MergeDNA: Context-aware Genome Modeling with Dynamic Tokenization Through Token Merging

Siyuan Li[1,2,3], Kai Yu[2], Anna Wang[2], Zicheng Liu[1,2,3], Chang Yu[2], Jingbo Zhou[1,2], Qirong Yang[3]*, Yucheng Guo[3], Xiaoming Zhang[3], Stan Z. Li[2]*

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

Abstract

Modeling genomic sequences faces two unsolved challenges: the information density varies widely across different regions, while there is no clearly defined minimum vocabulary unit. Relying on either four primitive bases or independently designed DNA tokenizers, existing approaches with naive masked language modeling pre-training often fail to adapt to the varying complexities of genomic sequences. Leveraging Token Merging techniques, this paper introduces a hierarchical architecture that jointly optimizes a dynamic genomic tokenizer and latent Transformers with context-aware pre-training tasks. As for network structures, the tokenization module automatically chunks adjacent bases into words by stacking multiple layers of the differentiable token merging blocks with local-window constraints, then a Latent Encoder captures the global context of these merged words by full-attention blocks. Symmetrically employing a Latent Decoder and a Local Decoder, MergeDNA learns with two pre-training tasks: Merged Token Reconstruction simultaneously trains the dynamic tokenization module and adaptively filters important tokens, while Adaptive Masked Token Modeling learns to predict these filtered tokens to capture informative contents. Extensive experiments show that MergeDNA achieves superior performance on three popular DNA benchmarks and several multi-omics tasks with fine-tuning or zero-shot evaluation, outperforming typical tokenization methods and large-scale DNA foundation models.

1 Introduction

Modeling genomic DNA sequences with foundation models (Ji et al. 2021) is an emerging frontier that promises to advance bioinformatics and precision medicine. DNA is often likened to a natural language carrying the "code of life" (Cooper 1981), yet it poses unique modeling challenges far beyond ordinary text. Firstly, genomic information is distributed unevenly. Only around 2% of the human genome consists of coding sequences (CDS), densely packed with functional information, whereas the vast majority is non-coding sequence (nCDS) with regulatory or unknown functions, which contains repetitive or less informative content (Nguyen et al. 2024a). Secondly, unlike natural languages with semantic words (Kudo and Richardson 2018), DNA has no inherent word boundaries or predefined vocabulary units (Zhou et al. 2023). The meaningful "units" of DNA vary by context: a biologically relevant motif might be 3 bases (as a codon) (Liu et al. 2025) or 6–10 bases (a transcription factor binding site), or even longer sequences (Dalla-Torre et al. 2023). This makes fixed tokenization schemes inadequate (Qiao et al. 2024). Third, DNA sequences are extremely long (Nguyen et al. 2024b), often spanning tens of thousands to millions of bases, requiring models that can capture both short-range motifs and long-range dependencies efficiently. And naive pre-training objectives (Radford et al. 2018; Devlin et al. 2019) may fail to focus on the truly important parts of these vast sequences. These factors collectively make DNA fundamentally distinct from human language and call for a new class of sequence modeling architectures.

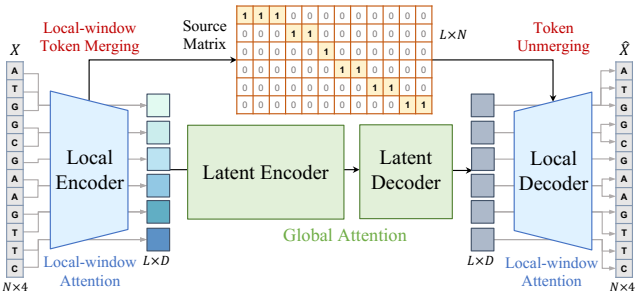


Figure 1: Overview of MergeDNA architecture. The Local Encoder & Decoder achieves adaptive DNA tokenization, while the Latent Encoder & Decoder learn contextual information with informative token masked modeling.

Recent studies have explored various facets of DNA foundation modeling. Long-sequence architectures such as linear-time state-space models (SSMs) (Nguyen et al. 2024b; Schiff et al. 2024), hierarchical Transformers (Shao and Yan 2024), and hybrid networks (Nguyen et al. 2024a; Ma et al. 2025) improve context length scalability with efficiency. Meanwhile, DNA tokenization strategies range from base-level encodings to k-mers (Ji et al. 2021; Wu et al. 2025) and learned vocabularies via BPE (Zhou et al. 2023) or vector quantization (van den Oord, Vinyals, and Kavukcuoglu 2017; Li et al. 2024). Pre-training objectives also vary, including masked modeling (Ji et al. 2021), autoregressive loss (Zhang et al. 2023), and advanced masking (Roy et al. 2023). However, most works optimize these dimensions in isolation and lack a unified mechanism to ad-

*Corresponding authors.

dress all three DNA modeling challenges. For example, the latest long-range models (Brixi et al. 2025) that still operate on single-base tokens may waste capacity on repetitive intergenic regions, while a learned tokenizer without a matching long-context encoder could miss global dependencies (Qiao et al. 2024). In this work, we argue that effective genome-scale modeling requires two core capabilities: (i) a context-sensitive tokenizer that learns to segment DNA into variable-length units based on local structure and semantics, and (ii) adaptive pre-training objectives that prioritize information-dense regions for representation learning. We try to address these jointly by leveraging token merging techniques (Bolya et al. 2023; Lee and Hong 2024) for end-to-end learnable token granularity and contextual abstraction.

This work presents **MergeDNA**, a context-aware genome modeling framework that dynamically adapts tokenization and pre-training to genomic context, as shown in Figure 1. The core idea of MergeDNA is a hierarchical autoencoder-style Transformer that learns to compress and reconstruct DNA sequences with a differentiable tokenizer and a long-range context model. Specifically, we design a Local Encoder composed of stacked local-window attention blocks with differentiable token merging, enabling the model to chunk adjacent bases into variable-length tokens based on local similarity. These merged tokens are then processed by a global-context Latent Encoder using full attention. On the decoder side, a symmetric Latent Decoder and Local Decoder reconstruct the input sequence. Two pre-training objectives jointly supervise the model: (i) Merged Token Reconstruction trains the tokenizer and encoder to preserve key information while filtering redundancies; and (ii) Adaptive Masked Token Modeling selectively masks and predicts important tokens identified through token merging, encouraging context-aware learning of functionally relevant patterns. Together, these components form a unified and scalable genome modeling pipeline that adapts both token resolution and attention allocation based on input complexity.

Our contributions are summarized as follows:

- **Unified Architectural Design:** We propose a novel hierarchical framework that tightly integrates a learnable DNA tokenizer with long-range sequence modeling. Leveraging differentiable token merging within local attention blocks, the Local Encoder captures irregular genomic patterns and determines where to merge as words.
- **Adaptive Context Modeling:** We propose context-aware pre-training tasks that adapt to varying information density in genomic sequences. Using token merging to select informative positions, the proposed Merged Token Reconstruction and Adaptive Masked Token Modeling allow the model to capture both local motif-level information and global long-range dependencies.
- **Strong Empirical Results:** MergeDNA achieves competitive performance across three major DNA benchmarks and shows excellent generalization to several RNA and protein downstream tasks, outperforming prior methods of DNA tokenization and foundation models in both short- and long-context settings.

