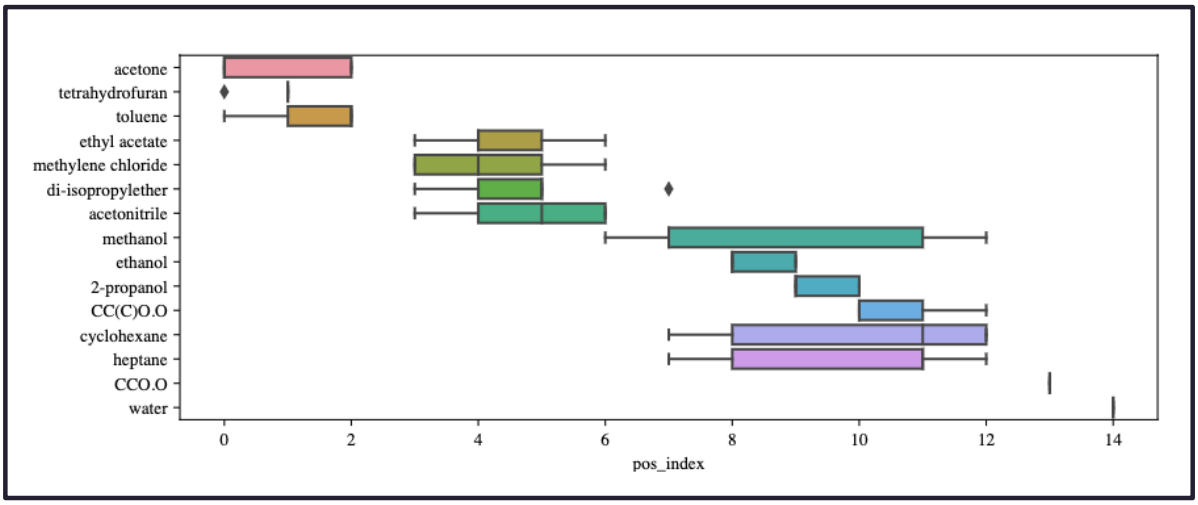


solvents:

ess default +

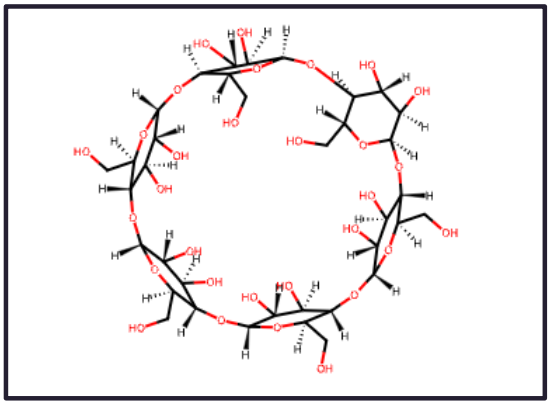
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#9

solute: Cyclodextrin A

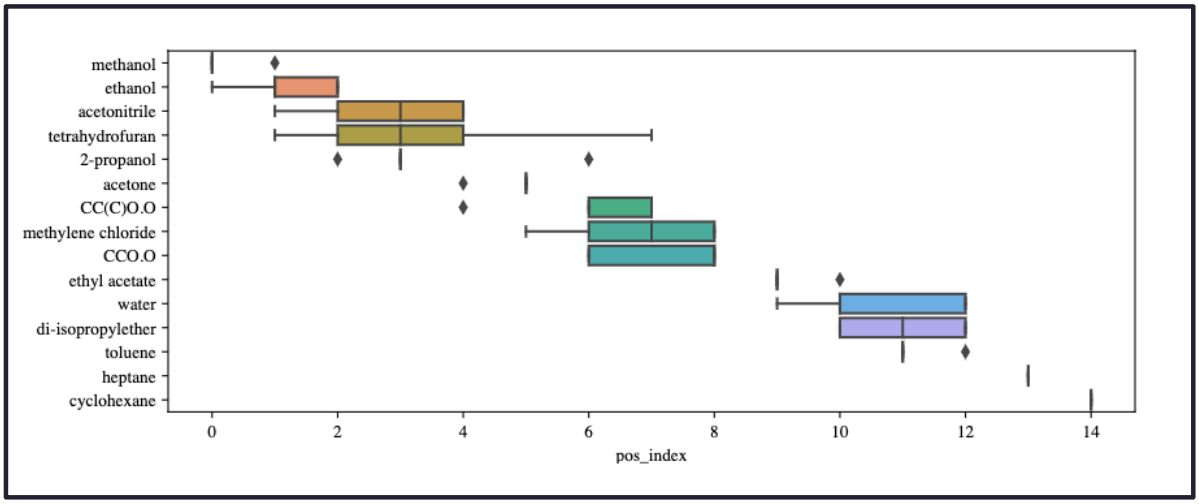


solvents:

