TrinityDNA: A Bio-Inspired Foundational Model for Efficient Long-Sequence DNA Modeling

Qirong Yang[1*], Yucheng Guo[1*], Zicheng Liu[1,2*], Yujie Yang[1], Qijin Yin[1], Siyuan Li[1,2], Shaomin Ji[1], Linlin Chao[1], Xiaoming Zhang[1†]

[1]BioMap Research, Beijing, China
[2]AI Lab, Research Center for Industries of the Future, Westlake University, China

Abstract

The modeling of genomic sequences presents unique challenges due to their long length and structural complexity. Traditional sequence models struggle to capture long-range dependencies and biological features inherent in DNA. In this work, we propose TrinityDNA, a novel DNA foundational model designed to address these challenges. The model integrates biologically informed components, including Groove Fusion for capturing DNA's structural features and Gated Reverse Complement (GRC) to handle the inherent symmetry of DNA sequences. Additionally, we introduce a multi-scale attention mechanism that allows the model to attend to varying levels of sequence dependencies, and an evolutionary training strategy that progressively adapts the model to both prokaryotic and eukaryotic genomes. TrinityDNA provides a more accurate and efficient approach to genomic sequence modeling, offering significant improvements in gene function prediction, regulatory mechanism discovery, and other genomics applications. Our model bridges the gap between machine learning techniques and biological insights, paving the way for more effective analysis of genomic data. Additionally, we introduced a new DNA long-sequence CDS annotation benchmark to make evaluations more comprehensive and oriented toward practical applications.

1 Introduction

The rapid advancements in large-scale, long-sequence modeling, particularly in the realm of Natural Language Processing (NLP) (Team et al. 2023; Achiam et al. 2023), have radically transformed the way we approach complex data. Deep learning models, such as Transformers (Vaswani et al. 2017), have achieved unprecedented success in tasks that span from language translation to text generation, revolutionizing not just NLP but a variety of other fields. These models have proven their capability to capture intricate dependencies in data, providing solutions to challenges that were once considered insurmountable. With these breakthroughs in NLP, there has emerged an exciting opportunity to extend the power of sequence modeling to a completely different domain—genomics—where data shares some key similarities, such as its sequential nature.

Genomic data, particularly DNA sequences, consists of extraordinarily long strings of information that encode the fundamental building blocks of life. Unlike the highly dense and structured data typically encountered in NLP, genomic sequences are sparse in nature, containing vast stretches of repetition and variability (Liu et al. 2024a). Despite this, they hold a rich repository of biological information that is crucial for understanding gene functions, regulatory mechanisms, and cellular processes. The ability to model DNA sequences deeply could lead to breakthrough applications in personalized medicine, genetic engineering, and the overall understanding of biological systems. However, effectively capturing the dependencies within such long, sparse sequences remains a significant challenge.

While the parallels between NLP and genomics are evident, directly applying traditional NLP models to genomic sequences proves difficult. The sparse, low-density nature of DNA sequences means that existing models often struggle to identify long-range dependencies and interpret the underlying biological structures(Mallet and Vert 2021; Zhou, Shrikumar, and Kundaje 2021). Moreover, the lack of biologically informed features in current models limits their effectiveness in genomic contexts. Many models trained on single-species data perform poorly when generalized to other species or broader biological contexts. As a result, the impact of these models on genomic research has been somewhat restricted, and their applicability to real-world challenges remains limited.

To address these issues, we introduce TrinityDNA, a novel DNA foundational model specifically designed to overcome the current limitations of genomic sequence modeling, as shown in Fig 1. TrinityDNA leverages the latest advancements in deep learning to create a model that is optimized for the unique challenges posed by DNA sequences while also incorporating key biological insights. The contributions are listed as follows:

- **Bio-inspired Design**: A multi-level architecture that is optimized for DNA sequences and leverages the Groove Fusion module and Reverse Complement (RC) fusion strategy. This design captures and exploits the unique structural properties of DNA, enabling long bi-directional genomic modeling.
- **Evolutionary Training Strategy**: A multi-species training regimen that spans a variety of organisms from prokaryotes to eukaryotes, enabling the model to generalize different genomic contexts and sequence lengths.
- **Comprehensive Large-Scale Data Integration**: Curated and integrated datasets from prominent genomic

*First three authors contribute equally.
†Corrsponding author.