2 Related Work

DNA Foundation Models. Adapting sequence modeling networks to genomics has demonstrated impressive transfer capacities to genomic applications, where a family of DNA foundation models (Benegas, Batra, and Song 2023) has merged with four lines of research. **(a) Long sequence modeling** is the most crucial technique for long DNA sequences. State-space models (SSMs) like HyenaDNA (Nguyen et al. 2024b; Thoutam and Ellsworth 2024) and Caduceus (Schiff et al. 2024) deliver linear complexity, while hierarchical attention (Shao and Yan 2024) or hybrid SSM-attention designs (Ma et al. 2025) capture both motifs and chromosome-level structure with moderate memory footprints. **(b) DNA tokenization** remains discussion with byte-level (Nguyen et al. 2024a), k-mers (Dalla-Torre et al. 2023), or learnable vocabularies (Ji et al. 2023; Qiao et al. 2024). **(c) Pre-training objectives** can be the BERT-style (Zhou et al. 2023; Li et al. 2025) or auto-regressive-like masked token modeling (Zhang et al. 2023; Zhu et al. 2024) for the encoder or decoder architectures, where some loss reweighing (Brixi et al. 2025) or tailored masking curricula (Roy et al. 2023; Roy, Sural, and Ganguly 2024) could be further beneficial. Only minor methods utilize contrastive learning (Zhou et al. 2025b) or cross-modality alignment tasks (Liu et al. 2025) to integrate multi-omic cues. **(d) Domains of pre-training and applications** are usually bound. While most models are pre-trained on the human reference (Nguyen et al. 2024b) or multiple species corpora (Zhou et al. 2023), specialized datasets confer niche expertise, *e.g.*, prokaryotic (Nguyen et al. 2024a), plant genomic domains (Mendoza-Revilla et al. 2023; Zhai et al. 2025), and metagenomes (Zhou et al. 2025a). Extending beyond monomodal DNA, multi-omics models aim to simulate the central dogma (Cooper 1981) within a unified architecture (Yang et al. 2024) with a gene-to-expression pipeline (Avsec et al. 2021; Yang, Zhu, and Su 2025) or the genome-to-protein pipeline (Song, Segal, and Xing 2024), leveraging shared structure across DNA, RNA, protein, and epigenome.

Byte-level Architectures. In NLP, early subword approaches like BPE (Sennrich, Haddow, and Birch 2015) and SentencePiece (Kudo and Richardson 2018) remain the default module in LLMs, and dynamic schemes like Dynamic Pooling (Nawrot et al. 2022; Liu et al. 2024) partially relax fixed vocabularies, yet they still require external preprocessing. Leveraging SSMs for linear-time attentions (Gu and Dao 2023), MegaByte (Yu et al. 2023), and MambaByte (Wang et al. 2024; Slagle 2024) demonstrated that multi-scale or SSM-based architectures without the subword tokenizer can model million-byte inputs end-to-end with great scale abilities (Ge et al. 2025) on text and other modalities (Wu et al. 2024). More recently, BLT (Pagnoni et al. 2024) introduces learned chunking with entropy-balanced patches, and HNet (Hwang, Wang, and Gu 2025) designs differentiable segmentation with jointly optimization. Similarly, DNA models with raw nucleotides as input are also byte-level architectures (Nguyen et al. 2024a). Meanwhile, classical tokenization strategies, such as BPE (Zhou et al. 2023) and k-mers (Dalla-Torre et al. 2023), as well as learnable dictionaries (Li et al. 2024), have also been explored.

3 Methodology

3.1 Preliminary

A DNA sequence can be seen as a string in the nucleotide alphabet $\mathcal{D} = \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$. We denote a sequence of length

N as $X = (x_1, x_2, \ldots, x_N) \in \mathcal{D}^N$, where each $x_i \in \mathcal{D}$. A DNA tokenizer $\mathcal{T} : \mathcal{D}^N \to \mathcal{V}^L$, segments X into a sequence of L tokens $Z_L = (z_1, \ldots, z_L)$ and maps to a vocabulary $\mathcal{V}$ with $N \geq L$. Given DNA sequences with a causal mask $M \in \{0,1\}^N$, a model f_θ with Attention blocks can be trained with an objective of masked token modeling (MTM):

$$\mathcal{L}_{MLM}(\theta) = -\frac{1}{L}\sum_{i=1}^{L} \log P(x_i \mid X * M;\, \theta), \qquad (1)$$

which encourages f_θ to infer each masked token x_i from its surrounding context to model the DNA context.

3.2 Architectural Overview

Adopting an autoencoder style, MergeDNA consists of four main components, which merge the fixed tokenizer and the sequence model into a hierarchical network in Figure 1.

Local Encoder for Tokenization. The Local Encoder $\mathcal{E}\phi$ serves as a learnable DNA tokenizer with local contexts, producing a tokenized sequence $Z_L \in \mathbb{R}^{L\times D}$ in the embedding dimension of D with a binary source matrix $\mathcal{S} \in \{0,1\}^{L\times N}$:

$$Z_L,\ \mathcal{S} \ =\ \mathcal{E}_\phi(X). \qquad (2)$$

Intuitively, Z_L is a context-dependent segmentation of X, and each row of $\mathcal{S}$ indicates which original positions in X were merged to form the corresponding token in Z_L. To adaptively chunk adjacent bases into informative tokens with local context, we implement the Local Encoder as a stack of local-window self-attention layers interleaved with differentiable token merging operations (described in Sec. 3.3), where the fixed local window size can also ensure linear-time computational complexity despite the long input. This learned tokenizer can be trained jointly with the rest of the model, allowing it to optimize token boundaries for the pre-training objective. Rather than using a fixed k-mer or requiring a byte-pair scheme, the Local Encoder can allocate shorter tokens (finer granularity) to dense information regions and longer tokens to repetitive regions, thereby addressing the varying information density of genomes.

Latent Context Modeling. Based on the tokenized sequence, the Latent Encoder $\mathcal{E}\psi$ is the main network for capturing long-range dependencies across the entire input, which can be implemented as a Transformer encoder with full attention (utilizing Flash Attention). As for inference, $\mathcal{E}\psi$ produces an output of the same length L:

$$Z'_L = \mathcal{E}_\psi(Z_L), \qquad (3)$$

where $Z'_L \in \mathbb{R}^{L\times D}$ are contextually enriched token embeddings. On top of the encoder, we include a lightweight Latent Decoder $\mathcal{E}_\omega$, which transforms Z'_L back toward the token space of $\hat{Z}_L$. The Latent Decoder has a symmetric architecture to $\mathcal{E}_\psi$, and outputs $\hat{Z}_L = \mathcal{E}_\omega(Z'_L)$, where $\hat{Z}_L$ can be seen as a reconstructed version of the Local Encoder's token sequence, containing the information needed to recover the original input. Together, $\mathcal{E}_\psi$ and $\mathcal{E}_\omega$ form an autoencoder on the token level. This design enables us to apply reconstruction-based training at the token level, providing learning signals to both the tokenizer and the context encoder. We emphasize that the Latent Decoder is used only during pre-training to assist the encoder and tokenizer.

