

---

# Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide Analysis

---

**Mohammdsaleh Refahi**  
Drexel University  
Philadelphia, PA

**Mahdi Abavisani**  
Dataminr  
New York, NY

**Bahrad A. Sokhansanj**  
Drexel University  
Philadelphia, PA

**James R. Brown**  
Drexel University  
Philadelphia, PA

**Gail Rosen**  
Drexel University  
Philadelphia, PA

## Abstract

Transformers have revolutionized nucleotide sequence analysis, yet capturing long-range dependencies remains challenging. Recent studies show that autoregressive transformers often exhibit Markovian behavior by relying on fixed-length context windows for next-token prediction. However, standard self-attention mechanisms are computationally inefficient for long sequences due to their quadratic complexity and do not explicitly enforce global transition consistency.

We introduce CARMANIA (Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide Analysis), a self-supervised pretraining framework that augments next-token (NT) prediction with a transition-matrix (TM) loss. The TM loss aligns predicted token transitions with empirically derived  $n$ -gram statistics from each input sequence, encouraging the model to capture higher-order dependencies beyond local context. This integration enables CARMANIA to learn organism-specific sequence structures that reflect both evolutionary constraints and functional organization.

We evaluate CARMANIA across diverse genomic tasks, including regulatory element prediction, functional gene classification, taxonomic inference, antimicrobial resistance detection, and biosynthetic gene cluster classification. CARMANIA outperforms the previous best long-context model by at least 7%, matches state-of-the-art on shorter sequences (exceeding prior results on 20/40 tasks while running  $\sim 2.5\times$  faster), and shows particularly strong improvements on enhancer and house-keeping gene classification tasks—including up to a 34% absolute gain in Matthews correlation coefficient (MCC) for enhancer prediction. The TM loss boosts accuracy in 33 of 40 tasks, especially where local motifs or regulatory patterns drive prediction. This enables more effective modeling of sequence-dependent biological features while maintaining robustness across non-coding and low-signal regions. Code available at <https://github.com/EESI/carmania>.

## 1 Introduction

Deoxyribonucleic acid (DNA), often referred to as the "language of life," encodes the genetic instructions essential for the development, functioning, and reproduction of all known living organisms and many viruses. The sequence of its four nucleotide bases—adenine (A), thymine (T), cytosine (C), and guanine (G)—forms the foundation of genetic information, dictating the synthesis of proteins and the regulation of various biological processes. Advancements in high-throughput sequencing technologies have exponentially increased genomic data availability.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Despite advances in sequencing, analyzing large-scale biological data remains challenging. Traditional methods—such as motif detection, sequence alignment, and Markov models—capture short-range dependencies well [Delcher et al. [2007], Steinegger and Söding [2017], Wood et al. [2019]], but often fail to represent complex or long-range patterns, especially in diverse or non-canonical sequences. Recent work has applied NLP techniques to biological sequences [Nguyen et al. [2024], Cahyawijaya et al. [2022], Schiff et al. [2024], Bo et al. [2025]], with large language models (LLMs) showing strong ability to capture complex dependencies. However, genomic sequences are often extremely long, making full attention computationally infeasible. To address this, transformer models use mechanisms like sliding window and sparse attention [Dong et al. [2025], Zaheer et al. [2020]], which expand the receptive field but may still struggle to capture long-range dependencies effectively in genomics.

In this work, we enhance transformers by incorporating Markovian priors, guiding them to follow long-range transition dynamics and better capture complex dependencies within the broader context of genomic sequences. Drawing on recent studies that interpret transformers' capabilities through the lens of Markovian processes [Hu et al. [2024], Zekri et al. [2024]], we propose a novel pretraining strategy that explicitly enforces these Markovian properties.

Our approach, named CARMANIA (Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide Analysis), supplements the standard next-token prediction objective with an auxiliary loss. This loss aligns the model's predicted first-order transition matrix with one derived from the full input sequence, encouraging consistency with its empirical token transition patterns. By doing so, CARMANIA captures the probabilistic structure of genomic sequences and improves context-awareness during next-token prediction.

To scale pretraining to long genomic sequences, we replace full attention with **sliding-window attention (SWA)**, a sparse mechanism that limits each token's receptive field to a fixed-size window of preceding tokens, reducing complexity from  $O(n^2)$  to  $O(kn)$  ( $k \ll n$ ) while preserving biologically meaningful local interactions. **TM loss** complements this by reinforcing short-range dependencies within the local window while promoting global consistency, since the transition matrix is computed over the entire input sequence. Together, these components enable a balance between local motif recognition and broader statistical alignment. *In practice, CARMANIA handles effective contexts up to 160 kbp, which makes it, to our knowledge, the longest-context transformer-based genomic language model to date.* An overview of our framework is illustrated in Figure 1.

In addition to improving context modeling, our pretraining approach supports efficient domain adaptation, as conserved transition patterns across species enable generalization with minimal fine-tuning [Bergeron et al. [2023]].

The main contributions of this work are:

- **A scalable architecture for long-context genomic modeling** that integrates sliding-window attention, RoPE-based positional encoding, FlashAttention, and a widened Transformer backbone, enabling efficient representation learning over extended DNA sequences.
- **A novel self-supervised objective based on Transition Matrix (TM) loss**, which regularizes the transformer model by aligning predicted token transitions with empirical n-gram distributions. This Markovian prior enhances memory retention, out-of-distribution generalization, and biological coherence across long genomic contexts.
- **Comprehensive evaluation across 40 genomic tasks**, including regulatory element prediction, functional gene classification, taxonomic inference, and biosynthetic gene cluster(BGC) prediction, where our model achieves consistent improvements—even in benchmarks where alternative inductive biases have previously dominated.

## 2 Related Work

### 2.1 Markovian Models in Genomics

Hidden Markov Models (HMMs) are widely used in genomics for modeling the probabilistic structure of biological sequences. They enable gene prediction, as in GeneMark [Lukashin and Borodovsky

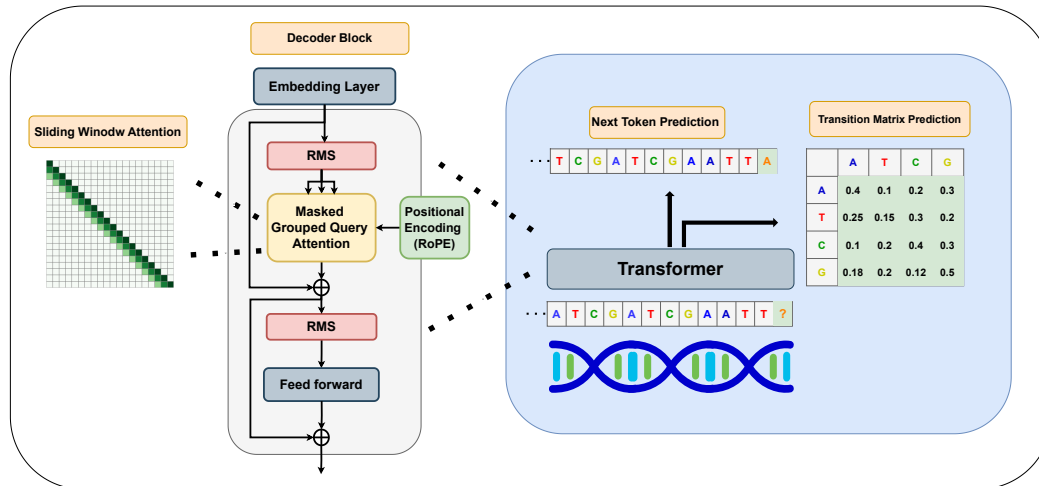


Figure 1: **Proposed Pretraining Framework.** We extend a LLaMA-style decoder with a transition matrix module to capture global nucleotide co-occurrence. The model uses sliding-window attention with rotary embeddings and local caching, reducing attention complexity from  $O(n^2)$  to  $O(n)$ . The transition matrix complements local attention by preserving long-range dependencies efficiently.

[1998], and other methods for identifying coding regions Delcher et al. [2007]. Profile HMMs are commonly applied to protein family analysis, capturing conserved regions and domain variability Potter et al. [2018]. HMMs also power microbial classification tools like Clasnip Chuan et al. [2023], and Aphmm Firtina et al. [2024] improves the speed and efficiency of profile HMM computations. Their effectiveness depends on model structure, sequence length, and the biological complexity of the task.

## 2.2 Language Models in Genomics

The first breakthrough in genomic language models came with DNABERT Ji et al. [2021], which employed a BERT architecture with k-mer tokenization and sequence masking to learn DNA sequence representations. The Nucleotide Transformer Dalla-Torre et al. [2024] scaled this concept, training a 2.5 billion parameter model on 850 genomes. Other models like HyenaDNA Nguyen et al. [2024] used long convolutional blocks for efficient long-sequence handling, while BigBird Refahi et al. [2025, 2023] applied sparse and sliding window attention to capture long-range dependencies. Caduceus Schiff et al. [2024], built on Mamba, introduced a bi-directional model preserving reverse complement symmetry.

Together, these advances in tokenization, architecture, and representation learning have led to increasingly effective models for genomic analysis Cahyawijaya et al. [2022], Celikkanat et al. [2024].

## 2.3 The Markovian Core of Autoregressive Transformers

There has been significant research Edelman et al. [2024], Zekri et al. [2024], Ildiz et al. [2024], Hu et al. [2024] aimed at establishing an equivalence between autoregressive language models based on transformers and Markov chains. Specifically, findings emphasize the role of statistical induction heads, which adaptively learn in-context patterns, forming a foundation for next-token predictions. This perspective introduces a compelling connection between autoregressive architectures and Markov chains, paving the way for a nuanced understanding of in-context learning (ICL). Zekri et al. [2024], Edelman et al. [2024]. Zekri et al. [2024] demonstrates that an autoregressive model with a vocabulary size  $T$  and a context window of size  $K$  can be represented as a Markov chain with a state space of size  $O(T^K)$ .

In an autoregressive model, the probability of predicting the next token  $x_{t+1}$  given the previous  $K$  tokens  $(x_t, x_{t-1}, \dots, x_{t-K+1})$  can be written as:

$$\Pr(x_{t+1} \mid x_t, x_{t-1}, \dots, x_{t-K+1}). \quad (1)$$

This conditional probability can be interpreted as a transition probability in a  $K$ -th order Markov chain, where each unique  $K$ -length token sequence defines a state. The transition matrix  $Q \in \mathbb{R}^{T \times T}$  is defined as:

$$Q_{s,s'} = \Pr(x_{t+1} = x_{s'} \mid x_t = x_s), \quad (2)$$

where  $s$  and  $s'$  index sequences of  $K$  tokens. The rows of  $Q$  sum to 1, satisfying the Markov property. Thus, the model's behavior approximates a Markov chain with  $T^K$  states, capturing dependencies within the fixed context window. Although self-attention is globally contextual and non-Markovian in theory, limited context in practice causes autoregressive Transformers to approximate Markovian behavior Ildiz et al. [2024].

