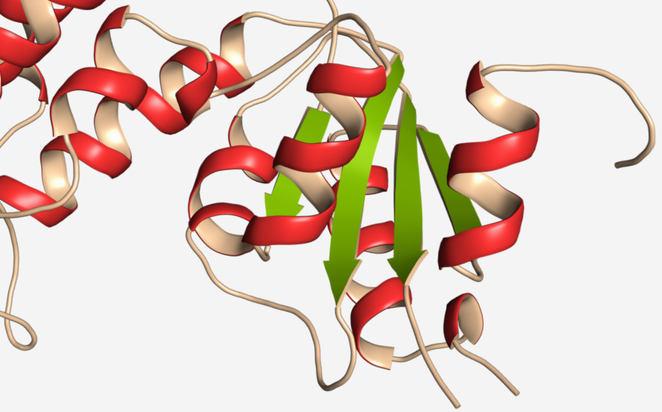
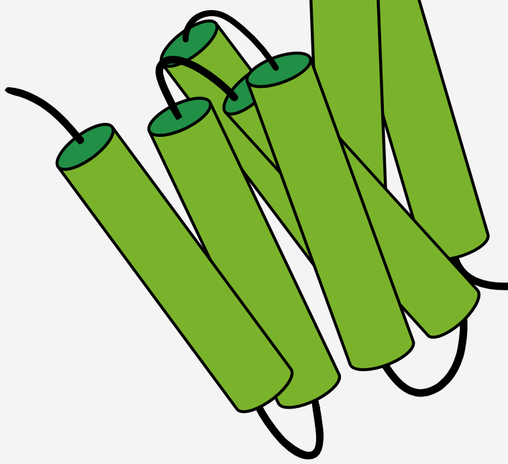
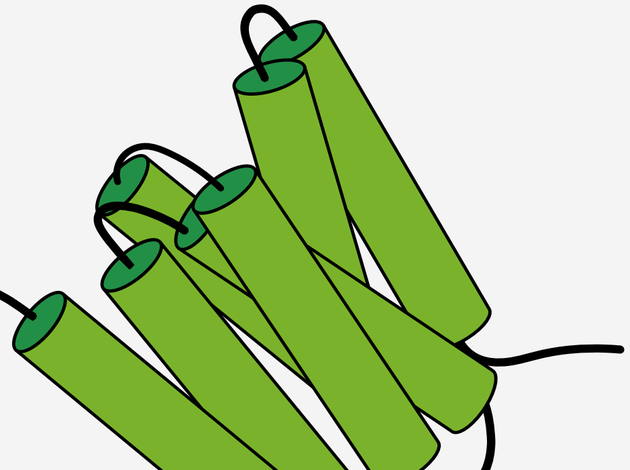
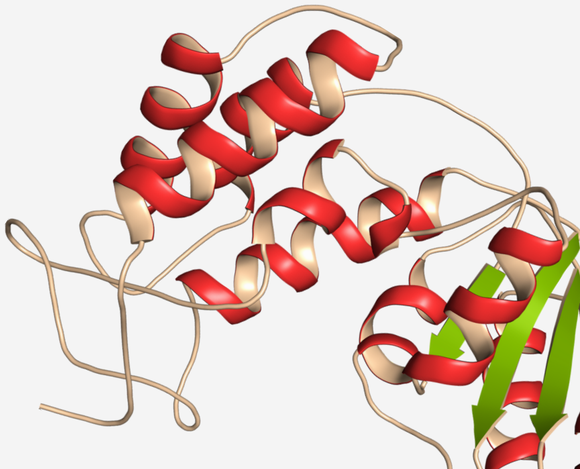


