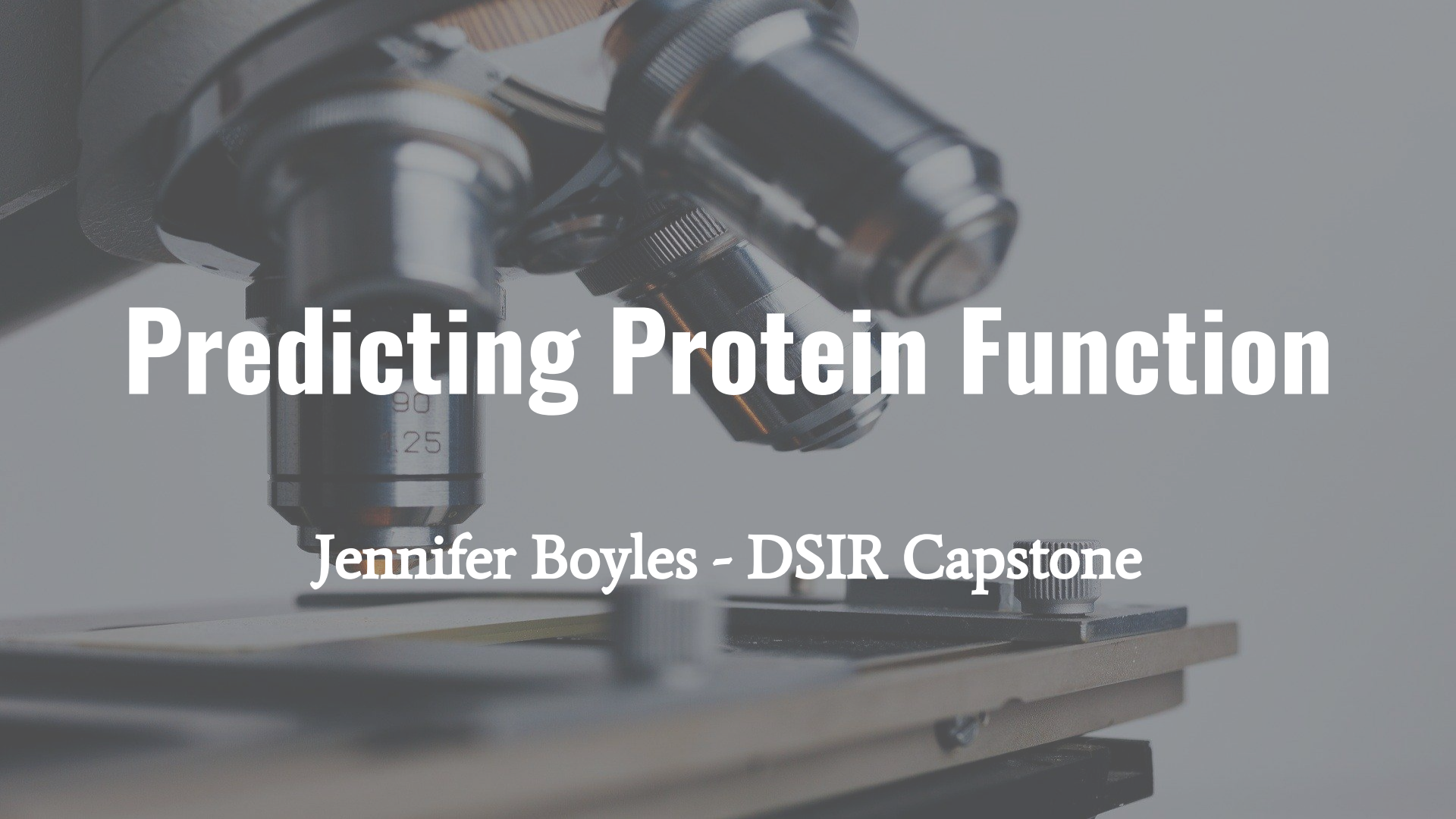
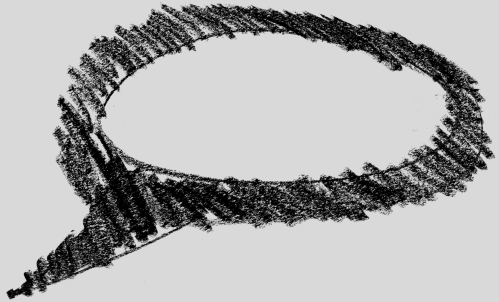


