Executive Summary (Team: Drug Discovery):

This project considers two datasets: one for classification task and one for regression task. Both the dataset are parts of TDC dataset and publicly available.

Classification Task: We consider `CYP P450 2C19 Inhibition` dataset. The name `CYP P450 2C19` is abbreviation of Cytochrome P450 2C19 which is an enzyme. Enzymes are proteins. In humans, it is the CYP2C19 gene that **encodes** the CYP2C19 protein. Protein-encoding genes specify the sequences of amino acids, which are the building blocks of proteins. In turn, proteins are responsible for orchestrating nearly every function of the cell. Both protein-encoding genes and the enzymes (proteins) that are their gene products are absolutely essential to life as we know it.

The CYP2C19 is a liver enzyme and is essential in the breakdown (metabolism) of various molecules and chemicals within cells, which is necessary for life. It acts on at least 10% of drugs in current clinical use, most notably the antiplatelet treatment clopidogrel (Plavix), drugs that treat pain associated with ulcers such as omeprazole, antiseizure drugs such as mephenytoin, the antimalarial proguanil, and the anxiolytic diazepam.

An enzyme inhibitor is a molecule that binds to an enzyme and blocks its activity. If a drug acts as an enzyme inhibitor and stops (inhibits) an enzyme, this would mean poor metabolism to this drug and other drugs, which could lead to drug-drug interactions and adverse effects.

The `CYP P450 2C19 Inhibition` dataset contains chemical nomenclatures of 12,665 drugs and whethere these drugs inhibit the activities of CYP2C19. Whether it inhibits or not is captured as binary (0/1) target variable. The dataset was more or less a balanced dataset. From the chemical nomenclatures of drugs, we extracted an extensive set of descriptors of these drugs by using **RDKit** toolkit. These descriptors are then used as **features** to build **XGBoost** based predictive model. We carried our detailed EDA and Spectral Embedding scheme to explore and dataset before building our ML model. The GridSearchCV scheme was employed to fine tune the **XGBoost** model, which yielded an accuracy of 0.81 for our **test data** set. We also presented the **top 20 most important features** based on our XGBoost model performance.

Regression Task: Here, we attempted to build a pipeline for ADME (Absorption, Distribution, Metabolism, Excretion) property prediction of drug molecules. The ADME properties of a drug molecule is essential to determine the viability of a drug during lead optimization in the pharmaceutical industry. Build models that can generalize the prediction on an unseen drug can save valuable time during clinical trials.

We used the `Caco2-wang dataset` to predict the Caco2 permeability. The features were generated using the **RDkit** and **DeepChem** packages. We particularly used two main classes of features- molecular descriptors (MD) as used in the classification project and molecular fingerprints (FP). The target property was **log(apparent permeability)**, `logPapp`, which has a range of values between -3 to -8.

Several models (namely, Random Forest, Linear regression and XGBoost) was considered with only MD, only FP and both MD + FP as feature vectors. **XGBoost** regressor turned out to be the best model for all three types of featurizations, resulting in an MAE of 0.55, 0.49 and 0.33 for MD, FP, and MD+FP respectively. Feature engineering approaches like similarity profiles using Tanimoto coefficients can likely improve the model performance.