### 3 Pretraining Long Sequences With Markovian Knowledge

We adopt a causal Transformer inspired by LLaMATouvron et al. [2023], where each layer computes hidden representations  $\mathbf{h}^{(l)}$  from input tokens  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  and the previous layer's outputs  $\mathbf{h}^{(l-1)}$ , using multi-head self-attention to predict the next token. To scale to long sequences, we replace full self-attention with SWA, where each token attends only to a fixed-size  $k$  window of preceding tokens,  $x_{t-k}, \dots, x_{t-1}$ , reducing complexity from  $O(n^2)$  to  $O(kn)$ . This not only improves efficiency but also reflects the local nature of most biological dependencies in genomic sequences. We further improve efficiency with FlashAttention-2 Dao [2023] and enhance positional encoding via Rotary Positional Embeddings (RoPE) Su et al. [2024], enabling better extrapolation to long contexts.

#### 3.1 Unified $n$ -th Order Transition Tensor Formulation

To complement the next-token prediction objective, we introduce a unified probabilistic framework that captures higher-order dependencies through a transition tensor. Let  $\mathbf{P}_t \in \mathbb{R}^V$  denote the model's predicted probability distribution over the vocabulary  $\mathcal{V}$  at position  $t$ , where  $[\mathbf{P}_t]_i = \mathbb{P}(x_t = i \mid x_{<t})$ .

We define the  $n$ -th order transition tensor  $\mathcal{T}^{(n)} \in \mathbb{R}^{V^n}$  as:

$$\mathcal{T}_{i_1 i_2 \dots i_n}^{(n)} = \frac{1}{B(L-n+1)} \sum_{b=1}^B \sum_{t=1}^{L-n+1} \prod_{s=0}^{n-1} [\mathbf{P}_{t+s}^{(b)}]_{i_{s+1}}, \quad (3)$$

where  $B$  is the batch size,  $L$  is the sequence length,  $V$  is the vocabulary size, and  $i_1, \dots, i_n$  index token values in the  $n$ -gram.

This formulation approximates the joint distribution over  $n$  consecutive model predictions, providing a global summary of sequential dynamics. In practice, we focus on the first-order case ( $n = 2$ ), which corresponds to modeling bigram transitions. This strikes a balance between expressiveness and computational efficiency, and directly supports the TM loss introduced in the next section.

#### 3.2 Pretraining Objective

Our pretraining objective combines two complementary components: next-token prediction and transition matrix alignment. The primary objective is the standard autoregressive language modeling loss, which minimizes the negative log-likelihood of predicting each token  $x_t$  given its preceding context:

$$\mathcal{L}_{\text{NT}} = - \sum_{t=1}^n \log P_{\theta}(x_t \mid x_1, \dots, x_{t-1}), \quad (4)$$

where  $P_{\theta}$  denotes the model's predicted token distribution.

To reinforce global sequence structure, we introduce a Transition Matrix (TM) loss that encourages the model to match the empirical token transition dynamics. Let  $p_{ij}$  and  $q_{ij}$  denote the smoothed  $n$ -gram distributions of the input and model outputs, respectively. The KL divergence between these distributions defines the TM loss:

$$\mathcal{L}_{\text{TM}} = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (5)$$

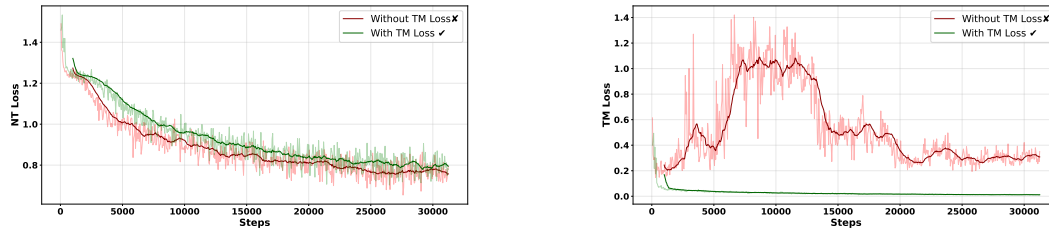


Figure 2: Comparison of training losses with and without explicit TM loss enforcement ( $\beta = 1$  vs.  $\beta = 0$ ). **Left:** Both models show similar reductions in next-token prediction loss. **Right:** Without TM loss ( $\beta = 0$ ), the model partially learns transition structure—TM loss rises then stabilizes—indicating implicit alignment with n-gram patterns.

Table 1: Ablation study showing the impact of TM loss and attention type on model performance (F1 Macro, BLEU) and compute cost. All models are trained on the Scorpio-Gene-Taxa dataset. F1 Macro is evaluated on the Antimicrobial Resistance (AMR) classification tasks, while BLEU and Relative FLOPs are computed on the Human Genome-Short dataset using 10,000 fragments from the GRCh38/hg38 reference genome. Relative FLOPs are reported with respect to the Sliding Window attention baseline, which has an absolute compute cost of approximately  $16.7 \times 10^{12}$  FLOPs per sequence.

| Model Setting              | F1 Macro ( $\uparrow$ ) | BLEU ( $\uparrow$ ) | Relative FLOPs ( $\downarrow$ ) |
|----------------------------|-------------------------|---------------------|---------------------------------|
| Transformer (SWA)          | $0.873 \pm 0.130$       | 0.73                | 1.00                            |
| + TM Loss                  | $0.882 \pm 0.131$       | 0.77                | 1.00                            |
| + Full Attention + TM Loss | $0.883 \pm 0.131$       | 0.71                | 1.58                            |

The full loss combines both terms:

$$\mathcal{L}_{\text{Full}} = \mathcal{L}_{\text{NT}} + \beta \mathcal{L}_{\text{TM}}, \quad (6)$$

where  $\beta$  is a tunable hyperparameter controlling the influence of the TM loss.

## 4 Experiments

In this section, we evaluate CARMANIA across a range of genomic tasks. We first describe the pretraining datasets, model architecture, and training setup, followed by ablation studies assessing the impact of the TM loss and attention mechanism on convergence, efficiency, and long-range retention. Finally, we present comprehensive benchmark results demonstrating the model’s adaptability, generalization, and superior performance over existing methods.

### 4.1 Experimental Setup

#### Pre-training Datasets

We pre-train our model on three large-scale genomic datasets: (i) GRCh38 GRCh38 [2013], which provides  $\sim 3\text{B}$  base pairs with 10 kbp and 160 kbp fragments; (ii) the Basic Genome Dataset Zhu et al. [2022], consisting of  $\sim 10\text{B}$  base pairs across 4,600+ genomes with 10 kbp fragments; and (iii) the Scorpio Gene-Taxa Dataset Refahi et al. [2025], comprising  $\sim 580\text{M}$  base pairs from 2,046 genomes with 4 kbp fragments. Additional details are in Appendix A.

#### Model and Training Details

**Architecture:** CARMANIA is a LLaMA-based causal Transformer tailored for genomic sequence modeling. It uses 16 attention heads (4 key-value), a window size of 128, and 5 custom Transformer layers with an embedding size of 1024 and intermediate dimension of 4608. The model uses SiLU activations and contains 83M parameters. We selected this wide architecture based on the comparative results in Supplementary Table 12, which demonstrate its superior performance and efficiency over deeper models in both in-domain and out-of-domain genomic tasks.

**Tokenization and n-gram Frequency Calculation:** We tokenize DNA sequences at the single-nucleotide level (A, T, C, G) to preserve fine-grained features such as SNPs and avoid disrupting open

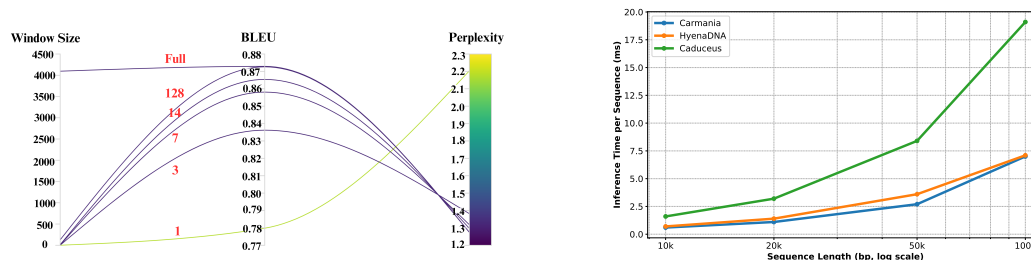


Figure 3: **Left:** Effect of window size on model performance. A window size of 128 achieves results comparable to full attention. **Right:** Inference time per sequence for CARMANIA(83M), HyenaDNA(1.6M), and Caduceus-PH(1.9M) across varying sequence lengths.

reading frames, as can occur with subword or non-overlapping k-mer methods. Additionally, during tokenization, we compute a normalized  $4 \times 4$  first-order transition matrix for each sequence based on bigram frequencies, where each row defines a probability distribution over possible next nucleotides. This sequence-specific matrix acts as a self-supervised training signal, serving as ground truth for the transition-matrix loss and guiding the model to learn biologically meaningful transition patterns.

**Training Parameters, and Hardware Specifics:** The model was trained for two epochs using PyTorch on an NVIDIA A100 GPU (80GB). To fit within memory constraints, batch sizes were set based on sequence length: 35 for 4 kbp, 19 for 10 kbp, and 1 for 160 kbp inputs. We used a cosine annealing schedule with an initial learning rate of  $5e-4$ . Additional training details, including optimizer settings and gradient clipping, are listed in Supplementary Table 13. The training setup yields an efficiency of approximately  $7.57 \times 10^{-9}$  GPU-hours per token.

## 4.2 Ablation Studies

### Effect of Transition Matrix Loss

We study the effect of adding first-order TM loss alongside the next-token (NT) loss by training on the Scorpio-Gen-Taxa dataset for 31k steps under two settings: with TM loss ( $\beta = 1$ ) and without it ( $\beta = 0$ ). In the  $\beta = 0$  case, TM loss is computed but not used during training, allowing us to observe the model's natural transition behavior. The left panel of Figure 2 shows NT loss decreases similarly for both models, indicating that adding TM loss does not hinder token-level learning. In contrast, the right panel reveals different TM loss patterns: for the  $\beta = 0$  model, the TM loss initially increases then decreases as the model partially and passively captures transition structure, while for the  $\beta = 1$  model, the TM loss consistently decreases from the start. However, as shown in supplementary results, convolution-based models like HyenaDNA require explicit TM regularization to exhibit this property.

