

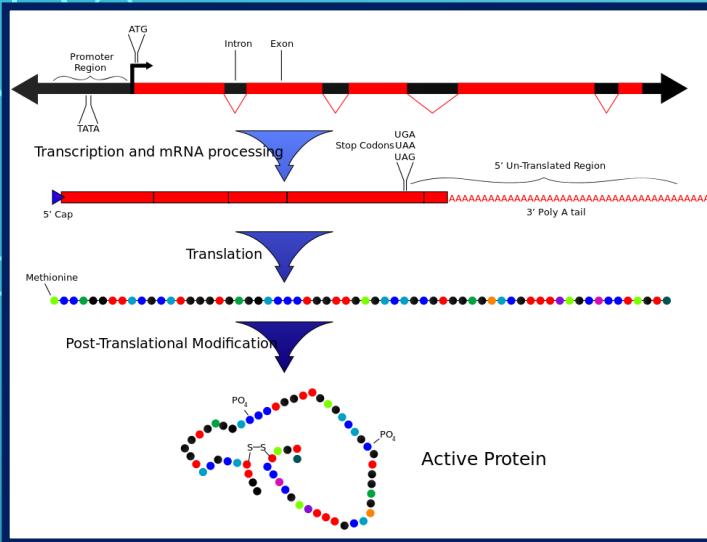
MODULE 5 CAPSTONE

SEETA RAJPARA

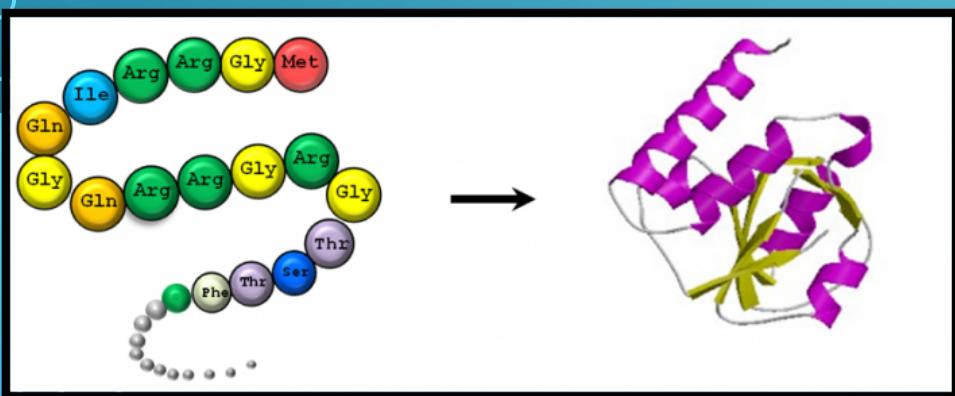
SELF-PACED DATA SCIENCE BOOTCAMP

SEPTEMBER 19TH, 2021

BACKGROUND AND PURPOSE

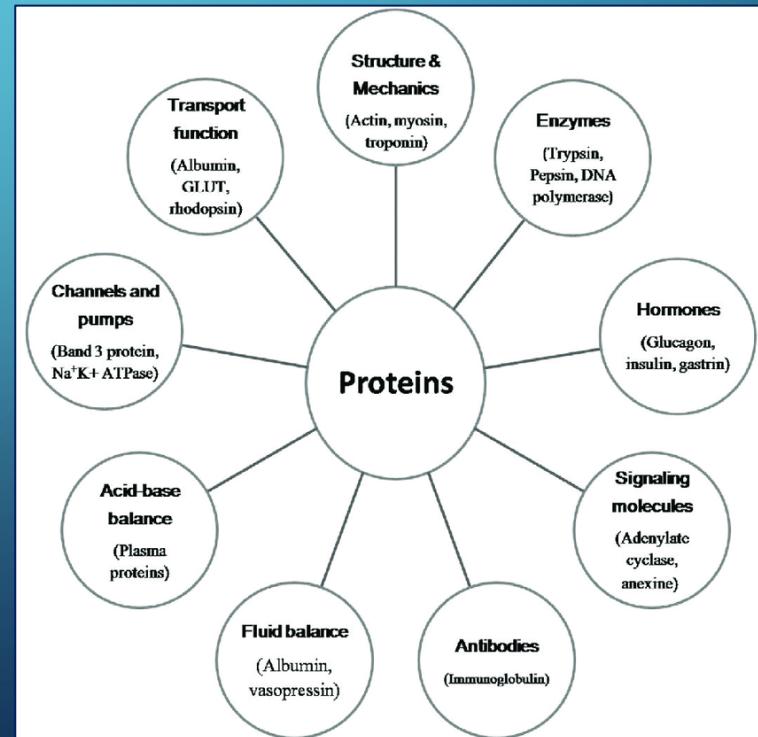


- DNA encodes the instructions for making proteins
- Proteins carry out all the functions of our cells
 - Made up of amino acid chains
 - 20 amino acids make up proteins found in the human body
