Intrinsic Evaluation of DNA Embeddings in Genome Language Models: Insights from Yeast Genomic Sequences

Ruhaib Muhammad [*] [1] [2] **Rajeeva Lokshanan Reguna Madhan** [*] [1] [2] **Roshan Balaji** [1] [2] **Nirav Pravinbhai Bhatt** [1] [2] [3]

Abstract

In this work, we present a task-independent evaluation of Genome Language Model (gLM) embeddings to understand what contextual and biological information they inherently capture. Through three novel experiments, we assess how well embeddings reflect sequence similarity, encode evolutionary context, and respond to synthetic point mutations using Yeast genomic sequences. Our findings reveal that embeddings correlate with sequence similarity, cluster by phylogenetic clade, and show differential robustness between coding and non-coding regions. These results offer new insights into the representational capabilities of gLMs and pave the way for principled interpretability and benchmarking of gLMs.

1. Introduction

Large Language Models (LLMs) have achieved unprecedented levels of success in handling natural languages (Liu et al., 2019; Sun et al., 2019; Achiam et al., 2023). In recent years, a new family of foundational models trained on genomic sequences instead of natural language, called Genomic Language Models (gLMs) has been developed to utilize these models' language understanding prowess on the language of life (Nguyen et al., 2023; Fishman et al., 2025; Zhou et al.). These encoder or decoder-based models use nucleotide sequences as inputs and learn genome languages using the concepts of LLMs. These learned embeddings are then used for various downstream tasks such as identifying promoters (Zhou et al.), detecting splice sites (Nguyen et al., 2023), and predicting transcription factors (Benegas et al., 2023), in which they have demonstrated remarkable accuracy.

Currently, gLMs are validated by their performance in predefined downstream tasks. This constitutes an *extrinsic* evaluation of the models (Marin et al., 2024), and one does not get insights into what knowledge the pre-training imbibes. Therefore, there is a need to perform an *intrinsic* evaluation of them - by characterizing them in isolation, independent of any downstream tasks, we can quantify their representational power. This will thoroughly assess the quality of the representations generated by the pre-trained gLMs based on the knowledge they contain, and help in not only gaining a better understanding of these models' learning process but also serve as a benchmark for gLMs. This work carries out an *intrinsic evaluation* of gLMs for biological information learned by gLMs. Particularly, the work investigates the questions of how well a set of open-source gLMs embeds the biological information for sequence similarity, contextualization, and robustness to mutation. We use well-studied genomic data from 1011 strains (Peter et al., 2018) of *Saccharomyces cerevisiae* to investigate our question.

2. Methodology

The main objective of this project is to evaluate gLMs intrinsically, regardless of downstream tasks. To do this, three criteria were devised:

1. **Sequence Similarity**: Similar sequences should yield similar embeddings.
2. **Embedding Contextuality**: Similar sequences in different contexts should have different embeddings.
3. **Embedding Robustness**: The effect of point mutations in sequences must be appropriately reflected in their embeddings.

2.1. Genome Language Models

To comprehensively evaluate the representational capabilities of gLMs, models of varying sizes, architectures (encoder and decoder), tokenization strategies, and pretraining data were selected as shown in Table 1. These models were used without any fine-tuning or modification, as the goal was to analyze the information and context already

[*]Equal contribution [1]Department of Biotechnology, Bhupat And Jyoti Mehta School Of Biosciences, Chennai–600036, India [2]Department of Data Science and AI, Wadhwani School of Data Science and AI, Chennai–600036, India [3]School of Engineering and Science, IIT Madras Zanzibar, Tanzania. Correspondence to: Nirav Bhatt <niravbhatt@iitm.ac.in>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

captured in their pre-trained embeddings. The diversity in model characteristics allowed us to systematically assess how these factors influence embedding quality. This allowed for a robust comparative analysis across the full spectrum of gLM design, purely based on their pretrained representations. The embedding vector, an output of the final layer of a particular gLM, was generated by providing each sequence as an input.

2.2. Data Sets: Yeast DNA Sequences

Genomic dataset from 1011 strains of *Saccharomyces cerevisiae* (baker's yeast), a well-characterized eukaryotic organism, was used (Peter et al., 2018). Since the yeast genome strikes a unique balance: it is simple and compact like bacterial genomes, yet retains the structural and regulatory complexity typical of eukaryotes, this dataset was selected for this study. This dataset also acts as a smaller snapshot of a greater biological diversity. Hence, results obtained on this dataset can be used to generalize to other biological systems. From each strain, we extracted specific individual gene sequences to serve as inputs. These genes are largely conserved across strains, with only minor nucleotide variations, and they provided an ideal testbed for analyzing how gLM embeddings reflect subtle sequence differences. This dataset was used to investigate sequence similarity and embedding contextuality criteria.

For the third criterion, embeddings pertaining to coding and non-coding sequences have to be compared. For the same, we made use of conserved non-coding sequences belonging to *Sarcopterygii matsunami*, sourced from Inoue & Saitou (2020).

