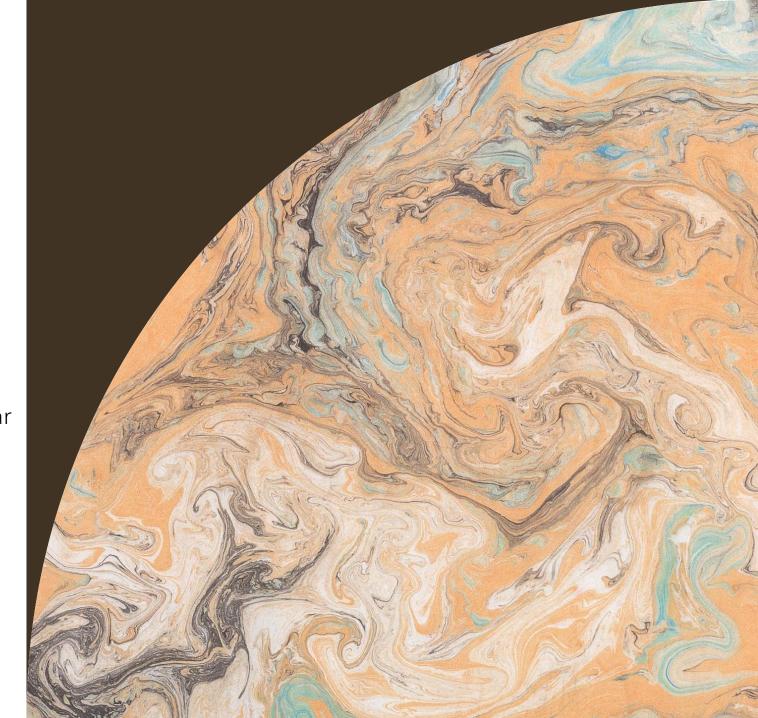
Application of Amino Acid Property Grouping Technique to Predict Protein Function Using Natural Language Processing Algorithms

By: Ritwik Katiyar



# **Purpose**

- Testing Amino Acid property
  grouping technique against some of
  the more common techniques and
  methodologies applied in the field of
  protein as a language.
- To test the viability of this technique and if it shows potential for future applications and testing.

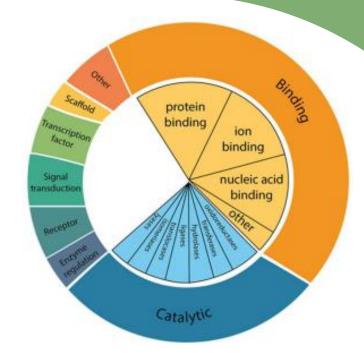


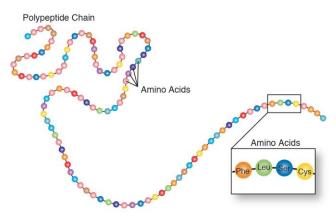


Introduction

### **Proteins and Functions**

- Life's fundamental building blocks consist of sequences of amino acids at the molecular level.
- Proteins exhibit a wide array of functions ranging from facilitating biochemical reactions to providing essential structural support.
- The functionality of proteins denotes their distinct biological roles within an organism.
- Further granularity exists within protein functions, as they can be subdivided into specific subfunctions.
- Proteins often serve multiple functions, showcasing their adaptability and versatility in biological processes.





### Amino Acids

Ala: Alanine Arg: Arginine Asn: Asparagine Asp:Aspartic acid Cys:Cysteine

Gln: Glutamine Glu: Glutamic acid Gly: Glycine

Lys: Lysine lle: Isoleucine Pro: Proline

Leu: Leucine

Ser: Serine Thr: Threonine Trp: Tryptophane Tyr: Tyrosisne

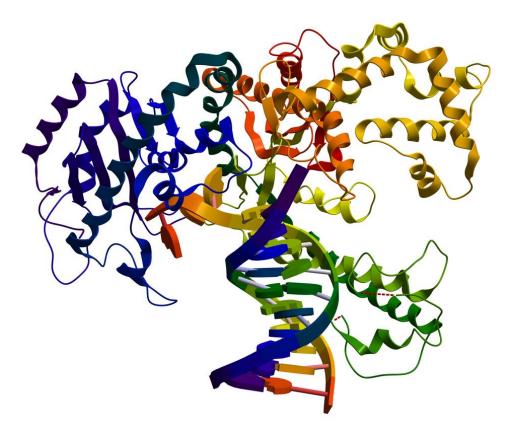
### Various Methods of Protein Function Determination

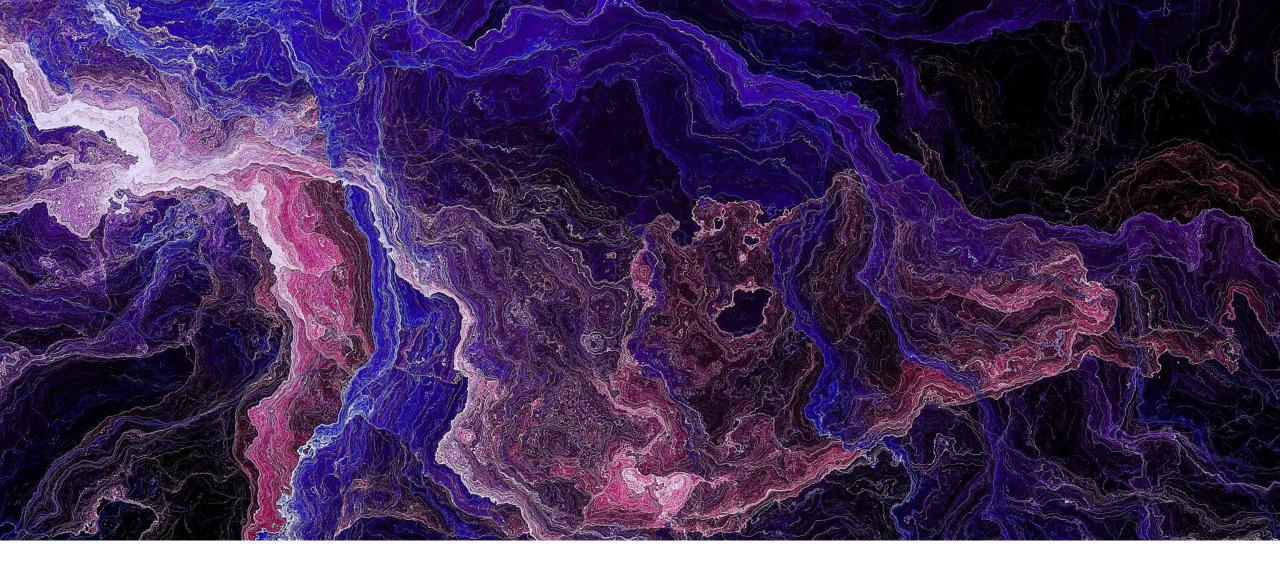
Various Approaches that can be used to obtain the function of the protein via the sequence. Some of the more popular techniques are:

- Conventional Methods via the use of experimental assays.
- Homology or Similarity Based Approaches (BLAST).
- Feature Embedding or Information Based.
- As a Language Approach.

