LDARNet

DNA Adaptive Representation Network

with Learnable Tokenization for Genomic Modeling

Daria Ledneva | Denis Kuznetsov

Moscow Independent Research Institute of Artificial Intelligence

ICML 2026

The Problem: Tokenization in Genomics

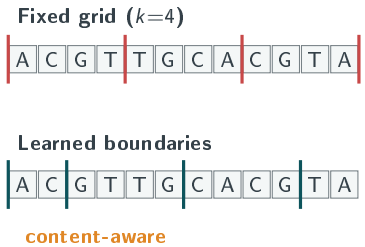
Genomic foundation models borrow LLM architectures but inherit **fixed tokenization**:

- ✎ ***k*-mers / BPE / single nucleotides**
- ✘ impose **arbitrary boundaries**
- ❧ no biological grounding

The question

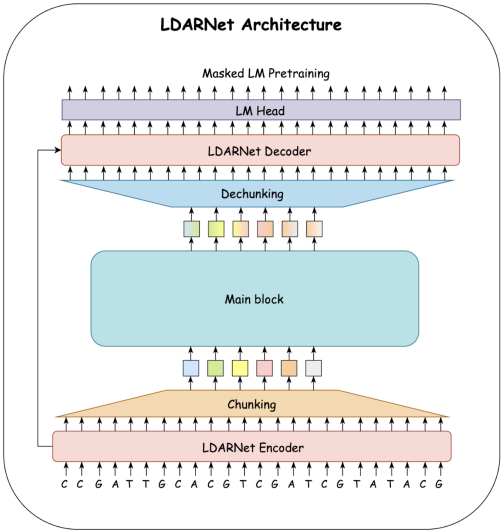
Can boundaries be **learned** from sequence — and do they capture functional structure?

H-Net (Hwang et al., 2025) learns this for *autoregressive* DNA. We bring it to *masked* modeling and downstream evaluation.



Fixed cuts split motifs; learned cuts can respect them.

LDARNet: H-Net Chunking → Masked Modeling



120M-param hierarchy

encoder–backbone–decoder

- **BiMamba-2** blocks (shared weights)
- one **local attention** layer in the encoder
- **dynamic chunking**: compress at learned boundaries (N=4)

Made bidirectional for DNA

Routing & dechunker read *both* directions — robust to `[MASK]`.

How Boundaries Are Learned

1. Bidirectional routing — low similarity between neighbors ⇒ a boundary:

$$\bar{s}_t = \tfrac{1}{2}(\cos(q_{t-1}, k_t) + \cos(q_t, k_{t-1}))$$

$$p_t = \tfrac{1}{2}(1 - \bar{s}_t), \qquad b_t = 1_{\{p_t > 0.5\}}$$

Averaging both orientations ⇒ robust to masked positions.

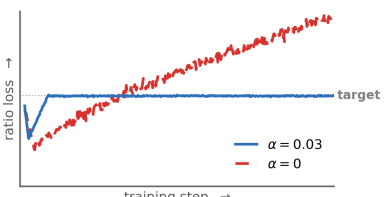
2. Bi-EMA dechunker restores resolution from *both* sides of each chunk.

3. Ratio loss keeps compression near target N:

Training objective

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \alpha\, \mathcal{L}_{\text{ratio}}$$

- ✔ no cost to reconstruction (MLM loss unchanged)
- ✔ without it, compression **drifts** freely



Result: Best Compact Model on 27 Tasks

11/18

NT wins among compact models

(<300M params; next best: 2)

Headline

SOTA on **5 histone tasks** — beating models up to **20× larger**.

Task (MCC)	DNA-BERT-2 117M	Hyena-DNA 55M	**LDAR Net** 120M	Gen. 1.2B 1.2B
H3K4me1	51.2	51.2	**58.3**	55.3
H3K4me2	33.3	45.5	**49.6**	42.4
H3K4me3	35.3	55.0	**57.6**	51.2
H3K79me3	61.5	66.9	**68.7**	67.0
H4ac	46.5	56.4	**62.3**	59.2
Enhancer	52.5	52.0	**57.7**	58.0
Splice all	95.0	91.7	94.2	97.8

Teal = best overall, incl. all large models.

Full 18-task NT + 9-task GB suites in paper.

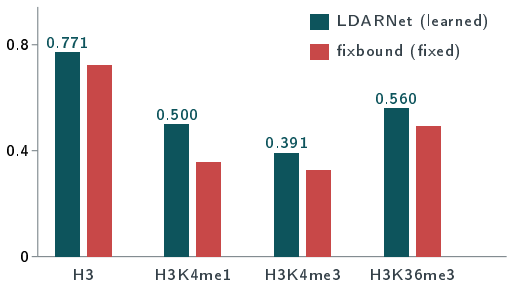
Why? Learned Routing — Isolated at Matched FLOPs

Same params, data, architecture, **same FLOP budget**. Only difference: **learned** vs **fixed** boundaries.

Histones: learned wins big

+14.3 pp on H3K4me1
+6.7 pp on H3K36me3, +4.8 pp on H3

Trade-off: fixed grid wins on *splice* (local GT/AG motifs) — tunable via N.

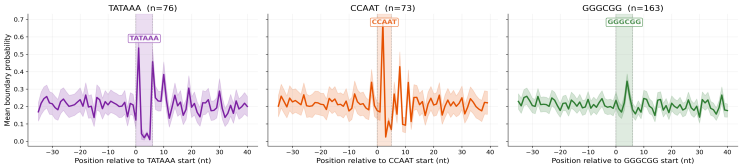


2.5M-param proxy, identical compute.

Why? Boundaries Land on Functional Elements

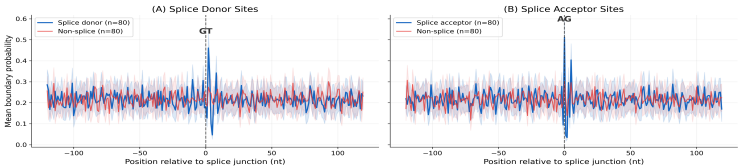
Promoter motifs

Figure R2: Learned boundaries around promoter motifs



Splice junctions

Figure R1: Learned boundary profiles at splice sites



Enriched, unsupervised

Boundary probability **peaks** at promoter motifs and at true GT/AG splice junctions vs. **matched controls**.

No motif supervision — the router treats binding sites & exon–intron transitions as **natural landmarks**.

Takeaways

What LDARNet shows

1. **Learnable tokenization** lets a 120M model match/beat models 4–20× larger.
2. Gains are **causal**: learned routing > fixed grid at matched FLOPs (up to +14.3 pp on H3K4me1).
3. Boundaries are **biologically interpretable** — aligned to promoter & splice landmarks.

Task-dependent benefit: learned routing dominates long-range epigenetics; fixed grids suit local motifs — tunable via N.

Next: multi-stage compression (>100 kb); zero-shot; multimodal (RNA-/ATAC-seq, Hi-C).

Adaptive > Scaling

Progress can come from **adaptive representations**, not only parameters.

Thank you!

Learned boundaries are biologically meaningful boundaries.

Daria Ledneva | a.ledn2026@gmail.com