Predicting Protein Function

Jennifer Boyles - DSIR Capstone

Agenda

- What is a Protein?
- The Problem
- Data & Data Cleaning
- EDA
- Null Model & Evaluation Metrics
- Final Model with Demo
- Conclusion & Recommendations
- Improvements



What is a protein?

- DNA Sequence (Gene) ====> Protein Sequence

- Proteins: central role in all biological processes
  - Hormones, Immune System, Metabolism

Amino Acid Sequence

- Linear polymer made up of amino acid monomers.
- 21 unique amino acids represented with 1-letter abbreviation

Sequence vs Function Disparity

- 1990 - 2003: Human Genome Project - 200 labs & 18 countries, sequence full human genome

- 1 day: Next-Generation Sequencing (Illumina, etc.)

- Protein functional assays & microarrays take weeks to perform

Problem Statement

- Identification of novel proteins functions have potential benefits in medicine, agriculture, nutrition, etc.

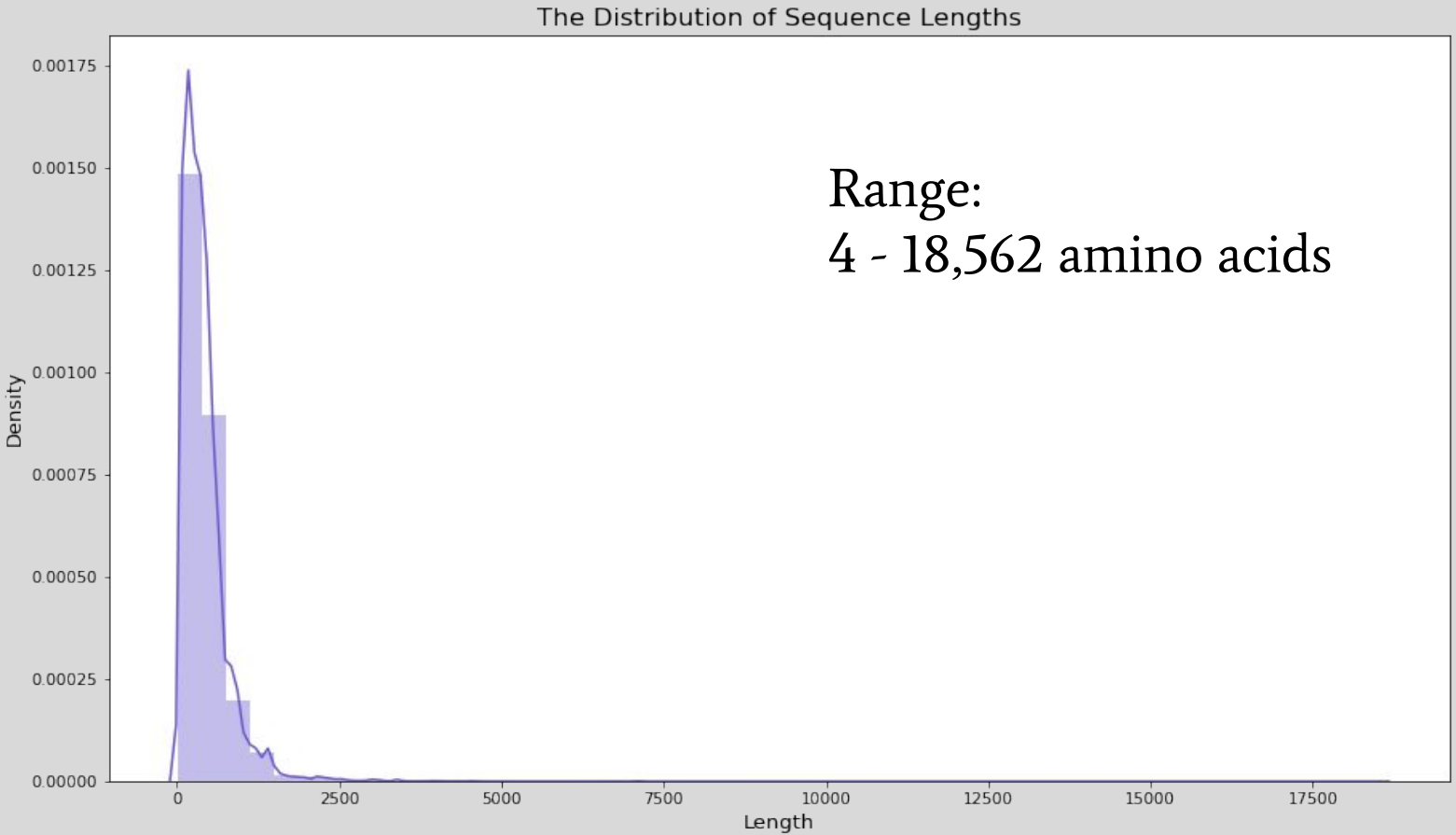
- Can we rapidly identify protein function based on amino acid sequence?