# **Importance of Discovering Protein Function**

- Facilitate and Streamline Drug Discovery processes.
- Enhance the understanding of biological systems
- Design of more targeted and effective therapeutic interventions.
- Protein engineering
- Bio-Manufacturing

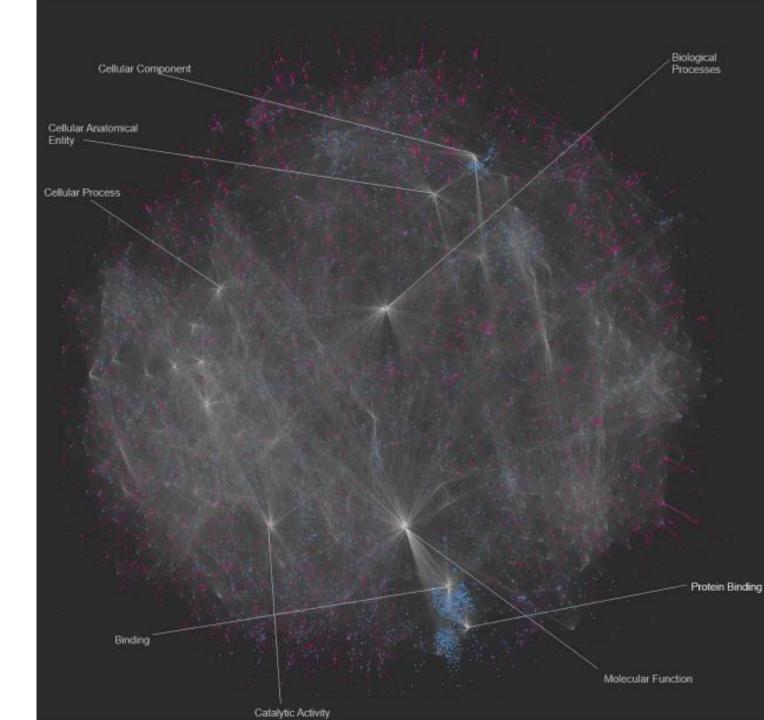




Background

# Complexity of the Relation Between Proteins and Functions.

The Bi-Partite Network showcases the complex relation between proteins (blue) and their respective functions (red) within our sample data of just 3131 proteins sequences.

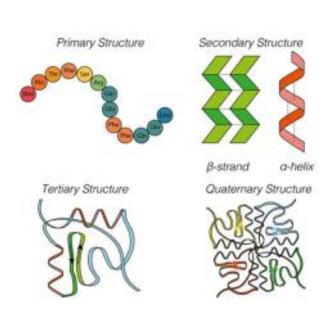


# Feature Embedding - Common Features

• Early applications employing the Term Frequency - Inverse Document Frequency to obtain the number of times an amino acid appears in a sequence [3]

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

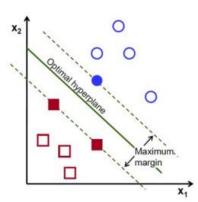
- Predicted secondary Structures (Elongated coils, Helixes, and Beta sheets)
- Tertiary structures (Domains and Motifs)
- pKa (Acidity of the amino acid)
- Hydrophobicity (Repealed by water molecules)



# **Support Vector Machine (SVM)**

- SVM is a supervised machine learning algorithm that is used for binary classification.
- SVM attempts to classify the data by finding the best possible hyperplane that classifies the data points.
- A standard SVM is designed to work best with liner data
- Polynomial Kernel traditionally used for DNA and Protein applications.
- The polynomial kernel uses a polynomial function to redefine the data into a higher dimension.

$$K(x_1, x_2) = (x_1^T x_2 + c)^d$$



### **Decision Tree & Random Forest**

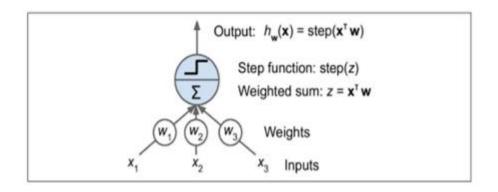
- Decision trees are non-parametric supervised learning methods that are commonly used for classification and regression.
- Most decision trees use the Gini index to determine the nodes it needs to select.
- The Gini score is a measure of the impurity of the node
- A node is considered pure if all training instances it applies to belong to the same class.

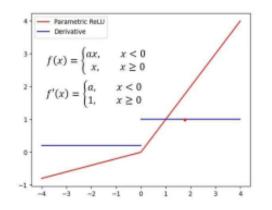
$$G_i = 1 - \sum_{k=1}^n P_{i,k}^2$$

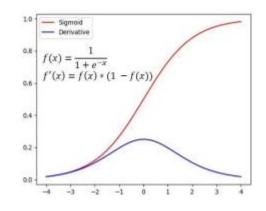
 Random Forest is largely considered over decision trees, as a random forest algorithm is simply a collection or an ensemble of decision trees

# **Deep Neural Network (DNN)**

- A neural network aims to mimic the human brain and how the neurons interact with one another
- Like neurons the neural networks are made up of perceptron's. Invented in 1957 by Frank Rosenblatt, perception uses a weighted sum to generate output from an input.
- The ReLU activation function primarily assists with the performance of the neural network by replacing negative values with 0 and only keeping only the positive value.
- The sigmoidal function restricts the values between 0 and 1. Essentially generating a probability.

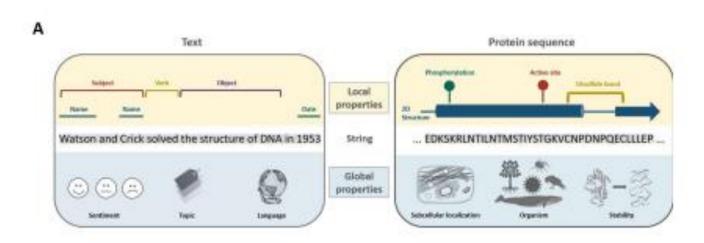






### A Shift From Feature Embedding

- Improved Computational Algorithms Resulting in the use of language within the need for embedding.
- Tokenization or Bag of Words technique.
- As a language approach.

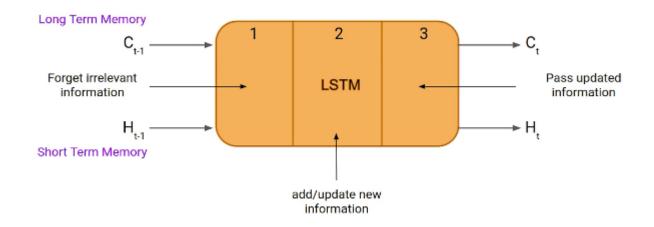


 String:
 The cat sat on the mat
 MSTIYSTGKVCNP...

 Possible tokenizations:
 [\*start\*] [T] [h] [e] [\*space\*] [c] [a] [t] ... [\*start\*] [M] [S] [T] [Y] [S] [T] [G] ... [\*start\*] [MS] [TI] [YS] [TG] ... [\*start\*] [MS] [TG] ... [\*start\*] [MS] [TG] [TG] ... [\*start\*] [MS] [TG] ... [\*start\*] [MS]

