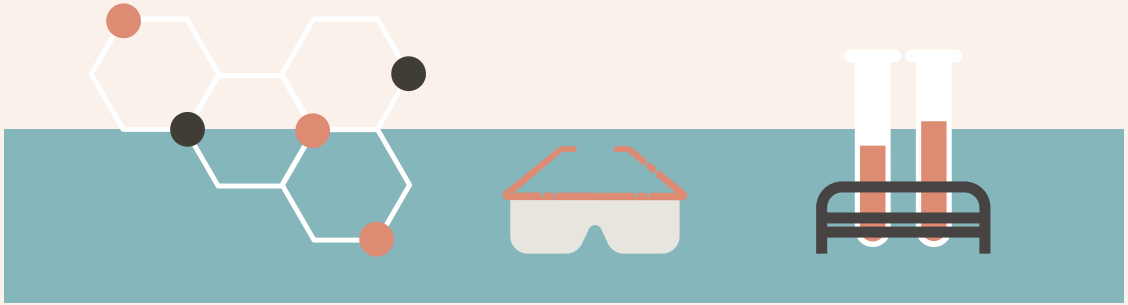
Protein Family Classification



Aarohi Chopra & Daniel Quintana

01

OVERVIEW

Why do we care about this?

02

BACKGROUND

Crash course on proteins!

03

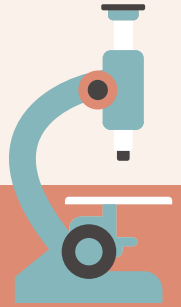
METHODS

How did we do what we did?

04

RESULTS

This is what we got!!! :D



01 OVERVIEW

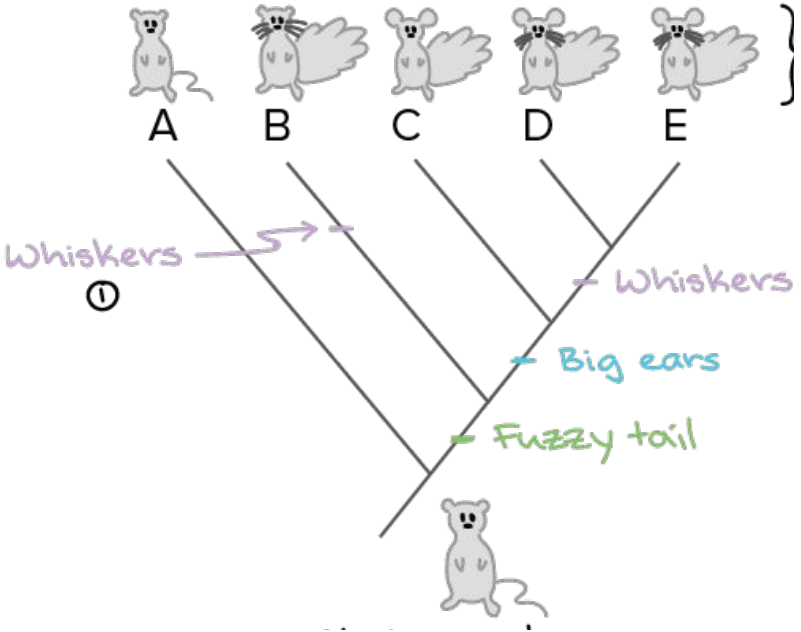
What we want to do!

Problem Statement

- Our goal is to predict what family a protein belongs to.

- Why?
  1. Evolutionary Relationships - Family Trees!

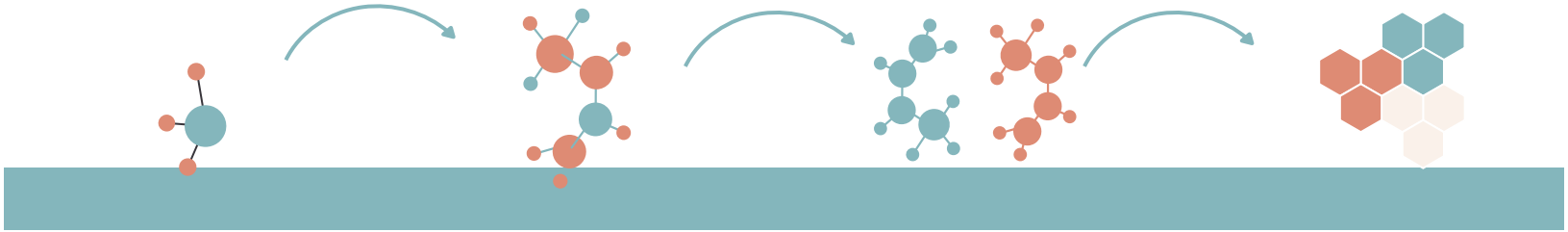
  2. Infer novel proteins' biological functions



