The 40th Annual AAAI Conference on Artificial Intelligence

JANUARY 20 — JANUARY 27, 2026 | SINGAPORE

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*]

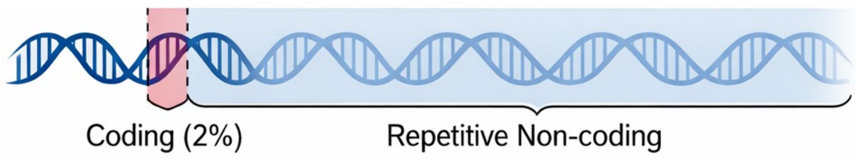
[1] Zhejiang University [2] Westlake University [3] Biomap Research

Introduction and Contributions

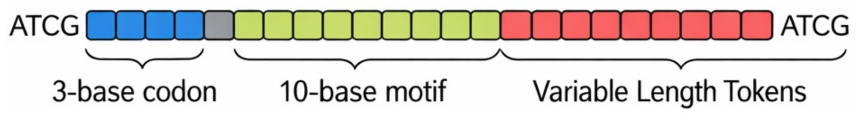
- **Unified Architectural Design:** A hierarchical framework that tightly integrates a learnable DNA tokenizer with adaptive sequence modeling. Leveraging differentiable token merging (ToMe) within local attention blocks, the Local Encoder captures irregular patterns and determines where to merge as words.
- **Adaptive Context Modeling:** Designing context-aware pre-training tasks to adapt information density in different genomics. Selecting informative positions, Merged Token Reconstruction and Adaptive Masked Token Modeling allow the model to capture both local motifs and global long-range dependencies.
- **Strong Empirical Results:** Achieving strong performance across three DNA benchmarks and shows generalization to several RNA and protein downstream tasks, outperforming works of DNA tokenization and foundation models.

