4/16/2020 Week 13 Module 2
Solubility

- This module is a lecture describing
  - the importance of solubility prediction
  - the application of QSPR to solubility prediction
    - based on the 2008 solubility challenge of J.M. Goodman [2, 3]
    - based on deep learning
  - free energy calculations of solubility
  - is based on a lecture by David Mobley [1]
- Even though it is a basic concept that you learned about in general chemistry, there are a lot of subtleties to solubility that
  - you probably never thought about!
  - I never thought about before preparing for this class
- Congratulations, this is the last lecture of Chem 456!

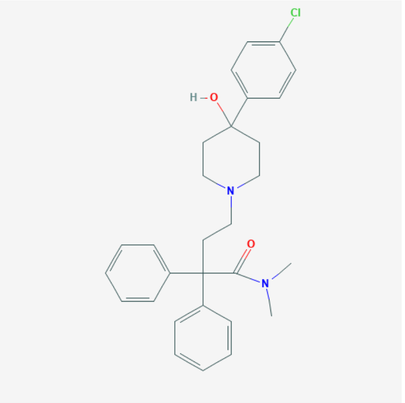
The importance of solubility prediction

Why predict solubility? [2]

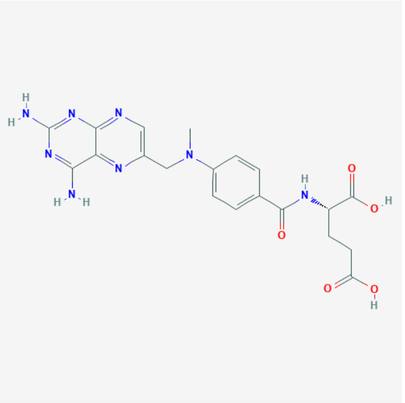
- Drug candidates should have sufficient solubility to be
  - tested in biological assays
  - stored
  - bioavailable
- Rules of thumb are unreliable

Poor solubility is a major problem [2]

- Pharmacokinetics
  - Up to 40% of drug discovery programs abandoned due to pharmacokinetics - the movement of drugs within the body
  - Low solubility is associated with problems with pharmacokinetics
- Precipitation
  - Loperamide (diarrhea drug) can sometimes precipitate when injected
  - Methotrexate (cancer drug) has kidney toxicity due to precipitation
- Polymorphs can affect solubility and have different bioavailability, e.g. ritonavir (HIV drug) had to be reformulated

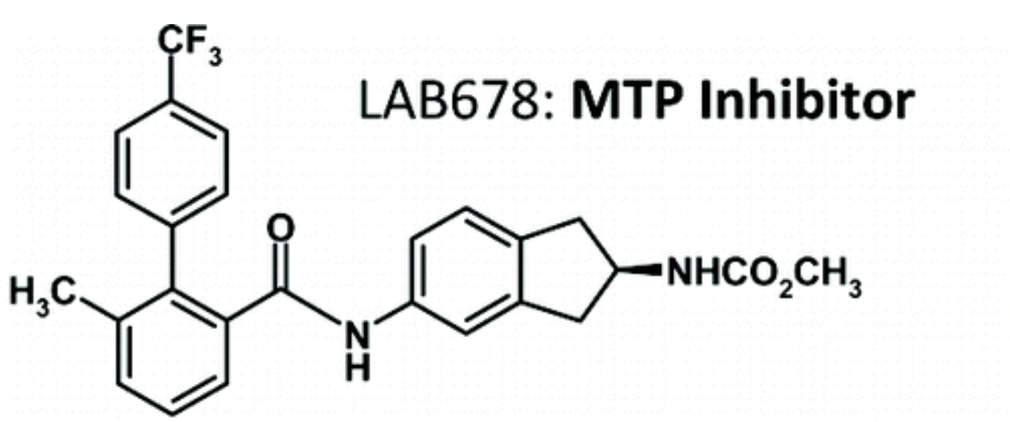


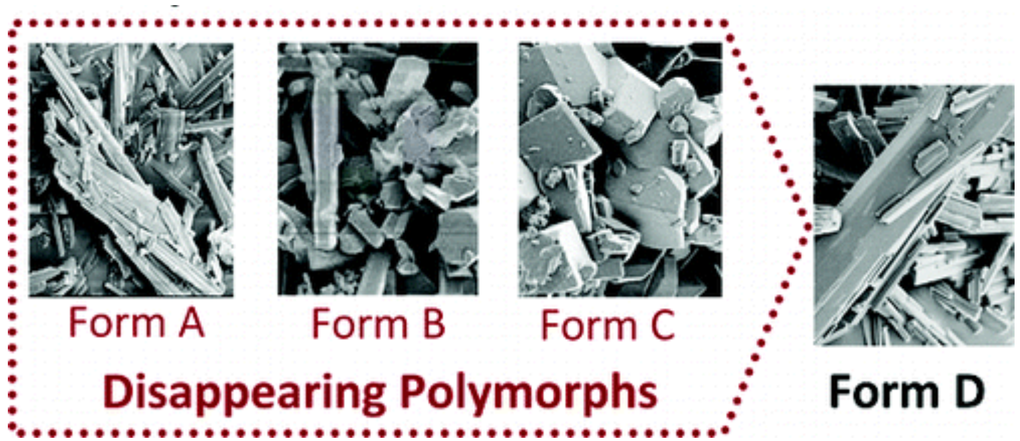
Loperamide



Methotrexate

Polymorphs are chemical engineering problem





Chemical structure and SEM images from [4]

- LAB687 (lower triglyceridess and LDL cholesterol)
- Early on, observed form A and B
- Later, found form C
- Form D observed in scale up, and began to predominate!