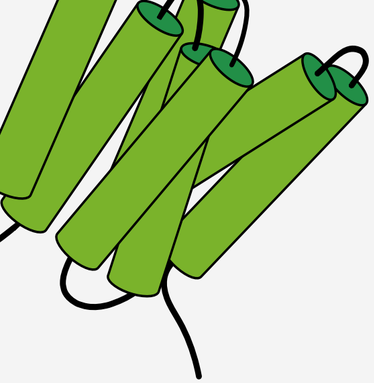


CLASSIFYING PROTEINS FROM THE AMINO ACID CODE

By Jediael Desir







ABOUT ME

Jediael Desir

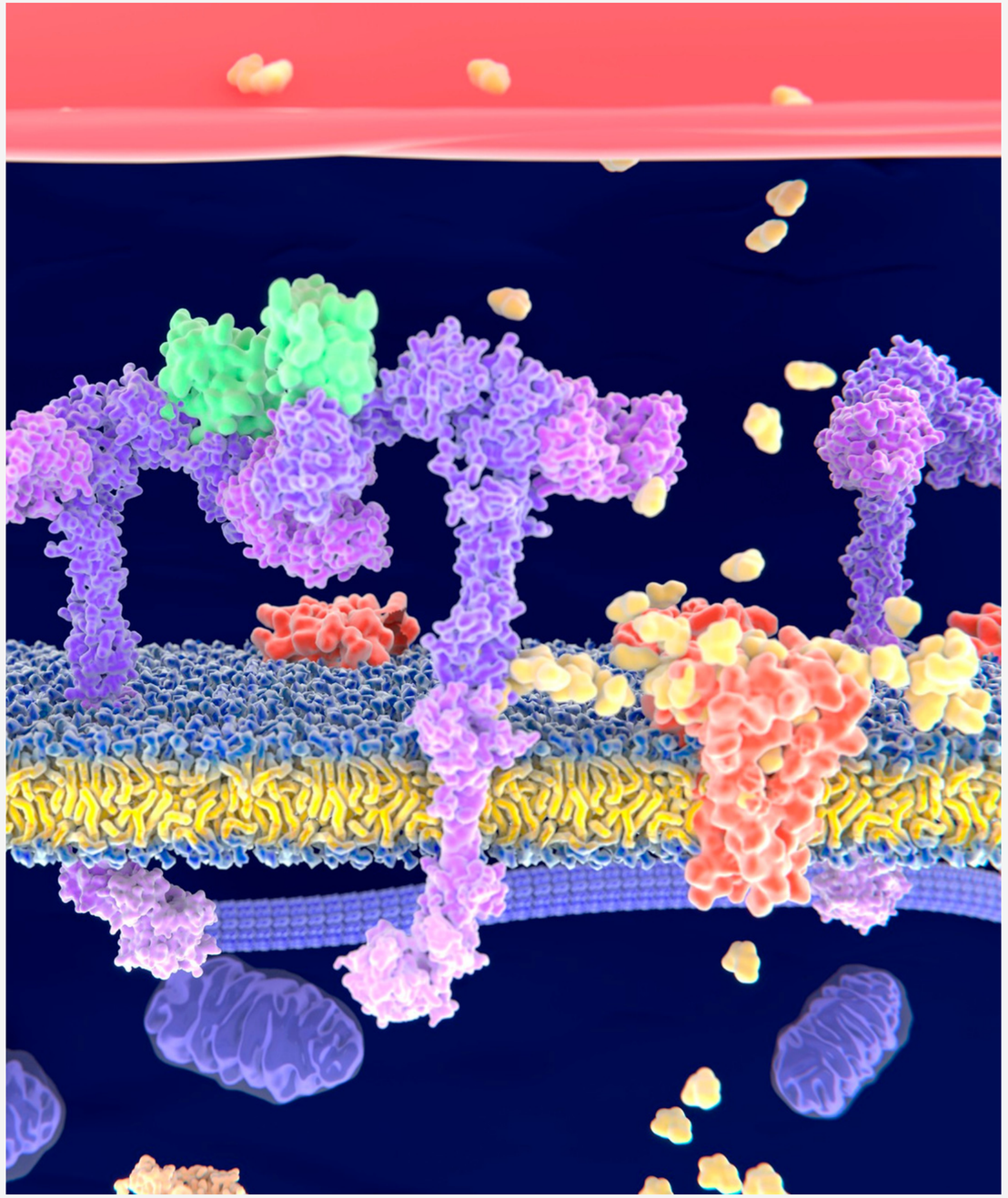
Data Scientist
B.A. in Biology

Backround: Microbiology

Github: jdthedatascientist
Email: jediaeldsr@gmail.com

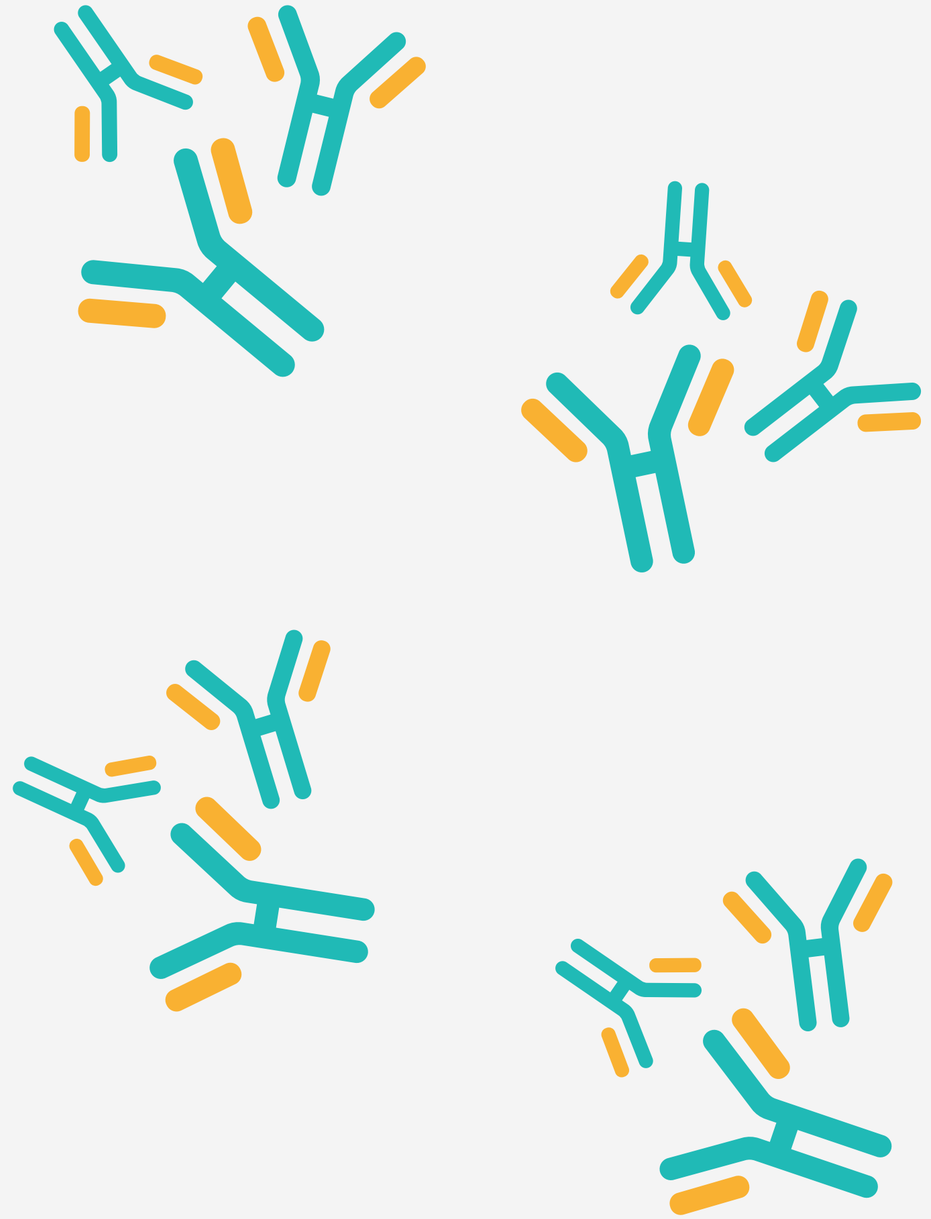
Olivia Wilson

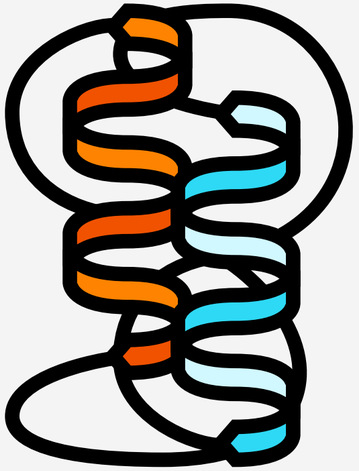
Contents



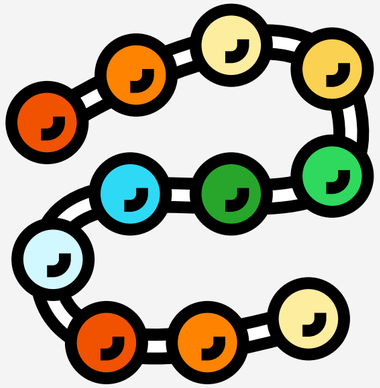
Business Problem

The continuous discovery of proteins presents challenges in their classification and functional understanding, particularly due to the inefficiency in swiftly assigning new proteins to their respective families. To address this, a computational solution leveraging amino acid sequence and residue data is proposed. Advanced modeling techniques will be employed to facilitate rapid and accurate classification of proteins. The primary goals of this initiative are to accelerate protein research by enabling efficient classification and functional elucidation, ultimately driving innovation and commercialization in industries dependent on protein-based technologies.

