Phylogenetics and Genomic Language Models

Catherine Felce
Department of Physics, Math and Astronomy
California Institute of Technology
Pasadena, CA 91125
cfelce@caltech.edu

1 Introduction: Biological language models

LLMs are increasingly being used to investigate viral evolution [1], and predict protein function and structure [2]. Protein [3] and DNA sequences [4] can be used to predict the evolutionary conservation of different sequences. DNA sequence trained models can also effectively predict the effects of nucleotide variants [5], as well as disentangling causality from linkage-disequilibrium (association of nearby chromosomal sites in meiosis).

Challenges include biases in the species represented in pre-training data, repetitive sequences which effectively cause overfitting to said sequences, differences in conservation between coding and non-coding regions, and bidirectionality in DNA sequences. The special challenges in generalizability for molecular sequencing deep learning models and the need for performance metrics which explicitly consider the overlap of specific features between training and test sets has been highlighted [6].

Since long-range DNA interactions can be significant, the quadratic scaling of attention in transformers can be prohibitive. This issue is aggravated by the fact that single-nucleotide polymorphisms (SNPs) can be significant, meaning that tokenization can obscure important interactions. Other approaches are being proposed to increase context lengths whilst maintaining single nucleotide resolution, including HyenaDNA [7], a convolutional long-context model.

Despite challenges, LLMs show enormous promise for analyzing biological sequences [8]. In particular, in this project, we explore potential applications of genomic (DNA) language models to the study of phylogenetics: the evolutionary relationships between organisms.

1.1 Genomic language models in phylogenetics

For phylogenetic reasoning, biologists are interested in relationships between aligned sequences across different organisms or strains. Desired tasks include reconstructing accurate phylogenetic trees [9], and inferring evolutionary dynamics from combined sequence information and known trees [10]. LMs are more effective at learning intra-sequence than inter-sequence patterns [11]. To address this, combined methodologies have emerged, attempting to integrate intra and inter-sequence reasoning to leverage LMs in phylogenetics [11, 12]. The approaches involve combining a masked language modeling (MLM) loss with a phylogenetic loss function based on a known tree or multiple sequence alignment. Phyla [11], for example, combines Mamba layers (state-space model [13]) with sparsified self-attention layers which attend only to within-sequence tokens.

However, such combined approaches have variable generalizability, and at best can only predict tree/MSA structure, without addressing the question about the mechanisms of evolution along such trees. In this project, we briefly explore the zero-shot transfer performance of one such combined model, Albors et al. [12]. We then propose a novel approach to identifying regions of evolutionary interest via analysis of LM vs MSA entropy, and present a proof-of-concept experiment. Thus we attempt to contribute to two research questions. Firstly, can current, MSA-informed gLMs successfully generalize to species beyond those given in the training alignment? And secondly, can the

Preprint. Under review.

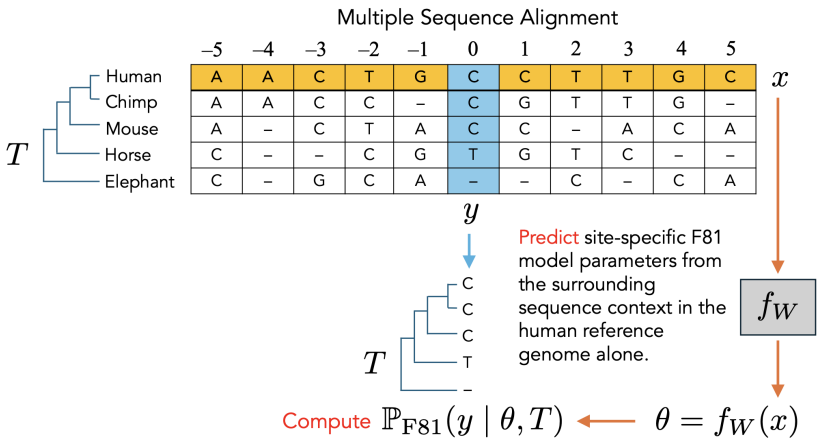

Fig. 1. Illustration of PhyloGPN's modeling framework. The input data consist of 481 bp windows from the human reference genome GRCh38 and the alignment columns are obtained from a whole-genome alignment of 447 mammalian species to GRCh38.

Figure 1: Figure 1 from Albors et al. [12]: the training methodology for PhyloGPN.

separate use of alignment-blind language models and known MSAs contribute to our understanding of how different genomic regions evolve?