Why work with low-solubility compounds?

- Practical reason
  - increased weight and lipophilicity of test compounds from high-throughput synthesis and screening
- Principled reasons
  - drugs usually need to permeate cell membranes
  - increased binding affinity - the molecule prefers the complex over the solvent - often associated with reduced solubility

How are solubility problems hidden?

- use of solvents, e.g. DMSO, in early-stage assays
- pH dependence
  - free acid/free base is less soluble than salt
  - intrinsic solubility is solubility of free acid/free base
- kinetic solubility - appearance or disappearance of precipitate at different concentrations than the thermodynamic saturating concentration (sub- or supersaturation)
- polymorphs not always known

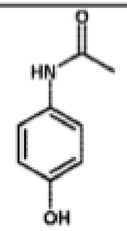
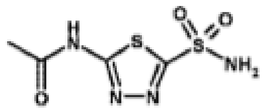
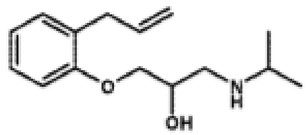
- These problems are also reasons why solubility is difficult to measure and predict

The Solubility Challenge

Solubility Challenge (2008) [2, 3]

- measured
  - intrinsic solubility (free acid/free base)
  - highly reproducible (thermodynamic versus kinetic solubility)
- published 100 solubilities
- challenged researchers to predict 32 more
- received >100 entries between July and September 2008

Examples from the dataset [2]

Structure	Name	MW Neutral Form	pKa	Kinetic Solubility µM µg/ml	Intrinsic Solubility µM µg/ml
	Acetaminophen	151.17	9.52 ± 0.01	161700 ± 7000 µM 24400 ± 1060 µg/ml	86300 ± 7000 µM 13000 ± 1060 µg/ml
	Acetazolamide	222.25	8.75 ± 0.02 7.31 ± 0.04	6100 ± 3840 µM 1360 ± 850 µg/ml	3670 ± 80 µM 816 ± 18 µg/ml
	Alprenolol	249.36	9.47 ± 0.01	5080 ± 50 µM 1266 ± 12 µg/ml	2320 ± 40 µM 580 ± 10 µg/ml
	Amantadine	151.25	10.48 ± 0.01	17300 ± 3960 µM 2620 ± 600 µg/ml	14000 ± 1180 µM 2120 ± 180 µg/ml