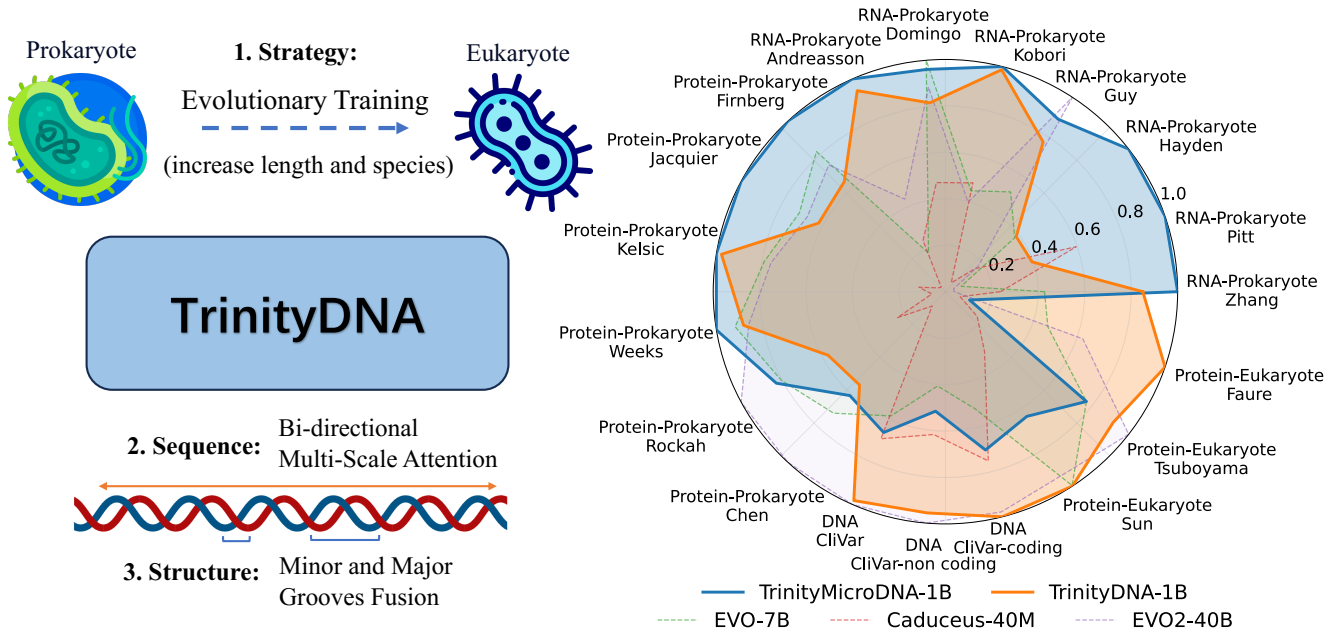

Figure 1: Overview of TrinityDNA Model: (Left) The evolutionary training strategy of TrinityDNA, progressing from prokaryotic DNA to multi-species eukaryotic DNA, and its DNA-targeted long-sequence modeling approach addressing structural features such as bidirectional complementarity and major/minor grooves. (Right) Radar chart illustrating the state-of-the-art performance on the zero-shot performance of our models versus popular models such as EVO and Caduceus.

databases such as GTDB, IMG, and RefSeq, ensuring a diverse and high-quality foundation for model training.

- **New Benchmark for Long-Sequence Inference**: We introduce a novel *CDS Annotation Benchmark* that focuses on gene-structure labeling in prokaryotic genomes, assessing both long-sequence modeling and practical annotation performance.

2 Background

DNA Terminology for Structures.

Basic Composition Deoxyribonucleic acid (DNA) is the hereditary material in most living organisms, consisting of long sequences of nucleotides. Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G). A DNA sequence can be represented as a string $S = (s_1, s_2, \ldots, s_N)$, where $s_i \in \{A, T, C, G\}$ and N denotes the sequence length.

Minor and Major Grooves. The double helix structure of DNA features two distinct grooves: the minor groove and the major groove. These grooves arise from the asymmetric positioning of the phosphate backbone relative to the base pairs. (1) *Major Groove*: wider and deeper, the major groove provides greater accessibility for protein binding and molecular interactions. It generally covers five to seven nucleotides. (2) *Minor Groove*: narrower and shallower, the minor groove presents a different arrangement of hydrogen bond donors and acceptors. While less accessible than the major groove, its length is three to five nucleotides.

Reverse Complement Strands. DNA molecules consist of two complementary strands running in opposite directions, a feature known as antiparallel orientation. For a DNA sequence $S = (s_1, s_2, \ldots, s_N)$, its reverse complement S^R:

$$S^R = (s_N^C, s_{N-1}^C, \ldots, s_1^C)$$

where s_i^C denotes the complementary base of s_i following the base-pairing rules: $A \leftrightarrow T$ and $C \leftrightarrow G$. This reverse complementarity is fundamental to DNA replication and transcription processes. Incorporating reverse complement information into computational models enhances their ability to capture symmetrical and complementary patterns, thereby improving predictions related to gene annotation and regulatory element identification.

DNA Long-Sequence Modeling

Structured State Space Models (SSMs). A prominent class of models for handling long-range dependencies is based on *Structured State Space Models (SSMs)* (Gu, Goel, and Ré 2022a, 2021, 2022b; Gupta, Schuurmans, and Ré 2022; Smith et al. 2022; Dao et al. 2022). These models emerge from discretizing a continuous-time linear system:

$$\boldsymbol{h}(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{B}\,x(t), \qquad y(t) = \boldsymbol{C}\,\boldsymbol{h}(t) + \boldsymbol{D}\,x(t),$$
$$\boldsymbol{h}_{t+1} = \overline{\boldsymbol{A}}\,\boldsymbol{h}_t + \overline{\boldsymbol{B}}\,x_t, \qquad y_{t+1} = \boldsymbol{C}\,\boldsymbol{h}_t + \boldsymbol{D}\,x_t,$$

where $\boldsymbol{h}(t)$ (or $\boldsymbol{h}_t$) is an internal state that can capture long-range dependencies in the input sequence $x(t)$ (or x_t). Through efficient convolution-based implementations of these recurrences, SSMs have shown strong performance

on very long sequences. However, conventional SSMs do not explicitly adapt their parameters to specific tokens or positions, which may limit expressivity for tasks such as DNA sequence modeling. Therefore, a line of SSMs are designed for long-sequence DNA.

