File 1: Predictions.csv

This file is fairly self explanatory. The first column is the SMILE code, the second column is the predicted probability that it is a "good" inhibitor, and the third column is a ranking of 1-70,000 for each compound.

Important Note: the files below cover information constructed from the holdout data which was constructed from the original data from the Chemists, and ran through one of the twenty ensemble models. It is merely meant to give a good representation of the data.

File 2: LiftTable.xlsx

There is a lot of information here, I will try to break it down as best I can. Column 1 corresponds to different cutoff values or thresholds in which a compound will be classified as "good", ranging from 1 down to zero. This table tries to answer the question: if I set a cutoff value at x%, how many compounds will I have to test and then about how many can I expect to be positive hits from that test? If you looked at all compounds that have a probability at that threshold or higher, you would have a certain number of "good" compounds in that sample (corresponding to column 2). Column 3 corresponds to how many total observations would be in that sample. The final column I want to highlight is called the Lift Ratio. In its simplest terms, it is a ratio of the proportion of good compounds in your sample to the overall proportion of good compounds in the entire data set.

Example: look at the row that has a cutoff value of 0.690234065. If we take all compounds that have a predicted probability of being “good” that is that value or higher, we will have a total of 3 compounds to test, of which 2 ended up being good. The original data had about 2.5% of compounds being good. In our new sample of 3 compounds, 66% of the compounds were good, giving us the lift ratio of 24.56652361. In other words, with a cutoff value of 0.690234065, the proportion of good compounds we will identify is approximately 24x larger than the overall proportion of good compounds in the original data.

File 3: LifePlot.png

This graphic shows visually what is shown in file 2. With different cutoff values on the x axis, and corresponding lift ratios on the y axis, you are able to see visually how changing the cutoff value impacts this ratio.

File 4: GainPlot.png

On the x axis, it plots the percent of compounds that are “tested”. On the y axis, it plots the percent of total “good” compounds found at that x value. For example: if you tested the compounds that fell in the top 20% in terms of highest predicted probability of being good, you would correctly identify about 65% of the total number “good” compounds.

**What this all means for testing:**

In our meeting on Friday, we discussed the idea of looking at the top 100 compounds for testing. Based on the information in file 2, we could expect approximately 40 of these compounds to be in fact “good”.

Hypothetical scenario: If we were to take the top 1000 compounds from the NIH database, this would correspond to a cutoff value of about .576 or higher. From the LiftTable, a sample taken with this cutoff value will be approximately 20% good compounds. So taking the top 1000 compounds in the prediction file, we can expect approximately 200 will be positive hits.