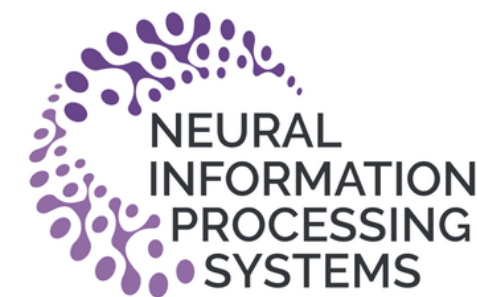


Context-Aware Regularization with Markovian Integration for Attention- Based Nucleotide Analysis

M.S. Refahi , M. Abavisani, B.A. Sokhansanj , J.R. Brown and G.L. Rosen



Motivation: DNA is a Language (with Challenges)

- **DNA as Language:** Genomic sequences can be treated like natural language
 - **Long-Range Dependencies:** Many genomic features depend on far-apart elements (enhancer–gene links have a median distance of **~28 kb** and can extend up to **~1 Mb**)
 - **Long-Context Information:** Some features require large genomic span (e.g., biosynthetic gene clusters).
 - **Repetitive Elements:** Huge portions of genomes are repeats (e.g. **~50% in humans' genome**)

Review Article | Published: 29 November 2011

Repetitive DNA and next-generation sequencing: computational challenges and solutions

[Todd J. Treangen](#) & [Steven L. Salzberg](#) ✉

[Nature Reviews Genetics](#) **13**, 36–46 (2012) | [Cite this article](#)

34k Accesses | 1536 Citations | 129 Altmetric | [Metrics](#)

Article | [Open access](#) | Published: 25 September 2023

Genome-wide enhancer–gene regulatory maps link causal variants to target genes underlying human cancer risk

[Pingting Ying](#), [Can Chen](#), [Zequn Lu](#), [Shuoni Chen](#), [Ming Zhang](#), [Yimin Cai](#), [Fuwei Zhang](#), [Jinyu Huang](#), [Linyun Fan](#), [Caibo Ning](#), [Yanmin Li](#), [Wenzhuo Wang](#), [Hui Geng](#), [Yizhuo Liu](#), [Wen Tian](#), [Zhiyong Yang](#), [Jiuyang Liu](#), [Chaoqun Huang](#), [Xiaojun Yang](#), [Bin Xu](#), [Heng Li](#), [Xu Zhu](#), [Ni Li](#), [Bin Li](#), ... [Xiaoping Miao](#) ✉

[+ Show authors](#)

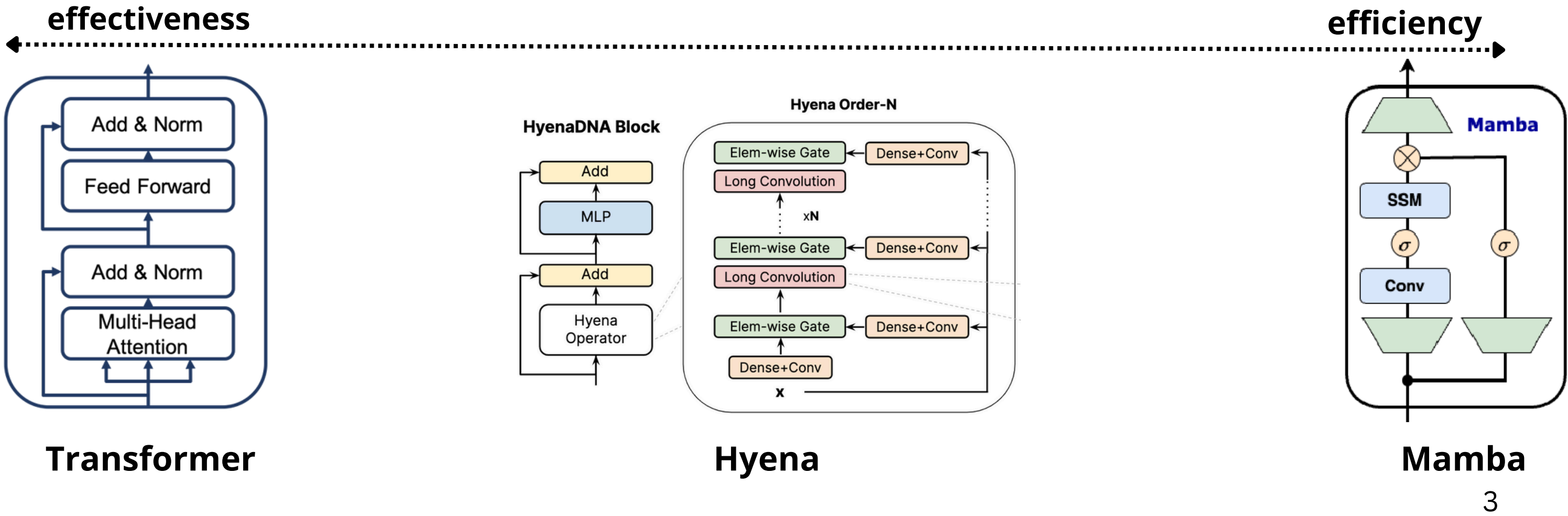
[Nature Communications](#) **14**, Article number: 5958 (2023) | [Cite this article](#)

21k Accesses | 24 Citations | 40 Altmetric | [Metrics](#)



