InstaDeep™

Leveraging State Space Models for Long Range Genomics

LMRL ICLR 2025

Anirudha Ramesh*, Matvei Popov*, Aymen Kallala*, Rick Gentry, Shivesh Khaitan, Narimane Hennouni, Alain-Sam Cohen

Problem & Motivation: The Long Range Genomics Challenge

Processing genomes requires **LONG** contexts while maintaining **single-nucleotide** resolution,

- ➔ 🧬 **Genomic input is massive —** Human Genome is **~3Gbp** long!
- ➔ 🔍 Key signals rely on **single-base precision** across long distances**.**

Existing SoTA are,

- ➔ Transformer based, **Nucleotide Transformer:** quadratic attention limitation **(12kbp)** & rely on explicit positional encodings.
- ➔ And often sacrifice on single nucleotide resolution, **DNABert**: k-mer tokenization

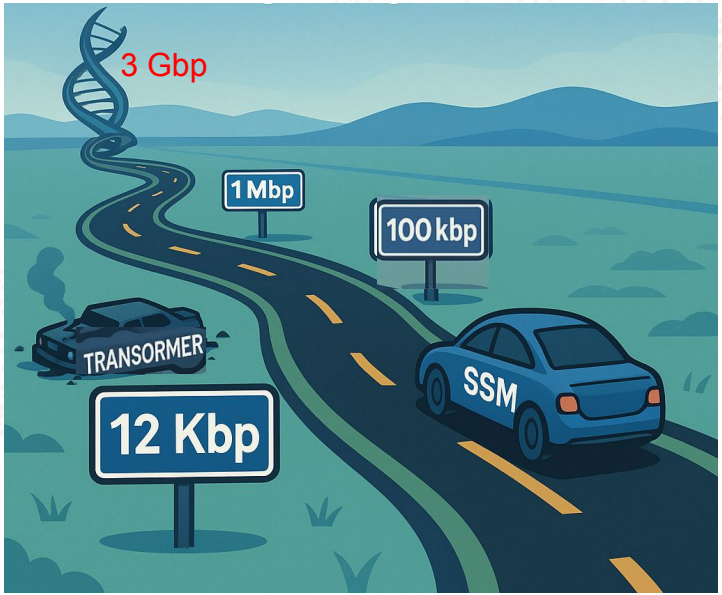
Ji et. all, 2020, Effective gene expression prediction from sequence by integrating long-range interactions
Dalla-Tore et. all, 2024, Nucleotide Transformer: building and evaluating robust foundation models for human genomics

SSMs: A Scalable Alternative To Transformers?

SSMs,

➔ Scale linearly with sequence length

➔ No rigid explicit position encodings



Could it be?

Source: ChatGPT

Objective: Answer 3 Key Questions

Can SSMs,

➔ Match transformers across genomics tasks?

➔ Zero-shot extrapolate well beyond training sequence lengths?

➔ Enable democratic access to ultra-long inference?

Experimental Setup

🧬 Datasets & Setup

📚 Pre-training

50M parameter **NTv2 (Transformer)**, **Caduceus & Hawk (SSMs)** on **InstaDeep's Multi-Species Dataset (MSD)**

🧪 Fine-Tuning & Evaluation

On **Genomics Long-Range Benchmark (GLRB)**

🚀 Zero-Shot Extrapolation

Tested directly on longer sequences from **GLRB**

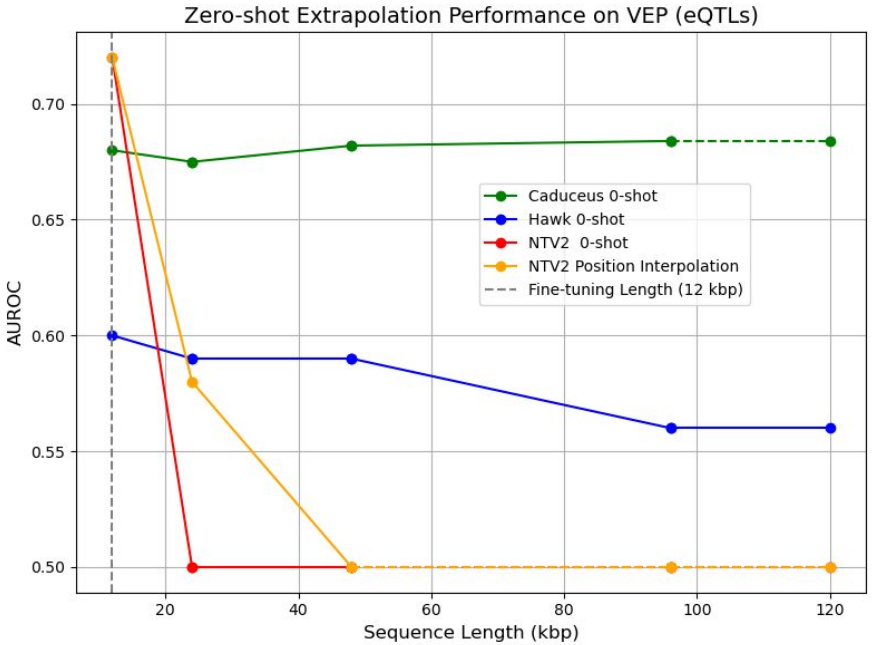
Dalla-Tore et. all, 2024, Nucleotide Transformer: building and evaluating robust foundation models for human genomics
Schiff et. all, 2024, Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling
De et. all, 2024, Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models
Trop et. all, 2024, The Genomics Long-Range Benchmark: Advancing DNA Language Models

