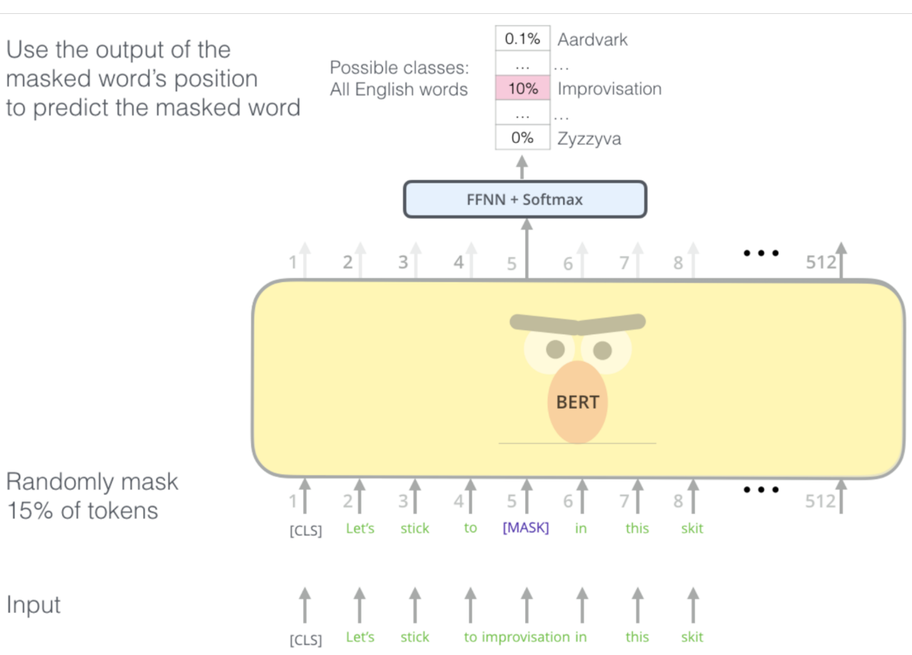
Deciphering the Black Box: Mastering the MSA Transformer for Phylogenetic Tree Reconstruction

Ruyi Chen[1], Sanjana Tule[1], Gabriel Foley[1], and Mikael Boden[1]
[1]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