Data Acquired & Cleaned

- UniProt: Freely accessible database
  - SwissProt - reviewed protein sequences

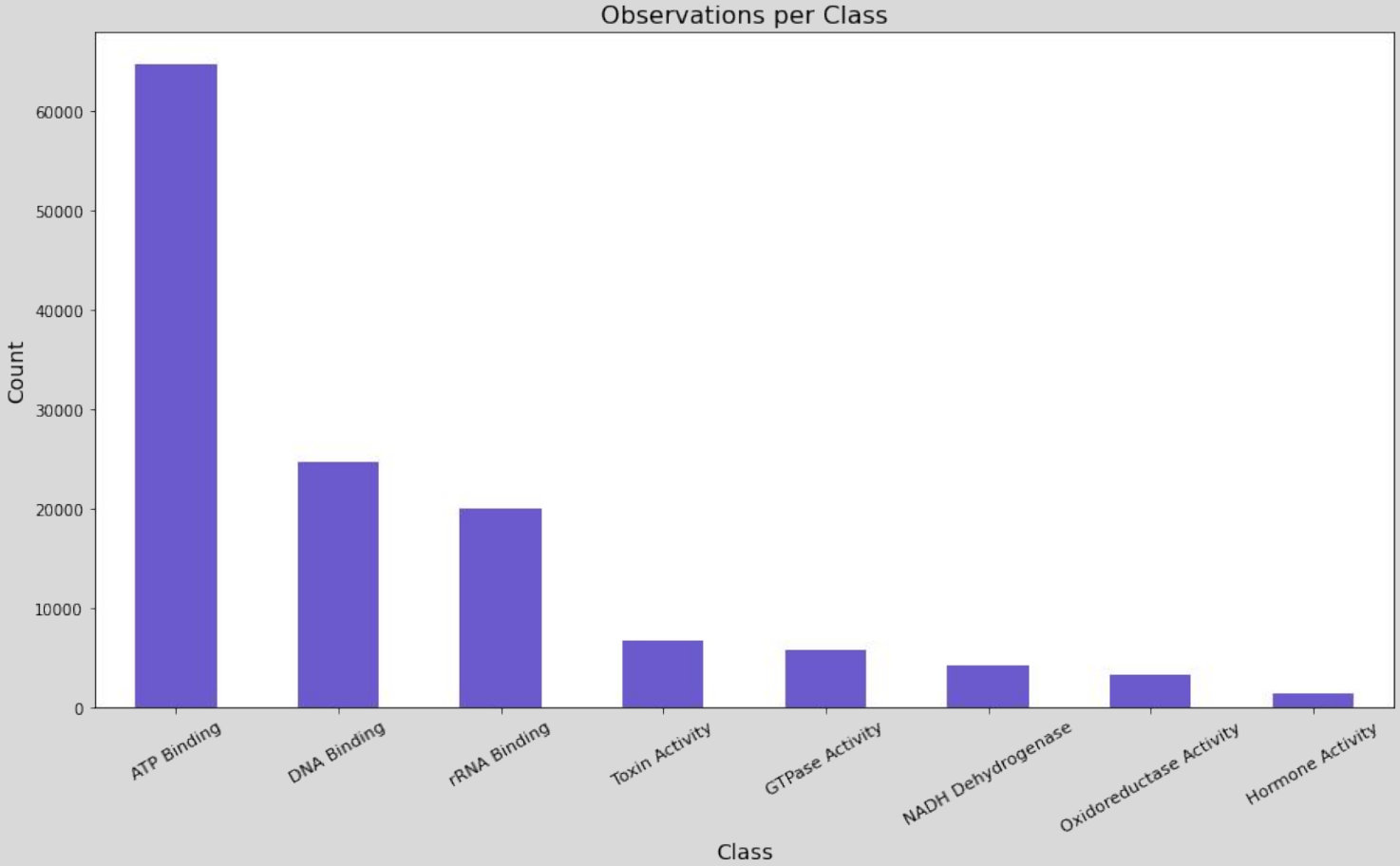
- 8 function classes formed with GO I.D.
  - rRNA binding/Ribosome structural component, DNA binding, ATP binding, Hormone, GTPase, NADH Dehydrogenase (ubiquinone & quinone), Oxidoreductase, Toxin