ess default +

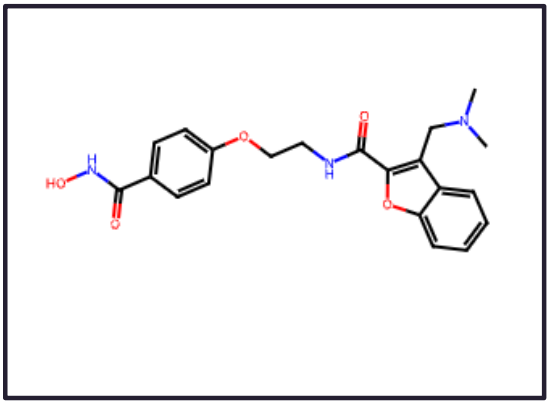
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#10

solute: Abexinostat

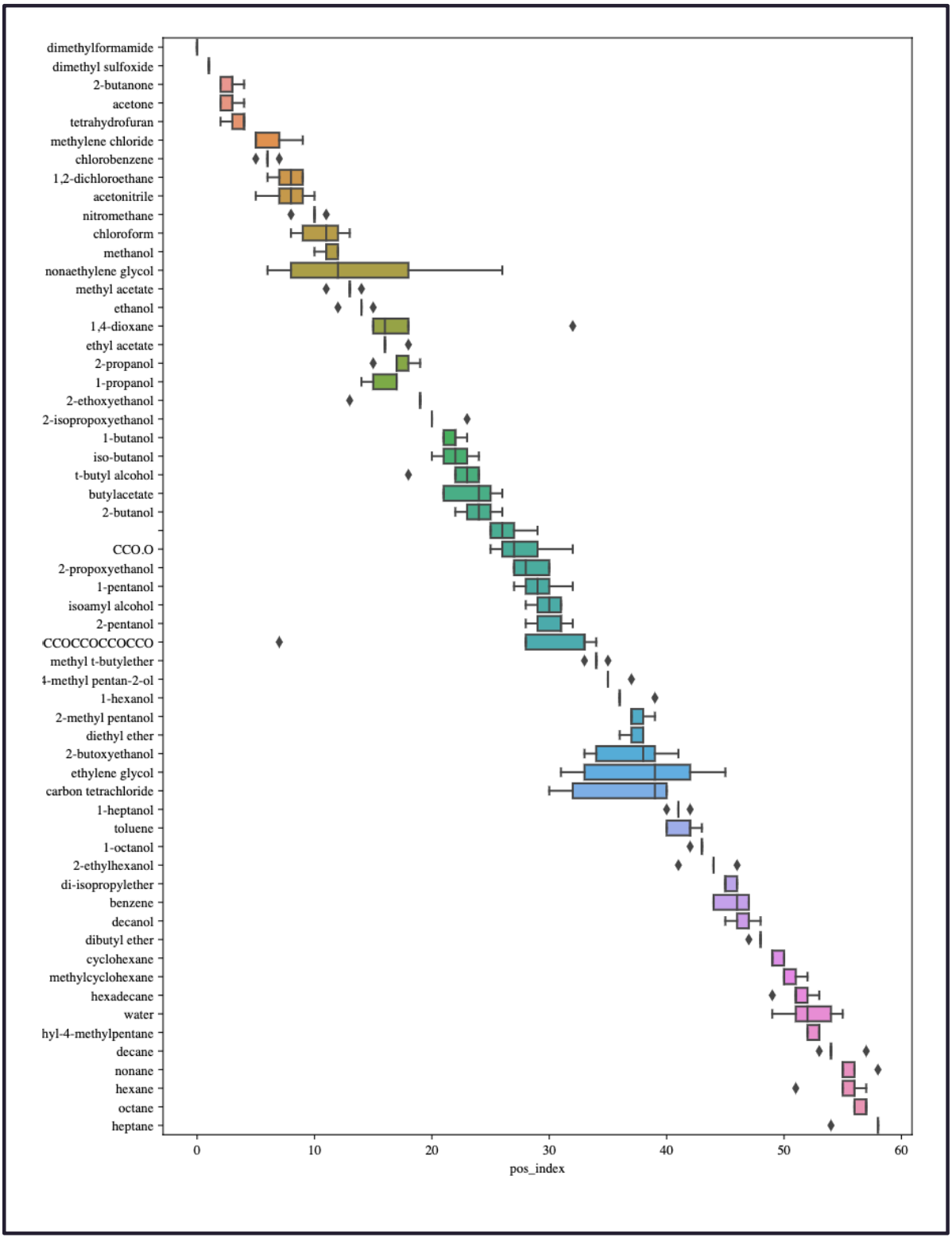


solvents:

ess default +

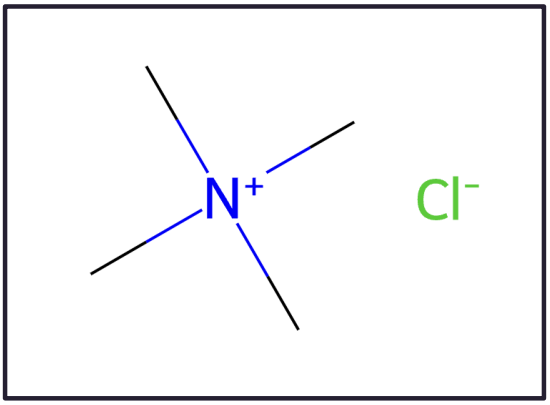
cyclohexane
octane
heptane
hexadecane
nonane
decane
2,2-dimethyl-4-methylpentane
hexane
methylcyclohexane
water
toluene
benzene
...

ranking:



#11