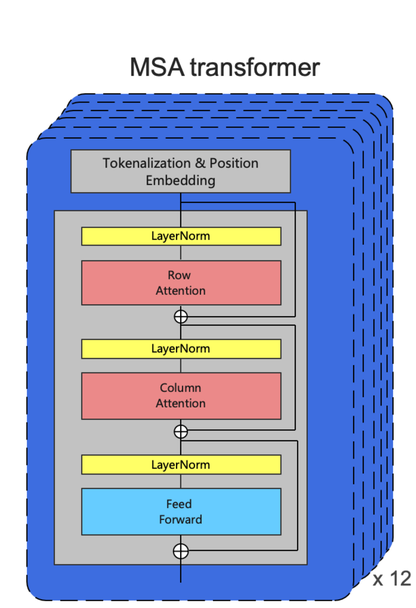
Abstract

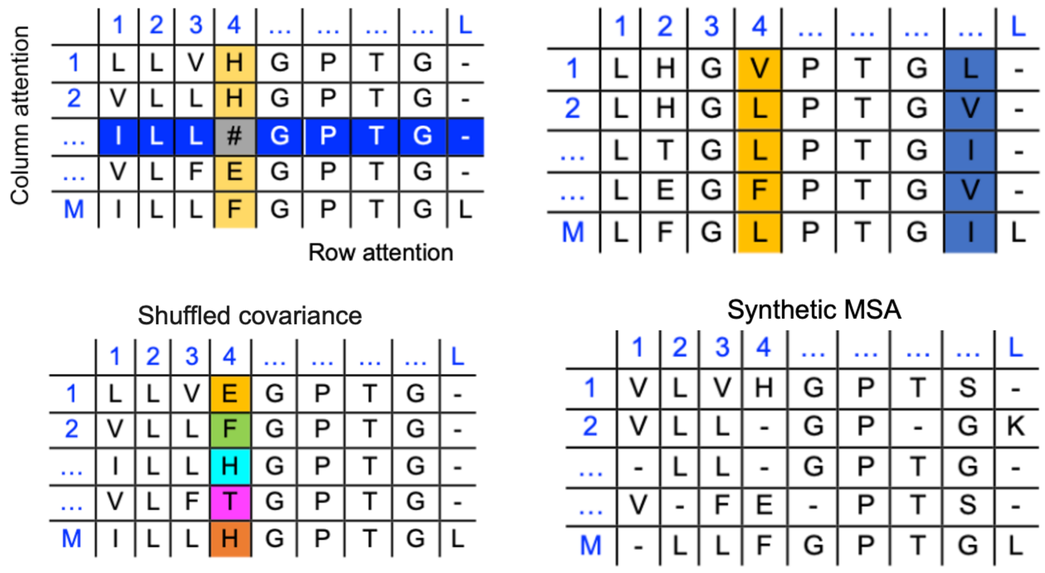
Proteins are composed of peptide bonds that link together amino acids in a sequence. By comparing protein sequences of different organisms, we can hypothesize their evolutionary relationships and shared ancestry, thereby shedding light on the functional significance and evolutionary pressures acting on those proteins. Classical methods of inferring phylogenetic relationships employ mathematical models, such as Maximum Likelihood and Bayesian inference coupled with continuous-time Markovian evolutionary models. Protein Language Models (PLMs) offer an alternative pathway to recover evolutionary relationships. Much like how natural language processing perceives sentences as chains of words, a protein sequence can be envisioned as a "sentence," with amino acids analogous to words. However, the "black box" attributes of neural networks can shroud the rationale behind their conclusions, complicating the use of PLMs in phylogenetic tree reconstruction. To this end, we illustrate how a PLM framed around a multiple-sequence alignment (MSA), the MSA transformer, encodes phylogeny despite of not being explicitly trained to recognise such, and provide a guide for phylogenetic tree reconstruction. Equipped with insights learned from the MSA transformer, we then reconstructed a phylogenetic tree for the RNA virus RNA-dependent RNA polymerase (RdRp) domain, demonstrating how both novel and previously known evolutionary relationships are available from a "non-classical" approach with different computational requirements. It is anticipated that PLMs will complement classical phylogenetic approaches to accurately piece together the evolutionary history of protein families.

