**Problem Definition:** Drugs are made of bioactive molecules. Thus, it would be very useful in the process of drug discovery if there were some means to predict whether a molecule is bioactive or not, using various molecular descriptors as the input.

**Methods:**

Data Pre-processing:

* We only consider molecules which have their standard bioactivity values in ‘IC50’ units.
* Class labels were given to molecules according to their standard values as ‘active’(<1,000), ‘intermediate’(1,000-10,000) and ‘inactive’(>10,000)
* The Cannonical smiles(representation of the molecular structure)had to be cleaned before it would be fed to the [RDkit](https://www.rdkit.org/docs/) class.

Finding Descriptors:

* We use [Lipinski Descriptors](https://en.wikipedia.org/wiki/Lipinski%27s_rule_of_five)(Molecular Weight, octanol-water partition coefficient(logP), #of Hydrogen bond donors, # of hydrogen bond acceptors) of a molecule as our features for further analysis

.

Analysis of Descriptors:

* The analysis of the descriptors was carried out using box plots and a statistical test called the [Mann-Whitney U test](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test).
* The box plots for each descriptor are inside the plots folder.
* If U score of the descriptors with the bioactivity class is less than 0.05, we conclude that the descriptor and bioactivity come from different distributions.
* The results of these tests can be found in the ‘U score test’ folder.

Model Fitting

* We used this article as a guide for selecting models for Regression
* We try using [Gaussian Process Regressor](https://scikit-learn.org/stable/modules/gaussian_process.html), [NuSVR](https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVR.html) and [ARDRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ARDRegression.html) for this dataset.
* NuSVR gave the best results with r2 score of 0.234.

**Conclusions:**

* ‘LogP’ fails to disprove the null hypothesis that the different bioactivity classes come from a different distribution. Thus, it is not a useful feature to predict bioactivity in this particular case.
* NuSVR gave the best regression score of 0.234 amongst the models tested.