2.3. Sequence Similarity

Sequences of the same gene from all 1011 strains were used to study sequence similarity criteria. Seven different genes of varying lengths (ranging from 600 to 4000 bases) were considered. These sequences are mostly conserved, and hence, they differ only at a few nucleotides. Their similarity is quantified by the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970). The lower Needleman-Wunsch score implies a more dissimilar sequence. The embeddings for these gene sequences were generated using the gLMs, and the similarity between embeddings was calculated using the Euclidean distance. Higher Euclidean distance implies a dissimilar sequence. Thus, if our hypothesis holds, then the Needleman-Wunsch alignment scores between these DNA sequences should be negatively correlated with the Euclidean distances between their respective embeddings.

2.4. Embedding Contextuality

This criterion evaluates the ability of the gLMs to capture certain sequence-specific characteristics/contexts. Context, for example, could be the clade of the species from which a sequence is obtained. An ideal gLM should be able to generate similar embeddings for sequences from the same clade. To test this, gene sequences of *S. cerevisiae* were obtained from Peter et al. (2018). Genes were chosen instead of entire genome sequences as the average yeast genome length is 12 million nucleotides, which is well beyond the context lengths of all of the models that were considered. Embeddings for each of the sequences were obtained. The k-means and agglomerative clustering algorithms were applied to embedding vectors to investigate the question of how the embeddings group with respect to one another.

2.5. Embedding Robustness

Mutations are nucleotide substitutions in DNA sequences that can often have far-reaching consequences, capable of drastically altering an organism's phenotype. Accordingly, understanding how gLM embeddings respond to such mutations is essential for assessing their reliability in applications involving genomic variation, such as variant effect prediction or disease mutation analysis. Further, it is important to analyse whether the embeddings pertaining to coding and non-coding sequences could be differentiated by gLMs. Hence, robust embeddings should reflect meaningful biological differences while remaining invariant to neutral or synonymous changes, especially in coding sequences where redundancy in the genetic code, i.e., codon degeneracy, often mitigates the impact of mutations.

Here, coding sequences from *S. cerevisiae* and non-coding sequences from *S. matsunami* were chosen for each of the models based on their context window, i.e., for each model, the sequences' lengths were equal to the maximum permissible input length. Synthetic mutations were generated in these sequences by randomly replacing nucleotides in increasing percentages, ranging from 5% to 50%, in steps of 5%. For each mutation percentage, 100 independently mutated versions of the original sequence were generated, ensuring randomness in mutation sites. After generating embeddings for each of these sequences, the Euclidean distances between each mutated sequence's embedding and its original (baseline) embedding were computed to measure their difference in context. The following questions are posed to investigate the robustness of embeddings.

- How sensitive are gLM embeddings to random point mutations in DNA sequences?
- Whether the type of sequence (coding vs. noncoding) affects embedding shifts?
- How do embedding distances change with an increase in the percentage of mutated bases?

Table 1. Summary of the Genome Language Models considered for this study.

Model	Architecture	Size	Tokenization	Context Length	Trained on
HyenaDNA(Nguyen et al., 2023)	Hyena	0.4M - 6.5M	Character-level	1k - 1M	Human
DNABERT-2 (Zhou et al.)	BERT	117M	Byte Pair Encoding	512	Human, Multispecies
DNABERT-S (Zhou et al., 2024)	BERT	117M	Byte Pair Encoding	512	Human, Multispecies
Nucleotide Transformer (Dalla-Torre et al., 2025)	BERT	500M - 2.5B	k-mer tokens	1k	Human, Multispecies
GENA-LM (Fishman et al., 2025)	BigBird	336M	Byte Pair Encoding	36k	Human, Multispecies
GPN (Benegas et al., 2023)	CNN-Transformer Hybrid	65M	Character-level	512	Human
GROVER (Sanabria et al., 2024)	BERT	0.4M - 6.5M	Byte-Pair Encoding	510	Human

3. Results

3.1. Sequence Similarity

The models exhibit strong negative correlations as shown in Figure 1, giving credence to our hypothesis that similar sequences have similar embeddings. As seen in Figure 1, HyenaDNA-large-1M performs better on longer gene sequences such as YAL026C (4068 bases), while the Nucleotide Transformer family of models performs better with shorter gene sequences such as YAL007C (648 bases). Note that HyenaDNA-large-1M and Nucleotide Transformer are trained on sequences of larger and shorter lengths, respectively. This suggests that the length of training data could play a role in embedding characteristics of these models. DNABERT-S, GROVER, and Gena-LM models perform equally well irrespective of the sequence size.

3.2. Embedding Contextuality

Several distinct clusters were formed for the embeddings from each clade as seen in Figure 3a in Appendix A.1, suggesting that these gLMs were able to pick up characteristic features pertaining to the entire strain from their genes' sequences alone. These results were consistent throughout the models as seen in Figure 3a. It can be noted that the shape of these clusters seemed to be conserved as well.