2 Assessment of zero-shot transfer in Albors et al.

One model which aims to combine phylogenetic insights with intra-sequence LM learning is PhyloGPN by Albors et al. [12]. The authors highlight the challenges in genomic zero-shot nucleotide variant prediction, and point out that providing MSAs as input has been shown to improve performance (GPN-MSA [14]). They further comment that requiring the MSA as input "makes it difficult to apply the model to other species or to regions where alignment to the human genome is poor. This undermines its utility for transfer learning." GPN-MSA also requires excluding human-proximal primates from the training data to avoid the model 'copying' from those sequences. Albors et al. instead use a CNN to predict a central nucleotide using the sequence context, minimizing a phylogenetic loss function calculated across the MSA. Their approach is summarized in Figure 1, taken directly from their paper [12]. However, since the model trains on the human genome, and all validation is also on human variant prediction, testing the model on a different species would help to assess the added utility of PhyloGPN.

2.1 Applying the model

PhyloGPN is tested on a clinical dataset of human genomic variants (ClinVar), where deleterious nucleotide variants are labeled as pathogenic. We weren't able to find the code used to produce the ROC for ClinVar variant prediction in Albors et al. [12], but we did achieve a AUROC of ~ 0.8 for the ClinVar deleterious variant task using PhyloGPN. See our results in Figure 2, and the corresponding code at `https://colab.research.google.com/drive/1X0dgyFV_SJJtK4oYhabXea6kgikhj4sh?usp=sharing`. We consider only $10,000$ variants, and without filtering for variants of a certain type, which could explain the discrepancy with the AUROC shown in [12].

Since PhyloGPN uses a multiple sequence alignment (MSA) as part of training, but not as input, its clearest use case is for species where genomes, but not alignments, are available. However, in [12], the model is tested exclusively on human sequences. As the next most well-annotated sequence, we attempted fitting the model on the corresponding sequences from the mouse genome. This should give an idea of whether the conservation level learned by PhyloGPN is really ubiquitous over the phylogenetic tree. This is an assumption of their model, which learns transition rates, λ_i, for $i \in \{A, C, G, T\}$, which are assumed to govern a Markovian process of nucleotide switching, with the λ_i constant across the entire tree.

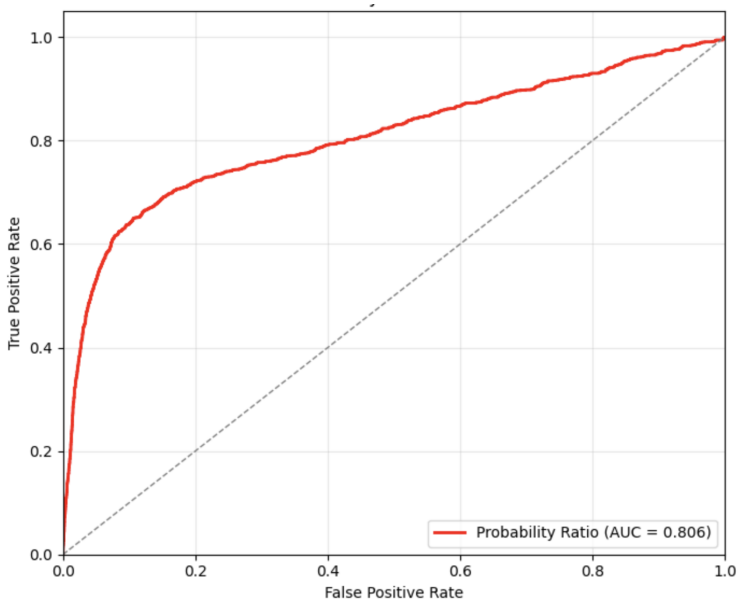

Figure 2: ROC for our test of PhyloGPN on ClinVar variants, attempting to reproduce Figure 3a from [12].

To perform this analyis, we tried two different tests. In the first, we calculated the log-likelihood ratio given by PhyloGPN between the human reference and variant alleles given in ClinVar. Although this seems unfair, given that we are testing on mouse, not human, sequences, it does seem consonant with the biological assumption made in [12], that λ_i, and hence the stationary proabilities for the different nucleotides, are constant over the phylogenetic tree. The loss function involves the probability of the full vector of nucleotides across the MSA, *given* the value in the given species. However, the results for this test are poor, as shown in Figure 3 (ROC ~ 0.39). This is not unsurprising, since the gLM learns from the sequence context within a species. In the second approach, we consider the LLR learnt from the mouse sequence between the value of the central nucleotide in the mouse, and the variant listed in ClinVar. The result for this approach is also shown in Figure 3, and is better than the previous approach (ROC ~ 0.62), although much worse than the performance on human sequences. This suggests that the model might have learned more from the surrounding sequence than from the MSA, and be effectively learning the probabilities of nucleotides within the human genome rather than their conservation across species. This may be a consequence of the training assumption that nucleotides are drawn from a stationary distribution across species.

3 Identifying informative regions via genomic language model entropy analysis

Having noted the limitations of language models in predicting cross-sequence patterns [11, 6], and highlighted the limited generalizability of one recent model in Section 2, we now consider an approach which could exploit these limitations to shed light on evolutionary mechanisms.