HyenaDNA (EVO). *HyenaDNA* (Poli, Suteu et al. 2023) is an SSM variant, a decoder-style model capable of handling long genomic sequences (e.g., hundreds of thousands of tokens). It leverages the *Hyena* operator, which replaces traditional attention with a fast convolution-based mechanism. Concretely, HyenaDNA blocks compute a Toeplitz convolution filter on projected input tokens, allowing processing of very large contexts (in $\mathcal{O}(L \log L)$ time) without the quadratic cost typically associated with attention.

Caduceus (MambaDNA). *MambaDNA* (Schiff et al. 2024)—also referred to as *Caduceus*—builds on the selective SSM approach of Mamba and incorporates *reverse-complement symmetry*, a core property of DNA sequences. In standard Mamba, the module processes sequences in a single direction. MambaDNA extends this design in two ways: (1) **BiMamba**: Instead of only left-to-right processing, BiMamba applies the Mamba block twice. (2) **RC Equivariance**: MambaDNA explicitly enforces RC symmetry by taking a sequence and its RC as inputs to the same SSM-based module. See more details in Appendix E.

3 TrinityDNA: Sequence, Structure, and Strategy for DNA Modeling

Preliminaries

Lost in the locality. While SSMs theoretically excel at handling long sequences, they inherently exhibit a *locality bias* (Wang et al. 2024). This issue is exacerbated in DNA sequence modeling, where dependencies across vast genomic regions must be captured to enable accurate biological interpretation. Specifically, in genomic data, long-range dependencies often span tens of thousands or even hundreds of thousands of base pairs. Existing models, such as SSMs, typically focus on local dependencies due to computational challenges in processing long sequences. Consequently, our empirical results in Figure 2 demonstrate that

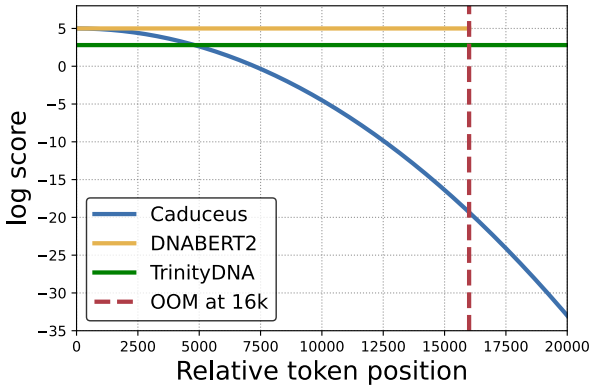

Figure 2: Comparison of log influential scores log $|\partial yt/\partial xs|$ versus distance $(t - s)$ on HG-38 (Nguyen et al. 2023).

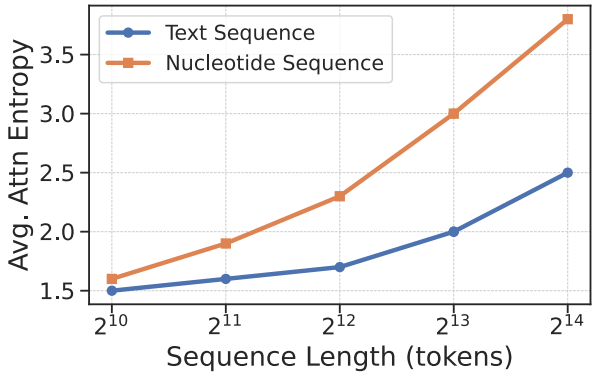

Figure 3: Average attention entropy of full self-attention models as sequence length increases.

the SSM-based model (Caduceus) lost their focus as the sequence length increased, while the full-attention-based model (DNABERT2) suffered from heavy computation. These limitations hinder modern DNA foundation models' ability to fully capture the inherent complexities of DNA sequences and their interconnections across genomic regions.

Oversmoothing in long attention. As sequences grow in training, full self-attention "flattens out": attention scores converge toward a uniform, high-entropy distribution in Figure 3, so every token is weighted almost equally, and useful signals vanish. This oversmoothing affects low-density data—images, DNA, and the hardest— where informative tokens are sparse and far apart. In shorter windows, the same model still shows specialized heads, *e.g.*, retrieval vs. induction (Xiao et al. 2024), but that diversity collapses at the kilobase scale. These findings motivate a multi-window, multi-head scheme that combines narrow local windows with broader global ones to reduce entropy and maintain head specialization. Detailed settings refers to the Appendix.

Groove Fusion Module

To account for the minor and major grooves in DNA sequences, we propose a Groove Fusion module that combines convolutional operations of varying kernel sizes. These grooves have distinct structural and functional roles in DNA, with the major groove offering greater accessibility for protein binding and the minor groove being involved in different molecular interactions. To model these differences, we perform tokenization on the DNA sequence using three convolutional kernels of sizes 3, 5, and 7. This multi-scale convolution approach enables the model to focus on different spatial features across the sequence, effectively capturing the structural nuances associated with the two grooves.

Formally, the Groove Fusion process can be defined as:

$$\text{GrooveFusion}(S) = \sum_{k \in \{3,5,7\}} \text{GELU}(\text{Conv}_k(S))$$

where Conv_k represents the convolution operation with kernel size k, and S is the input DNA sequence. The output of each convolution operation is fused to capture the multi-scale contextual information necessary for interpreting the structural variations between the major and minor grooves.

