iPRD

UNIVERSITY OF LEEDS

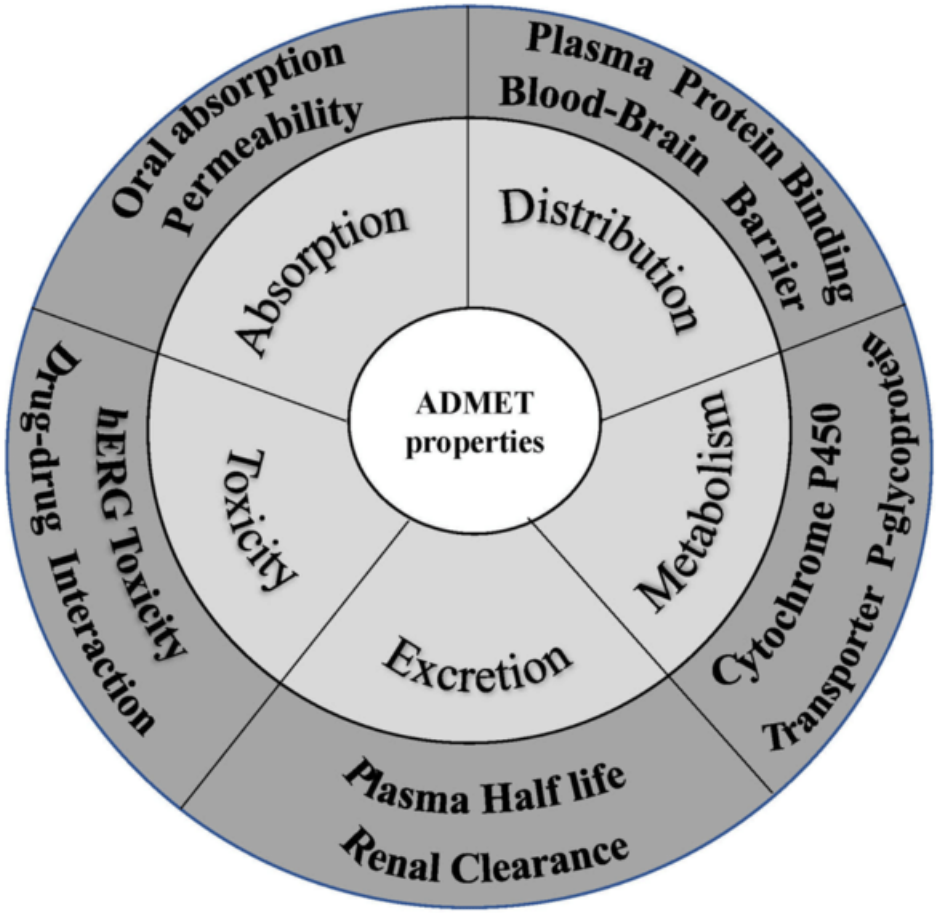
Machine Learning and Solubility Prediction

Dr Bao N. Nguyen

Dec 2021

Why solubility?

- ***Aqueous solubility*** is an important ADMET property for drug candidates (linked to bioavailability).



- It’s surprisingly difficult to predict!

The challenges of solubility prediction

Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?

Antonio Llinàs*, Robert C. Glen and Jonathan M. Goodman*

View Author Information

Cite this: *J. Chem. Inf. Model.* 2008, 48, 7, 1289-1303
Publication Date: July 15, 2008
https://doi.org/10.1021/ci800058v

Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets

Antonio Llinas and Alex Avdeef*

Cite this: *J. Chem. Inf. Model.* 2019, 59, 6, 3036-3040
Publication Date: May 1, 2019
https://doi.org/10.1021/acs.jcim.9b00345

The challenges of solubility prediction

Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules?

David S. Palmer*† and John B. O. Mitchell*‡
† Department of Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, U.K.
‡ Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, University of St. Andrews, Purdie Building, North Haugh, St. Andrews, Scotland KY16 9ST, U.K.

