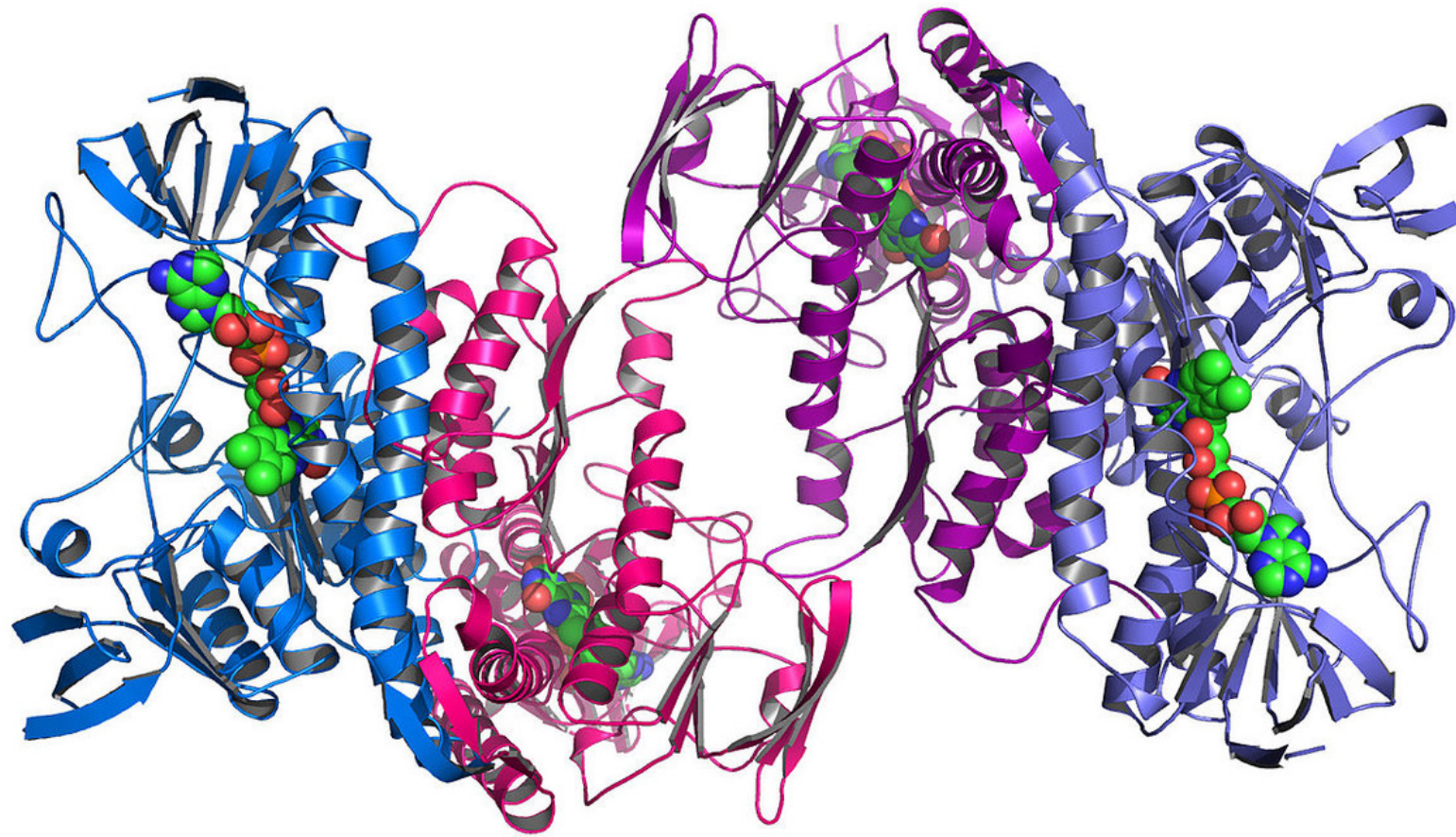


# ***Protein Classification***

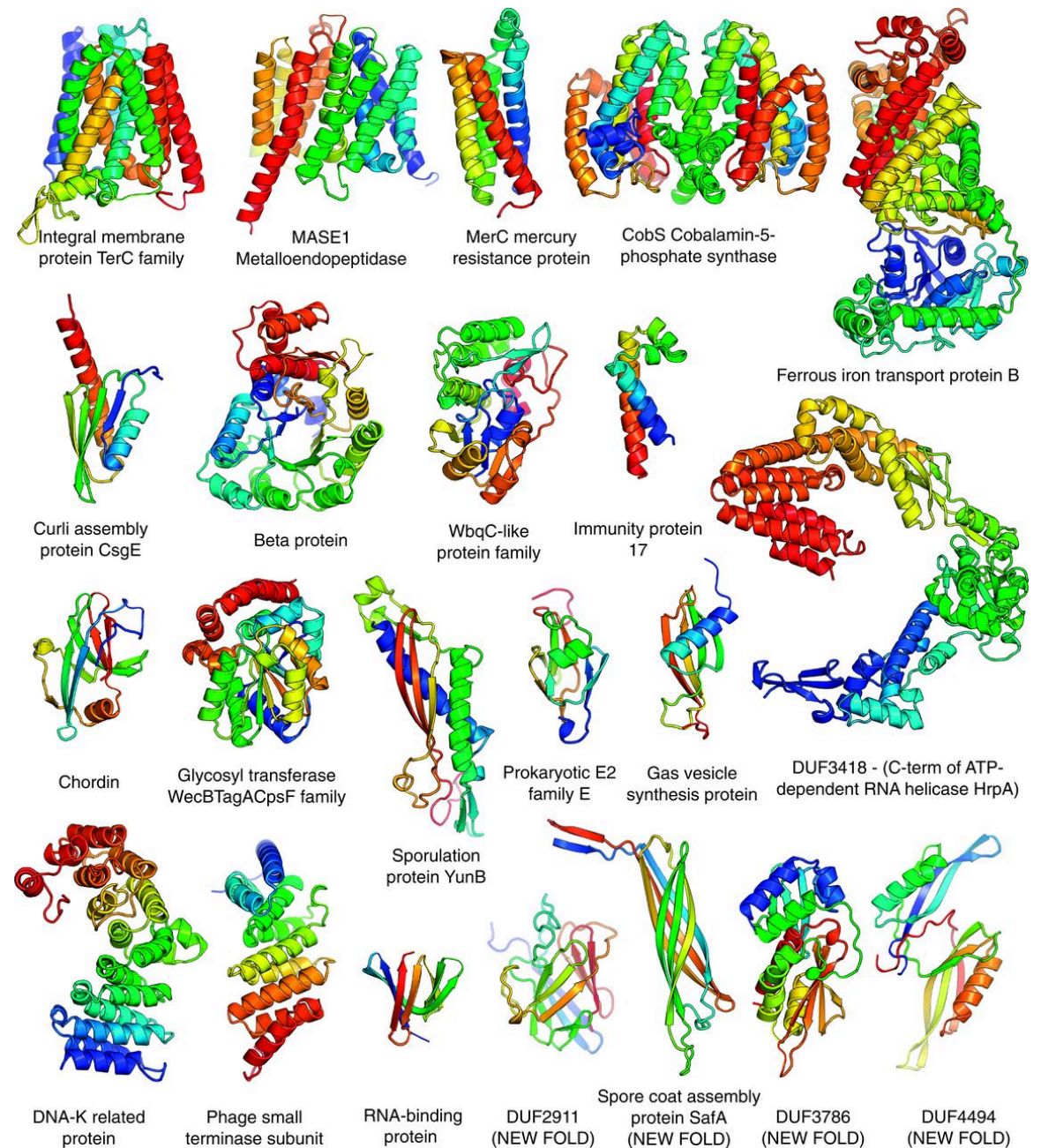
***Functional & Cellular Location Prediction***

***Scott W. Lew***



# PROTEINS

- **Biological Polymers**
- **Made of 20 naturally occurring amino acids**
- **Essential for life: enzymes, immune system defenders, necessary for thought, sensation, digestion, breathing, .....**



# Outline

- **Supervised Machine Learning Classification *using only the Protein Amino Acid Sequence***
- **Predict the function: *what does the protein do?***
- **Predict the cellular location: *where is the protein inside the cell?***



# Motivation

- *A scientist in biotech industry or academia discovers a novel Protein that plays a vital role in some disease*
- *How to determine what it does? Or where it is located?*
- *One solution: apply Machine Learning Methods to predict the function and location  
....Save time & energy*



