**Capstone Project Milestone Report**

**Introduction**

In recent years, high-throughput screening has been used in the exploratory phase of the drug-discovery process. In this stage, thousands of small molecule compounds are tested for inhibitory capacity against a biological target in order to find a lead target molecule. Any molecule found to be inhibiting a biological function is further optimized to be more potent. Also, at this period, a structure-activity relationship diagram can be developed in which molecules having the same scaffold are analyzed for degrees of affect based on physical characteristics. The target audience of the data analysis would be the biochemists and organic chemists that work together to synthesize and optimize a new drug-like small molecule.

**Data Source**

One of the projects I worked on in my dissertationwas to find inhibitors targeting the non-structural 3 (NS3) helicase protein of the Dengue virus genome. Dengue virus (DENV) is one of the most significant human viral pathogens transmitted by mosquitoes. It causes 50 million or more cases of infection worldwide each year, resulting in around 24,000 deaths. Control of dengue virus through the use of vaccination or a direct-acting antiviral has proven to be elusive.

The data I’ll be using came from the high-throughput screening I performed using an automated malachite green-based colorimetric assay to detect compounds that directly inhibit helicase-catalyzed ATP hydrolysis of the Dengue virus. My hypothesis is predicated upon the important role played by NS3 helicase protein in the creation of the Dengue viral genome. Therefore, by inhibiting the function of the NS3 helicase, the growth of the Dengue virus is also inhibited. In this study, I screened over 4,000 samples containing small organic molecules for its potential to inhibit the ATPase activity of the DENV helicase protein. I also merged the assay results for Hepatitis C Virus (HCV) that another postdoc in our lab performed using the same set of compounds, this is represented in column “HCVHDAInh.”

This data set will have the following fields:

CdId : Compound ID

MolWeight : Molecular Weight

MolFormula : Molecular Formula

LotNumber : Lot Number

DaughterPlate : Daughter Plate

PlateMap : Plate Map

PercentPurity : Percent Purity

HCVHDAInh : HCV HDA Assay Percent Inhibition

DENVATPaseInh: Dengue ATPase Assay Percent Inhibition

IUPACName : IUPAC Name

SMILES : SMILES

LogP : Log P

LogD : Log D

HBondDonors : Hydrogen Bond Donors

HBondAcceptors: Hydrogen Bond Acceptors

RotatableBonds : Rotatable Bonds

LipinskiRuleof5 : Lipinski Rule of 5

**Data Limitations**

The main limitation of this data set is its relatively small size. It’s only composed of 4,079 rows which represent the number of molecules I screened. This data also assumes that the ATPase assay I used is an appropriate measure of biochemical inhibition of the target protein. While I can argue and vouch for the assay validity and robustness, it should still be noted that it can be subject to bias.

**Data Wrangling**

The initial data was converted to an excel spreadsheet file by the microplate reader machine used for screening. I further merged it with the molecule file that came from the vendor which contains molecular information such as SMILES (1D chemical representation). I then imported it to Instant Jchem software. This software further calculated molecular descriptors based on the chemical structure, such as LogP, LogD, # of H bond donors, # of H bond acceptors, # of Rotatable bonds, and whether it satisfies Lipinski’s Rule of 5.

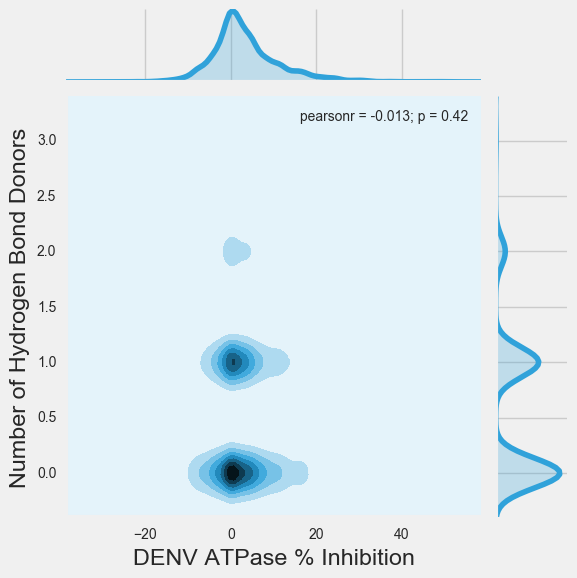
**Preliminary Exploration**

I would like to find out if certain patterns of behavior of the molecule correlate to its structure. For example:

1. Do bigger-sized molecules more/less potent inhibitors?
2. Does the number of benzene rings increase/decrease inhibitory capability?
3. Does the octanol/water index (LogP) affect ligand activity?