Introduction



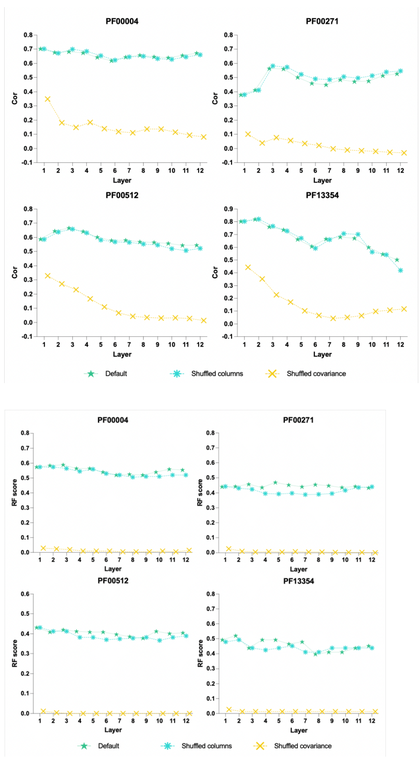
Methods

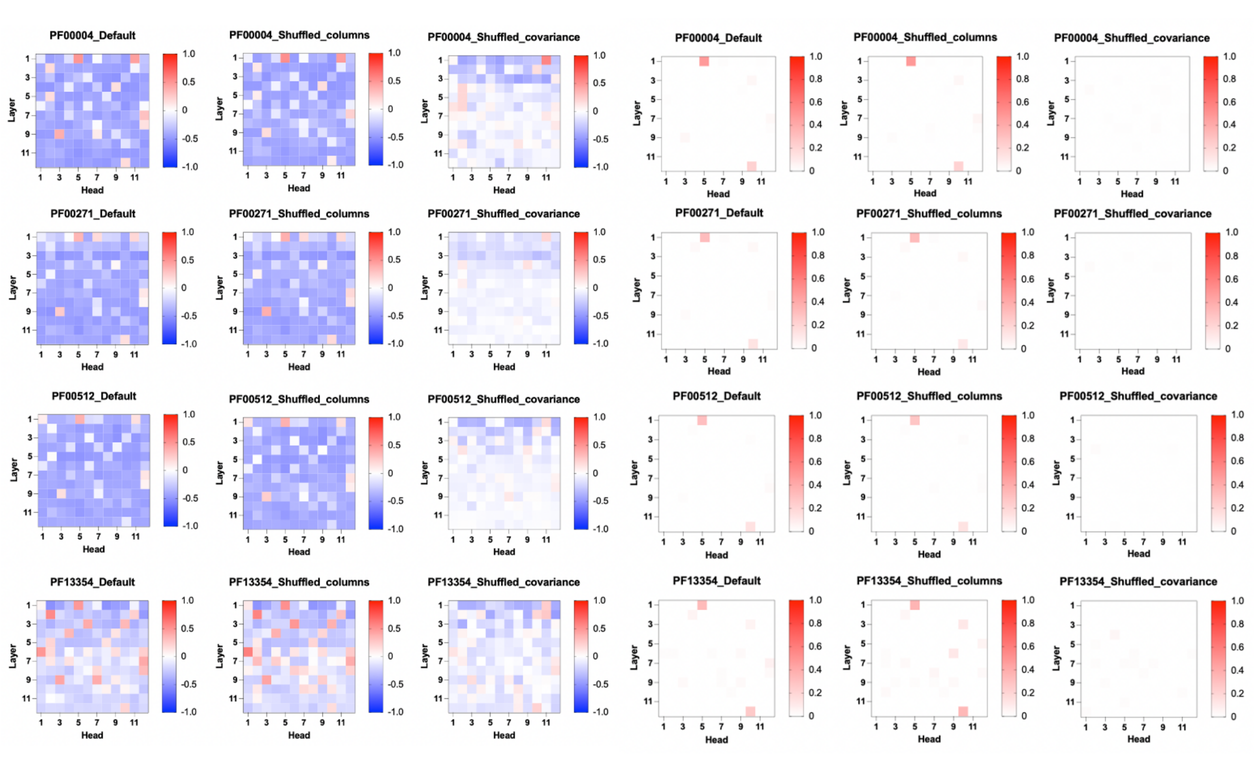




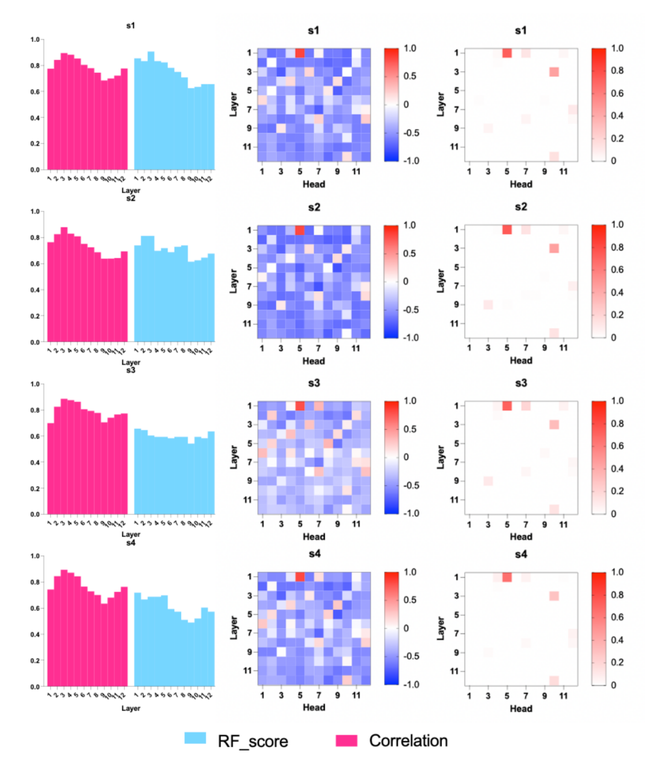
- 3 kinds of MSAs generated from different protein domains and Synthetic MSAs
- Extracting **embeddings** and **column attention** from the MSA transformer layer by layer
- Neighbor-joining trees curated in the Pfam