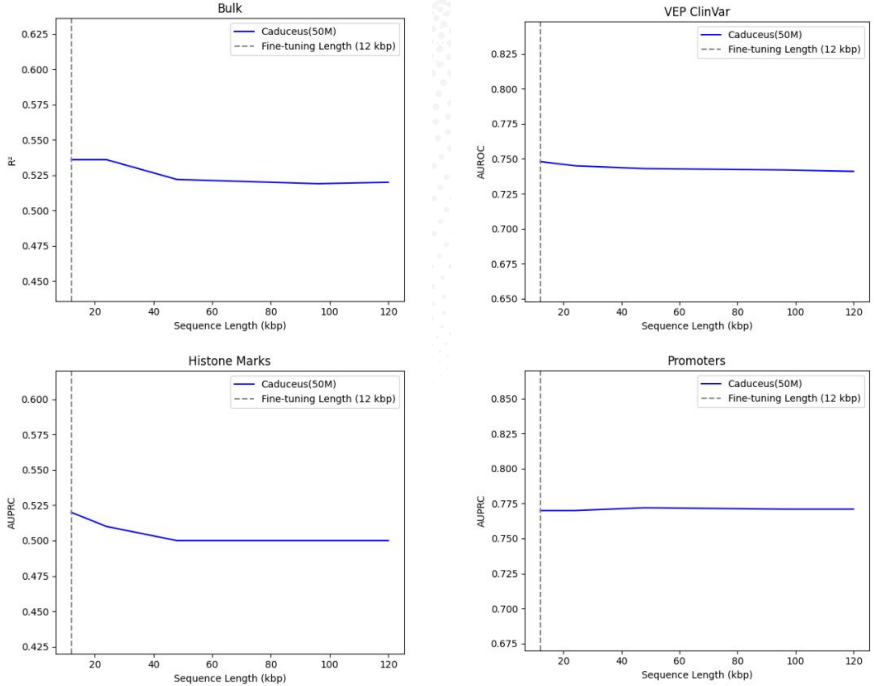
Results: SSMs Can Perform Comparably Across Multiple Tasks!

Task	NTv2	Caduceus
Bulk RNA (R^2)	0.52	**0.53**
VEP eQTL (AUROC)	**0.72**	0.68
VEP ClinVar (AUROC)	**0.75**	**0.75**
Histone Marks (AUPRC)	0.34	**0.52**
Promoters (AUPRC)	0.75	**0.77**
Enhancers (AUROC)	**0.78**	0.75

Caduceus can match or outperform NTv2 on GLRB

Results: SSMs Show Remarkable Zero-Shot Extrapolation Across Tasks!





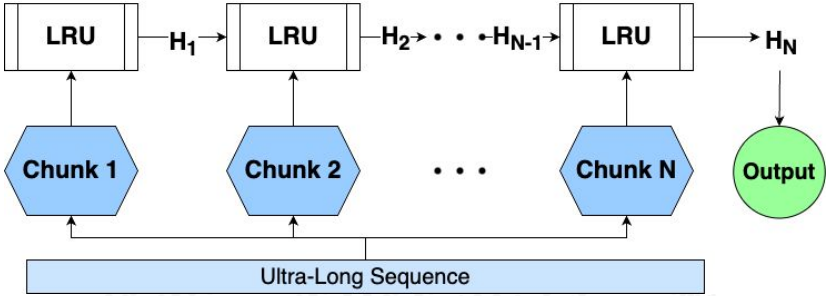
From 12Kbp to 120Kbp Without Retraining

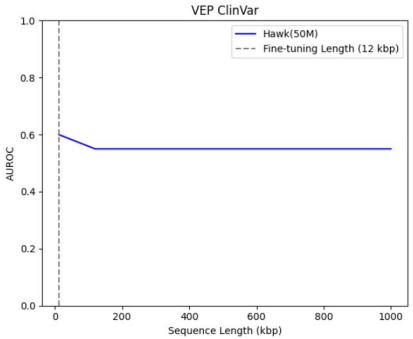
SSMs generalize, Transformers collapse

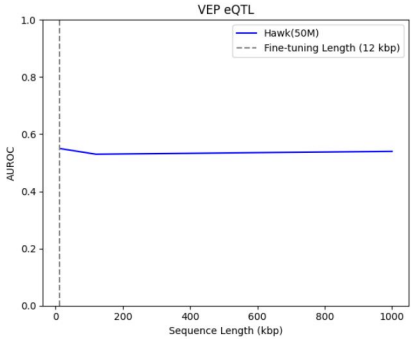
Enables scaling beyond training during inference!

InstaDeep™

Results: SSMs Enable Efficient 1 Million Base-Pair Inference!







💡 **Single-GPU, Ultra-Long Sequence Inference**

Chunk-wise hidden state transfer + parallel scan within chunks

→ Enables linear-time inference w reduced constant v/s full linear scan

→ Scales efficiently to 1Mbp+ on a single A100 GPU (40GB)

🚀 Efficient, scalable, and democratized — long-range genomics without quadratic cost!

Conclusions, and Looking Ahead

SSMs: A New Frontier for Genomics Modeling.

Effective, **scalable**, **democratic** and ready for whole-genome applications!

Next Steps

➔ Better utilize longer contexts available during inference, to actively improve in tasks which could benefit from it.