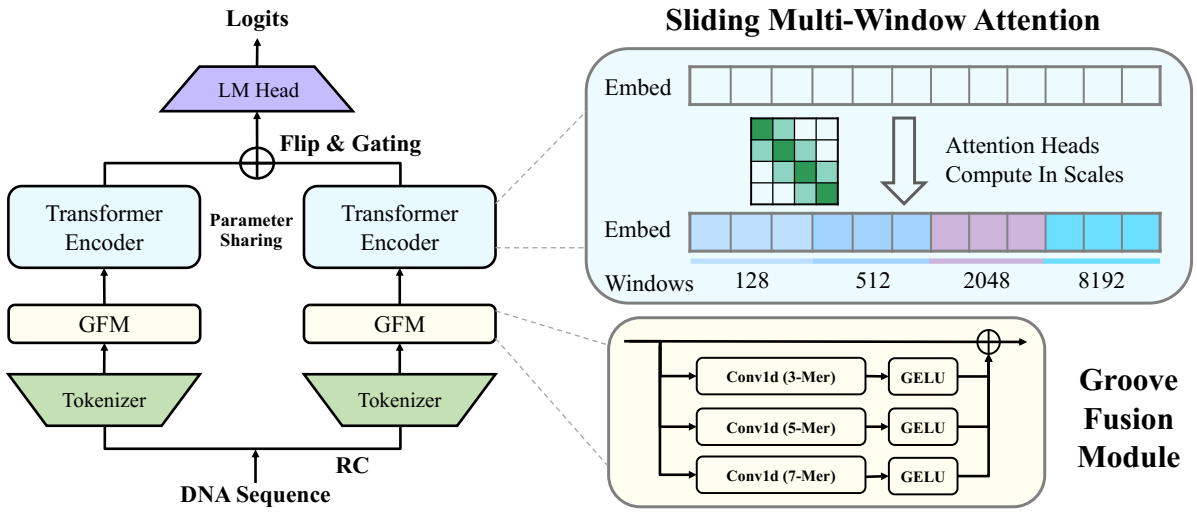

Figure 4: Model Architecture of TrinityDNA: The model integrates DNA sequences and structural features by considering its grooves and reverse complementary sequence with shared parameters.

Sliding Multi-Window Attention

To mitigate locality bias and attention oversmoothing, we revisit the design of Multi-Head Attention (MHA), focusing on varying attention window sizes across heads. In traditional MHA, each head typically attends to the entire sequence using a fixed attention window, which can limit the model's ability to capture dependencies at different scales. To address this, we introduce a multi-scale attention grouping strategy named SMWA. In this approach, we assign different attention heads to capture dependencies at different scales in the DNA sequence, enabling the model to specialize in local or global dependencies, depending on the feature scale. Formally, we define the window sizes for each attention head h as $L_h \in \mathbb{N}$, with L_h representing the length of the attention window for head h. For head h, the attention mechanism is computed using a sliding window of size L_h over the sequence, allowing the model to focus on scales:

$$\text{Attn}_h(S_i) = \text{Softmax}\left(\frac{Q_h(i)K_h(i+[-L_h, L_h])^T}{\sqrt{d_k}}\right) V_h(i+[-L_h, L_h])$$

where $Q_h \in \mathbb{R}^{N\times d_k}$ is the query matrix for head h, $K_h \in \mathbb{R}^{N\times d_k}$ is the key matrix, $V_h \in \mathbb{R}^{N\times d_v}$ is the value matrix, N is the sequence length, and d_k and d_v are the dimensionality of the key and value vectors, respectively. i is the index of the sequence, and $[-L_h, L_h]$ represents the range of indices for the sliding window around i. This enables each head to specialize in attending to either short-range or long-range dependencies by adjusting L_h. The final output of the multi-head attention layer is the concatenation of all heads' outputs, followed by a linear transformation:

$$\text{SMWA}(S) = \text{Concat}(\text{Attn}_1, \text{Attn}_2, \ldots, \text{Attn}_H)W_O$$

where $W_O \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ is a learned output weight matrix, H is the number of attention heads, and d_{out} is the output dimensionality. This multi-scale attention mechanism allows the model to simultaneously capture both local and global dependencies by allocating different heads to focus on different sequence scales. Thus, shorter sequences can be captured by heads with smaller windows, while longer-range dependencies can be modeled by heads with larger windows, enabling the model to capture the hierarchical nature of DNA sequences better.

Gated Reverse Complement

To leverage the reverse complementarity inherent in DNA sequences, we introduce a novel Gated Reverse Complement (GRC) mechanism. This mechanism employs a shared Transformer module to process both the forward and reverse complement sequences in parallel. The reverse complement of a DNA sequence $S = (s_1, s_2, \ldots, s_N)$ is defined as $S^R = (s_N^C, s_{N-1}^C, \ldots, s_1^C)$, where s_i^C denotes the complementary base of s_i (with base-pairing rules $A \leftrightarrow T$, $C \leftrightarrow G$). The GRC mechanism works by feeding both the forward and reverse complement sequences into a shared SMWA-equipped Transformer network f_θ, where the outputs are gated using a linear gating mechanism to combine the two representations effectively. The final layer is:

$$\text{Output} = \text{GRC}(S, S^R) = f_\theta(S) + \sigma(W_G \cdot f_\theta(\text{Flip}(S^R)))$$

where W_G are learned weights, σ represents the identity function, and the Flip operator means to reverse the sequence in the original order. This allows the model to learn the representations of both sides simultaneously.

