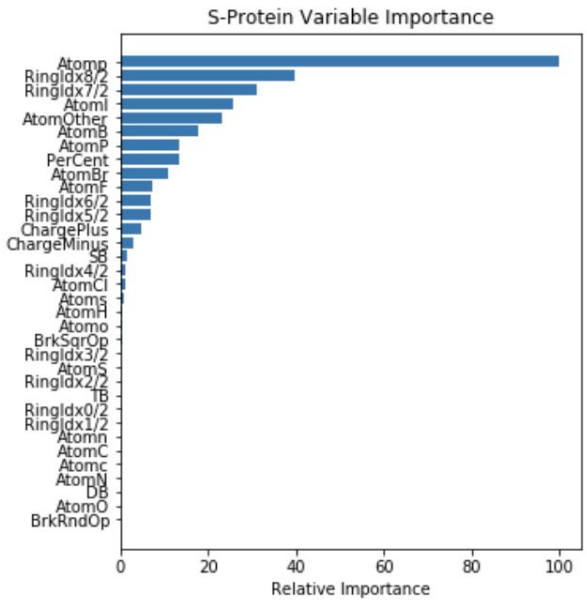
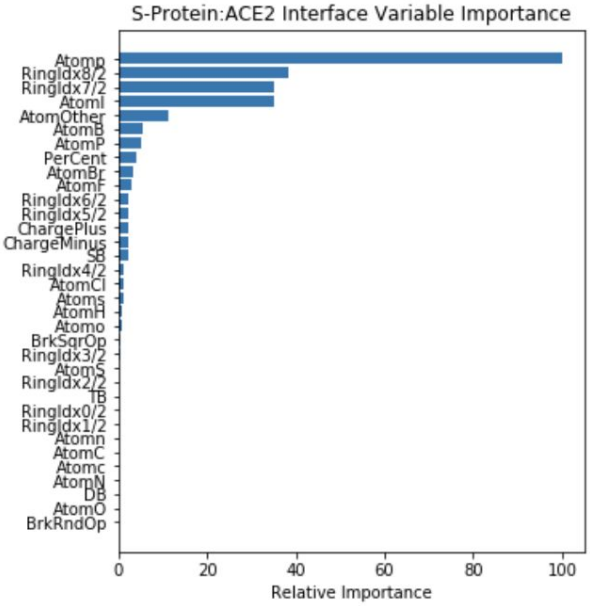
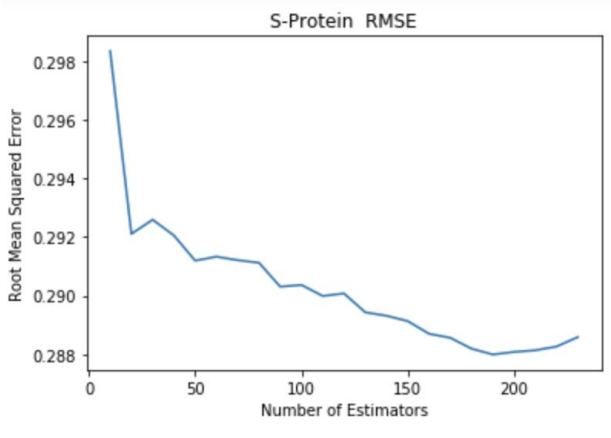
Victoria Lloyd
COVID19 Data Analytics
Final Project Progress
Due Wed, May 6, 2020

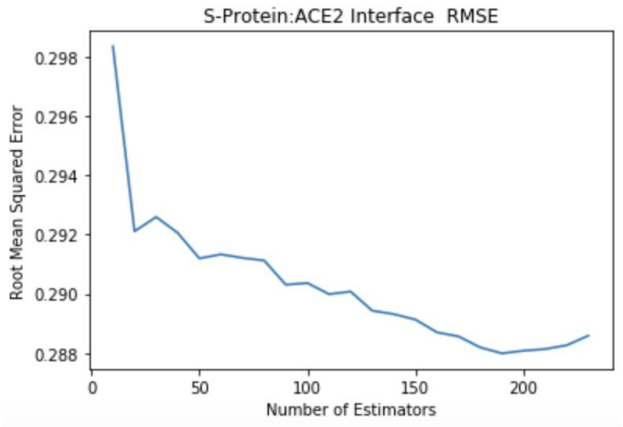
I decided to change topics for my final project since there was not enough publicly available data for the analysis I wanted to do. For my new project, I have decided to look into the use of machine learning techniques to screen small molecules to find treatments for COVID-19. In this project I have been expanding on the work of Smith et al. in *Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface* and Batra et al. in *Screening of Therapeutic Agents for COVID-19 using Machine Learning and Ensemble Docking Simulations*. In this project, Batra et al. used databases containing information about thousands of small molecules and used machine learning techniques to try to predict Vina scores from geometric and chemical information about the molecules. Using individual RF models, Batra discovered which small molecules had the lowest Vina scores, and would therefore be most able to bond to the outer spikes of the coronavirus exterior (S-Proteins) and to the interface between these protein and their human receptors, the Angiotensin-converting enzyme 2 (ACE2). After talking to Dr. Batra about his work on this project and accessing the data he used, I noticed that in his screening process he chose to use an RF model without verifying that gave the best results. In this project I decided to investigate other popular regression models and compare them to the RF models to verify that Batra used the best model type for his investigation.
I started by cleaning the data, removing all molecules with negative Vina scores in order to reduce the skewness of the model since we are hoping to identify molecules with low Vina scores. I created a function which takes in the SMILES representation of a molecule and creates a list of their geometric and chemical properties, or molecular fingerprint. I measured the variable importance on the S-Protein and the S-Protein:ACE2 interface, shown below:



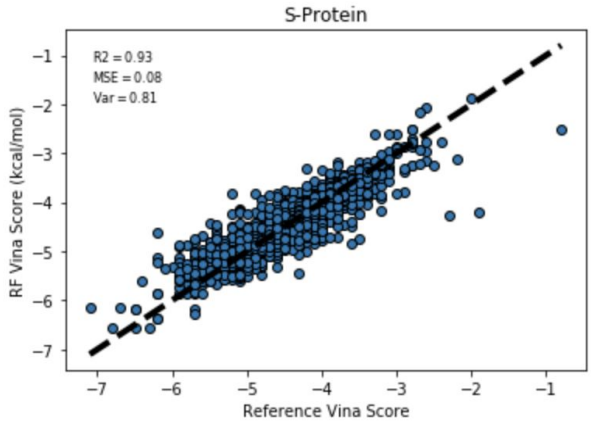


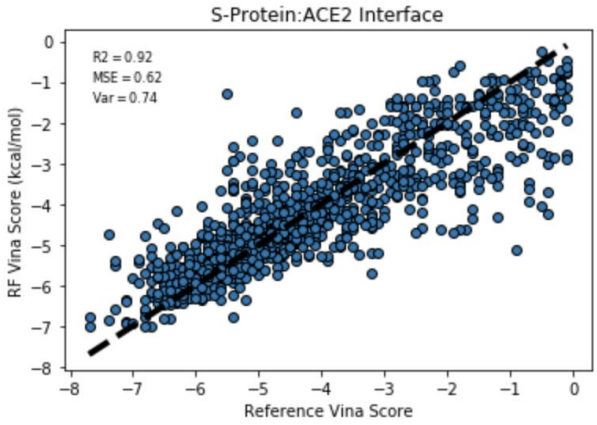
I then trained a RF model to estimate the Vina scores from the molecular fingerprint. To optimize the number of estimators to use in each model (one for the S-Protein and one for the S-Protein:ACE2 interface) I plotted the root mean square error for each model against the number of estimators:



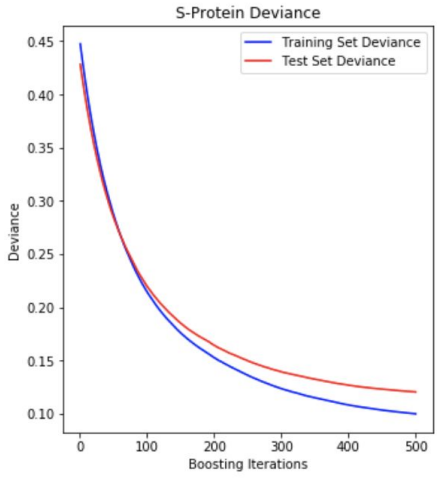


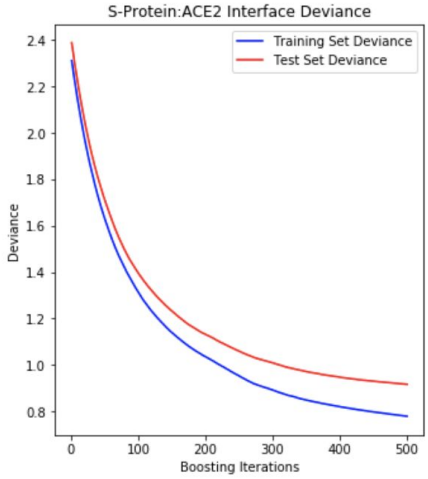
Looking at these plots I decided to set the number of estimators to 190 for both models. They yielded the following results:



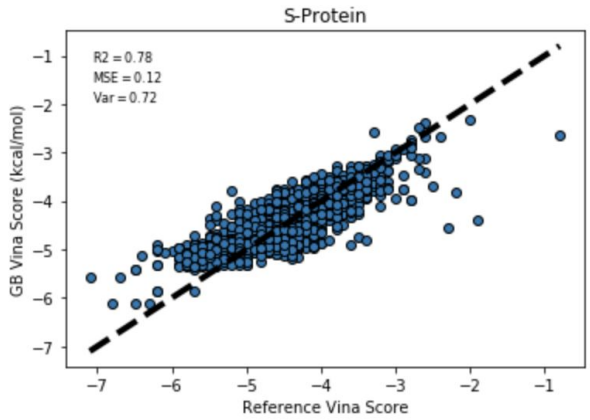


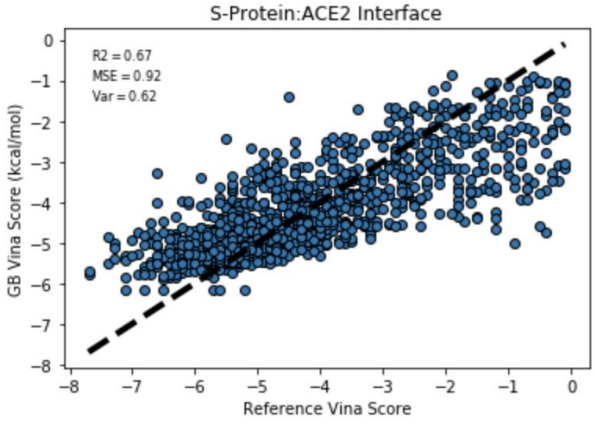
Similarly for the gradient boosting model plotted the deviance against the number of boosting iterations for both models:



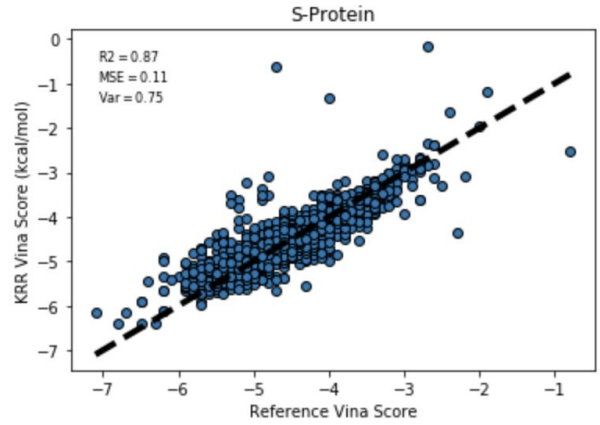


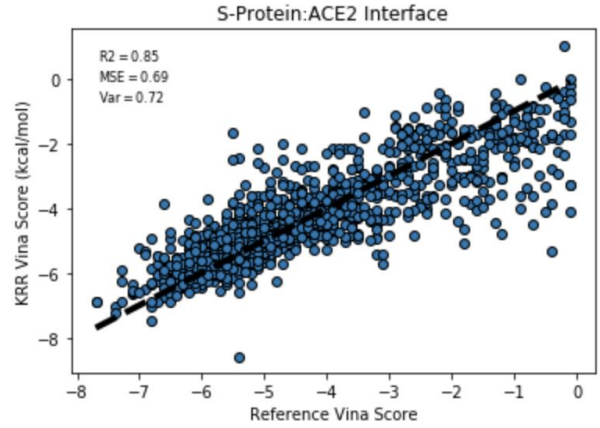
And created each model with the following results



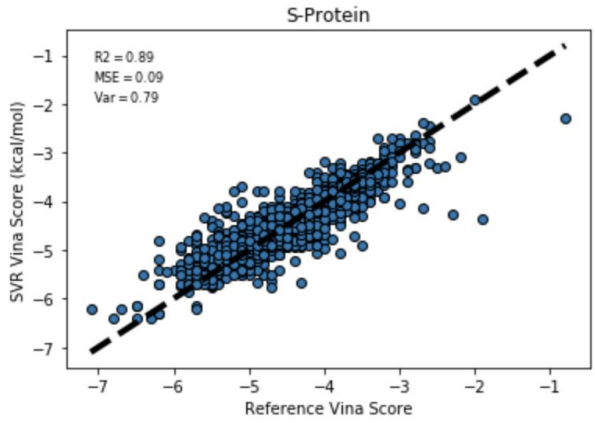


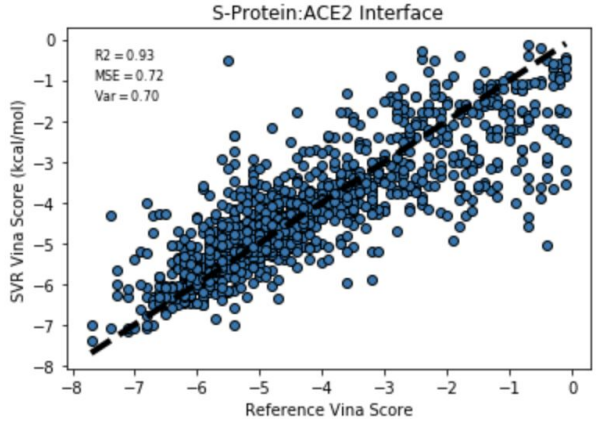
For the SVR and KRR models I used scikit-learn's grid search function to optimize each model, which generated the following results for KRR:



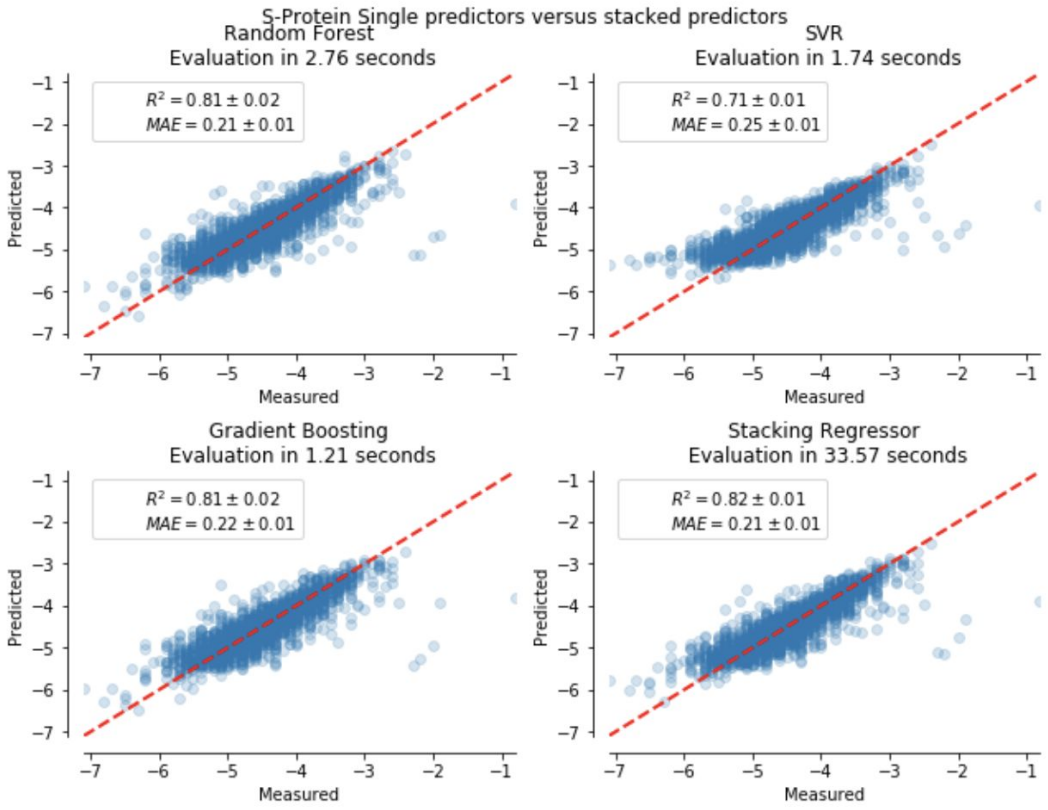


And for SVR:

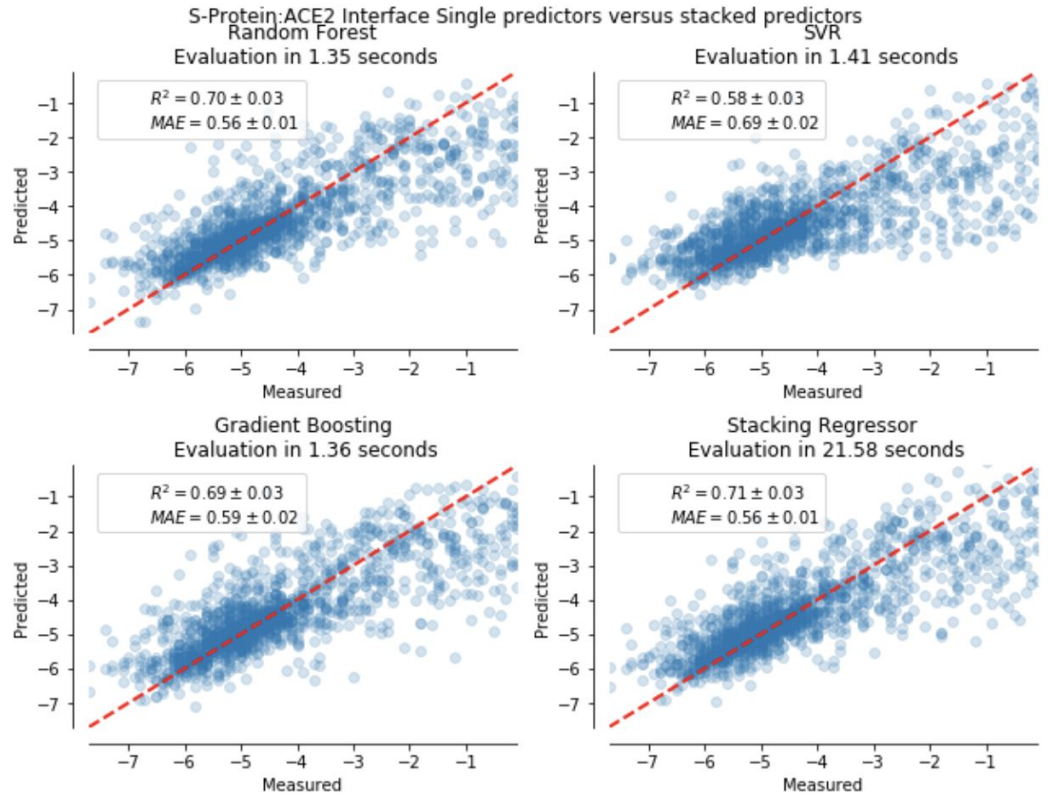




Lastly I used scikit-learn's stacking model to compare the predictions of each individual model - gradient boosting, SVR, RF - against the stacked predictions of these base estimators for the S-protein:



And for the S-Protein:ACE2 interface:



For the small differences between the results of the individual components of the stacking algorithms and my models I attribute this to differences in optimization. From this analysis we can see that for both the S-Protein the RF does have the highest R^2 value, the lowest mean squared error, and the shortest runtime. For the S-Protein:ACE2 interface, the stacked regressor has the lowest mean squared error but one of the lowest R^2 values, while the SVR has the highest R^2 value but the second highest error. The RF model seems to perform best for this model, with almost as high an R^2 value as SVR and an error only moderately above that of the stacking model while also having a much faster runtime. This analysis appears to confirm that the RF model was the best choice for the analysis performed by Batra et al.