# ***Functional Class***

*What does the protein do?*

**Some Functions:**

***Hydrolase (a type of enzyme)***

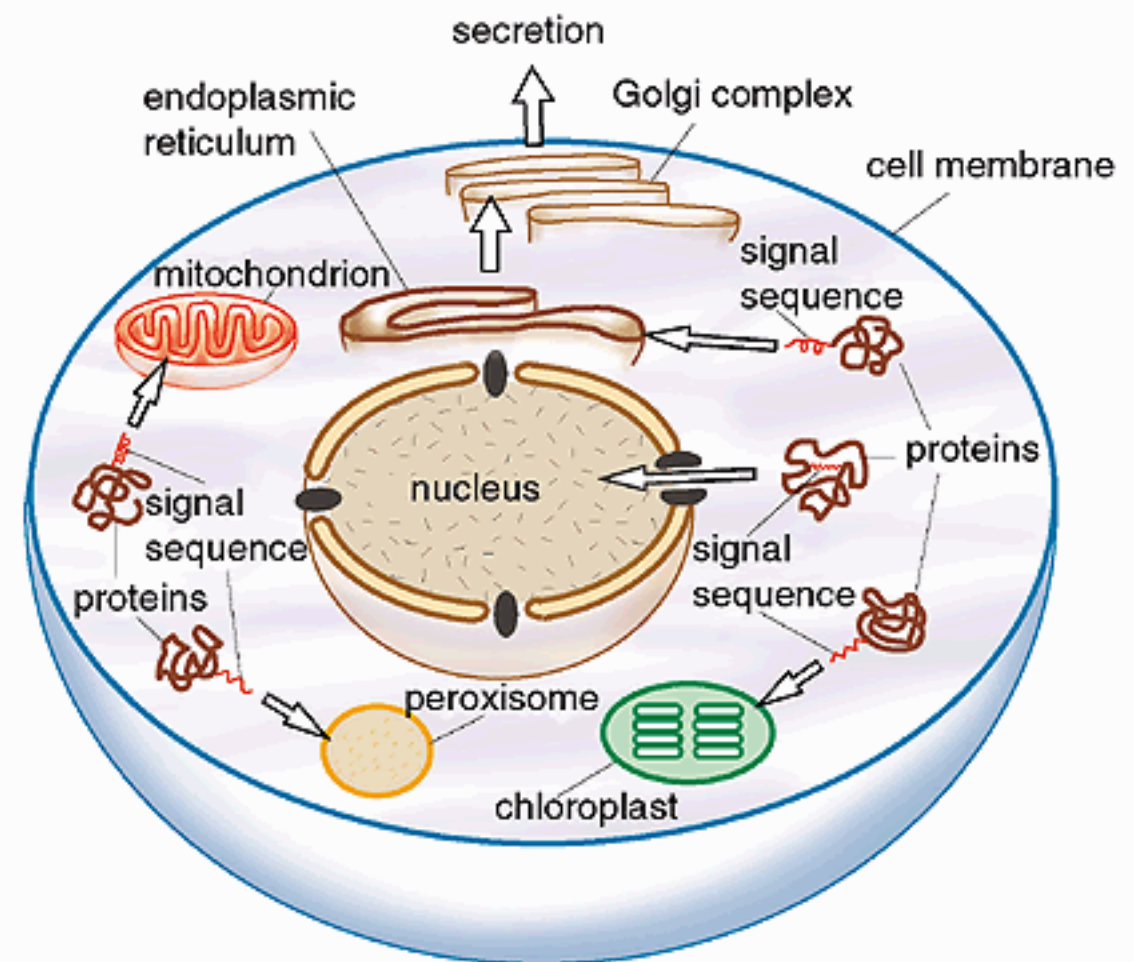
***DNA binding***

***Transporter***

***Immune system***

# *Cellular Location*

- Cell is divided into compartments
- Each protein has its own location inside the cell
- Predict where the protein is:  
Nucleus? ER?  
mitochondrion?....etc, etc



# Peptide Count:

*Sliding Window*

**W A L A L E U**

**W A L A L E U**

**W A L A L E U**

**W A L A L E U**

**W A L A L E U**

**W A L A L E U**

WAL	ALA	LAL	ALE	LEU	PAW
1	1	1	1	1	0

One-Hot Encoding

# Count Vectorizer

***String is converted into a Vector***  
*Using Peptide (**Substring**) Count with Sliding Window*

MVTVGNYCEAGPSEALAVGP...

*Protein Sequence String*



[1 0 1 0 0 0 1 1 1 0 0...]