- The sequence of amino acids influence how the proteins fold, which dictates the function of the protein
- Protein function is a vast area of research in biotechnology, and understanding this further is critical for developing therapeutics and precision diagnostics
- Dataset:
 - We have 2 .csv files
 - Physical properties
 - Amino acid sequences
 - Focusing on only top 20 classifications



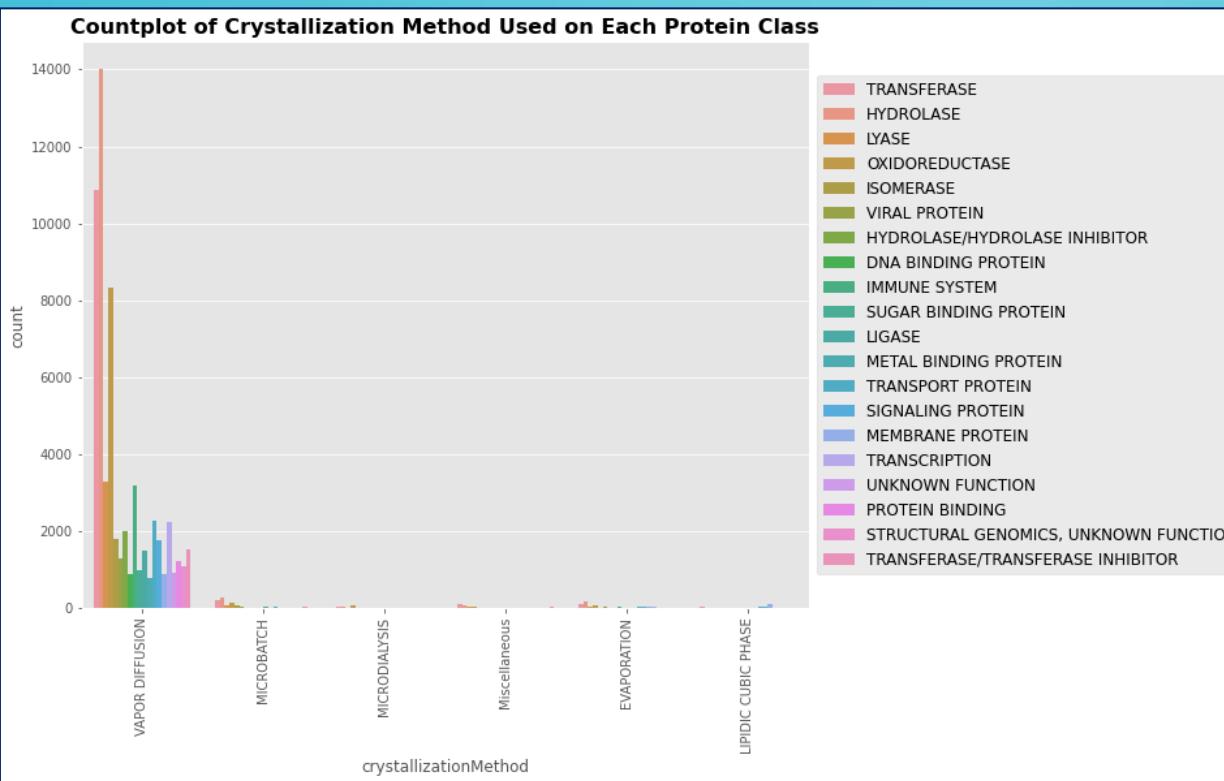
<https://www.ncbi.nlm.nih.gov/books/NBK557845/>

