Protein Function Prediction Using Deep Neural Networks

Project Purpose

- **Motivation:** Address the challenge of protein function annotation.
- **Objective:**
  - Develop multi-label classification model to enhance biological insights using custom-built neural networks.
  - Determine optimal embedding and model combination for functional annotation.
- **Novelty:**
  - Develop our original NN models to explore embedding-annotation relationships.
  - Integrate additional biological information for enhanced model accuracy.

Protein Functions and Gene Ontologies (GO)

- **3 GO categories (**Ashburner et al., 2000**):**
  - **Molecular Function (MF):** Protein's physical activities
    - Ex. Sequence-specific DNA binding
  - **Biological Process (BP):** Biological goals led by a protein's MF(s)
    - Ex. Cell death
  - **Cellular Component (CC):** Protein's location
    - Ex. Nucleus

*Accurate prediction aids research in disease mechanism identification, drug discovery, and evolutionary studies.

Data Overview

- **Protein Sequences:** Textual representation of amino acid sequences.
- **InterPro Annotations:** Categorize proteins into families, domains, and functional sites. A single protein can belong to multiple InterPro categories.
- **GO Labels:** Multi-label annotations linking proteins to GO terms.

	index	proteins	accessions	genes	sequences	annotations
57336	400055	RPA12_YEAST	P32529; D6VWN4;	853526	MSVVGSLIFCLDCGDLLENPNAVLGSNVECSQCKAIYPKSQFSNLK...	[GO:0005736|IDA, GO:0003676|IEA, GO:0001054|ID...

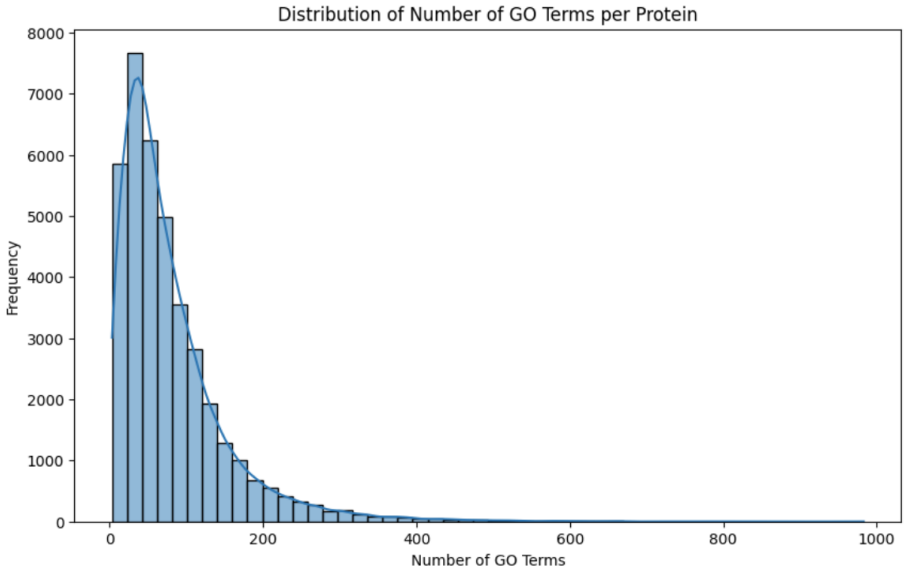
Protein **Protein Sequence**

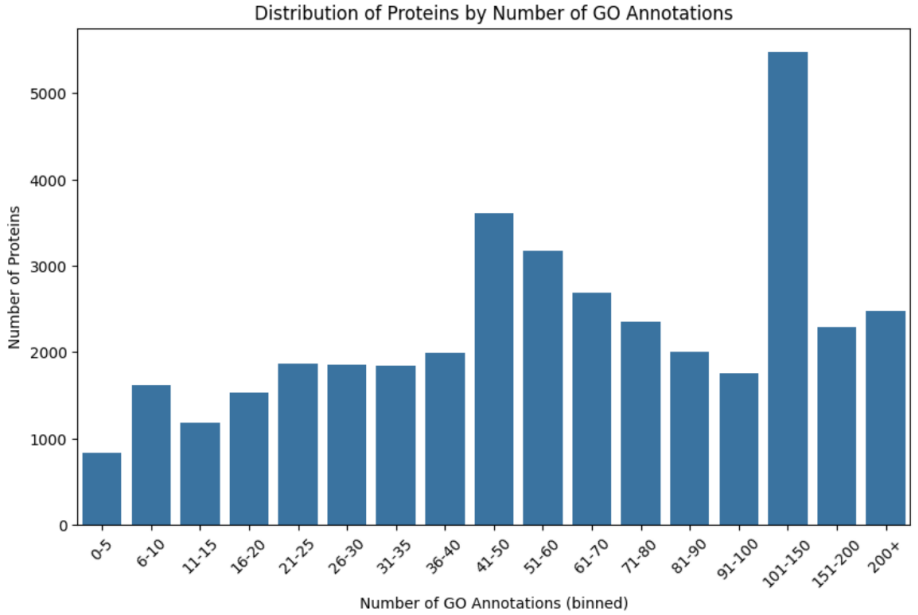
string_ids	orgs	interpros	exp_annotations	prop_annotations	cafa_target
[4932.YJR063W]	559292	[IPR019761, IPR001529, IPR012164, IPR034004, I...	[GO:0005736, GO:0001054, GO:0061629, GO:000012...	[GO:1901576, GO:0006353, GO:0044271, GO:005125...	True