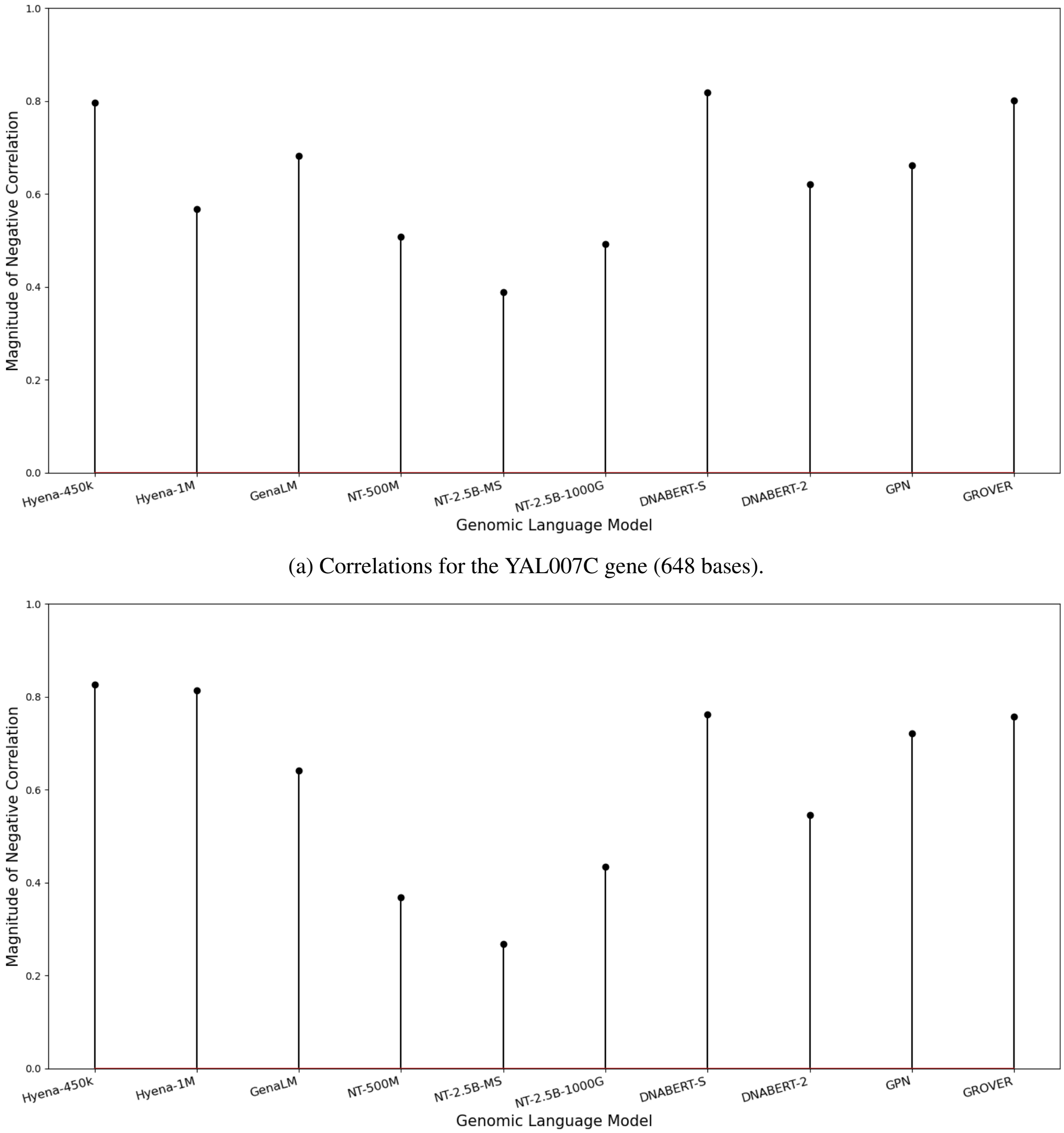
The clustering algorithms, k-means and agglomerative clustering, were able to identify 25-30 clusters, which is close to the total number of clades, i.e., 30. By observing the variation in the quality of these clusterings (measured by metrics such as the silhouette score and the adjusted Rand score) with clustering algorithm parameters, the highest scores were obtained when the number of clusters parameter was close to the total number of clades in the case of the agglomerative clustering, and when the parameter for the number of neighbors within a cluster was close to the median number of strains in each clade, i.e., 17, in the case of k-means clustering (refer to Figure 3 in Appendix A.1). Thus, the embeddings belonging to one clade tended to be closer to each other, complying with our hypothesis for the same.

3.3. Embedding Robustness

The magnitude of distances increases as the number of mutations in the sequence increases, as depicted in Figure 4 in Appendix A.2, agreeing with our hypothesis. One of the most striking patterns observed in our mutation robustness experiments was the consistently larger Euclidean distances between embeddings of mutated and original sequences in non-coding DNA compared to coding DNA across most genomic language models. This suggests that point mutations in non-coding sequences result in more significant shifts in embedding space, implying a higher contextual sensitivity to such changes.