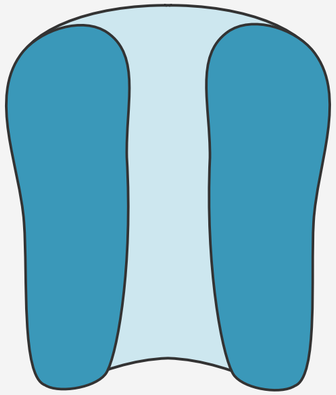


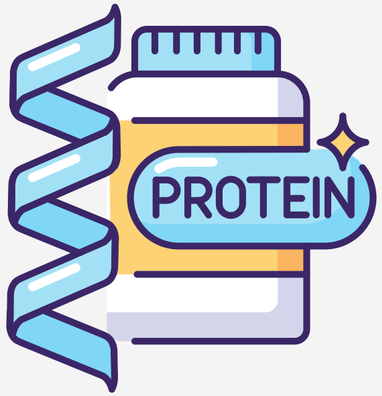


Data Overview

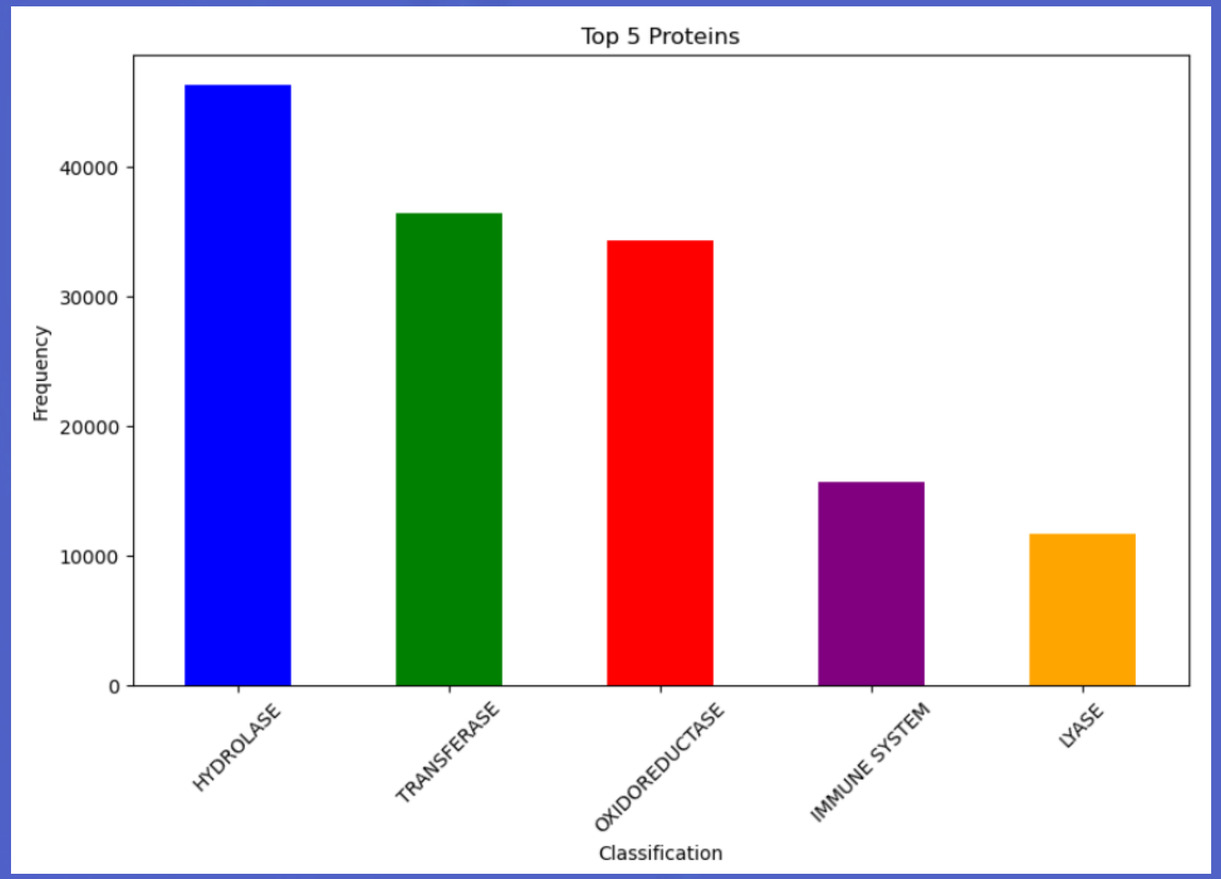


- Data downloaded from Kaggle and sourced from (RCSB) Protein Data Bank (PDB)
- Data Set included 2 dataframes with 141,401 and 467,304 entries, respectively
- After data manipulation 144,378 entries remained

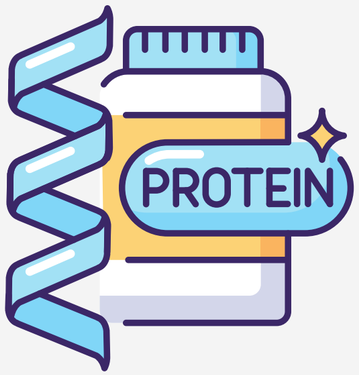


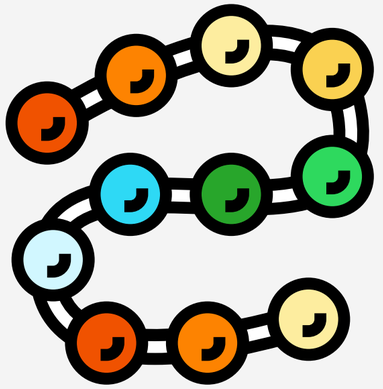


DATA PREP



- Merged data frames