Local Decoder for Reconstruction. The final stage is the Local Decoder $\mathcal{E}_\zeta$, which maps the Latent Decoder's output $\hat{Z}_L$ back to the original base space and plays the role of "detokenizer". We first apply a token unmerging operation $\mathcal{U}(\cdot,\cdot)$ using the source matrix $\mathcal{S}$ to upsampling the L-length decoded tokens to length N, $\bar{Z}_N = \mathcal{U}(\hat{Z}_L, \mathcal{S})$. $\bar{Z}_N \in \mathbb{R}^{N\times D}$ denotes an unmerged sequence, where each position corresponds to an original base in X. In matrix form, if $\mathcal{S}_{ij} = 1$ indicates the position i covers original position j, then $\bar{Z}_N = \mathcal{S}^\top \hat{Z}_L$. After unmerging, $\mathcal{E}_\zeta$ applies a stack of local attention (as the reverse of the Local Encoder) to refine local details and output the reconstructed sequence $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_N)$:

$$\hat{X} \ =\ \mathcal{E}_\zeta(\bar{Z}_N). \qquad (4)$$

The Local Decoder thus completes the autoencoder by learning to fill in base-level information that may have been abstracted away by the Local Encoder. Meanwhile, the Local Encoder can be encouraged to produce merge groupings that are easy to invert, as the source matrix preserves positional information that enables accurate reconstruction.

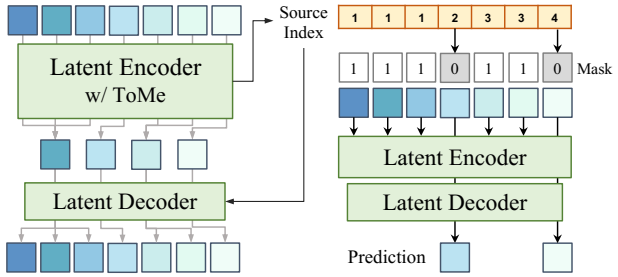
Training *vs.* Inference. During pre-training, we can optimize all modules $\theta = \{\phi, \psi, \omega, \zeta\}$ end-to-end by applying learning objectives of reconstruction and prediction tasks (detailed in Sec. 3.4) between the final output $\hat{X}$ and the original input. As for inference, MergeDNA can be truncated or reconfigured depending on the task types, which can function as a typical encoder-only model for representation learning, or as an encoder–decoder model for generative purposes. For generative tasks or any task requiring output at the nucleotide level, we can use the entire autoencoder, or fine-tune the Local Decoder for the specific output prediction. For classification or regression tasks at the sample level, we can discard both decoders and use the Latent Encoder output directly with a fine-tuned head.

3.3 MergeDNA Tokenization

Local-window Token Merging. At the heart of the Local Encoder is a differentiable token merging mechanism that learns to segment the sequence. We build upon ToMe (Bolya et al. 2023), which progressively fuses similar tokens to reduce sequence length, but adapt it for local, fine-grained chunking. Each Local Encoder layer consists of a standard local self-attention followed by a token merging module. Given the l-th layer, supposing the input sequence length is N_{l-1}, the merging module will select r_l pairs of tokens within each window to compute the average, reducing the sequence by r_l tokens. We denote this operation as:

$$\mathcal{S}^{(l)}, Z^{(l)}_{N_l} = \text{LocalToMeAttn}^{(l)}\Big(Z^{(l-1)}_{N_{l-1}}, \mathcal{S}^{(l-1)}, r_l\Big), \quad (5)$$

where $\mathcal{S}^{(l-1)}$ is the source matrix carried from the previous layer (with $\mathcal{S}^{(0)} = I_{N_0}$ as an identity matrix at input), and $\mathcal{S}^{(l)}$ is updated to reflect the new merges at the l-th layer. In implementation, we compute a similarity score for each pair of tokens in a local window (using a lightweight *grouping* embedding as in DTEM (Lee and Hong 2024)). The top-r_l most similar token pairs in each window are selected to merge. We then perform a *soft merging*: one token ("keeper") absorbs the other ("merger") by adding their



(a) Token Merge Reconstruction (b) Adaptive Masked Token Modeling

Figure 2: Pre-training of MergeDNA for (a) Local Encoder & Decoder and (b) Latent Encoder & Decoder.

representations, or a weighted average, and we mark this in $\mathcal{S}^{(l)}$, where the merger token's source positions are assigned to the keeper. Tokens not selected for merging pass through unchanged to the next layer. This continuous relaxation of token merging ensures the operation is differentiable, allowing gradients to tune both the token embeddings and the merging criteria.

Token Unmerging and Reconstruction. To optimize the end-to-end tokenization capacity, we introduce a *Merged Token Reconstruction (MTR)* objective $\mathcal{L}_{MTR}$ that forces the network to reconstruct the original sequence from compressed tokens, which can be computed as the cross-entropy between $\hat{X}$ and X:

$$\mathcal{L}_{MTR}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log P(\hat{X}_i \mid X_i;\ \theta). \tag{6}$$

During training, we use a compression ratio sampling strategy, which randomly chooses the number of tokens to retain each iteration. For example, if the average goal is $L \approx \frac{N}{2}$, we might sample L from a Gaussian distribution centered at $\frac{N}{2}$ (with the variance to ensure $L \in [0.4N, 0.6N]$). This strategy exposes $\mathcal{E}\phi$ to a wider range of segmentation during training, improving its generalization ability and ensuring that the Local Encoder does not overfit to a particular compression rate.

3.4 Adaptive Context Modeling

As discussed in Sec. 3.2, the Latent Encoder $\mathcal{E}\psi$ processes L tokens uniformly during inference. However, genomic sequences often contain long stretches of low-information content (*e.g.*, repetitive DNA) where modeling every token is unnecessary. We aim to help the model find out and focus most informative tokens and design two steps to improve the naive MLM object, as shown in Figure 2.