02 BACKGROUND

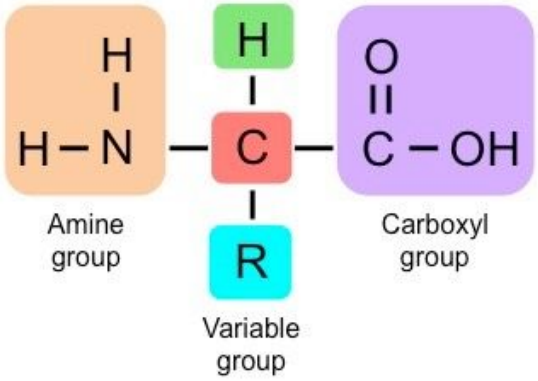
Protein Time!!! :D

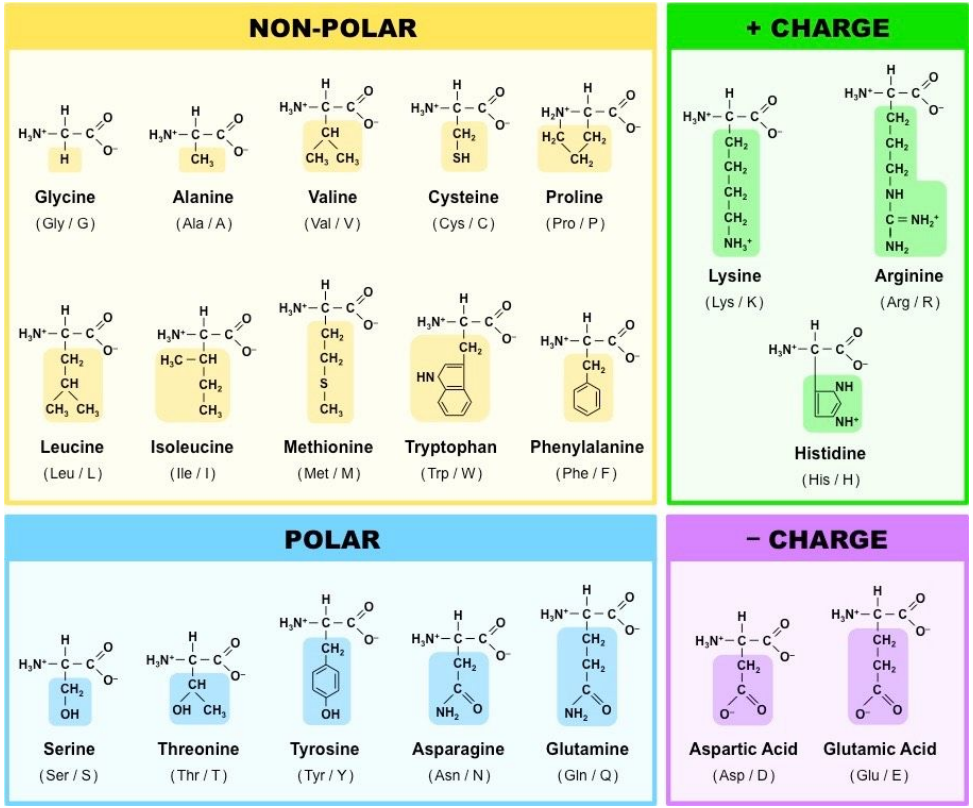
Background



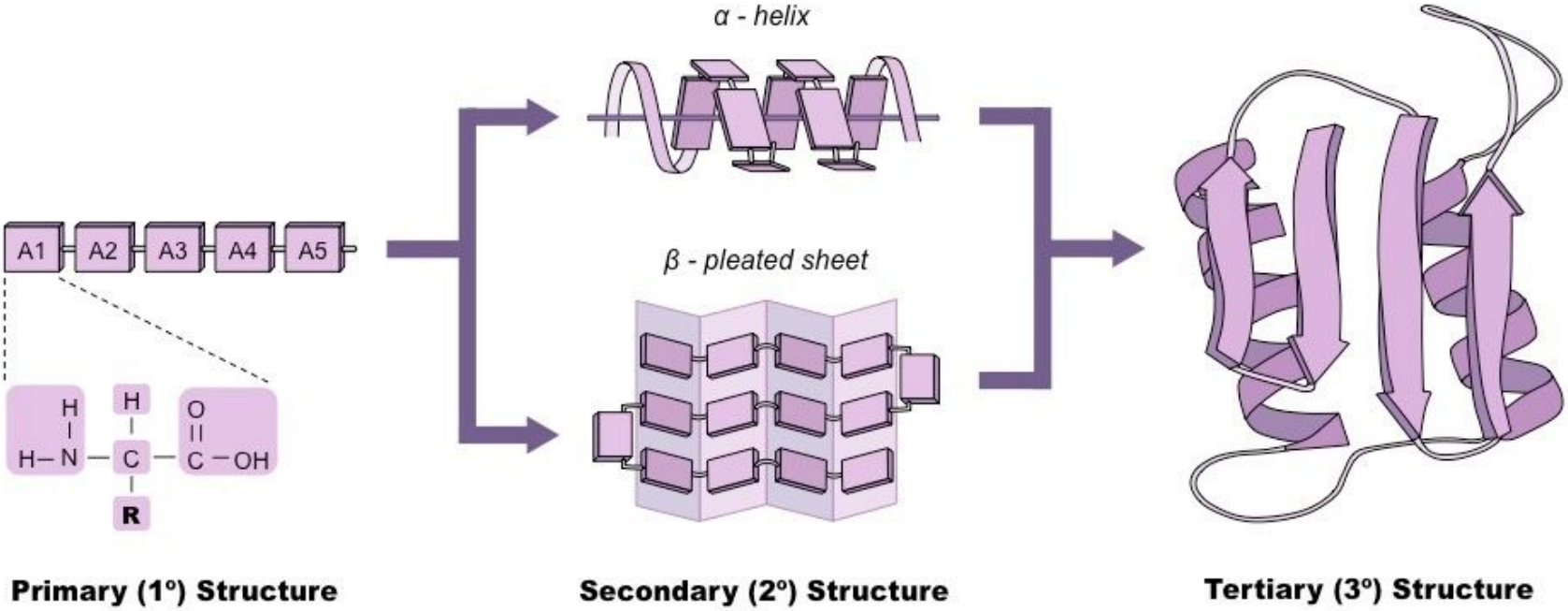
Amino Acids Proteins Classification Familys