While this was true for most models, the HyenaDNA models were an exception, with coding sequences having marginally higher Euclidean distances in most of them. GPN also showed higher distances for coding sequences. Another peculiar case was that of the Nucleotide Transformer models; the 500M parameter model and 2.5B parameter model trained on Human genome data showed significantly higher distances for non-coding sequences whereas the 2.5B parameter model trained on multi-species genome data had very little difference between the coding and non-coding distances, despite all of them receiving the same input sequences due to their shared context length.

A biologically plausible explanation for this phenomenon lies in the concept of codon degeneracy. In protein-coding DNA, the genetic code is redundant as multiple codons can encode the same amino acid. Consequently, embeddings of coding sequences may remain relatively stable under such



(a) Correlations for the YAL007C gene (648 bases).

(b) Correlations for the YAL026C gene (4068 bases).

Figure 1. Magnitudes of the overall negative correlations between the gene sequences' pairwise Needleman-Wunsch alignment scores and the Euclidean distances between their corresponding embeddings, for each model.

mutations. In contrast, non-coding DNA lacks this redundancy. Even a single-nucleotide substitution in a non-coding region can disrupt regulatory motifs or structural elements that are crucial for gene expression and genome architecture. Hence, point mutations in non-coding sequences are more likely to cause shifts in functional meaning, which in turn is reflected in larger embedding distances.

In order to further confirm these findings on a larger scale, we repeated the experiment with 1000 different sequences for each mutation percentage, each of which was randomly mutated 100 times so as to ensure consistency in the results. The results of this experiment are shown in Figure 5 in Appendix A.2, where we have compared the distribution of all distances pertaining to each type of sequence. Using the Kolmogorov-Smirnov test on each pair of distributions yielded a p-value of zero, confirming that they were significantly different.

3.3.1. Coding/Non-coding Classification

To see if these embeddings are indeed different for coding and non-coding sequences, we generated embeddings for 20000 coding and 20000 non-coding sequences from the human reference genome, in the hopes that this would be a fairer evaluation as as all the selected gLMs contain human sequences in their pretraining data. Using these embeddings, we trained a random forest classifier on two different train-test splits, of 80-20 (one-shot) and 1-99 (pseudo zero-shot).

Across both settings, the classifier achieved significantly higher-than-chance accuracy for nearly all models as depicted in Figure 6 in Appendix A.2, with some (e.g., HyenaDNA variants) reaching accuracies above 90% in the one-shot setting and maintaining robust performance even in the pseudo zero-shot regime. These results indicate that the embeddings produced by gLMs inherently encode features that distinguish coding from non-coding sequences, without requiring any fine-tuning. This finding is consistent with the mutation robustness analysis, where coding and non-coding sequences showed systematically different embedding behaviors. Together, these experiments suggest that gLMs are not only capable of capturing localized sequence patterns but may also be encoding functional genomic context that differentiates regulatory and protein-coding regions.

4. Discussion

In this preliminary study, a suite of experiments was proposed to intrinsically evaluate the embeddings of the pretrained gLMs. This study sheds light on how the gLMs represent the DNA sequences. These gLMs models considered in this study accurately group similar sequences and dissimilar sequences, even when the difference is only a change in a small number of nucleotides. Further analysing the effect of the mutations on the representational capabilities of the gLMs, it is observed that many of these models can distinguish coding and non-coding sequences without explicitly being trained on these data separately. This points to the fact that these models are able to understand the biological differences between the different types of sequences. Further, this analysis will aid in understanding how exactly these models learn the differences between these sequences, independent of downstream tasks. In conclusion, our work serves as a stepping stone towards better understanding genomic language models.

5. Impact Statement

This work investigates an important question of the biologically relevant of embeddings generated by a set of genome language models (gLMs) on Yeast genomes. In contrast to the existing studies, which are based on biologically relevant downstream tasks, the study emphasises the intrinsic evaluations of gLMs through a set of experiments. Hence, this study will be useful in establishing intrinsic benchmarks for evaluating the learning of gLMs. However, one of the main limitations of this study is a smaller set of experiments, and these experiments have to be validated at a larger scale to support the stronger conclusions made in this work.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Benegas, G., Batra, S. S., and Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023. doi: 10.1073/pnas.2311219120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311219120.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Fishman, V., Kuratov, Y., Shmelev, A., Petrov, M., Penzar, D., Shepelin, D., Chekanov, N., Kardymon, O., and Burtsev, M. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1310. URL https://doi.org/10.1093/nar/gkae1310.

Inoue, J. and Saitou, N. dbCNS: A new database for conserved noncoding sequences. *Molecular Biology and Evolution*, 38(4):1665–1676, November 2020. ISSN 1537-1719. doi: 10.1093/molbev/msaa296. URL http://dx.doi.org/10.1093/molbev/msaa296.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Marin, F. I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., and Boomsma, W. BEND: Benchmarking DNA Language Models on biologically meaningful tasks. *The Twelfth International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:265308711.