Selection and Reconstruction. During pre-training, we modify the latent encoder $\mathcal{E}_\psi$ to forward an additional round and select a smaller number K of salient tokens. Technically, we replace the standard attention in $\mathcal{E}\psi$ with a ToMe-style Attention that merges tokens at the global scale (as opposed to local windows). Formally, we obtain $(Z'_K, \mathcal{S}') = \mathcal{E}\psi(Z_L, \mathcal{S})$, where $Z'_K \in \mathbb{R}^{K\times D}$ with $K < L$. Note that $\mathcal{S}'$ identifies the K most essential tokens among the Z_L: the merging algorithm preferentially fuses tokens that appear redundant or less salient, while preserving distinct tokens that carry unique information. We then feed Z'_K into the Latent Decoder to to produce $\hat{Z}_L$, but first we upsample it back to length L with the unmerge operation, $\bar{Z}_L = \mathcal{U}(Z'_K, \mathcal{S}')$. This distributes each of the K latent tokens back to its original token positions. Finally, $\hat{Z}_L$ is passed through the Local Decoder to produce $\hat{X}$, and we compute a reconstruction loss as $\mathcal{L}_{MTR}$. We refer to this loss as the latent MTR loss, $\mathcal{L}_{MTR}(\theta \setminus \{\phi\})$, since it trains the latent models to recover context from the selected tokens while the Local Encoder is held fixed (ϕ is not updated in this step). Intuitively, this task forces the latent transformer to not rely on having every token available – it must learn to encode the sequence in such a way that even if nearly $L - K$ tokens worth of information are dropped, the remaining K still capture the essential context to rebuild the sequence. This pushes $\mathcal{E}_\psi$ to generate a more compact, salient representation.

Adaptive Masked Token Modeling Beyond reconstruction, we also devise a masking strategy to predict the informative tokens. We leverage the latent merging outcome $\mathcal{S}'$ to decide which tokens to mask. The key idea is to assign a higher masking probability to tokens deemed important (those not merged heavily) and lower probability to tokens that were aggressively merged (low information). Given $\mathcal{S}' \in 0, 1^{K\times L}$ from the Latent Encoder, we compute an importance probability for each of the L local tokens. Let $g_i = \sum_{j=1}^{L} S' * i, j$ be the number of original tokens (out of L) that were grouped into the i-th latent token. We assign each latent group i a weight inversely proportional to its size, *e.g.*, $w_i = \frac{1}{g_i}$. For each token j that belongs to group i, we set $P_L(j) \propto \frac{w_i}{g_i}$, and choose the normalizing constant such that $\sum_{j=1}^{L} P_L(j) = 1$. This yields a probability vector $P_L \in \mathbb{R}^L$ over the local tokens, where tokens in large merged groups (large g_i) receive low probability and tokens in singleton or small groups receive higher probability. We then sample exactly K tokens without replacement according to P_L to mask. Letting $M_L \in 0, 1^L$ be the mask indicator, we map this mask back to the input space via the source matrix, $M_N = \mathcal{U}(M_L, \mathcal{S}) \in 0, 1^N$. In other words, if a merged token is selected to be masked, all of its constituent base positions in X will be masked out. Finally, we feed the masked sequence $X * M_N$ through the entire network without latent token merging, and get an output $\tilde{X}$. We define an Adaptive Masked Token Modeling (AMTM) loss as:

$$\mathcal{L}_{AMTM}(\theta) = -\frac{1}{K}\sum_{i:\, M_N(i)=1} \log P(\hat{X}_i \mid X * M_N; \theta). \tag{7}$$

This is essentially a masked language modeling loss focused on the K high-information tokens (and their base positions), ignoring the easy/redundant tokens. Overall, our full pre-training objectives involve three losses computed in three forward passes, which can be computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{MTR}(\theta) + \lambda \mathcal{L}_{MTR}(\theta \setminus \{\phi\}) + \mathcal{L}_{AMTM}(\theta), \tag{8}$$

where λ denotes a down-weighting factor, we set $\lambda = 0.25$ in practice, which ensures that the model learns to recover dropped information without overweighting the low-information content in its training signal.

Method	HyenaDNA	Caduceus-16	DNABERT	DNABERT2	GENA-LM	NT-500M	VQDNA	MxDNA	ConvNova	GENERator	**MergeDNA**
Date	NeurIPS'23	ICML'24	Bioinfo'21	ICLR'24	NAR'23	NM'24	ICML'23	NeurIPS'24	ICLR'25	arXiv'25	**Ours**
# Params	6.6M	7.9M	86M	117M	113M	500M	93M	100M	1.7M	1.3B	380M
Architecture Type	byte+SSM	byte+SSM	6-mer+A	BPE+A	BPE+A	6-mer+A	VQ+A	DC+A	byte+CNN	6-mer+A	byte+A
Pre-training Task	AR	AR	BERT	BERT	BERT	BERT	BERT	BERT	BERT	AR	MTR+AMTM
Enhancers (3 tasks)	80.88	79.96	80.14	82.81	83.22	84.56	82.37	82.79	80.90	84.87	**85.11**
Species Classification (2 tasks)	93.61	94.65	94.74	95.49	95.11	96.64	95.79	96.46	95.50	**96.95**	96.84
Regulatory Elements (3 tasks)	88.89	85.97	83.42	86.33	87.89	89.05	87.62	90.57	87.30	90.30	**90.66**
Average (8 tasks)	87.07	85.89	85.02	87.30	87.94	89.26	87.69	89.12	86.95	90.71	**90.87**

Table 1: Comparison on Genomic Benchmarks. Top-1 accuracy (%) averaged over several similar tasks is reported for popular DNA foundation models with SFT evaluation. The best and the second best results are marked as the **bold** and underlined types.

Method	HyenaDNA	Caduceus-PS	DNABERT	GROVER	DNABERT2	NTv2-500M	MxDNA	ConvNova	GENERator	**MergeDNA**
Date	NeurIPS'23	ICML'24	Bioinfo'21	bioRxiv'23	ICLR'24	NM'24	NeurIPS'24	ICLR'25	arXiv'25	**Ours**
# Params (M)	6.6M	1.9M	86M	87M	117M	500M	100M	1.7M	1.2B	380M
H3	78.14	80.48	77.41	76.80	79.31	78.17	82.78	81.50	80.60	**82.95**
H3K4me1	44.52	52.83	43.83	46.10	48.34	51.64	56.15	**56.60**	55.30	56.24
H3K4me2	42.68	49.88	32.38	40.30	43.02	37.24	55.59	**57.45**	42.40	55.67
H3K4me3	50.41	56.72	31.49	45.80	45.43	50.30	63.68	**67.15**	51.20	64.10
H3K9ac	58.50	63.27	52.55	62.60	60.04	61.05	64.78	**68.10**	61.20	65.01
H3K14ac	56.71	60.84	46.51	54.80	54.49	57.22	68.27	**70.71**	60.50	68.51
H3K36me3	59.92	61.12	50.98	56.30	57.58	60.50	67.05	68.31	65.70	**68.19**
H3K79me3	66.25	67.17	60.48	58.10	64.38	65.78	**74.29**	72.08	67.00	74.23
H4	78.15	80.10	79.60	76.90	78.18	79.87	81.18	81.12	**81.50**	81.06
H4ac	54.15	59.26	41.53	53.00	51.80	55.22	**67.65**	66.10	59.20	67.26
Enhancer	53.13	55.20	79.13	51.60	52.50	54.51	**79.90**	57.60	58.00	79.84
Enhancer Types	48.16	47.17	54.73	43.30	44.32	43.36	60.50	49.75	47.70	**60.62**
Promoter All	95.57	96.65	97.05	92.60	96.23	96.82	97.16	96.82	96.20	**97.40**
Promoter Non-TATA	95.86	96.31	97.02	92.50	97.17	**97.45**	97.24	96.76	96.20	97.35
Promoter TATA	95.88	96.21	96.22	89.10	**96.99**	96.53	96.01	96.34	94.80	96.70
All	94.05	92.87	97.83	91.90	93.75	98.15	98.14	96.33	97.80	**98.35**
Accpetor	96.98	94.21	97.81	91.20	97.49	97.99	98.01	96.23	98.10	**98.67**
Donor	95.27	94.69	98.43	88.80	94.33	98.50	98.10	96.62	97.80	**98.93**
Average (18 tasks)	70.24	72.50	68.61	67.32	69.74	71.13	78.14	76.42	72.84	**78.39**

