



Predicting Non-Aqueous Solubility with Molecular Descriptors and Message Passing Neural Networks

Jackson W. Burns ^{1,**} and Lucas Attia ^{1,**}

¹Massachusetts Institute of Technology, Cambridge, MA

^{**}Equal Contribution

March 21, 2024

Notes from 4/3 - Look for other temperature dependent models - Look for any models that have directly trained on Florence’s dataset - Consider other approaches to interact solute and solvent in a tabular/molecular descriptor feature rep.

Scientific Goal:

The solubilities of drug-like molecules in non-aqueous organic solvents are crucial properties for drug substance and drug product manufacturing. Experimentally measuring non-aqueous solid solubility requires notoriously tedious experiments which are both time-consuming and resource-intensive. Thus, predicting organic solubility of drug-like molecules *a-priori* based on their structure alone has been an active and robust area of academic and industrial research. The traditional approach relies on empirical solubility models like the Abraham Solvation model to estimate solubility. However, these empirical approaches are incapable of extrapolation by their nature, limited by the experimental data from which they are derived. We will extend recent work on applying molecular Machine Learning (ML) to this problem, which in theory could learn the underlying physics dictating the solubility and thus generalize.

Problem Formulation and Background

This a supervised learning problem with many open questions related to molecular representation and overall model structure. For the former, previous literature has focused primarily on applying learned representations via message-passing graph neural networks. We will include this as a reference point but will instead primarily focus on the application of descriptor-based models via **fastprop** ([GitHub.com/JacksonBurns/fastprop](https://github.com/JacksonBurns/fastprop)). This comparison between representation approaches should prove to be informative.

For the latter point, prior literature has also usually tried to enforce physics constraints in the model architecture. Vermeire et al. [1] compiled one of the largest publicly available datasets of thermodynamic quantities of drug-like molecules as well as a testing set of non-aqueous solubility for the same. Using the former they trained a combination of three Directed-Message Passing (Graph) Neural Networks (D-MPNN) models (via Chemprop [2, 3]) to predict different thermodynamic quantities, which in turn predicted solubility using a thermocycle. Another work by Yashaswi and coauthors [4] used an ‘interaction block’ - an intermediate layer in their network which performed a row-wise multiplication of the solute and solvent learned representations which was then passed to an FNN. This approach is analogous to training the model to map the structures to abraham-like solubility parameters, which are then weighted and combined for prediction. This question of appropriately enforcing physics is likely the most challenging aspect of this project (see Challenges).

Data

We will be using the aforementioned solubility dataset published by Vermeire et al. [1], which is made available via a machine-readable data format on Zenodo. This dataset contains 6261 solubility datapoints, with solute and solvent SMILES, solubility (logS), and temperature (K) as features. The original collators performed extensive data curation, so the reported solubility values are already well-sanitized and on a unified scale. We may apply standard scaling, log scaling, or power scaling to the values to simplify prediction though this will ultimately be decided based on the performance.

Challenges

We anticipate that the small amount of data and its highly imbalanced nature will require us to build physics in to our models. The reference study which aggregated this data enforced physics by never directly training on the solubility and instead creating models to predict other molecular properties used to calculate it. Our naive initial fastprop model will simply ingest the solute, solvent, and temperature as inputs to an FNN, effectively assuming that there is some ethereal non-linear mapping which can be learned between these and the solubility with no physics knowledge. The challenge is that there is likely some ‘intermediate’ between these two ideas which includes a sufficient amount of physics so as to assist the model in learning complex relationships but not so much that it becomes inflexible. By ‘interacting’ the solute and solvent representation (learned or descriptor-based) via element-wise multiplication, for example, we could force the model to learn a latent representation which is analogous to an abraham-like multiplicative solubility coefficient. Finding which ‘interaction’ between these representations is the most effective will require creativity and extensive experimentation.

Cited Works

1. Vermeire FH, Chung Y, Green WH (2022) Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. *Journal of the American Chemical Society* 144:10785–10797. <https://doi.org/10.1021/jacs.2c01768>
2. Yang K, Swanson K, Jin W, et al (2019) Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* 59:3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>
3. Heid E, Greenman KP, Chung Y, et al (2024) Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling* 64:9–17. <https://doi.org/10.1021/acs.jcim.3c01250>
4. Pathak Y, Mehta S, Priyakumar UD (2021) Learning atomic interactions through solvation free energy prediction using graph neural networks. *Journal of Chemical Information and Modeling* 61:689–698. <https://doi.org/10.1021/acs.jcim.0c01413>