Leveraging State Space Models in Long-Range Genomics

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

InstaDeep

* Equal contribution † Correspondence: a.ramesh@instadeep.com

Motivation

- Regulatory interactions in the genome often span hundreds of Kbp to Mbp. Human Genome is ∼3 Billion Base Pairs long!
- Existing models mostly operate on the scale of tens of thousands of base pairs.
  - Transformer models are constrained by quadratic attention, limiting practical context to ∼ 10kbp.
  - Training directly on $>$ 100 kbp sequences is computationally prohibitive for most labs, and data available is limited.
- **Goal**: Find a way to process longer sequences effectively, without needing to train on longer sequences.

State Space Machines (SSMs) for Genomics

- Linear complexity vs. quadratic for transformers.
- Implicit positional encoding through hidden states.

This would allow processing longer genomic sequences, without finetuning, as positional embeddings don't go out of distribution!

Experiments

Models & Training. We pre-train 3 models with **50 M params** on masked-language-modeling task using **300 B nucleotides** from InstaDeep's multi-species dataset.

- **NTv2** – 12-layer Transformer (current GLRB SOTA)
- **Caduceus** – bi-directional Mamba SSM (+ RC-equivariance)
- **Hawk** – Linear Recurrence Unit (LRU) SSM (+ added bi-directionality)

We finetune and evaluate on the **Genomics Long-Range Benchmark (GLRB)**. We find that SSMs,

- **Perform comparably to transformer baselines across multiple tasks. → Fig-1.**
- **Can 0-shot extrapolate 10-100× its training context length (12Kbp to 120Kbp) across multiple tasks, sustaining performance, compared to NTv2 which collapses. → Fig-2, 3, 5.**
- **Can process ultralong sequences (1 Mbp) on a single A100 GPU, aided by a hidden-state transfer mechanism. → Fig-4, 5.**

Fig-1: GLRB Tasks Results (12 kbp)

Task	NTv2	Caduceus	Hawk
Bulk RNA (R2)	0.52	**0.53**	-
VEP eQTL	**0.72**	0.68	0.60
VEP ClinVar	**0.75**	**0.75**	0.55
Histone Marks	0.34	**0.52**	-
Promoters	0.75	**0.77**	-
Enhancers	**0.78**	0.75	-

SSMs perform comparably to SoTA transformers across a range of genomics tasks!

Fig-2: Zero-Shot Extrapolation (12 to 120 Kbp)