Table 2: Comparison on NT Benchmark. Matthews Correlation Coefficient (MCC) (%) or F1 score (%) is reported for 18 subtasks with SFT evaluation. The best and the second best results are marked as the **bold** and underlined types.

4 Experiments

4.1 Experimental Setup

Implementations. Following the Transformer architecture as LLaMA (Touvron et al. 2023), MergeDNA (380M parameters) adopts an embed dimension of $D = 1024$ and the local window size of 16. The Local Encoder and Decoder stack 4 and 2 Local ToMeAttention blocks, while the Latent Encoder and Latent Decoder use 20 and 4 Transformer blocks. Following DNABERT-2 (Zhou et al. 2023), we pre-train MergeDNA on the `Multi-Species Genomes` corpus using AdamW optimizer (Loshchilov and Hutter 2019) for 100K iterations with a base learning rate of 1×10^{-4} and the sequence length of 4096. Hierarchical compression yields a local encoder output length $L{=}\frac{N}{2}$ and a latent encoder length $K{=}\frac{L}{2}$, effectively modeling long-range context with reduced complexity. For downstream tasks, we adhere to the benchmark-specific SFT protocols. On sequence-level tasks, we discard both decoders and fine-tune a classification head on the latent encoder's output. For token-level (base-resolution) tasks, we retain the Local Decoder to recover sequence resolution and fine-tune a new token-level prediction head. All experiments are conducted with PyTorch and NVIDIA A100-80G GPUs for three trials.

Comparison Baselines. We compare MergeDNA with state-of-the-art genomics models across four architecture paradigms: (1) sequence modeling architectures with SSMs have HyenaDNA (Nguyen et al. 2024b) and Caduceus (Schiff et al. 2024), (2) Standard Transformers have DNABERT (Ji et al. 2021), DNABERT-2 (Zhou et al. 2023), NTv1/v2 (Dalla-Torre et al. 2023), GROVER (Sanabria, Hirsch, and Poetsch 2023), and GenSLM (Zvyagin et al. 2022)), (3) Hybrid models have Evo variants (Nguyen et al. 2024a) and HyriDNA (Ma et al. 2025)), and (4) CNN is ConvNova (Bo et al. 2025). As for the tokenizer, four popular types are compared in Table 1: (1) Byte-level like Evo, (2) k-mer like NTv2, (3) BPE like DNABERT2, (4) DNA dynamic tokenizer methods have VQDNA (Li et al. 2024) and MxDNA (Qiao et al. 2024). All baseline foundation models were pre-trained with standard masked language modeling (BERT) or autoregressive (AR) objectives. As for downstream tasks, typical specialist models are also included.

4.2 Comparison Results on Genomic Benchmarks

Genomic Benchmarks. We first evaluate on eight representative tasks from the Genomic Benchmark suite (Grešová et al. 2023), covering enhancer identification, species classification, and regulatory element prediction. All models are fine-tuned on each task, and we report top-1 accuracy following the GenBench protocol. As shown in Table 1, MergeDNA achieves the highest overall accuracy (90.87%), outperforming all prior DNA foundation models. Notably, it yields state-of-the-art results on the enhancer tasks (85.11%

Method	HyenaDNA	Caduceus-PS	DNABERT	NT-multi	DNABERT2	VQDNA	MxDNA	ConvNova	HybriDNA-7B	**MergeDNA**
Date	NeurIPS'23	ICML'24	Bioinfo'21	NM'24	ICLR'24	ICML'24	NeurIPS'24	ICLR'25	arXiv'25	**Ours**
# Params (M)	6.6M	1.9M	86M	2.5B	117M	93M	100M	1.7M	7B	380M
Epigenetic Marks Prediction (10)	58.94	58.39	49.08	58.06	55.98	57.95	67.29	68.91	63.05	**68.82**
Human TF Detection (3)	61.74	–	64.17	63.34	70.11	70.56	–	–	**72.89**	72.24
Mouse TF Detection (3)	64.37	–	56.43	67.02	67.99	69.80	–	–	**78.02**	73.21
Core Promoter Detection (3)	69.22	–	71.81	71.63	70.53	73.37	–	–	71.37	**73.41**
Promoter Detection (3)	80.14	–	81.69	**88.15**	84.21	86.58	–	–	85.53	87.73
Splice Site Reconstructed (1)	77.76	–	84.07	89.35	84.99	89.53	–	–	**90.09**	89.95
Virus Covid Classification (1)	25.88	–	55.50	73.04	71.02	74.32	–	–	74.02	**74.41**
Average (24 tasks)	62.58	58.39	60.53	67.23	66.43	68.51	67.29	68.91	76.42	**77.11**

Table 3: Comparison on GUE Benchmark. Matthews Correlation Coefficient (MCC) (%) or F1 score (%) averaged across 24 sub-tasks is reported with SFT evaluation. The best and the second best results are marked as the **bold** and underlined types.

Method	SpliceAI	DNABERT2	NT-500M	Caduceus	Evo2-7B	**MergeDNA**
# Params	3.5M	117M	500M	7.9M	7B	380M
Donor	57.4	63.5	55.7	64.2	**64.5**	64.4
Acceptor	69.1	70.7	72.2	74.0	74.3	**74.5**
Mean	63.2	67.1	63.9	69.1	69.2	**69.8**

Table 4: Comparison on Splicing Prediction on the SpliceAI dataset, where the AUROC score is reported.

Method	GenSLM-2.5B	NT-2500M	ESM2-650M	Evo-7B	Evo2-7B	**MergeDNA**
# Params	2.5B	2.5B	650M	7B	7B	380M
Bacteria	24.7	9.4	51.2	45.30	**45.85**	42.72
Human	6.9	4.7	37.5	11.10	**36.9**	20.58

Table 5: Comparison on Protein Fitness Prediction. Zero-shot SRCC (%) is reported on DMS datasets.

vs 84.87% by the second best) and regulatory element tasks, while maintaining competitive performance on species classification (second only to a larger model).

