# CHEM361 SPRING2025
# HOMEWORK 1

**Due date: 5pm, Feb 13, 2025**

**Instructions:** Use this PDF file as a template to complete your homework. Submit your work on time as a ZIP file to **Canvas**, including:

(1) A single PDF or Word document containing your answers (handwritten or typed).

(2) Your code in a Jupyter Notebook.

You may refer to the provided `homework1_reference.ipynb` file or online resources to help you solve the questions; If needed, you can use Google Colab (https://colab.research.google.com/) instead of a local Python environment. For Python package functionality, consult the official documentation for detailed explanations and examples.

There are 4 homework assignments; the 3 highest scores will each contribute 20% to your final grade ***Late homework submissions will not be accepted.***

## 1 Calculate Molecular Representations (50 points)

**1.1** Draw the chemical structure of Acetaminophen using RDKit. (Hint: Download the SMILES of Acetaminophen from PubChem: https://pubchem.ncbi.nlm.nih.gov/) **(10 points)**

**1.2** Please find the quantitative estimation of drug-likeness (QED) value for Acetaminophen calculated using RDKit. **(5 points)**

**1.3** Compute the Morgan fingerprint of Acetaminophen with a radius of 2 and a length of 256, print the fingerprint, and count the number of 1-bits. **(10 points)**

**1.4** Please calculate the number of bit collisions in the 1024-bit Morgan fingerprint of Acetaminophen (radius = 2). (Hint: You can find the value by comparing to its 2048-bit fingerprint, 4096-bit fingerprint, etc.) **(15 points)**

**1.5** Please visualize the subgraphs in the 2048-bit Morgan Fingerprint of Acetaminophen. **(10 points)**

## 2 Cluster Molecules and Dimensionality Reduction (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 tasks. Please use the feature vectors containing properties of "Molecular Weight", "Number of H-Bond Donors", "Number of Rings" and "Number of Rotatable Bonds" and Euclidean distance to perform clustering and dimensionality reduction (Hint: Data normalization is critical).

**2.1** Cluster the chemical compounds in the Delaney dataset using the K-means algorithm with $K = 20$. Repeat the clustering five times and report the size of the largest cluster for each run. If the results are inconsistent, please explain why. **(10 points)**

**2.2** Perform K-means clustering using

$$K = 10, 20, 30, 40, 50$$

For each K, calculate the loss function (sum of squared distances or SSD). Visualize the loss function (SSD) as a function of $K$. Based on the plot, determine the optimal number of clusters and justify your choice. **(10 points)**

**2.3** Please cluster molecules in the Delaney dataset using DBSCAN with the parameter of $eps = 0.5$ and $min\_samples = 5$. Please report the number of noise points (outliers) after clustering. **(10 points)**

**2.4** Perform PCA on the Delaney dataset to project the chemical compounds onto the first two principal components. Report PC1 and PC2. **(10 points)**

**2.5** Building on Q2.4, plot the compounds from the Delaney dataset on a 2D plot using the reduced dimensions PC1 and PC2. Additionally, include the compound 'Acetaminophen,' which is not part of the Delaney dataset, on the same plot. Use a distinct color to differentiate the Acetaminophen data point from the others. **(10 points)**