Finally, as reported in Table 1, models trained with TM loss achieve higher macro F1 scores on downstream Antimicrobial Resistance (AMR) classification tasks (detailed in Supplementary Table 18) and improved BLEU scores on the human genome benchmark. We also evaluated this effect across all 40 downstream benchmarks and found that adding TM loss led to improved performance in 33 of them.

**Higher-Order Transition Matrices.** As shown in Supplementary Tables 10 and 11, applying second-order TM loss led to performance degradation across tasks compared to first-order TM or no TM. This drop is likely due to the sparsity of higher-order transitions in biological sequences, which makes the loss signal unstable. In contrast, first-order TM loss consistently improves performance by capturing reliable and biologically meaningful pairwise dependencies.

### Effect of Windowed vs. Full Attention

We assess the effect of attention window size by training on a 100M bp subset of the Scorpio-Gen-Taxa dataset. As shown in Figure 3, larger windows improve BLEU scores and reduce perplexity, with full attention as an upper bound. Notably, a window size of 128 performs comparably to full attention, indicating that TM loss enables smaller windows to capture complex dependencies efficiently. Figure 3 also illustrates inference time comparisons across models on input sequences of varying lengths, up to 100k bases. Despite having 83M parameters, our model runs approximately

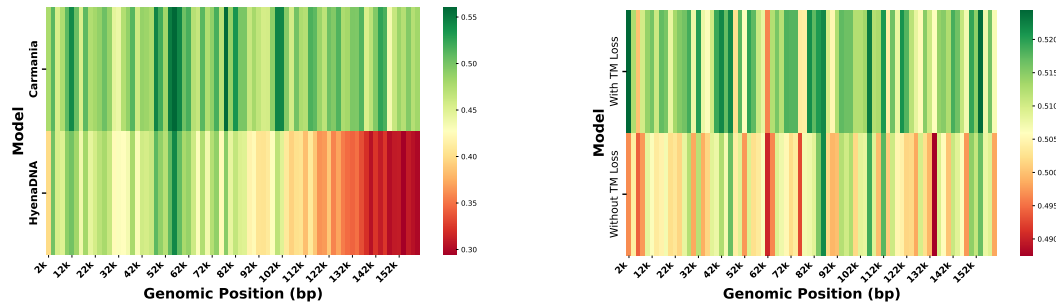


Figure 4: Heatmap of average sequence similarity across 50 independent 160 kbp genomic segments in 100 bp windows at 2000 bp intervals. **Left:** CARMANIA consistently maintains high similarity across all regions, demonstrating superior long-range memory, whereas HyenaDNA shows reduced sequence coherence in later segments. **Right:** Incorporating the TM loss improves memory retention in CARMANIA, yielding higher sequence similarity across extended contexts.

2.5 times faster than Caduceus becomes significantly slower as sequence length increases, while HyenaDNA (1.6M) maintains relatively fast inference but underperforms on long sequences and remains slower than our model. This demonstrates the practical efficiency of our architecture, and also highlights how windowed attention enables faster inference in Transformer-based models.

Table 1 compares the performance of models trained with full attention and windowed attention. While full attention achieves slightly better performance on downstream tasks, it results in a lower BLEU score. In addition, we evaluate the computational efficiency of windowed attention. As shown in Table 1, full attention requires 58% more computational resources compared to a window size of 128. This demonstrates the trade-off between accuracy and efficiency, as windowed attention reduces complexity from  $O(n^2)$  in full attention to  $O(k^2n)$ , making it a more practical and scalable choice for large-scale sequence modeling.

### 4.3 Long-Range Sequence Retention

We evaluate each model's ability to preserve long-range sequence structure by measuring internal consistency across extended human genome regions. Specifically, we analyze 50 independent 160 kbp genomic segments. Within each segment, we extract 100 bp windows using a fixed stride of 2000 bp and compare the model-predicted sequences against the corresponding regions in the original input sequence using Hamming similarity. This produces a position-wise similarity profile, which we average across all segments to assess how well each model maintains fidelity over increasing genomic distance.

As shown in Figure 4, HyenaDNA shows a marked drop in sequence similarity at distal positions. This likely reflects the limitations of its fixed-depth recurrence architecture, where information must propagate layer by layer without explicit token-to-token interaction Nguyen et al. [2024], making it harder to retain detailed signals over long spans. In contrast, CARMANIA maintains high similarity across the entire sequence due to its sliding-window attention and the TM loss, which encourages consistency in token transitions. The effect of the TM loss is further illustrated in Figure 4, where models trained without it show noticeably shorter retention spans, suggesting its role in supporting full-sequence memorization.

### 4.4 Downstream Task Evaluations

#### Genomics Benchmark: Supervised Adaptation

To evaluate task-specific adaptation, we fine-tuned our model on diverse genomics classification datasets. The Genomic Benchmarks collection includes tasks such as regulatory element prediction, enhancer detection, and binary species classification Grešová et al. [2023]. For a fair comparison, we fine-tuned the human-pretrained CARMANIA model using 5-fold cross-validation and compared its performance to other state-of-the-art models trained on the human genome. As shown in Table 2, CARMANIA achieves top performance in 4 out of 8 tasks and matches the best performance in 2 others, demonstrating its strong generalization ability in supervised downstream settings.



Table 2: Top-1 accuracy ( $\uparrow$ ) across 5-fold cross-validation on Genomic Benchmark tasks. Results for CNN, HyenaDNA, Mamba, and both Caduceus variants are taken from the original Caduceus paper [Schiff et al., 2024]. Underlined values indicate second-best results.

| Task                    | CNN               | HyenaDNA                            | Mamba             | Caduceus-PH                         | Caduceus-PS                         | CARMANIA                            | Baseline w/o TM                     |
|-------------------------|-------------------|-------------------------------------|-------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Mouse Enhancers         | 0.715 $\pm$ 0.087 | <u>0.780 <math>\pm</math> 0.025</u> | 0.743 $\pm$ 0.054 | 0.754 $\pm$ 0.074                   | <b>0.793 <math>\pm</math> 0.058</b> | 0.761 $\pm$ 0.019                   | 0.748 $\pm$ 0.020                   |
| Coding vs. Intergenic   | 0.892 $\pm$ 0.008 | 0.904 $\pm$ 0.005                   | 0.904 $\pm$ 0.004 | 0.915 $\pm$ 0.003                   | 0.910 $\pm$ 0.003                   | <b>0.935 <math>\pm</math> 0.001</b> | <u>0.930 <math>\pm</math> 0.001</u> |
| Human vs. Worm          | 0.942 $\pm$ 0.002 | 0.964 $\pm$ 0.002                   | 0.967 $\pm$ 0.002 | <b>0.973 <math>\pm</math> 0.001</b> | <u>0.968 <math>\pm</math> 0.002</u> | <u>0.968 <math>\pm</math> 0.003</u> | 0.966 $\pm$ 0.001                   |
| Human Enhancers Cohn    | 0.702 $\pm$ 0.021 | 0.729 $\pm$ 0.014                   | 0.732 $\pm$ 0.029 | <b>0.747 <math>\pm</math> 0.004</b> | <u>0.745 <math>\pm</math> 0.007</u> | 0.724 $\pm$ 0.005                   | 0.699 $\pm$ 0.005                   |
| Human Enhancer Ensembl  | 0.744 $\pm$ 0.122 | 0.849 $\pm$ 0.006                   | 0.862 $\pm$ 0.008 | 0.893 $\pm$ 0.008                   | <u>0.900 <math>\pm</math> 0.006</u> | <b>0.916 <math>\pm</math> 0.002</b> | 0.892 $\pm$ 0.001                   |
| Human Regulatory        | 0.872 $\pm$ 0.005 | 0.869 $\pm$ 0.012                   | 0.814 $\pm$ 0.211 | 0.872 $\pm$ 0.011                   | 0.873 $\pm$ 0.007                   | <b>0.895 <math>\pm</math> 0.002</b> | <u>0.893 <math>\pm</math> 0.002</u> |
| Human OCR Ensembl       | 0.698 $\pm$ 0.013 | 0.783 $\pm$ 0.007                   | 0.815 $\pm$ 0.002 | <b>0.828 <math>\pm</math> 0.006</b> | <u>0.818 <math>\pm</math> 0.006</u> | 0.775 $\pm$ 0.002                   | 0.763 $\pm$ 0.002                   |
| Human NonTATA Promoters | 0.861 $\pm$ 0.009 | 0.944 $\pm$ 0.002                   | 0.933 $\pm$ 0.007 | 0.946 $\pm$ 0.007                   | 0.945 $\pm$ 0.010                   | <b>0.963 <math>\pm</math> 0.002</b> | <u>0.961 <math>\pm</math> 0.003</u> |

## Nucleotide Transformer Tasks: Supervised Adaptation

We benchmarked our model on 18 diverse genomics classification tasks introduced by Dalla-Torre et al. [2024], spanning histone modification prediction, regulatory annotation, and splice site detection. Following their evaluation protocol, we performed 10-fold cross-validation with early stopping, reporting the mean and standard deviation across different random seeds.

As shown in Table 3, CARMANIA achieves top-1 performance on 5 out of 18 tasks and outperforms the average of all baselines by over 3% across metrics. Notably, in enhancer classification tasks, CARMANIA achieves up to 34% absolute improvement in MCC over previous models Zhou et al. [2023], Avsec et al. [2021], Schiff et al. [2024], Nguyen et al. [2024], Dalla-Torre et al. [2024]. These results demonstrate the model’s strong ability to capture regulatory signals from sequence. A notable gain is observed on the H4ac histone mark, where adding the TM loss yields over 40.9% relative improvement in MCC. The benefit of the TM loss, however, is not uniform across all genomics tasks. In enhancer prediction, where positive and negative classes differ strongly in their local motif distributions (KL divergence = 9.28), the TM loss provides no measurable improvement ( $\Delta = 0.0\%$ ). In splice-site donor prediction (KL = 0.99), it yields a modest 5.2% gain, and in the H4ac histone mark dataset (KL = 0.64), a substantial 40.9% improvement. These results demonstrate that TM-driven performance gains inversely track inter-class bigram separability, confirming that TM loss is most effective when local sequence features are ambiguous and long-range dependencies become critical. Biologically, this aligns with our understanding that enhancer regions possess clear, unique motifs that NT loss can easily learn, whereas splice sites rely on short, dispersed sequence cues. In contrast, H4 acetylation involves multiple lysine sites (K5, K8, K12, K16) on the H4 N-terminal tail that act cooperatively to regulate chromatin accessibility Dion et al. [2005], Ma et al. [2024]. Instead of strong individual motifs, these sites follow diffuse and combinatorial sequence patterns, making TM’s global co-occurrence modeling particularly advantageous.