Genomic Language Models

Model	Time Complexity	Max Seq Length	Models
Transformer	$O(n^2)$	<4-10k bp	NT, DNABERT2 , glm2, MetaBerta
Hyena	$O(n \log n)$	160k-1m bp	HyenaDNA, EVO
Mamba	$O(n)$	131k bp	Caduceus



LLMs Don't Count ! But Biology Needs Counting

ChatGPT 5 ▾

dna|

1/8

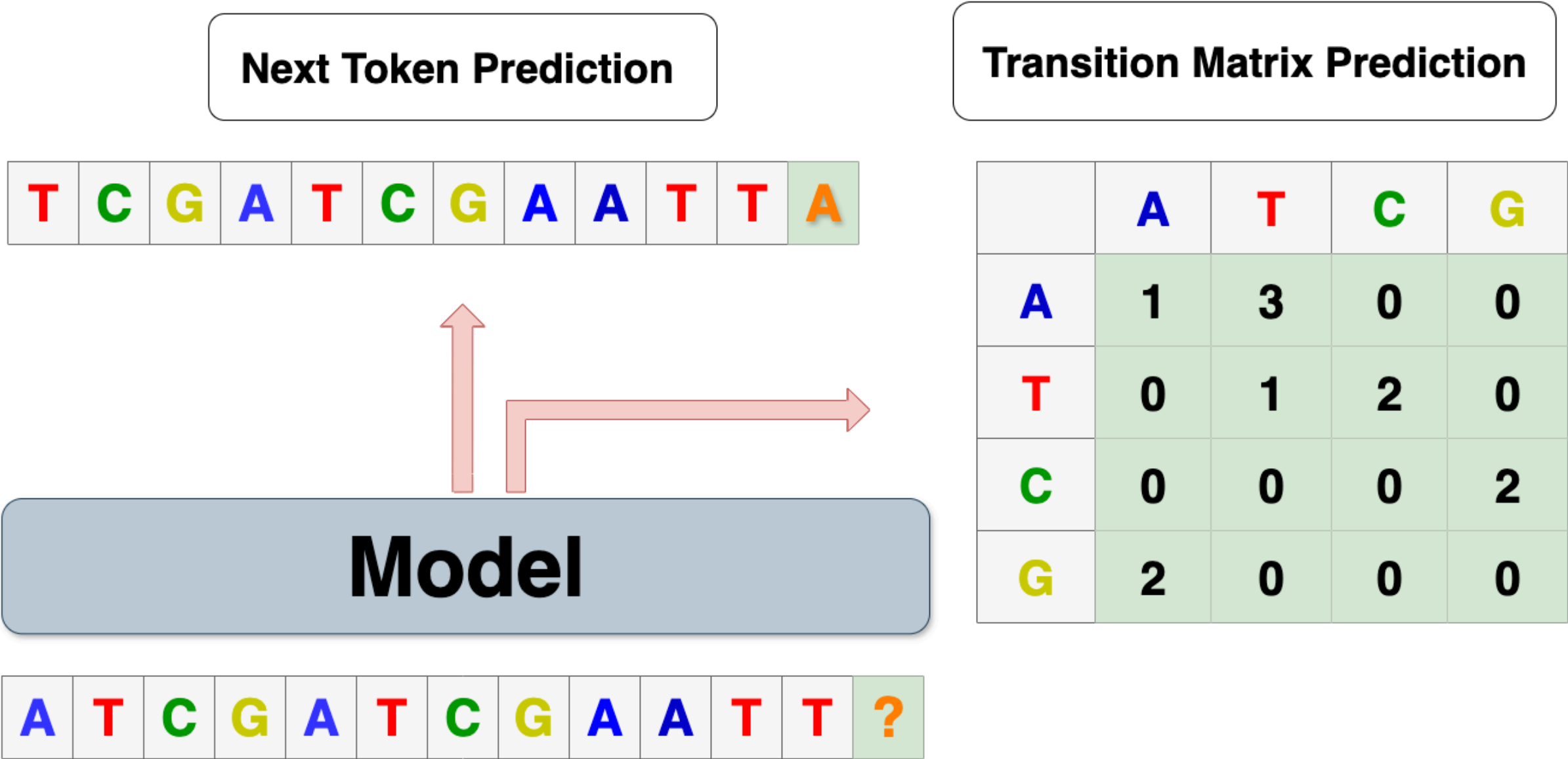


Write a meaningful paragraph that exactly includes at least 10 **DNA** ,
10 protein , and 10 RNA .

In modern biology, **DNA** serves as the foundational blueprint for life, encoding the instructions that guide the synthesis of **RNA** and **protein**. Within every cell, **DNA** replication ensures the faithful transmission of genetic material, while **RNA** transcription translates these **DNA** sequences into functional messages. Messenger **RNA**, transfer **RNA**, and ribosomal **RNA** each play crucial roles in interpreting **DNA** information and guiding **protein** assembly. The resulting **protein** molecules fold into intricate shapes that define their biological functions. Enzymatic **protein** complexes regulate **DNA** repair and **RNA** processing, ensuring stability across generations. Structural **protein** frameworks provide cellular integrity, while signaling **protein** molecules control responses to environmental cues. Ultimately, the interplay among **DNA**, **RNA**, and **protein** forms a dynamic triad—each **DNA**, **RNA**, and **protein** interaction sustaining the continuity and adaptability of life.