- Dropped irrelevant columns and null values.
- The data was subsetted on 'macromolecule_Type_x'
- Five most frequent proteins in the dataset left for modeling
- Initially 4,468 macromolecules present in dataset
- Label encoding used to encode sequence data into numerical values

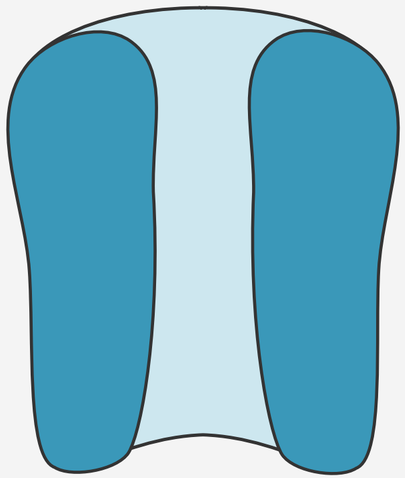


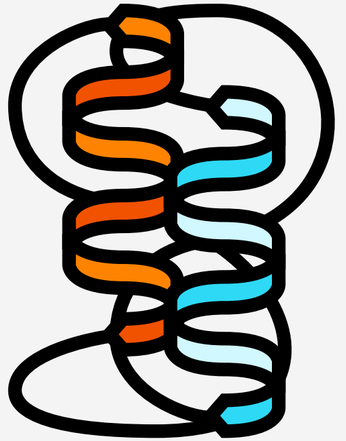


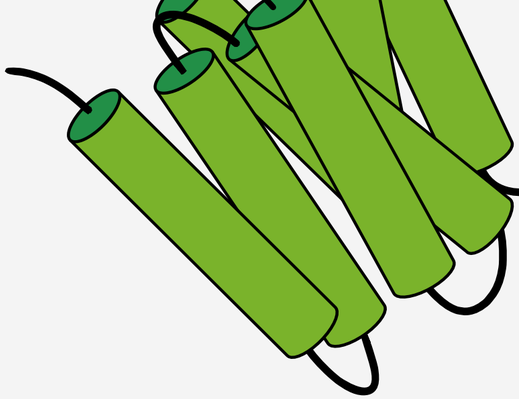
Modeling

Classification Models utilized were:

