Assignment 2

Problem 1: Predicting Molecular Toxicity from SMILES (50 points)

Toxicity prediction is an essential task in drug discovery, as **toxic molecules** can cause harmful effects in biological systems. The **Tox21 dataset** provides molecular toxicity data across **12 biological targets**, with molecules represented as **SMILES strings**.

Please refer to the course slides and DeepChem tutorials, solve the tasks:

1. Load the Tox21 dataset from **MoleculeNet** and extract SMILES strings.
2. Convert SMILES strings into ECFP (Extended Connectivity Fingerprints) using RDKit. (Please do not load the dataset by using `featurizer='ECFP'`, convert SMILES into ECFP)
3. Train a binary classification model (e.g., Random Forest or Logistic Regression) to predict molecular toxicity.
4. Train a custom built Graph Convolution model to predict molecular toxicity.
5. Compare the result using metrics such as ROC-AUC (Area Under Curve), accuracy, precision recall scores and training time. What's your opinion about the Graph Convolution model?

Problem 2: Proteins Secondary Structure Prediction (50 points)

Protein secondary structures (alpha-helices, beta-sheets, etc.) play a key role in biological functions. The **CB513 dataset** contains **amino acid sequences** with their corresponding secondary structures.

Please refer to DeepChem tutorials and solve the tasks:

1. Load the CB513 dataset from DeepChem, and display 10 sequences.
2. Convert protein sequences into one-hot encodings or physicochemical feature vectors. (Similar to Problem 1, please do not load by using `featurizer='one-hot'`)
3. Train a deep learning model (e.g., LSTM, CNN) to predict secondary structure labels.
4. Evaluate performance using Accuracy and F1-score.
5. Perform a small ablation study—remove one feature type (e.g., one-hot encoding or physicochemical features) and observe how it affects model accuracy.