The Search for Phylogenetic-Aware Embeddings

Mukhilshankar Umashankar
mu9@illinois.edu
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA

Mohsin Ansari
mansari5@illinois.edu
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

Phylogenetic trees are essential tools for understanding evolutionary relationships among proteins. Traditionally, these trees are constructed using multiple sequence alignment (MSA) tools such as Clustal Omega and MAFFT, followed by distance-based methods like neighbor joining. However, MSA methods are computationally expensive and scale poorly for large datasets. This paper explores the use of protein embeddings derived from pre-trained models, namely ESM-2, as an alternative approach to infer phylogenetic relationships. Our results demonstrate that while raw embeddings capture some evolutionary signals, they fail to replicate the accuracy of MSA-based methods. To address this, we propose a neural network-based refinement process that creates phylogenetic-aware embeddings. We then align the embedding-based trees with ground truth MSA-based trees, achieving significant improvements. This work highlights the promise of embedding-based methods for scalable phylogenetic analysis.

ACM Reference Format:
Mukhilshankar Umashankar and Mohsin Ansari. 2024. The Search for Phylogenetic-Aware Embeddings. In . ACM, New York, NY, USA, 14 pages.

1 INTRODUCTION

Phylogenetic trees provide insights into the evolutionary relationships among proteins, genes, and species. Traditional methods rely on multiple sequence alignment (MSA) followed by tree inference techniques such as neighbor joining. However, as the size of protein datasets grow, these approaches become computationally prohibitive, particularly for large-scale analyses like metagenomics or biodiversity studies.

Recent advances in protein sequence embeddings, such as those generated by Evolutionary Scale Modeling (ESM) [1], offer an alternative pathway. These embeddings encode rich sequence-level features and are computationally efficient. However, their ability to accurately capture evolutionary relationships and replicate phylogenetic tree topologies has not been thoroughly investigated.

This study compares embedding-based phylogenetic trees of two oncological-relevant protein families, Ras and TP53, with those generated using traditional MSA methods. Furthermore, we propose a framework to refine protein embeddings in *phylogenetic-aware embeddings* that better align with ground truth trees based on MSA. This approach aims to combine the computational efficiency of embeddings with the evolutionary accuracy of traditional methods.

2 METHODS

2.1 Traditional Phylogenetic Tree Construction

The relationship between multiple sequence alignment (MSA) and phylogenetic tree construction is foundational to evolutionary biology. MSA aligns homologous sequences by arranging residues into columns, accounting for insertions, deletions, and substitutions. This alignment forms the basis for calculating evolutionary distances, as conserved regions and observed substitutions between sequences provide insights into their relationships. From this alignment, a distance matrix is constructed to quantify the evolutionary divergence between sequence pairs. Phylogenetic trees are then inferred using hierarchical distance-based clustering methods, such as neighbor joining, which clusters sequences based on their calculated evolutionary distances.

In this study, we use Clustal Omega [3] and MAFFT [4] as the tools for generating MSA-based phylogenetic trees as candidates to serve as the ground truth for comparison with the embedding technique we introduce here. These tools were chosen for their widespread use, robust algorithms, and complementary strengths. Both Clustal Omega and MAFFT are compatible with the BLOSUM62 substitution matrix [5], a scoring system derived from observed amino acid substitutions in aligned protein blocks. We transform these similarity

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

University of Illinois at Urbana-Champaign, September 2024, Urbana, Illinois, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

scores into distance values, generating a matrix that encodes biologically meaningful relationships while ensuring the alignments accurately reflect evolutionary connections.

Clustal Omega employs a progressive alignment strategy, which constructs a guide tree to determine the order in which sequences are aligned. Its scalability makes it particularly well-suited for large datasets, aligning thousands of sequences efficiently while maintaining biological accuracy. In contrast, MAFFT introduces computational innovations such as fast Fourier transform (FFT), which maps sequences into a numerical property space to rapidly calculate pairwise similarities. MAFFT also offers iterative refinement options, such as L-INS-i and G-INS-i modes, which enhance alignment accuracy for datasets with complex evolutionary relationships.

Despite their different algorithmic approaches, Clustal Omega and MAFFT often produce consistent phylogenetic trees under many conditions. For our analysis, these tools were used independently to generate MSAs, from which we computed pairwise distance matrices and constructed phylogenetic trees using the Neighbor Joining algorithm [6]. These trees serve as the baseline against which we compare embedding-based approaches. The consistency between results from these established tools helps validate the reliability of our ground truth phylogenetic trees.

To quantify the consistency between these methods, we analyzed the Ras protein family using a set of 20 sequences retrieved from UniProt, the comprehensive protein sequence database [7]. The Frobenius norm of the difference between the distance matrices generated by MAFFT and Clustal Omega was 1.1933, representing approximately 217% of the mean pairwise distance between sequences. While this indicates notable differences between the methods, visual inspection of the resulting trees in Figure 1 and Figure 5 reveals that the core evolutionary relationships are preserved. Based on its established track record in phylogenetic analysis, we selected Clustal Omega-derived trees as our ground truth for subsequent comparisons. With Clustal Omega established as our ground truth method, we investigate how embedding-based approaches compare in analyzing evolutionary relationships.

2.2 Protein Language Models

The emergence of protein language models (PLMs) represents a significant advancement in computational biology, applying transformer-based deep learning architectures to protein sequence analysis. These models, trained on vast databases of protein sequences, learn to capture complex patterns and relationships in amino acid sequences without requiring explicit structural alignment. PLMs process protein sequences as "sentences" of amino acids, learning contextual relationships between residues through self-attention mechanisms and masked prediction tasks.

Unlike traditional sequence analysis methods that rely on pairwise alignments, PLMs encode proteins into high-dimensional vector spaces where evolutionary and functional relationships may be captured through geometric relationships. These models learn from the co-evolution patterns present in large protein sequence databases, potentially capturing both local and long-range interactions that might be missed by traditional sequence alignment approaches.

ESM (Evolutionary Scale Modeling), introduced by Meta AI Research [1], marked a significant advancement in protein language modeling. The original ESM model demonstrated that transformer architectures trained on protein sequences could learn meaningful biological properties and evolutionary relationships. ESM was trained using masked language modeling on UniRef50, learning to predict randomly masked amino acids based on their sequence context.

ESM-2 [2], the successor model, incorporates several architectural improvements and was trained on a significantly larger dataset. Trained on over 65 million protein sequences from UniRef50, ESM-2 uses a deeper transformer architecture (we use the smaller 12-layer variant with 35 million parameters in this study) and employs improved tokenization and training strategies. The model demonstrates enhanced performance across various protein modeling tasks, including structure prediction, mutation effect prediction, and sequence family classification [8].