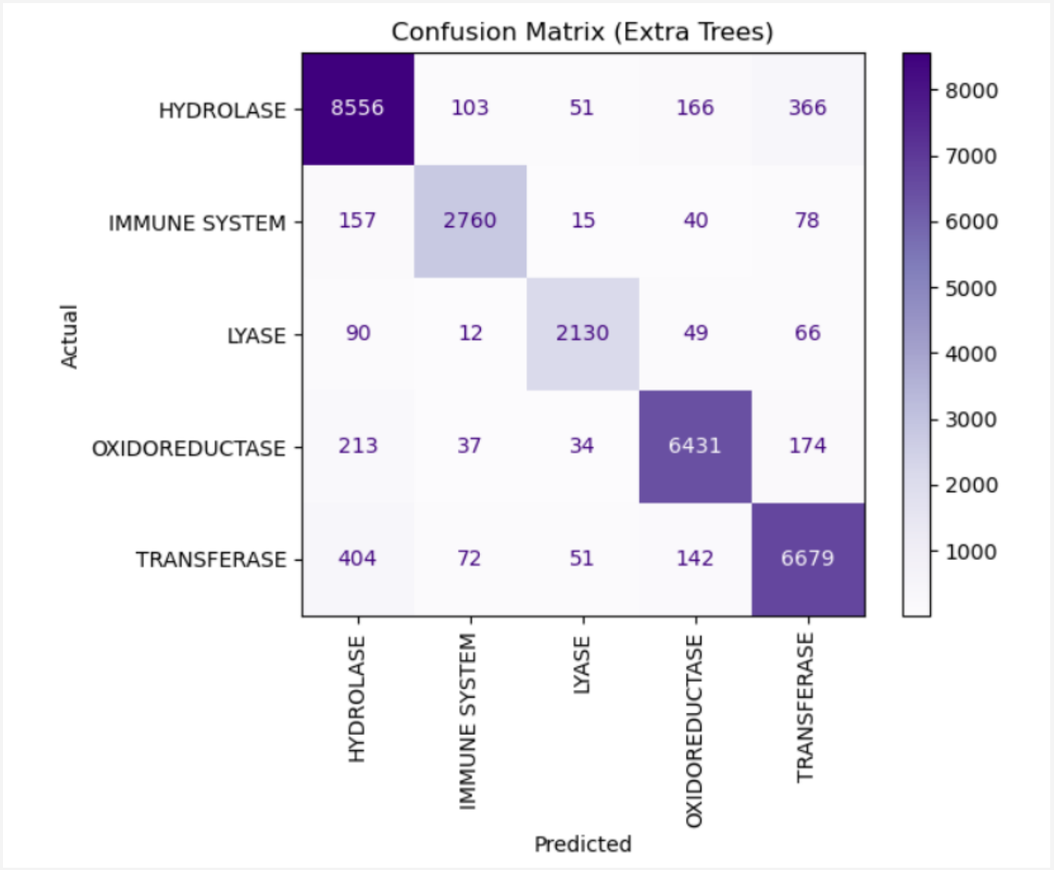
- KNN
- Descision Trees
- Random Forests
- **Extra Trees (Best Model)**







