HATAX-ESM: HIERARCHICAL ATTENTION-BASED PHYLOGENETIC CLASSIFICATION OF PROTEINS

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Understanding the evolutionary relationships between protein sequences is crucial for phylogenetic classification, mutation prediction, and functional annotation. The NCBI taxonomy database (Sayers et al., 2022) contains over one million distinct taxa, but many classes have very few representative sequences, creating extreme class imbalance. Traditional sequence similarity-based methods like BLAST are widely used for taxonomic classification but are computationally expensive and ineffective for de novo sequences without close homologs.

Recent deep learning approaches, such as PhyloTransformer (Wu et al., 2021) and TEMPO (Zhou et al., 2023), have leveraged transformer-based architectures for phylogenetic tasks. However, these models do not impose explicit hierarchical constraints, limiting their ability to ensure phylogenetic consistency. Inspired by phylogenetic tree-guided learning (Chandar et al., 2024; Yax et al., 2024), we introduce a model that combines a frozen ESM feature extractor with attention pooling and hierarchical softmax-based classification.

HATax-ESM improves computational efficiency while enforcing structured taxonomic predictions. By conditioning classification probabilities at each level, our model ensures predictions follow valid phylogenetic lineages, making it a robust alternative to traditional similarity-based methods. This structured approach allows for better generalization, particularly for underrepresented taxa, enhancing protein classification and evolutionary inference.

2 METHODS

2.1 DATASET PREPARATION AND PHYLOGENETIC SAMPLING

We curated a data set of 180 million UniProt protein sequences, filtering for sequences less than or equal to 500 amino acids. A total of 20% of the data was reserved for testing, ensuring a balanced split across taxonomic levels. Inspired by phylogenetic tree-based sampling methods (Zhou et al., 2023), we selected test samples by starting from the leaf nodes and moving up the tree, maintaining a consistent 20% representation per parent node at each hierarchical level. This ensured that each parent node had representation in the test set while preventing over-sampling.

For validation, we selected 940 representative sequences, ensuring that at least one child per parent node was included for the first six levels of the phylogenetic tree. This hierarchical sampling approach was motivated by previous studies in phylogenetic-guided deep learning (Wu et al., 2021; Chandar et al., 2024) and ensures an informative subset for evaluation.

2.2 MODEL ARCHITECTURE

The model consists of three primary components:

ESM Feature Extractor: We used the ESM2-t33-650M-UR50D model (Rives et al., 2021), a frozen transformer-based model pretrained on large protein sequence datasets, to extract meaningful embeddings.

Attention Pooling Layer: Given that protein sequences vary in length, we applied scaled dot-product attention (Vaswani et al., 2023) to produce a fixed-dimensional representation for downstream classification.

Hierarchical Softmax Classifier: The final classification layer implements a hierarchical softmax, where the probability distribution at each level is conditioned on the previous level. This approach improves efficiency and enforces phylogenetic structure (Mohammed & Umaashankar, 2018).

Unlike PhyloTransformer (Wu et al., 2021), which employs a self-attention mechanism for mutation prediction, our model explicitly encodes phylogenetic constraints into the classification process. Hierarchical softmax has been successfully applied in language models (Mohammed & Umaashankar, 2018), and our work extends this paradigm to protein classification.

3 RESULTS

After training on 9 million sequences, our model achieved an average classification accuracy of 35% across the validation set. The accuracy was computed based on correctly predicted tree nodes over the total number of nodes in the ground-truth tree. If the model predicted a child node not belonging to a valid lineage, inference was terminated at that point. This evaluation strategy aligns with methodologies used in phylogenetic-informed AI models (Wu et al., 2021; Yax et al., 2024).

An overview of the dataset structure, model architecture, and training performance is provided in **Appendix B**. The dataset follows a hierarchical phylogenetic structure spanning 36 levels, with Cellular Organisms and Viruses as the two primary top-level categories. The class distribution across levels shows that Cellular Organisms exhibit a broader taxonomic diversity compared to Viruses. Training performance trends indicate a steady improvement in classification accuracy as the model processes more sequences.

To assess the difficulty of the classification task, we computed the expected accuracy under andom guessing. The expected full-lineage accuracy—the probability of randomly predicting an entire lineage correctly—was found to be 5.6×10^{-5}. The expected fraction of correct nodes per lineage, which represents the probability of correctly guessing an individual node at any level, was 2.98%. The formulation for these calculations is provided in **Appendix A**, where we describe the probabilistic approach used to estimate these values.

Our experiments were conducted on two AMD MI-250 GPUs, leveraging mixed-precision and distributed training with PyTorch Lightning. The model consisted of 2.5 billion parameters, with 1.8 billion trainable parameters coming from the hierarchical classifiers.