Amino Acids

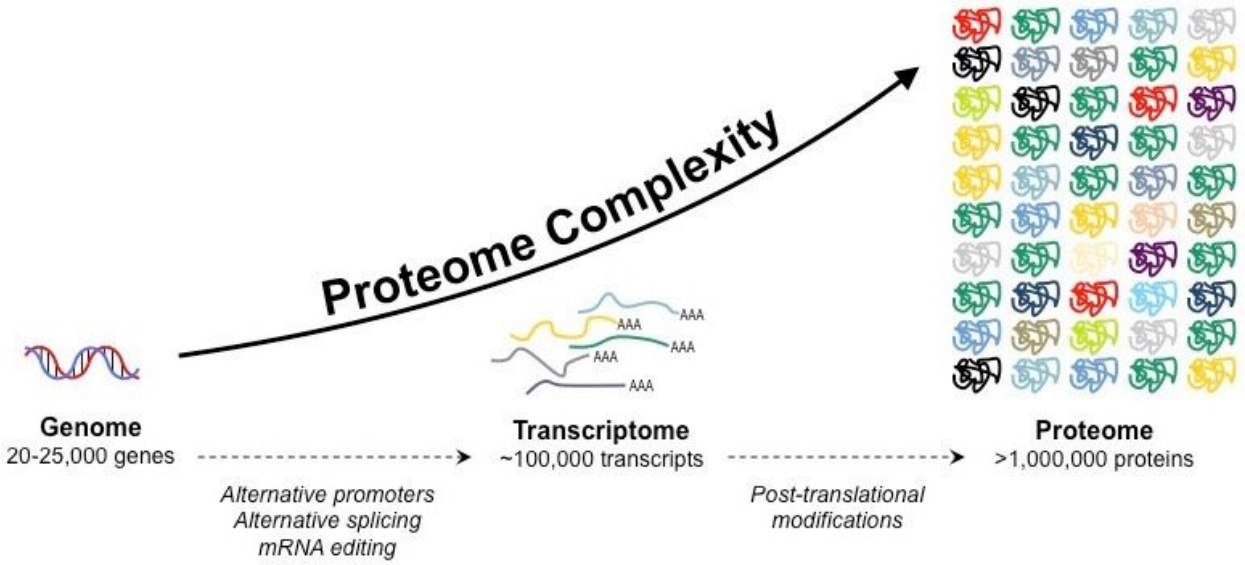

H
H–N
C
O
C–OH
Amine group
Carboxyl group
R
Variable group


NON-POLAR
Glycine (Gly / G)
Alanine (Ala / A)
Valine (Val / V)
Cysteine (Cys / C)
Proline (Pro / P)
Leucine (Leu / L)
Isoleucine (Ile / I)
Methionine (Met / M)
Tryptophan (Trp / W)
Phenylalanine (Phe / F)
+ CHARGE
Lysine (Lys / K)
Arginine (Arg / R)
Histidine (His / H)
POLAR
Serine (Ser / S)
Threonine (Thr / T)
Tyrosine (Tyr / Y)
Asparagine (Asn / N)
Glutamine (Gln / Q)
– CHARGE
Aspartic Acid (Asp / D)
Glutamic Acid (Glu / E)

