Capstone Project Proposal

Problem statement formation
How CYP2C19 Inhibition can be predicted from a SMILE string with greater than 0.713 AUPRC before July 31st?

Context
"CYP P450 genes are essential in the breakdown (metabolism) of various molecules and chemicals within cells. A drug that can inhibit these enzymes would mean poor metabolism to this drug and other drugs, which could lead to drug-drug interactions and adverse effects" (ADME - TDC).

A SMILE is a variable length string that represents the chemical structure of a molecule.
Example smiles: (*Clc1ccccc1-c1nc(-c2ccccc2)n[nH]1,*
COc1ccccc1C(c1nnnn1C(C)(C)C)N1CCN(Cc2ccncc2)CC1)

The sample space of molecules is 10^60 and a major bottleneck in medical research is in deciding what drugs to test. This model will make the decision of what drugs to test in a lab more informed.

Criteria for Success
A model that can take as input an SMILE drug string and returns a boolean for if that drug will inhibit CYP2C19 with an average AUPRC score of greater than .714.. A well commented and reproducible github repository where users can **easily** duplicate and extend my work.

Scope of Solution Space
Unsupervised Feature Extraction Techniques.
Adapting the CASTER Drug-Drug interaction model
Adapting the DeepPurpose single feature prediction model.
Deep Learning models, this problem is not well suited to traditional ML approaches.

Constraints
Unbalanced training data, 2/3 of the labeled data is negative. The model training needs to take place on the cloud.

Stakeholders
Kexin Huang: My contact at Therapeutic Data Commons (TDC)
Anna Klimovskaia: A Sweedish Data Scientist advising me on the technical details.

Data sources
~12.6k labeled SMILEs from TDC (provided via API).
~ 200k unlabeled SMILEs from DrugBank.
~ 50k unlabeled SMILEs from BIOSNAP.

Approaches:

Extract N common patterns in the massive DrugBank and BIOSNAP datasets. One-hot encode the presence of those patterns into a boolean vector and feed that into some kind of Neural Network.

The CASTER model predicts if interaction will occur between two drugs. I can take the molecule for CYP2C19 and each molecule in the labeled dataset and use that data to train a variation on CASTER

Kexin recommended I check out the DeepPurpose library for single feature prediction of molecules.

There are several different “message passing” paradigms where information about the “local neighborhood” of a node in a graph is encoded into a vector representing the relationship between it and it’s neighbors. I don't know much about this but it seems promising.

RDKit lets you extract molecule level features from SMILE strings.

Deliverables:
A model object with a high AUPRC score.
A reproducible github repo of my code.
A report of the methodology and outcomes of the models.
A series of Notebooks that explore different techniques.
A slide deck of the model architecture, outcomes and accuracy.
My model and code on the leaderboard at TDC.