Table 3: Performance comparison ( $\uparrow$ ) on 10-fold cross-validation across 18 tasks, including histone marks, regulatory elements, and splice site prediction. We report MCC for histone/enhancer tasks, F1 for promoter/splice site, and accuracy for the “All” splice task. Baselines follow Caduceus Schiff et al. [2024] and Dalla-Torre et al. [2024]. CARMANIA achieves top performance on 5 tasks and remains competitive elsewhere. Underlined values denote second-best results.

|                               | ENFORMER<br>252M  | DNABERT-2<br>117M | NT-V2<br>500M                     | HYENADNA<br>1.6M  | Caduceus-PH<br>1.9M               | Caduceus-PS<br>1.9M               | CARMANIA<br>83M                     | Baseline w/o TM<br>83M              |
|-------------------------------|-------------------|-------------------|-----------------------------------|-------------------|-----------------------------------|-----------------------------------|-------------------------------------|-------------------------------------|
| <b>Histone Markers</b>        |                   |                   |                                   |                   |                                   |                                   |                                     |                                     |
| H3                            | 0.719 $\pm$ 0.048 | 0.785 $\pm$ 0.033 | 0.784 $\pm$ 0.047                 | 0.779 $\pm$ 0.037 | <b>0.815<math>\pm</math>0.048</b> | <u>0.799<math>\pm</math>0.029</u> | 0.782 $\pm$ 0.023                   | 0.785 $\pm$ 0.011                   |
| H3K14AC                       | 0.288 $\pm$ 0.077 | 0.516 $\pm$ 0.028 | 0.551 $\pm$ 0.021                 | 0.612 $\pm$ 0.065 | <b>0.631<math>\pm</math>0.026</b> | 0.541 $\pm$ 0.212                 | <u>0.627<math>\pm</math>0.021</u>   | 0.631 $\pm$ 0.022                   |
| H3K36ME3                      | 0.344 $\pm$ 0.055 | 0.591 $\pm$ 0.020 | 0.625 $\pm$ 0.013                 | 0.613 $\pm$ 0.041 | 0.601 $\pm$ 0.129                 | 0.609 $\pm$ 0.109                 | <b>0.632 <math>\pm</math> 0.011</b> | <u>0.629 <math>\pm</math> 0.017</u> |
| H3K4ME1                       | 0.291 $\pm$ 0.061 | 0.511 $\pm$ 0.028 | <b>0.550<math>\pm</math>0.021</b> | 0.512 $\pm$ 0.024 | <u>0.523<math>\pm</math>0.039</u> | 0.488 $\pm$ 0.102                 | 0.515 $\pm$ 0.017                   | 0.516 $\pm$ 0.017                   |
| H3K4ME2                       | 0.211 $\pm$ 0.069 | 0.336 $\pm$ 0.040 | 0.319 $\pm$ 0.045                 | 0.455 $\pm$ 0.095 | 0.487 $\pm$ 0.170                 | 0.388 $\pm$ 0.101                 | <b>0.502 <math>\pm</math> 0.025</b> | <u>0.500 <math>\pm</math> 0.038</u> |
| H3K4ME3                       | 0.158 $\pm$ 0.072 | 0.352 $\pm$ 0.077 | 0.410 $\pm$ 0.033                 | 0.549 $\pm$ 0.056 | 0.544 $\pm$ 0.045                 | 0.440 $\pm$ 0.202                 | <u>0.565 <math>\pm</math> 0.012</u> | <b>0.586 <math>\pm</math> 0.011</b> |
| H3K79ME3                      | 0.496 $\pm$ 0.042 | 0.613 $\pm$ 0.030 | 0.626 $\pm$ 0.026                 | 0.672 $\pm$ 0.048 | <u>0.697<math>\pm</math>0.077</u> | 0.676 $\pm$ 0.026                 | <b>0.699 <math>\pm</math> 0.013</b> | 0.695 $\pm$ 0.018                   |
| H3K9AC                        | 0.420 $\pm$ 0.063 | 0.542 $\pm$ 0.029 | 0.562 $\pm$ 0.040                 | 0.581 $\pm$ 0.061 | <b>0.622<math>\pm</math>0.030</b> | 0.604 $\pm$ 0.048                 | <u>0.615 <math>\pm</math> 0.013</u> | 0.608 $\pm$ 0.010                   |
| H4                            | 0.732 $\pm$ 0.076 | 0.796 $\pm$ 0.027 | <u>0.799<math>\pm</math>0.025</u> | 0.763 $\pm$ 0.044 | <b>0.811<math>\pm</math>0.022</b> | 0.789 $\pm$ 0.020                 | 0.772 $\pm$ 0.010                   | 0.769 $\pm$ 0.019                   |
| H4AC                          | 0.273 $\pm$ 0.063 | 0.463 $\pm$ 0.041 | 0.495 $\pm$ 0.032                 | 0.564 $\pm$ 0.038 | <b>0.621<math>\pm</math>0.054</b> | 0.525 $\pm$ 0.240                 | <u>0.606 <math>\pm</math> 0.014</u> | 0.197 $\pm$ 0.011                   |
| <b>Regulatory Annotation</b>  |                   |                   |                                   |                   |                                   |                                   |                                     |                                     |
| ENHANCER                      | 0.451 $\pm$ 0.108 | 0.516 $\pm$ 0.098 | 0.548 $\pm$ 0.144                 | 0.517 $\pm$ 0.117 | 0.546 $\pm$ 0.073                 | 0.491 $\pm$ 0.066                 | <b>0.880<math>\pm</math>0.013</b>   | <u>0.878<math>\pm</math>0.013</u>   |
| ENHANCER TYPES                | 0.309 $\pm$ 0.134 | 0.423 $\pm$ 0.051 | 0.424 $\pm$ 0.132                 | 0.386 $\pm$ 0.185 | 0.439 $\pm$ 0.054                 | 0.416 $\pm$ 0.095                 | <b>0.724<math>\pm</math>0.013</b>   | <u>0.724<math>\pm</math>0.010</u>   |
| PROMOTER: ALL                 | 0.954 $\pm$ 0.006 | 0.971 $\pm$ 0.006 | <b>0.976<math>\pm</math>0.006</b> | 0.960 $\pm$ 0.005 | <u>0.970<math>\pm</math>0.004</u> | 0.967 $\pm$ 0.004                 | 0.963 $\pm$ 0.001                   | 0.960 $\pm$ 0.001                   |
| NOTATA                        | 0.955 $\pm$ 0.010 | 0.972 $\pm$ 0.005 | <b>0.976<math>\pm</math>0.005</b> | 0.959 $\pm$ 0.008 | <u>0.969<math>\pm</math>0.011</u> | 0.968 $\pm$ 0.006                 | 0.962 $\pm$ 0.002                   | 0.960 $\pm$ 0.003                   |
| TATA                          | 0.960 $\pm$ 0.023 | 0.955 $\pm$ 0.021 | <b>0.966<math>\pm</math>0.013</b> | 0.944 $\pm$ 0.040 | 0.953 $\pm$ 0.016                 | <u>0.957<math>\pm</math>0.015</u> | 0.942 $\pm$ 0.008                   | 0.940 $\pm$ 0.008                   |
| <b>Splice Site Annotation</b> |                   |                   |                                   |                   |                                   |                                   |                                     |                                     |
| ALL                           | 0.848 $\pm$ 0.019 | 0.939 $\pm$ 0.009 | <b>0.983<math>\pm</math>0.008</b> | 0.956 $\pm$ 0.011 | 0.940 $\pm$ 0.027                 | 0.927 $\pm$ 0.021                 | <u>0.967<math>\pm</math>0.008</u>   | 0.967 $\pm$ 0.004                   |
| ACCEPTOR                      | 0.914 $\pm$ 0.028 | 0.975 $\pm$ 0.006 | <b>0.981<math>\pm</math>0.011</b> | 0.958 $\pm$ 0.010 | 0.937 $\pm$ 0.033                 | 0.936 $\pm$ 0.077                 | <u>0.958<math>\pm</math>0.005</u>   | 0.943 $\pm$ 0.010                   |
| DONOR                         | 0.906 $\pm$ 0.027 | 0.963 $\pm$ 0.006 | <b>0.985<math>\pm</math>0.022</b> | 0.949 $\pm$ 0.024 | 0.948 $\pm$ 0.025                 | 0.874 $\pm$ 0.289                 | <u>0.973<math>\pm</math>0.002</u>   | 0.921 $\pm$ 0.018                   |



Table 4: Performance comparison on the Scorpio-Gene-Taxa dataset.

|                 | Test         |              |              |              |              | Gene out     |              |              |              | Taxa out     |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                 | Phylum       | Class        | Order        | Family       | Gene         | Phylum       | Class        | Order        | Family       | Gene         |
| MetaBERTa       | 0.712        | 0.584        | 0.426        | 0.322        | 0.282        | <b>0.640</b> | <b>0.471</b> | <b>0.290</b> | <b>0.204</b> | 0.074        |
| HyenaDNA        | 0.764        | 0.636        | 0.447        | 0.292        | 0.846        | 0.449        | 0.265        | 0.113        | 0.058        | 0.674        |
| Baseline w/o TM | 0.860        | 0.765        | 0.589        | 0.417        | 0.845        | 0.547        | 0.401        | 0.222        | 0.145        | 0.608        |
| CARMANIA        | <b>0.861</b> | <b>0.768</b> | <b>0.596</b> | <b>0.419</b> | <b>0.909</b> | 0.469        | 0.316        | 0.159        | 0.094        | <b>0.728</b> |

### Task-Adaptive Pre-Training on the Scorpio-Gene-Taxa Dataset

We evaluated task-adaptive pre-training on the *Scorpio-Gene-Taxa* dataset [Refahi et al., 2025], which includes three evaluation splits: *Test* (seen genes and taxa), *Gene-out* (unseen gene labels), and *Taxa-out* (unseen phyla). All models were trained on the same data and evaluated using FAISS-based [Douze et al., 2024] embedding similarity to retrieve the best match from the training set.

As shown in Table 4, CARMANIA achieves the highest accuracy on both the main test set and the *Taxa-out* split, while also showing strong generalization on the *Gene-out* set. In contrast, MetaBERTa performs well on *Gene-out* but struggles on *Taxa-out*, suggesting it may rely more on taxonomic memorization than on learning a transferable mapping from gene sequence to taxonomy. This highlights a fundamental challenge in this dataset: there is often a trade-off between generalizing across gene identities and taxonomic groups. Most models struggle to do both simultaneously unless they explicitly capture the hierarchical structure linking genes to taxa [Refahi et al., 2025].

To further examine these results, we visualized the learned embeddings using t-SNE for the 10 most frequent genes (Figure 5). CARMANIA forms compact clusters that align well with both gene identity and taxonomic structure, suggesting it captures both levels of organization. HyenaDNA produces well-separated clusters based on gene identity but lacks alignment with taxonomic labels. In contrast, MetaBERTa embeddings show stronger alignment with taxonomy but weaker separation by gene, as also seen in Supplementary Figure 8.