Overall performance in the solubility challenge

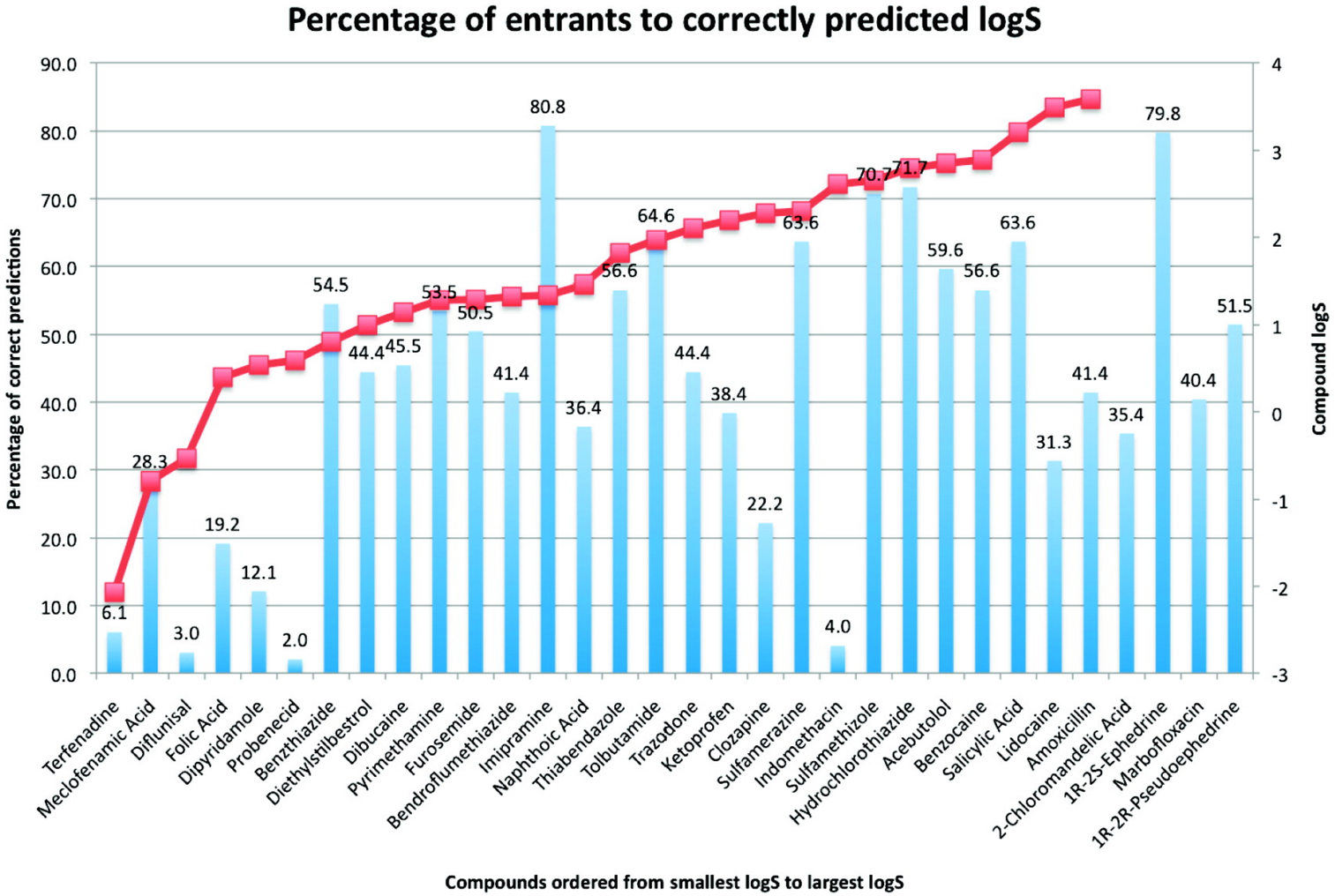


Figure 2 from ref [3]
“Correct” is within 0.5 logS
Blue bars are percentage correct
Red is actual logS

Stated major findings [3]

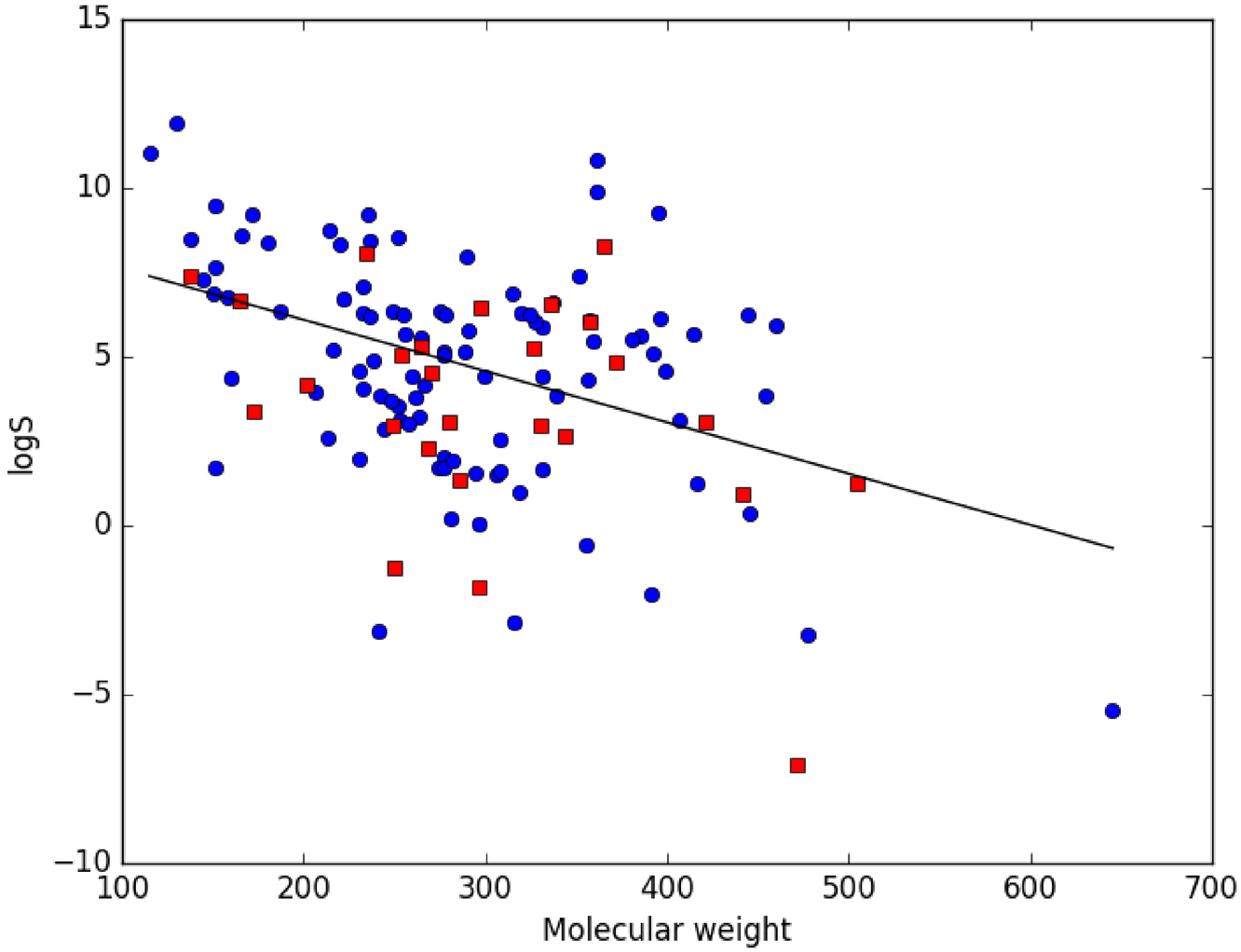
- For measured set of 28 compounds, percent correct on
  - absolute S (±10%) ranges from 0%-18%
  - logS (±0.5 logS) ranges from 10-61%
  - Worst two compounds 2% and 4% correct by logS
- R^2 ranges from 0 to 0.65
- No polymorph predictions
- Compounds with “normal” solubility (logS = 0.5 to 3) were easiest
- Many methods, but none clearly the best