Nucleotide Transformer Benchmarks. We also compare on the popular Nucleotide Transformer (NT) benchmark (Dalla-Torre et al. 2023) as summarized in Table 2. This benchmark includes a diverse mix of epigenomic classification (measured by MCC or F1) and core promoter/splice site detection tasks. MergeDNA attains the best overall performance with an average score of 78.39, slightly surpassing the dynamic tokenizer MxDNA (78.14) and substantially higher than other baselines. In particular, our model consistently ranks at or near the top on most individual tasks (*e.g.*, yielding the best MCC on 10 out of 18 tasks).

GUE Benchmarks. We further evaluate on the Genome Understanding Evaluation (GUE) suite introduced by DNABERT2 (Zhou et al. 2023), which aggregates 24 short-range subtasks grouped into seven practical genomic applications: Epigenetic Mark Prediction (Yeast), Transcription Factor (TF) binding site detection (Human and Mouse), Promoter and Core Promoter Detection, Splice Site Prediction, and Virus Genomic Classification. We use Matthews Correlation Coefficient (MCC) or F1 score, as in prior work, and baseline results are taken from DNABERT-2 or the original papers for consistency. As summarized in Table 3, MergeDNA delivers the highest mean performance (77.11%), edging out the much larger HybriDNA-7B (76.42%) and outperforming all other foundation models. These results highlight that MergeDNA's dynamic tokenization and dual-context pre-training yield broad improvements across heterogeneous genomic prediction tasks.

4.3 Multi-omics Downstream Tasks

RNA Splicing Site Prediction. Pre-mRNA splicing is a crucial step in gene expression, and we evaluate our model on the SpliceAI dataset (Jaganathan et al. 2019), which provides long pre-mRNA sequences labeled with donor and acceptor splice sites. We treat this as a binary sequence classification (site vs. non-site) and report the area under the ROC curve (AUROC). In Table 4, MergeDNA achieves a mean AUROC of 69.8, substantially outperforming the classic SpliceAI model (63.2) and all prior DNA foundation models. MergeDNA nearly matches the 7B-parameter Evo2 on donor site prediction and exceeds it on acceptor sites.

Long-range Expression Prediction. We next consider two challenging expression quantitative trait tasks from the Genomics Long-Range Benchmark (LRB) (Trop et al. 2024) that demand modeling of kilobase-scale contexts. (i) Causal eQTL Effect Prediction: given a genomic locus and a candidate variant, predict if the variant alters gene expression (evaluated by AUROC). (ii) Bulk RNA Expression Prediction: predict gene expression levels from the surrounding DNA sequence (evaluated by R^2 correlation). As shown in Table 6, MergeDNA attains new state-of-the-art results on both tasks: an AUROC of 0.75 for eQTL (*vs* 0.74 by the best baseline) and an R^2 of 0.62 for bulk expression (*vs* 0.60).

Protein Fitness Prediction. Finally, we further evaluate MergeDNA in a strict zero-shot setting on protein fitness prediction tasks using Deep Mutational Scanning (DMS) data (Notin et al. 2022). Here, models must predict the functional fitness of protein variants (amino acid mutations) directly from the DNA coding sequence, without any fine-tuning on protein data. Table 5 reports Spearman's rank correlation coefficient (SRCC) between predicted and actual fitness on two representative DMS datasets (one bacterial protein and one human protein). A specialized protein language model (ESM2 (Lin et al. 2022), gray in the table) achieves the highest scores as expected. Among DNA-based models, MergeDNA shows strong cross-omics generalization: for the bacterial protein, it obtains 42.7% SRCC—on par with the 7B Evo model and only slightly behind the multi-omics Evo2 (45.9%). On the human protein, MergeDNA (20.6% SRCC) substantially outperforms earlier DNA models like GenSLM (6.9%) and the original Evo (11.1%), though it trails Evo2, which leverages direct protein training.

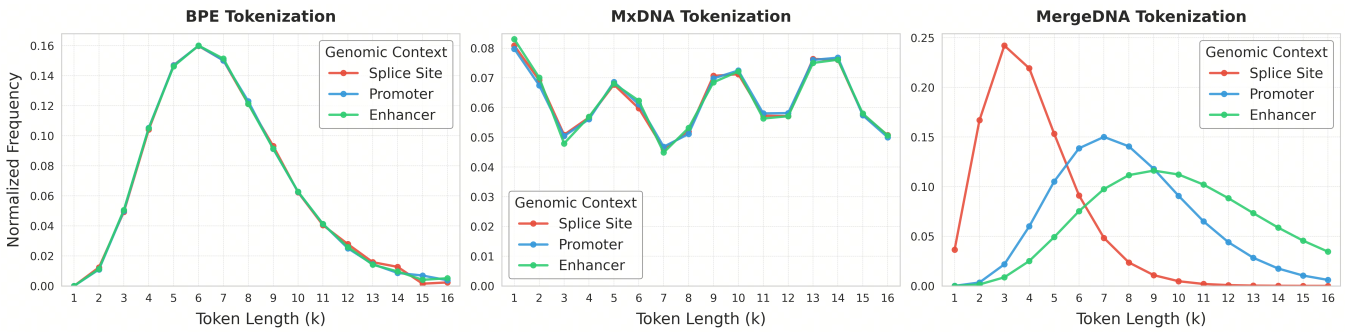


Figure 3: Visualization of Token Length Distributions for (a) BPE (Zhou et al. 2023), (b) MxDNA (Qiao et al. 2024), and (c) MergeDNA across different genomic contexts. Baseline tokenizers show a static, context-agnostic distribution, while MergeDNA adaptively changes its tokenization strategy based on the sequence type, demonstrating strong context-awareness.

Method	DNABERT2	DNABERT-S	NTv2-500M	HyenaDNA-160K	Caduceus-131K	HybridDNA-131K	Evo2-7B	**MergeDNA**
# Params	117M	117M	500M	12.9M	7.7M	300M	7B	380M
Causal eQTL (AUROC)	0.72	0.73	0.72	0.71	0.68	0.74	0.74	**0.75**
Bulk RNA (R^2)	0.51	0.52	0.60	0.46	0.52	0.52	0.60	**0.62**

Table 6: Comparison on LRB Benchmark with Causal eQTL Variant Effect Prediction and Bulk RNA Expression Prediction.

4.4 Empirical Analysis of Tokenization

To understand how MergeDNA learns to parse genomic sequences, we analyze the vocabularies learned by comparing MergeDNA with two representative baseline tokenizers, **BPE** and **MxDNA**. We visualize the normalized frequency of token lengths (from 1-mer to 16-mer) across different genomic contexts: promoters, enhancers, and splice sites. The BPE tokenizer in Figure 3(a) produces a fixed, long-tailed distribution that peaks around a token length of 6, regardless of the underlying sequence type. Similarly, the MxDNA tokenizer in Figure 3(b) yields a relatively uniform token length distribution that also shows minimal variation across different genomic contexts. This context-agnostic behavior limits their ability to adaptively capture functionally relevant motifs of varying lengths. Contrast in Figure 3(c), MergeDNA's local encoder shows strong context-awareness by learning to produce different token length distributions tailored to the biological properties of the input sequence. For longer and more complex regulatory regions like promoters and enhancers, the distributions shift towards longer tokens (peaking at $k = 7$ and $k = 9$, respectively). This data-driven tokenization allows MergeDNA to dynamically capture genomic motifs at their relevant biological scales.