This is further supported by additional results in Supplementary Table 5, where CARMANIA outperforms all other models on AMR gene classification tasks without fine-tuning, highlighting its strong domain adaptation capability.

### Domain Adaptation and Generalization

We evaluated several pre-trained genomic models on the AMR dataset without fine-tuning. The dataset, derived from MEGARes Bonin et al. [2023] and CARD Jia et al. [2016], includes three classification tasks: Gene Family, Resistance Mechanism, and Drug Class Yoo et al. [2024]. Labels were assigned based on the closest match (best hit) in embedding space using FAISS [Douze et al. [2024] on frozen representations.

As shown in Table 5, CARMANIA achieves the highest F1 Macro scores across all tasks, indicating that it produces more informative embeddings for AMR classification than prior models, including HyenaDNA [Nguyen et al. [2024], Caduceus [Schiff et al. [2024], MetaBERTa [Refahi et al. [2025], and Nucleotide Transformers [Dalla-Torre et al. [2024].

Table 5: F1 Macro scores for AMR classification on Gene Family, Resistance Mechanism, and Drug Class using FAISS-based best-hit embedding retrieval.

| Model                  | GeneFamily   | Resist-Mech  | DrugClass    |
|------------------------|--------------|--------------|--------------|
| MetaBERTa (35.2M)      | 0.650        | 0.898        | 0.886        |
| HyenaDNA (1.6M)        | 0.623        | 0.840        | 0.880        |
| Caduceus-PH (1.9M)     | 0.597        | 0.823        | 0.839        |
| NucleotideTrans (2.5B) | 0.611        | 0.856        | 0.859        |
| w/o TM (83M)           | 0.728        | 0.974        | 0.931        |
| CARMANIA (83M)         | <b>0.733</b> | <b>0.975</b> | <b>0.942</b> |

### Long Context Task: Biosynthetic Gene Cluster Classification

To evaluate the ability of genomic language models to capture functional patterns across extended DNA regions, we performed a large-scale classification task on biosynthetic gene clusters (BGCs) using the MiBiG database [Kautsar et al. [2020], Liu et al. [2022], where each cluster is labeled by its

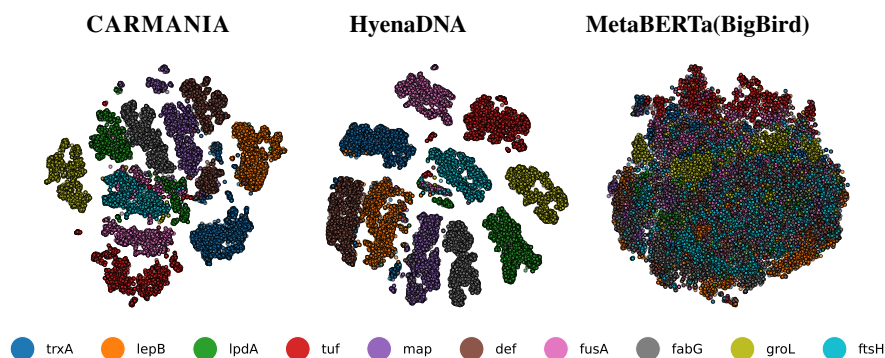


Figure 5: t-SNE visualization of the 10 most common genes in the Scorpio-Gene-Taxa dataset. CARMANIA effectively clusters genes while maintaining taxonomic coherence, leading to superior gene-to-taxonomy classification performance.

associated secondary metabolite class. This task was framed as multi-class classification using raw DNA sequences—without relying on protein translation or domain-level annotations. Since BGCs vary in length (average 377k bp), we truncated all sequences to 100k bp for model compatibility.

Unlike previous approaches that depend on external tools like Prodigal and HMMER Hannigan et al. [2019], Liu et al. [2022], our method directly operates on nucleotide sequences. We evaluated performance using 5-fold cross-validation. As shown in Table 6, CARMANIA, which incorporates TM loss, achieves the highest accuracy (48.4%), outperforming HyenaDNA by over 7% and significantly surpassing other baselines.

Table 6: 5-fold CV accuracy and standard deviation for BGC classification on 100,000 DNA sequences.

| Model              | Caduceus-Ph       | HyenaDNA          | Baseline w/o TM   | CARMANIA                            |
|--------------------|-------------------|-------------------|-------------------|-------------------------------------|
| Accuracy $\pm$ Std | 0.326 $\pm$ 0.119 | 0.412 $\pm$ 0.130 | 0.410 $\pm$ 0.240 | <b>0.484 <math>\pm</math> 0.330</b> |

## 5 Conclusions

In this paper, we introduced CARMANIA, a novel pre-trained genomic model designed for long-sequence analysis and nucleotide-level processing. Unlike existing models, CARMANIA effectively captures both local and global sequence dependencies, making it highly suitable for diverse genomic tasks. Our evaluations demonstrate superior performance across multiple benchmarks, including AMR classification, gene-taxa association, regulatory element prediction, species classification, and biosynthetic gene cluster prediction.

**Limitations and Future Work:** While our method improves sequence representation learning, one limitation lies in the fixed scaling parameter  $\beta$  used to balance the next-token and TM losses. We set  $\beta = 1$  and observed that the losses naturally aligned in scale, making this choice effective in practice. However, tasks involving rare but biologically important motifs may require more fine-grained or dataset-specific tuning. In cases where  $H_{\text{emp}}(X) \neq H_{\text{model}}(X)$ , a fixed  $\beta$  may lead to overfitting on common patterns while missing informative outliers. Future work could explore adaptive scaling strategies that dynamically adjust  $\beta$  based on the training loss distribution. Although the TM loss is computationally efficient due to its matrix-based formulation, memory usage increases with larger vocabulary sizes and batch dimensions. This can limit scalability, particularly in domains such as chemistry (e.g., SMILES representations) or natural language, where token sets are larger and sequence variability is higher. To address this, future work could explore memory-efficient approximations, sparse matrix operations, or gradient checkpointing to reduce the computational footprint. In addition, we aim to scale up both the model and the diversity of training datasets to improve generalization across a broader range of genomic and functional prediction tasks. Finally, we note that, like other predictive genomic models, this approach may raise concerns around potential misuse in synthetic biology or privacy-sensitive contexts, which should be addressed in future deployments.

## References

- Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.
- Lucie A Bergeron, Søren Besenbacher, Jiao Zheng, Panyi Li, Mads Frost Bertelsen, Benoit Quintard, Joseph I Hoffman, Zhipeng Li, Judy St. Leger, Changwei Shao, et al. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951):285–291, 2023.
- Yu Bo, Weian Mao, Yanjun Shao, Weiqiang Bai, Peng Ye, Xinzhu Ma, Junbo Zhao, Hao Chen, and Chunhua Shen. Revisiting convolution architecture in the realm of dna foundation models. *arXiv preprint arXiv:2502.18538*, 2025.
- Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S Morley, et al. Megares and amr++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic acids research*, 51(D1):D744–D752, 2023.
- Samuel Cahyawijaya, Tiezheng Yu, Zihan Liu, Tiffany TW, Xiaopu Zhou, Nancy Y Ip, and Pascale Fung. Snp2vec: Scalable self-supervised pre-training for genome-wide association study. *BioNLP 2022@ ACL 2022*, page 140, 2022.
- Abdulkadir Celikkanat, Andres Masegosa, and Thomas Nielsen. Revisiting k-mer profile for effective and scalable genome representation learning. *Advances in Neural Information Processing Systems*, 37:118930–118952, 2024.
- Jiacheng Chuan, Huimin Xu, Desmond L Hammill, Lawrence Hale, Wen Chen, and Xiang Li. Clasnip: a web-based intraspecies classifier and multi-locus sequence typing for pathogenic microorganisms using fragmented sequences. *PeerJ*, 11:e14490, 2023.
- Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pages 1–11, 2024.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- Arthur L Delcher, Kirsten A Bratke, Edwin C Powers, and Steven L Salzberg. Identifying bacterial genes and endosymbiont dna with glimmer. *Bioinformatics*, 23(6):673–679, 2007.
- Michael F Dion, Steven J Altschuler, Lani F Wu, and Oliver J Rando. Genomic characterization reveals a simple histone h4 acetylation code. *Proceedings of the National Academy of Sciences*, 102(15):5501–5506, 2005.
- Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, ZIJIA CHEN, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=A1ztozypga>.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.
- Benjamin L Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

- Can Firtina, Kamlesh Pillai, Gurpreet S Kalsi, Bharathwaj Suresh, Damla Senol Cali, Jeremie S Kim, Taha Shahroodi, Meryem Banu Cavlak, Joël Lindegger, Mohammed Alser, et al. Aphmm: Accelerating profile hidden markov models for fast and energy-efficient genome analysis. *ACM Transactions on Architecture and Code Optimization*, 21(1):1–29, 2024.
- Ensembl GRCh38. p13 (genome reference consortium human build 38), insdc assembly, 2013.
- Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.
- Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, Michael Wurst, Jakub Kotowski, Dan Chang, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic acids research*, 47(18):e110–e110, 2019.
- Jiachen Hu, Qinghua Liu, and Chi Jin. On limitation of transformer for learning hmms. *arXiv preprint arXiv:2406.04089*, 2024.
- M. Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers, 2024. URL <https://arxiv.org/abs/2402.13512>.
- Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- Baofeng Jia, Amogelang R Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K Tsang, Briony A Lago, Biren M Dave, Sheldon Pereira, Arjun N Sharma, et al. Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, page gkw1004, 2016.
- Satria A Kautsar, Kai Blin, Simon Shaw, Jorge C Navarro-Muñoz, Barbara R Terlouw, Justin JJ Van Der Hooft, Jeffrey A Van Santen, Vittorio Tracanna, Hernando G Suarez Duran, Victòria Pascal Andreu, et al. Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research*, 48(D1):D454–D458, 2020.
- Mingyang Liu, Yun Li, and Hongzhe Li. Deep learning to predict the biosynthetic gene clusters in bacterial genomes. *Journal of Molecular Biology*, 434(15):167597, 2022.
- Alexander V Lukashin and Mark Borodovsky. Genemark. hmm: new solutions for gene finding. *Nucleic acids research*, 26(4):1107–1115, 1998.
- Wendong Ma, Xiaofan Ding, Jiajia Xu, and Terence Chuen Wai Poon. Chhm: a manually curated catalogue of human histone modifications revealing hotspot regions and unique distribution patterns. *International Journal of Biological Sciences*, 20(10):3760, 2024.
- Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.
- Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. Hmmer web server: 2018 update. *Nucleic acids research*, 46(W1):W200–W204, 2018.
- Mohammadsaleh Refahi, Bahrad A Sokhansanj, Joshua C Mell, James R Brown, Hyunwoo Yoo, Gavin Hearne, and Gail L Rosen. Enhancing nucleotide sequence representations in genomic analysis with contrastive optimization. *Communications Biology*, 8(1):517, 2025.
- MS Refahi, BA Sokhansanj, and GL Rosen. Leveraging large language models for metagenomic analysis. In *2023 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6. IEEE, 2023.

- Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.
- Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13, 2019.
- Tao Wu and David F Gleich. Retrospective higher-order markov processes for user trails. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1185–1194, 2017.
- Hyunwoo Yoo, Bahrad Sokhansanj, James R Brown, and Gail Rosen. Predicting anti-microbial resistance using large language models. *arXiv preprint arXiv:2401.00642*, 2024.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.
- Oussama Zekri, Ambroise Odonnat, Abdelhakim Benechehab, Linus Bleistein, Nicolas Boullé, and Ievgen Redko. Large language models as markov chains. *arXiv preprint arXiv:2410.02724*, 2024.
- Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.
- Qiyun Zhu, Shi Huang, Antonio Gonzalez, Imran McGrath, Daniel McDonald, Niina Haiminen, George Armstrong, Yoshiki Vázquez-Baeza, Julian Yu, Justin Kuczynski, et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. *Msystems*, 7(2):e00167–22, 2022.

## A Supplementary Material

### A.1 Effect of TM Loss on Other Domains: Protein Modeling

While CARMANIA was primarily developed for genomic sequences, we further examined the generality of the proposed **TM loss** in other biological domains. In theory, TM loss should benefit any modeling task that relies on capturing long-range dependencies. To validate this hypothesis, we adapted the CARMANIA architecture for amino acid sequences and trained it on the **Scorpio-Gene-Taxa** protein dataset. Specifically, we constructed a  $20 \times 20$  bigram frequency matrix representing pairwise transition statistics between standard amino acids and incorporated TM loss into the training objective.

As summarized in Table 7, incorporating TM loss consistently accelerated convergence and improved predictive accuracy across all hierarchical levels. These results confirm that TM loss generalizes beyond nucleotide modeling and effectively enhances protein-sequence representations.

Table 7: Performance comparison on the **Scorpio-Gene-Taxa** dataset using protein sequences.

| Model           | Test Set |       |       |        |      | Gene-Heldout |       |       |        | Taxa-Heldout |
|-----------------|----------|-------|-------|--------|------|--------------|-------|-------|--------|--------------|
|                 | Phylum   | Class | Order | Family | Gene | Phylum       | Class | Order | Family | Gene         |
| Without TM Loss | 91.1     | 84.9  | 70.6  | 51.3   | 99.5 | 19.8         | 10.9  | 3.3   | 1.1    | 98.4         |
| With TM Loss    | 91.4     | 85.5  | 71.0  | 51.5   | 99.6 | 20.5         | 11.6  | 3.8   | 1.5    | 98.7         |

### A.2 Effect of TM Loss on Convolution-Based Architecture: HyenaDNA

We trained HyenaDNA on the Scorpio-Gene-Taxa dataset with and without TM loss to investigate its effect on model performance. As shown in Figure 6, the model without TM loss demonstrates better convergence, with the training loss decreasing more effectively compared to the model with TM loss. Notably, the model without TM loss appears to struggle with learning the bigram frequency (transition matrix) inherent to the Markovian model. This observation suggests that convolution-based models, even in an autoregressive manner, may not capture Markovian dependencies as effectively as transformer-based models, which leverage attention mechanisms to model long-range dependencies more efficiently.

Additionally, we evaluated the models' performance on the AMR detection tasks using the macro F1 score across all tasks. The results, Figure 7, indicate that incorporating TM loss improves performance in CARMANIA; however, the HyenaDNA model does not exhibit a similar performance gain. This suggests that convolution-based models may inherently struggle to learn embeddings and representations for tasks that rely on Markovian properties, highlighting a potential architectural limitation in such models for sequence modeling tasks.

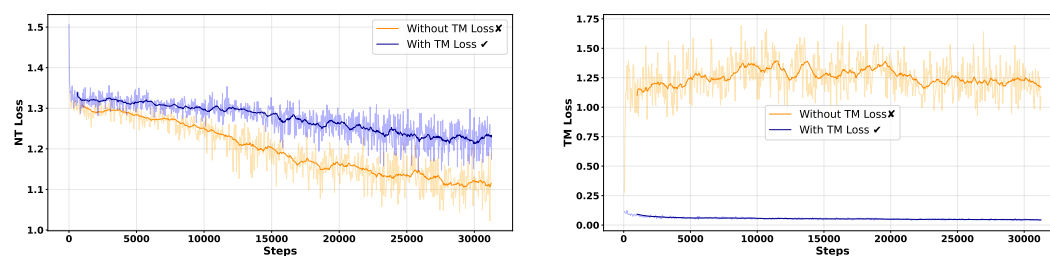


Figure 6: Left: Next-token loss; Right: TM loss for HyenaDNA. The convolution-based model struggles to learn the TM effectively, unlike the attention-based model, which better captures Markovian dependencies.

### A.3 Effect of TM Loss on Mamba-Based Architecture: Caduceus

Caduceus Schiff et al. [2024] is a bi-directional, reverse-complement equivariant DNA language model built upon the MambaDNA architecture, which leverages selective state-space modeling for long-range sequence processing. While Caduceus is pretrained using a masked language modeling (MLM) objective distinct from our causal next-token prediction (NT) setup we incorporated the proposed TM loss into the MLM objective while keeping all other configurations unchanged. This modification allows us to assess whether TM loss provides additional benefits beyond the CARMANIA backbone.

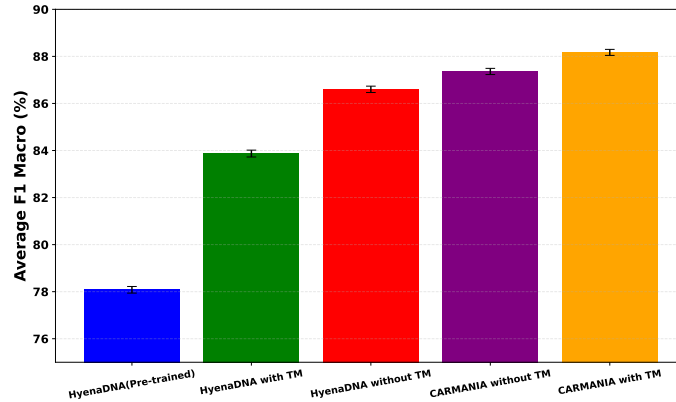


Figure 7: Effect of Transition Matrix Loss on HyenaDNA and CARMANIA : Pre-trained on Scorpio-gene-taxa

For this experiment, we used the largest available pretrained variant of Caduceus (approximately 7.7M parameters) and fine-tuned it on the **Scorpio Gene–Taxa** dataset with and without TM loss. Incorporating TM loss led to a notable performance boost, particularly on the gene classification task, yielding a +12% gain on the test set and +6% on the held-out set. These results suggest that TM loss provides a meaningful signal even in state-space models like Caduceus, although its impact remains more pronounced in transformer-based architectures such as CARMANIA.

Table 8: Effect of TM Loss on Caduceus Fine-Tuned on the Scorpio-Gene–Taxa Dataset. Performance (accuracy, %) across hierarchical taxonomic levels.  $\Delta$  denotes the relative improvement when adding TM loss.

| Model                | Test Set |       |       |        |       | Gene-Heldout |       |       |        | Taxa-Heldout |
|----------------------|----------|-------|-------|--------|-------|--------------|-------|-------|--------|--------------|
|                      | Phylum   | Class | Order | Family | Gene  | Phylum       | Class | Order | Family | Gene         |
| $\Delta$ (TM w/o TM) | +6.2     | +7.0  | +5.8  | +4.0   | +12.4 | +0.3         | +0.1  | 0.1   | 0.0    | +6.6         |
| Caduceus w/o TM      | 40.0     | 24.9  | 13.4  | 7.7    | 26.2  | 30.3         | 13.6  | 4.1   | 1.5    | 15.9         |
| Caduceus + TM        | 46.2     | 31.9  | 19.2  | 11.7   | 38.6  | 30.6         | 13.7  | 4.0   | 1.5    | 22.5         |
| CARMANIA             | 86.1     | 76.8  | 59.6  | 41.9   | 90.9  | 46.9         | 31.6  | 15.9  | 9.4    | 72.8         |

#### A.4 Sensitivity Analysis: Effect of $\beta$

The sensitivity analysis evaluates the impact of the hyperparameter  $\beta$  on model performance. As shown in Table 9, increasing  $\beta$  from 0 to 1.0 improves the macro F1 score (87.36% to 88.17%) and BLEU (0.73 to 0.77), suggesting that the TM loss helps capture Markovian dependencies. However, when  $\beta$  is increased to 5.0, performance significantly degrades, indicating that excessive TM loss disrupts the next-token objective. The best performance is achieved with  $\beta = 1.0$ , demonstrating a balanced contribution from both objectives.

Table 9: Comparison of  $\beta$  values with F1 Macro, BLEU, and Perplexity scores.

| $\beta$ | Average F1 Macro | BLEU | Perplexity |
|---------|------------------|------|------------|
| 0       | 0.873            | 0.73 | 3.6        |
| 0.5     | 0.874            | 0.76 | 3.6        |
| 1.0     | 0.882            | 0.77 | 3.6        |
| 5.0     | 0.836            | 0.85 | 3.8        |

#### A.5 Higher-Order Transition Matrix

We explored the impact of incorporating TM loss at different Markov orders in CARMANIA, with results shown in Tables 10 and 11. Surprisingly, while the first-order TM consistently improves downstream classification performance across tasks, the second-order TM performs worse than both the first-order TM and the baseline without TM.

This performance degradation in the second-order setting can be attributed to the increased sparsity of higher-order transition patterns, particularly in biological sequence data. Since second-order



transitions require co-occurrence of triplets, many such patterns are infrequent or entirely missing from the training set, leading to unstable or noisy estimations. This sparsity reduces the model’s ability to generalize and can hinder the convergence of the transition-based objective Wu and Gleich [2017]. On the other hand, first-order transitions strike a balance by enforcing local smoothness and capturing robust pairwise dependencies that are both statistically reliable and biologically meaningful. This results in more stable training and improved generalization across domains.