QSPR prediction of solubility

Historically informative descriptors [5]

- Molecular size or surface area - cost of forming cavity in water
- Octanol:water partition coefficient (P) - solubility relative to a non polar phase
- Melting point - higher means more stable solid
- Hydrogen bonding - favors water solubility
- Atom/group contributions
- Molecular connectivities

Molecular weight has some predictive power



From Ref [1]
Blue: training
Red: test
Black: predictions

QSPR in the solubility challenge

- Hewitt et al tried four QSPR models [5]
  - Multiple linear regression (426 descriptors, genetic algorithm, no more than 5 used at once)
  - Artificial neural network
  - Category-specific models based on H bond ability
  - Various commercial QSPR models
  - Consensus, mean of four models
- The best was a three-descriptor linear regression
  - based on log P, boiling point, and R maximal autocorrelation of lag 2 (related to size and connectivity).
  - $R^2_{train} = 0.74, R^2_{test} = 0.51$
- "none of the other modeling approaches used in this study was able to improve upon the predictions made by the MLR model"

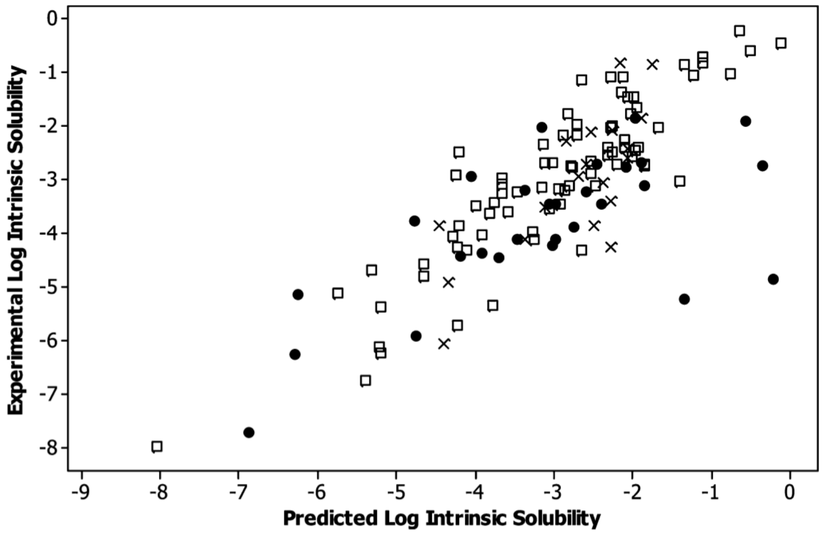


Figure 1. Plot of experimental versus predicted log(intrinsic solubility) using the MLR model (eq 2) for (□) the training set, (×) the validation set, and (●) the test set.

Can deep learning do better? [6]

Table 5

Prediction performance and standard deviations using 10- fold cross validation on the Solubility Challenge Dataset (125 molecules)