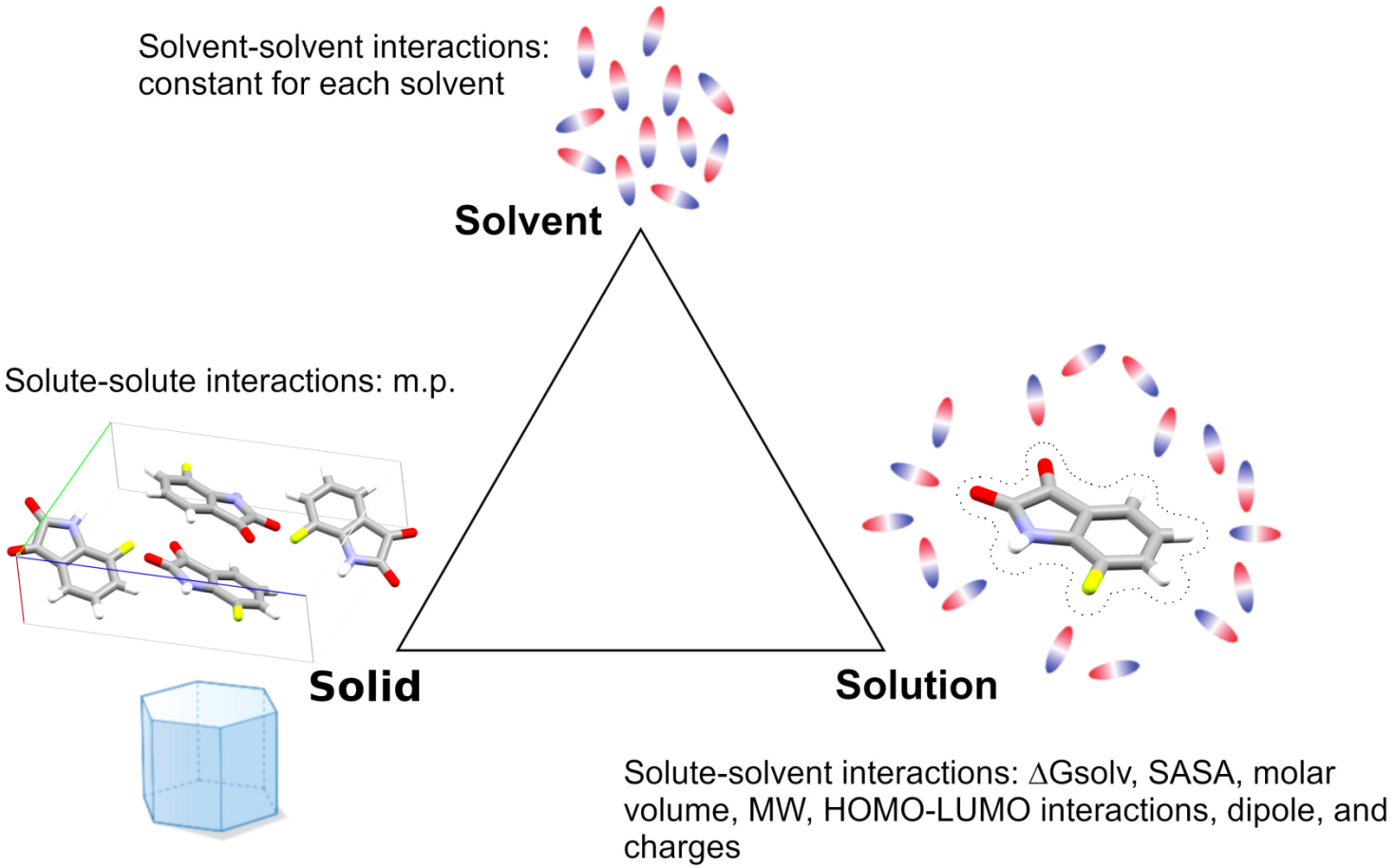
Mol. Pharmaceutics, **2014**, *11* (8), pp 2962–2972
DOI: 10.1021/mp500103r
Publication Date (Web): June 11, 2014