Evolutionary Training Strategy

The Evolutionary Training Strategy (ETS) approach leverages a two-stage, evolution-inspired training strategy to progressively address the varying complexities of genomic data. In the first stage, the model is trained on prokaryotic genomes, which are relatively straightforward in terms of their regulatory architectures (Xu and Jin 2006; Nguyen

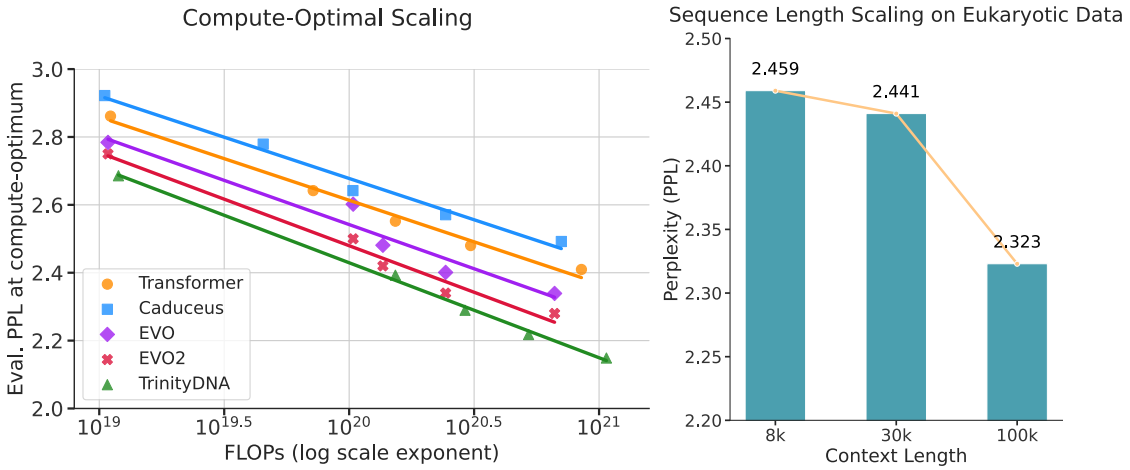

Figure 5: **Scaling Behaviors of Our Proposed Model.** *(Left)* Evaluation perplexity (PPL) against total FLOPs across multiple architectures, showing consistent improvements to various baselines. *(Right)* Impact of increasing context length (8k, 30k, 100k) on a eukaryotic dataset, where PPL steadily decreases with longer context windows.

et al. 2024). Through this foundational phase, the model captures essential DNA sequence motifs and organizational patterns. Subsequently, the second stage introduces eukaryotic genomes, known for their intron-exon structures and gene lengths that can span tens of kilobases (Dalla-Torre et al. 2024). Alongside this transition, the model's context window is enlarged from 8K to 100K base pairs, accommodating multiple co-expressed genes and regulatory elements.

4 Experiments

Pre-training

Data. We adopt masked language modeling (MLM) (Devlin et al. 2018) and character-level tokenization for all our DNA models. Following our *Evolutionary training* strategy, we conduct pre-training in two phases: Stage 1 (Short-Sequence Pre-training): We use prokaryotic (bacterial and archaeal) DNA data from the OpenGenome dataset introduced in (Nguyen et al. 2024), training on sequences of length 8k to learn fundamental nucleic acid patterns in shorter contexts. Stage 2 (Multi-species Post-training): We then continue pre-training on a multi-species collection, the Multispecies dataset, presented in (Dalla-Torre et al. 2024). The sequences in this stage can be as long as 100k, spanning archaebacteria, fungi, vertebrate genomes, and more. This step exposes the model to a rich spectrum of evolutionary signals, enabling it to handle much longer contexts and to capture the diverse structural intricacies of eukaryotic DNA. Hence, we propose two main models, `TrinityMicroDNA` and `TrinityDNA`, each with 1 billion (1B) parameters. The former is trained solely on prokaryotic data, while the latter builds upon the former by post-training on eukaryotic data. By transitioning from bacterial genomes to longer eukaryotic genomes, our method achieves broad coverage of genomic features across scales. More details in Appendix A.

Scaling Laws. Figure 5 illustrates three key aspects of our model's scaling behavior across parameter sizes, pre-training context lengths (8k, 30k, 100k). *(Left)* We plot the compute-perplexity frontier against total FLOPs for several architectures, demonstrating that our approach consistently outperforms baseline methods (Transformer, Caduceus, EVO, and EVO2) at every compute level in different parameter sizes (6M to 1B). *(Right)* We examine the effect of increasing context window sizes on a eukaryotic benchmark, finding a steady drop in perplexity when moving from 8k to 30k, and a further substantial improvement at 100k.

Ablation Study and Analysis

(1) Effectiveness. Table 1 shows that GFM consistently lowers perplexity by modeling the spatial 'groove' features of DNA sequences. Likewise, incorporating GRC (captures reverse-complement patterns) yields a notable drop in PPL, reflecting the importance of complementary-strand information. Meanwhile, SMWA enables multi-scale context handling, trading off some computational overhead for competitive perplexity across local and longer-range dependencies. Although SMWA incurs some performance loss, it is still evident that it saves resources for long sequence training.

(2) Efficiency. The left panel of Figure 6 contrasts token-throughput as we sweep both sequence length and micro-batch size on 1B scale. Across every setting, TrinityDNA remains clearly at the top, retaining more than 80% of its short-sequence throughput even at 64k tokens. This robustness stems from its sliding multi-window attention and optimized fused kernels, which maintain memory traffic at a

Components	W/O	W	W/O	W
GRC	2.731	2.599 (-0.132)	-	-
GFM	2.599	2.534 (-0.065)	-	-
SMWA	2.534	2.544 (+0.010)	64.5	44.5 (-31%)

Table 1: Effect of components on pre-training perplexity (left side) and computational cost (left side, TFLOPs).

