Chem361 Spring2025
Sample Homework

Instructions: Use this PDF file as a template to develop your homework. Submit your homework on time as a ZIP file to **Canvas** containing:

(1) A single PDF or Word document with your answers to the questions.

(2) Your code within a Jupyter Notebook.

Homework can be completed in any format, including handwritten answers. There will be 4 sets of homework assignments, and 3 out of 4 with highest scores will each contribute 20% to the final grade. Please submit homework electronically via the Canvas system before the due date. ***Late homework submissions will not be accepted.***

1 Processing Chemical Compound Dataset Using Python (50 points)

Fetch the Delaney dataset (`https://raw.githubusercontent.com/deepchem/deepchem/master/datasets/delaney-processed.csv`), which contains solubility data for a set of chemical compounds, and complete the following five tasks. This problem requires the use of Python, along with the Pandas and RDKit packages.

Hint: i) You may refer to the provided `homework0_reference.ipynb` file or other internet resources to help you solve these questions; ii) You may use Google Colab (`https://colab.research.google.com/`) if you don't have a local Python environment. iii) If you're unsure how to use a specific functionality of a Python package, refer to its official documentation website for detailed explanations and examples.

1.1 Please calculate the total number of chemical compounds present in the dataset. **(10 points)**

1.2 Please provide the minimum and maximum molecular weights of chemical compounds in the dataset. **(10 points)**

1.3 Please calculate the number of chemical compounds with at least two rings in the dataset. **(10 points)**

1.4 Please select any chemical compound with at least two rings and provide its Compound ID along with its SMILES string. **(10 points)**

1.5 Please visualize your selected chemical compound using RDkit package. **(10 points)**

2 Predicting Compound Solubility Using Linear Regression (50 points)

Please use your fetched Delaney dataset to complete the following four tasks. This problem requires using Python and the scikit-learn package.

Hint: i) You may refer to the provided `homework0_reference.ipynb` file or other internet resources to help you solve these questions; ii) You may use Google Colab (`https://colab.research.google.com/`) if you don't have a local Python environment. iii) If you're unsure how to use a specific functionality of a Python package, refer to its official documentation website for detailed explanations and examples.

2.1 Please perform a 70:30 train-test split on the dataset. Fit a linear regression model using "Molecular Weight" as the predictor to estimate the compound's log solubility on the training set. Provide the slope and intercept of the fitted linear regression model. **(10 points)**

2.2 Please report the mean squared error (MSE) for both the training and test sets. **(10 points)**

2.3 Please visualize the distribution of chemical compounds in the molecular weight vs. log solubility space for both the training and test sets, and overlay the fitted linear regression line on the plots. **(10 points)**

2.4 Please fit another linear regression model using the property "Polar Surface Area" as the predictor. Report the mean squared error (MSE) for both the training and test sets. Visualize the distribution of chemical compounds in the polar surface area vs. log solubility space for both the training and test sets, and overlay the fitted linear regression line on the plots. Please briefly explain whether molecular weight or polar surface area is a better feature for prediction of log solubility. **(20 points)**