MSE 465/1065, Fall 2024

Project title: Predicting Solubility of Small Molecules in Solvent Mixtures

**Project motivation**

Small molecules (such as active pharmaceutical ingredients) are the cornerstone of most materials systems. Handling these molecules in industrial and laboratory processes often requires their solubilization in a carrier solvent. Determining the appropriate combination of solvents in the absence of one good solvent is both challenging and sometimes necessary (to support objectives such as green synthesis). Further, solubility experiments are time consuming and laborious, so we would like to minimize the number of failed experiments.

**Project objective**

Students will work with a real dataset of solubility data of small drug-like molecules in binary solvent systems at different temperatures. The goal is to use machine learning models to predict the solubility of a given small molecule in a new combination of binary solvents or at a previously un-tested temperature.

**Project outline**

1. **Data preparation:** Start with understanding the data in the provided CSV file and cleaning the data to ensure consistency
2. **Data exploration:** Use techniques like clustering, chemoinformatics tools (RDKit) or chemical knowledge to understand and describe the datset’s diversity. Represent this data in simpler forms such as reduced dimensions or histograms
3. **Feature engineering:** Use tools like RDKit to add more features to the dataset, making it richer for analysis
4. **Model training:** Build and train models using methods like Random Forest, Gradient Boosting Machines. Use cross-validation to check the model’s performance.
5. **Model evaluation:** Choose the right metrics to see how well the model works on new data. Conduct feature importance analysis to determine which elements of the model are most important and how certain the model is about its predictions.
6. **Advanced task for graduate students:**
   1. **Feature generation for arbitrary molecules**: expand the model’s applicability to *in silico* molecules by incorporating additional property predictions to generate input features for arbitrary molecules.