3.1 The entropy framework

As a proof of concept, we follow the approach of Lytras et al. [3]. The authors use the concept of protein language model (pLM) entropy to predict the level to which amino acids are conserved/variable in a given protein sequence. This LM entropy represents the model uncertainty over the prediction of an amino acid. The traditional approach to this problem is to use multiple sequence alignments (MSAs), to directly calculate the entropy at a certain sequence position by comparing amino acids across sequences from homologous proteins in different species/strains. However, the LM approach allows us to analyze a single protein sequence, using the site's surrounding context and leveraging

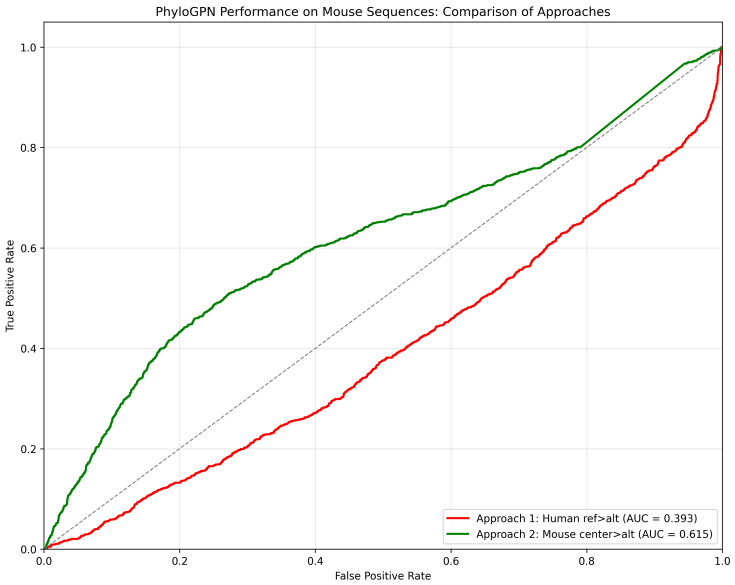

Figure 3: ROC curves from testing on mouse data. **Red curve:** ROC curve using the LLR between human reference and alternate alleles in ClinVar variants. **Green curve:** ROC curve using the LLR between the realized mouse nucleotide and the alternate allele in the ClinVar variants. The same $10,000$ variants as in Figure 2 are used.

understanding of the common structure of proteins from the training data. Whilst Lytras et al. focus on protein sequences, the results can be readily applied to genomic sequences.

As a reminder, the expression for entropy is given by:

$$H = -\sum_i p_i \log p_i, \tag{1}$$

where p_i are the probabilities of different outcomes in a probability distribution. In this context, and in the analysis which follows, for the MSA entropy calculation, the p_i values are the proportion of sequences in the alignment with a given nucleotide/sequence of nucleotides at a given genomic position. For example, at a specific single-nucleotide genomic position, p_A would be given by $\frac{N_A}{N}$, for N_A the number of sequences in the alignment with nucleotide A at that position, and N the total number of sequences in the alignment. In the LM context, at a given token position, the p_i represent the probabilities assigned by the model to the different possible tokens in the model vocabulary, with each token indexed by i.

Lytras et al. [3] perform a comparison between the LM and MSA entropies for viral protein sequences. The lack of viral sequences represented in common protein databases presents an issue for this application of a pLM, and introduces biases towards species heavily represented in these databases (e.g. human). To overcome this, the authors use fine-tuning on sequences from the relevant family of viruses, which they call 'evotuning'. Since MSA entropies provide an independent validation of the pLM entropies, this measure can be used as an assessment of the improvement to the model provided by evotuning. The authors use this to confirm that evotuning has improved their two base models (ESM-2 and protT5MLM,) for application to sequences from the IAV HA surface protein, reporting good correlation between the two entropy measures. However, they note that these results depend on the variety of serotypes included in the 'evotuning' sequence set.

3.2 The evolutionary hypothesis

Regions of functional importance can be identified through conservation across an MSA, resulting in low MSA entropy. Given that a particular genomic region is highly conserved, this could be due to an immediate, structural selection pressure which makes any mutation deleterious to the function of the