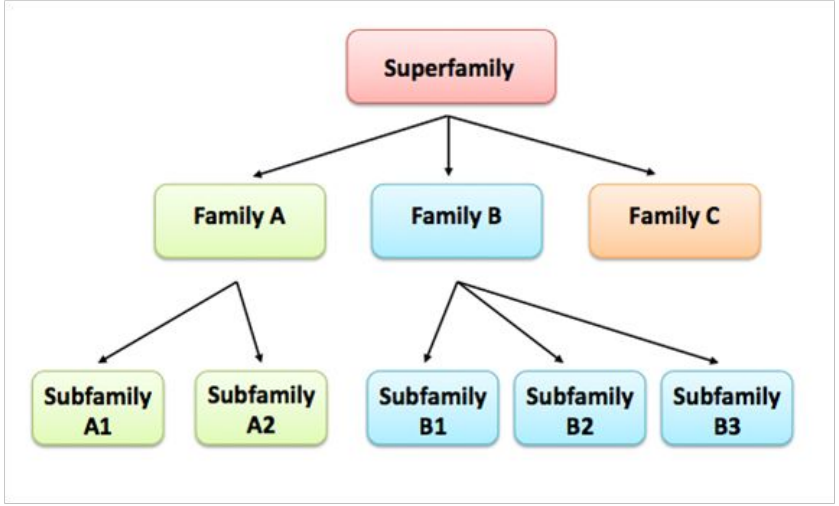
Proteins

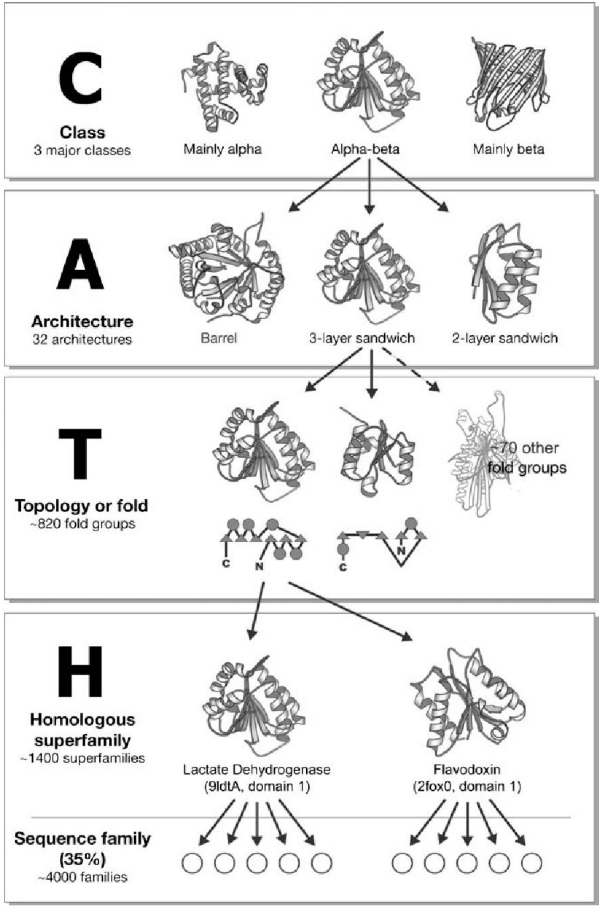

α - helix
β - pleated sheet
A1
A2
A3
A4
A5
H
H–N
C
R
O
C–OH
Primary (1°) Structure
Secondary (2°) Structure
Tertiary (3°) Structure

Classification and Effects



Familys


Superfamily
Family A
Family B
Family C
Subfamily A1
Subfamily A2
Subfamily B1
Subfamily B2
Subfamily B3

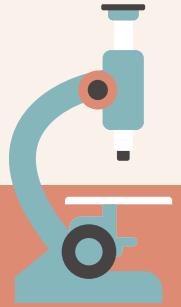

C
Class
3 major classes
Mainly alpha
Alpha-beta
Mainly beta
A
Architecture
32 architectures
Barrel
3-layer sandwich
2-layer sandwich
T
Topology or fold
~820 fold groups
~70 other fold groups
H
Homologous superfamily
~1400 superfamilies
Lactate Dehydrogenase (9ldtA, domain 1)
Flavodoxin (2fox0, domain 1)
Sequence family (35%)
~4000 families

Problem Statement

- Our goal is to predict what family a protein belongs to.
  - Saves time!

How?

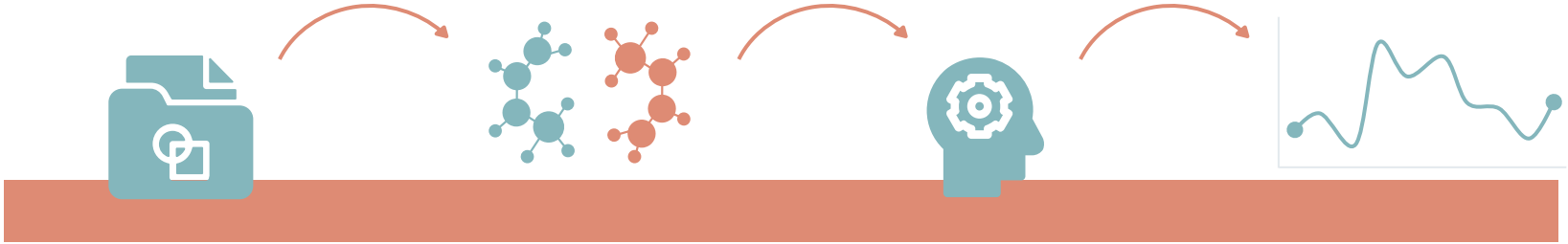
- Graph Classification!



03 METHODS

Here is what we did!

METHODS



Data Graph Model Tuning

Data

- Kaggle Data “Deep Protein Sequence family Classification”
  1. Metadata

```
structureId
chainId
sequence
residueCount_x
classification
experimentalTechnique
macromoleculeType_y
residueCount_y
resolution
structureMolecularWeight
crystallizationMethod
crystallizationTempK
densityMatthews
densityPercentSol
pdbxDetails
phValue
```

	structureId	classification	experimentalTechnique	macromoleculeType	...	:entSol	pdbxDetails	phValue	publicationYear
0	100D	DNA-RNA HYBRID	X-RAY DIFFRACTION	DNA/RNA Hybrid	...	30.89	pH 7.00, VAPOR DIFFUSION, HANGING DROP	7.0	1994.0
1	101D	DNA	X-RAY DIFFRACTION	DNA	...	38.45	NaN	NaN	1995.0
2	101M	OXYGEN TRANSPORT	X-RAY DIFFRACTION	Protein	...	60.20	3.0 M AMMONIUM SULFATE, 20 MM TRIS, 1MM EDTA, ...	9.0	1999.0

