Final Project Proposal

Bookmarks	
Class Link	
Date	
LOs and HCs	
Overview	
Reviewed Bookmarks	□
Unit	

Dataset

https://www.kaggle.com/xiaotawkaggle/inhibitors

Description

This is a dataset of molecules that can inhibit the action of cancer-related proteins. The task is to predict if the test set of molecules would inhibit a certain protein.

What will I do?

This will be a classification task where I identify whether a new molecule "looks similar" to the molecule-inhibitors of a certain protein. One way to look at this would be to model the attributes (features) as items and all the molecules-inhibitors as counts. For each attribute, we will have a count of how many times it appeared in the inhibitor dataset per protein. Thus, likelihood would be a multinomial, and a prior would come from a Dirichlet-categorical distribution, for which I am going to have to find the posterior parameters for.

In this case, a sample from Dirichlet distribution would indicate which attributes are more likely to be present in an inhibitory molecule. Then, for classification, I can generate many samples and check the p-value of a new molecule and see if it is from the same process.

Concerns

I still don't really understand how exactly this RDkit attribute encoding of molecule works — I can add some structural features to the model that would consider physical/chemical properties. I will ask Prof. Kern for some advice on this!

Adding to the first one, there are definitely interaction terms in here (maybe when attribute 1 and 4 are active, a molecule is an inhibitor, but if at least one of them is inactive, then it is not). Therefore, I will need to come up with something more complicated than just individual counts. A dumb idea is to consider all subset combinations, but since we have about 8k features, this seems implausible.