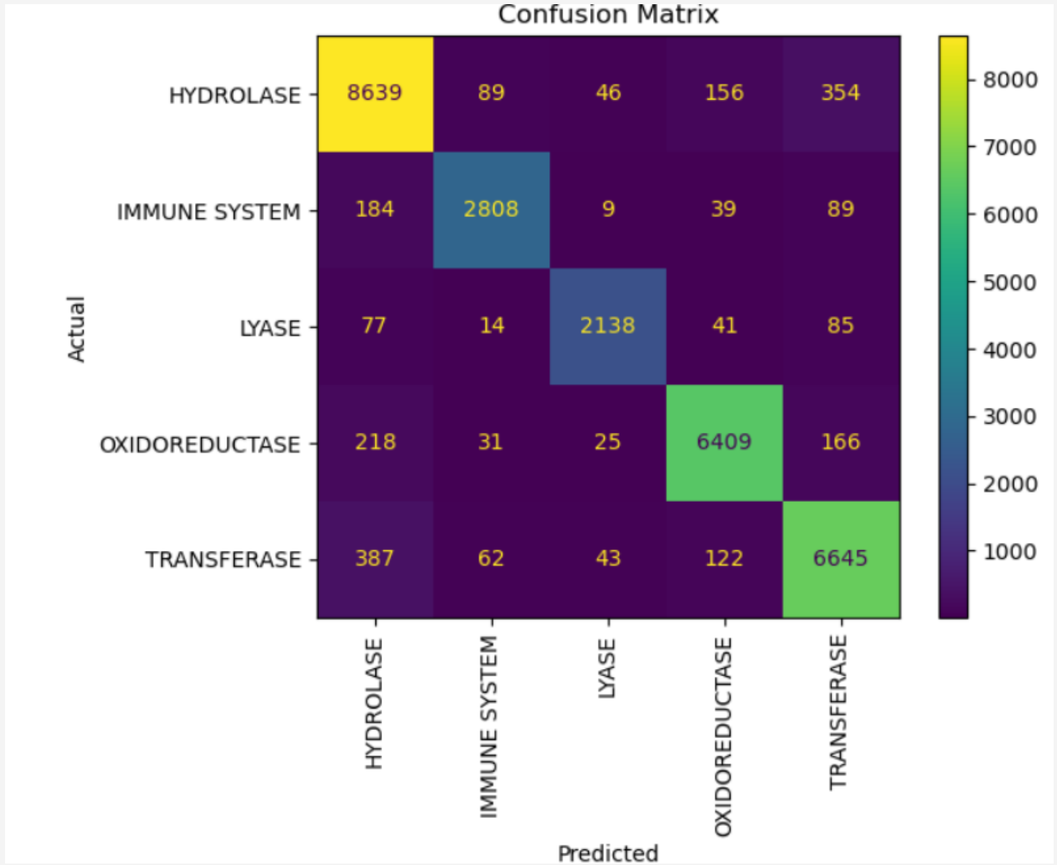
Modeling



Training Accuracy: 99.8 %

Validation Accuracy: 92.0 %

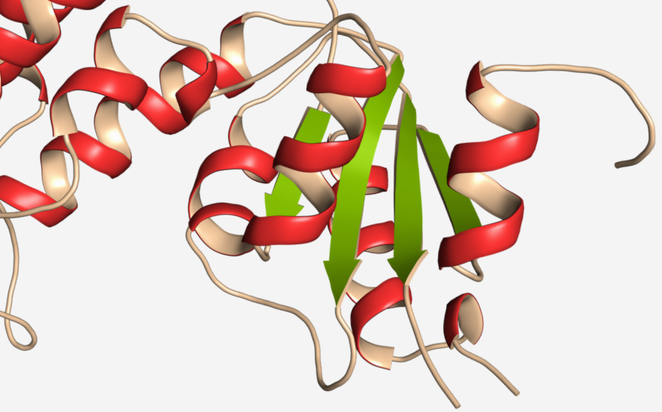
Evaluation

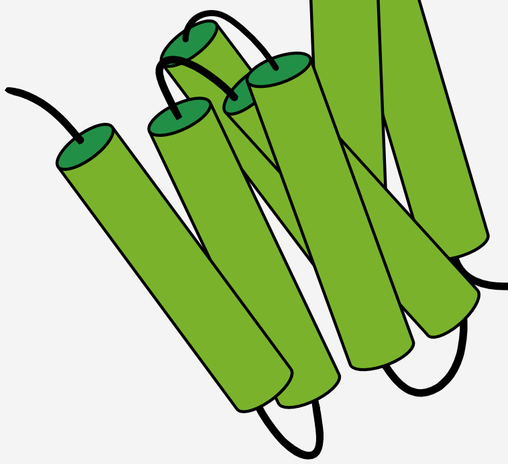


Accuracy Used

Validation Set: 92.1%

Test Set: 92.3%





Next Steps

- Do a comparative analysis between protein sequences
- Use features other than sequence or residue data to build a predictive model.
- unsupervised learning analysis of sequence data