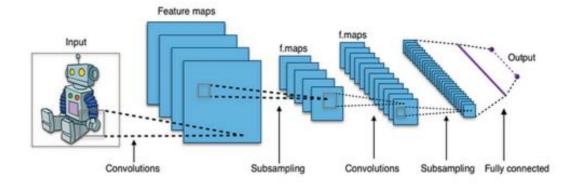
## **Recurrent Neural Network (RNN)**

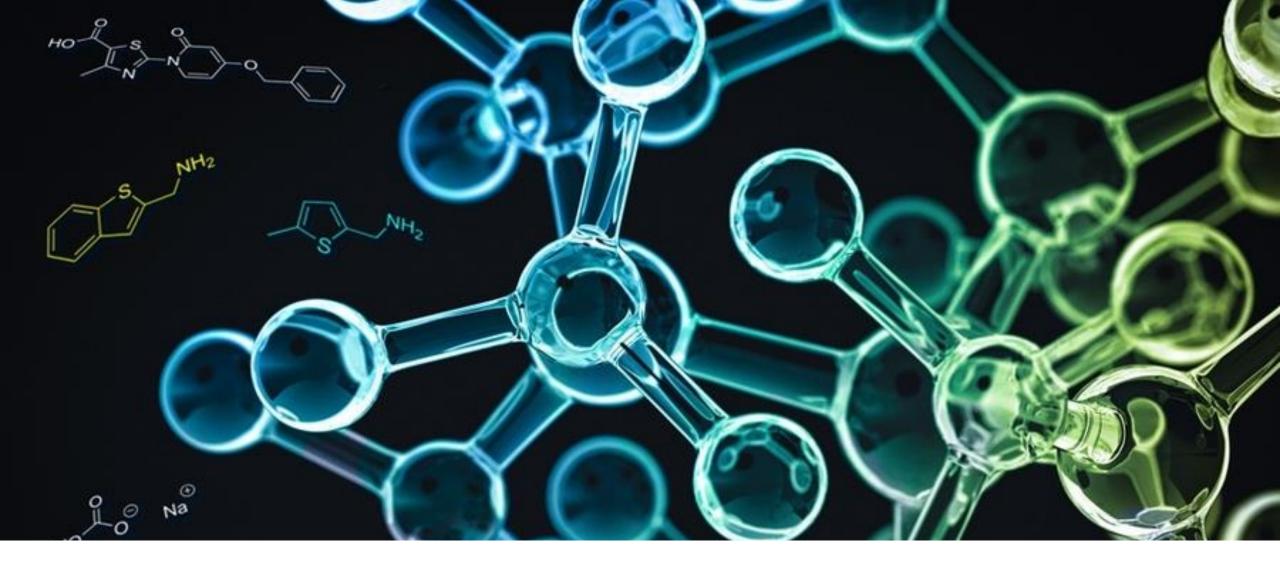
- RNN's can account for the sequential nature of language and process language while accounting for context.
- Fundamentally, an RNN comprises of Long Short-Term Memory (LSTM) layers.
- At each iteration, the LSTM meticulously evaluates the current word, the carry, and the cell state
- ProLanGo [4]



# Convolutional Neural Network (CNN)

- The application of CNNs to textual data becomes feasible through tokenization and the conversion of language into two-dimensional numerical arrays.
- CNN architectures encompass convolutional layers comprising specialized perceptrons, which can be either pooled or fully connected
- ProtConv [9]





Data Background, Exploratory Data Analysis, and Methodology

# **Data Background**

- The Gene Ontology knowledgebase provides a computational representation of our current scientific knowledge about the functions of genes.
- The GO ontology: The logical structure describing the full complexity of the biological functions of various kinds.
- Includes experimental findings from over 150,000 published papers and 700,000 experimentally supported annotations. [5]
- Join over 16M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies
- The data itself was obtained from Kaggle referred to as CAFA-5 [2]

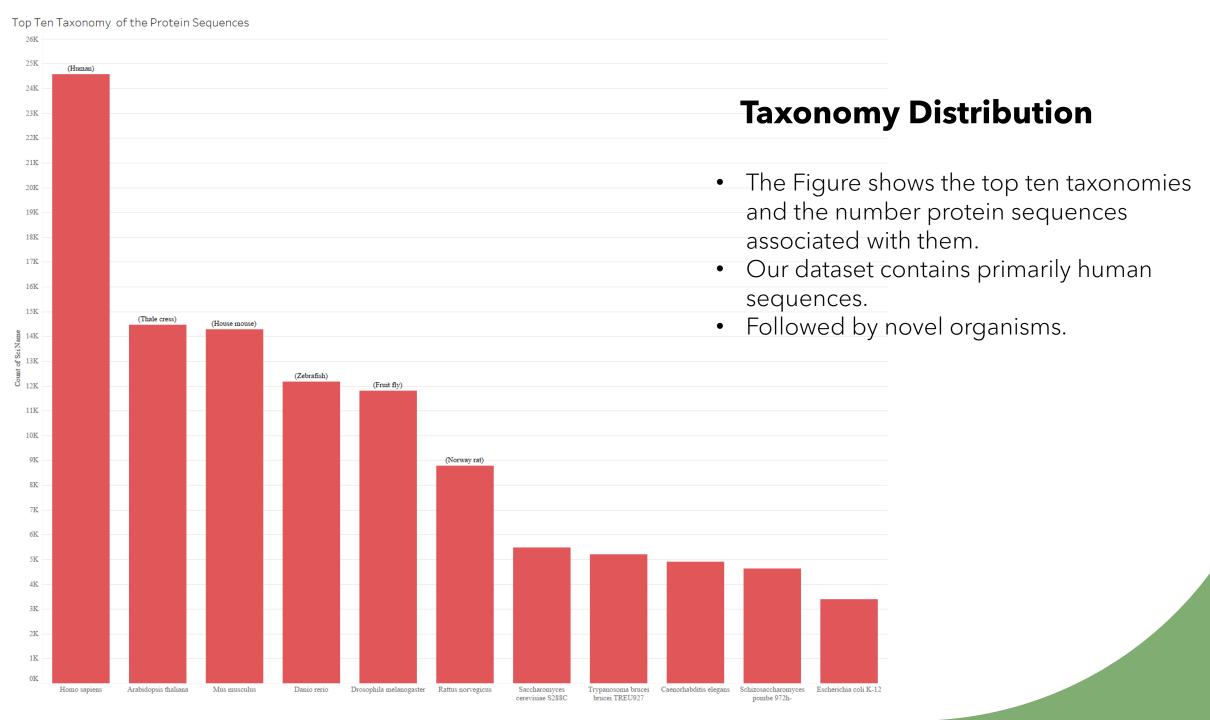
# Kaggle



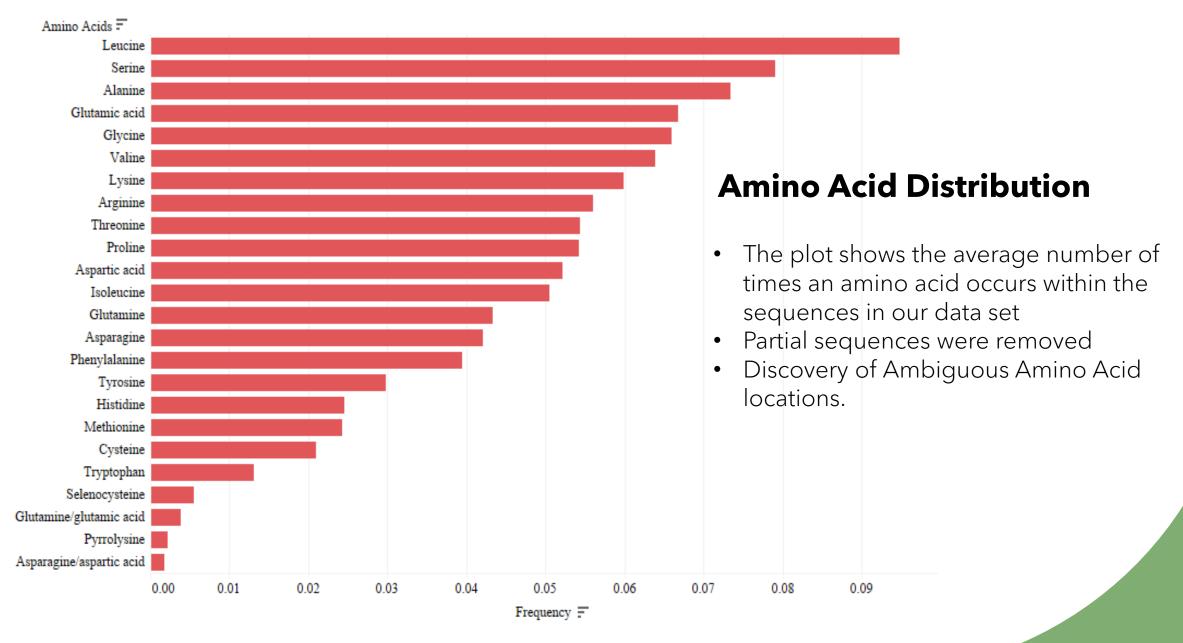
### Taxonomy of the Protein Sequences Mus musculus (House mouse) Homo sapiens (Human) Saccharomyces cerevisiae S288C 24.582 5.469 Escherichia coli K-12 3,402 Gallus gallus (Chicken) (Cattle) SC5314 Dictyostelium Aspergillus (Zebrafish) 12,181 Trypanosoma brucei brucei TREU927 Arabidopsis thaliana (Thale cress) 14,457 Rattus norvegicus (Fruit fly) 11,809 (Norway rat) Orvza sativa pombe 972h

