“A novel framework for the identification of drug target proteins: Combining stacked auto-encoders with a biased support vector machine” - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409512/>

This dataset contains biochemical statistics of ~5400 proteins classified as either a druggable target protein (DTP) or a non-druggable protein targets (NDTP). The features in this data set are calculated suing EMBOSS PepStat Calculator from the European Bioinformatics Institue. The calculator creates this dataset based off of the amino acid sequence of the protein alone. For this Project I would like to recreate the work from this paper using different machine learning models, and attempting to employ auto encoders.

Drug discovery data - <https://www.kaggle.com/datasets/ruanrodriguez/drug-discovery-data>

This data set has drug discovery data that is provided by Laboratory innovation Science at Harvard. The purpose of the data is to analyze the probablility of discovering new treatments for mycobacterium Tuberculosis. The data is representative of experimental results.

Structural Protein Sequences- <https://www.kaggle.com/code/basu369victor/attention-based-protein-structure-prediction/data>

This is a protein data set retrieved from Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). Understanding of protein structure and access to proper crystallographic structures is essential for rationally derived in silico drug design methods. This is a data set that can be interrogated and used to develop a machine learning model that ca predict certain characteristics about proteins. I would like to use it to predict protein structural components.