<https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/>



EXPLORATORY ANALYSIS

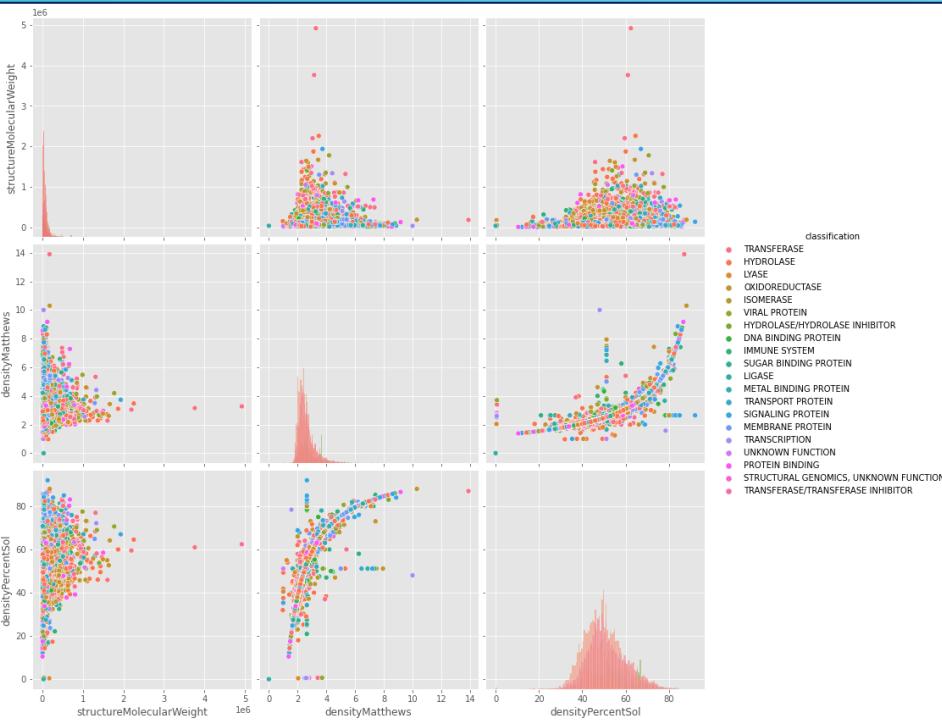
CATEGORICAL VARIABLE



- Various crystallization methods were used for purifying proteins from their aqueous media
- As demonstrated by this count plot, however, methods related to vapor diffusion were primarily used to generate the data in this study
- This won't be critical for us in classifying protein type