2.3 ESM-2 Embeddings

For our analysis, we utilize the 12-layer, 35M-parameter variant of ESM-2, trained on over 65 million protein sequences from UniRef50, to generate embeddings. We generate embeddings for the set of 20 Ras protein sequences previously analyzed using traditional MSA approaches. ESM-2 employs a transformer architecture to learn complex patterns in protein sequences. We use these learned representations to generate fixed-dimensional embeddings that capture evolutionary and functional relationships.

Specifically, our implementation extracts embeddings from layer 12 of the model, computing the mean over all sequence token embeddings while excluding special tokens. Mean pooling is chosen as it aggregates information across all sequence tokens, providing a balanced representation of local and global features while mitigating potential biases introduced by special tokens. To construct phylogenetic trees from these embeddings, we compute the distance matrix from the pairwise Euclidean distances between the sequence vectors and apply the Neighbor Joining algorithm, enabling direct comparison with our MSA-based approaches.

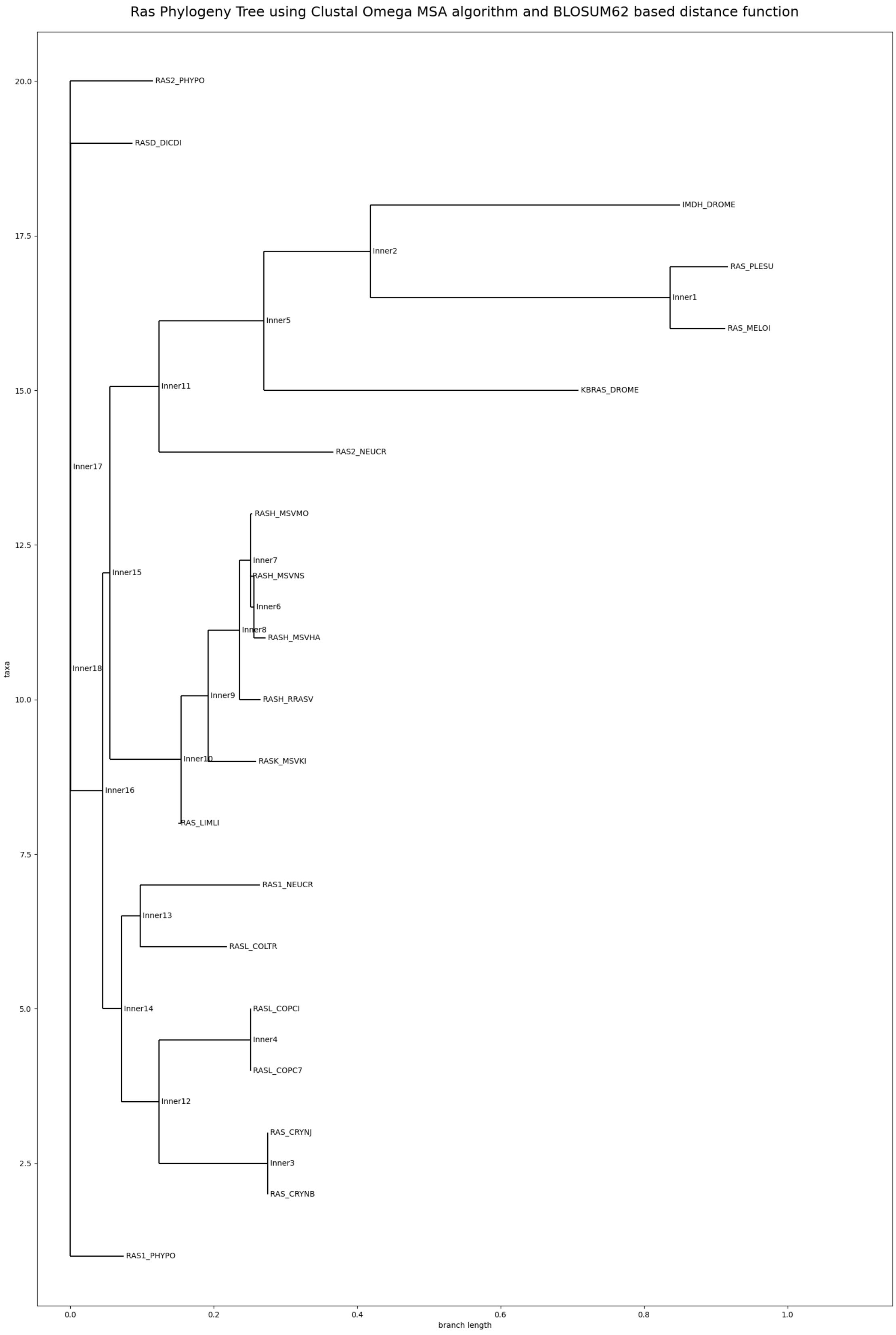

Figure 1: Ras phylogenetic Tree based on ClustalOmega.

