POSTER 20

A Unified AI Framework for Solubility Prediction Across Organic Solvents

Antony D. Vassileiou[a], Murray N. Robertson[b], Bruce G. Wareham[c], Mithushan Soundaranathan[c], Sara Ottoboni[b], Alastair J. Florence[b], Thoralf Hartwig[d], Blair F. Johnston[a,b,e]

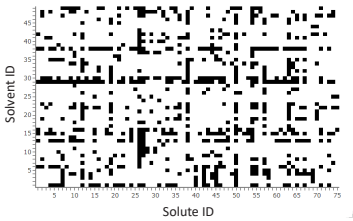
[a] EPSRC ARTICULAR, University of Strathclyde, Glasgow, G1 1RD
[b] EPSRC CMAC Future Manufacturing Hub, University of Strathclyde, Glasgow, G1 1RD
[c] Doctoral Training Centre in Continuous Manufacturing and Crystallisation, University of Strathclyde, Glasgow, G1 1RD
[d] GlaxoSmithKline, GSK Medicines Research Centre, Gunnels Wood Road, Stevenage, SG1 2NY, UK
[e] National Physical Laboratory, Teddington, TW11 0LW

antony.vassileiou@strath.ac.uk

Introduction

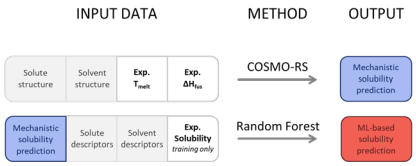
We report on the use of a single, unified dataset for machine learning (ML)-driven solubility prediction across the chemical space. This was a departure from the solvent-specific datasets more commonly used.

Working with a modest dataset of 714 experimental data points spanning 75 solutes and 49 solvents (81% sparsity), a random forest (RF) was trained to enhance an initial solubility prediction provided by Conductor like Screening Model for Real Solvents (COSMO-RS), an industry-standard model based on thermodynamic laws.



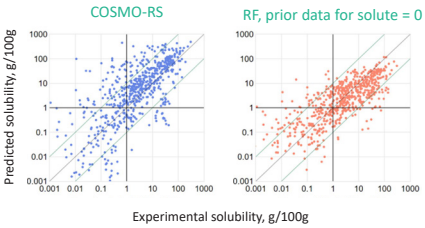
■ = solubility present in dataset

For each data point, a mechanistic prediction was generated from COSMO-RS. This was used with a standard set of molecular descriptors for each solute/solvent as input for ML, which were trained to improve upon the initial result.



Results

An initial ML-based model with a prediction RMSE of 0.75 log S for unseen solutes was produced. This compares favourably with COSMO-RS, which yielded a prediction RMSE of 0.97 for the same dataset.

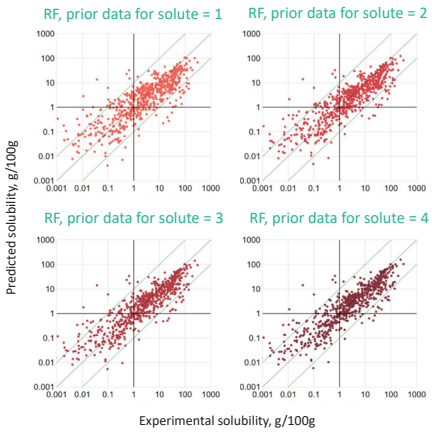


Adopting a multi-solute-multi-solvent data structure enabled the exploitation of valuable relational information between systems.

The effect was major, with even a single experimental measurement of a solute in a different solvent being enough to significantly improve predictions on it, and successive ones improving them further.

Error reduced to a mean RMSE of 0.65 when one instance of the solute (in a different solvent) was included in the training data; this iteratively reduced further to 0.60, 0.57 and 0.56 when two, three and four instances were available, respectively.

Model	COSMO-RS	RF 0	RF 1	RF 2	RF 3	RF 4
RMSE	0.97	0.75	0.65	0.60	0.57	0.56
R^2	0.30	0.58	0.69	0.73	0.76	0.77



Conclusion

- With an improved baseline error rate over COSMO-RS, the application of this framework is a low-risk and overall enhancement in predictive capability
- The standard of performance meets or exceeds those of alternative ML-based solubility models insofar as they can be compared, and does so while spanning a wide range of solvents
- The most emphatic benefits are realised when performing predictions on the same solute in successive solvents: prior knowledge can be incorporated into the training set thanks to its multi-solute-multi-solvent structure, improving performance
- The dataset is greatly extensible, accepting solubility data for any single solute/solvent system and tolerating sparsity
- The framework is modular with respect to the mechanistic model, the molecular descriptor set and the ML algorithm used: operators may freely swap any of them for preferred/available techniques

Ref	Model Domain	RMSE	R^2	No. of same-solvent data points in set	No. of same-solute data points in set	Total data points in set
This study	**49 solvents**	**0.75-0.56**	**0.58-0.77**	**max 51, median 10**	**0-4**	**714**
1	water	0.67	0.81	829	0	829
2	water	0.88	0.45	85	0	85
3	water (narrow range)	0.71	0.76	560	0	560
3	water (wide range)	0.71	0.93	900	0	900
3	ethanol	0.79	0.53	695	0	695
3	benzene	0.54	0.75	464	0	464
3	acetone	0.83	0.42	452	0	452

Comparison of model performance and data requirements with current literature

1. Lovrić, M. et al. (2021) *J Chemom.* 35(7–8), p3349. doi: 10.1002/CEM.3349.
2. Palmer, D. S. and Mitchell, J. B. O. (2014) *Mol Pharm.* 11(8), p2962–2972. doi: 10.1021/mp500103r.
3. Boobier, S. et al. (2020) *Nat Commun.* 11(1), p5753. doi: 10.1038/s41467-020-19594-z.