  2. Sequence

	structureId	chainId	sequence	residueCount	macromoleculeType
0	100D	A	CCGGCGCCGG	20	DNA/RNA Hybrid
1	100D	B	CCGGCGCCGG	20	DNA/RNA Hybrid
2	101D	A	CGCGAATTCGCG	24	DNA
3	101D	B	CGCGAATTCGCG	24	DNA
4	101M	A	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...	154	Protein

Data

- Exploratory Data Analysis:

1. Removed non-proteins

	structureId	macromoleculeType
0	100D	DNA/RNA Hybrid
1	100D	DNA/RNA Hybrid
2	101D	DNA
3	101D	DNA
4	101M	Protein

	structureId	macromoleculeType_y
0	1TQF	Protein
1	1Y4W	Protein
2	5MJY	Protein
3	3ENP	Protein
4	1NPZ	Protein

2. Removed duplicate ID and Sequence

```
4V98    160
6EKC    160
4V6B    144
4V46    120
4NWR     96
        ...
1L9V      1
4K6R      1
4K6N      1
1ZCY      1
4EQ5      1
Name: structureId,
```

```
1A04    1
4IAQ    1
4IBZ    1
4IBY    1
4IBX    1
       ..
3AVR    1
3AVO    1
3AV8    1
3AV4    1
6FAH    1
```

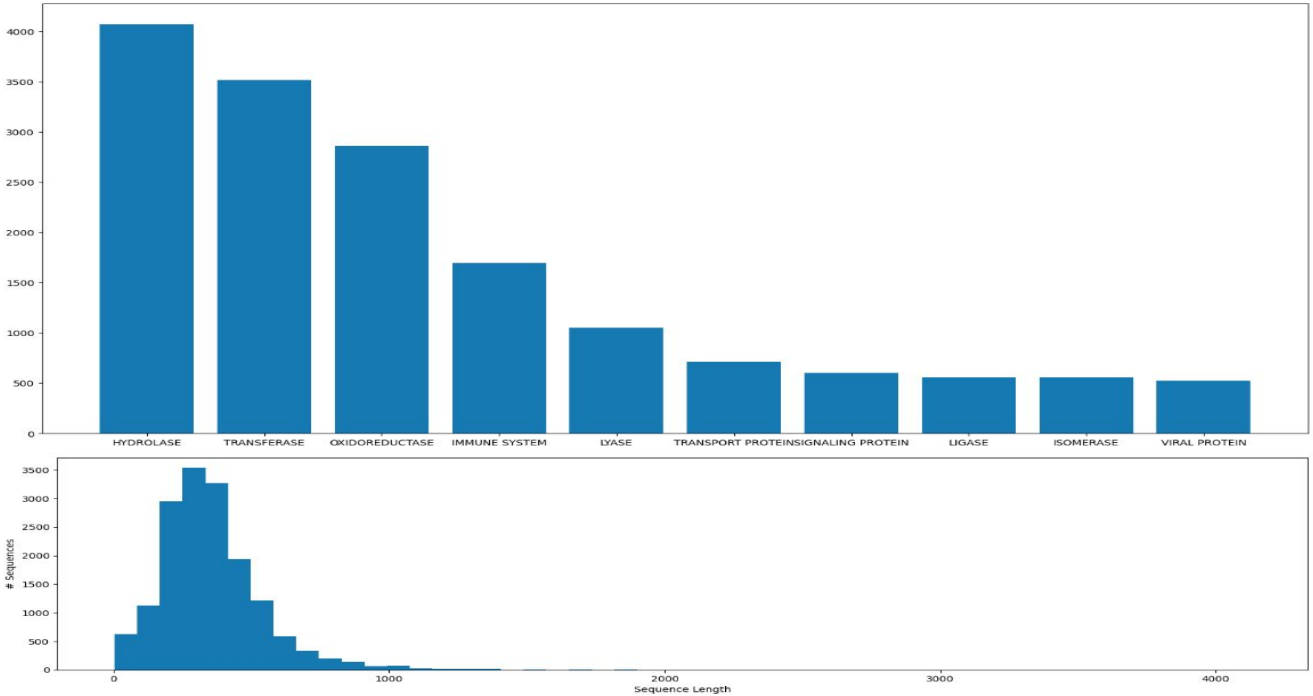
sequence
CCGGCGCCGG
CCGGCGCCGG
CGCGAATTCGCG
CGCGAATTCGCG

1. Removed short sequences

3	101D	B	CGCGAATTCGCG	24	DNA
4	101M	A	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...	154	Protein

Data

- Exploratory Data Analysis:
  - Only picked top 10 families, balanced data (all to 500)