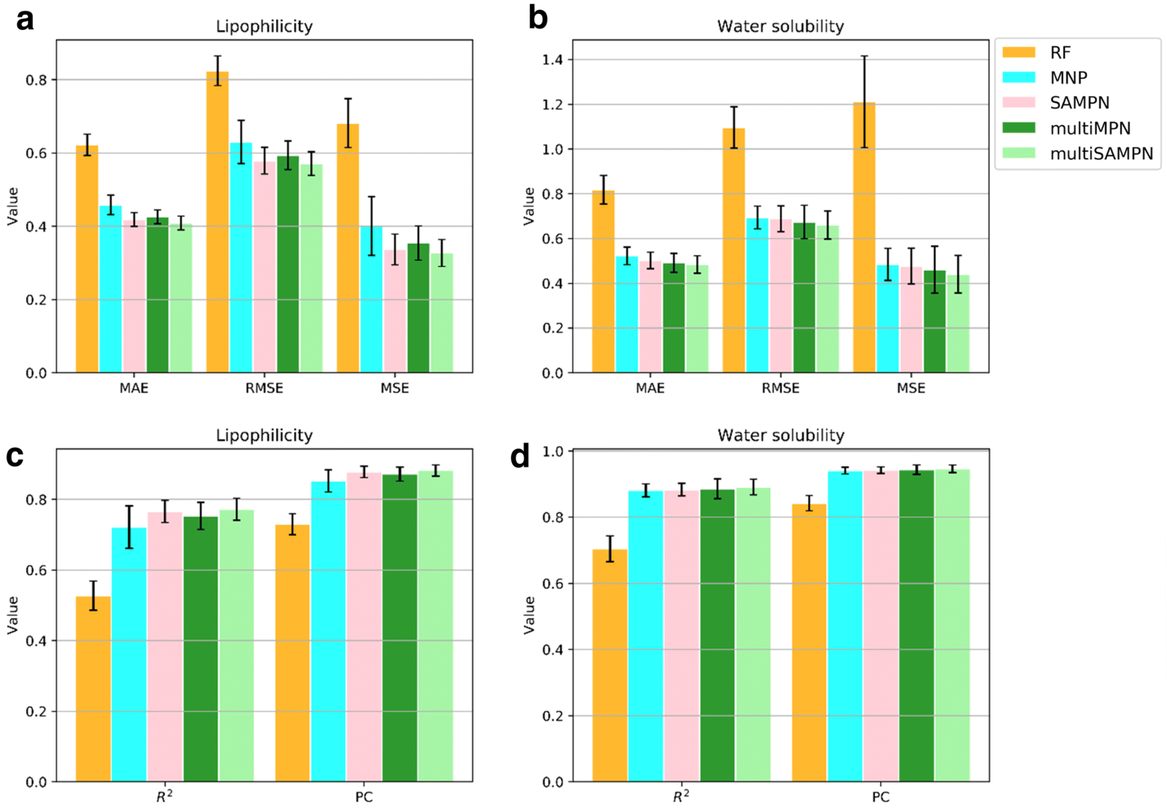
Models	R^2	std R^2	RMSE	std RMSE	AAE	std AAE
UG-RNN	0.32	0.03	1.41	0.12	1.08	0.10
UG-RNN-LogP	0.45	0.04	1.27	0.13	1.03	0.11
UG-RNN-CR-LogP	0.44	0.09	1.28	0.18	1.03	0.16
UG-RNN-Huusk	0.43	0.02	1.16	0.03	0.93	0.03
UG-RNN-Huusk-Sub	0.48	0.02	1.11	0.03	0.84	0.01
UG-RNN-LogP-Huusk	0.54	0.02	1.00	0.03	0.82	0.03
UG-RNN-LogP-Huusk-Sub	0.60	0.02	0.94	0.02	0.71	0.02
UG-RNN-CR-LogP-Huusk	0.62	0.03	0.96	0.06	0.83	0.06
UG-RNN-CR-LogP-Huusk-Sub	0.67	0.03	0.90	0.06	0.74	0.05
NN-Sol-Chal[11]	0.40	-	1.51	-	-	-
MLR-Sol-Chal[11]	0.51	-	0.95	-	0.77	-
New *in silico* consesus[11]	0.60	-	0.90	-	0.68	-

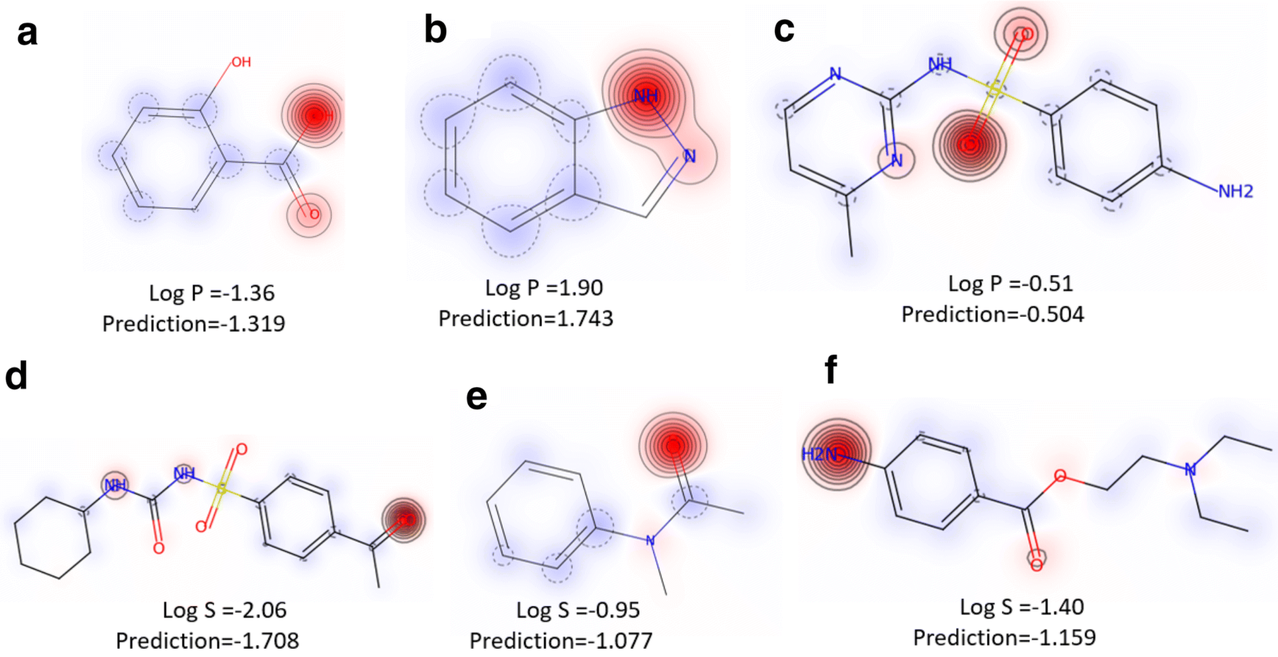
Better than simple neural network, but worse than MLR