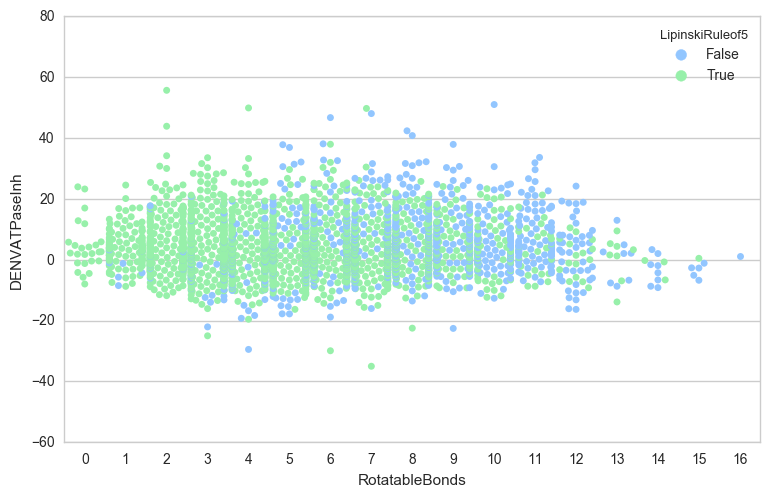
I have performed some simple data analysis to visualize the data. The ultimate goal for this project is to run some machine learning methods on the data such as linear regression and probably a discriminant or tree-based classification algorithm.

1. For now, I wanted to find out if there are relationships between a molecule’s physical characteristic and its inhibitory capability. So I first plotted the DENV helicase percent inhibition of the molecule against the number of hydrogen bond donors in the molecule. The two columns # hydrogen bond donors and # hydrogen bond acceptors can affect the molecule’s ability to form weak bonds with the biological target. And I wanted to see if it has any impact in the set of molecules I assayed.

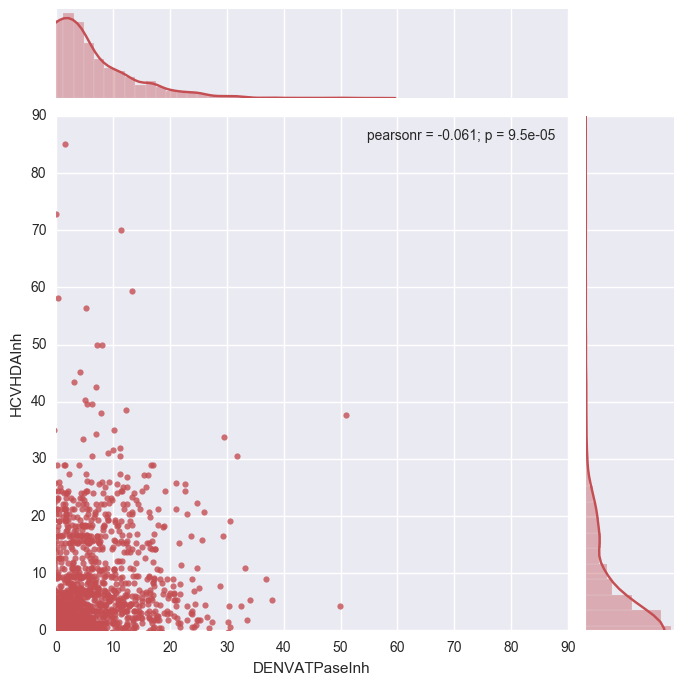


In this kdeplot, it appears that the number of hydrogen bond donors do affect the % inhibition of the molecule in the ATPase assay, such that the lack of it correlate with a higher inhibitory percentage.

1. Another relationship I wanted to detect is whether the number of rotatable bonds affect the inhibitory capability and whether the molecule satisfies the Lipinski Rule of 5. I used a scatter plot for this. The plot shows that the number of rotatable bonds somewhat affect the inhibitory capacity with a higher number of it correlating with a higher inhibitory rate.



1. I also wanted to find out if there are molecules that have inhibitory capability targeting both Hepatitis C and Dengue viruses. It shows one molecule that can inhibit both with greater than 50% effecicacy for DENV and close to 40% effectiveness for HCV.



**Bibliography**

1. Lock, E. F., Abdo, N., Huang, R., Xia, M., Kosyk, O., O'Shea, S. H., Rusyn, I. (2012). Quantitative high-throughput screening for chemical toxicity in a population-based in vitro model. Toxicological Sciences, 126(2), 578-588. DOI: [10.1093/toxsci/kfs023](http://dx.doi.org/10.1093/toxsci/kfs023)
2. Guzman M. G. *et al*. Dengue: A continuing global threat. *Nature Reviews Microbiol*ogy **8**, S7–S16 (2010).