Graphs

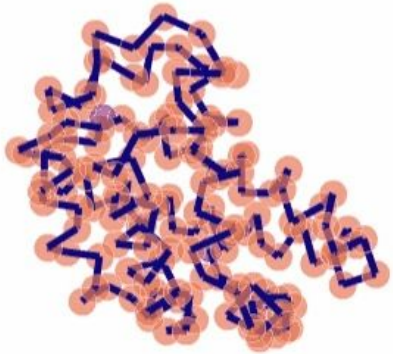
- Queried the “RCSB Protein Data Bank” to for the aforementioned proteins pdb files

```
Downloading PDB files:  11%|▉          | 530/5000 [13:47<1:58:00,  1.58s/it]Warning: Failed to download PDB file with ID 4V7G
Downloading PDB files:  17%|▊          | 869/5000 [22:07<2:20:27,  2.04s/it]Warning: Failed to download PDB file with ID 5BP4
Downloading PDB files:  49%|████▊      | 2426/5000 [1:00:56<57:49,  1.35s/it]  Warning: Failed to download PDB file with ID 5EUJ
Downloading PDB files:  91%|████████▏  | 4549/5000 [1:52:28<10:52,  1.45s/it]Warning: Failed to download PDB file with ID 4V96
Downloading PDB files: 100%|█████████| 5000/5000 [2:04:53<00:00,  1.50s/it]Total successful downloads: 4996
```

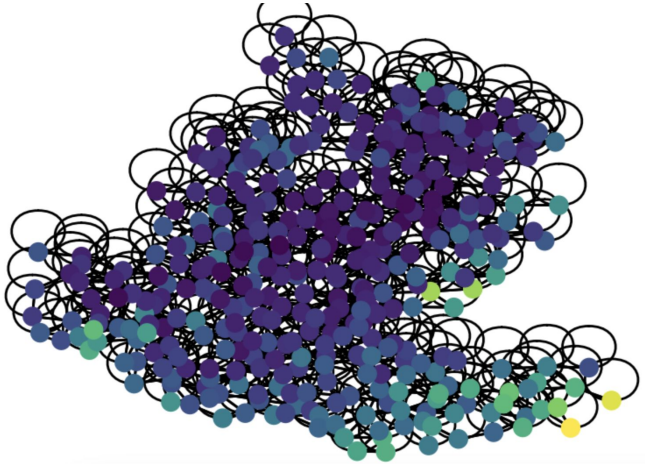
Graphs

- Turned pbd's into graphs and labels using their respective family

1. Graphein



2. NetworkX

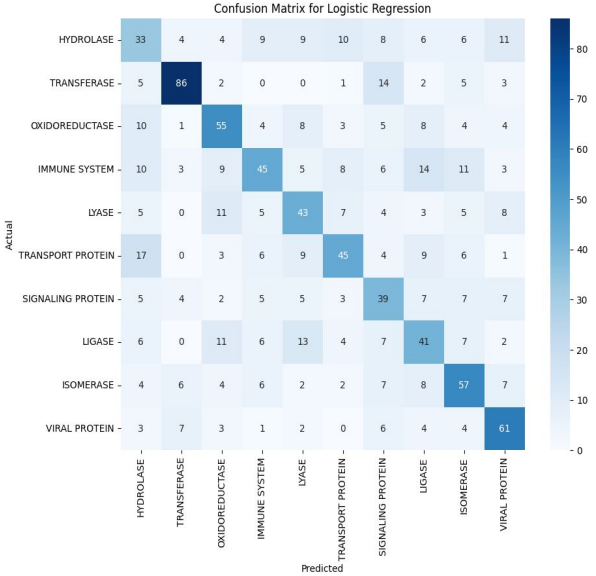


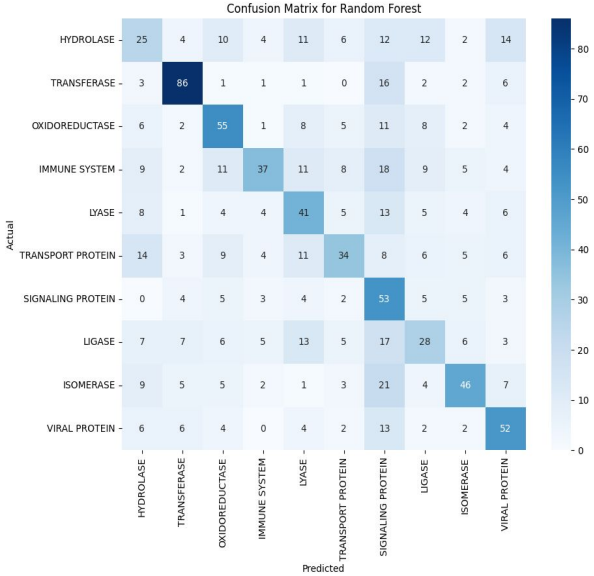
Models

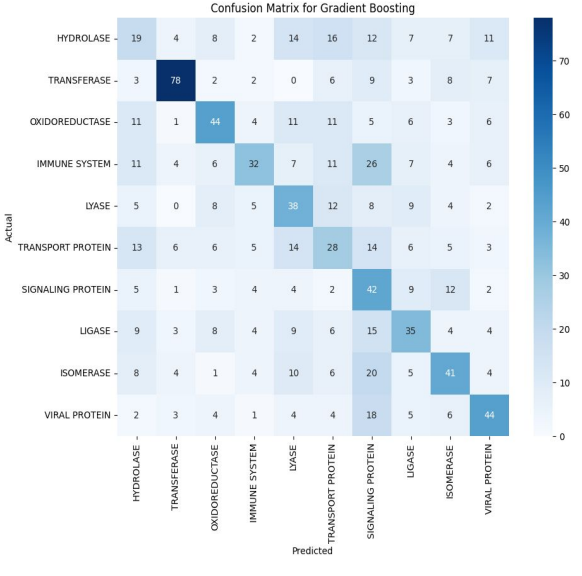
- Now we have three data sets!
    1. Metadata + Sequence
    2. Sequence
    3. Graphs