Can We Force an LLM to Behave like a Word Counter?




Transition Matrix Loss: Teaching the Model to Count

- **Ground-truth transition matrix:** Row-normalized bigram frequency matrix from the actual sequence.

$$p_{ij}$$

	A	T	C	G
A	5	5	5	5
T	0	1	1	1
C	5	1	2	2
G	1	1	2	1



	A	T	C	G
A	0.25	0.25	0.25	0.25
T	e	0.33	0.33	0.33
C	0.5	0.1	0.2	0.2
G	0.2	0.2	0.4	0.2

Transition Matrix Loss: Teaching the Model to Count

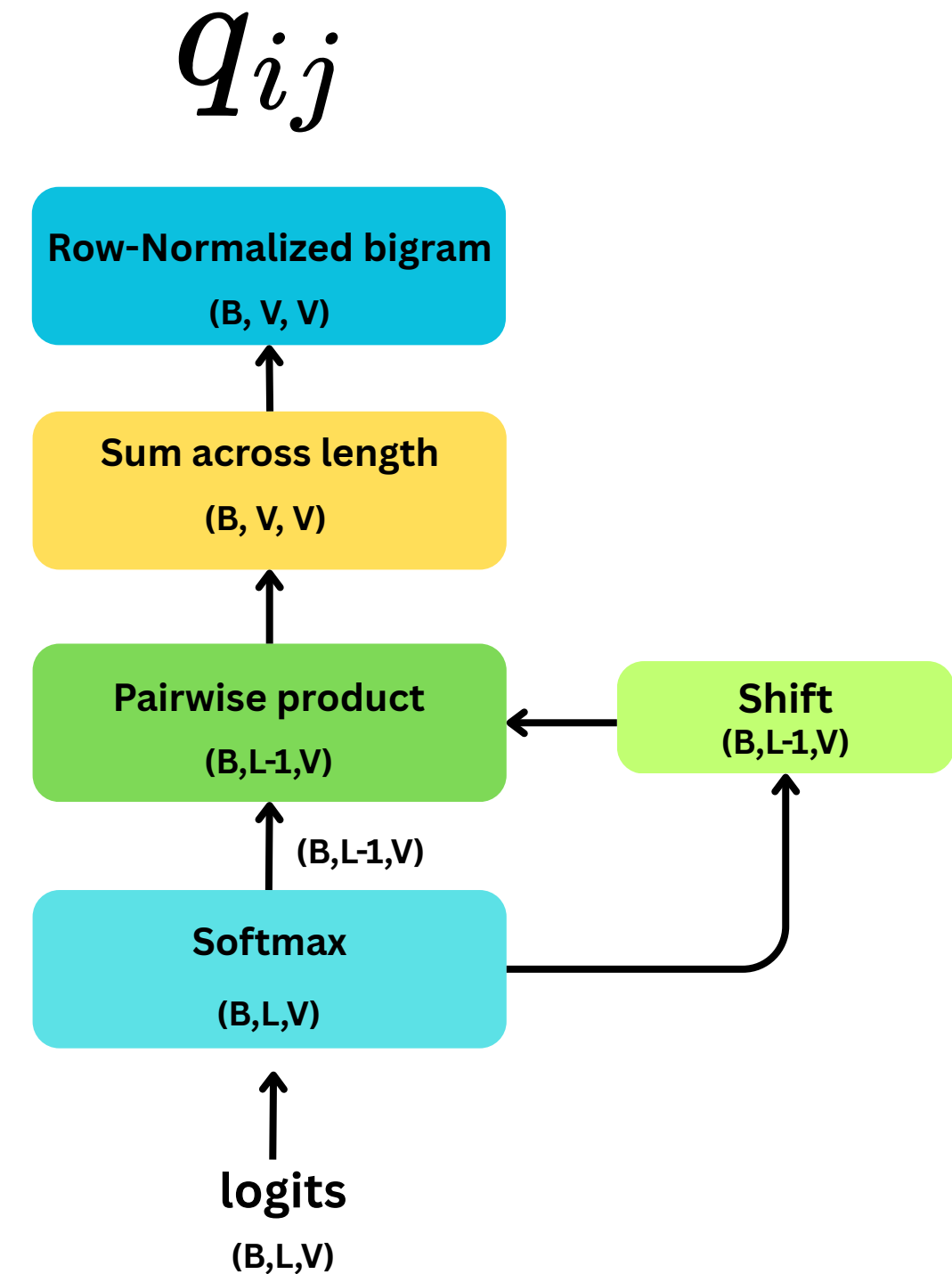
- **Ground-truth transition matrix:** Row-normalized bigram frequency matrix from the actual sequence.
- **Predicted transition matrix::** Computed from model outputs via pairwise probability products.