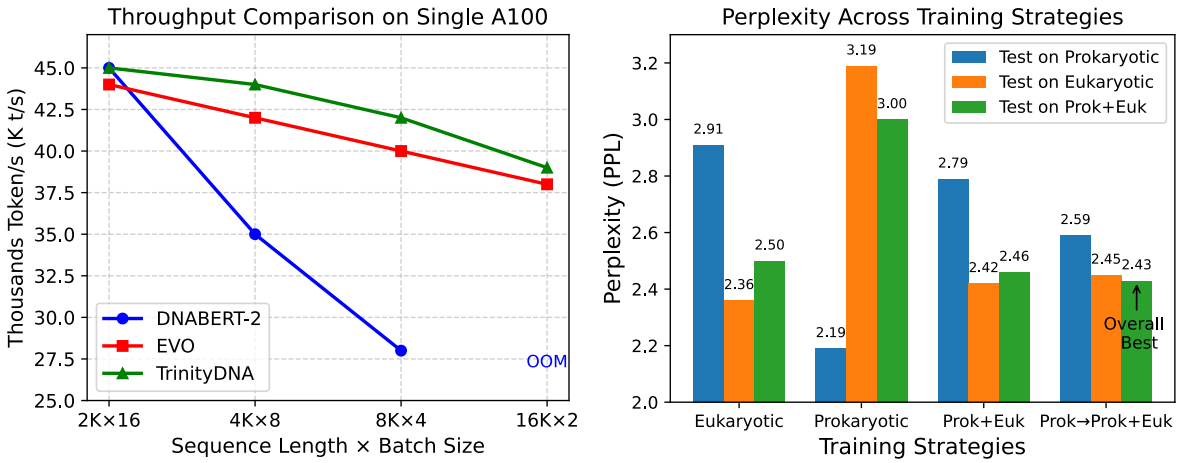

Figure 6: Comprehensive ablation study and efficiency analysis on TrinityDNA.

Tasks **# Params**	**DNABERT** 86M	**NT** 2.5B	**DNABERT2** 117M	**Caduceus** 40M	**HyenaDNA** 1.6M	**TrinityDNA** 1B
H3	0.731 ±0.015	0.788 ±0.012	0.783 ±0.014	0.799 ±0.029	0.779 ±0.037	**0.814** ±0.014
H3K14ac	0.401 ±0.018	0.562 ±0.015	0.526 ±0.019	0.541 ±0.212	0.612 ±0.065	**0.694** ±0.016
H3K36me3	0.473 ±0.017	0.620 ±0.012	0.569 ±0.015	0.609 ±0.109	0.613 ±0.041	**0.692** ±0.014
H3K4me1	0.414 ±0.012	0.553 ±0.011	0.505 ±0.015	0.488 ±0.102	0.512 ±0.024	**0.611** ±0.015
H3K4me2	0.323 ±0.014	0.365 ±0.010	0.311 ±0.013	0.388 ±0.101	0.455 ±0.095	**0.480** ±0.013
H3K4me3	0.278 ±0.015	0.403 ±0.011	0.363 ±0.014	0.440 ±0.202	0.549 ±0.056	**0.522** ±0.015
H3K79me3	0.612 ±0.013	0.647 ±0.010	0.674 ±0.012	0.676 ±0.026	0.672 ±0.048	**0.741** ±0.014
H3K9ac	0.512 ±0.015	0.560 ±0.013	0.556 ±0.017	0.604 ±0.048	0.581 ±0.061	**0.662** ±0.014
H4	0.793 ±0.012	0.817 ±0.015	0.807 ±0.011	0.789 ±0.020	0.763 ±0.044	**0.829** ±0.012
H4ac	0.372 ±0.016	0.491 ±0.018	0.504 ±0.019	0.525 ±0.240	0.564 ±0.038	**0.632** ±0.013
Human TF	0.642 ±0.012	0.633 ±0.015	0.701 ±0.020	-	-	**0.714** ±0.009
Mouse TF	0.564 ±0.018	0.670 ±0.014	0.680 ±0.015	-	-	**0.786** ±0.012
Promoter	0.768 ±0.015	0.799 ±0.012	0.774 ±0.019	-	-	**0.803** ±0.012
Splice Recon.	0.841 ±0.010	0.894 ±0.014	0.850 ±0.020	-	-	**0.927** ±0.009
Virus Covid	0.555 ±0.017	**0.730** ±0.012	0.710 ±0.014	-	-	0.706 ±0.015
Overall Avg	0.552	0.636	0.621	0.586	0.610	**0.708**

Table 2: GUE Benchmark performance comparison. Metrics are single/multi-label binary classification (MCC).

nearly constant level as the context grows.

(3) Training Strategies. The right plot in Figure 6 shows a comprehensive ablation study separating the contributions of dataset size and evolutionary training strategy. The table shows perplexity scores across different model configurations. Key finding: Models initialized with weights from prokaryotic pre-training and then fine-tuned on combined data show better performance than models trained from scratch on the combined dataset. This validates both the importance of large, diverse datasets and our EST.

Downstream Tasks

We employ `TrinityMicroDNA` and `TrinityDNA` with 1B parameters as our base model for downstream evaluation, and we use LoRA tuning except for the zero-shot task.

GUE Benchmark We evaluate our models on a comprehensive Genomic Understanding Evaluation (GUE) benchmark comprising tasks from *Genomics Benchmark* (Zhou et al. 2023) and *Nucleotide Transformer* tasks (Dalla-Torre et al. 2024) by standard LoRA fine-tuning (Hu et al. 2021). For baselines, results are used in the original paper.