4.5 Ablation Study

We conducted ablation experiments on the Genomic Benchmark tasks to quantify the contribution of each component in MergeDNA. Table 7 summarizes the average top-1 accuracy on the Genomic Benchmark 8-task under various configurations. First, we examine the impact of hierarchical architecture. Replacing the first 4 Transformer layers with our Local Encoder (merging tokens in windows) improves performance by +0.39 with the same parameter budget, confirming the benefit of local token merging. Next, we ablate the pre-training objectives. Training with only the naive masked token modeling (MTM) objective results in suboptimal performance; adding the Merged Token Reconstruction ($\mathcal{L}MTR$) objectives targeting the tokenizer's output provides a large gain (+1.03), and further introducing the Adaptive MTM on the filtered tokens pushes the improvement to +1.57 over the baseline. We also find that scaling down the loss weight λ for the latent $\mathcal{L}MTR$ (*i.e.*, not directly updating tokenizer parameters) to 0.25 is crucial for better generalization, yielding the best result. Finally, we vary the depth of the Local Encoder: using only 2 local merging layers (with correspondingly more latent decoder layers) degrades performance, indicating that a deeper tokenizer (4 merging blocks) is important for capturing rich subword representations.

Tokenizer	Latent Enc.	Local Dec.	Pre-training	Acc.
Byte	24 layers	2 layers	$\mathcal{L}_{MTM}$	89.30
Local Enc. (4)	20 layers	2 layers	$\mathcal{L}^{\theta}_{MTR} + \mathcal{L}_{MTM}$	+0.39
Local Enc. (4)	20 layers	2 layers	$\mathcal{L}^{\theta}_{MTR} + \mathcal{L}^{\theta\setminus\{\phi\}}_{MTR} + \mathcal{L}_{AMTM}$	+1.03
Local Enc. (4)	20 layers	2 layers	$\mathcal{L}^{\theta}_{MTR} + \lambda\mathcal{L}^{\theta\setminus\{\phi\}}_{MTR} + \mathcal{L}_{AMTM}$	**+1.57**
Local Enc. (2)	20 layers	4 layers	$\mathcal{L}^{\theta}_{MTR} + \lambda\mathcal{L}^{\theta\setminus\{\phi\}}_{MTR} + \mathcal{L}_{AMTM}$	+1.21

Table 7: Ablation Study of Life-Code and pre-training tasks with DNA, protein, and central dogma (CD) tasks. Note that the blue background denotes the selected setups.

5 Conclusions

This paper introduces MergeDNA, a context-aware DNA foundation model that addresses fundamental challenges in genome modeling: heterogeneous information density, ambiguous sequence tokenization, and long-range dependencies. MergeDNA unifies a differentiable local tokenizer and a global latent Transformer through a hierarchical architecture and two complementary pre-training tasks, *i.e.*, Merged Token Reconstruction and Adaptive Masked Token Modeling. Extensive experiments on three standard DNA benchmarks and multi-omics tasks demonstrate that MergeDNA achieves state-of-the-art performance with strong generalization across modalities, offering a scalable and principled approach to genome-scale representation learning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Project No. 624B2115, 623B2086, and U21A20427), the Science & Technology Innovation 2030 Major Program (Project No. 2021ZD0150100), the Center of Synthetic Biology and Integrated Bioengineering at Westlake University (Project No. WU2022A009), and the Westlake University Industries of the Future Research Program (Project No. WU2023C019). This work was done when Siyuan Li interned at BioMap Research. We thank the GPU support from BioMap Research and the AI station of Westlake University.

References

Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.

Benegas, G.; Batra, S. S.; and Song, Y. S. 2023. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44): e2311219120.

Bo, Y.; Mao, W.; Shao, Y.; Bai, W.; Ye, P.; Ma, X.; Zhao, J.; Chen, H.; and Shen, C. 2025. Revisiting Convolution Architecture in the Realm of DNA Foundation Models. *arXiv preprint arXiv:2502.18538*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *International Conference on Learning Representations*.

Brixi, G.; Durrant, M. G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G. A.; King, S. H.; and et al. 2025. Genome modeling and design across all domains of life with Evo 2. *Arc Institute Manuscripts*.

Cooper, S. 1981. The central dogma of cell biology. *Cell biology international reports*, 5(6): 539–549.

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.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186.

Ge, H.; Feng, J.; Huang, Q.; Fu, F.; Nie, X.; Zuo, L.; Lin, H.; Cui, B.; and Liu, X. 2025. ByteScale: Efficient Scaling of LLM Training with a 2048K Context Length on More Than 12,000 GPUs. *ArXiv*, abs/2502.21231.

Grešová, K.; Martinek, V.; Čechák, D.; Šimeček, P.; and Alexiou, P. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1): 25.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *ArXiv*, abs/2312.00752.

Hwang, S.; Wang, B.; and Gu, A. 2025. Dynamic Chunking for End-to-End Hierarchical Sequence Modeling. *arXiv preprint arXiv:2507.07955*.

Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J. F.; Darbandi, S. F.; Knowles, D.; Li, Y. I.; Kosmicki, J. A.; Arbelaez, J.; Cui, W.; Schwartz, G. B.; Chow, E. D.; Kanterakis, E.; Gao, H.; Kia, A.; Batzoglou, S.; Sanders, S. J.; and Farh, K. K.-H. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176(3): 535–548.e24.

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.

Ji, Z.; Zhang, H.; Huang, J.; et al. 2023. GENA-LM: Multimodal Pre-training for the Central Dogma. *bioRxiv*.

Kudo, T.; and Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Conference on Empirical Methods in Natural Language Processing*.

Lee, D. H.; and Hong, S. 2024. Learning to Merge Tokens via Decoupled Embedding for Efficient Vision Transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Li, Q.; Wu, W.; Zhu, Y.; Feng, F.; Ye, J.; and Wang, Z. 2025. GENERanno: A Genomic Foundation Model for Metagenomic Annotation. *bioRxiv*.

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)*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

Liu, Y.; Ji, T.; Sun, C.; Wu, Y.; and Wang, X. 2024. Generation with Dynamic Vocabulary. In *Conference on Empirical Methods in Natural Language Processing*.