	A	T	C	G
A	5	5	5	5
T	0	1	1	1
C	5	1	2	2
G	1	1	2	1



	A	T	C	G
A	0.25	0.25	0.25	0.25
T	e	0.33	0.33	0.33
C	0.5	0.1	0.2	0.2
G	0.2	0.2	0.4	0.2

p_{ij}



Transition Matrix Loss: Teaching the Model to Count

- **Ground-truth transition matrix:** Row-normalized bigram frequency matrix from the actual sequence.
- **Predicted:** Computed from model outputs via pairwise probability products.
- **KL divergence loss:**

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

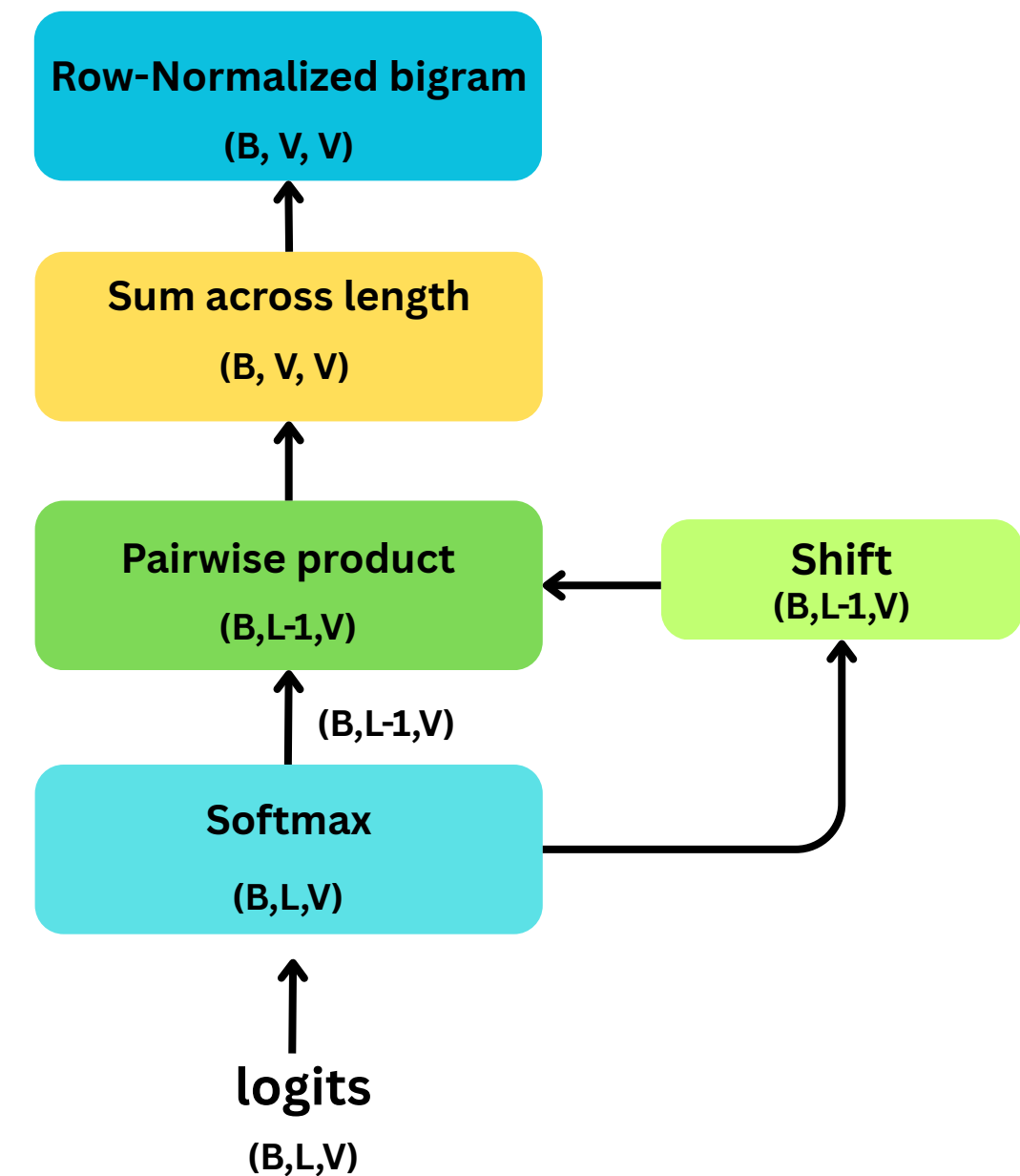
p_{ij}

	A	T	C	G
A	5	5	5	5
T	0	1	1	1
C	5	1	2	2
G	1	1	2	1



	A	T	C	G
A	0.25	0.25	0.25	0.25
T	e	0.33	0.33	0.33
C	0.5	0.1	0.2	0.2
G	0.2	0.2	0.4	0.2

q_{ij}



Markovian Nature of In-Context Learning in Transformers

Key Point :

In-context learning in Transformers often follows a Markovian pattern .

The model primarily depends on **a few preceding tokens**, not the entire sequence history.

Enables use of **sliding-window attention** instead of full attention ?!

The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains

Ezra Edelman*
University of Pennsylvania
ezrae@cis.upenn.edu

Nikolaos Tsilivis*
New York University[†]
nt2231@nyu.edu

Benjamin L. Edelman
Harvard University
bedelman@g.harvard.edu

Eran Malach
Harvard University
emalach@g.harvard.edu

Surbhi Goel
University of Pennsylvania
surbhig@cis.upenn.edu

From Self-Attention to Markov Models: Unveiling the Dynamics of Generative Transformers