Huusk results are based on more data.
Sub results use different solubility values.

In summary, deep learning can do better if there is more data.

Opening the black box?



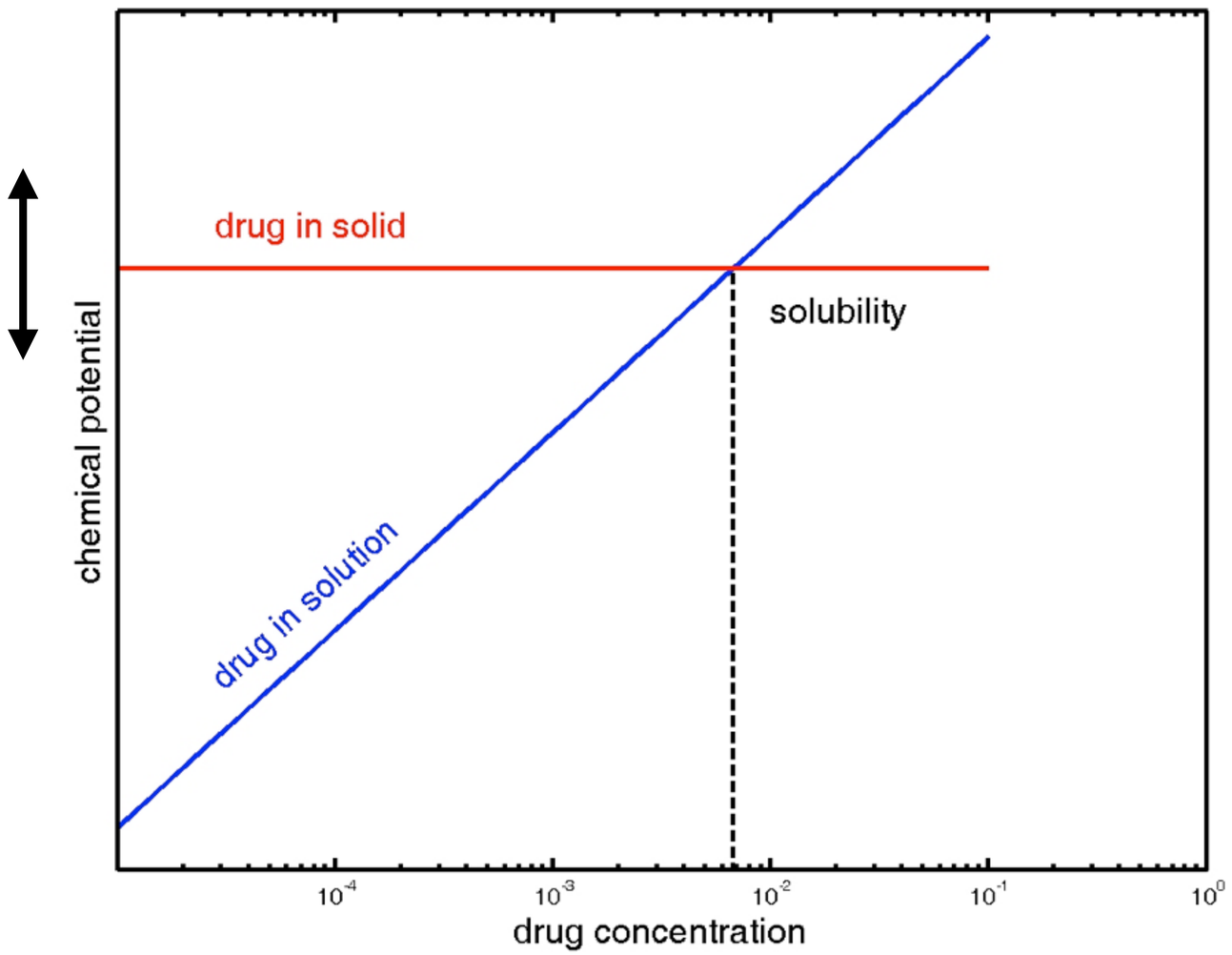


Recent deep learning method seems to perform well in solubility prediction and also highlights moieties that contribute to (red) or detract from (blue) solubility. See Ref. [7].