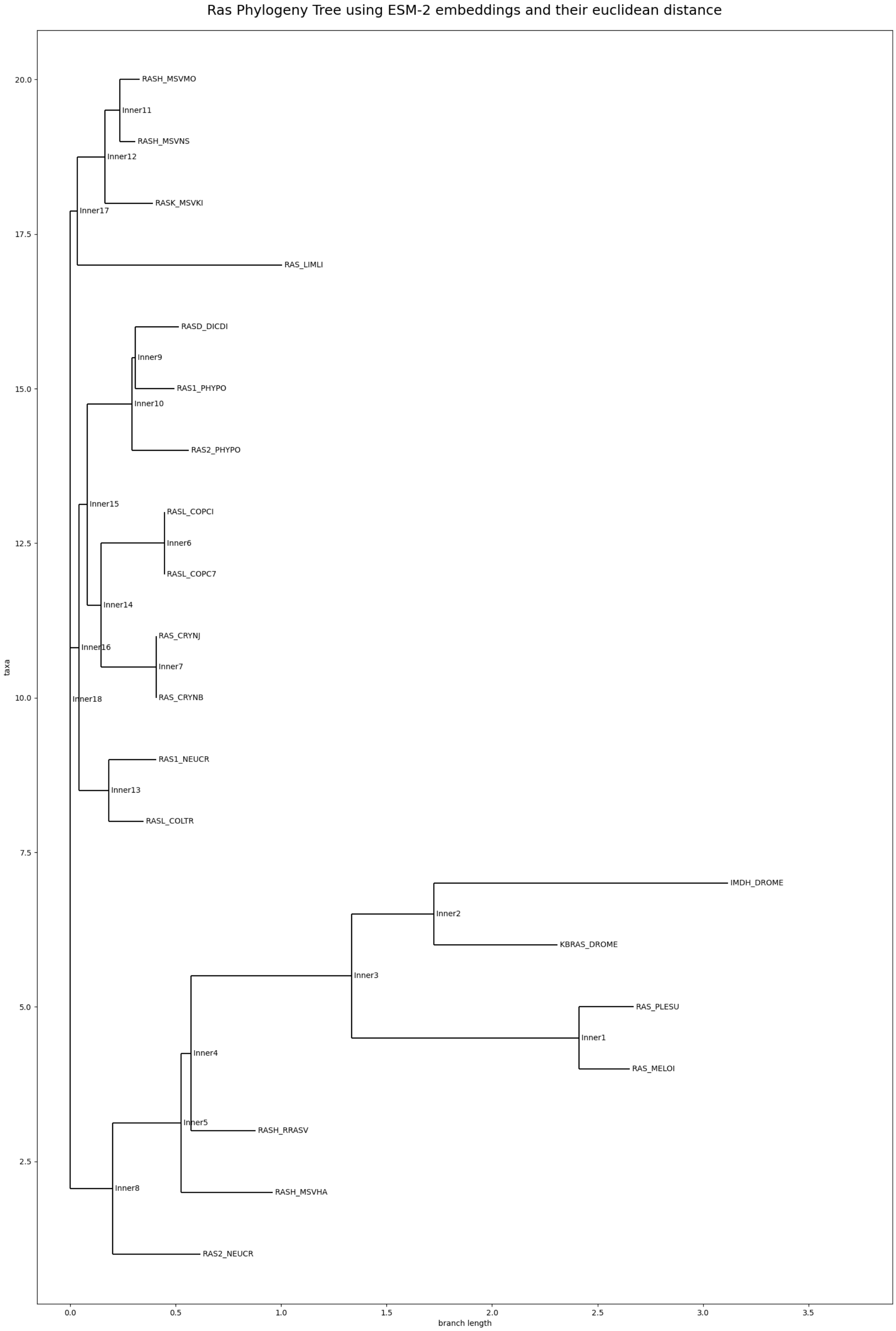

Figure 2: Ras phylogenetic Tree based on ESM-2 Embeddings.

This embedding-based approach offers a fundamentally different perspective on sequence relationships compared to traditional MSA methods. Unlike MSA, which explicitly aligns sequences to identify evolutionary relationships, embeddings encode these relationships implicitly, potentially capturing non-linear patterns and higher-order interactions that may be overlooked in traditional alignment-based approaches. The resulting initial ESM-2 embedding phylogenetic tree is shown in Figure 2.

2.4 Comparative Analysis of Distance Matrices

Our analysis of the Ras protein family (n=20) reveals striking differences in how sequence relationships are captured by traditional MSA-based methods versus embedding-based approaches. While MAFFT and Clustal Omega show relatively consistent results (Frobenius norm 216.82% of mean pairwise distance), their comparison with ESM-2 embeddings reveals dramatic differences in the encoded relationships (Table 1).

Table 1: Comparison of Distance Matrices Across Methods for Ras phylogenetic tree construction

Method Comparison	Mean Pairwise Distance	Frobenius Norm	Relative Frobenius (%)
MAFFT vs. Clustal Omega	0.5504	1.1933	216.82
MAFFT vs. ESM-2	1.1226	28.1518	2507.63
Clustal Omega vs. ESM-2	1.1427	27.1628	2376.99

The similarity between MAFFT and Clustal Omega is reflected in their mean pairwise distance (0.5504) and moderate Frobenius norm (1.1933). However, when comparing either MSA-based method to ESM-2, we observe Frobenius norms that are more than 20 times larger. Notably, while the mean pairwise distances remain relatively stable across all methods (0.5504-1.1427), the large Frobenius norms in ESM-2 comparisons (>2300% of mean pairwise distance) indicate fundamental differences in how specific sequence relationships are encoded.

These results suggest that while traditional MSA-based approaches and embedding-based methods may capture similar overall evolutionary distances, they differ substantially in their representation of specific sequence relationships. This divergence may reflect the fundamentally different approaches to sequence analysis: MSA methods explicitly align sequences to identify evolutionary relationships, while ESM-2 embeddings capture these relationships implicitly through learned representations of protein sequence space.

These quantitative differences are further reflected in the phylogenetic trees generated from each method. The ESM-2-based tree exhibits markedly different topological structure from both MSA-based trees, with notably longer branch lengths that suggest greater inferred evolutionary distances between sequences. These substantial differences in both the distance matrices and the resulting tree structures, while highlighting ESM-2's distinct perspective on sequence relationships, motivates us to explore whether we could bridge this gap. Specifically, we investigate whether fine-tuning the embeddings could bring the ESM-2 representations into closer alignment with the established phylogenetic relationships captured by traditional MSA-based approaches.

2.5 Fine-Tuning ESM-2 Embeddings for constructing phylogenetic tree

While ESM-2 embeddings capture rich sequence-level features, they often fail to explicitly reflect evolutionary relationships, leading to discrepancies when used for phylogenetic analysis. To address this issue, we developed a fine-tuning approach that adjusts the embeddings to make them more "phylogenetic-aware." Our method employs a single-layer neural network designed to transform the 480-dimensional ESM-2 embeddings while preserving their dimensionality. The goal is to align the pairwise Euclidean distances in the embedding space with the evolutionary distances captured by Clustal Omega alignments.

The network architecture consists of a linear transformation followed by a ReLU activation function, expressed as:

$$f(x) = \text{ReLU}(Wx)$$

where x represents the original 480-dimensional ESM-2 embedding and W is a learned 480×480 weight matrix. This simple architecture was chosen to avoid overfitting while allowing meaningful transformations of the embedding space.