Cite this: *Mol. Pharmaceutics* 2014, 11, 8, 2962-2972

- Typical experimental error is LogS $\pm$ 0.5-0.7
- Traditional metrics R^2 and RMSE can be misleading given the errors in the training data
- Two predictive thresholds (%LogS $\pm$ 0.7 and %LogS $\pm$ 1.0) are used to evaluate the models

Approach in Nguyen group

- DFT (B3LYP/6-31+G(d)) was used to generate the descriptors for 900 compounds



Nature Commun. **2020**, *11*, 5753

Initial model metrics

- ***Water_set_wide*** (LogS = -12 – 2)

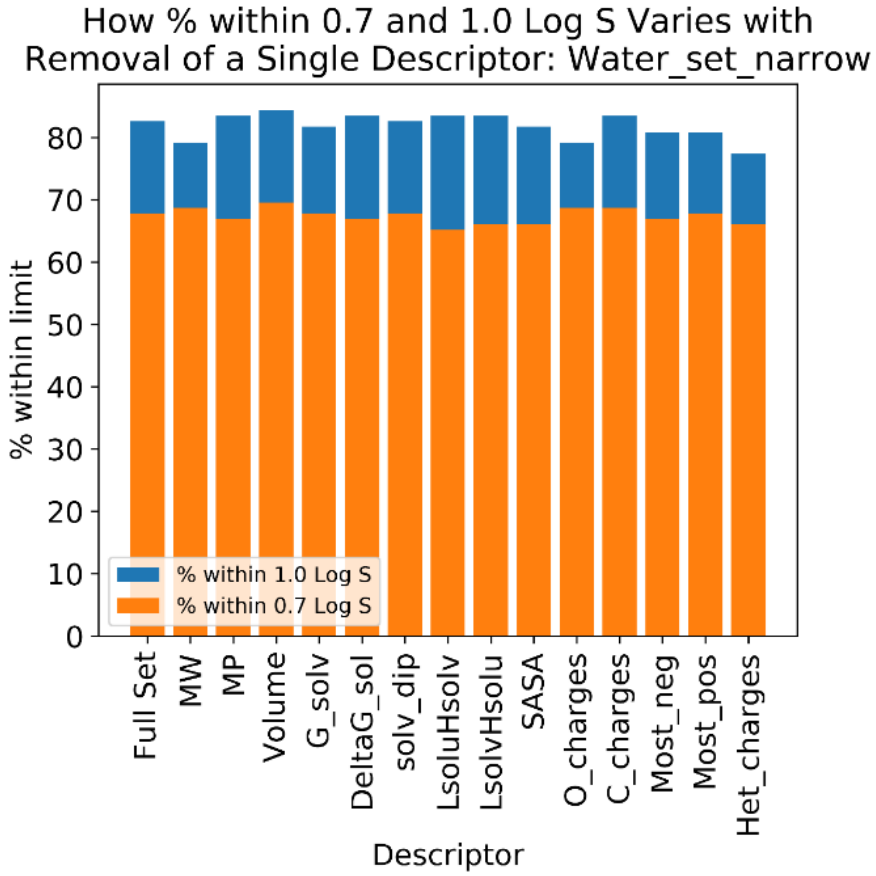
Metric	MLR	PLS	ANN	SVM	GP	RF	ET	Bag	Stdev
R^2	0.80	0.80	0.90	0.89	0.88	0.90	**0.92**	0.90	0.01
RMSE	1.15	1.16	0.83	0.85	0.89	0.81	**0.73**	0.81	0.05
%LogS±0.7	50.5	52.6	58.9	**71.6**	68.4 (91.6)	58.9	63.2	57.9	5.69
%LogS±1.0	66.3	67.4	77.9	80.0	74.7 (94.7)	**82.1**	**82.1**	**82.1**	1.88