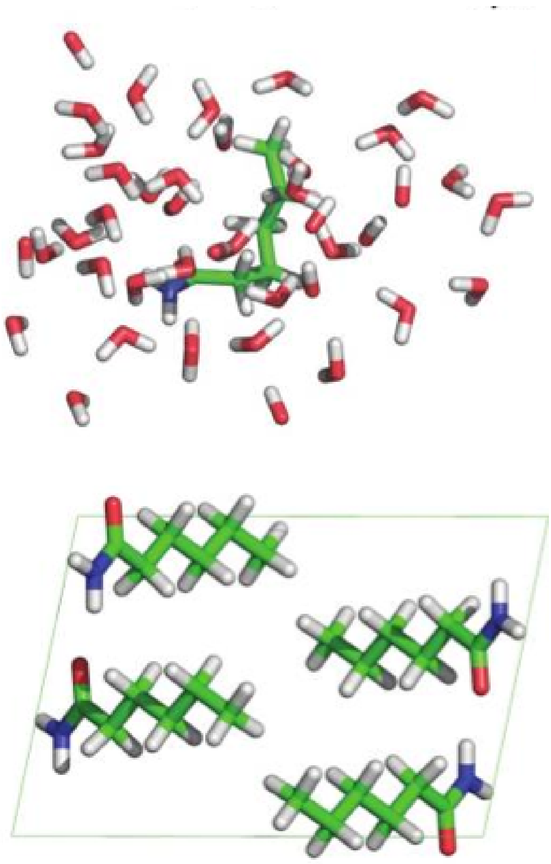
Free energy calculations of solubility

Chemical potential determines solubility

Polymorphs have different chemical potential

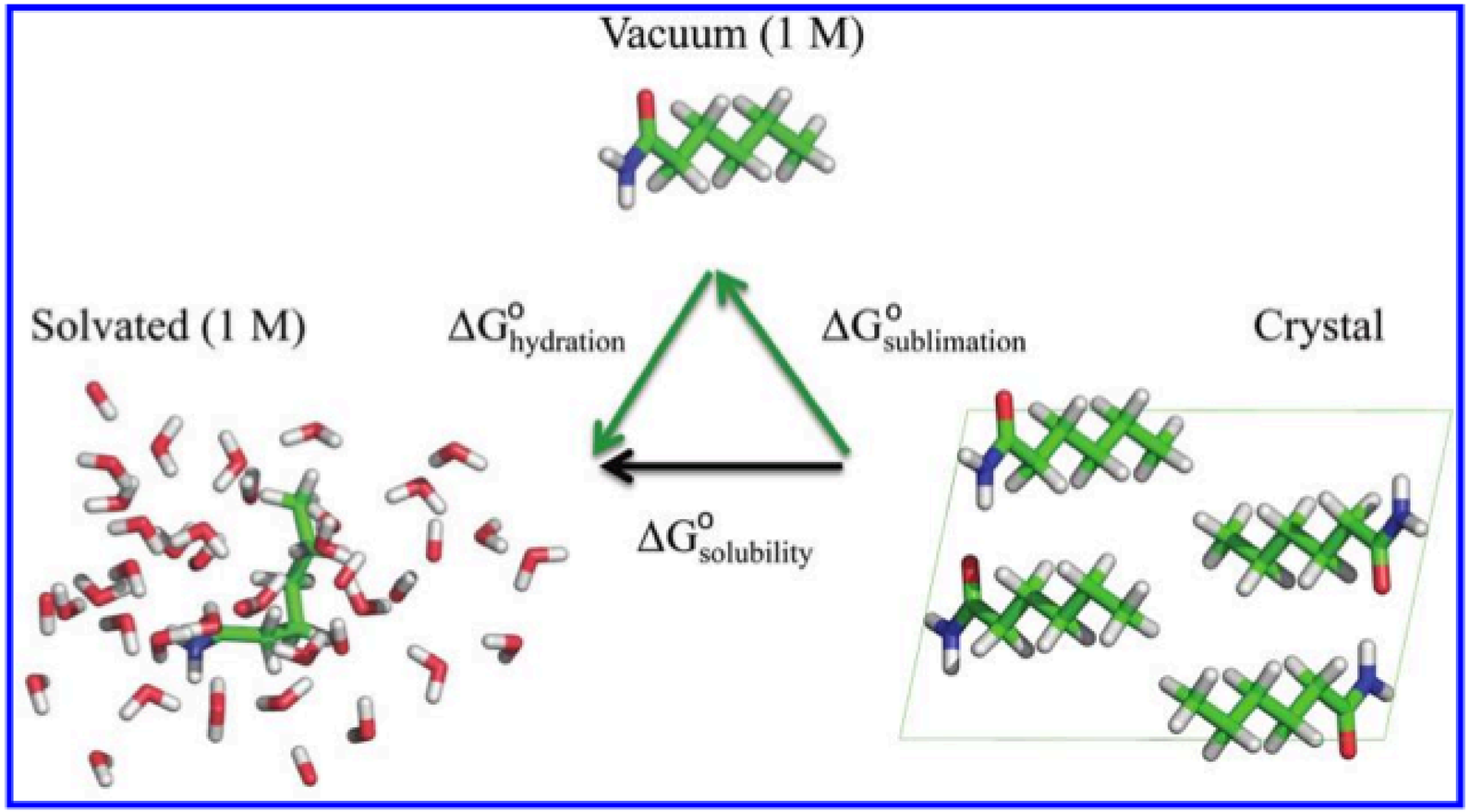


From Ref [1]



From Ref [8]

A thermodynamic cycle for solubility



From Ref [8]. Note that solubility ≠ solvation or hydration.
Hydration + sublimation free energies calculated with alchemical free energy methods.

