Come for the drug, stay for the solvent!

Ioana Sovago, Peter Wood

The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB1 2EZ, UK;
isovago@ccdc.cam.ac.uk

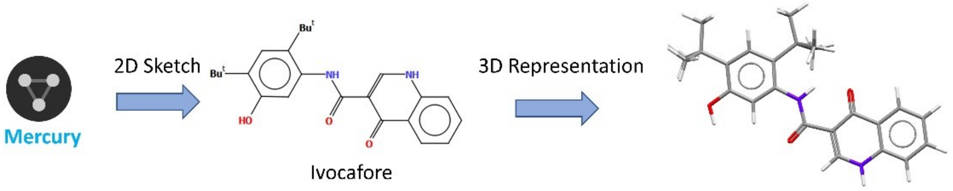
The ability to predict physicochemical properties starting from 2-dimensional molecular information is of paramount importance within the crystal engineering discipline, finding applications in industries as diverse as pharmaceuticals, agrochemicals, and pigments.

Within the CCDC, we have been developing a suite of predictive methods to help scientists assess the likely properties of a given small molecule.

The large amount of data available and the fast-growing Artificial Intelligence (AI) field can now facilitate the development of software tools allowing such predictions. Due to the rise of new and easy to implement Machine Learning (ML) algorithms in recent years[1–3] multiple scientific questions have been answered by applying AI approaches. It is now possible to quickly predict NMR spectra using ML models based on quantum calculations[1] which help with the interpretation of experimental NMR spectra. Space groups can be predicted solely based on Pair Distribution Functions,[2] and for the first time a new antibiotic was identified using ML, thus significantly reducing the number of experiments required.[3]

We have developed a method that provides an early-stage assessment of the likelihood of solvate formation, so that this can be factored into target compound selection and experimental solid form screening can be planned more effectively. Using a sophisticated machine-learning approach we can predict solvate formation quickly using only 2D molecular information. The addition of effective assessment of the likelihood of solvate formation to our solid form design toolbox takes us a big step closer towards more a complete understanding of the behaviour of compounds in the solid state as well as the ability to factor in prediction of solid-state properties in the design stage of a project.

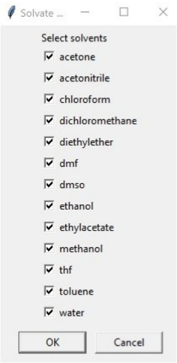
1.



2.

CSD-Python API

solvate_prediction.py



Select solvents

☑ acetone
☑ acetonitrile
☑ chloroform
☑ dichloromethane
☑ diethylether
☑ dmf
☑ dmso
☑ ethanol
☑ ethylacetate
☑ methanol
☑ thf
☑ toluene
☑ water

select among 13 solvents to predict

Prediction for Ivocafore

Solvent	Probability
acetone	0.79
acetonitrile	0.74
chloroform	0.8
dichloromethane	0.53
diethylether	0.93
dmf	0.86
dmso	0.64
ethanol	0.79
ethylacetate	0.88
methanol	0.74
thf	0.66
toluene	0.45
water	0.83

[1] W. Gerrard, L. A. Bratholm, M. J. Packer, A. J. Mulholland, D. R. Glowacki and C. P. Butts, *Chem. Sci.*, 2020, **11**, 508–515.

[2] C. H. Liu, Y. Tao, D. Hsu, Q. Du and S. J. L. Billinge, *Acta Crystallogr. Sect. A Found. Adv.*, 2019, **75**, 633–643.

[3] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackerman, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **180**, 688-702.e13.

Keywords: Machine Learning, Solvent Prediction, CSD, big data