InterPros **Gene Ontology Labels**

Exploratory Data Analysis

- **Multi-Label Structure:** Proteins associated with several GO terms.
- **Project Challenge:** Handle multi-label structure accurately for reliable predictions.





Feature Engineering

- **Protein Embeddings**:

  Transformed protein sequences into numerical vectors for model input.
  - Generated fixed-size embeddings using pretrained models (ESM2, ProtT5, TAPE, and ProtBERT) from Hugging Face.
  - Preprocessed sequences by replacing rare amino acids with 'X' and adding spaces between residues and tokenized using pre trained model's tokenizer
  - Mean pooled hidden states from the last layer to create fixed-size vector representations.

- **InterPro Integration**: Multi-hot encoded features capturing protein families, domains, and functional sites.

esm2	prott5	tape	protbert
[0.018639212, 0.032441415, -0.057677362, -0.06...	[-0.044096634, -0.050428215, -0.07790137, 0.10...	[0.4024663, -0.09638704, 0.41344175, -0.911272...	[0.04877469, -0.021751985, -0.015491902, -0.02...

Generated Embeddings

Model Architecture and Variants

Baseline Neural Network Architecture:

1. **Input Layers:** Accepts embeddings (e.g. ESM2, ProtT5, TAPE) or combined embeddings + InterPro.
2. **Hidden Layers:**
    - Fully connected layers (1024 → 512), ReLU activation.
    - Dropout layers (30%) to improve generalization.
3. **Output Layer:** Sigmoid activation for multi-label classification.

Model Variants:

1. **Embedding-Only Models:** Input: Protein sequence embeddings.
2. **Embedding + InterPro (Concatenated):** Input: Concatenated embeddings and multi-hot encoded InterPro annotations.
3. **Separate Processing Models:** Independent processing for embeddings and InterPro annotations. Features merged before final layers, allowing specialized processing.

Custom Logical Loss Function

- **Binary Cross-Entropy (BCE) Loss:** Handles multi-label classification.
- **Logical Loss Function:** Enforces biological consistency in predictions by incorporating GO axioms as constraints.
  - **A Implies B**: If term A is predicted, term B must also be predicted. Loss: Penalizes when P(A) > P(B)
  - **Disjointness:** Mutually exclusive terms cannot co-occur. Loss: Penalizes if ∑P(disjoint terms) > 1
  - **A and B Imply C:** If A and B are predicted, C must also be predicted. Penalizes when min(P(A),P(B)) > P(C)
- **Benefits**
  - Ensures predictions are biologically valid.
  - Improves interpretability and performance for complex multi-label tasks.

Evaluation Metrics & Results

Metrics Used:

- **F1-Score, Precision, Recall:** Assess balance between true positives and false positives/negatives.
- **Hamming Loss:** Fraction of incorrectly predicted labels.
- **ROC AUC:** How well the model ranks true positives higher than false positives across all possible thresholds.

MF:

