SURP Nomination Assignment

Predicting Molecular Properties

Your task is essentially to use the dataset given below, containing SMILES representation of certain molecules and their respective log solubility (target variable). Design an optimised model to predict accurately the target value for a given molecule in its SMILES format.

Key Tasks:

- Data Preprocessing (converting SMILES to Morgan Fingerprinting)
- Exhaustive EDA to understand the fingerprinting data
- Model architecture definition for a simple Baseline Model (Random Forest, XGBoost, etc) (Train-Test Split : 75-25)
- **BONUS:** Design a simple GNN model using DeepChem, compare the performance of GNN vs Baseline Model
- One pager on the workflow, model choice and performance analysis

Deliverable

Jupyter Notebook with markup cells for any sub-block created and 1-page report (PDF format)

Hint List (helpful libraries)

- Rdkit
- Pandas
- Scikit-learn
- Matplotlib
- xgboost/lightgbm
- Torch, torch_geometric (optional)

Dataset

Chem Dataset (Target: measured log solubility in mols per litre)

Submission Link:

https://forms.gle/yRYSunoiHWvJxUSd6