Training optimizes the transformation to minimize the L1 loss between the pairwise Euclidean distance matrix of the transformed embeddings and the distance matrix resulting from Clustal Omega:

$$Loss = \sum_{i,j} \left| d_{ij}^{\text{Transformed Embeddings}} - d_{ij}^{Clustal} \right|$$

where $d_{ij}^{Clustal}$ represents the evolutionary distance between sequences i and j in the Clustal Omega distance matrix and $d_{ij}^{\text{Transformed Embeddings}}$ represents the evolutionary distance between sequences i and j in the ESM-2 transformed embedding distance matrix. We use the Adam optimizer with a learning rate of 1×10^{-5}, training the model for 3,000 epochs. During training, we monitor the Frobenius norm between the original and transformed embeddings, as well as the norm between their respective distance matrices, to evaluate the magnitude of change in the learned transformation and its effect on the alignment of the distance matrix. Ideally, we want to avoid excessively altering the original ESM-2 embeddings, as this could diminish other meaningful features they capture. A balance should be struck between enhancing phylogenetic-awareness and preserving other important features.

This fine-tuning process adjusts the embedding space to better reflect phylogenetic relationships while preserving the rich sequence-level features learned by ESM-2. The transformed embeddings enable the construction of phylogenetic trees that align more closely with traditional MSA-based approaches, bridging the gap between deep learning representations and classical evolutionary biology.

3 RESULTS AND CONCLUSION

In the interest of space, we display the resulting phylogenetic trees for the TP53 protein family in the Appendix. To conserve space, we also reuse Figure 1 to illustrate the Ras phylogenetic tree produced by the transformed embeddings as the two trees were perfectly identical. This is a result of the large number of epochs and overfitting.

Figure 3 captures the degree of overfitting. We see that at around 500 epochs the Frobenius norm between the derived distance matrices of the transformed embeddings and the Clustal Omega algorithm is about 1.0. This is already lower than the Frobenius norm of the distance matrices derived from MAFFT and the Clustal Omega algorithm.

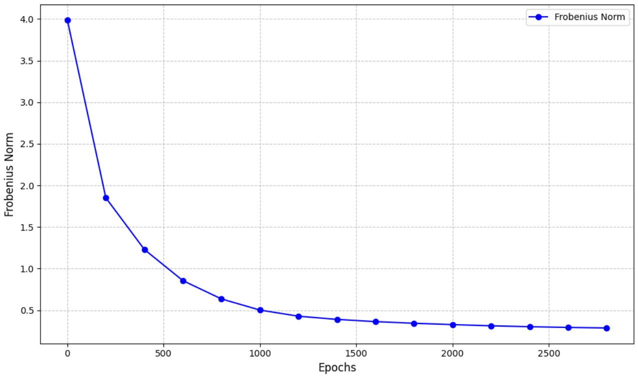

Figure 3: Change over epochs of the Frobenius norm difference between the transformed embedding distance matrix and the Clustal Omega distance matrix for Ras

While overfitting could imply the loss of representation of other features in the original ESM-2 embeddings, the utility of the transformed embeddings is dependent on the downstream task it will be used for. If the downstream task requires highly-sensitive phylogenetic-aware embeddings, but is indifferent to other representations, then overfitting is beneficial. However, if the downstream task requires a balanced representation of multiple features while also being phylogenetic-aware then the transformed embeddings should not be extensively trained.

The degree of training can be informed by Figure 4.

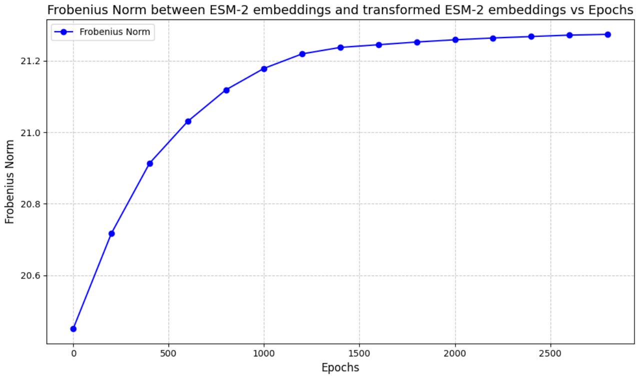

Figure 4: Frobenius norm between ESM-2 embeddings and transformed ESM-2 embeddings vs. Epochs for Ras

Figure 4 captures the magnitude of change between the original ESM-2 embeddings and the transformed embeddings. Training for fewer epochs, will preserve the original representations more.

The balance between the original feature representation and the phylogenetic-aware representation can be coded into the loss function. In this way, the balance can be optimized for.

4 FUTURE WORK

This study demonstrates the potential of fine-tuned ESM-2 embeddings to align with traditional phylogenetic relationships, offering a scalable alternative to MSA-based methods. However, several limitations and opportunities for further exploration remain.

First, our results revealed overfitting due to the lack of regularization during training. Without constraints, the model excessively deviated from the original embedding space, potentially losing some of the rich sequence-level features encoded by the pre-trained ESM-2 model. Future work could address this by incorporating regularization techniques, such as L2 regularization or embedding preservation loss terms, to penalize large deviations from the original embedding distances. These techniques would enable the model to balance preserving original features while aligning with the Clustal Omega distance matrix ground truth.

Second, while this study focused on phylogenetic tree construction, future work should explore the utility of these embeddings for downstream tasks where both evolutionary and functional relationships are critical. Tasks such as protein function prediction, binding site identification, or structure prediction could benefit from embeddings fine-tuned to capture phylogenetic relationships. Additionally, the choice of loss functions could be tailored to specific tasks, such as triplet loss for clustering proteins based on function or contrastive loss for distinguishing evolutionary clades.

Third, the architecture employed in this study—a single-layer neural network—was intentionally simple to minimize complexity and computational overhead. However, this simplicity may have limited the model's capacity to capture non-linear relationships in the embedding space. Future research could explore more advanced architectures, such as multi-layer neural networks, attention-based models, or graph neural networks, to better capture higher-order relationships. Parameter-efficient fine-tuning approaches, such as adapters or fine-tuning specific layers of the ESM-2 model, could also be investigated to address scalability concerns for large datasets.

Finally, future studies could explore methods to evaluate the generalizability of these embeddings across diverse datasets and protein families. By testing the embeddings on unseen evolutionary clades or incorporating more diverse sequence data, researchers can better understand their robustness and potential limitations. These efforts would extend the applicability of fine-tuned embeddings to a broader range of biological problems.

By addressing these challenges, future work can mitigate overfitting, enhance generalizability, and unlock the full potential of embedding-based approaches for protein sequence analysis.