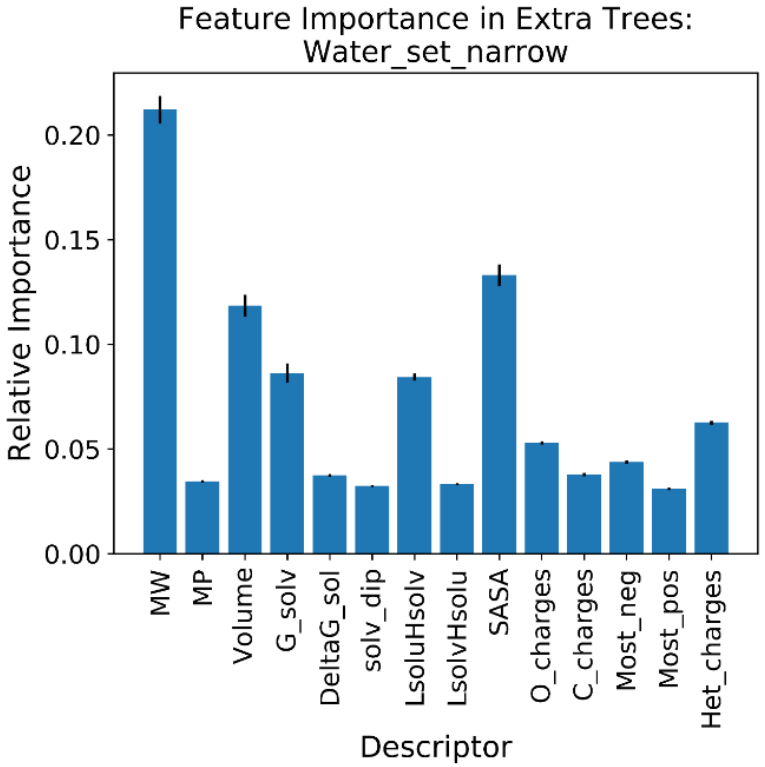
- The predictions using non-linear methods are very similar, and depends on the training data and descriptors far more than on the ML method

Nature Commun. **2020**, *11*, 5753

Importance of each descriptor

- Impact of skipping descriptors in H_2O





- Important descriptors: *MW*, *molar volume*, ΔG_{sol}, *SASA*, *most_neg*, *Het_charges*
- Which one can be more accurately calculated?