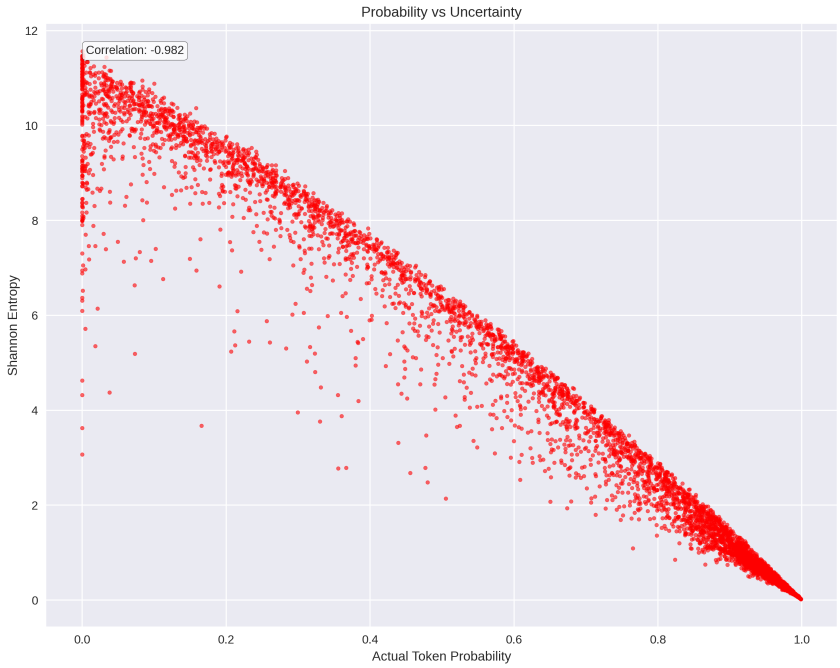

Figure 4: Token entropy vs actual token probability according to the nucleotide-transformer LM, pre-trained on human sequences and applied to the coronavirus genome.

entire sequence. We would expect this scenario to result in a low LM entropy, since the LM would reliably be able to predict each token along the length of the structurally integrated region. However, the conservation could also be due to 'higher level' selection effects, which are primarily to do with specific functions, and not just structural considerations. Thus these low MSA, high LM entropy regions could be interesting sequences for further investigation, as they represent regions which 'could have been otherwise' but have been 'actively selected' to be conserved across strains/species.

For example, a long and not particularly crucial protein sequence may be highly conserved, because small mutations destroy the function of the protein. However, a region full of short enhancers, whose sequences may not be predictable from the surrounding context, might be conserved because each enhancer plays a crucial role in the organisms survival. This second, functionally rich region, would have higher LM entropy, and would be more interesting to study.

3.3 Model fitting and entropy calculation

For ease of interpretation, we use the readily available MSA for coronaviruses, given at the GISAID database [15]. We use the `nucleotide-transformer-500m-human-ref` [16] for our foundational language model, a transformer pre-trained on human sequences. We assessed the performance of the model and calculated model entropies for each token in only the first sequence in the MSA. The tokenization is in nucleotide k-mers of up to length six (e.g. 'ACGTTA'). The code for the LM fitting and subsequent analyses are included here: `https://colab.research.google.com/drive/1p0i5MvQmaiOlqoiOVvXAHbTZYrHSVXrR?usp=sharing`.

Model performance on the coronavirus genome was promising, with a mean correct probability of 62%, and a median correct probability of 73%, and the correlation between confidence and accuracy is demonstrated in a plot of entropy vs actual token probability in Figure 4. The Pearson correlation between the entropy and the actual token probability is -0.982. In addition, for the 5083 tokens in the $\sim 30,000$ nucleotide genome, 30.3% were correctly predicted with very high $> 90\%$ confidence, and 65.2% were correctly predicted with $> 50\%$ confidence. The distribution of actual probabilities is shown in Figure 5, and the average actual token prediction across positions in the genome is shown in Figure 6.

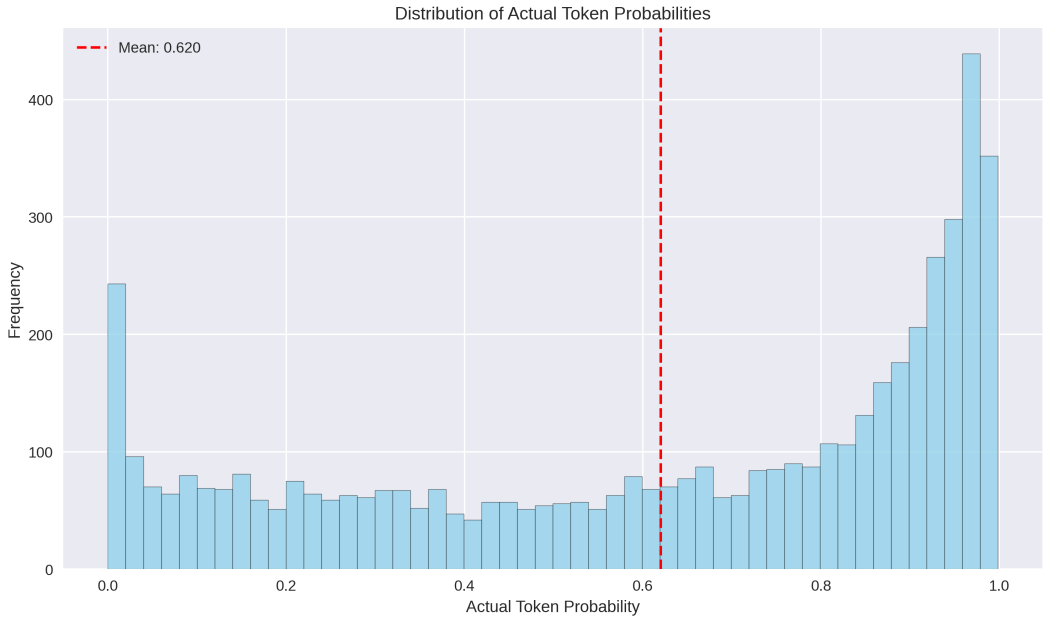

Figure 5: Distribution of actual token probabilities of the nucleotide-transformer LM across the coronavirus genome.