Results In Table 2, we compare models DNABERT, Nucleotide Transformer (NT), DNABERT2, and `TrinityDNA`. Overall, ours outperforms prior methods across many metrics. We observe large improvements in tasks that demand recognition of extended promoter regions or higher-order structural features, aligning with our architectural design for multi-window attention and GRC-based reverse complement awareness.

Zero-shot Performance We evaluated our two 1B models on 19 zero-shot downstream tasks—covering DNA pathogenicity (ClinVar), seven RNA DMS benchmarks and fifteen protein-fitness benchmarks across prokaryotic and eukaryotic (Table 3). The prokaryote-focused TrinityMicroDNA-1B dominates the prokaryotic regime, winning 8 of 13 prokaryotic tasks and attaining the high-

Task Type	Task	TrinityMicroDNA	TrinityDNA	EVO	EVO2	EVO2	Caduceus
	# Params	1B	1B	7B	40B	1B	40M
RNA DMS (*Prokaryote*)	Zhang	**0.560**	0.476	0.239	0.021	0.152	0.133
	Pitt	**0.294**	0.116	0.021	0.011	0.040	0.175
	Hayden	**0.365**	0.141	0.138	0.065	0.010	0.059
	Guy	0.370	0.321	0.214	**0.417**	0.354	0.019
	Kobori	**0.569**	0.561	0.255	0.226	0.317	0.275
	Domingo	0.438	0.372	**0.456**	0.403	0.315	0.215
	Andreasson	**0.292**	0.276	0.053	0.127	0.036	0.070
Protein DMS (*Prokaryote*)	Firnberg	**0.673**	0.433	0.552	0.499	0.621	0.018
	Jacquier	**0.659**	0.409	0.471	0.446	0.529	0.023
	Kelsic	0.406	0.397	0.321	0.309	**0.463**	0.047
	Weeks	**0.573**	0.505	0.526	0.492	0.561	0.034
	Rockah	0.592	0.411	0.574	**0.715**	0.657	0.168
	Chen	0.383	0.344	0.448	**0.630**	0.534	0.053
DNA	ClinVar	0.629	0.933	0.555	**0.950**	0.927	0.657
	ClinVar-non coding	0.503	0.931	0.397	**0.974**	0.920	0.601
	ClinVar-coding	0.654	**0.930**	0.484	0.910	0.898	0.699
Protein DMS (*Eukaryote*)	Sun	0.202	0.315	0.314	0.295	**0.334**	0.097
	Tsuboyama	0.595	0.708	0.595	**0.773**	0.717	0.134
	Faure	0.067	**0.609**	0.284	0.381	0.482	0.041
Average Performance (Prokaryote)		**0.475**	0.366	0.328	0.335	0.353	0.099
Average Performance (Eukaryote)		0.404	**0.699**	0.415	0.667	0.670	0.314

Table 3: Zero-shot performance across DNA, RNA, and protein DMS tasks. ClinVar tasks are binary classification (AUC); others are RNA/protein fitness regression (Spearman).

Category	Models	Exact Match			75% Match		
		Recall	Precision	F1	Recall	Precision	F1
Pre-trained Models	TrinityMicroDNA-1B	0.775	**0.740**	**0.754**	0.826	**0.788**	0.803
	TrinityMicroDNA-470M	0.743	0.623	0.692	0.803	0.693	0.755
	TrinityMicroDNA-6M	0.592	0.333	0.488	0.723	0.445	0.524
	Caduceus-40M	0.149	0.148	0.140	0.194	0.189	0.180
Classical Pipelines	Prodigal	**0.832**	0.666	0.725	**0.909**	0.765	**0.829**
	GENSCAN	0.721	0.681	0.702	0.810	0.774	0.799
	Glimmer	0.704	0.663	0.688	0.802	0.760	0.780

Table 4: Results of CDS Annotation Benchmark on the filtered *RefSeq* test set.

est prokaryotic average (0.475), whereas the multi-species TrinityDNA-1B excels on eukaryotic protein-fitness prediction, delivering the top score and the highest eukaryotic average (0.699), even surpassing the 40B EVO2 model (0.667). TrinityDNA also excels in state-of-the-art DNA pathogenicity performance, leading in CliVar-coding and ranking second overall. These complementary strengths underscore the benefit of ETS, and a UMAP projection of genome-level embeddings for ten representative clades (Appendix D) reveals clear taxonomic clustering, indicating that both models learn rich species signals without fine-tuning.

CDS Annotation Benchmark We also introduce a novel CDS Annotation Benchmark, aiming to assess long-sequence inference capabilities, practical utility for gene annotation in real-world genomes. From RefSeq, we collect all prokaryotic reference genomes and parse the GenBank annotation files for gene positions/types. This yields token-level labels indicating whether each token belongs to a coding sequence (CDS) with 20k sequence length and, if so, in which strand/direction it is transcribed. The detailed data statistics are described in Appendix D.

Results The results are shown in Table 4. While Prodigal shows strong recall performance, TrinityMicroDNA-1B delivers the best precision and F1 scores for exact matches, highlighting the model's strong generalization capabilities across diverse datasets compared to classical pipelines.