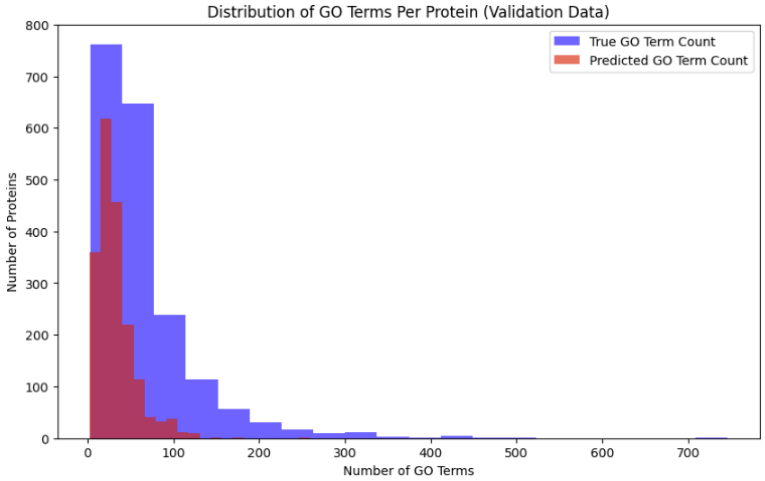
Embedding	Processing	ROC AUC	Hamming Loss	Subset Accuracy	Precision	Recall	F1-score
ESM2	Concatenated	0.974165	0.002054	0.000000	0.563993	0.364484	0.442804
ESM2	Embedding Only	0.981189	0.001925	0.000526	0.625882	0.349409	0.448459
ESM2	Separate Processing	0.965748	0.002060	0.000000	0.571113	0.321793	0.411645
PROTBERT	Concatenated	0.964469	0.002053	0.001052	0.577853	0.309413	0.403025
PROTBERT	Embedding Only	0.979214	0.001967	0.000526	0.622771	0.308872	0.412941
PROTBERT	Separate Processing	0.961134	0.002042	0.000000	0.584215	0.305257	0.400992
PROTT5	Concatenated	0.968845	0.002086	0.000000	0.554528	0.347569	0.427309
PROTT5	Embedding Only	0.981339	0.001933	0.000000	0.634537	0.323173	0.428241
PROTT5	Separate Processing	0.968403	0.002028	0.000000	0.581215	0.338531	0.427855
TAPE	Concatenated	0.969968	0.001968	0.000526	0.673571	0.235151	0.348601
TAPE	Embedding Only	0.969953	0.001999	0.000000	0.644737	0.239242	0.348986
TAPE	Separate Processing	0.955930	0.002042	0.000000	0.593779	0.278996	0.379622

Model is conservative in predicting positives. i.e. predicts fewer GO terms to minimize incorrect predictions

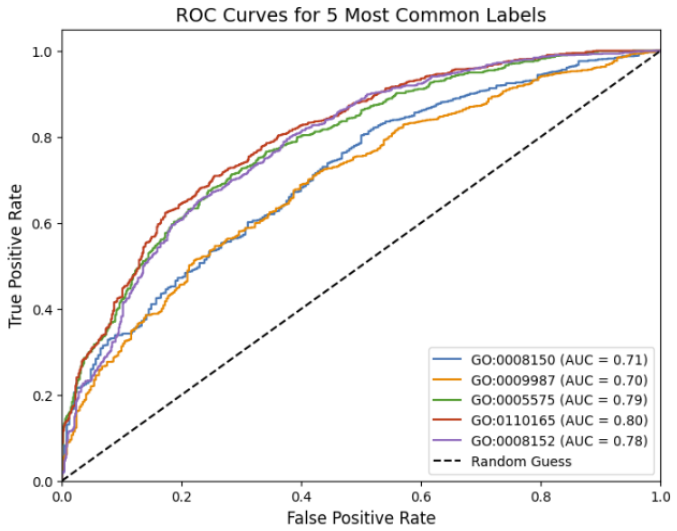
- Avoids false positives (helping Hamming Loss).
- Misses true positives, resulting in lower Recall and Precision.

Results on Validation data (Best performing model)

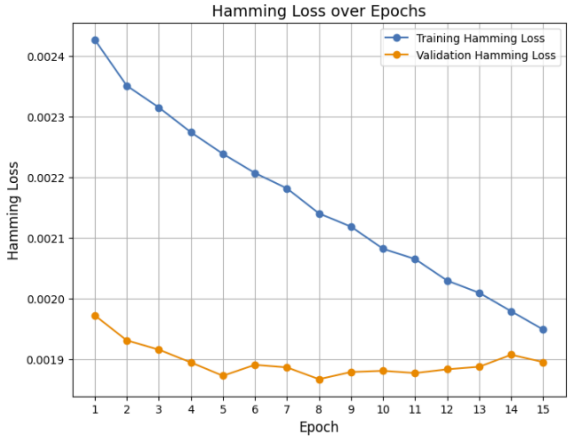
Distribution of True vs Predicted GO terms per protein