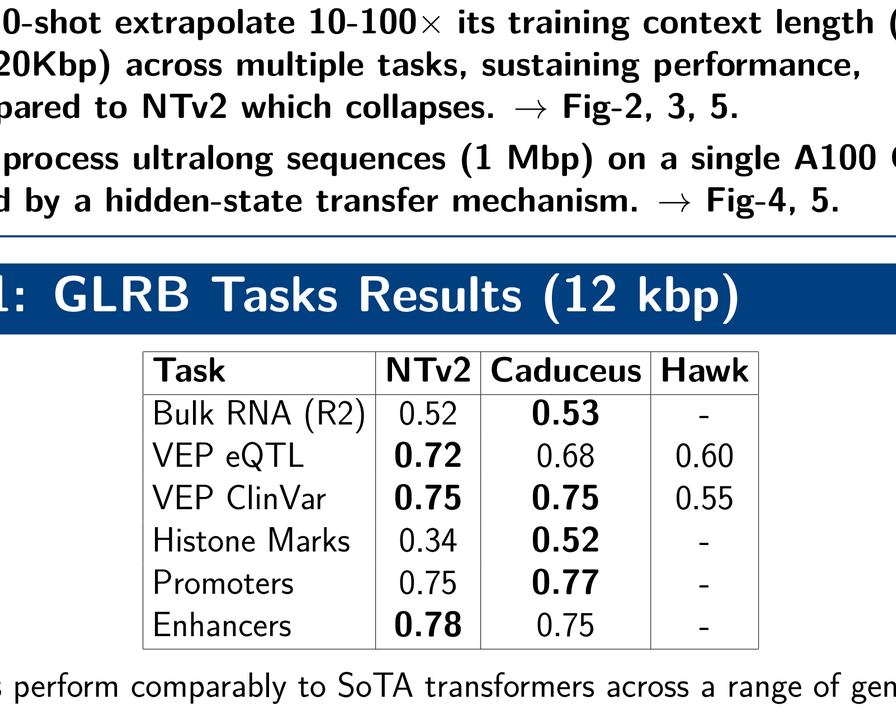


Figure: AUROC on VEP eQTLs. Dotted line = 12 kbp training length. SSMs maintain performance with increase in inference sequence length, transformers collapse owing to distribution shift in positional embeddings.

Fig-3: Zero-Shot Extrapolation (12 to 120 Kbp)

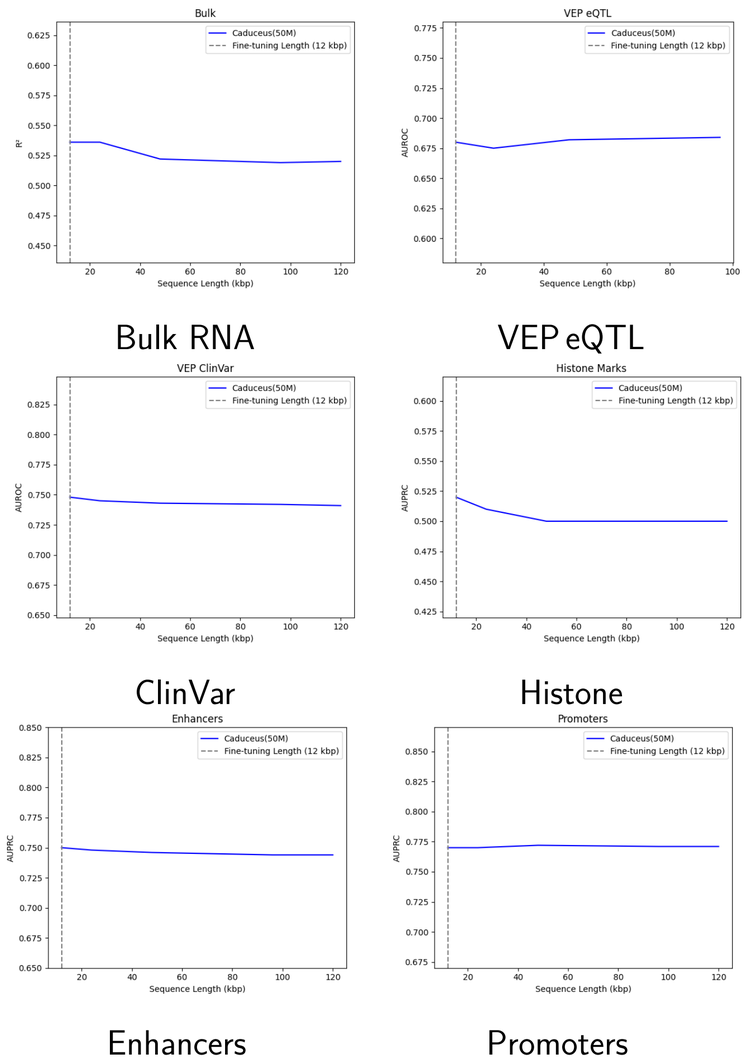


Figure: Zero-shot extrapolation up to 120 kbp across multiple GLRB tasks. Zero-shot extrapolation capabilities are true across a range of tasks.