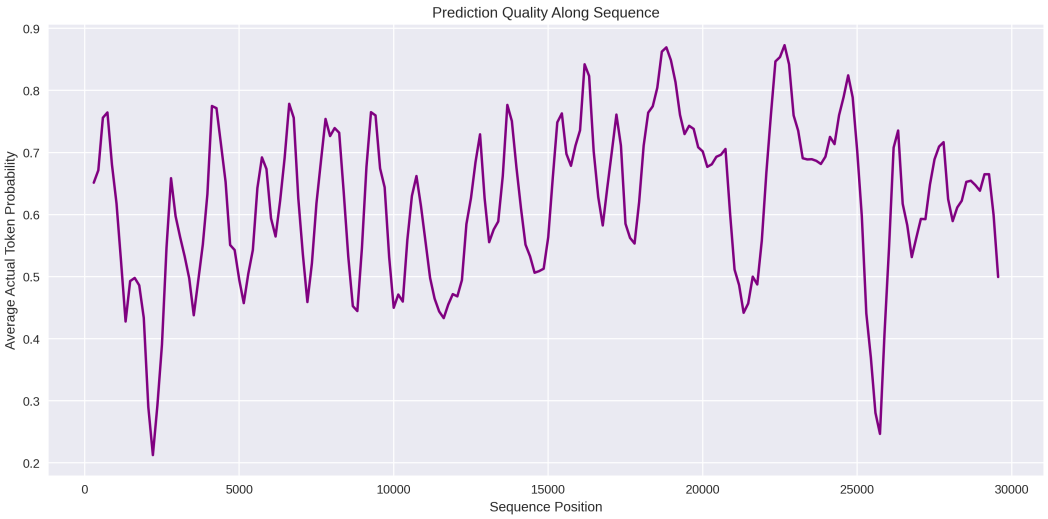

Figure 6: Average performance of the nucleotide-transformer LM at positions across the coronavirus genome, using a sliding window of 100 tokens.

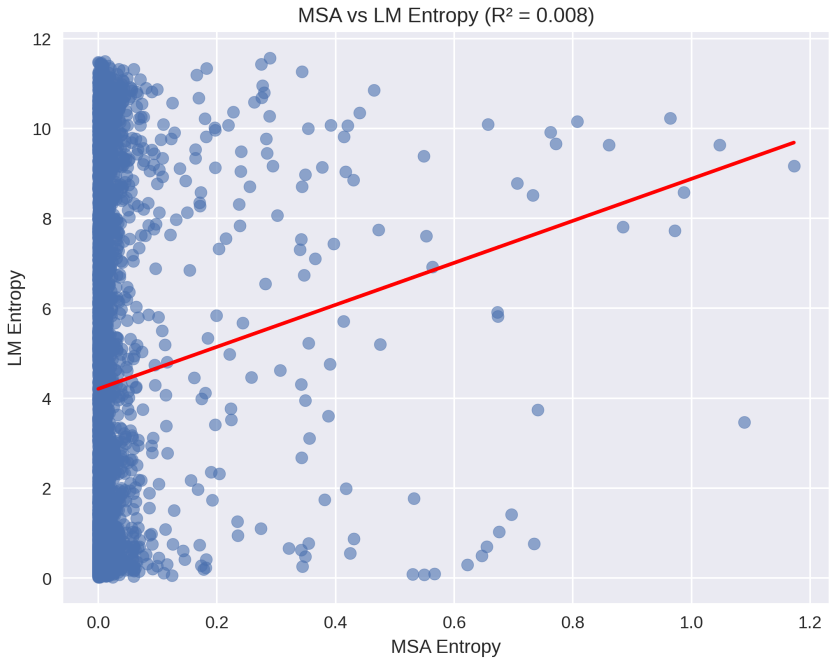

Figure 7: LM vs MSA entropy for each token in a coronavirus genome.

We then calculated the 'MSA entropy' for each token, by considering the distribution of sequences at each token position across the entire dataset. The goal was to see whether the distribution of entropies revealed any significant regions of interest in the genome.

3.4 Results

The overall correlation between LM and MSA entropy at each token position was low, with $R = 0.008$, as shown in Figure 7. However, taking non-overlapping windows with varying numbers of tokens did reveal significant correlations (see Table 1), with p-values < 0.05 for windows of size $2, 5$ and 10.