5 CODE AVAILABILITY

You can find the code for this project here: Github link

REFERENCES

[1] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021. https://doi.org/10.1073/pnas.2016239118

[2] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022, p. 500902. https://doi.org/10.1101/2022.07.20.500902

[3] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, no. 1, p. 539, 2011. https://doi.org/10.1038/msb.2011.75

[4] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, pp. 3059–3066, 2002. https://doi.org/10.1093/nar/gkf436

[5] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992. https://doi.org/10.1073/pnas.89.22.10915

[6] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987. https://doi.org/10.1093/oxfordjournals.molbev.a040454

[7] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, Database issue, pp. D115–D119, 2004. https://doi.org/10.1093/nar/gkh131

[8] D. Chen, P. Hartout, P. Pellizzoni, C. Oliver, and K. Borgwardt, "Endowing protein language models with structural knowledge," *arXiv*, arXiv:2401.14819v1 [q-bio.QM], Jan. 26, 2024. https://arxiv.org/abs/2401.14819

6 APPENDIX

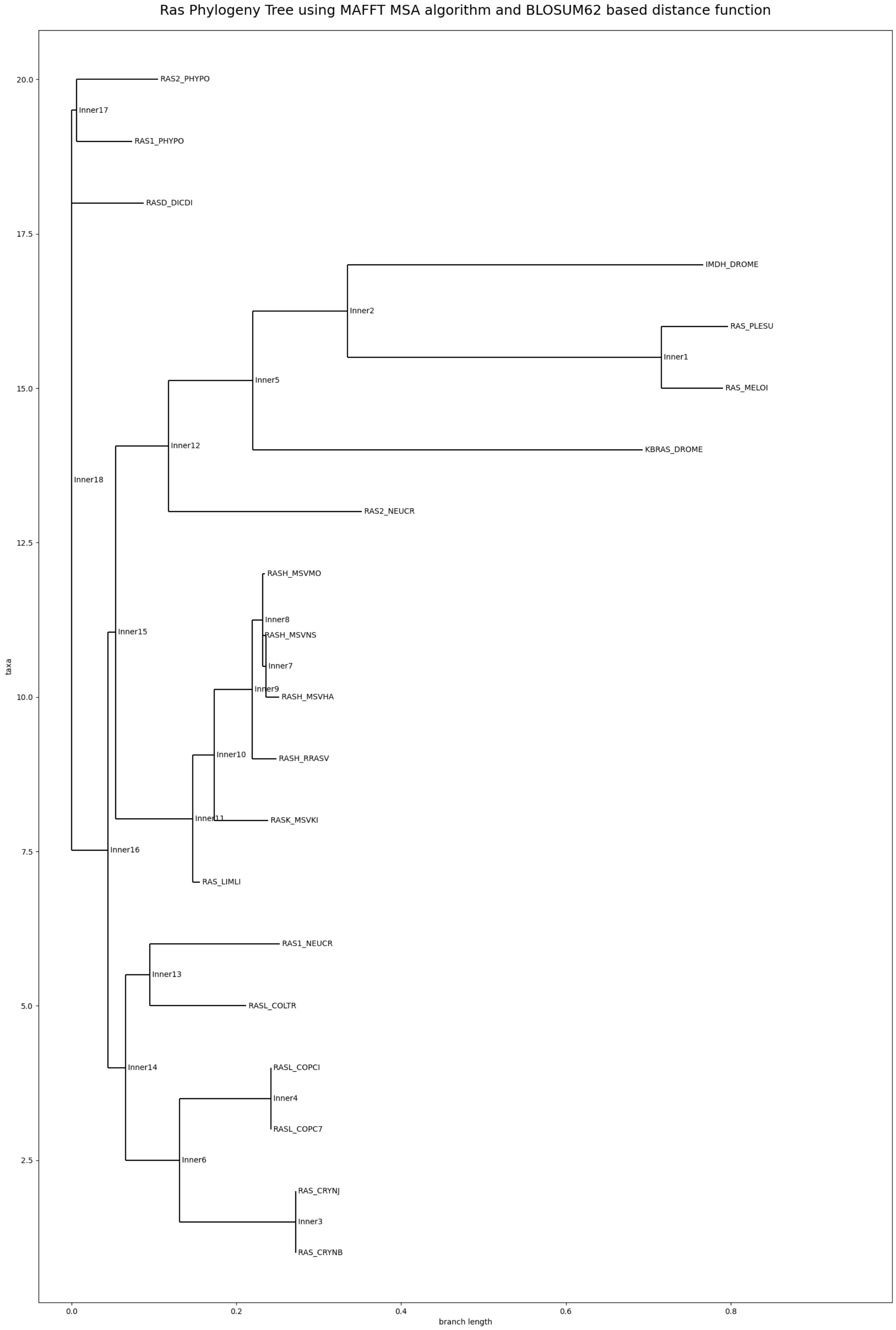

Figure 5: Ras phylogenetic Tree based on MAFFT.

