1. Scrape raw data from the Tox21 Public database
   Google drive link to raw data
2. Generate outcome matrix and feature matrix using RDKit
   File outcome matrix feature matrix
   -> 208 features
3. Feature selection using R file preprocessing -> **114 features**
   a. define functions to remove features and run models
   b. run defined functions on tox21
4. **Perform lasso regression to select top 40 features**
   File lasso regression
   Output of 40 features, bal_acc for 5, 10, 15, 20, 25, 30, 40, 50, 60, 114
5. Generate DS1 -> assay by assay with 40 features
6. Generate DS2 -> stacked 50 assays with 40 features
7. Generate DS3 -> stacked 50 assays with 40 features and gender and organism
8. Run ridge, naïve Bayesian and HBM for three datasets
   a. Ridge
      i. DS1
      ii. DS2
      iii. DS3
   b. Naive Bayes
      i. DS1
      ii. DS2
      iii. DS3
   c. HBM
      i. DS1
         1. training
         2. testing + validation
      ii. DS2
      iii. DS3
9. Use histogram to look at feature importance
   a. Ridge DS1
   b. Ridge DS2
   c. Ridge DS3
   d. Naive DS1
   e. Naive DS2
   f. Naive DS3
   g. HBM DS1
   h. HBM DS2
   i. HBM DS3