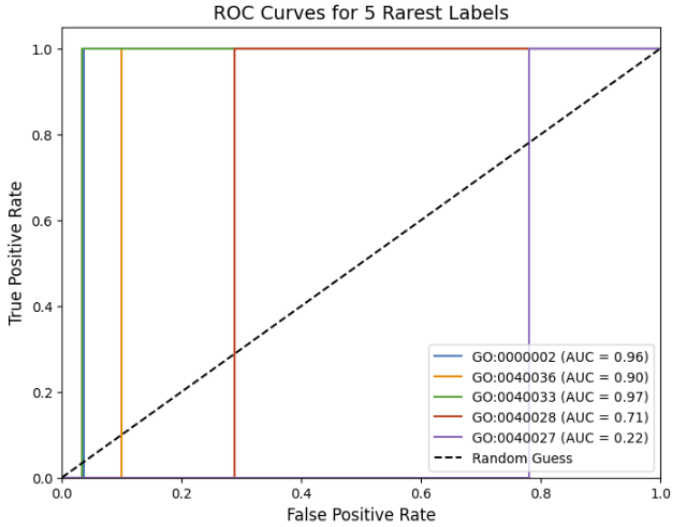


ROC Curves for 5 most common and rarest labels



Hamming Loss over Epochs





Model is conservative in predicting positives

- Avoids false positives (helping Hamming Loss).
- Misses true positives, resulting in lower Recall and Precision.

Conclusion

- ESM2 with embedding only model performed the best based on F1 score
- Logical loss enforcing biological consistency improved predictions
- Achieved high ROC AUC and low Hamming Loss and good F1 score and precision.
- Model was conservative, recall remained lower than desired.
- Predicting exact GO term combinations is challenging.

Possible Improvements

- Use Protein-protein interactions (PPIs) as additional feature
- Use Per-Residue Embeddings
  - Generate embeddings for each amino acid instead of a fixed-size vector.
  - Sequence-based models (e.g., Transformers, RNN) can be used
  - Capture both local features and long-range dependencies in protein sequences.

Q&A

Appendix
Results: Biological Process (BP)

Embedding	Processing	ROC AUC	Hamming Loss	Subset Accuracy	Precision	Recall	F1-score
ESM2	Concatenated	0.977431	0.001944	0.001045	0.587627	0.299012	0.396345
ESM2	Embedding Only	0.981682	0.001896	0.000000	0.608950	0.312121	0.412706
ESM2	Separate Processing	0.968781	0.001982	0.001742	0.571605	0.283828	0.379311
PROTBERT	Concatenated	0.965645	0.002039	0.000697	0.539161	0.307429	0.391580
PROTBERT	Embedding Only	0.979728	0.001917	0.000697	0.637310	0.236405	0.344880
PROTBERT	Separate Processing	0.963647	0.002037	0.001394	0.542610	0.289818	0.377830
PROTT5	Concatenated	0.969801	0.001990	0.000348	0.561381	0.307999	0.397766
PROTT5	Embedding Only	0.981373	0.001929	0.000697	0.593436	0.305783	0.403601
PROTT5	Separate Processing	0.968860	0.002004	0.001045	0.551236	0.327874	0.411179
TAPE	Concatenated	0.973319	0.001931	0.000000	0.697833	0.167874	0.270641
TAPE	Embedding Only	0.970104	0.001951	0.000000	0.652370	0.184105	0.287169
TAPE	Separate Processing	0.962678	0.002021	0.001394	0.565209	0.229036	0.325978

Results: Cellular Component (CC)

Embedding	Processing	ROC AUC	Hamming Loss	Subset Accuracy	Precision	Recall	F1-score
ESM2	Concatenated	0.970639	0.001876	0.012146	0.583501	0.279332	0.377803
ESM2	Embedding Only	0.978053	0.001788	0.019231	0.646021	0.271800	0.382620
ESM2	Separate Processing	0.962261	0.001932	0.010796	0.550706	0.284571	0.375241
PROTBERT	Concatenated	0.953937	0.001905	0.011808	0.569883	0.266987	0.363620
PROTBERT	Embedding Only	0.976463	0.001810	0.017544	0.653400	0.238665	0.349624
PROTBERT	Separate Processing	0.956806	0.001885	0.012146	0.586596	0.255308	0.355772
PROTT5	Concatenated	0.962816	0.001955	0.007422	0.536917	0.296238	0.381814
PROTT5	Embedding Only	0.975998	0.001814	0.016532	0.629905	0.266805	0.374841
PROTT5	Separate Processing	0.959671	0.001892	0.013158	0.575033	0.274706	0.371796
TAPE	Concatenated	0.967131	0.001817	0.006073	0.694298	0.194601	0.303997
TAPE	Embedding Only	0.964226	0.001843	0.004723	0.676111	0.183929	0.289187
TAPE	Separate Processing	0.951920	0.001900	0.008097	0.596590	0.209765	0.310393