*Vector*

M V T V G ✓

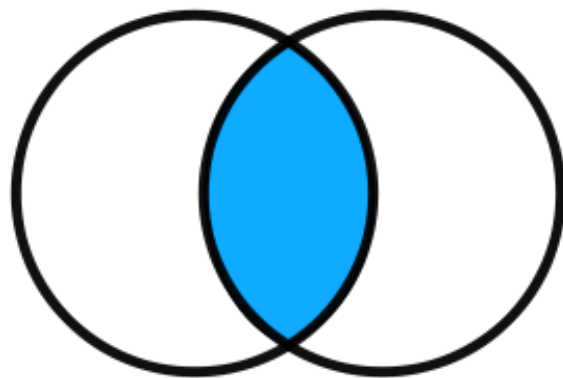
V T V G N ✓

P S E R P ✗

.....



***“all models are flawed, but some are useful”***

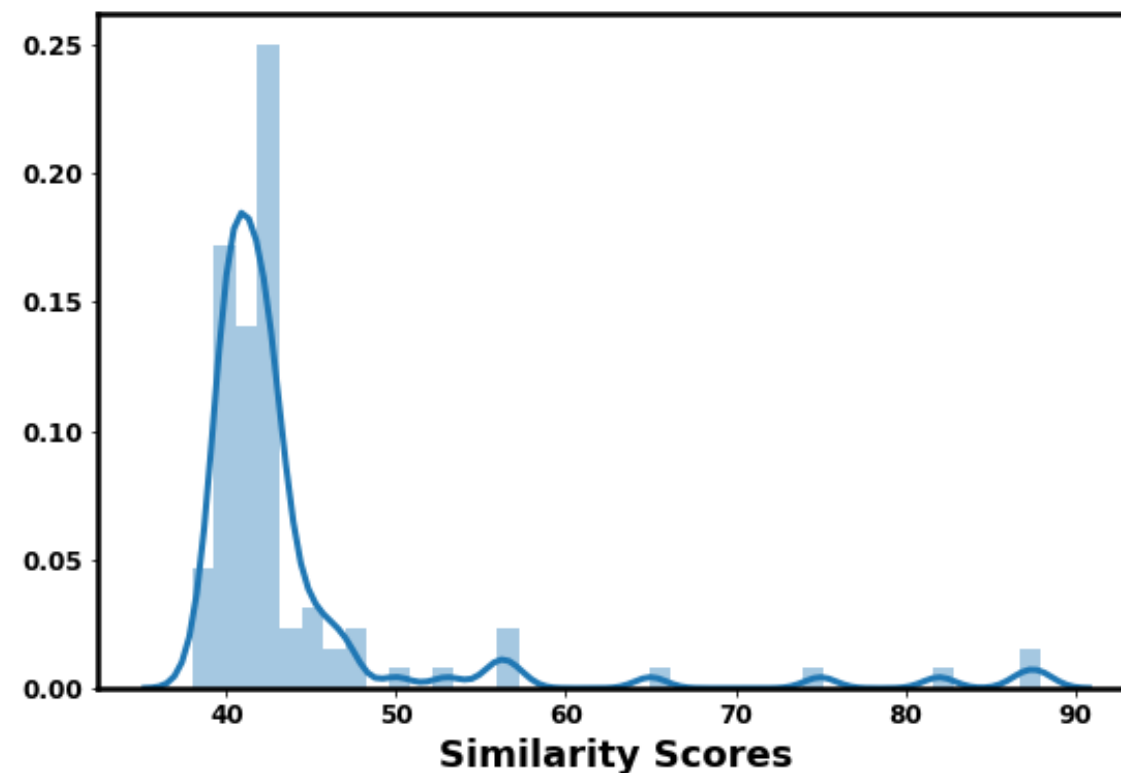


**Fuzzy Wuzzy Analysis  
for String Comparison**

**100 Hydrolase sequences  
were compared with  
Transporter Sequences  
using Fuzzy Wuzzy**

**Most Hydrolases  
have < 50% similarity  
with Transporter sequences.**

**But, some have more than  
70% similarity.....  
Overlapping categories**