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970. ISSN 0022-2836. doi: 10.1016/0022-2836(70)90057-4. URL http://dx.doi.org/10.1016/0022-2836(70)90057-4.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36: 43177–43201, 2023.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., and Schacherer, J. Genome evolution across 1, 011 saccharomyces cerevisiae isolates. *Nature*, 556(7701):339–344, April 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0030-5. URL http://dx.doi.org/10.1038/s41586-018-0030-5.

Sanabria, M., Hirsch, J., Joubert, P. M., and Poetsch, A. R. DNA language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00872-0. URL http://dx.doi.org/10.1038/s42256-024-00872-0.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*.

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., Wang, Z., and Liu, H. DNABERT-S: Pioneering species differentiation with species-aware DNA embeddings, 2024. URL https://arxiv.org/abs/2402.08777.

A. Appendix

A.1. Clade Clustering Plots

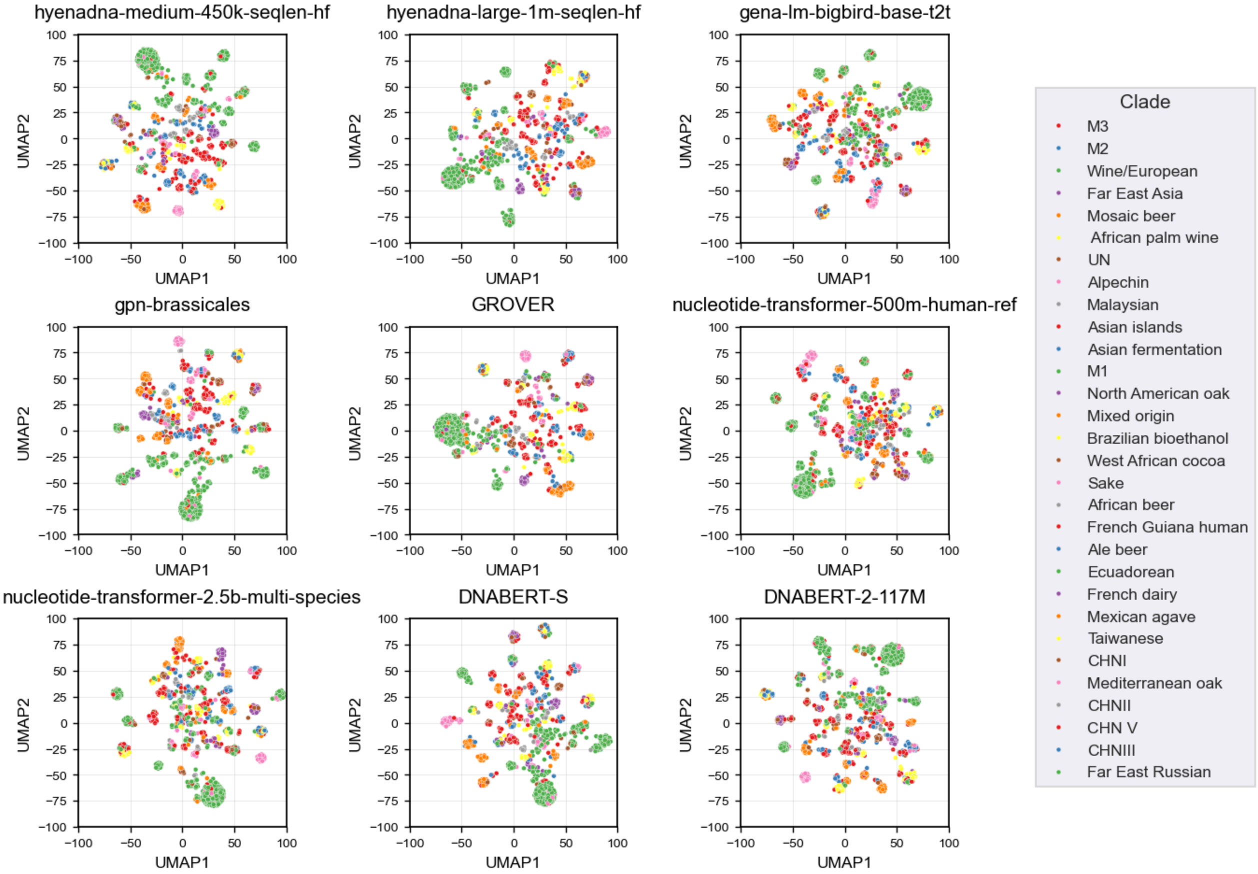
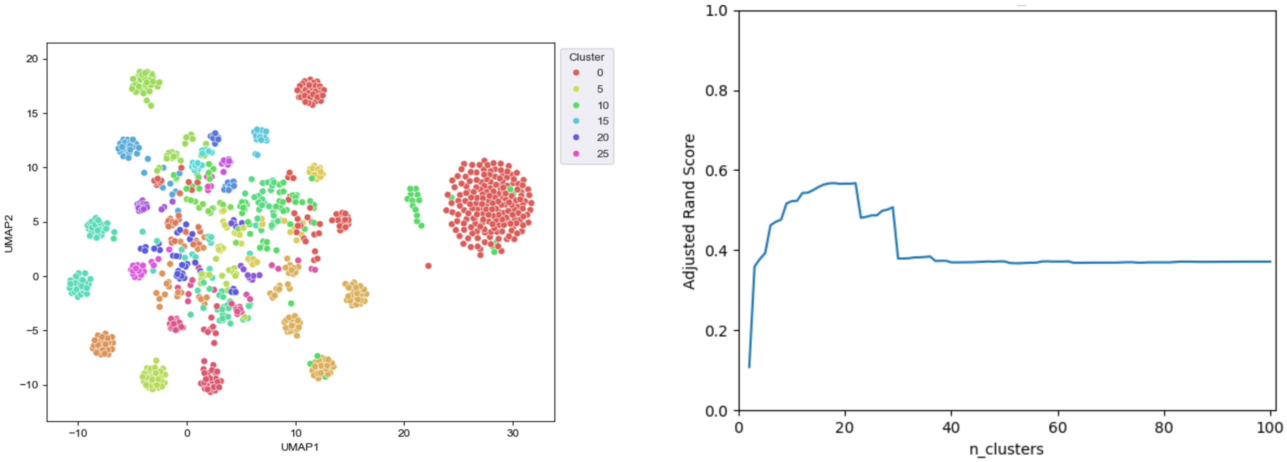


