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

Jackson W. Burns [1,**] and Lucas Attia [1,**]

[1] Massachusetts Institute of Technology, Cambridge, MA
[**] Equal Contribution

March 12, 2024

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. Recent work has instead turned to molecular Machine Learning (ML) which in theory could learn the underlying physics dictating the solubility and thus generalize.

Given the experimental challenges, datasets of organic solubility are highly dispersed in the literature. Previous efforts by 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. This coupling of an ML workflow to thermodynamics modeling achieved smooth gradients of solubility with respect to temperature.

While the reference model achieves a low error across most of chemical space, we have observed extremely poor performance and non-physical predictions in the limit of high solubility (>1 mol/L). We believe this is due to the imbalanced nature of the training data, which has relatively few examples of highly soluble compounds. By simplifying the learning task and instead directly predicting solubility from the input structures, we believe model performance can be improved.

We propose to train a Chemprop model which learns a representation for both the solute and solvent and then ingests the temperature by concatenating it to that representation. Additionally, we will compare an alternative model architecture `fastprop`(GitHub.com/JacksonBurns/fastprop), which has been shown to outperform Chemprop on other prediction tasks using only classical molecular descriptors for the solute and solvent. We will also quantify the performance of our models in the limit of high solubility to observe if direct prediction reduces the occurrence of non-physical predictions. This approach could provide improved solubility predictions with greater interpretability, which could be a helpful contribution to solubility prediction.

Data

We will be using the solubility dataset published by Vermeire et al. [1]. This dataset contains 6261 solubility datapoints, with solute and solvent SMILES, solubility (logS), and temperature (K) as features.

Supervisor

Given this is a molecular property prediction task on pharmaceutically-relevant small molecules, we would greatly appreciate Professor Coley's expertise on our project.

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