EXPLORATORY ANALYSIS

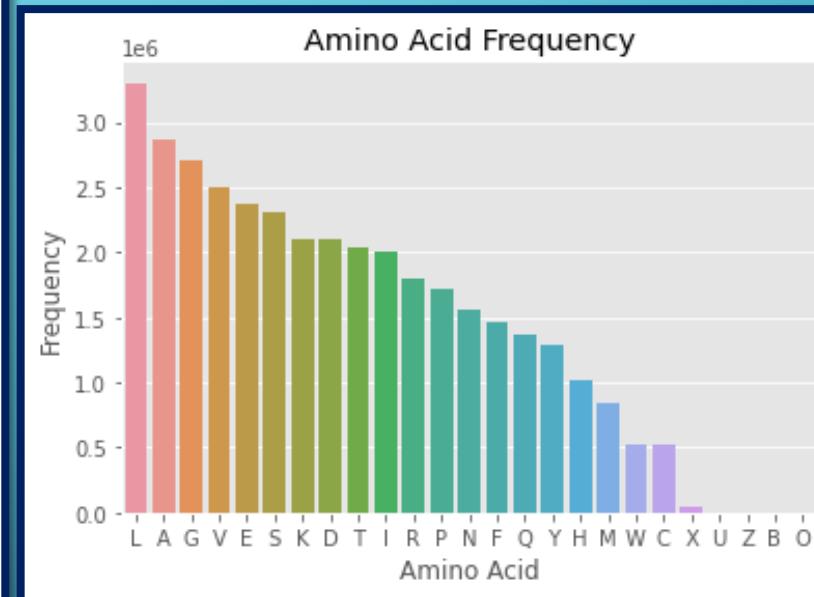
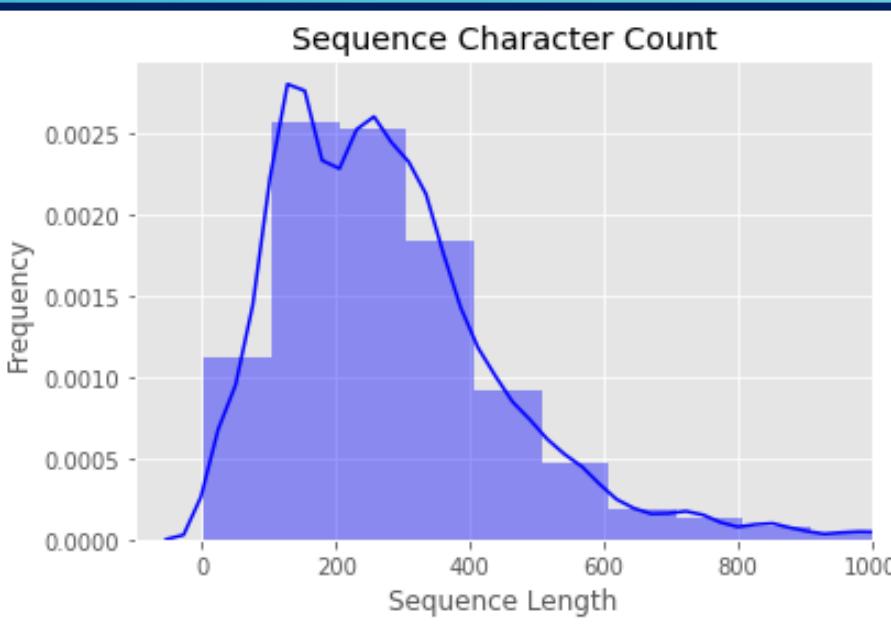
NUMERICAL VARIABLES



- These numerical variables don't seem to exhibit a specific pattern over each of the protein classes
- Can incorporate this in some sort of regression model, but sequences of amino acids will likely be the best predictor of protein classification