Figure 2. Scatter plots of the models' UMAP-reduced embeddings for the YAL001C gene sequence from each of the 1011 strains, labelled according to the clade they belong to.



(a) Agglomerative clustering of UMAP-reduced embeddings.

(b) Adjusted rand scores v/s no. of clusters for agglomerative clustering.

Figure 3. Variation in clustering quality with parameters.

A.2. Mutation Sensitivity Plots

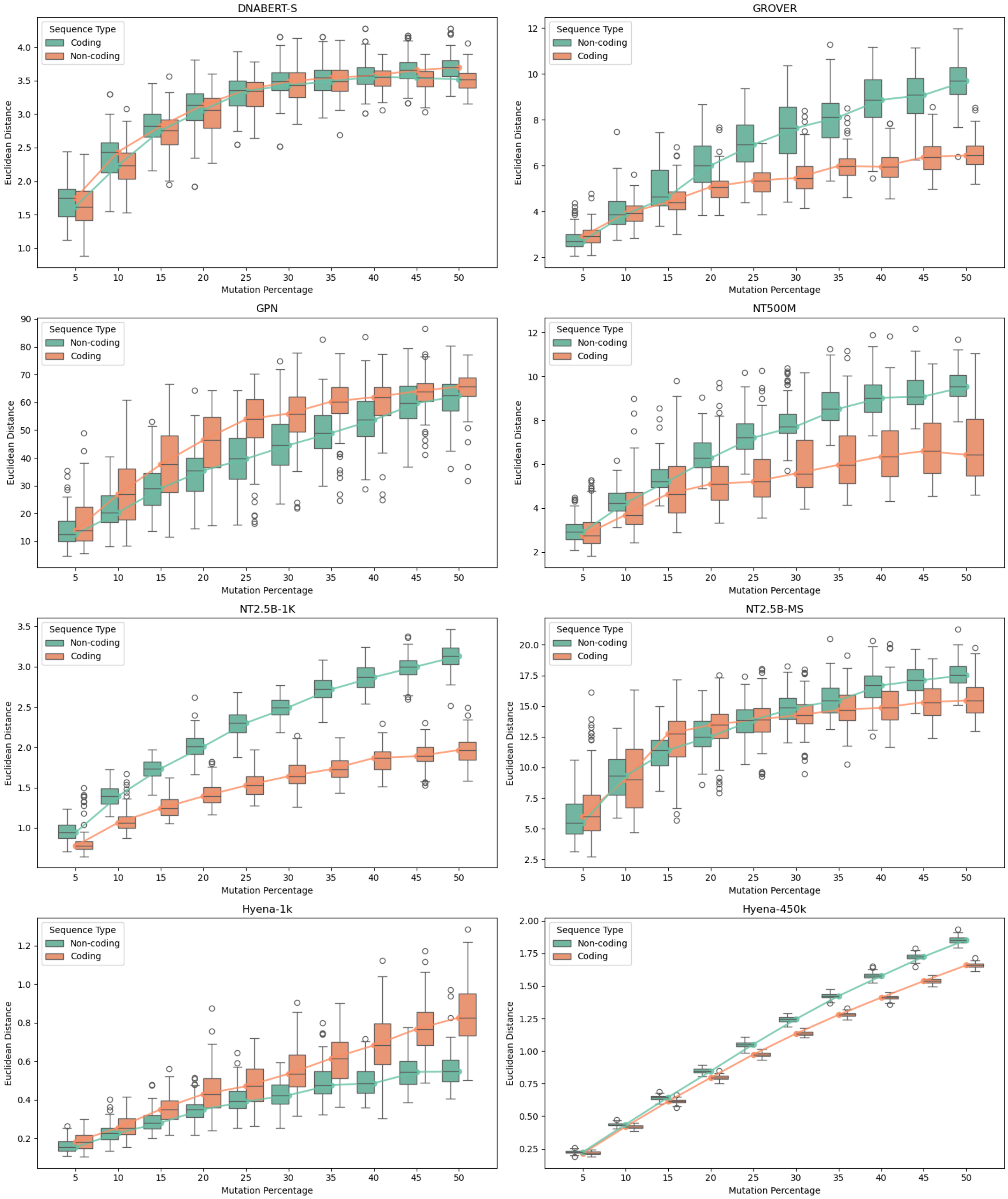


Figure 4. Results from our third experiment. Here, each boxplot corresponds to the distribution of Euclidean distances between the embeddings of the 100 randomly mutated sequences at that value of mutation percentage and the original sequence's embedding.