Liu, Z.; Li, S.; Chen, Z.; Xin, L.; Wu, F.; Yu, C.; Yang, Q.; Guo, Y.; Yang, Y.; and Li, S. Z. 2025. Life-Code: Central Dogma Modeling with Multi-Omics Sequence Unification. *ArXiv*, abs/2502.07299.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Ma, M.; Liu, G.; Cao, C.; Deng, P.; Dao, T.; Gu, A.; Jin, P.; Yang, Z.; Xia, Y.; Luo, R.; Hu, P.; Wang, Z.; Chen, Y.; Liu, H.; and Qin, T. 2025. HybriDNA: A Hybrid Transformer-Mamba2 Long-Range DNA Language Model. *ArXiv*.

Mendoza-Revilla, J.; Trop, E.; Gonzalez, L.; Roller, M.; Dalla-Torre, H.; de Almeida, B. P.; Richard, G.; Caton, J.; Lopez Carranza, N.; Skwark, M.; et al. 2023. A Foundational Large Language Model for Edible Plant Genomes. *bioRxiv*, 2023–10.

Nawrot, P.; Chorowski, J.; Łańcucki, A.; and Ponti, E. M. 2022. Efficient Transformers with Dynamic Token Pooling. arXiv:2211.09761.

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. 2024a. Sequence modeling and design from molecular to genome scale with Evo. *Science*, eado9336.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Wornow, M.; Birch-Sykes, C.; Massaroli, S.; Patel, A.; Rabideau, C.; Bengio, Y.; et al. 2024b. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36.

Notin, P.; Dias, M.; Frazer, J.; Marchena-Hurtado, J.; Gomez, A. N.; Marks, D. S.; and Gal, Y. 2022. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *ArXiv*, abs/2205.13760.

Pagnoni, A.; Pasunuru, R.; Rodriguez, P.; Nguyen, J.; Muller, B.; Li, M.; Zhou, C.; Yu, L.; Weston, J. E.; Zettlemoyer, L. S.; Ghosh, G.; Lewis, M.; Holtzman, A.; and Iyer, S. 2024. Byte Latent Transformer: Patches Scale Better Than Tokens. *ArXiv*, abs/2412.09871.

Qiao, L.; Ye, P.; Ren, Y.; Bai, W.; Liang, C.; Ma, X.; Dong, N.; and Ouyang, W. 2024. Model Decides How to Tokenize: Adaptive DNA Sequence Tokenization with MxDNA. In *Conference on Neural Information Processing Systems*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training.

Roy, S.; Sural, S.; and Ganguly, N. 2024. Unlocking Efficiency: Adaptive Masking for Gene Transformer Models. In *European Conference on Artificial Intelligence*.

Roy, S.; Wallat, J.; Sundaram, S. S.; Nejdl, W.; and Ganguly, N. 2023. GeneMask: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning. In *European Conference on Artificial Intelligence*.

Sanabria, M.; Hirsch, J.; and Poetsch, A. R. 2023. The human genome's vocabulary as proposed by the DNA language model GROVER. *bioRxiv*, 2023–07.

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 preprint arXiv:2403.03234*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural Machine Translation of Rare Words with Subword Units. *ArXiv*, abs/1508.07909.

Shao, B.; and Yan, J. 2024. A long-context language model for deciphering and generating bacteriophage genomes. *Nature Communications*, 15(1): 9392.

Slagle, K. 2024. SpaceByte: Towards Deleting Tokenization from Large Language Modeling. *ArXiv*, abs/2404.14408.

Song, L.; Segal, E.; and Xing, E. P. 2024. Toward AI-Driven Digital Organism: Multiscale Foundation Models for Predicting, Simulating and Programming Biology at All Levels. *ArXiv*, abs/2412.06993.

Thoutam, V.; and Ellsworth, D. 2024. MSAMamba: Adapting Subquadratic Sequence Models To Long-Context DNA MSA Analysis. *arXiv preprint*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.

Trop, E.; Schiff, Y.; Marroquin, E. M.; Kao, C. H.; Gokaslan, A.; Polen, M.; Shao, M.; de Almeida, B. P.; Pierrot, T.; Li, Y. I.; et al. 2024. The Genomics Long-Range Benchmark: Advancing DNA Language Models. *arXiv preprint*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *ArXiv*.

Wang, J.; Gangavarapu, T.; Yan, J. N.; and Rush, A. M. 2024. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660*.

Wu, S.; Tan, X.; Wang, Z.; Wang, R.; Li, X.; and Sun, M. 2024. Beyond Language Models: Byte Models are Digital World Simulators. arXiv:2402.19155.

Wu, W.; Li, Q.; Li, M.; Fu, K.; Feng, F.; Ye, J.; Xiong, H.; and Wang, Z. 2025. GENERator: A Long-Context Generative Genomic Foundation Model. *arXiv preprint*.

Yang, Z.; Fan, X.; Lan, M.; Li, X.; You, Y.; Tian, L.; Church, G.; Liu, X.; and Gu, F. 2024. Multiomic foundation model predicts epigenetic regulation by zero-shot. *bioRxiv*.

Yang, Z.; Zhu, J.; and Su, B. 2025. SPACE: Your Genomic Profile Predictor is a Powerful DNA Foundation Model. In *International Conference on Machine Learning (ICML)*.

Yu, L.; Simig, D.; Flaherty, C.; Aghajanyan, A.; Zettlemoyer, L.; and Lewis, M. 2023. MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers. *ArXiv*, abs/2305.07185.

Zhai, J.; Gokaslan, A.; Schiff, Y.; Berthel, A.; Liu, Z.-Y.; Lai, W.-Y.; Miller, Z. R.; Scheben, A.; Stitzer, M. C.; Romay, M. C.; et al. 2025. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proceedings of the National Academy of Sciences*, 122(24): e2421738122.

Zhang, D.; Zhang, W.; Zhao, Y.; Zhang, J.; He, B.; Qin, C.; and Yao, J. 2023. DNAGPT: A generalized pre-trained tool for versatile DNA sequence analysis tasks. *arXiv preprint arXiv:2307.05628*.

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*.

Zhou, Z.; Riley, R.; Kautsar, S.; Wu, W.; Egan, R.; Hofmeyr, S.; Goldhaber-Gordon, S.; Yu, M.; Ho, H.; Liu, F.; et al. 2025a. GenomeOcean: An Efficient Genome Foundation Model Trained on Large-Scale Metagenomic Assemblies. *bioRxiv*, 2025–01.

Zhou, Z.; Wu, W.; Ho, H.; Wang, J.; Shi, L.; Davuluri, R. V.; Wang, Z.; and Liu, H. 2025b. DNABERT-S: Pioneering species differentiation with species-aware DNA embeddings. *Bioinformatics*, 41: i255–i264.

Zhu, X.; Qin, C.; Wang, F.; Yang, F.; He, B.; Zhao, Y.; and Yao, J. 2024. CD-GPT: a biological foundation model bridging the gap between molecular sequences through central dogma. *bioRxiv*, 2024–06.

Zvyagin, M. T.; Brace, A.; Hippe, K.; Deng, Y.; Zhang, B.; Bohorquez, C. O.; Clyde, A.; Kale, B.; Perez-Rivera, D.; Ma, H.; et al. 2022. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv*.