5 Conclusion

We present `TrinityDNA`, a foundational DNA model that integrates biologically inspired modules—*Groove Fusion* and *Gated Reverse Complement*—alongside a *multi-scale attention mechanism* to capture scales of genomic context effectively, built upon an ETS that transitions from prokaryotic to eukaryotic genomes. `TrinityDNA` gains strong generalization for diverse genomic prediction tasks. We also introduce the *CDS Annotation Benchmark*, which evaluates coding sequence identification across organisms and provides a realistic standard for genome-scale annotation.

References

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

Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; and Sayers, E. W. 2012. GenBank. Nucleic Acids Research, 41(D1):D36–D42.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023–01.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Lopez Carranza, N.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2024. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 1–11.

Dao, T.; Jain, S.; Rudra, A.; and Ré, C. 2022. Hungry Hungry Hippos: Towards Language Model Scaling via Backpropagation in Space. arXiv preprint arXiv:2212.14052.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dreos, R.; Ambrosini, G.; Perier, R. C.; and Bucher, P. 2013. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. Nucleic Acids Research.

Gu, A.; Goel, K.; and Ré, C. 2021. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers. arXiv preprint arXiv:2103.10897.

Gu, A.; Goel, K.; and Ré, C. 2022a. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations (ICLR)*.

Gu, A.; Goel, K.; and Ré, C. 2022b. Parameterization and Initialization of Diagonal State Spaces for Sequence Modeling. arXiv preprint arXiv:2208.04933.

Gupta, H.; Schuurmans, D.; and Ré, C. 2022. Diagonally-Scalable Structured State Space Models for Long Sequence Modeling. arXiv preprint arXiv:2211.07225.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Khare, S.; Gurry, C.; Freitas, L.; Kelly, B. B.; Maurer-Stroh, S.; et al. 2021. GISAID's role in pandemic response. China CDC Weekly, 3(49):1049–1051.

Li, S.; Wang, Z.; Liu, Z.; Wu, D.; Tan, C.; Zheng, J.; Huang, Y.; and Li, S. Z. 2024. VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling. In *International Conference on Machine Learning (ICML)*.

Liu, Z.; Li, J.; Li, S.; Zang, Z.; Tan, C.; Huang, Y.; Bai, Y.; and Li, S. Z. 2024a. GenBench: A Benchmarking Suite for Systematic Evaluation of Genomic Foundation Models. arXiv:2406.01627.

Liu, Z.; Li, S.; Wang, L.; Wang, Z.; Liu, Y.; and Li, S. Z. 2024b. Short-Long Convolutions Help Hardware-Efficient Linear Attention to Focus on Long Sequences. arXiv:2406.08128.

Liu, Z.; Wang, L.; Li, S.; Wang, Z.; Lin, H.; and Li, S. Z. 2024c. LongVQ: Long Sequence Modeling with Vector Quantization on Structured Memory. arXiv:2404.11163.

Mallet, V.; and Vert, J.-P. 2021. Reverse-complement equivariant networks for DNA sequences. *Advances in Neural Information Processing Systems*, 34: 13511–13523.

Nebrao, M. J.; Kung, V. T.; Rowe, K.; Amaral, G. R.; Azad, R. K.; and Cambray, G. 2021. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. Bioinformatics.

Nguyen, E.; Poli, M.; Durrant, M. G.; Kang, B.; Katrekar, D.; Li, D. B.; Bartie, L. J.; Thomas, A. W.; King, S. H.; Brixi, G.; Sullivan, J.; Ng, M. Y.; Lewis, A.; Lou, A.; Ermon, S.; Baccus, S. A.; Hernandez-Boussard, T.; Ré, C.; Hsu, P. D.; and Hie, B. L. 2024. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723): eado9336.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Notin, P.; Kollasch, A.; Ritter, D.; Van Niekerk, L.; Paul, S.; Spinner, H.; Rollins, N.; Shaw, A.; Orenbuch, R.; Weitzman, R.; et al. 2023. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36: 64331–64379.

Poli, M.; Suteu, I.; et al. 2023. Hyena: A Quasi-Attention Approach to Long-Range Language Modeling. arXiv preprint arXiv:2302.10866.

Schiff, Y.; Kao, C.-H.; Gokaslan, A.; Dao, T.; Gu, A.; and Kuleshov, V. 2024. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. arXiv:2403.03234.

Smith, S.; Gu, A.; Roberts, D. A.; and Ré, C. 2022. Simplified State Space Layers for Sequence Modeling. arXiv preprint arXiv:2209.12951.

Stamatoyannopoulos, J. A.; Snyder, M.; Hardison, R.; Ren, B.; Gingeras, T.; Bernstein, B. E.; et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biology, 13(8):418.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Cai, R.; Wang, Y.; Zhu, J.; Srivastava, P.; Wang, Z.; and Li, P. 2024. Understanding and Mitigating Bottlenecks of State Space Models through the Lens of Recency and Over-smoothing. arXiv:2501.00658.

Wang, R.; Bai, X.; and Chen, K. 2019. SpliceFinder: Differential splicing detection using RNNs. Bioinformatics, 35(14):2373–2380.

Xiao, G.; Tang, J.; Zuo, J.; Guo, J.; Yang, S.; Tang, H.; Fu, Y.; and Han, S. 2024. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. *arXiv*.

Xu, Z.; and Jin, L. 2006. Genome-wide average gene length is highly conserved but recombination rate varies among prokaryotes and eukaryotes.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33: 17283–17297.

Zhou, H.; Shrikumar, A.; and Kundaje, A. 2021. Towards a better understanding of reverse-complement equivariance for deep learning models in regulatory genomics. *BioRxiv*, 2020.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.