Springboard Data Science Capstone Two: Predicting the Solubility of Small Organic Molecules in Water from its Chemical Structure

Grace Nye

Problem Statement: How can we predict the solubility of a small organic molecule in water from only its chemical structure with an average absolute error (AAE) within range of existing techniques (0.75) in the next month by first predicting the melting point to inform our solubility calculation?

Context:

When designing small molecule drugs, it is essential to understand their chemical properties, including their solubility in aqueous solution. This property is also important for chemists working in many other industries, including the agrochemical industry. Therefore, it is worthwhile to have a fast and simple method for estimating the aqueous solubility of a given small molecule. The General Solubility Equation (GSE) is able to provide an accurate description of solubility given the melting point (MP) of the molecule. If that data is not readily available, however, it is much harder to predict solubility. Existing datasets for training ML models for predicting solubility directly from the chemical structure are small and highly variable. Here we propose a ML model for estimating aqueous solubility of a small organic molecule (MW < 200) by first generating a model for predicting the melting point from the CCDC melting point dataset, containing ~58,000 molecules, and then using the General Solubility Equation (GSE) or ESOL to calculate solubility.

Criteria for Success:

Success for this project will be defined by developing a model for predicting the molar solubility of a small organic molecule in water in the next month with an average absolute error (AAE) of 0.75.

Scope of the Solution Space:

Solution space will consider an existing dataset available from the Cambridge Crystallographic Data Centre (CCDC) containing molecular descriptors and experimentally determined melting points for ~58,000 molecules. Using several machine learning methods, such as a Random Forest Regressor and Linear Regression, a model will be trained to predict the melting point of a molecule from its chemical structure and molecular descriptors that can easily be calculated from the structure only. Then, the solubility of that molecule will be calculated using the GSE.

Constraints within solution space:

Several factors that may constrain the success of this project may include the small size of the dataset, the uncertainty associated with the experimental methods used to determine melting point, as well as the accuracy of the GSE.

Stakeholders to provide key insight:

Key stakeholders include my Springboard mentor, Blake Arensdorf, as well as other members of the Springboard community.

Information and datasets are sourced from Cambridge Crystallographic Data Centre (CCDC).

Key data sources:

- Cambridge Crystallogrpahic Data Centre (CCDC) - https://www.ccdc.cam.ac.uk/support-and-resources/downloads/ (under validation, Sutro melting point training data).
- Zhu, X. *et al.* Building Machine Learning Small Molecule Melting Points and Solubility Models Using CCDC Melting Points Dataset. *J. Chem. Inf. Model.* **2023**, 63, 10, 2948–2959. https://doi.org/10.1021/acs.jcim.3c00308
- Jain, N., Yalkowsky, S.H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *Journal of Pharmaceutical Sciences.* **2001,** 90, 2, 234-252. https://doi.org/10.1002/1520-6017(200102)90:2<234::AID-JPS14>3.0.CO;2-V
- Delaney, J.S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 3, 1000–1005. https://doi.org/10.1021/ci034243x