Nature Commun. **2020**, *11*, 5753

Improvement of solvation energy

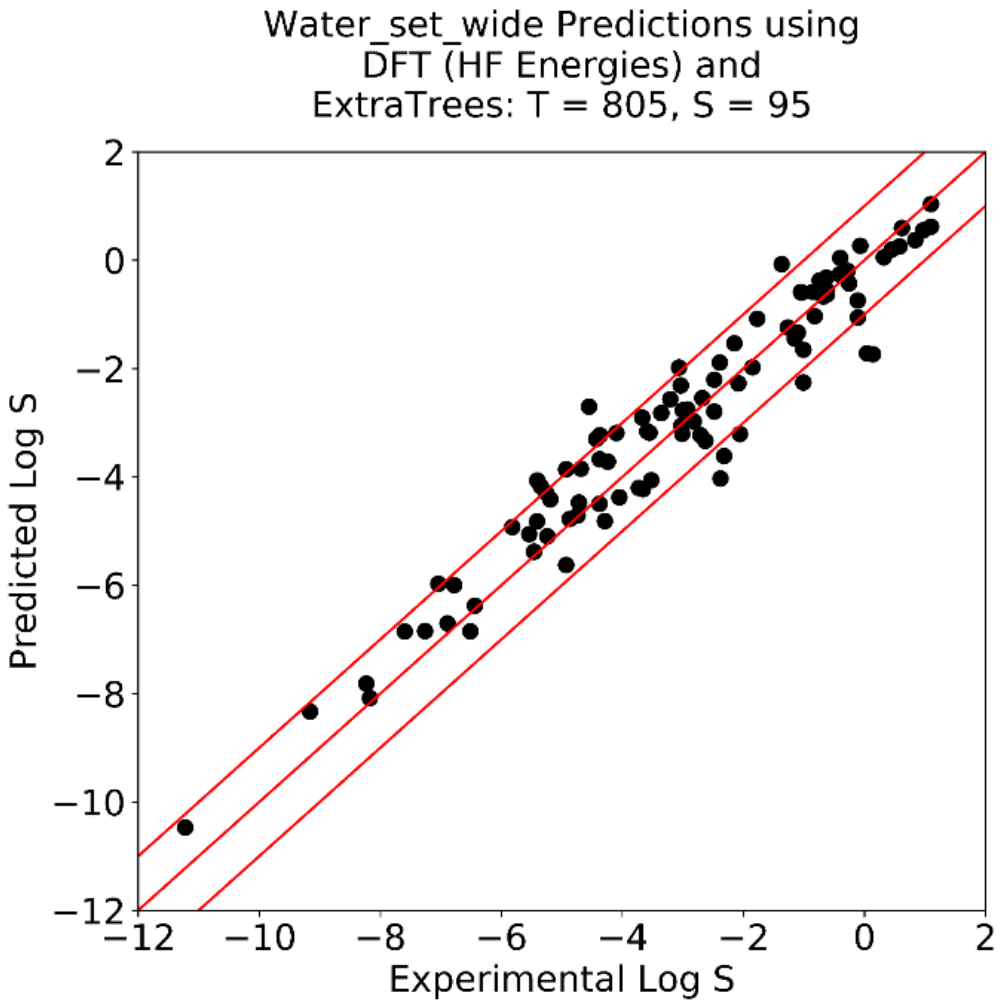
Method	R^2	RMSE	%LogS ±0.7	%LogS ±1.0	R^2	RMSE	%LogS± 0.7	%LogS ±1.0
	Old descriptors (PCM solvation model)				**New descriptors (HF SMD solvation model)**			
ANN	0.90	0.84	**58.9**	**78.9**	0.91	0.81	**68.4**	**84.2**
SVM	0.89	0.85	**71.6**	**78.9**	0.90	0.82	**72.6**	**83.2**
RF	0.90	0.83	**60.0**	**75.8**	0.90	0.82	**63.2**	**80.0**
ET	0.93	0.71	**66.3**	**84.2**	0.93	0.70	**69.5**	**84.2**
Bag	0.90	0.82	**57.9**	**76.8**	0.90	0.83	**65.3**	**81.1**
GP	0.88	0.89	**68.4**	**73.7**	0.90	0.80	**70.5**	**82.1**

- ***Water_set_wide*** (LogS = -12 to 2)
- PCM = Polarizable continuum model (*Chem. Rev.* **2005**, *8*, 2999)
- SMD = Solvent model based on density (*J. Phys. Chem. B* **2009**, *113*, 6378)