# Represented Taxonomies

- 140674 Sequences
- Average Sequence Length 553.56
- Max Sequence Length 35375 (This complex protein is from the Norway rat. The protein seems to be responsible for the deployment of muscle)
- Min Sequence Length 3 (The protein sequence is from a type of a mushroom. The protein is generally responsible for regulation and inhibition)
- 3131 protein sequences obtained from plants, animals, and micro-organisms (bacteria and viruses)
- Removed Partial Sequences



### Average Frequency of Amino Acids Within Sequences

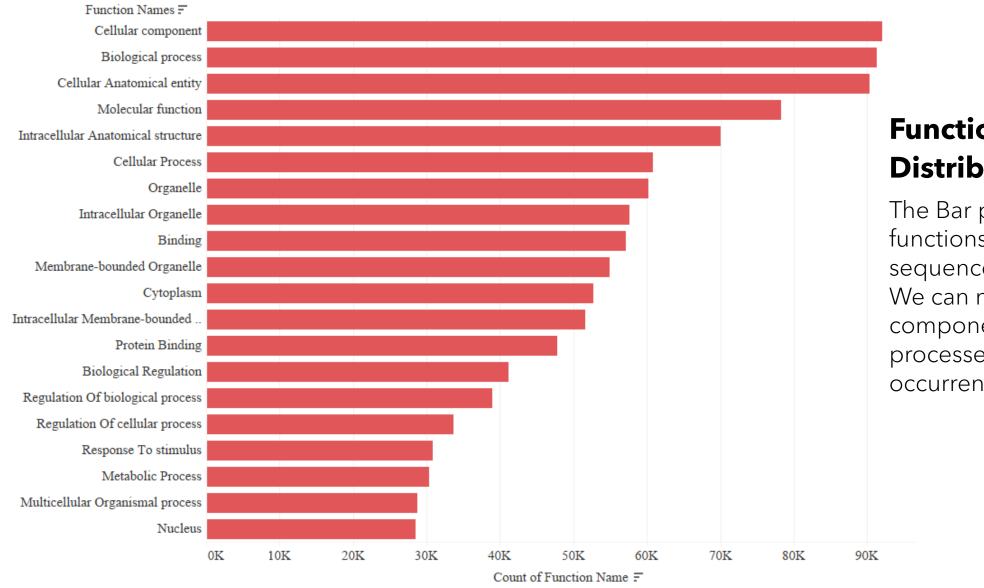


Function Distribution																
Cellular component	Binding	Nucleus														
		Developmental														
		Developmental Process		-	$\dashv$		++	+++	++	+++	+++		$\dashv \vdash$		+++	
	Membrane-bounded Organelle					<u> </u>	$\bot \bot$		$\perp \downarrow \downarrow$						$\perp \downarrow \downarrow \downarrow$	
Biological process		Organic Substance		-												
		metabolic process														
	Cytoplasm	Anatomical													ЩЦ	
		Structure development						<del>                                     </del>	$\Box$		++++	++++		++++	<del>                                     </del>	
Cellular Anatomical entity																
	T. 11.1	Membrane		-												
	Intracellular Membrane-bounded organelle			-												
	organization of the state of th	Catalytic Activity														
										<u>-                                     </u>						
Molecular function	Protein Binding	Cellular Metabolic process									<u> </u>					
					_											
		Primary Metabolic process														
Intracellular Anatomical	Biological Regulation					_										
Intracellular Anatomical structure		Nitrogen Compound														
	Regulation Of biological	metabolic process	-													
	process	Multicellular Organism														
Cellular Process		development Coll Parinham														
	Regulation Of cellular process	Cell Periphery														
		Cellular														
Organelle	Response To stimulus	Component														
		Regulation Of metabolic process														
	Metabolic Process	Cellular														
Intracellular Organelle		Component														3383333311
	Multicellular Organismal process	Positive Regulation of														

# Function Distribution

- This visualization illustrates the vast magnitude of various functions within our data set.
- There are over all a 43,248 Different Functions

### **Top Twenty Functions**



### **Function Distribution**

The Bar plot shows the top 20 functions that the protein sequences are associated with. We can note that Cellular components and Biological processes have the highest occurrences.

# Sampling

- SMOTE Technique or Synthetic Minority Oversampling Technique: Such a single sequence form every single taxon is represented even taxon that may have only one sequence.
- SMOTE was utilized since it provides a diverse number of proteins with unique functions. However, this could influence the accuracy of our model.
- SMOTE was chosen specifically to have a high variation in our sample data. A high variation.
- Sampled 3131 Sequences. One for each Taxonomy.

### Function Distribution - Sample

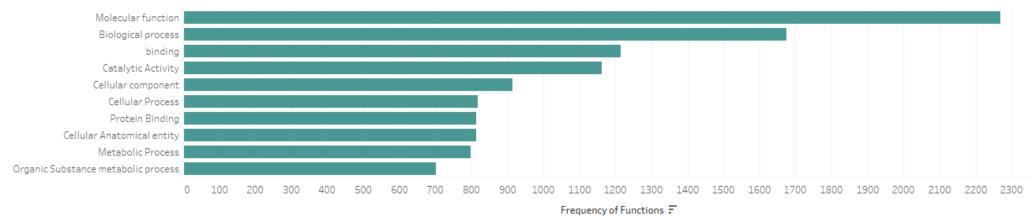
Molecular function	Protein Binding	Extracellular Region	external biotic stimulus	Heterocycle Metabolic process  Metal Ion
		Nitrogen Compound metabolic process	Response To other organism	binding
	Metabolic Process		Response To stress	Viral Process Viral Process
		Organic Substance biosynthetic process	SUCSS	Intracellular
			Macromolecule Metabolic process	Organelle
Biological process		Cellular Biosynthetic process	Cellular	Defense Response
	Organic Substance metabolic process		Aromatic compound	Cellular Lipid metabolic
		Oxidoreductase Activity	Organic Cyclic compound binding	process  Regulation Of
			Heterocyclic	metabolic process
	Cellular Metabolic process	Regulation Of cellular process	Compound binding	Cellular Nitrogen compound
binding		Transferase Activity	Catabolic Process	
	Primary Metabolic process	Identical Protein binding	Biological Process involved	positive
			in symbiotic	Regulation of
Catalytic Activity	Biological Process involved in interspecies interaction between organisms	Organic Cyclic compound metabolic process	Regulation of biological	Regulation Of Control
			organelle	Defense Response to
	Biological Regulation	Ion Binding	Biological Process involved	Lyase Activity
Cellular component	Response To stimulus	Organonitrogen Compound metabolic process	in interaction	Intracellular
			process	
	Hydrolase Activity	Response To external stimulus	Modulation Of process of	Lipid Biosynthetic
Cellular Process	nyarorase Activity	Intracellular Anatomical	Cation Binding	Regulation Of nitrogen
		structure		