4 DISCUSSION AND FUTURE WORK

The results indicate that even with limited training data (˜10% of total sequences), the model can learn meaningful phylogenetic relationships. This highlights the utility of hierarchical softmax for structured classification problems in biology. Compared to PhyloTransformer (Wu et al., 2021) and TEMPO (Zhou et al., 2023), which focus on mutation prediction, our approach excels in taxonomy-aware classification, ensuring that predictions remain biologically interpretable.

Beyond classification, our model has significant applications in functional annotation. Many newly discovered protein sequences lack well-defined taxonomic labels, making annotation difficult. By providing a structured prediction framework, our model can infer evolutionary relationships and assign meaningful classifications to previously uncharacterized proteins. Additionally, the model's hierarchical nature makes it compatible with generative models, offering a means of enforcing taxonomic constraints in protein design. For instance, a generative model could sample new protein sequences while ensuring that their evolutionary placement remains consistent with known biological taxonomies.

Future work will extend training to the entire dataset, exploring methods to further refine hierarchical modeling. Additional improvements include integrating contrastive learning techniques for enhanced representation learning and adapting the model to predict functional annotations alongside taxonomic labels. Moreover, we aim to explore the integration of the classifier with generative protein sequence models, allowing evolutionary constraints to guide sequence generation in protein design applications. This model serves as a crucial evaluation step for generative models, ensuring that the generated protein sequences align with the intended taxonomy and maintain the correct phylogenetic structure.

MEANINGFULNESS STATEMENT

Language serves as a fundamental tool for mapping concepts to words in our daily lives. Similarly, in the realm of biological sciences, proteins can be understood as a "language" of life, encoding functional and structural information. This work explores how deep learning models can bridge these two representations by leveraging protein sequences to generate meaningful taxonomic lineage mappings. By doing so, we aim to enhance our understanding of the underlying biological relationships and contribute to the broader effort of translating complex biological data into interpretable and structured knowledge.

REFERENCES

Vijay Chandar, Vishnu G Nair, and Rajeswari D. Phylogenetic-informed generative adversarial network for predicting mutations in sars-cov-2. In *2024 2nd International Conference on Networking and Communications (ICNWC)*, pp. 1–7, 2024. doi: 10.1109/ICNWC60771.2024.10537523.

Abdul Arfat Mohammed and Venkatesh Umaashankar. Effectiveness of hierarchical softmax in large scale classification tasks. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1090–1094. IEEE, September 2018. doi: 10.1109/icacci.2018.8554637. URL http://dx.doi.org/10.1109/ICACCI.2018.8554637.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jocelyn Chan, Donald C. Comeau, Rory Connor, Karen Funk, Cole Kelly, Scott Kim, Tom Madej, Aron Marchler-Bauer, Christopher J. Lanczycki, Steven Lathrop, Zhiyong Lu, Francine Thibaud-Nissen, Thomas Murphy, Loc Phan, Yuliya Skripchenko, Tanya Tse, Jian Wang, Robin Williams, Ben W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D20–D26, 2022. doi: 10.1093/nar/gkab1112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Yingying Wu, Shusheng Xu, Shing-Tung Yau, and Yi Wu. Phylotransformer: A discriminative model for mutation prediction based on a multi-head self-attention mechanism, 2021. URL https://arxiv.org/abs/2111.01969.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. Phylolm : Inferring the phylogeny of large language models and predicting their performances in benchmarks, 2024. URL https://arxiv.org/abs/2404.04671.

Binbin Zhou, Hang Zhou, Xue Zhang, Xiaobin Xu, Yi Chai, Zengwei Zheng, Alex Chichung Kot, and Zhan Zhou. Tempo: A transformer-based mutation prediction framework for sars-cov-2 evolution. *Computers in Biology and Medicine*, 152:106264, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.106264. URL https://www.sciencedirect.com/science/article/pii/S0010482522009726.

A ACCURACY CALCULATIONS

A.1 EXPECTED ACCURACY UNDER RANDOM GUESSING

To evaluate the difficulty of the phylogenetic classification task, we computed the expected accuracy assuming random selection of nodes at each level. Given that each taxonomic level contains a varying number of possible children per parent node, the probability of correctly guessing a lineage depends on the branching structure of the taxonomy.