Challenges

Uneven Information Density in DNA sequences

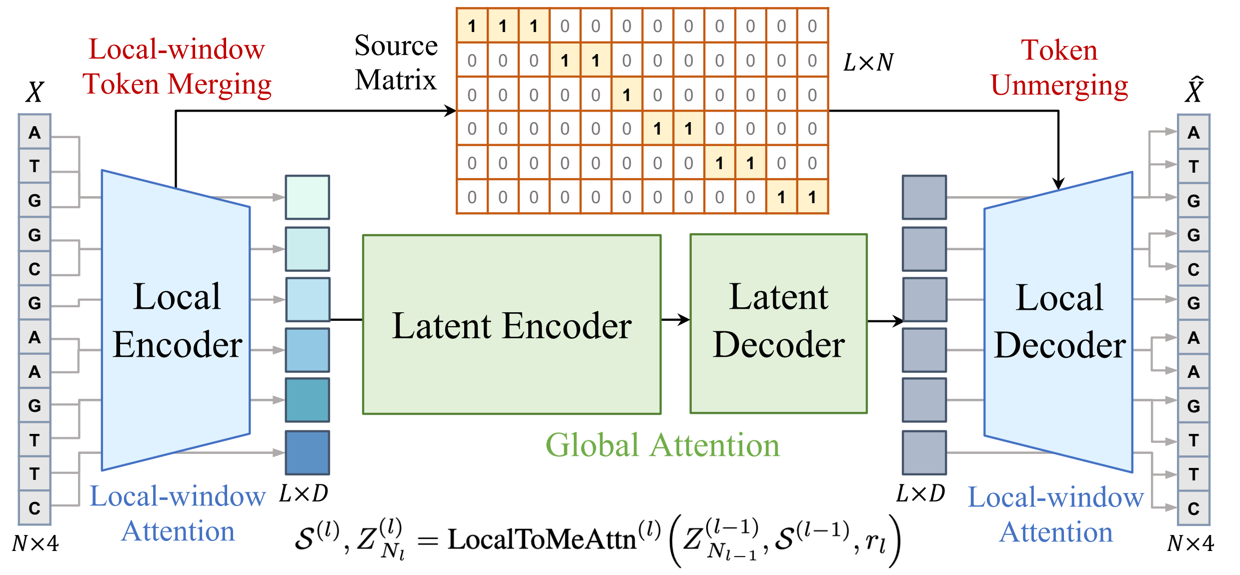


No Inherent "Words" of Genomics



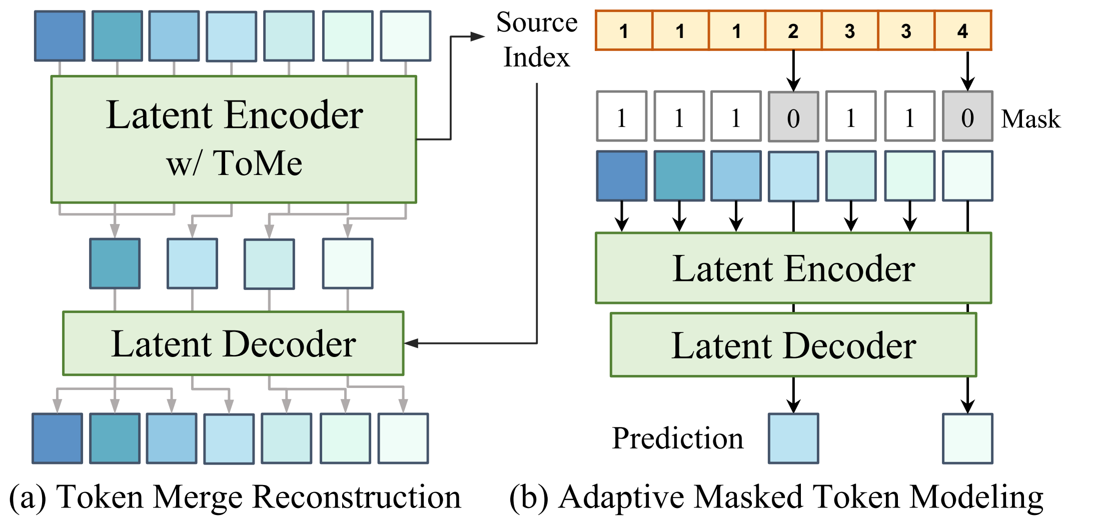
Extreme Sequence Length, capturing dependencies across millions of bases requires highly efficient Models

Method: The MergeDNA Architecture



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

Method: Content-aware Pre-training Tasks



(a) Token Merge Reconstruction

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

(b) Adaptive Masked Token Modeling

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

Comparison Experiments

- **DNA tasks:** Following Genomic, NT, and GUE benchmarks with SFT evaluation.
- **RNA tasks:** RNA Splicing Site Prediction on SpliceAI dataset and Long-range tasks in LRB (e.g., Causal eQTL Effect Prediction).
- **Protein tasks:** Protein Fitness Prediction on Deep Mutational Scanning (DMS) data with zero-shot evaluation.

Table 1: **Genomic Benchmarks**. Top-1 accuracy over similar tasks is reported with SFT evaluation.

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 2: **GUE Benchmark**. Matthews Corr. Coeff. (MCC) or F1 score are shown over 24 tasks with SFT evaluation.

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**

Empirical Study of Vocabulary

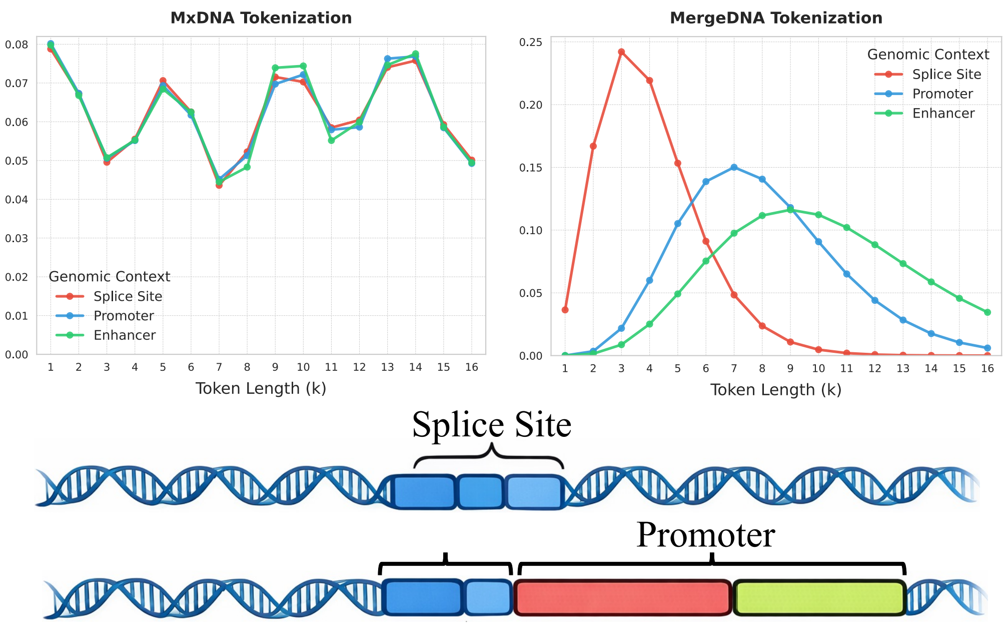


Table 3: **NT Benchmark**. MCC or F1 score are reported over 18 tasks with SFT evaluation.

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	ul65.01
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**