# Sample Function Distribution

- A similar
  Representation of the magnitude of functions. However, this visualization uses our sample data.
- We can note that our sample has a lot less functions than our original data set
- 5206 Different Functions

# **Sample Functions - Top Ten**

### Function Distribution - Sample

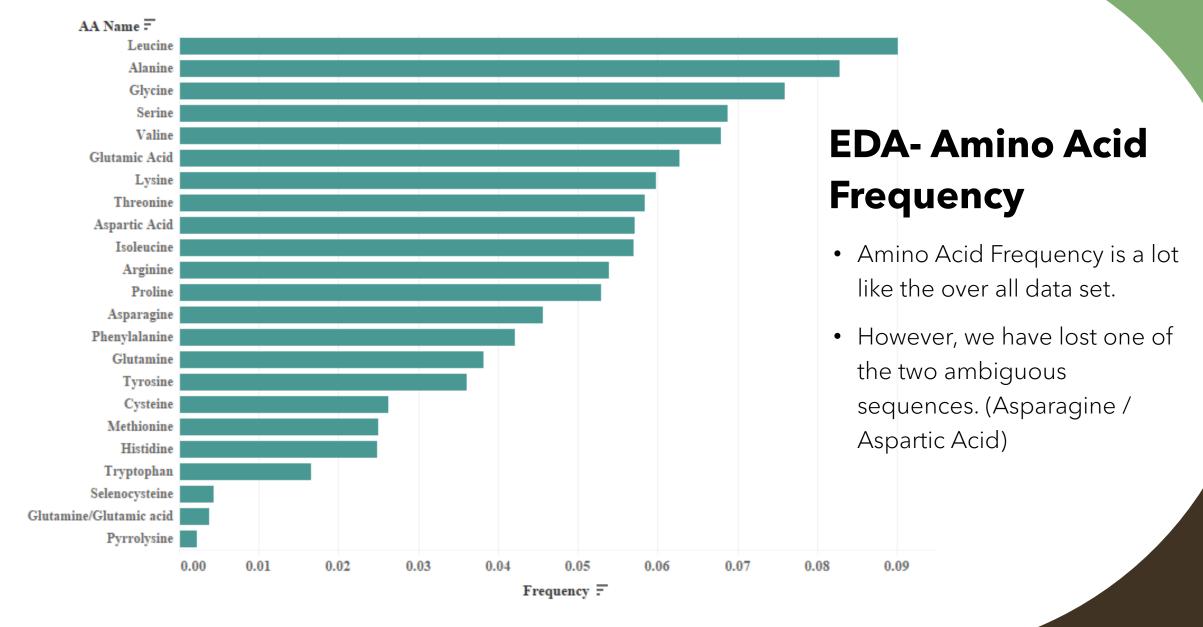


- The minimum sequence length within this dataset remains three (pertaining to sequences from the same fungal species discussed earlier).
- The maximum sequence length has extended to 6,872 amino acids.
- The average sequence length stands at 474 amino acids.
- The plot shows the top 10 functions within the now sampled data. We can note that now cellular component is no longer the most dominant function
- The Distribution has changed. However, the goal is to test accuracy of our model. These changes shouldn't be a cause for concern.
- However, we can note that our data is highly imbalanced.

# **Functions Explained**

- Molecular Function Molecular level activities performed by gene products such as catalysis or transportation as opposed to non-gene product activities.
- Biological Process A proteins involvement in a larger multimolecular activity
- Binding & catalytic activity Border terms used to describe proteins that perform said functions
- Cellular Component The protein belongs to a specific or a larger cellular component.
- Cellular Anatomical entities Implies gene product proteins involved in cellular structures.
- Protein binding & Metabolic Processes Proteins which bind with other proteins or are involved with the metabolic pathways
- Organic Substance Metabolic Processes Proteins that are involved in chemical reactions and pathways which require organic acids

### Sample - Mean Amino Acid Frequency



# **Sample - Secondary Structure Prediction**

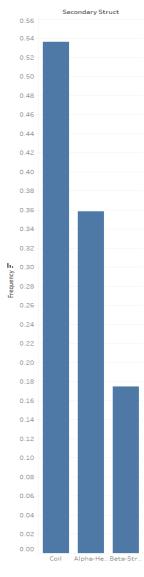
- S4pred is a prediction tool that uses the sequence to predict secondary structure. [8]
- The tool uses a deep semisupervised learning framework to obtain its results
- The tool boasts an about 75%
   accuracy on predictions. And is
   regarded as one of the top
   secondary structure prediction tools.
- The tool is also open source can be downloaded from git-hub

sec_seq_id	sec_seq
O87540	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
B8ZV93	СССССЕЕССССССССНИНИНИНИНИНИНИНИНИНИНИНИН
Q1JUQ0	ССССССССССССССССССССССССССССНИНИНИНИНННИННСССЕЕЕЕСССССНССССНИНИНИНИ
P17452	ссссссссссссссснинининннесссссссснинининнннн
P19508	ссссссссссссснинининининининининининини
P31621	ССССССССССССНИННИННССССССССССССНИННИННСССССС
P10271	сссссснининининининсссссснининининин
Q8UN00	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
P07567	сссссснининининнинссссссснининининн
P17863	ССССССССССССССССССССССССССССССССССССССС
P25020	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
P03354	сснинининининссссссинининининин
P25058	ссссссссссссснинининннесссссссснинининнннесссссссс
O39198	CCCEEEECCCCCCCEECCCCCEEEEECCCCCEEEECCCCC

# **EDA - Secondary Structure**

- alpha-helix, which is a regular state denoted by an 'H'.
- beta-strand, which is a regular state denoted by an 'E'.
- coil region, which is an irregular state denoted by a 'C'.

### Mean Secondary Strucutre Rate



# Sample - Tertiary Structure

- Tertiary structure information was obtained from Conserved domains database (CDD)
   [10]
- CDD uses RPS-BLAST a variation on the PSI-BLAST algorithm to find regions of similarity and obtain hits.
- Only the concise hits with values higher than 0.0001 E-values were selected. To Guarantee the best results
- Due to high traffic the limit was set to 1000 sequence per run.
- Average run time was around thirty minutes to an hour.



### Conserved Domains and Protein Classification

OVERVIEW

### Conserved Domain Database (CDD)

Resources

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAMs).

Search How To Help News FTP Publications

### CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as specific hits. The CD-Search Help provides additional details, including information about running CD-Search locally.

Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 4,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

CD-Search (Help & FTP) Batch CD-Search (Help) Publications

### SPARCLE: Protein Classification

The Subfamily Protein Architecture Labeling Engine (SPARCLE) is a resource for the functional characterization and labeling of protein sequences that have been grouped by their characteristic domain architecture. To use SPARCLE, you can either: (1) enter a query protein sequence into CD-Search, which will display a "Protein Classification" on the results page if the query protein has a hit to a curated domain architecture in the SPARCLE database (example, using NP\_337887 as the query sequence), or (2) search the SPARCLE database by keyword to retrieve domain architectures that contain the term(s) of interest in their descriptions (example, searching for the words "chloride" and "channel" in the domain architecture record, and limiting the results to curated architectures). With either approach, the corresponding SPARCLE record(s) will display the name and functional label of the architecture, supporting evidence, and links to other proteins with the same architecture.