Results on an initial series appear promising without any tuning

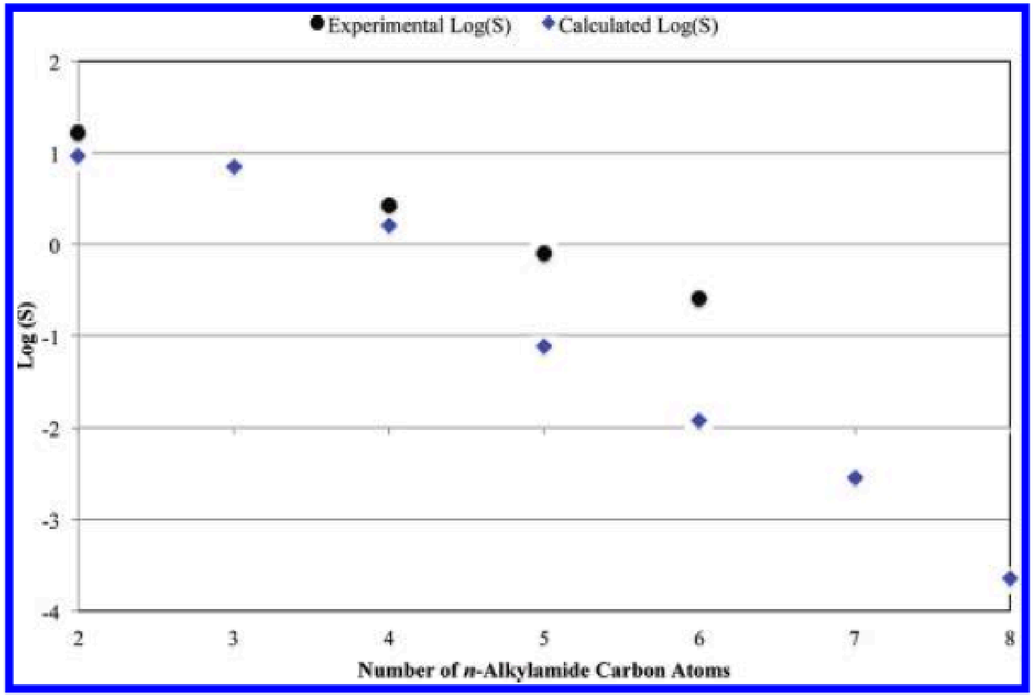


Figure 5. Shown are experimental and calculated log(*S*) values for the *n*-alkylamides (*S* has units of mol/L) from acetamide to octanamide. There is a monotonic trend in both the experimental and calculated values toward lower solubility with each additional CH_2 group due to increasingly favorable deposition and to a lesser extent from unfavorable solvation.

From Ref [8]

References

- [1] Many parts of today's lecture were adapted from a lecture by David Mobley (https://github.com/MobleyLab/drug-computing/tree/master/uci-pharmsci/lectures/free_energy_basics) under the CC BY 4.0 license. The lecture is part of the Drug Discovery Computing Techniques course (PharmSci 175/275) at UC Irvine.
- [2] Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? J. Chem. Inf. Model. 2008, 48 (7), 1289–1303. https://doi.org/10.1021/ci800058v.
- [3] Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. J. Chem. Inf. Model. 2009, 49 (1), 1–5. https://doi.org/10.1021/ci800436c.
- [4] Prashad, M.; Sutton, P.; Wu, R.; Hu, B.; Vivelo, J.; Carosi, J.; Kapa, P.; Liang, J. Process Research and Development of a MTP Inhibitor: Another Case of Disappearing Polymorphs upon Scale-Up. Org. Process Res. Dev. 2010, 14 (4), 878–882. https://doi.org/10.1021/op100115u.

References

- [5] Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. J. Chem. Inf. Model. 2009, 49 (11), 2572–2587. https://doi.org/10.1021/ci900286s.
- [6] Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. J. Chem. Inf. Model. 2013, 53 (7), 1563–1575. https://doi.org/10.1021/ci400187y.
- [7] Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A Self-Attention Based Message Passing Neural Network for Predicting Molecular Lipophilicity and Aqueous Solubility. J Cheminform 2020, 12 (1), 15. https://doi.org/10.1186/s13321-020-0414-z. Figures adapted under the CC BY 4.0 license.
- [8] Schnieders, M. J.; Baltrusaitis, J.; Shi, Y.; Chattree, G.; Zheng, L.; Yang, W.; Ren, P. The Structure, Thermodynamics, and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. J. Chem. Theory Comput. 2012, 8 (5), 1721–1736. https://doi.org/10.1021/ct300035u.