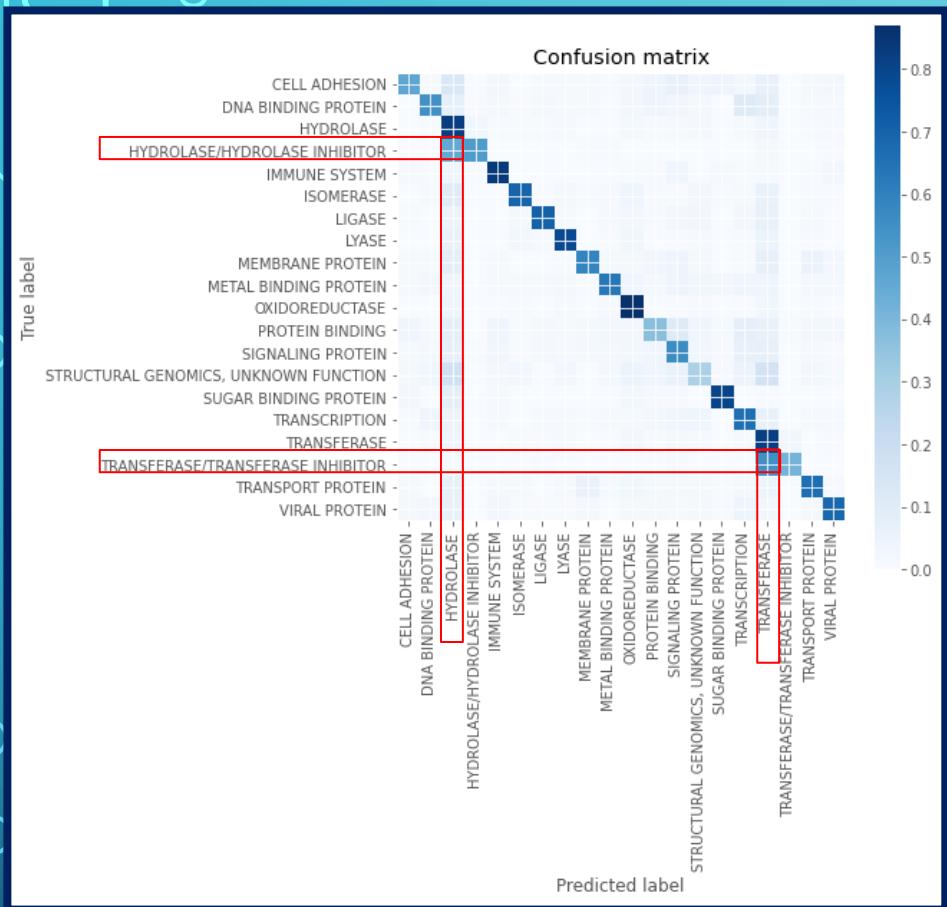
SEQUENCE ANALYSIS

	sequence	classification
4	MVLSEGEWQLVLHVWAKVEADVAGHQQDILIRLFKSHPETLEKFDR...	OXYGEN TRANSPORT
7	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIYGHLLTKSPSLNAAKSE...	HYDROLASE(O-GLYCOSYL)
8	MVLSEGEWQLVLHVWAKVEADVAGHQQDILIRLFKSHPETLEKFDR...	OXYGEN TRANSPORT
11	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIYGHLLTKSPSLNSLDAAK...	HYDROLASE(O-GLYCOSYL)
12	MVLSEGEWQLVLHVWAKVEADVAGHQQDILIRLFKSHPETLEKFDR...	OXYGEN TRANSPORT

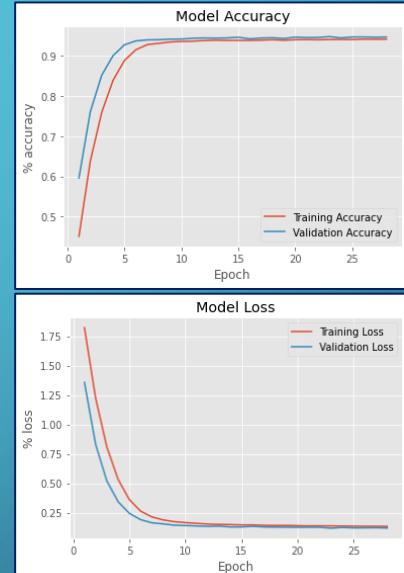


- Average length of each sequence ~285.9 AA's
- Most frequent AA is Leucine (L) which is involved in the biosynthesis of proteins
- A good model for classification using character sequences would be a recurrent neural network (RNN)

MODEL PERFORMANCE



	precision	recall	f1-score	support
CELL ADHESION	0.56	0.47	0.51	259
DNA BINDING PROTEIN	0.56	0.55	0.55	278
HYDROLASE	0.77	0.82	0.80	4025
HYDROLASE/HYDROLASE INHIBITOR	0.64	0.51	0.57	497
IMMUNE SYSTEM	0.86	0.84	0.85	786
ISOMERASE	0.76	0.70	0.73	519
LIGASE	0.75	0.71	0.73	401
LYASE	0.81	0.79	0.80	852
MEMBRANE PROTEIN	0.66	0.59	0.62	324
METAL BINDING PROTEIN	0.66	0.63	0.65	274
OXIDOREDUCTASE	0.88	0.87	0.87	2477
PROTEIN BINDING	0.50	0.37	0.43	408
SIGNALING PROTEIN	0.49	0.56	0.52	538
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.27	0.29	0.28	318
SUGAR BINDING PROTEIN	0.80	0.80	0.80	296
TRANSCRIPTION	0.65	0.66	0.65	711
TRANSFERASE	0.75	0.82	0.79	3015
TRANSFERASE/TRANSFERASE INHIBITOR	0.58	0.41	0.48	332
TRANSPORT PROTEIN	0.79	0.66	0.72	680
VIRAL PROTEIN	0.72	0.69	0.70	488
accuracy			0.74	17478
macro avg	0.67	0.64	0.65	17478
weighted avg	0.74	0.74	0.74	17478