Sequence Lengths



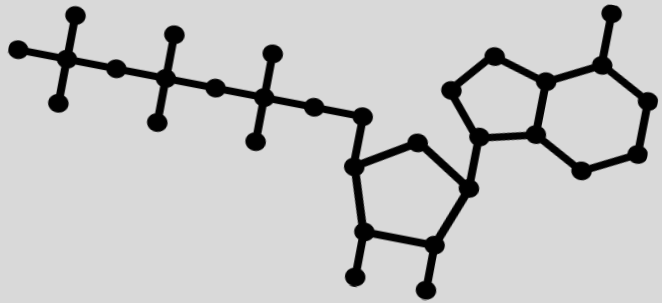
Range:
4 - 18,562 amino acids

Imbalanced Classes



Null Model

- Baseline comparison to deep learning classifiers
- Based on most frequent class value
- Null model accuracy: 49.4%



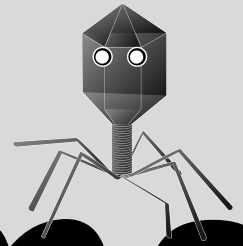
Evaluation Metrics

- Balanced Accuracy: Imbalanced data
- F1-Score: Balance recall & precision

Final Model

- RNN: Oversampled minority classes
  - Embedding Layer
  - 3 Convolutional Layers
  - Bidirectional GRU of 72 Nodes
  - 2 Hidden Layers: 160, 80 Nodes
- Balanced Accuracy: 94.3%
- F1-Score: 95.7%

Demo



Conclusion & Recommendations

- Strong predictive power with amino acid sequence
  - Did not utilize sequence alignment or homology

- Narrow scope of research

Improvements

- Create comprehensive classes with consistent hierarchical representation
- Continue to train as sequence data grows
  - Focus on underrepresented classes

Thank you!

Questions?

Citations

- All Images: No Attribution Required from Pixabay