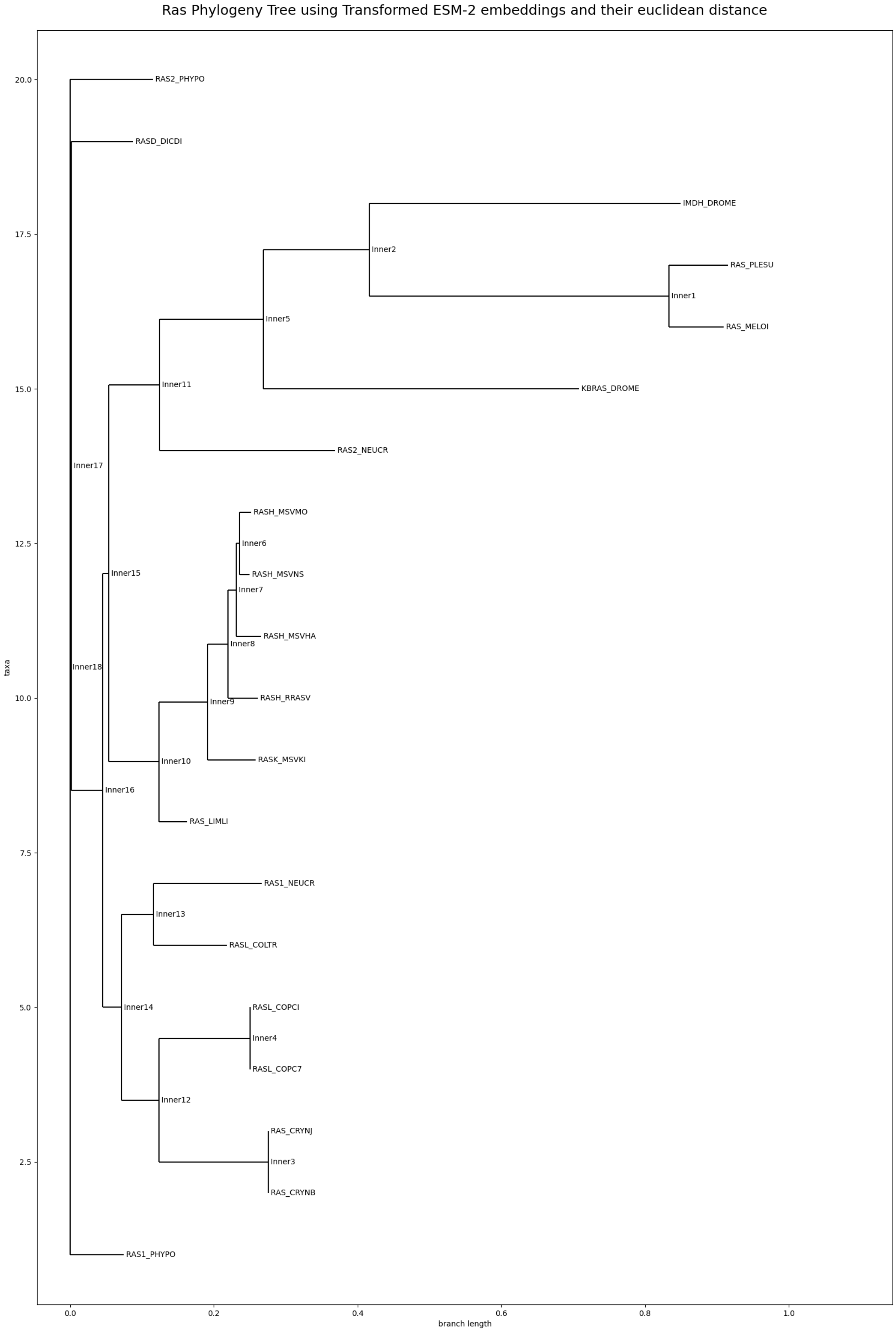

Figure 6: Ras phylogenetic Tree based on fine-tuned ESM-2 Embeddings

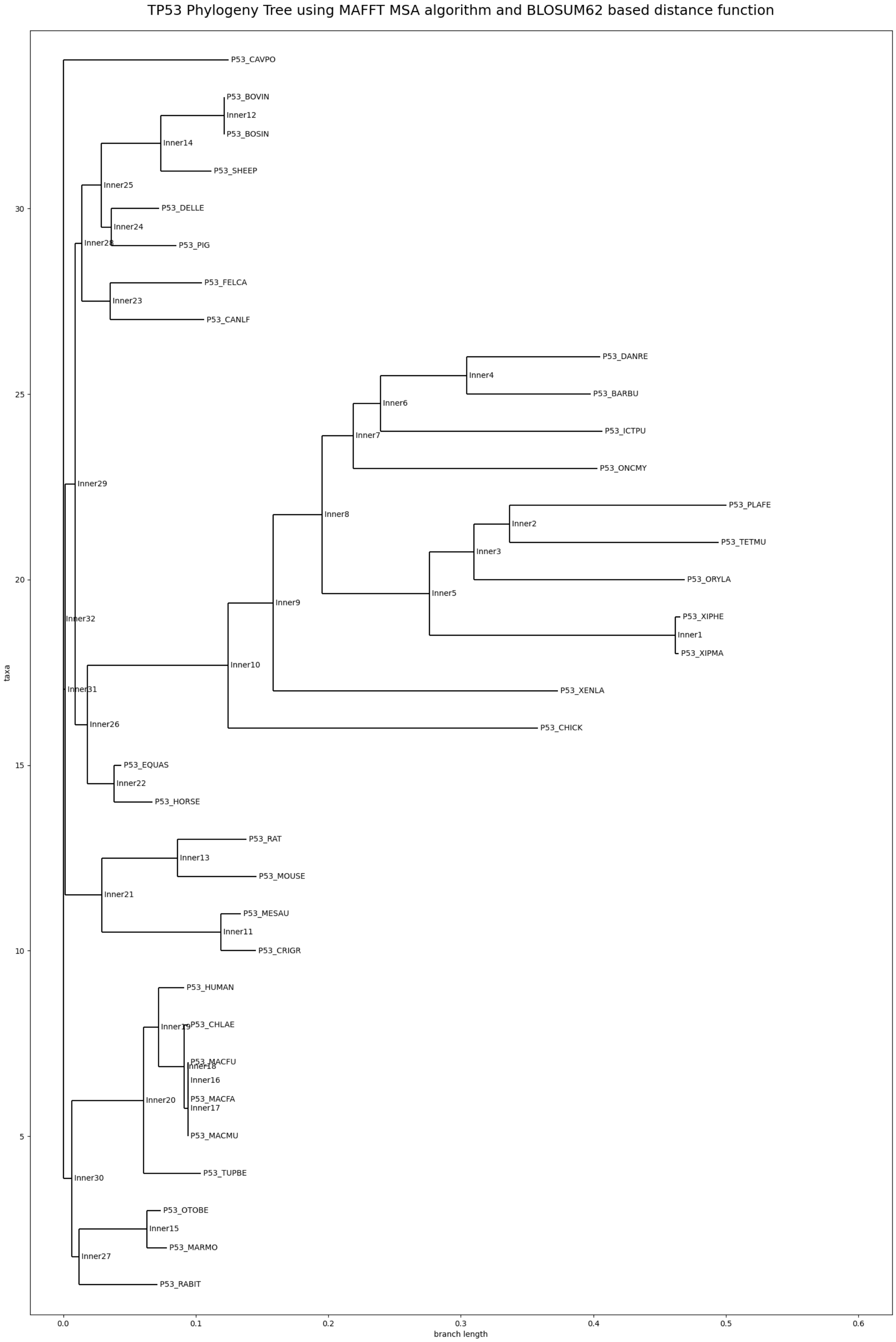

Figure 7: TP53 phylogenetic Tree based on MAFFT.