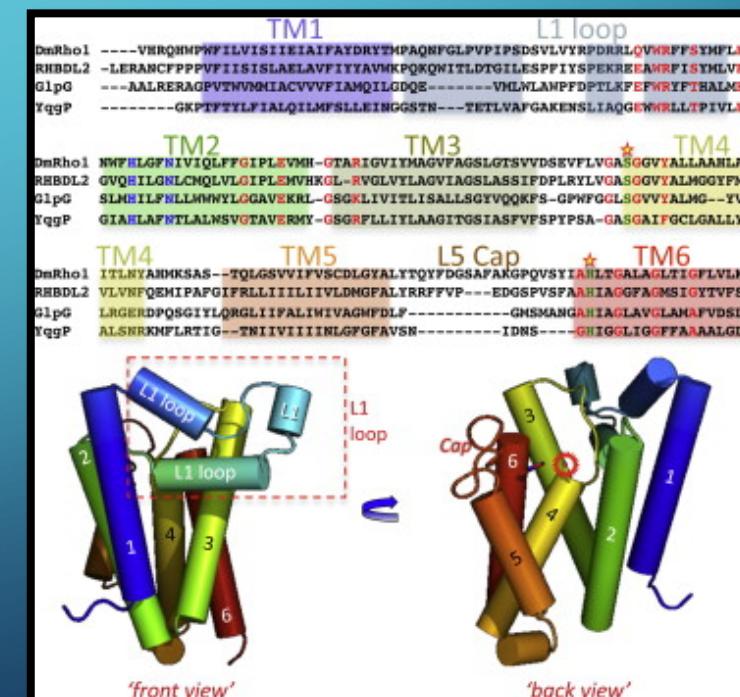
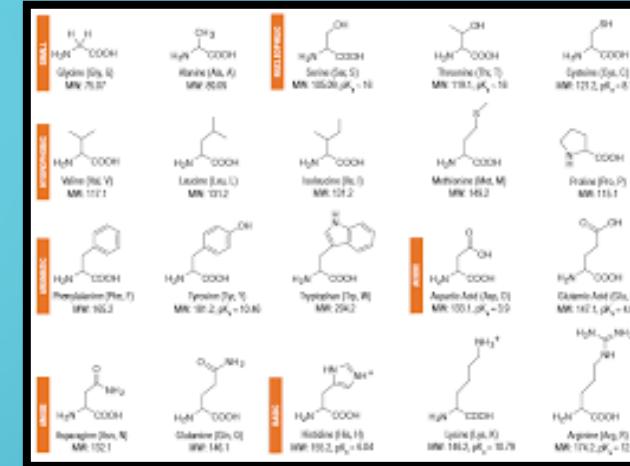
- With a specific type of RNN model (long short term memory, LSTM), we are able to predict protein classification with relatively high precision as plotted above
- This shows that the sequence of amino acids inform protein function to a high degree and this can be used as key to understanding protein function

CONCLUSIONS AND FUTURE STEPS

- 1) Protein classification is influenced by amino acid sequence
- 2) Many physical properties are shared by proteins of all classes, but granular data can be extracted from these features
- 3) Experimental techniques applied to each protein classification are often the same, but building more general classes for each protein type specified in the dataset could help mitigate the issues with data notation

Next Steps

- Try focusing only on the top 10 protein classifications instead of top 20, might get better results with fewer labels
- Include numerical variables into a mixed model, somehow incorporating regression and RNN models to predict with high precision





THANK YOU!

APPENDIX