Real MSAs

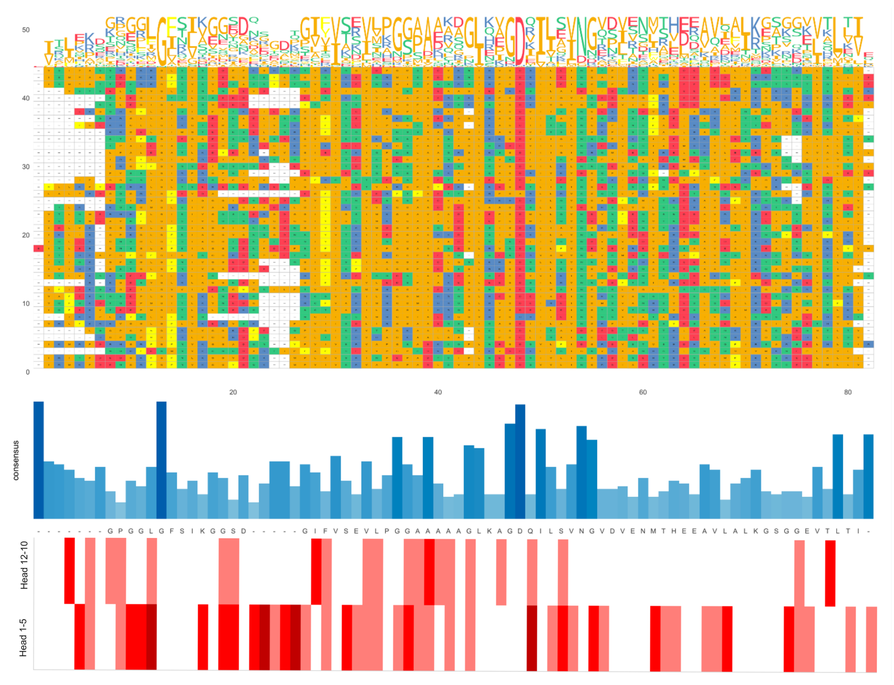




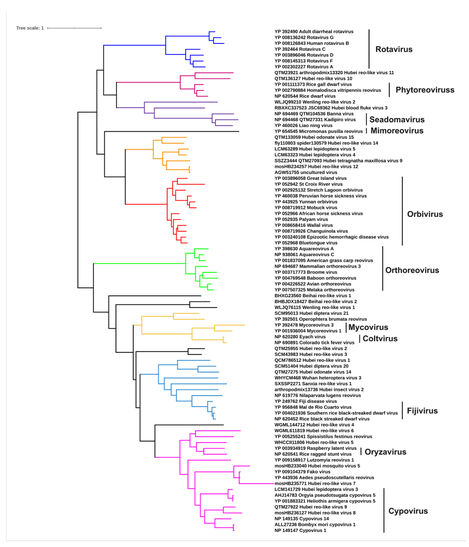
Synthetic MSAs

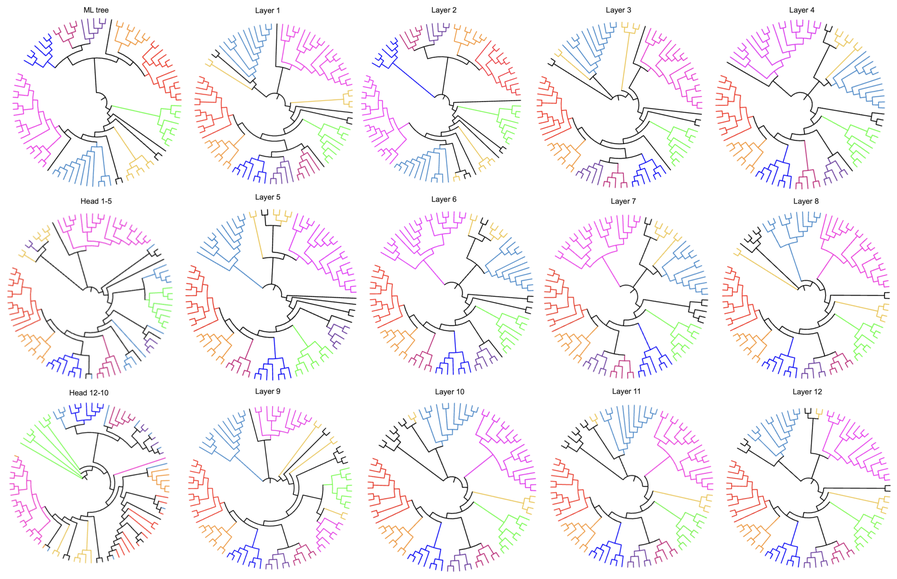


Each column's effect on Phylogeny



Ancestral relationships present in the embedding trees





The RNA-dependent RNA polymerase(RdRp) is the only conserved-sequence domain across all RNA viruses and was therefore used for phylogenetic inference.

Conclusion

- The Euclidean distances of embeddings and column attention heads exhibit a strong correlation with the evolutionary distances from phylogenetic trees
- The attention trees built from two attention heads invariably have some similarity to neighbor trees across all protein families
- Phylogenetic signals are dependent on the covariance of information among aligned sites
- Each column possesses a certain capability to reconstruct phylogenetic trees
- Ancestral relationships present in the embedding trees

Contact us

Emails: ruyi.chen@uq.edu.au
Lab twitter: @LabBoden