# *ABC Transporters*

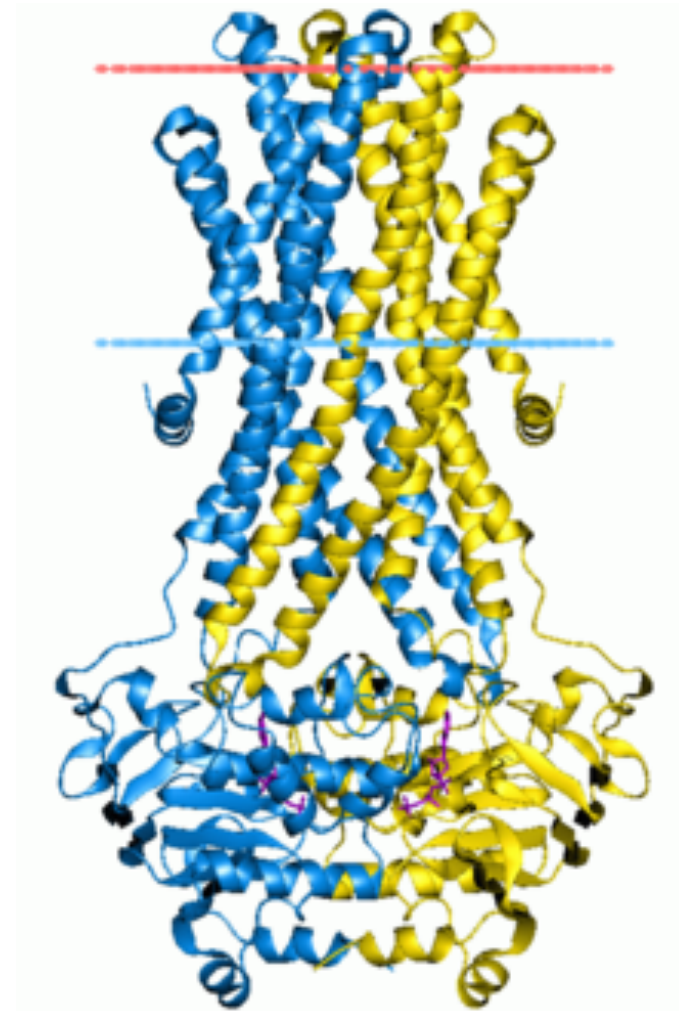
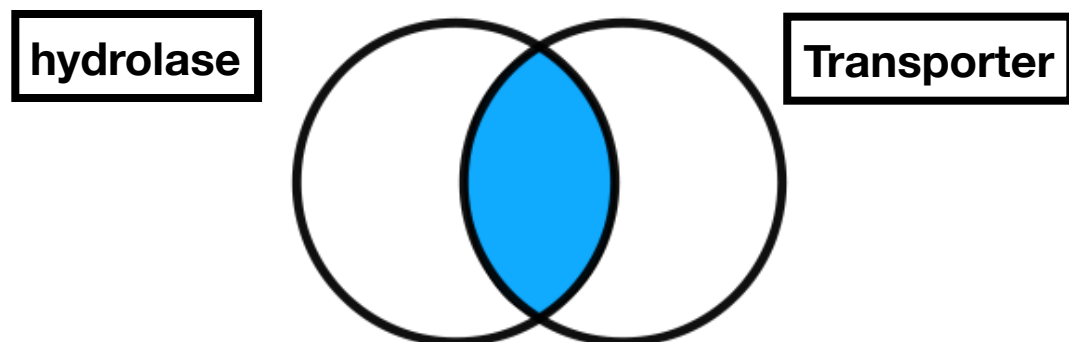
ABC Transporters are proteins that can be classified as

**BOTH**:

*a hydrolase (an enzyme)!*

**AND** *a Transporter!*

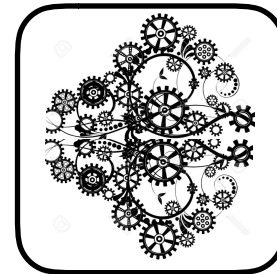
***Belong to 2 Functional Classes!***



# *Machine Learning Classification*

MVTVGNYCEAEGPVGP...

*Function*



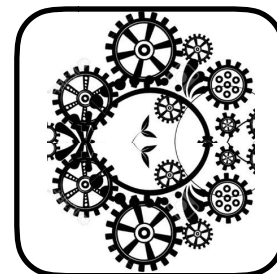
DNA Binding

Ligase

Hydrolase

MFDLEYQLKNLPDKPGV...