About Help Input protein sequence | Search by text word | FTP

### CDART: Domain Architectures

Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive domain profiles rather than direct sequence similarity. Proteins similar to the query are grouped and scored by architecture. You can search CDART directly with a query protein sequence, or, if a sequence of interest is already in the Entrez Protein database, simply retrieve the record, open its "Links" menu, and select "Domain Relatives" to see the precalculated CDART results (illustrated example). Relying on domain profiles allows CDART to be fast and, because it relies on annotated functional domains, informative.

About Search Help FTP Publications

### CDTree

CDTree is a helper application for your web browser that allows you to interactively view and examine conserved domain hierarchies curated at NCBI. CDTree works with Cn3D as its alignment viewer/editor, it is used in the CDD curation process and is a both classification and research tool for functional annotation and the study of protein and protein domain families.

About Install Publications

# **Tertiary Structure - Data**

- We can observe various motifs and domains discovered within a protein sequence.
- A motif refers to a conserved pattern associated with distinct functions.
- Domain is a 3-D structure, Hence a larger pattern associated with a function.
- A specific Motif can occur multiple times within a sequence.

```
,Query,Accession,Short name
0,087540,TIGR03507,decahem S01788
1,B8ZV93,pfam00221,Lyase aromatic
2,052237,COG0642,BaeS
3,052237,pfam08448,PAS 4
4,052237,c134674,RocR superfamily
5,Q0VZ68,pfam00221,Lyase aromatic
6,Q9KIZ4,cd20630,P450 epoK-like
7,050078,PRK06498,PRK06498
8,Q66YX3,NF033729,borfam54 2
9,B8YNY4,NF033596,denti PrcB
10,C0J1Q3,pfam02368,Big 2
11,C0J1Q3,smart00635,BID 2
12,C0J1Q3,pfam02368,Big 2
13,C0J1Q3,smart00635,BID_2
14,C0J1Q3,pfam02368,Big 2
15,C0J1Q3,smart00635,BID 2
16,C0J1Q3,smart00635,BID
17,C0J1Q3,c102708,Big ~
18,C0J1Q3,c102708
```

tertiary struct.csv

# Single letter code for Amino Acids

Amino acid	1 letter code	3 letter code
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic acid	D	Asp
Cysteine	C	Cys
Glutamic acid	E	Glu
Glutamine	Q	Gln
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val
Selenomethionine*	X	SeM

### id, amino prop

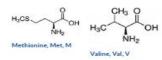
087540, AAPAPBPBAAAAAAAAAAAAPAAPAAPUPUUPPUPPUPCUPCUUCUAUPAPPAP AUCUAPUUUPPAPRPUAUPPBPRACPBPUPRPRBRCARCPPBARPAPAPPBRPAAPA BACUCBARBPAPCACPPAPPBPPCRACUBUBUBAARPBAABPABPAPBARUBCPBAU CACRAAUBUBPPPACPPPPAAPBPPCRPACABPABPPAPBPAAPPRUACPPPPAPPC PAURPUBCPARAABPUAACUBAPAPCAUBAARPPPBCABAUPPUACPCPARPRAURP PPBBACPAPAPAPAPPBPACRCABBUBPBAURAAPCPAPBPPCAPUBAAAUACAPAP CRPACPRBBBUAAPPAAAACBPUARPPUAPAPPPPBPAUPPRAABPBCPACPRUAA B8ZV93, ACARAAAAUBAUACACAAPACBABAPBABAPURUACBRUARUAPPURUCA APPAABURPUAPUCAAPAAAPAAPBUABUAAUAPUPAUAPUCAAUAPBAAAUAAUCU UAAPAAARBABBAABAPAAAPACAAPAAPAPPBURPBPUPAABPBUUPAAAABAABC BPAUPAAUAAACPAPAPPBRACCCAPPAPCPUAAACPUACPRBUAPRBUPRAAAAPC PURCUUPRAAPPAAPCPACAAAUPPABPAUPPUPPPCAAPAUAAAABBPAAAPCPAU CCBUAPPAAAPABUAAAPCPUAAAAPPAAPACAPUCAPACUBA

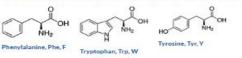
052237, AABPUPAUUBACAPPUBUAPBPPUPCUCCUAPCAAAAABPABPAAUUAAC PABAAPBBAUCAABAPBPPUAUPARAAABAPUPCCUAAARACCAPCBAAACCAABBC AAUBPCACAAUACBAAUPAABCBCAAPPUBPAPRCUACUUUUPCBARBPPBUAABBU AUPUAAAABAPPAAAAABCPAUUCABBAAABACUPACBAABAABPAACRPBABAAUU RCCCBAUPAAPPAAUPAAPBPUCUPPABABAAABCPARPBACABPUUUUAUCPABUB AAPPACPUPPRPAPAUBAPBUAUUUUUBPA

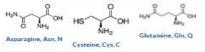
Q0VZ68, ABAPUPPAPARCAACAPABBAPACACUPPACBAAAABCBPPARUCAPBUA CAABAAAAABAPPAABURPUAPACPABAAACRAPBUABUAAUPPUPAUAPUCAPUAP PAAPUPPAAPUAAPAAAUBARBARBAAAAAPACRAPPAUUPPUURCCBUBAUBPBPU UARAPCARPABAAUPAAUUAACPACRABBAACCCAPPPPCPUAARCAUCPPRBUAPR RAAPABPUAAPURCUUPRAAPPAAPCPACAAAUPPABPAUPPUPPPCAAPAUPPPAB CAPBPAUARPCCPUARCARPPABARAUUCURBABAAPBACAAAPPAPP

Q9KIZ4, APPCPAPPPCPBUARCRBURAUURACCURUAACBABCAPUARRRCCUBPR BRUARUAUUCCBABABBAAPUPRPPBAACAABACAPBPACPAACABPUPCCRCAABC BACCCPBPAAAPAPCUAAAABUAACCBBBPUACPCAAPAAAPACACUPBAPPBCAAA BPAACCAABRCPAABAUPABRABPCACRPUAPABBUCAARAAAUPAABCUPARPBUC RUCABABCPUARURBUARBPACPAPAAABUPBAU

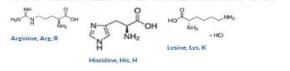
### AMINO ACIDS WITH HYDROPHOBIC SIDE CHAIN - ALIPHATIC

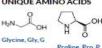






### AMINO ACIDS WITH ELECTRICALLY CHARGED SIDE CHAINS - ACIDIC



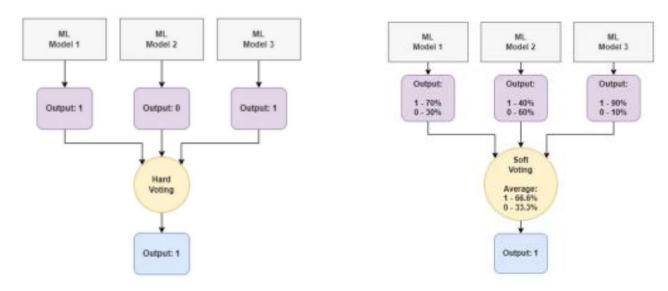


# **Amino Acid -Properties**

- Changed the amino acid sequence and grouped amino acids based on their properties
- Sigma Aldrich. [1]
- (A,I,L,M,V) → Aliphatic Amino Acids (A)
- (F,W,Y) → Aromatic (R)
- (N,C,Q,S,T) → Polar Neutral (P)
- (D,E) → Acidic (C)
- (R,H,K) → Basic (B)
- (G,P) → Unique (U)
- (O,U,Z) → Rare (S)

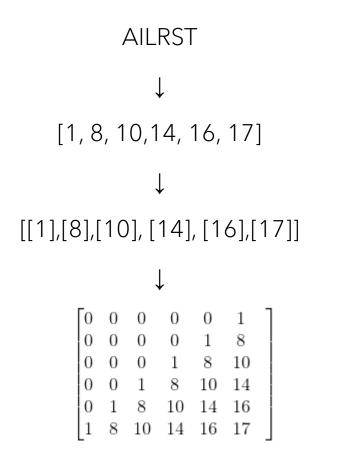
# **Ensemble Techniques**

- Hard and Soft Voting
- In hard voting, a majority voting mechanism among multiple algorithms is employed to determine the final predictions.
- Soft voting leverages the probability estimates from each algorithm to formulate its predictions.