Traditional Models

- Machine Learning on: Sequence & Metadata
  - Logistic Regression - max iteration = 50
  - Random Forest - n_estimators = 100
  - Gradient Boosting - n_estimators = 100

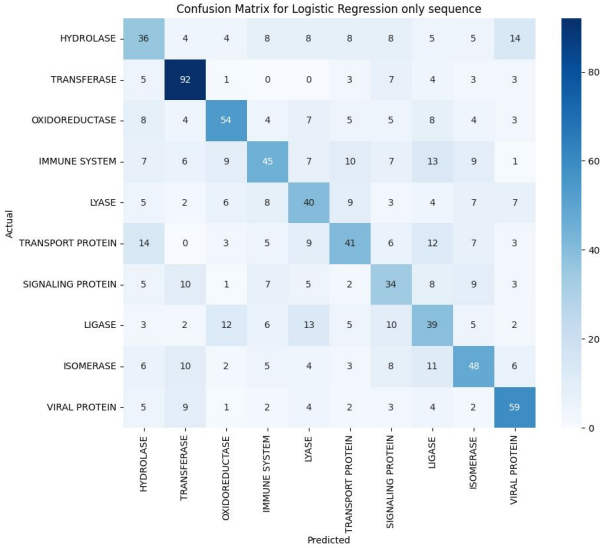


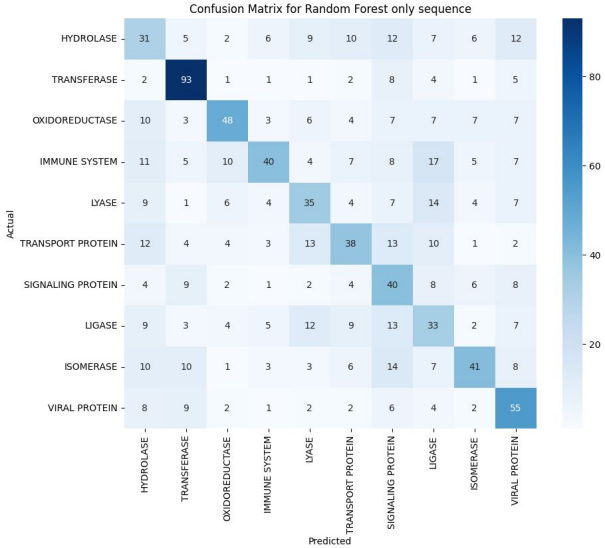


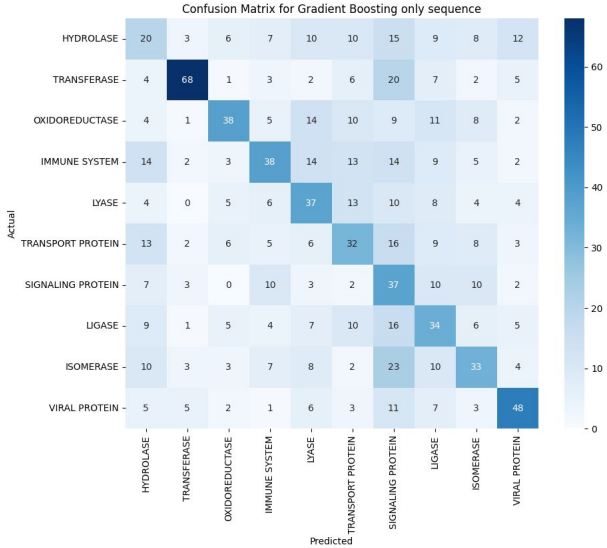


Traditional Models

- Machine Learning on: Sequence Only, using CountVectorizer
  - Logistic Regression - max iteration = 50
  - Random Forest - n_estimators = 100
  - Gradient Boosting - n_estimators = 100