*Location*



Nucleus

Golgi

Mitochondria

# *Classifiers*

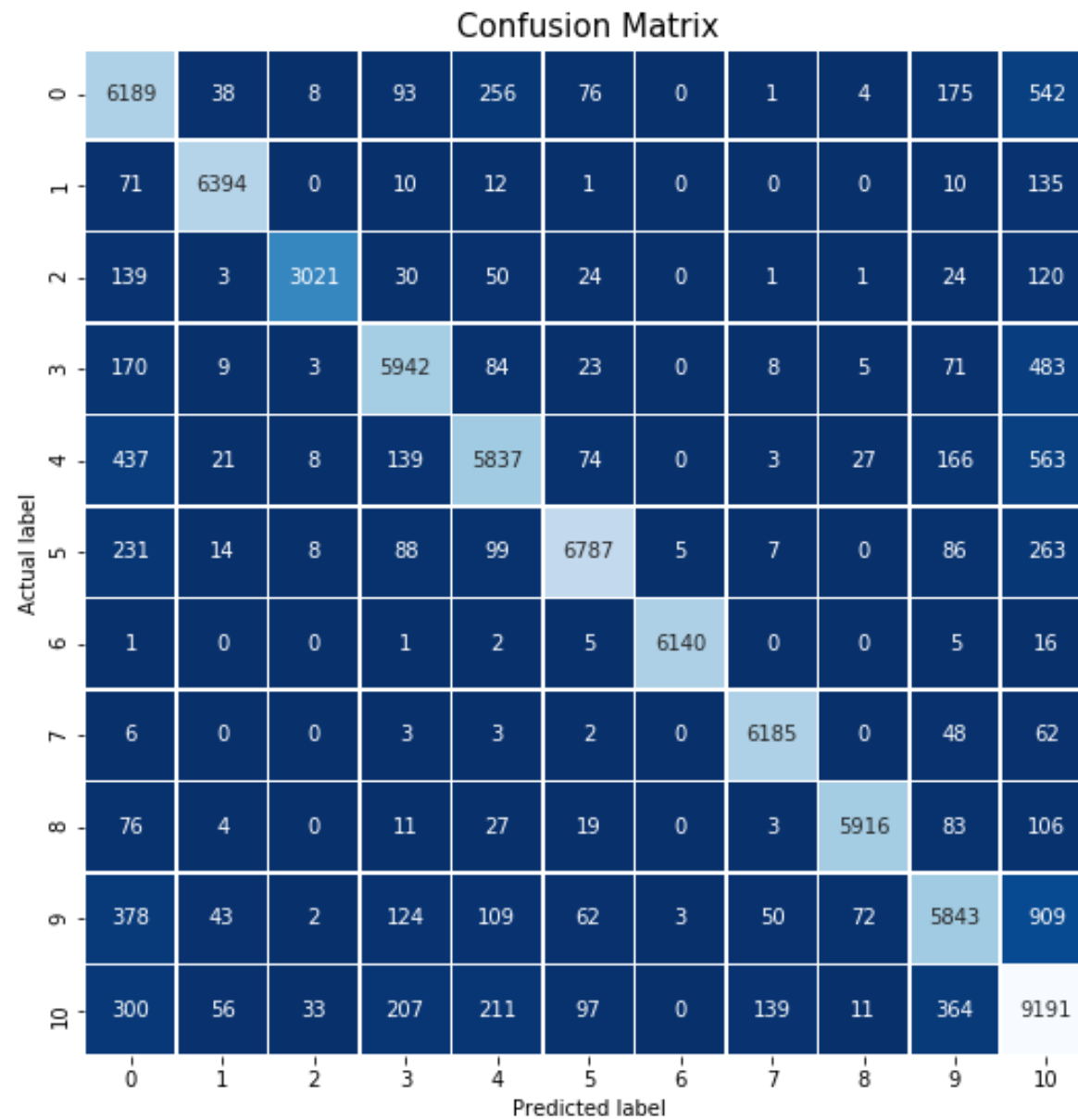
- **Function Prediction:** Majority Voting Classifier made of *Linear Model SGD, Passive Aggressive Classifier, & Multinomial Naive Bayes*
- **Location Prediction:** Majority Voting Classifier made of *Linear Model SGD, Passive Aggressive Classifier, Perceptron, & Multinomial Naive Bayes*