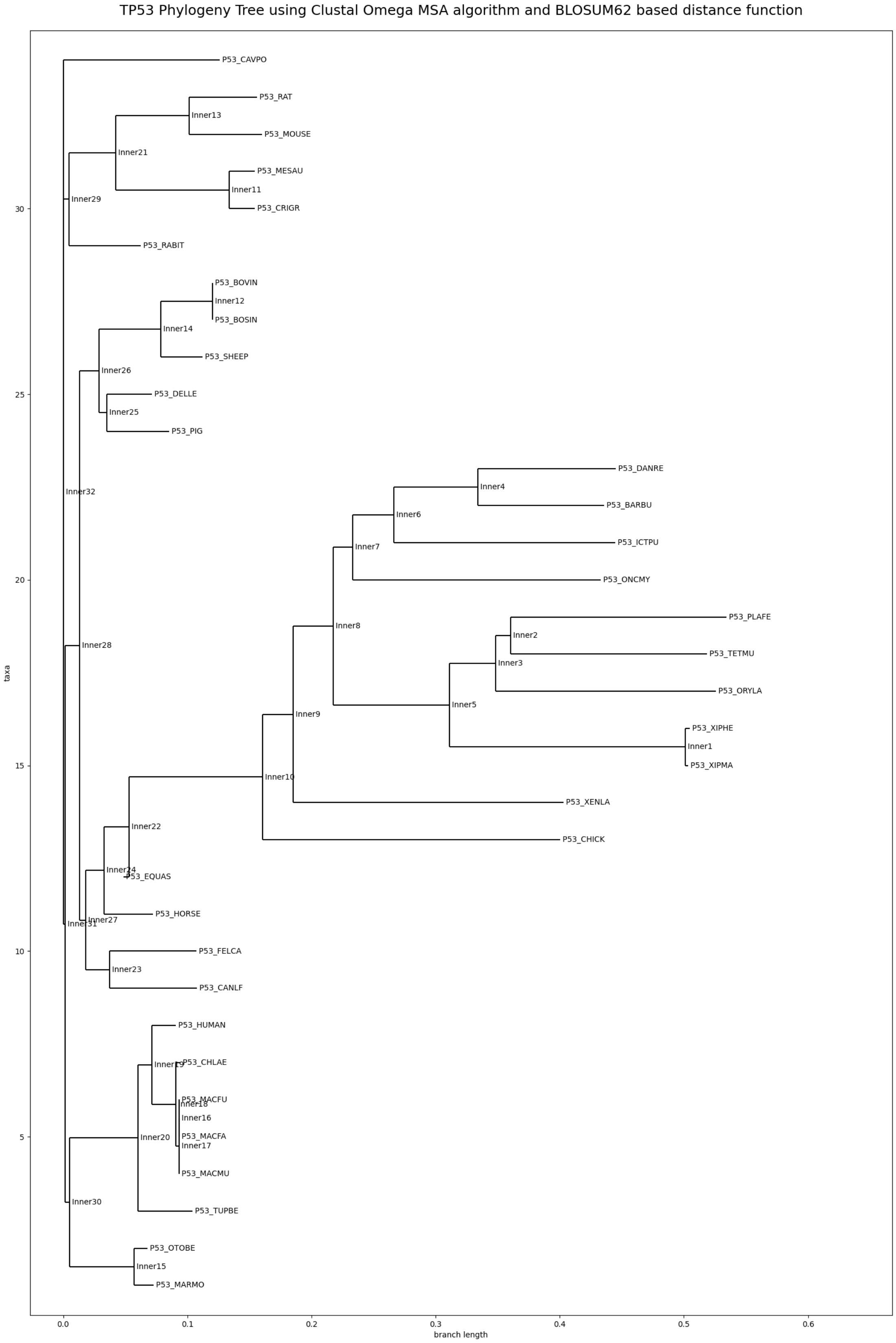

Figure 8: TP53 phylogenetic Tree based on Clustal Omega.