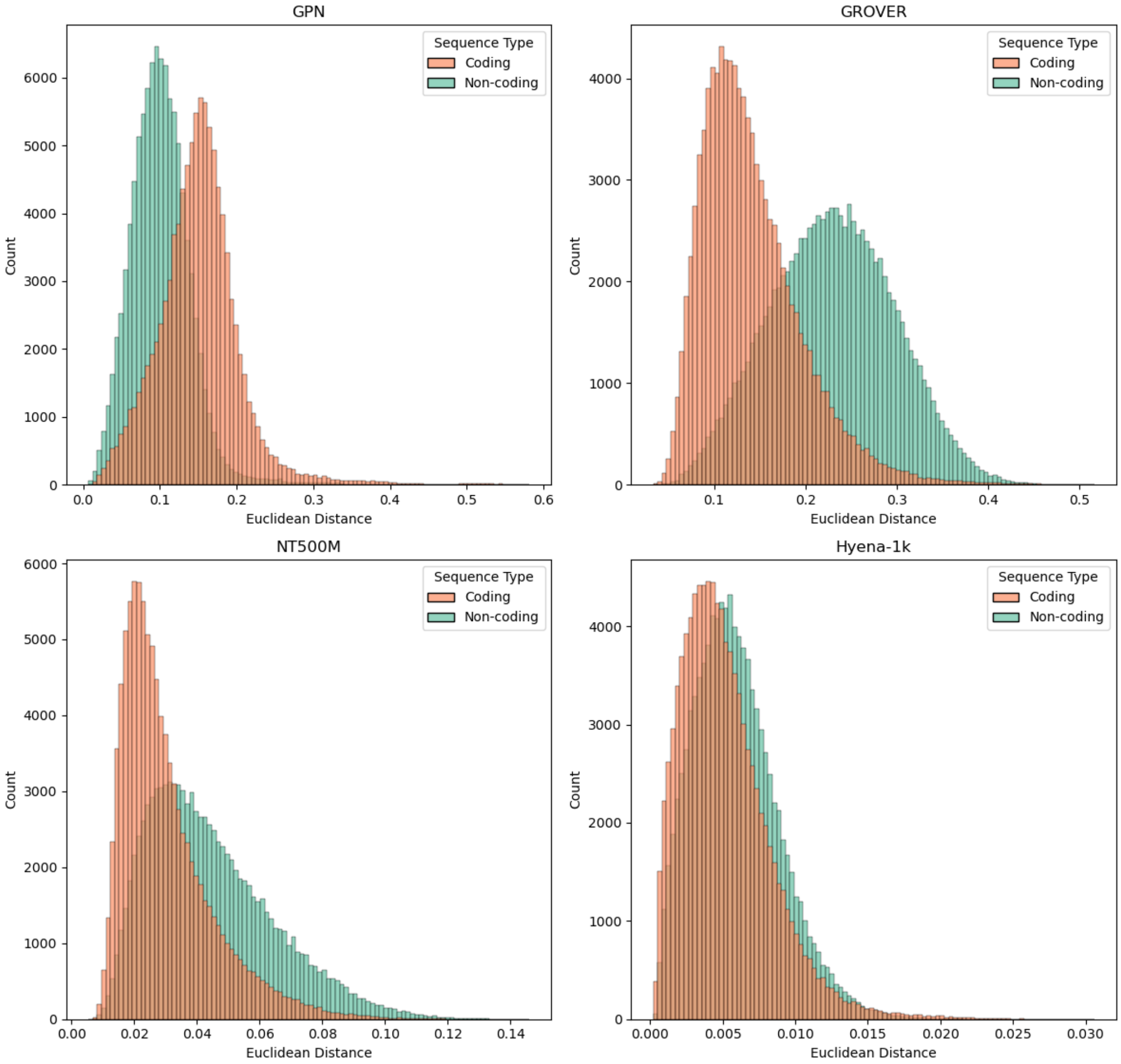
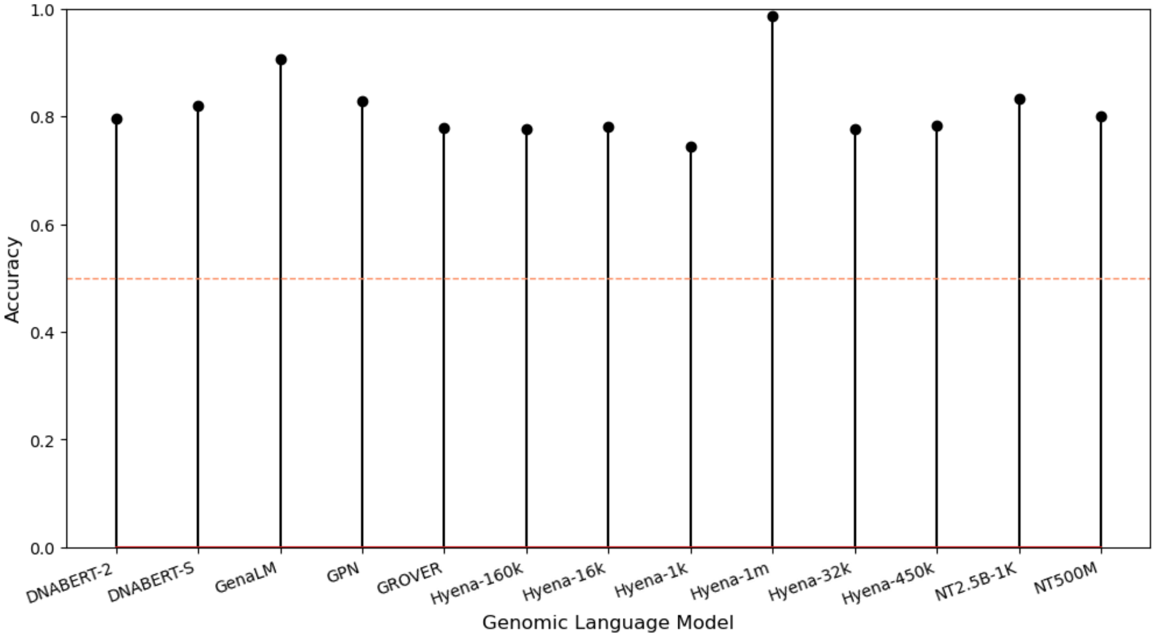
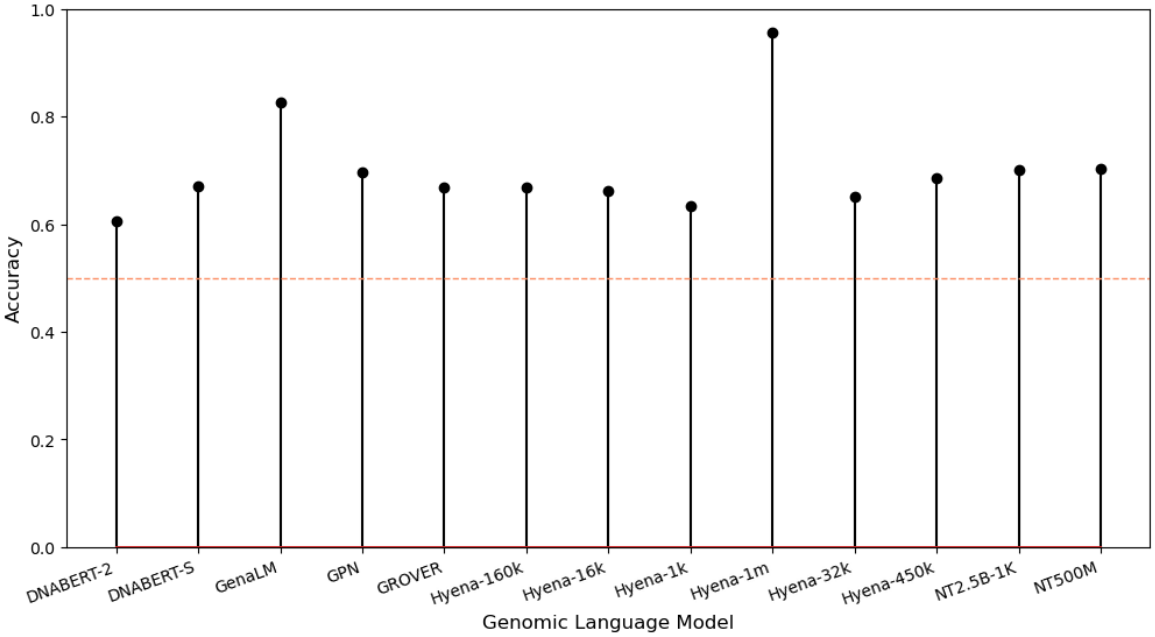


Figure 5. Distributions of all distances for coding and non-coding sequences, for 1000 input sequences at each mutation percentage level.



(a) One-shot classification, with a train-test split of 80-20.



(b) Pseudo Zero-shot classification, with a train-test split of 1-99.

Figure 6. Classification accuracies of a random forest classifier trained on human coding and non-coding sequence embeddings.