J. Chem. Phys. **149**, 104102 (2018)

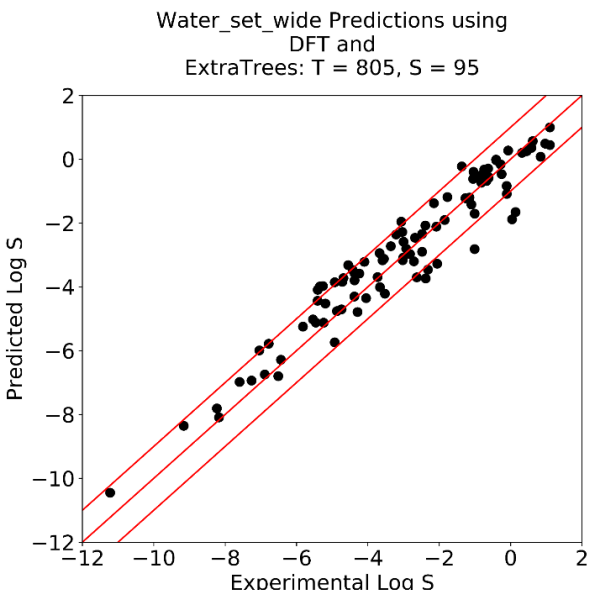
Improvement of solvation energy

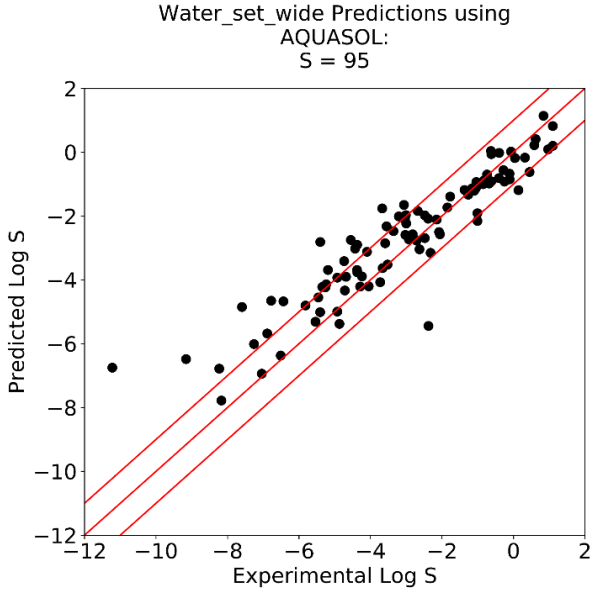
- ***Water_set_wide***
  (LogS = -12 to 2)

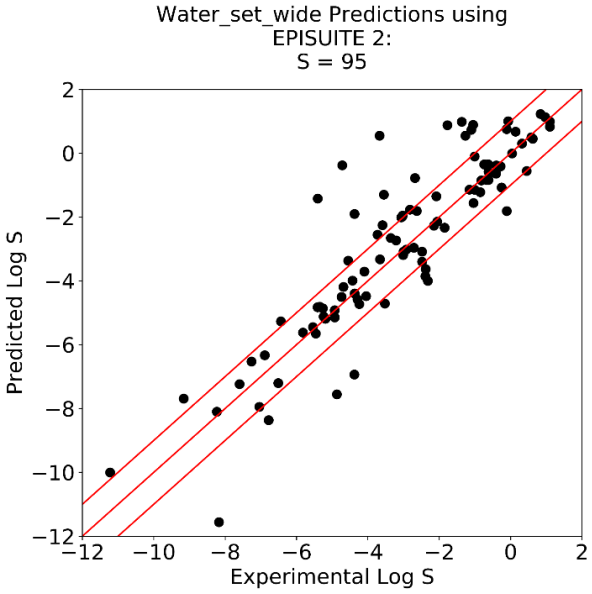


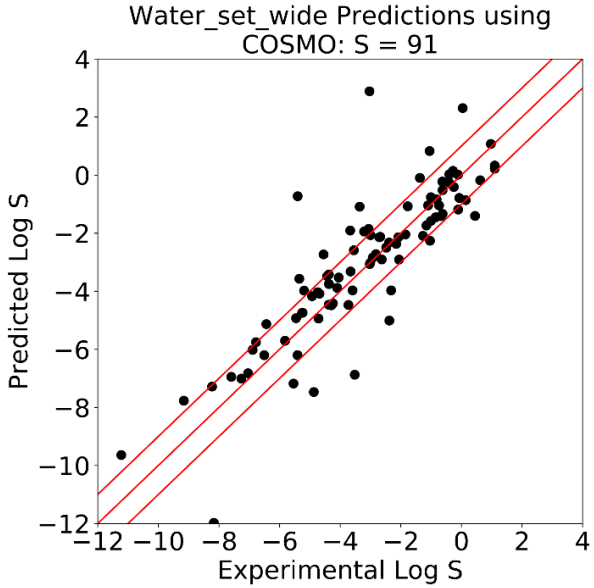
Nature Commun. **2020**, *11*, 5753

Benchmarking in H_2O









- AQUASOL and EPISUITE are standard FDA tools for solubility prediction

Nature Commun. **2020**, *11*, 5753

Project details

- We want to explore other approaches to interpretable Machine Learning models for solubility prediction
- You will be given datasets generated by molecular modelling on the same 900 compounds
- Details on algorithms, code (R, Python) will be guided by Dr Gusnanto and Dr Cutillo