Table 10: Comparison of F1 Macro scores across three AMR classification tasks with different TM orders. First-order TM improves performance across all tasks, while second-order TM leads to degradation.

| Model Variant | GeneFamily   | Resist-Mech  | DrugClass    |
|---------------|--------------|--------------|--------------|
| Without TM    | 0.728        | 0.974        | 0.931        |
| 1st-Order TM  | <b>0.733</b> | <b>0.975</b> | <b>0.942</b> |
| 2nd-Order TM  | 0.701        | 0.943        | 0.915        |

Table 11: Accuracy comparison on the Scorpio-Gene-Taxa test set across different taxonomic levels. First-order TM leads to consistent improvements, especially at the gene level. Second-order TM degrades performance across all levels.

| Model Variant | Phylum       | Class        | Order        | Family       | Gene         |
|---------------|--------------|--------------|--------------|--------------|--------------|
| Without TM    | 0.860        | 0.765        | 0.589        | 0.417        | 0.845        |
| 1st-Order TM  | <b>0.861</b> | <b>0.768</b> | <b>0.596</b> | <b>0.419</b> | <b>0.909</b> |
| 2nd-Order TM  | 0.810        | 0.703        | 0.520        | 0.371        | 0.690        |

## A.6 Wide vs. Deep: Architectural Trade-offs for DNA Modeling

As shown in Table 12, our wide architecture reduces the number of layers while significantly increasing the hidden and intermediate sizes. This results in a network that is both more expressive and more efficient.

Most notably, the wide model achieves a substantially higher accuracy on the Scorpio-Gene-Taxa classification task ( $71.05\% \pm 18.07$  vs.  $63.87\% \pm 19.16$ ), demonstrating its superior ability to capture biologically meaningful patterns from genomic sequences. Furthermore, the wide model also outperforms the deep model on the AMR prediction task—an out-of-domain setting—with a higher macro F1 score ( $88.32\% \pm 13.12$  vs.  $85.22\% \pm 14.66$ ). These consistent gains across both in-domain and out-of-domain datasets highlight the robustness and generalization ability of the wide configuration.

The observed improvements can be attributed to three key factors:

First, wider layers with large hidden and MLP dimensions better capture motif-level and co-occurrence patterns, which are prevalent in DNA sequences. This capacity is especially critical for tasks such as gene classification and function prediction, where local patterns carry more signal than hierarchical abstractions.

Second, the reduced depth allows for better GPU utilization and faster convergence. With fewer sequential operations, the wide model completes training in nearly half the time while requiring fewer steps to achieve peak performance.

Third, by concentrating representational power within each layer, the wide network is better equipped to model long-range dependencies and subtle biological variations without needing deep hierarchical stacks. Given these advantages, we adopt the wide model as the backbone for the rest of our study. Its performance and efficiency make it a strong choice for DNA modeling tasks, especially when scaling to large or diverse datasets.

## A.7 Model Size Scalability

We evaluate the effect of model scaling on performance and compute cost using the Basic Genome dataset. Table 14 compares a small 4M-parameter model trained on a 0.5B-token subset with the full 83M-parameter model trained on 10B tokens. Results show substantial gains in both F1 and BLEU scores, demonstrating that increased model capacity and training data size improve representation quality without disproportionately increasing relative FLOPs.

Table 12: Comparison of model configurations between deep and wide architectures. The wide model uses fewer layers but larger hidden and intermediate sizes.

| Configuration                      | Deep Model (80M)       | Wide Model (83M)  |
|------------------------------------|------------------------|-------------------|
| Number of Layers                   | 24                     | 5                 |
| Hidden Size                        | 512                    | 1024              |
| Intermediate (MLP) Size            | 1664                   | 4608              |
| Attention Heads                    | 8                      | 16                |
| Key/Value Heads                    | 4                      | 4                 |
| Attention Window                   | 128                    | 128               |
| Activation Function                | SiLU                   | SiLU              |
| <b>Training Steps</b>              | 109,500                | 31,000            |
| <b>Epoch Time</b>                  | 6.1 hours              | 3.2 hours         |
| Task                               | Performance Comparison |                   |
| Scorpio-Gene-Taxa Test (Accuracy%) | 63.87 $\pm$ 19.16      | 71.05 $\pm$ 18.07 |
| AMR (F1 Macro%)                    | 85.22 $\pm$ 14.66      | 88.32 $\pm$ 13.12 |

Table 13: Parameter Ranges for Model Training

| Parameter         | Range / Value |
|-------------------|---------------|
| Optimizer         | AdamW         |
| Batch Size        | 1, 19, 35     |
| Num Epochs        | 2             |
| Learning Rate     | 5e-4          |
| Weight Decay      | 0.2           |
| Adam Epsilon      | 1e-6          |
| Adam Betas        | (0.9, 0.999)  |
| Warmup Steps      | 400           |
| LR Scheduler Type | Cosine        |
| Max Grad Norm     | 0.85, 2       |

## A.8 Impact of Sequence Length Scaling

To evaluate the effect of input sequence length on model performance, we conducted an ablation study using the GRCh38 dataset, comparing models trained on 10 kbp and 160 kbp fragments. As shown in Table 15, increasing sequence length improves both macro F1 and BLEU score, while maintaining the same computational footprint (FLOPs). This suggests that access to longer genomic contexts enables the model to capture more distant dependencies, leading to more robust and informative sequence representations.

## A.9 Comparison with Other Convolution-Based Models (ConvNova)

To contextualize CARMANIA’s performance with recent convolution-based models, we compared it against ConvNova Bo et al. [2025], a recently proposed architecture for long DNA modeling. Following their evaluation protocol, we used five random seeds on the same test benchmarks to ensure consistency. As shown in Table 16, CARMANIA achieves comparable or superior results on several tasks, demonstrating that its attention-based and Markovian design effectively captures long-range dependencies.

## A.10 Pre-training Dataset

The pre-training phase utilizes large-scale genomic sequences to establish robust sequence embeddings, ensuring scalability, diversity, and compatibility with existing models trained on domain-specific datasets. We incorporate multiple datasets to support comprehensive learning and facilitate direct performance comparisons.

**The Human Reference Genome (GRCh38 - hg38):** This dataset comprises approximately 3 billion base pairs GRCh38 [2013]. From this genome, we extracted two non-overlapping sets of sequences to ensure coverage across diverse genomic regions. One set consists of long-range sequences with 160 kbp fragments, while the other contains shorter sequences of 10 kbp fragments.

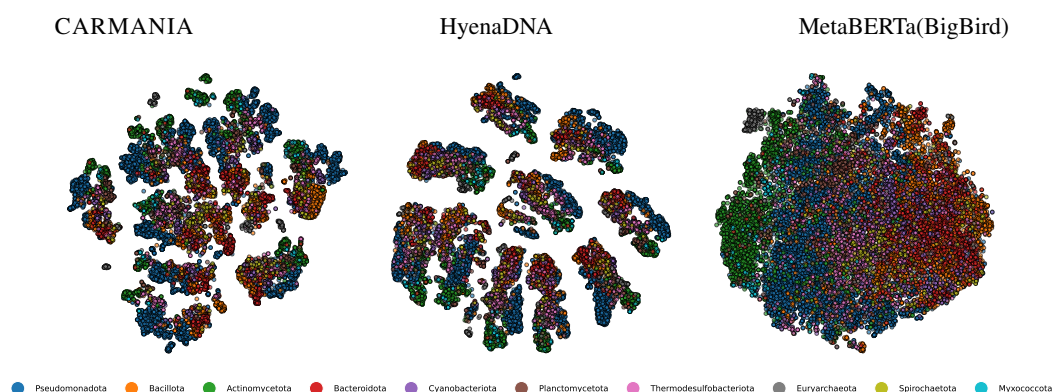


Figure 8: t-SNE visualization of the 10 most common phyla in the Scorpio-Gene-Taxa dataset. CARMANIA effectively captures the structure of inter-cluster taxonomy. As shown in Figure 5, the genes exhibit a clear global structure, with phylogenetic groups positioned near each other when colored based on phyla. This demonstrates that our model outperforms even MetaBERTa(BigBird) in preserving the taxonomy signal.

Table 14: Impact of model size on performance and compute cost.

| Model Size / Dataset    | F1 Macro(AMR) ( $\uparrow$ ) | BLEU (Human dataset) ( $\uparrow$ ) | Relative FLOPs ( $\downarrow$ ) |
|-------------------------|------------------------------|-------------------------------------|---------------------------------|
| 4M params / 0.5B tokens | $0.809 \pm 0.114$            | 0.41                                | 0.07                            |
| 83M params / 10B tokens | $0.860 \pm 0.130$            | 0.82                                | 1.00                            |

**The Basic Genome Dataset:** Introduced by Zhu et al. [2022], this dataset includes 10 billion base pairs from 4,634 bacterial, archaeal, viral, and eukaryotic genomes. Sequences were extracted as 10 kbp fragments from each genome, ensuring representation across various species and evolutionary lineages.

**The Scorpio Gene-Taxa Dataset:** Developed by Refahi et al. [2025], this dataset spans 580 million base pairs from 2,046 bacterial and archaeal species, encompassing 497 distinct gene types. To maintain consistency, fragments of 4 kbp were extracted, with shorter sequences padded as needed. Table 17 summarizes the key statistics of pre-training datasets.

### A.11 Downstream Datasets

We used five distinct datasets for fine-tuning and evaluation, covering a range of genomic classification tasks and label spaces. These include regulatory element prediction, gene function classification, taxonomic inference, antimicrobial resistance (AMR) detection, and biosynthetic gene cluster (BGC) prediction. Table 18 provides detailed statistics for each dataset.

**AMR:** The Antimicrobial Resistance classification dataset used in our evaluation is based on the benchmark introduced by Yoo et al. [2024]. It integrates sequences from two major sources: **MEGARes** Bonin et al. [2023], a manually curated database of antimicrobial resistance genes annotated by gene family and resistance mechanism; and **CARD** Jia et al. [2016], which provides detailed molecular annotations of AMR genes, mechanisms, and associated drugs.

**BGC Dataset.** To evaluate functional classification over extended DNA regions, we used the biosynthetic gene cluster (BGC) classification dataset derived from the MiBiG database Kautsar et al. [2020], Liu et al. [2022], where each cluster is labeled with a secondary metabolite class. Since BGCs vary in length (average 377k bp), we truncated all sequences to 100k bp for model compatibility.

**Nucleotide Transformer Tasks.** We evaluated our model on 18 genomic classification tasks introduced by Dalla-Torre et al. [2024], covering histone mark prediction, regulatory element annotation, and splice site detection. These tasks span a wide range of sequence-based regulatory functions, offering a comprehensive benchmark for assessing model generalization.