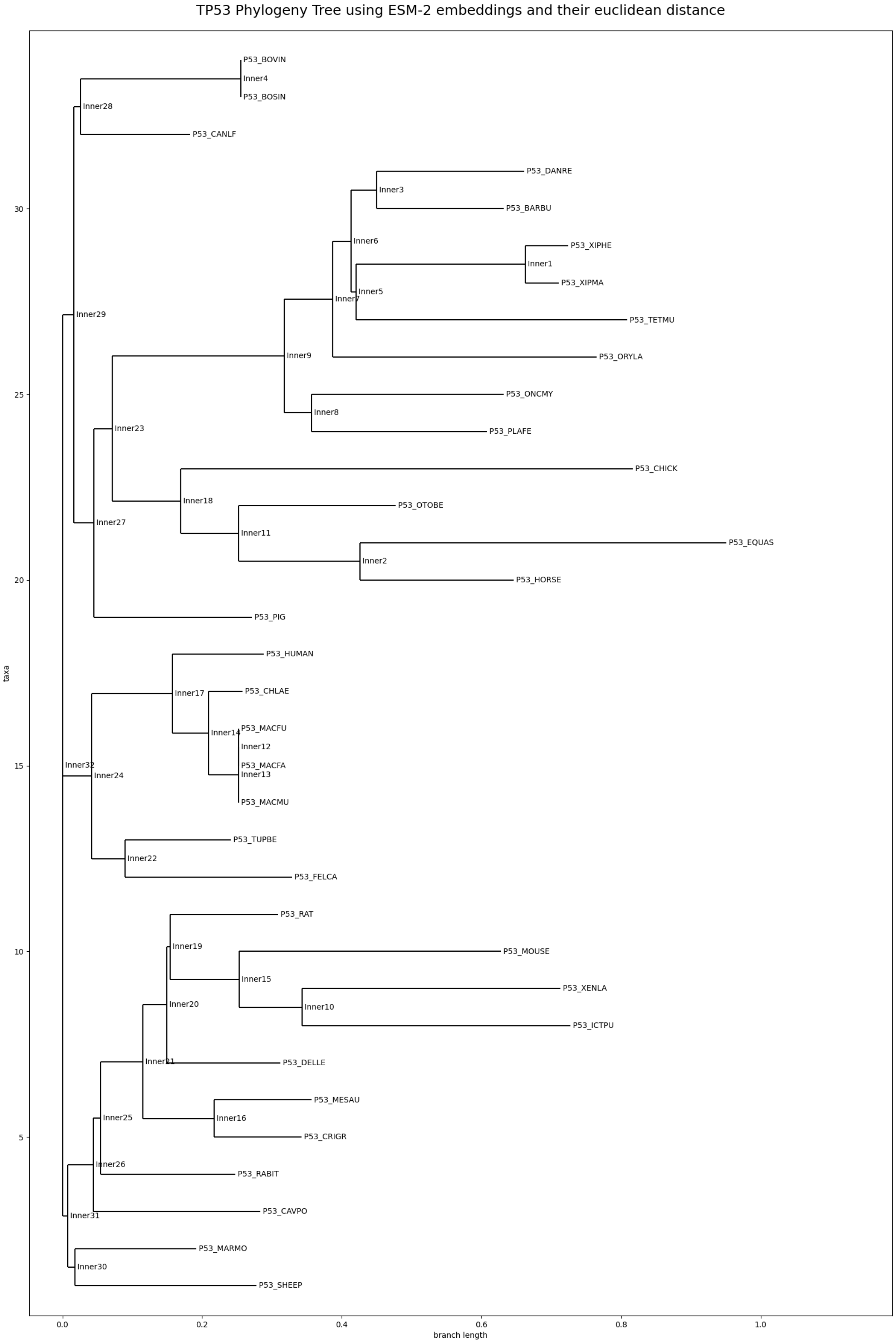

Figure 9: TP53 phylogenetic Tree based on ESM-2 embeddings

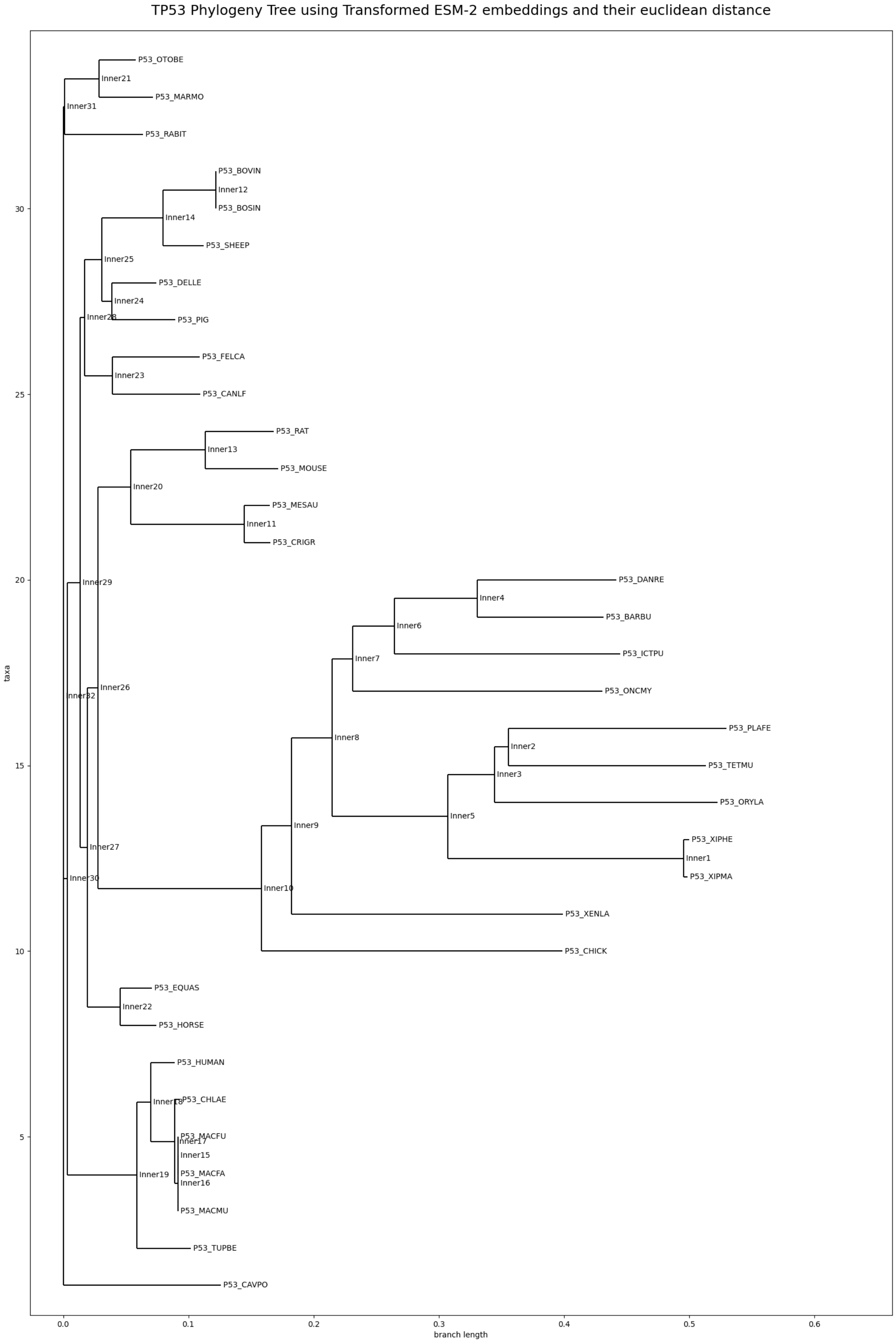

Figure 10: TP53 phylogenetic Tree based on transformed ESM-2 embeddings

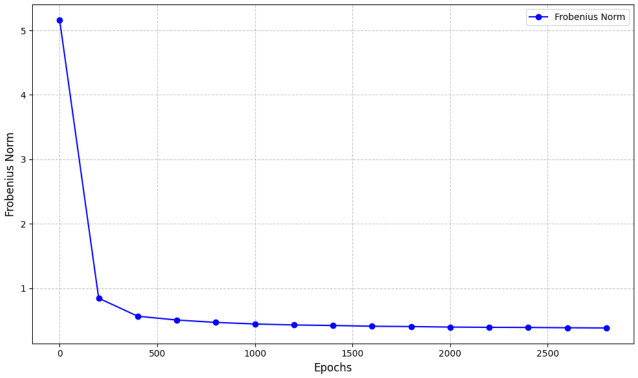

Figure 11: Change over epochs of the Frobenius norm difference between the transformed embedding distance matrix and the Clustal Omega distance matrix for TP53

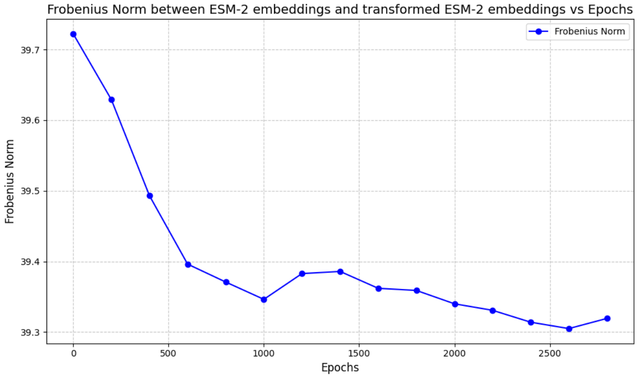

Figure 12: Frobenius norm between ESM-2 embeddings and transformed ESM-2 embeddings vs. Epochs for TP53