Let C_ℓ represent the number of possible child classes at level ℓ. The probability of correctly selecting the true lineage at each level is given by:

$$P_{\text{correct},\ell} = \frac{1}{C_\ell} \tag{1}$$

Since classification involves making predictions across multiple levels of the hierarchy, the probability of randomly selecting the entire correct lineage is computed as the product of the independent probabilities at each level:

$$P_{\text{full-lineage}} = \prod_{\ell=1}^{L} P_{\text{correct},\ell} = \prod_{\ell=1}^{L} \frac{1}{C_\ell} \tag{2}$$

where L is the total depth of the lineage. Using this approach, we estimated:

- **Expected full-lineage accuracy (random guessing):** 5.677858×10^{-5}
- **Expected fraction of correctly predicted nodes in a lineage:** 2.98%

These values indicate that random guessing would result in an extremely low probability of correctly classifying an entire lineage, reinforcing the difficulty of the task.

A.2 Validation Accuracy Calculation

In contrast to the expected accuracy under random guessing, the validation accuracy of our model is computed by measuring the fraction of correctly predicted nodes in a lineage. Given a ground truth taxonomy and a predicted taxonomy, we define accuracy as:

$$\text{Validation Accuracy} = \frac{\text{Number of correctly predicted nodes}}{\text{Total number of nodes in ground truth lineage}} \tag{3}$$

For example, consider the following ground truth and predicted lineages:

- **Ground truth:** Cellular Organisms, Eukaryota, Opisthokonta, Fungi, Dikarya, Basidiomycota, Agaricomycotina, Agaricomycetes, Agaricomycetes Incertae Sedis, Cantharellales, Ceratobasidiaceae, Ceratobasidium, Unclassified Ceratobasidium, Ceratobasidium sp. AG-Ba.
- **Predicted:** Cellular Organisms, Eukaryota, Opisthokonta, Fungi, Dikarya, Ascomycota.

Here, the model correctly predicts 5 out of 14 nodes (*up to Dikarya*), but incorrectly classifies the next taxonomic level. Using the formula:

$$\text{Accuracy} = \frac{5}{14} = 0.3571 \tag{4}$$

This demonstrates how validation accuracy is computed, ensuring that only the correctly predicted hierarchical nodes contribute to the accuracy measure.

B Model Architecture and Training Performance

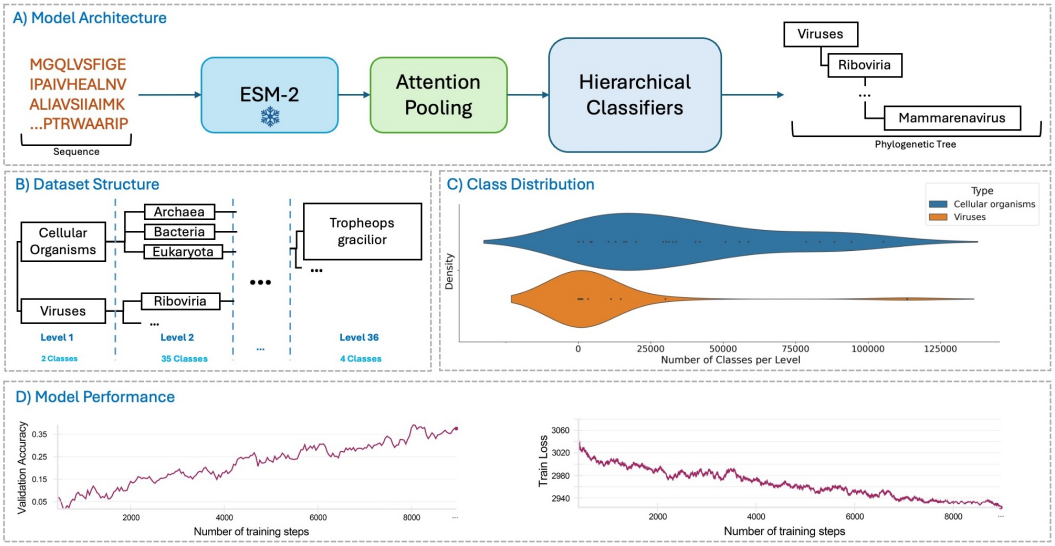

Figure 1: Overview of the hierarchical phylogenetic classification model. **(A) Model Architecture:** Protein sequences are first processed using a frozen ESM-2 model to generate embeddings, which are then refined through attention pooling before being passed into hierarchical classifiers for taxonomic prediction. **(B) Dataset Structure:** The dataset follows a hierarchical phylogenetic tree structure with a total of 36 levels, where Cellular Organisms and Viruses form the two major top-level categories. Lower levels capture finer taxonomic classifications down to species. **(C) Class Distribution:** The number of unique classes per level for Cellular Organisms and Viruses. Cellular organisms exhibit a much broader class diversity at each level compared to viruses. **(D) Model Performance:** Validation accuracy and training loss across the training steps. The validation accuracy steadily improves with training, reaching 35% after processing 9 million records, while the training loss consistently decreases, indicating effective learning.