Window Size	Independent Correlation	P-value	N Windows	Is Significant
2	0.096725	0.000001	2541	True
5	0.102746	0.001039	1016	True
10	0.111445	0.011955	508	True
20	0.114915	0.067478	254	False
30	0.145405	0.059258	169	False
40	0.163219	0.066726	127	False
50	0.074562	0.458665	101	False
100	-0.005643	0.968976	50	False
500	0.438484	0.204935	10	False
1000	0.488490	0.403742	5	False

Table 1: A significance summary for LM and MSA average entropies correlation, using non-overlapping windows of given sizes. 'Independent Correlation' refers to the Pearson correlation between the average MSA entropy and average LM entropy within each non-overlapping window.

Although these entropy correlations are lower than those found in [3], the goal of this analysis was to explore the strategy of using *disparities* between the LM and MSA entropies to reveal interesting genomic regions. To this end, we examined regions of the coronavirus genome which were enriched for low MSA, high LM entropy tokens. Our hypothesis was that these regions are enriched with sequences which are evolutionarily constrained, despite not being inherently prevented from mutation by structural considerations. The results of this analysis are shown in Figure 8.

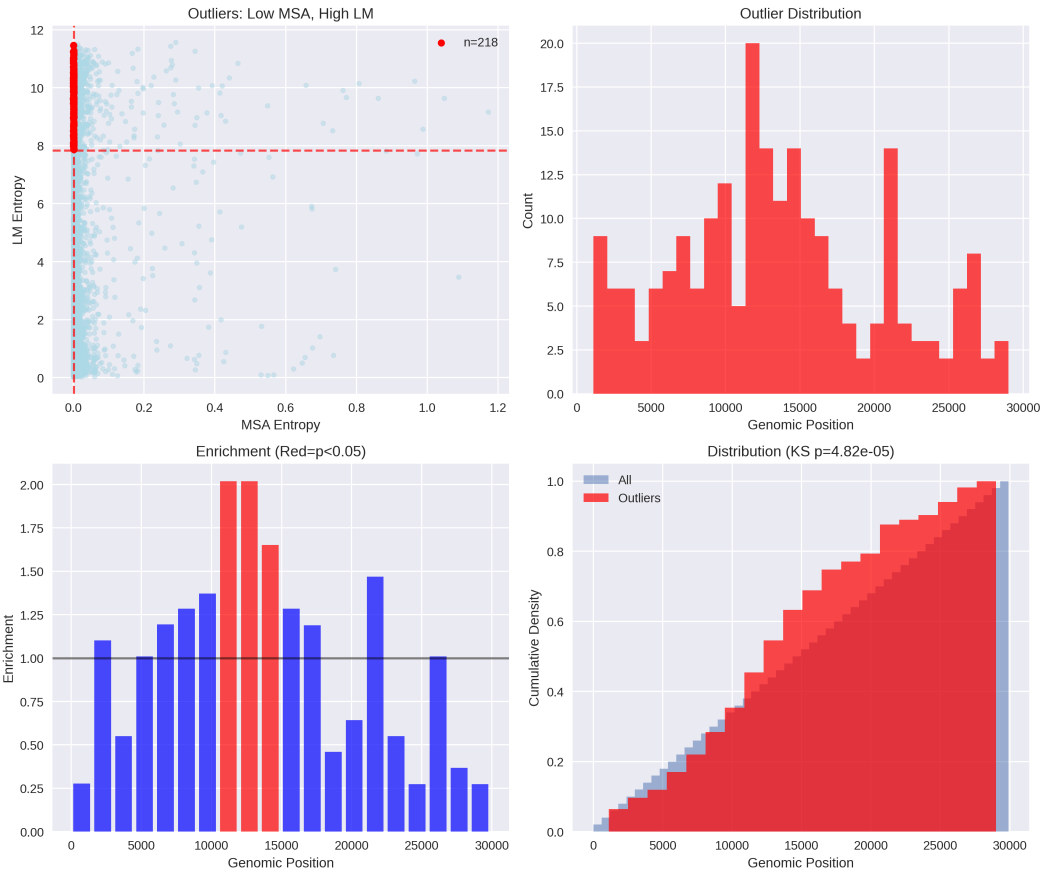

Figure 8: Enrichment of regions for tokens with low MSA entropy (bottom 25th percentile) and high LM entropy (top 25th percentile). **Top left:** Outliers highlighted in a scatter plot of tokens, with the LM entropy vs MSA entropy for each token. **Top right:** Distribution of outlier tokens across the genome. **Bottom left:** Regions enriched at the $p < 0.05$ level for outlier tokens. **Bottom left:** Cumulative distribution of outliers across the genome.

