Overview:

We have Chemical ID and Expectedcolumns in train data. Expected shows whether a chemical is toxic or not, we have to predict this for chemicals in test data and generate a submission file which will be updated on kaggle(This is a private competition) to test the accuracy.

Code overview:
We will generate features/descriptors for these chemicals and apply ML algorithms to predict the toxicity for test data chemicals

Train and test datasets:
They are attached in the mails train and test data sets- train-II, test-II
I also provided sample code for features generation and using a model-> I used XGboost

Target:

I want you to perform better feature selection and improve my accuracy which is currently 82% on kaggle