1. Scrape raw data from the Tox21 Public database

[Google drive link to raw data](https://drive.google.com/drive/u/0/folders/1JvxDnfZMClksJhCvKHzeMAHZ88KJQ19C)

1. Generate outcome matrix and feature matrix using RDKit

File [outcome matrix](https://drive.google.com/file/d/1gk0XK_ws1cVARk674RbKwuLFsgNeZghO/view?usp=drive_link) [feature matrix](https://colab.research.google.com/drive/1nAGFFK4-0QJApG_DNfAHxQVTtg7dDehx)

**-> 208 features**

1. Feature selection using R file preprocessing -> **114 features**
   1. [define functions to remove features and run models](https://drive.google.com/file/d/1h_Wy-MS0KMGGOMnE5hbAs9DAzYqCfJkP/view?usp=drive_link)
   2. [run defined functions on tox21](https://drive.google.com/file/d/1ipFi5ZBwjBBSJEhLw6TVRgJh69BZ4owF/view?usp=drive_link)
2. **Perform lasso regression to select top 40 features**

File [lasso regression](https://colab.research.google.com/drive/1ySBEn7Hb5YVyLRk70wqjzAJV84Yq0ZSS#scrollTo=nvPs6mDq5iro)

Output of 40 features, bal\_acc for 5, 10, 15, 20, 25, 30, 40, 50, 60, 114

1. Generate DS1 -> assay by assay with 40 features
2. Generate DS2 -> stacked 50 assays with 40 features
3. Generate DS3 -> stacked 50 assays with 40 features and gender and organism
4. Run ridge, naïve Bayesian and HBM for three datasets
   1. Ridge
      1. [DS1](https://colab.research.google.com/drive/1LrUTaOegKs2r2Ize9c8iWsssE_YPzAfm#scrollTo=lmfc0KJ-cwNK)
      2. [DS2](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
      3. [DS3](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   2. Naive Bayes
      1. [DS1](https://colab.research.google.com/drive/1_K0ExiRQO7Ds8iwvnmMwZpvQ2-xC621s#scrollTo=whjZgl-RFlqj)
      2. [DS2](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
      3. [DS3](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   3. HBM
      1. DS1
         1. [training](http://localhost:8888/notebooks/Downloads/hierarchical_model/Pymc_train_no_meta_assay_by_assay.ipynb)
         2. [testing + validation](http://localhost:8888/notebooks/Downloads/hierarchical_model/Tox21_pred_no_meta_assay_by_assay.ipynb)
      2. [DS2](http://localhost:8888/notebooks/Downloads/hierarchical_model/Pymc_train_5_fold.ipynb)
      3. [DS3](http://localhost:8888/notebooks/Downloads/hierarchical_model/Pymc_train_5_fold.ipynb)
5. Use histogram to look at feature importance
   1. [Ridge DS1](https://colab.research.google.com/drive/1WJ-YOEhAEjNO1mx-oj4nU-I6mu_fNye6#scrollTo=N0tQ-kROfsGq)
   2. [Ridge DS2](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   3. [Ridge DS3](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   4. [Naive DS1](https://colab.research.google.com/drive/1WJ-YOEhAEjNO1mx-oj4nU-I6mu_fNye6#scrollTo=N0tQ-kROfsGq)
   5. [Naive DS2](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   6. [Naive DS3](https://colab.research.google.com/drive/1cB34_pe_z1_tU65yXeaRbSxOP995KEan#scrollTo=xxaccSqpGce7)
   7. [HBM DS1](http://localhost:8888/notebooks/Downloads/hierarchical_model/Feature_imp_HBM_DS1.ipynb)
   8. [HBM DS2](http://localhost:8888/notebooks/Downloads/hierarchical_model/Feature_imp_HBM_DS2.ipynb)
   9. [HBM DS3](http://localhost:8888/notebooks/Downloads/hierarchical_model/Feature_imp_HBM_DS3.ipynb)