The analysis shows that high LM entropy, low MSA entropy tokens are enriched in the replicase region of the virus ($\sim 11500 : 16000$bp). The enrichment information is shown in table 2. These regions of the coronavirus encode viral proteins essential for replication, including RdRp and nsp6-nsp12. Comparing to the results for which regions are enriched for tokens selected purely for low MSA entropy (Figure 9), we see that enrichment analysis on low MSA entropy alone also highlights the spike protein region of the coronavirus ($\sim 21500 : 25400$bp). So, by comparing the LM and MSA entropies of the tokens, we are able to identify a difference between the two critically important regions of the coronavirus genome. The spike protein region, known to be highly conserved, is enriched for low MSA entropy tokens as expected, but not for low MSA-high LM entropy tokens. However, a section of the replicase region is enriched for both sets, indicating that the observed conservation within this region may be of special interest. Evolutionary pressures may act here, which force conservation even though structurally other sequences would be permissible.

These results are consistent with the hypothesis that it is the density of independent functionally important features which is being detected in this approach. Whereas the section of the replicase region which is implicated here is made up of several, short, replication proteins ($\sim 200 - 1000$bp), the spike protein region is made up of solely 2, ~ 2000bp, spike proteins. Thus, whereas both the replicase and spike protein regions are conserved across their entirety, the high LM entropy indicates that several, functionally important elements, which are structurally distinct, are gathered together in the same region in the case of the viral replicase region.

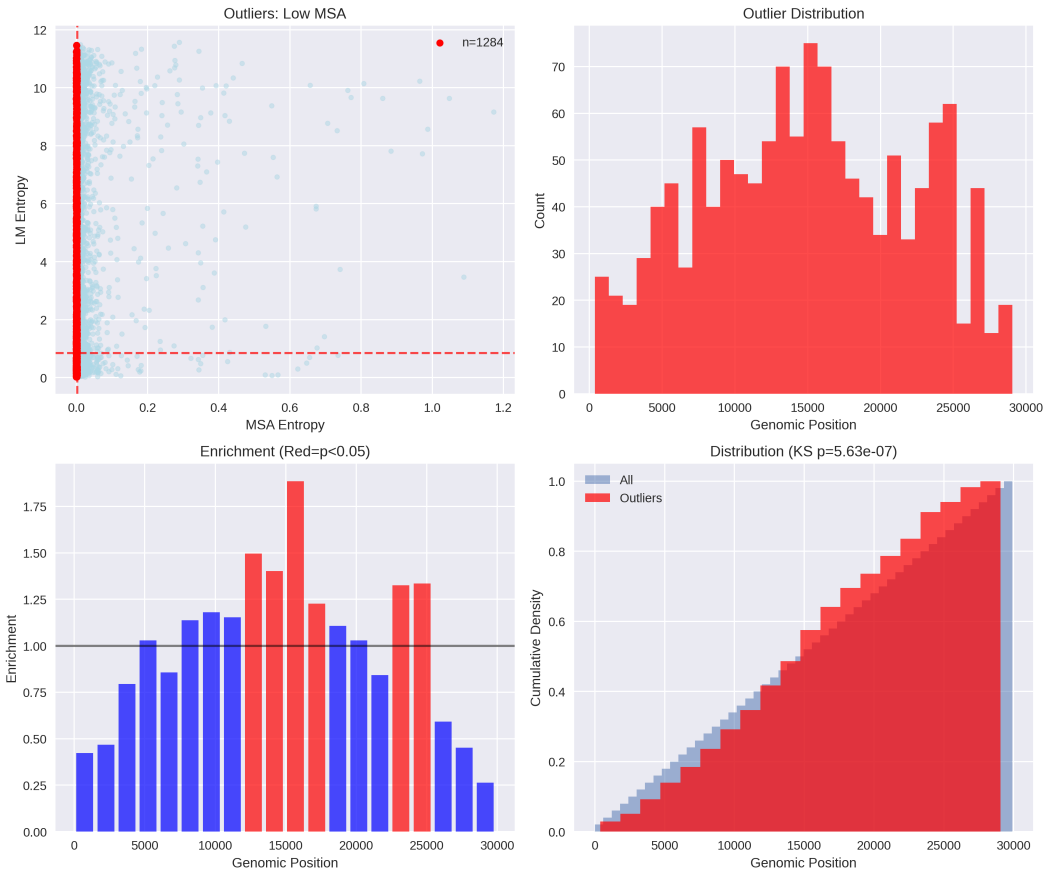

Figure 9: Enrichment of regions for tokens with low MSA entropy (bottom 25th percentile). **Top left:** Outliers highlighted in a scatter plot of tokens, with the LM entropy vs MSA entropy for each token. **Top right:** Distribution of outlier tokens across the genome. **Bottom left:** Regions enriched at the $p < 0.05$ level for outlier tokens. **Bottom left:** Cumulative distribution of outliers across the genome.