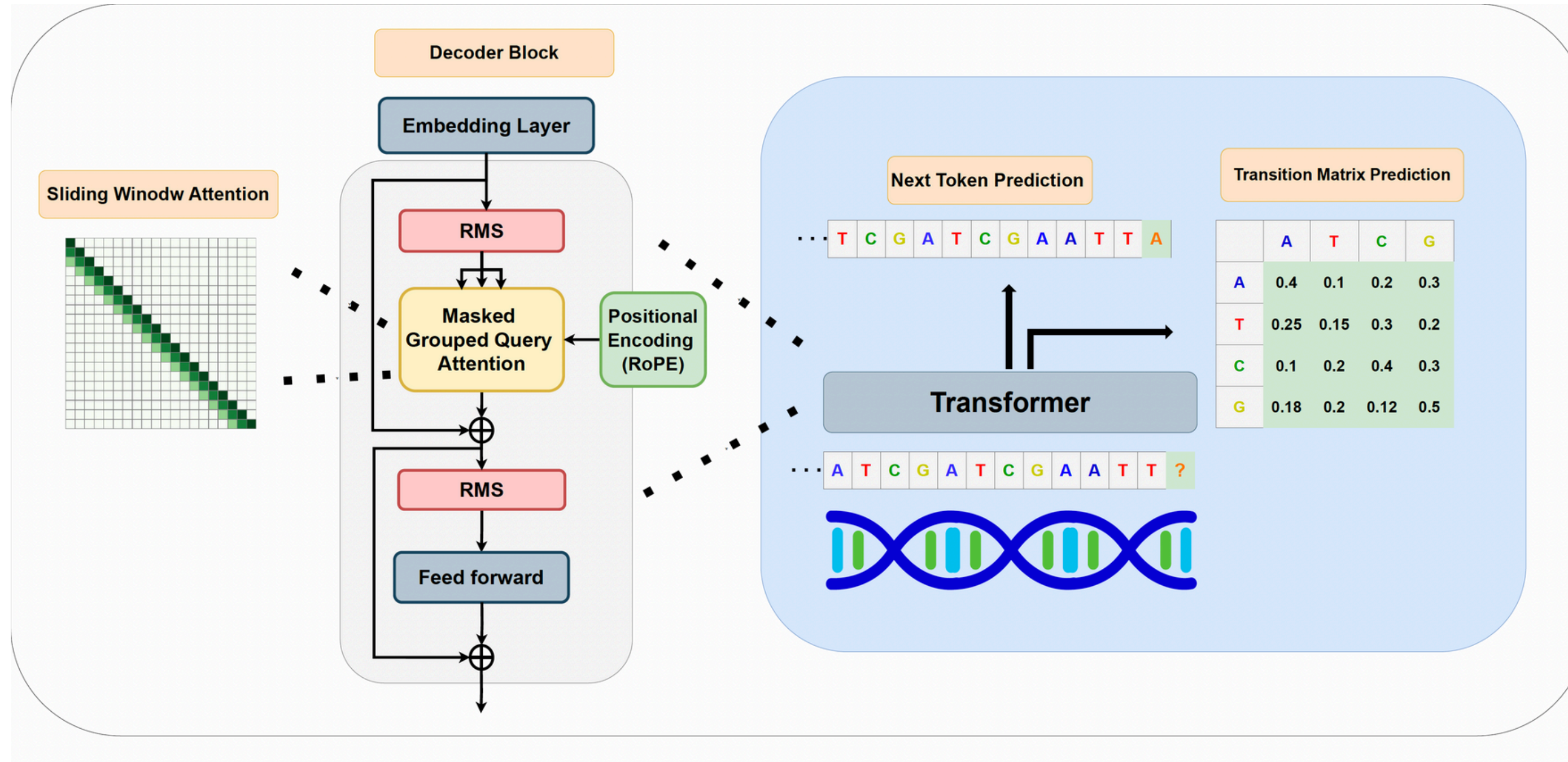
M. Emrullah Ildiz¹ Yixiao Huang¹ Yingcong Li¹
Ankit Singh Rawat² Samet Oymak¹

¹ University of Michigan, Ann Arbor
{eildiz,yingcong,oymak}@umich.edu, yixiao.huang@my.cityu.edu.hk