Table 15: Effect of sequence length on performance. Longer input improves both F1 and BLEU without increasing FLOPs.

| Training Setup            | F1 Macro ( $\uparrow$ ) | BLEU ( $\uparrow$ ) | Relative FLOPs ( $\downarrow$ ) |
|---------------------------|-------------------------|---------------------|---------------------------------|
| Trained on 10k Sequences  | $0.765 \pm 0.120$       | 0.82                | 1.00                            |
| Trained on 160k Sequences | $0.800 \pm 0.138$       | 0.84                | 1.00                            |

Table 16: Genomics Benchmark Results. Top-1 accuracy ( $\uparrow$ ) is reported for pretrained HyenaDNA, Caduceus-Ph, ConvNova, and CARMANIA, alongside the CNN baseline. The best result is in bold, and the second-best is underlined.

| Task                          | CNN               | HyenaDNA          | Caduceus-Ph                         | ConvNova                            | CARMANIA                            |
|-------------------------------|-------------------|-------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| <i>Enhancers</i>              |                   |                   |                                     |                                     |                                     |
| Mouse Enhancers               | $0.730 \pm 0.032$ | $0.779 \pm 0.013$ | $0.754 \pm 0.074$                   | $0.784 \pm 0.009$                   | <b><math>0.795 \pm 0.002</math></b> |
| Human Enhancers (Cohn)        | $0.702 \pm 0.021$ | $0.718 \pm 0.008$ | <b><math>0.747 \pm 0.004</math></b> | <u><math>0.743 \pm 0.005</math></u> | $0.725 \pm 0.002$                   |
| Human Enhancers (Ensembl)     | $0.744 \pm 0.122$ | $0.832 \pm 0.006$ | $0.893 \pm 0.008$                   | <b><math>0.900 \pm 0.004</math></b> | <u><math>0.895 \pm 0.002</math></u> |
| <i>Species Classification</i> |                   |                   |                                     |                                     |                                     |
| Coding vs. Intergenic         | $0.892 \pm 0.008$ | $0.904 \pm 0.008$ | $0.915 \pm 0.003$                   | <b><math>0.943 \pm 0.001</math></b> | <u><math>0.930 \pm 0.001</math></u> |
| Human vs. Worm                | $0.942 \pm 0.002$ | $0.961 \pm 0.002$ | <b><math>0.973 \pm 0.001</math></b> | $0.967 \pm 0.002$                   | <u><math>0.970 \pm 0.005</math></u> |
| <i>Regulatory Elements</i>    |                   |                   |                                     |                                     |                                     |
| Human Regulatory              | $0.872 \pm 0.005$ | $0.862 \pm 0.004$ | $0.872 \pm 0.011$                   | <u><math>0.873 \pm 0.002</math></u> | <b><math>0.894 \pm 0.002</math></b> |
| Human Non-TATA Promoters      | $0.861 \pm 0.009$ | $0.887 \pm 0.005$ | $0.946 \pm 0.007$                   | <u><math>0.951 \pm 0.003</math></u> | <b><math>0.965 \pm 0.002</math></b> |
| Human OCR (Ensembl)           | $0.698 \pm 0.013$ | $0.744 \pm 0.019$ | <b><math>0.828 \pm 0.006</math></b> | <u><math>0.793 \pm 0.004</math></u> | $0.778 \pm 0.002$                   |

**Genomics Benchmark Tasks.** We evaluated supervised adaptation on the Genomic Benchmarks collection Grešová et al. [2023], which includes a range of classification tasks such as regulatory element prediction, enhancer detection, and binary species classification.

**The Scorpio Gene-Taxa Dataset:** Developed by Refahi et al. [2025], this dataset spans 580 million base pairs from 2,046 bacterial and archaeal species, encompassing 497 distinct gene types. To maintain consistency, fragments of 4 kbp were extracted, with shorter sequences padded as needed. The dataset features *Test*, *Gene Out*, and *Taxa Out* splits, carefully designed to assess both memorization and generalization—by holding out specific genes or entire taxonomic groups (e.g., phyla) while preserving hierarchical relevance. This enables robust evaluation of gene-level and taxonomy-level generalization.

## B Confirming the Transition Matrix Implementation

The first-order transition probabilities for the next token given the current token are defined as:

$$P(w_{i+1} | w_i) = [P(A | w_i), P(T | w_i), P(C | w_i), P(G | w_i)]$$

Similarly, for the token after next:

$$P(w_{i+2} | w_{i+1}) = [P(A | w_{i+1}), P(T | w_{i+1}), P(C | w_{i+1}), P(G | w_{i+1})]$$

In our implementation, the bigram transition matrix for position  $i$  is computed as:

$$\mathbf{T}_i = P(w_{i+1} | w_i) \times P(w_{i+2} | w_{i+1})^T$$

Expanding this matrix explicitly:

$$\mathbf{T}_i = \begin{bmatrix} P(A | w_i)P(A | w_{i+1}) & P(A | w_i)P(T | w_{i+1}) & P(A | w_i)P(C | w_{i+1}) & P(A | w_i)P(G | w_{i+1}) \\ P(T | w_i)P(A | w_{i+1}) & P(T | w_i)P(T | w_{i+1}) & P(T | w_i)P(C | w_{i+1}) & P(T | w_i)P(G | w_{i+1}) \\ P(C | w_i)P(A | w_{i+1}) & P(C | w_i)P(T | w_{i+1}) & P(C | w_i)P(C | w_{i+1}) & P(C | w_i)P(G | w_{i+1}) \\ P(G | w_i)P(A | w_{i+1}) & P(G | w_i)P(T | w_{i+1}) & P(G | w_i)P(C | w_{i+1}) & P(G | w_i)P(G | w_{i+1}) \end{bmatrix}$$

Summing over the entire sequence of length  $N$ , we obtain the predicted (but not yet normalized) bigram matrix:

Table 17: Statistics of the Pre-training datasets.

| Dataset                                | Number of Samples | Number of Tokens | Length range (bp)            |
|----------------------------------------|-------------------|------------------|------------------------------|
| Human Genome-long GRCh38 [2013]        | 19,029            | 3B               | 160k                         |
| Human Genome-short GRCh38 [2013]       | 303,921           | 3B               | 10k                          |
| Basic Genome Zhu et al. [2022]         | 1,010,237         | 10B              | 10k                          |
| Scorpio-Gene-Taxa Refahi et al. [2025] | 547,523           | 580M             | 114–13,227 (4k for training) |

Table 18: Statistics of the fine-tuning datasets.

| Dataset                                                    | Number of Tasks | Classes                  | Length Range (bp)                    |
|------------------------------------------------------------|-----------------|--------------------------|--------------------------------------|
| Genomics Benchmark Grešová et al. [2023]                   | 8               | Binary (one multi-class) | 70–4,776                             |
| Scorpio-Gene-Taxa Refahi et al. [2025]                     | 10              | Multi-class (497–1,929)  | 114–13,227                           |
| Antimicrobial Resistance Prediction Yoo et al. [2024]      | 3               | Multi-class              | 211–5,274                            |
| Nucleotide Transformer Tasks Dalla-Torre et al. [2024]     | 18              | Binary (one multi-class) | 200–500                              |
| Biosynthetic Gene Cluster Prediction Kautsar et al. [2020] | 1               | 8                        | 204–8M (avg 377k; Truncated to 100k) |

$$\mathbf{T}_{Pred} = \sum_{i=1}^N \mathbf{T}_i$$

To confirm equivalence with the bigram frequency matrix, consider the specific cell ( $T \rightarrow A$ ):

$$\mathbf{T}_{Pred}(T \rightarrow A) = \sum_{i=1}^N P(T | w_i) P(A | w_{i+1})$$

If we replace probabilities with actual sequence counts, where  $P(w_i = T) = 1$  if  $w_i = T$  and 0 otherwise, and normalize each row to represent the transition probabilities, we get:

$$\mathbf{T}_{Pred}(T \rightarrow A) = \frac{\sum_{i=1}^N \mathbb{I}(w_i = T) \cdot \mathbb{I}(w_{i+1} = A)}{\sum_{x \in \{A, T, C, G\}} \sum_{i=1}^N \mathbb{I}(w_i = T) \cdot \mathbb{I}(w_{i+1} = x)}$$

This aligns with the actual bigram frequency:

$$P_{actual}(T \rightarrow A) = \frac{\text{count}(T \rightarrow A)}{\sum_{x \in \{A, T, C, G\}} \text{count}(T \rightarrow x)}$$

Both matrices are row-normalized, ensuring the sum of each row is 1:

$$\sum_{x \in \{A, T, C, G\}} \mathbf{T}_{Pred}(T \rightarrow x) = \sum_{x \in \{A, T, C, G\}} P_{actual}(T \rightarrow x) = 1$$

Thus, when probabilities are replaced with actual sequence values, the predicted transition matrix  $\mathbf{T}_{Pred}$  is equivalent to the actual bigram frequency matrix  $P_{actual}$ , confirming that our method is inherently inspired by bigram frequency calculations while extending them with learnable model probabilities.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We carefully reviewed the abstract and introduction, and they do not contain any claims that are not justified in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In the paper, we discuss both the strengths and limitations of our approach in the Experimental Results and Conclusion sections, particularly regarding model generalizability and long-range retention boundaries.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We clearly describe the assumptions behind our approach and provide a full derivation of the first-order transition tensor and loss in Section 3.1 and 3.2. Due to space constraints, extended derivations and higher-order analysis are included in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We use publicly available datasets and describe all model architectures, hyper-parameters, and training procedures in detail. The code will be released upon acceptance to further support reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: All datasets used are publicly available and cited appropriately. We will release the code and detailed instructions for reproducing the main experimental results in the supplementary material and a public repository upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The paper provides detailed descriptions of data splits, optimizer settings, training schedules, and hyperparameters in Section 4 and Appendix A. These include the batch size, learning rate, number of epochs, and evaluation metrics used across all benchmarks.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: For datasets with official splits, we follow the provided train/test partitions without additional resampling. For other datasets, we adopt the evaluation protocol from prior work, including 5-fold or 10-fold cross-validation where applicable. Additionally, we report



average performance across tasks when appropriate to summarize model generalization, as detailed in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We report the GPU type, memory configuration, and training time estimates in Section 4.1 and Appendix A. These include hardware specifications such as A100 GPUs and per-task training durations to support reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [\[Yes\]](#)

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We briefly discuss the potential applications of our method in genomic discovery and functional annotation, which could support advancements in health, biotechnology, and microbial ecology. We also acknowledge possible risks related to dual-use, privacy (in human genomics), or misuse of predictive models, as noted in the Broader Impact section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The datasets and models used in this work are not considered high risk for misuse; they are based on publicly available genomic data and do not involve sensitive personal or dual-use information.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: All external datasets and models used in this work are properly cited, and we rely exclusively on publicly released assets with appropriate licenses, as noted in the references and Appendix A.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: This work does not introduce any new datasets, models, or software assets beyond those already publicly available and cited.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Large language models were not used as part of the core methodology or experimental components of this research. Any use of LLMs was limited to minor editing and formatting support and does not impact the scientific contributions.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.