An obvious limitation of this example is the training of this nucleotide-transformer model on human genomic sequences. Despite this, the approach did allow us to distinguish two functionally distinct regions in the coronavirus genome, and could be a useful measure of the density of structurally independent features within different regions of the genome.

Table 2: High LM Entropy, Low MSA Entropy Token Enrichment in Replicase Region

Region	Enrichment	k	n	P-value
11561-13006	2.1×	22	245	0.000899
13006-14451	1.9×	20	245	0.004475
14451-15896	1.7×	18	246	0.019046

4 Conclusion

Noting the current limitations of language models in analyzing cross-species sequences, and therefore in drawing conclusions about evolutionary mechanisms, in this work we have both further highlighted these limitations, and suggested a new approach for leveraging these very limitations.

In Section 2, we perform a case study on a recent paper attempting to incorporate known phylogenetic models into cross-species genomic sequence learning. With a cross-species application, we highlight potential limitations of the training framework (phylogenetic loss function with constant transition rates), and the generalizability of the model. The use case for such models, which use multiple

sequence alignments in training, is unclear unless the model can be reliably used on sequences from unaligned genomes. However, if the model performance decreases significantly when moving from the human to the well-annotated mouse genome, which was included in the MSA used for training, this raises questions about the cross-species generalizability of the model.

In Section 3, we suggest an alternative approach, which accepts the difficulty inherent in inferring cross-species relationships from a (potentially biased) set of intra-species sequences, and tries to harness the limitations of genomic language models to make evolutionary inferences. We follow the entropy framework of [3], transferring it from the protein to the DNA sequence setting, and attempt to ask, instead of 'What can a gLM teach us?', 'What can we learn from what it cannot?'. With a simple analysis, we were able to use the combined MSA/LM level of constituent genomic tokens to differentiate between two crucial functional regions of the coronavirus genome. Thus we have highlighted the promise of this approach, which could be refined to reveal different evolutionary dynamics between different, equally conserved regions.

5 Data and code availability

The code for analyses in Sections 2 and 3 is available at `https://github.com/CatherineFelce/phylogenetic-gLMs`.

SARS-CoV-2 virus genome sequence data were obtained from the GISAID Database. The multiple alignment data can be assessed through FigShare.[15], at `https://doi.org/10.6084/m9.figshare.20486178.v1`.

Acknowledgments and Disclosure of Funding

Thank you to Matt Pennell for encouraging me to think about how to make evolutionary inferences from LLMs.

6 Other cool approaches

Lu et al. [17] use contrastive learning to inform their embedding of genomic sequences. The 'views' in their approach correspond to different homologs from the same ancestral sequence, and the learned embedding should encode functionally conserved regions between the two homologs.

References

[1] Brian Hie, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, jan 2021.

[2] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(723), 2019.

[3] Spyros Lytras, Adam Strange, Jumpei Ito, and Kei Sato. Inferring context-specific site variation with evotuned protein language models. *bioRxiv*, feb 2025. Preprint.

[4] Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Ana Berthel, Zong-Yan Liu, Wei-Yun Lai, Zachary R Miller, Armin Scheben, Michelle C Stitzer, M. Cinta Romay, Edward S Buckler, and Volodymyr Kuleshov. Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained dna language model. *PMC*, Jun 2024. Copyright and License information PMC.

[5] Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023.

[6] Y. Ektefaie, S. Amiri, and M. Gerstein. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6:1512–1524, 2024.

[7] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution, 2023.

[8] M. E. Consens, C. Dufault, M. Wainberg, et al. Transformers and genome language models. *Nature Machine Intelligence*, 7:346–362, 2025.

[9] David Posada and Keith A. Crandall. Felsenstein phylogenetic likelihood. *Journal of Molecular Evolution*, 89(3):134–145, Apr 2021. Epub 2021 Jan 13.

[10] Alexander L. Cope et al. Macroevolutionary divergence of gene expression driven by selection on protein abundance. *Science*, 387:1063–1068, 2025.

[11] Andrew Shen, Yasha Ektefaie, Lavik Jain, Maha Farhat, and Marinka Zitnik. Phyla: Towards a foundation model for phylogenetic inference. *bioRxiv*, 2025. Preprint.

[12] Carlos Albors, Jianan Canal Li, Gonzalo Benegas, Chengzhong Ye, and Yun S. Song. A phylogenetic approach to genomic language modeling, 2025.

[13] Meenakshi K. Doma and Roy Parker. Rna quality control in eukaryotes. *Cell*, 131(4):660–668, 2007. Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Review.

[14] Gonzalo Benegas, Clara Albors, Jiale Aw, et al. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 2025. Published online with open access.

[15] Boon WX and Chong Han Ng. MSA (SARS-CoV-2). 8 2022.

[16] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01, 2023.

[17] Amy X. Lu, Alex X. Lu, and Alan Moses. Evolution is all you need: Phylogenetic augmentation for contrastive learning, 2020.