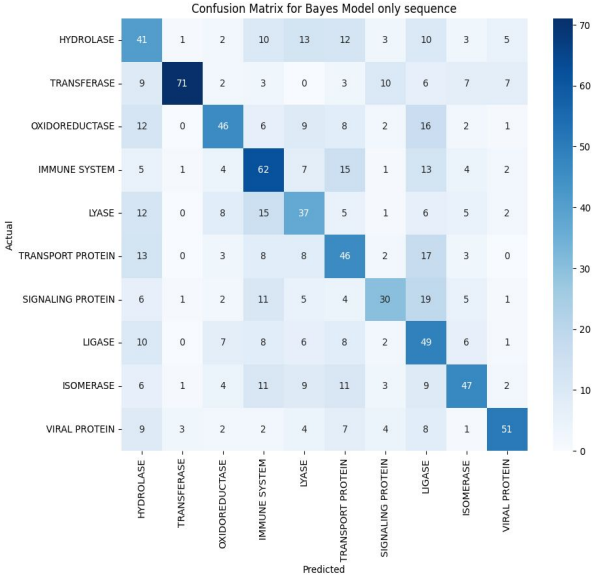


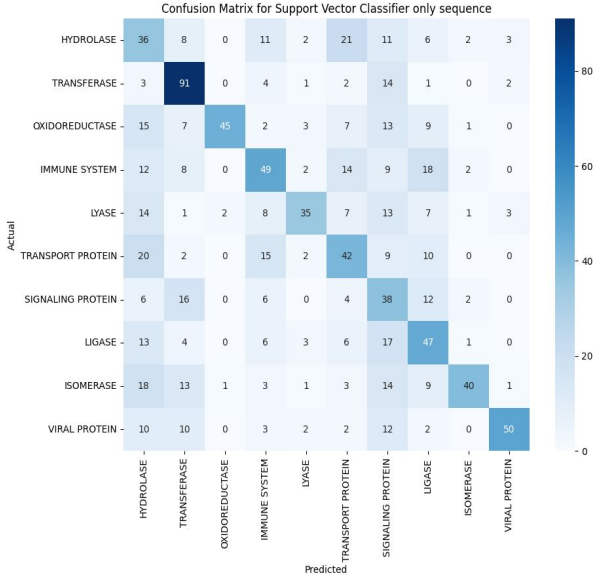




Traditional Models

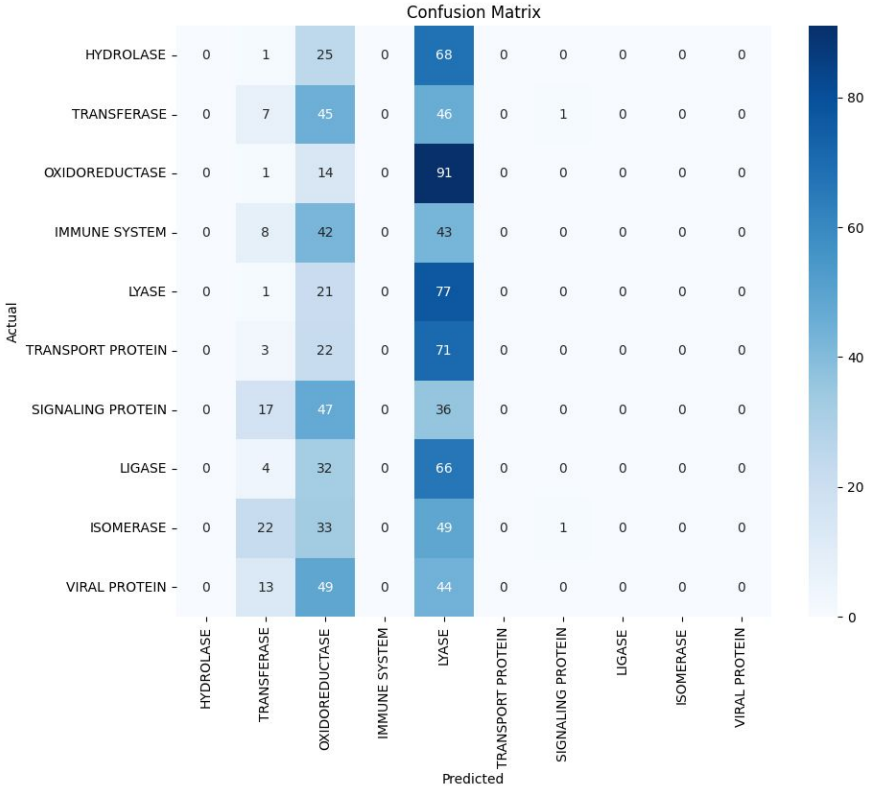
- Machine Learning on: Sequence Only
  - Naive bayes
  - SVM - kernel = rbf



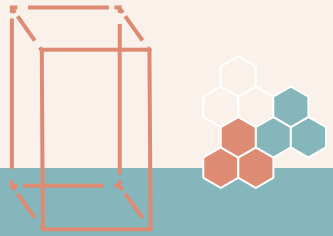


Graph Models + Tuning

- GCN
  - 20 Epocs
  - Hidden size: 64
  - Learning rate 0.001 & 0.0005
  - Dropout rate 0.3 & 0.5
  - ADAM optimizer
  - Training Time: 6+ hrs



04 RESULTS



Look at the pretty chart! :O

Results

Dataset	Model	Best Protein	Best Protein Accuracy	Overall Accuracy	Overall Ranking!
Kaggle Person	CNN	IMMUNE SYSTEM	0.59	0.22	8th
	CNN v2	LYASE	0.72	0.59	1st
Sequence + Metadata	Logistic Regression	TRANSFERASE	0.86	0.50	3rd
	Random Forest	TRANSFERASE	0.86	0.53	2nd
	Gradient Boosting	TRANSFERASE	0.78	0.40	6th
Sequence Only	Logistic Regression	TRANSFERASE	0.92	0.49	4th
	Random Forest	TRANSFERASE	0.93	0.45	5th
	Gradient Boosting	TRANSFERASE	0.68	0.39	7th
	Naive Bayes	TRANSFERASE	0.71	0.49	4th
	SVM	TRANSFERASE	0.91	0.49	4th
Graphs	GCN	LYASE	0.77	0.15	9th
	GATs			TBD	
	GraphSAGE			TBD	

Takeaways

- Kaggle people only got 59% accuracy!!!! :o
- Our Baseline was 50% but we were using less data

- We first thought the model worked better with Sequence Only, we were wrong :(

- Graphs were not as good as just using all the data

Next Steps

- Improving GCNs
  - B-factor
- GATs
- GraphSAGE
- Graph that includes metadata

Q&A

:)