# *Accuracy & F1 Scores*

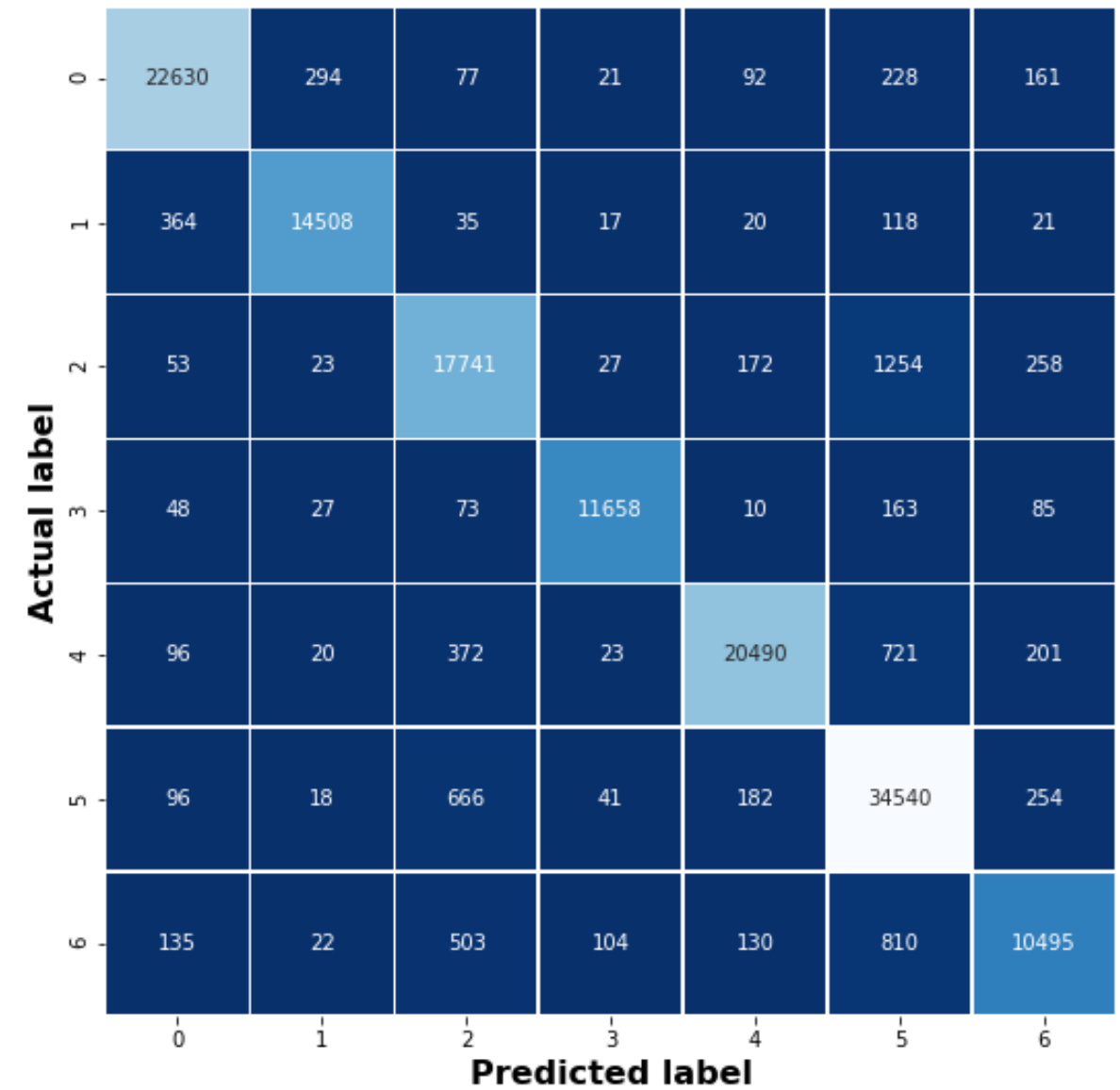
- *Accuracy on Test Data For Function Classification Model: 89%*
- *F1 Score Range for Function Classification Model: 0.81-1.0*
- *Accuracy on Test Data For Location Classification Model: 94%*
- *F1 Score Range for Location Classification Model: 0.89-0.97*



# Confusion Matrices



*Function Prediction*



*Location Prediction*

# CONCLUSIONS

- *Machine Learning can predict both the function & location of a novel protein*
- *ML Models can predict 11 Functions & 7 Cellular Locations*
- *Overlapping protein classes can be tricky to predict*