solute: tetramethylammonium chloride

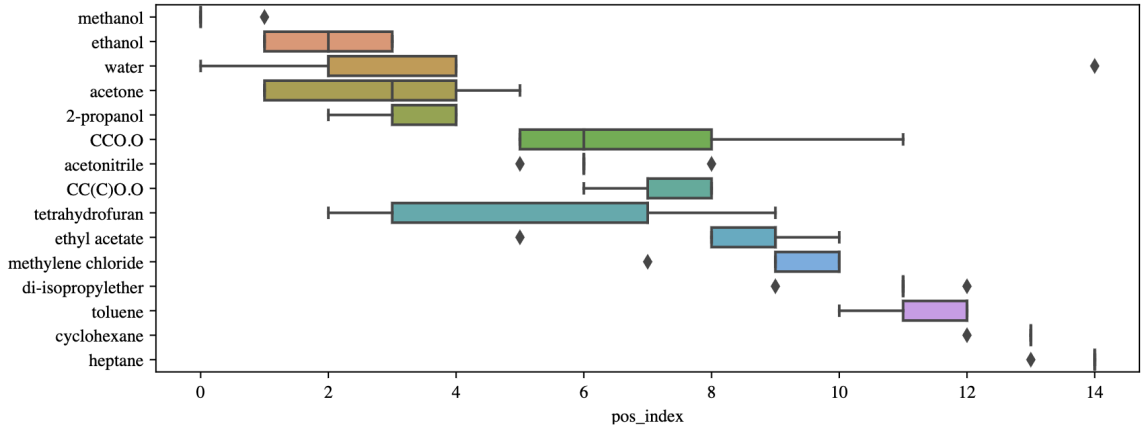


solvents:

ess default +

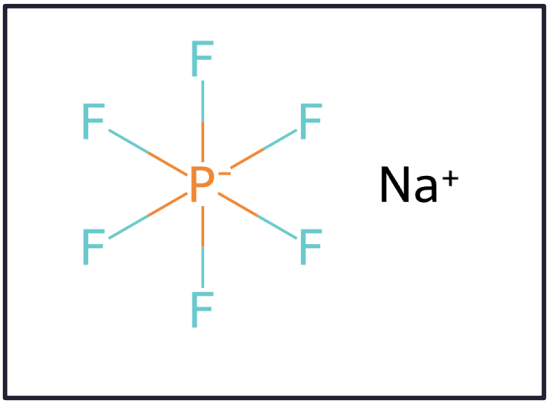
n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking:



#12

solute: sodium hexafluorophosphate



solvents:

ess default +

n-Heptan
Cyclohexan
Diisopropylether
Toluol
Tetrahydrofuran
Aceton
Ethylacetat
ACN
2-Propanol
Ethanol
EtOH / Wasser 1:1
Methanol

ranking: