Machine Learning for Protein-Protein Binding Affinity Prediction

CHW
CENTER FOR HEALTH AND WELLBEING
PRINCETON UNIVERSITY

Emmanuel Mhrous, '25, COS
Oswaldo Cruz Foundation

Funded by CHW under the GHP Internship Program

Introduction

- High-throughput drug discovery allows many different structural variations of a drug to be created
- However, how effective a given structure is is difficult to test without producing the drug in a lab
- Machine learning allows for the prediction of the binding affinity to select the most promising drugs

Objective of the Study

We hoped to test various different Machine Learning methods to see if there was a way to accurately predict protein-protein binding affinity.

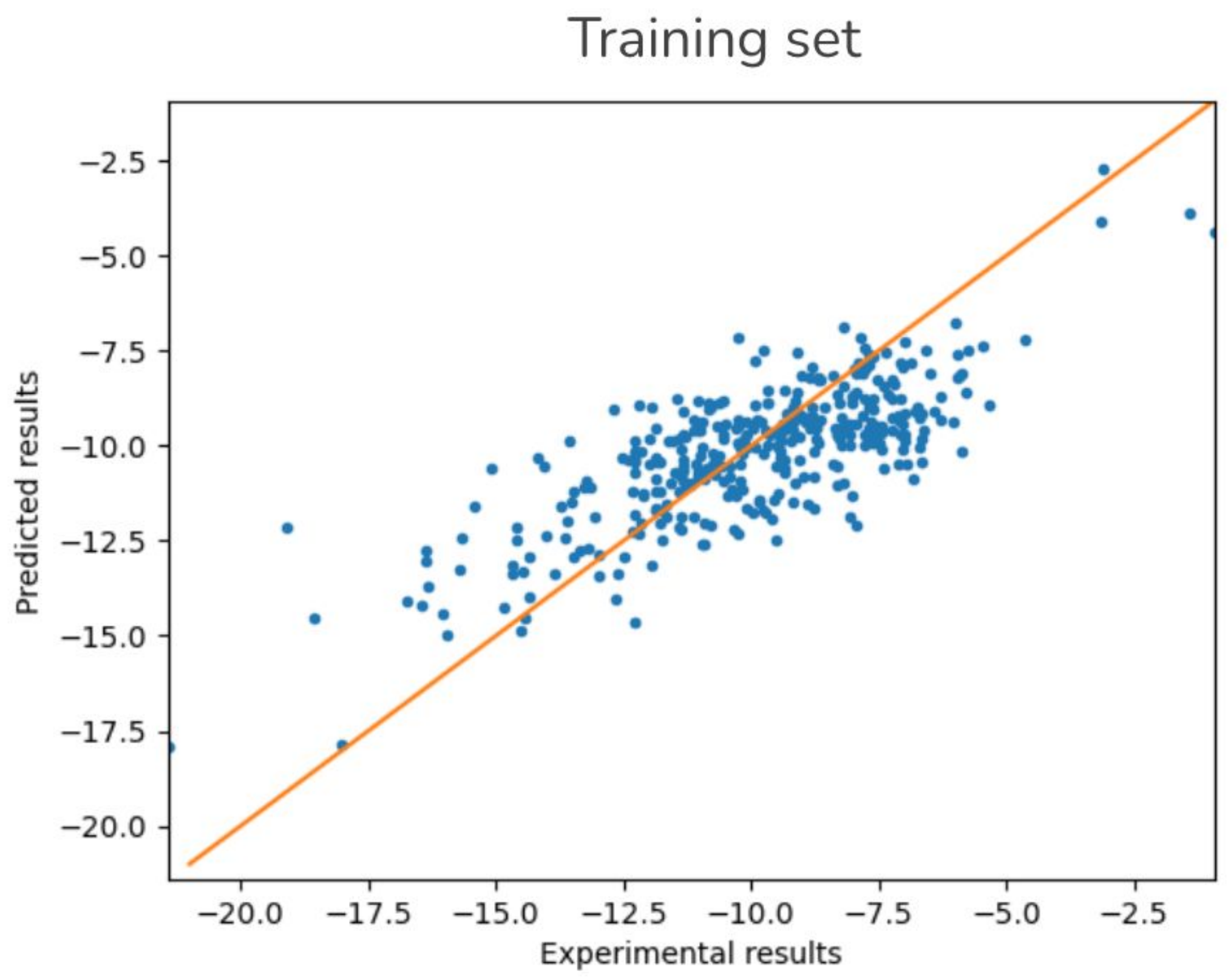
Methods

- Calculate structural properties using Rosetta, and split into training and testing set
- Research and apply various different Machine Learning methodologies to training set with k-means validation
- Apply most promising result to testing set

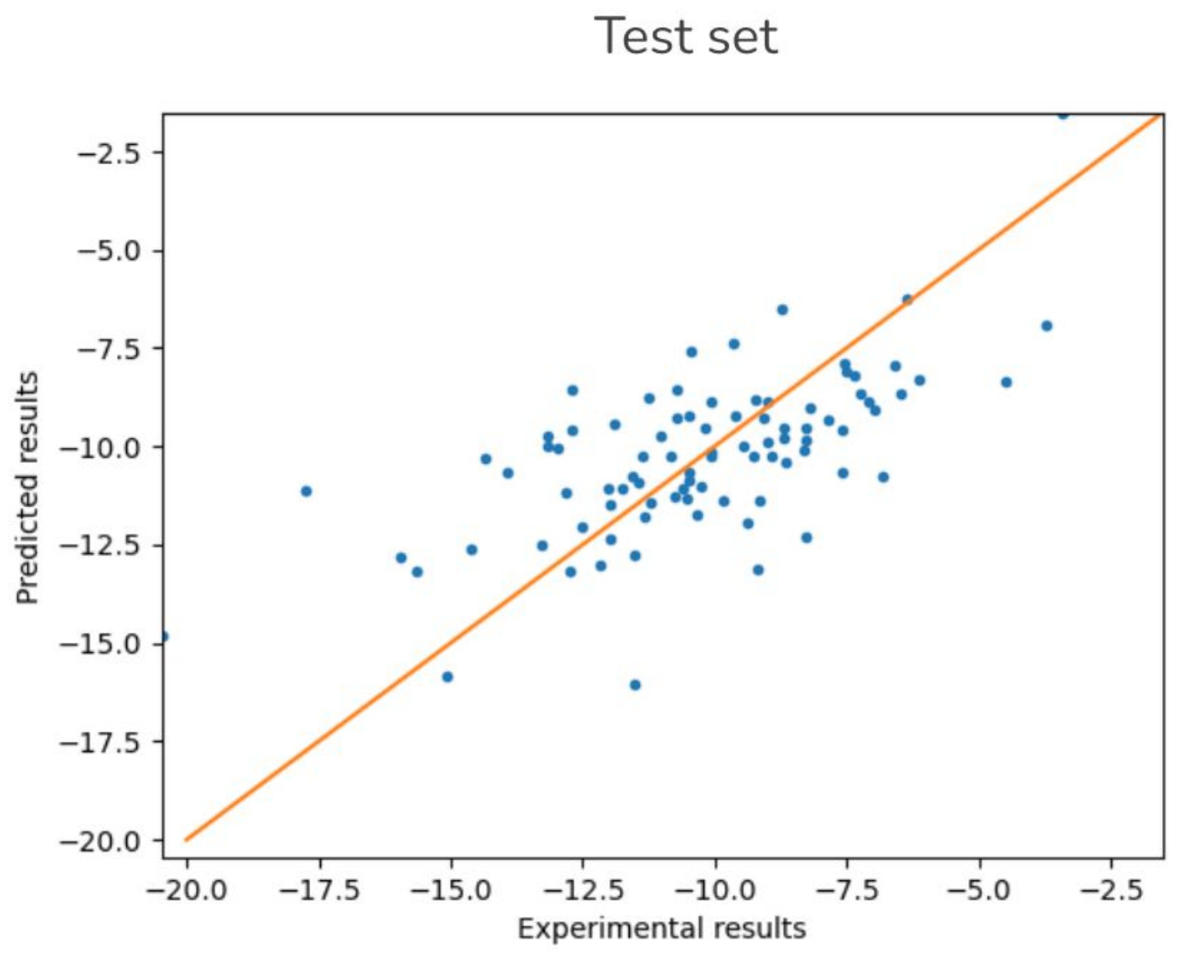
Results

- We attempted using different dimensionality methods (PCA, UMAP, LDA) in combination with the regression methods below; it was found that the best results were found using the entire dataset
- The parameters were calculated using Rosetta, which uses protein structures to make calculations of interactions between and within the proteins.
- We experimented with different calculated parameters until we again found a molecular simulation which provided the best results

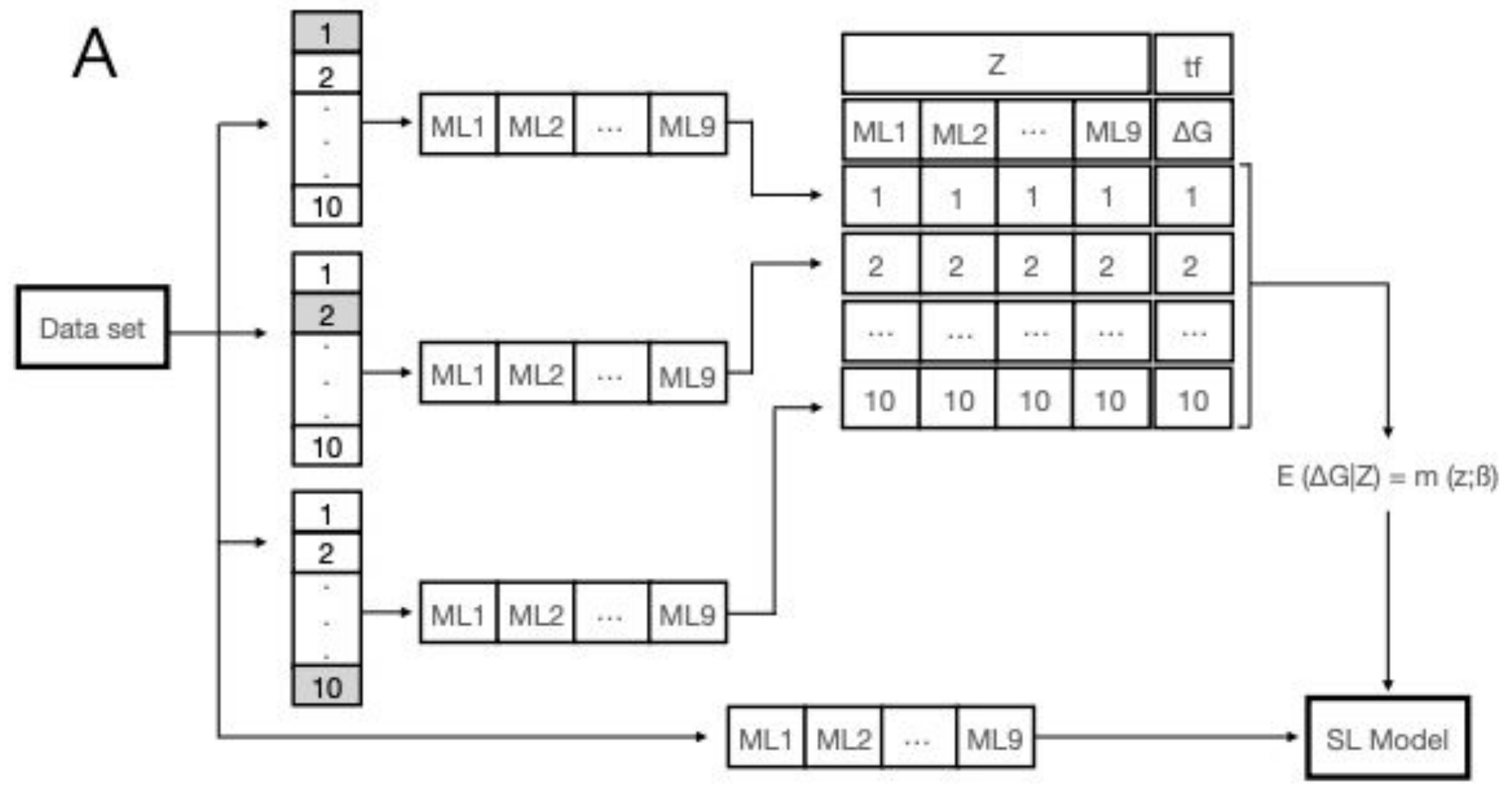
- We used a hyperparameter tuning to determine the best set of combinations in a neural network. The hyperparameters we tuned included dropout rate, regularization, dropout function, number of hidden layers, and size of hidden layers.
- We tuned using Bayesian Hypertuning, which uses the results of previous neural network iterations to predict the optimal set of parameters



RMSE=1.734, R=0.775



RMSE=2.151, R=0.655



- The hyperparameter tuning looked to minimize the mean squared error of k-fold cross validation, a method allowing us to see if our model accurately predicted unseen data
- Ultimately, the results of the neural network were combined with other machine learning methods to make a "super-learner" model

Discussion

- There is much more testing and research to be done in this field to help bring the optimal results
- Our model's shortcomings can be seen in its high RMSE - in the future, more accurate (and complex) methods and tools can help with lowering this
- The algorithms are also "black box" - we don't know why they work the way they do

Conclusion

We used the features of previously-solved protein-protein complexes to calculate features of these complexes, then used these features to create a neural network to predict the results within 2.3 kJ/mol

Acknowledgements

Thank you to the CHW and its staff for being so helpful with this process (especially Ms. Tohill), the Fiocruz Pernambuco birology department, and especially Prof. Lins, Elton, Matheus, and Manasses both for their tremendous help and support in this project, and for taking the time to show me around Brazil.