### Standard RNN Language Encoder

### Amino Acid Grouping RNN Encoder



# Singular Value Decomposition (SVD)

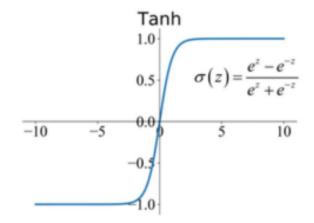
- Dimensionality reduction algorithm
- Condense the sparse array to a size of 28x28.
- The SVD is a matrix can be interpreted as the factorization of a matrix of three other matrices.

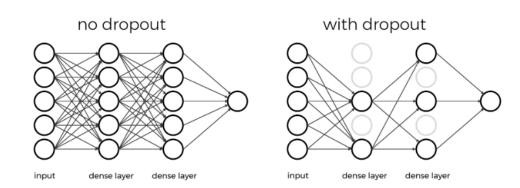
$$A = U \sum V^T$$

- U is the mxm matrix of the orthonormal eigenvectors of AA<sup>T</sup>
- $V^T$  is the transpose of nxn matrix containing the orthonormal eigenvectors of  $A^TA$
- $\Sigma$  is the diagonal matrix with r elements equal to the root of the positive eigenvalues  $AA^T$  or  $A^TA$

# **RNN Application**

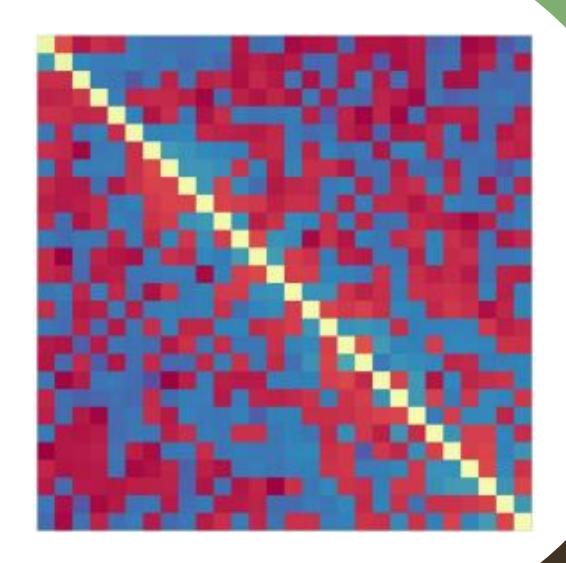
- RNNs are computationally intensive, rendering them more efficient when executed on a Graphics Processing Unit.
- Sequences exceeding 500 amino acids, truncation to the first 500 amino acids was necessary due to computational intensity
- For GPU execution, the activation function is modified from Rectified Linear Unit (ReLU) to Hyperbolic Tangent (Tanh)
- Drop Out layers





### **CNN - QR Code of Life**

- A heatmap represents the matrix as an image
- Enhance data visualization, natural logarithm transformation is applied
- ColorNet: Investigating the importance of color spaces for image classification. In *Computer Vision—ACCV 2018:* RGB coloring was used as the study dictates that RGB yielded better results compared to other color formats. [6]
- Ada-Delta optimizer [7]
- Pooling Layers





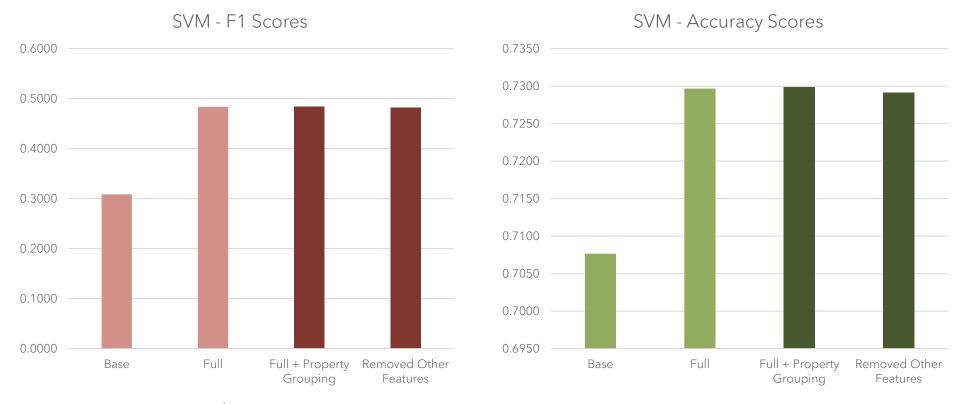
Results

# **SVM Results**



- Base Amino Acid Counts
- Full pKa, Amino Acid TFIDF, Secondary Structure TFIDF, Tertiary Structures, Hydrophobicity
- Full + Property Grouping Addition of Property grouping to the previous data set

# **SVM Results**

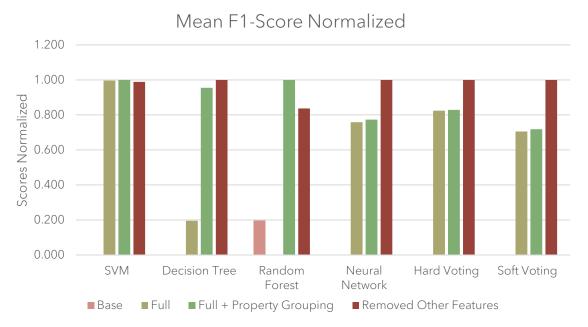


- Base Amino Acid Counts
- Full pKa, Amino Acid TFIDF, Secondary Structure TFIDF, Tertiary Structures, Hydrophobicity
- Full + Property Grouping Addition of Property grouping to the previous data set.
- Removed Amino Acid TFIDF, Tertiary Structures, and Property Grouping

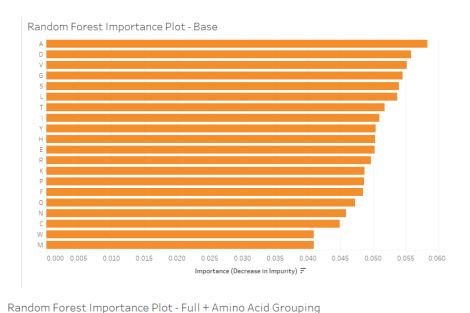
#### **Mean F1- Scores**

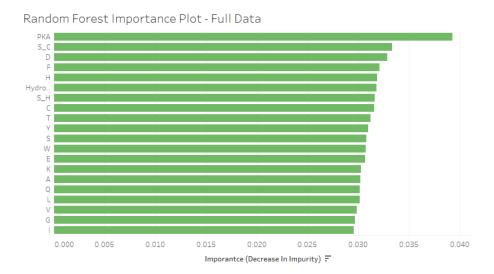
- We can observe the min-max normalization was applied to the scores for each algorithm, facilitating a clearer visualization
- SVM and Random Forest were the sole algorithms that exhibited improved performance with the full dataset and property grouping
- Other algorithms exhibited a stronger preference for the removed features