Fig-4: Processing UltraLong Sequences

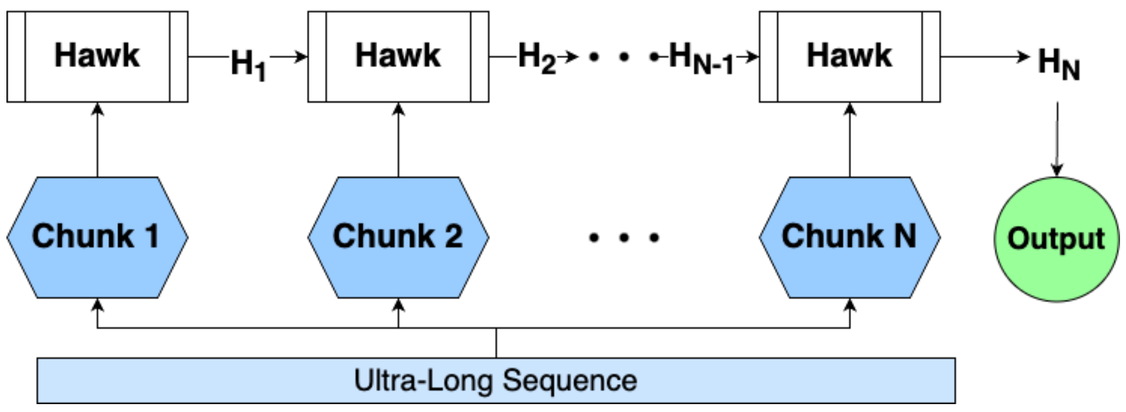
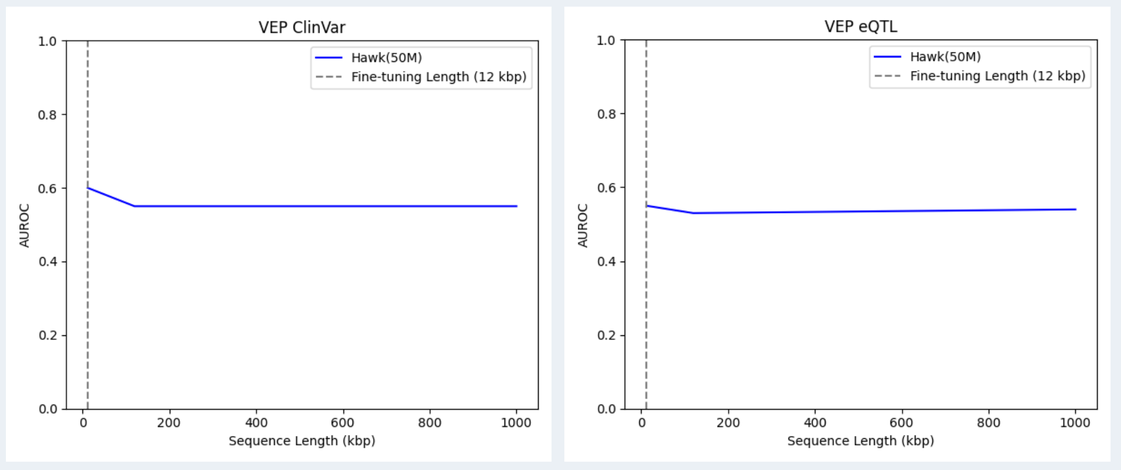


Figure: Hidden-state transfer enables chunk-wise inference while preserving global context. We perform a linear-scan across our chunks, while performing a parallel-scan within each chunk. This enables efficient processing in memory constrained environments across even ultra-long sequences, allowing >1Mbp .

Fig-5: Zero-Shot Extrapolation to 1 Mbp



ClinVar eQTL

Figure: SSMs can zero-shot extrapolate to input sequences of length 1 Mbp, despite training on 12 Kbp ($100\times$ input sequence length)!

Implications & Future Work

- SSMs unlock single-pass inference over entire genomic regions (> 1 Mbp).
- Commodity 40–80 GB GPUs suffice, democratizing long-range genomics.
- **Next steps:**
  - Utilize longer contexts available during inference more effectively.
- SSMs can serve as a foundation for large-scale, democratized comprehensive genome-scale modeling.