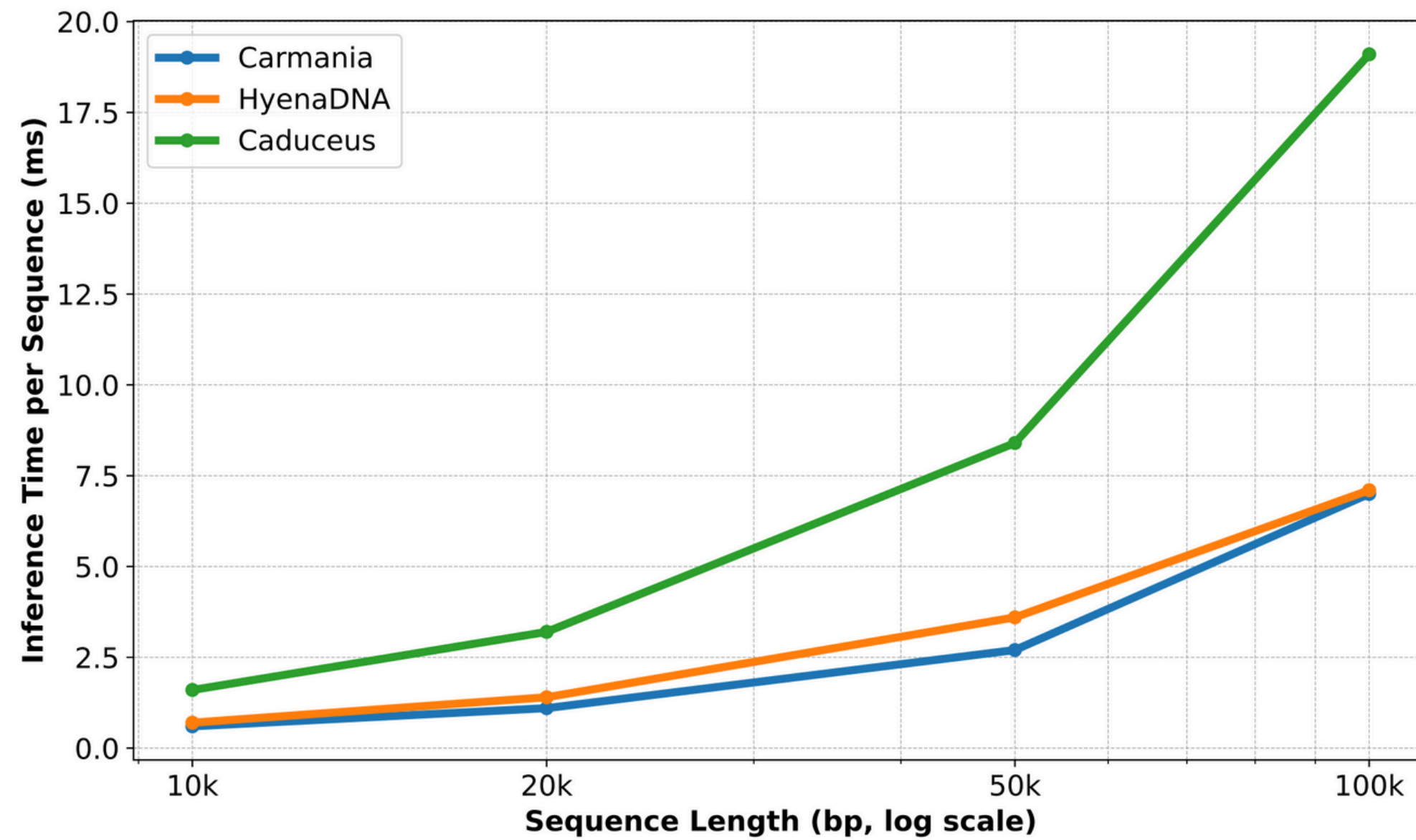
² Google Research NYC
ankitsrawat@google.com

CARMANIA: Our Long-Context Genomic LM

- CARMANIA is a 160k-context genomic Transformer language model.
- This Markovian nature of Transformers aligns with **transition-matrix objectives (TM-Loss)**.
- Enables use of **sliding-window attention** instead of full attention.
- Reduces complexity from $O(N^2) \rightarrow O(kN)$ -