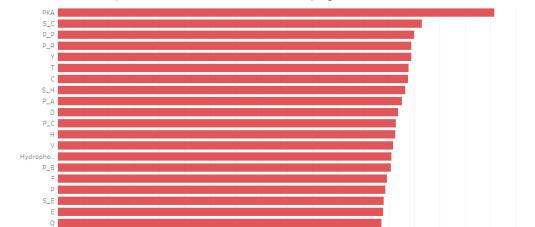




# **Random Forest Importance Plots**

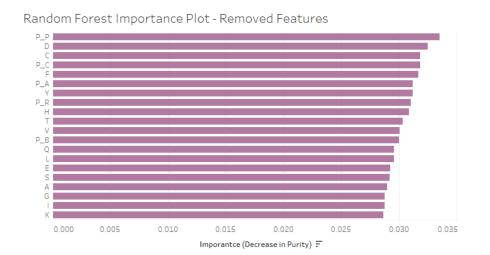






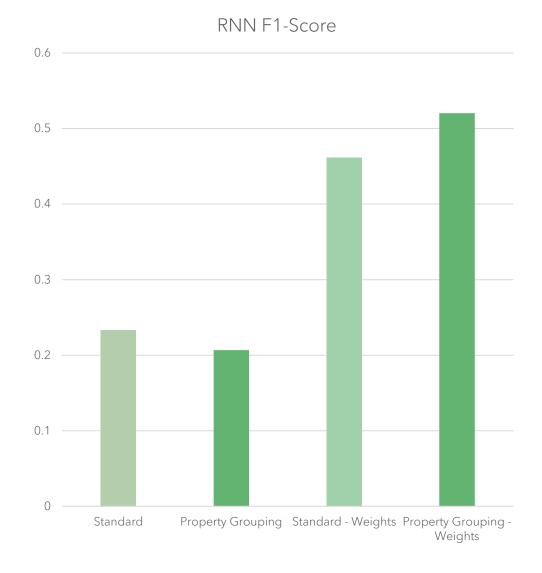
0.002 0.004 0.006 0.008 0.010 0.012 0.014 0.016 0.018 0.020 0.022 0.024 0.026 0.028 0.030 0.032 0.034

Imporantce (Decrease in Purity)



# **Standard Encoder**

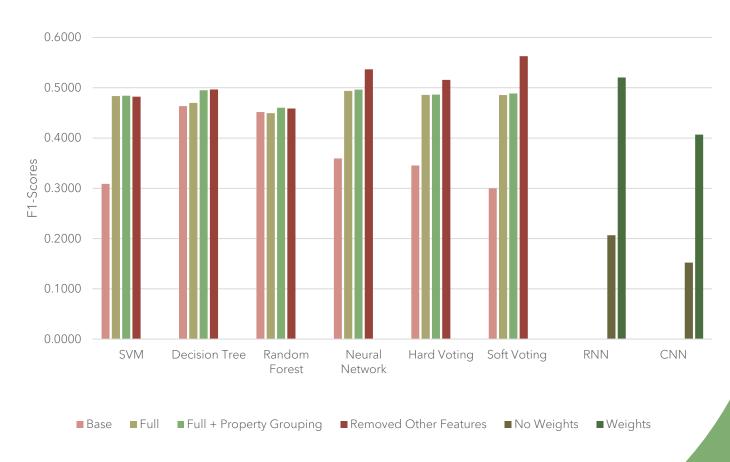
- We can observe that our Property Grouping Encoder outperformed the standard encoder.
- Hence, the property encoder was used for the CNN.
- Addition of weights helped obtain better F1-Scores.



# Mean Scores For RNN and CNN

- F1-scores indicated suboptimal performance, potentially attributed to overgeneralization and insufficient data for neural network training.
- Considering the imbalance in our dataset, class weights were introduced to enhance balance and consistency
- The assertion that the imbalanced dataset contributed to performance disparities F1-scores for each protein function. For example, Molecular function consistently achieved the highest accuracy and F1-scores across all algorithms (0.86).

Mean F1-Score





**Conclusion and Future Work** 

## **Conclusions**

- Amino acid property grouping technique shows promise for predicting protein sequence function.
- Deliberate introduction of high variance in sample data is crucial, impacting the significance of each feature.
- Despite intentional variance, property grouping technique demonstrates minimal loss in predictive power.
- SVM and decision tree algorithms exhibit capacity for property grouping without substantial decline in predictive ability.
- Neural network and voting algorithms show improved scores with the introduction of amino acid property grouping.
- Recurrent neural networks (RNN) and convolutional neural networks (CNN) achieve comparable F1-scores to feature-embedded algorithms with class weights, but technical constraints may contribute to suboptimal performance.

# **Future Work**

- Further analyses are needed before confirming the significance of this property grouping.
- Alternative sampling techniques should be explored to assess model performance under reduced sequence variance.
- Sampling strategies ensuring a balanced distribution of functions could provide additional insights.
- High-performance computation can be leveraged to utilize the entire dataset, mitigating concerns about data imbalance.
- The Transformer algorithm, with its multidirectional processing and attention mechanisms, holds promise for protein function prediction. [11]
- Recent publications explore Transformer architectures for protein sequences, but none align with the concept of amino acid property grouping.
- A specialized encoder based on amino acid grouping could be integrated into a Transformer encoder for enhanced protein function prediction.

# Motivation and Inspiration

- Background in Bioinformatics
- A unique look at a problem
- This work was inspired by several researches within the field
- An attempt to try and make a difference in an immerging field of research.
- Still a lot more work to be done.
  However, through our preliminary work
  we have shown that there is some
  promise to the property binning
  technique which requires further
  investigation.





### References

- 1. Amino acids reference chart. (n.d.). https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/protein-structural-analysis/amino-acid-reference-chart
- 2. CAFA 5 protein function prediction. Kaggle. (n.d.). https://www.kaggle.com/competitions/cafa-5-protein-function-prediction/data
- 3. Cai, C. Z., Wang, W. L., Sun, L. Z., & Department of the Company of the Compan
- 4. Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., & Chen, Z. (2017). ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules (Basel, Switzerland)*, 22(10), 1732. https://doi.org/10.3390/molecules22101732
- 5. "Gene Ontology Resource." *Gene Ontology Resource*, geneontology.org/. Accessed 9 Dec. 2023.
- 6. Gowda, S. N., & Yuan, C. (2019). ColorNet: Investigating the importance of color spaces for image classification. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14* (pp. 581-596). Springer International Publishing.
- 7. Kandel, I., Castelli, M., & Popovič, A. (2020). Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images. Journal of imaging, 6(9), 92. https://doi.org/10.3390/jimaging6090092
- 8. Moffat, L., & D. T. (2021). Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. Bioinformatics, 37(21). https://doi.org/https://doi.org/10.1093/bioinformatics/btab491
- 9. Sara, S. T., Hasan, M. M., Ahmad, A., & Samp; Shatabda, S. (2021). Convolutional neural networks with image representation of amino acid sequences for protein function prediction. Computational Biology and Chemistry, 92(1476–9271), 107494. https://doi.org/10.1016/j.compbiolchem.2021.107494
- 10. U.S. National Library of Medicine. (n.d.). Conserved domains database (CDD) and resources. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
- 11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30