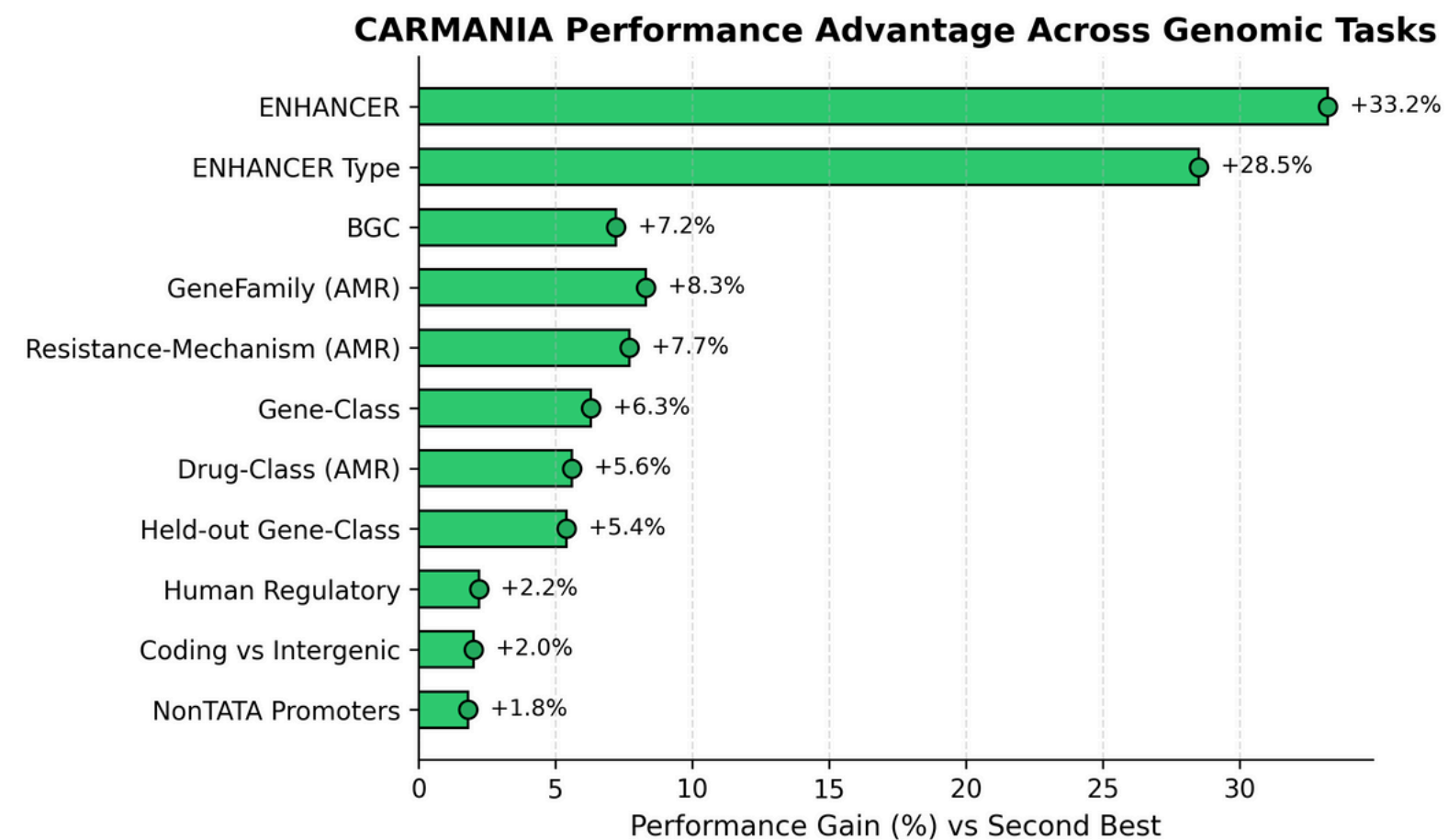


Fast Inference Time !



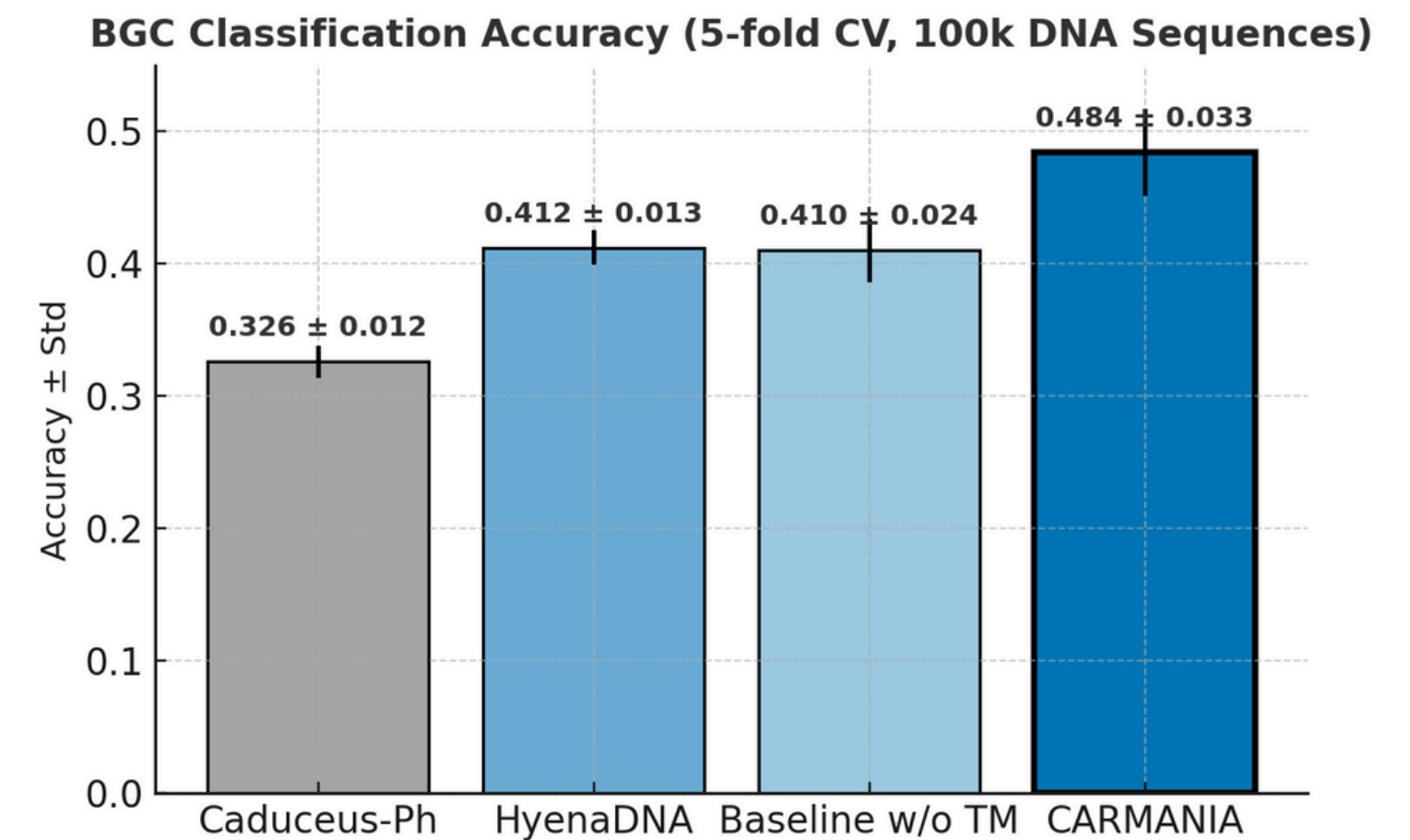
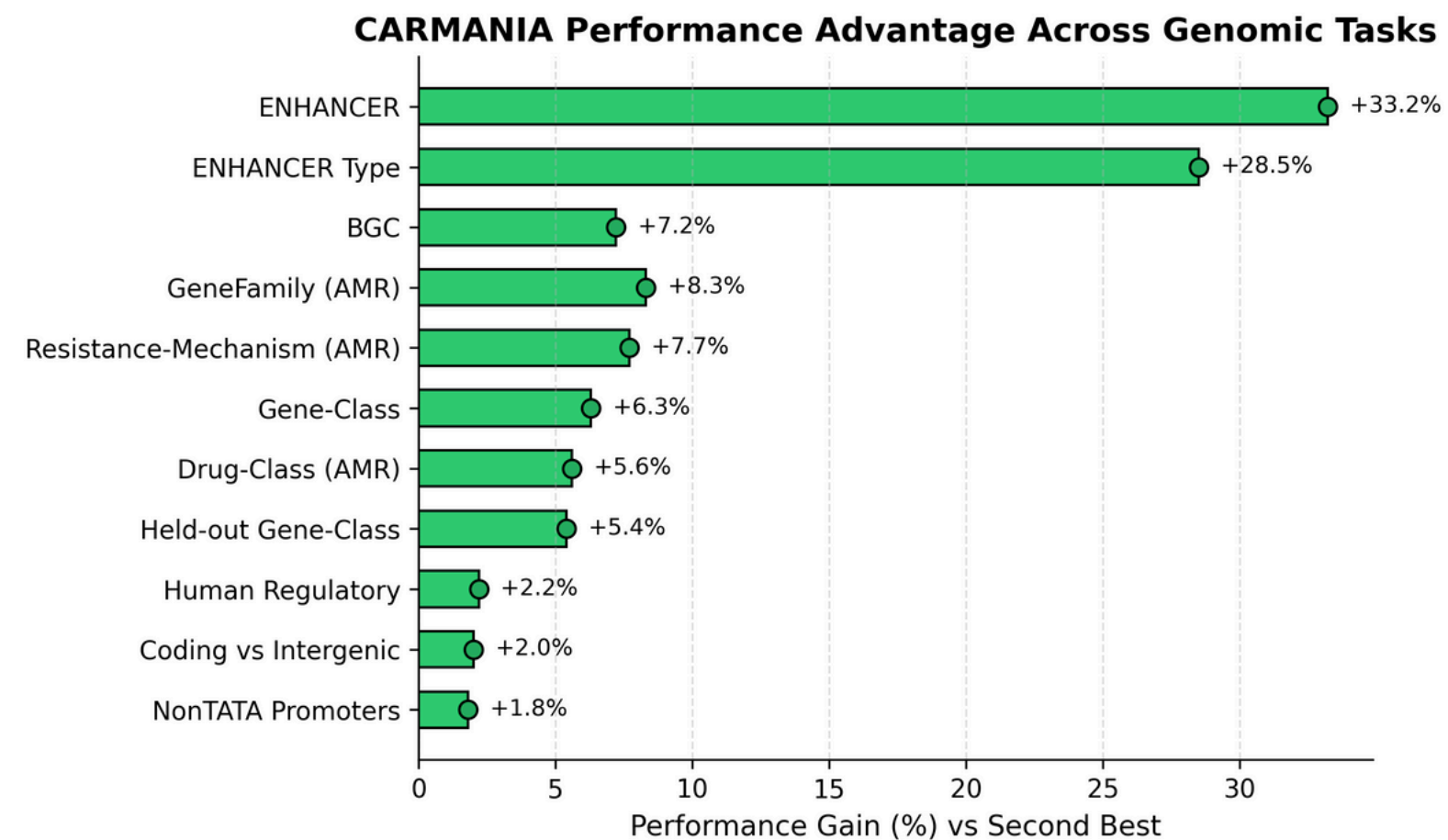
Genomic Task Performance Gains

- **Scope:** Compared CARMANIA with top models on 40 genomic tasks.
- **Wins:** Outperformed competitors on 20 out of 40 tasks.
- **Gains:** Achieved up to 33% improvement, setting a new state-of-the-art benchmark across all evaluations.



Genomic Task Performance Gains

- **Scope:** Compared CARMANIA with top models on 40 genomic tasks.
- **Wins:** Outperformed competitors on 20 out of 40 tasks.
- **Gains:** Achieved up to 33% improvement, setting a new state-of-the-art benchmark across all evaluations.
- CARMANIA outperforms prior models on **BGC classification** (100,000 DNA sequences).



Summary



- **160k context: Longest-context genomic Transformer to date**
- **TM-loss: Markovian regularization improves global consistency**
- **Fast inference: 2.5× speedup on long sequences**

Thank you!



Power your pipeline with **CARMANIA's** 160k-context understanding


MsAlEhR

/carmania-160k-seqlen-human



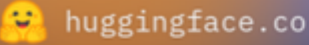
MsAlEhR/carmania-160k-seqlen-human · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.



MsAlEhR

/carmania-big-10k-prok-genome




MsAlEhR/carmania-big-10k-prok-genome · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.



EESI/carmania

Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide Analysis [NeurIPS2025]




1Contributor

0Issues


4Stars

1Fork



EESI/carmania: Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide...

Context-Aware Regularization with Markovian Integration for Attention-Based Nucleotide Analysis [NeurIPS2025] - EESI/carmania

 GitHub