ZERO-SHOT EXTRAPOLATION IN STATE-SPACE MODELS FOR LONG-RANGE GENOMICS

Matvei Popov* **Aymen Kallala*** **Anirudha Ramesh***

InstaDeep

ABSTRACT

Long-range dependencies are crucial for interpreting genomic structure and function, yet conventional transformer-based genomics models often fail to generalize beyond their training window even when employing sophisticated positional embeddings. We show that State-Space Models (SSMs) can zero-shot extrapolate two orders of magnitude beyond their original context length, thus capturing distal regulatory interactions required for gene expressions without specialized fine-tuning. With our hidden-state transfer mechanism, we can efficiently process ultralong genomic sequences (1Mbp) on a single GPU—providing a scalable, generalizable, and resource-efficient alternative to transformers.

1 INTRODUCTION

Modeling long-range interactions is crucial for genomic tasks which require tens or even hundreds of kilobases (kbp) of context (Nguyen et al., 2023; Schiff et al., 2024; Kao et al., 2024; InstaDeep.). For instance, studying structural variants or mapping distal enhancers to their target genes often demands analyzing hundreds of kilobases or more in a single pass. Training on such ultralong genomic data is very expensive and prohibitive. Training on shorter sequences (thereby reducing compute costs) and leveraging partial-genome datasets, models can subsequently extrapolate to full-length contexts without extensive retraining or specialized hardware, substantially broadening accessibility for research labs with limited computational resources.

Although transformers excel at capturing local sequence patterns (Ji et al., 2021; Dalla-Torre et al., 2024), they struggle to maintain both ultralong context coverage and high positional resolution when the input length far exceeds their training window. Their quadratic scaling with the sequence length serves as another obstacle. These limitations are problematic when subtle differences in distant nucleotides have functional significance. While specialized positional embeddings like sinusoidal (Vaswani, 2017), RoPE (Su et al., 2023), YaRN (Peng et al., 2023), and SelfExtend (Jin et al., 2024) improve extrapolation, many still require fine-tuning or can lead to a loss of positional resolution. By contrast, State-Space Models (SSMs) maintain linear-time complexity with respect to sequence length, enabling them to scale to ultralong contexts without incurring prohibitive memory or computational costs (Gu et al., 2022). In this paper, we demonstrate that SSM-based architectures can zero-shot extrapolate 100x beyond their training window on representative genomic tasks, preserving single-nucleotide resolution. We also introduce a hidden-state transfer mechanism that allows models to process sequences of up to 1 million base pairs (1Mbp) on a single NVIDIA A100 GPU—highlighting how SSMs offer a more practical solution for real-world genomic studies compared to attention-based methods.

2 BACKGROUND AND RELATED WORK

Transformers use positional encodings to inject sequence order information. Classic sinusoidal embeddings (Vaswani, 2017) are fixed functions of position, allowing relative offsets to be inferred. However out-of-distribution input lengths produce embeddings leading to phase mismatches and performance degradation. RoPE (Su et al., 2023) refines sinusoidal encodings by rotating queries

*equal contribution; corresponding author - a.ramesh@instadeep.com / anirudha.ramesh50@gmail.com

and keys according to position, enabling more flexible relative attention. Extending RoPE to much larger contexts still often requires re-tuning the rotational frequencies or partial fine-tuning. Further extrapolation strategies include YaRN (Peng et al., 2023), which adapts RoPE-based models by additional training at larger context windows, and SelfExtend (Jin et al., 2024), which modifies the attention mechanism during inference to handle extended contexts, but may lose local positional details for ultralong sequences due to grouping positions to keep the embeddings in-distribution.

Training on ultralong genomic sequences necessitates a large GPU vRAM which can be expensive. Chunking helps work around this by managing memory footprint, either by selectively processing portions of the input or by processing the input across multiple steps. Methods like SimCAS (Xie et al., 2024) and CItruS employ selective processing of portions of the sequence (Bai et al., 2024), while Recurrent Chunking Mechanisms (Gong et al., 2020) processes the input across multiple steps employing the hidden state propagation between segments, akin to our method.

State-Space Models view sequence modeling as a learned linear recurrence with convolution-like operations (Gu et al., 2022), granting them $\mathcal{O}(n)$ complexity and long-range coverage. Recent SSM variants, such as the "Caduceus" and "Hawk" architectures (Schiff et al., 2024; De et al., 2024), demonstrate strong performance on benchmarks ranging from standard text to genomic tasks. Crucially, SSMs bypass the need for positional embeddings in favor of hidden-state propagation. This removes the need for out-of-distribution positional embedding corrections, thus allowing better zero-shot extrapolation to contexts far exceeding their training length.

3 Experiments

3.1 Zero-Shot Extrapolation

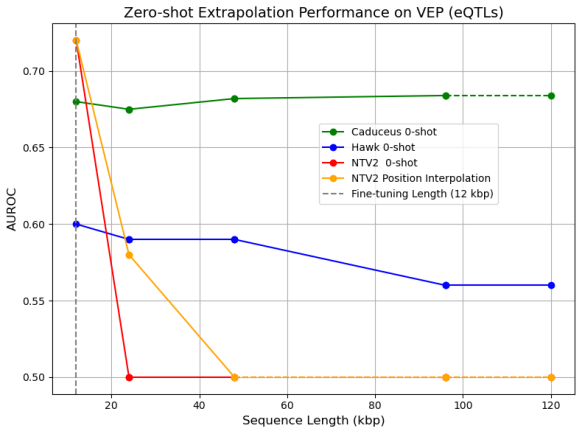

Figure 1: Zero-shot extrapolation on VEP eQTL (AUROC). All models trained at 12 kbp. Transformers (NTv2) collapse at 24 kbp+ lengths, even with position interpolation. SSMs (Caduceus, Hawk) remain stable up to 120 kbp. Dotted lines indicate unmeasured values, extrapolated from trends.

We assess zero-shot extrapolation by testing on downstream tasks at sequence lengths up to two orders of magnitude beyond the pretraining window. We evaluate our models on two tasks from the Genomics Long-Range Benchmark (Kao et al., 2024; InstaDeep.), Variant Effect Prediction (VEP) on eQTL and ClinVar. We use SSM-based models (e.g. Caduceus, Hawk), as well as Nucleotide Transformer baseline Dalla-Torre et al. (2024) for which we employ (i) RoPE (Su et al., 2024), or (ii) RoPE with Positional Interpolation (Chen et al., 2023) for better extrapolation. Each model is designed to be 50M parameters and pretrained on multi-species genomic dataset on 300 billion nucleotides as demonstrated in Dalla-Torre et al. (2024). All models are pretrained on 12 kbp and their scores on different sequence lengths from the same testing dataset are recoded. Figure 1 shows that while attention-based models degrade sharply beyond 24 kbp, SSMs remain stable up to 120 kbp without additional fine-tuning.

3.2 HIDDEN-STATE TRANSFER FOR 1MBP SEQUENCES

We also push SSMs to process sequences up to 1Mbp. To address memory constraints on a single GPU, we divide the input into manageable chunks (e.g. 100 kbp each). After processing one chunk, the model's final hidden state is passed forward as the initial state for the next chunk, ensuring continuity across the entire 1Mbp. This preserves global context while remaining memory-friendly, effectively emulating a full forward pass on a hypothetical large-memory device (Figure 2). Our implementation is based on Hawk's linear scan (De et al., 2024).

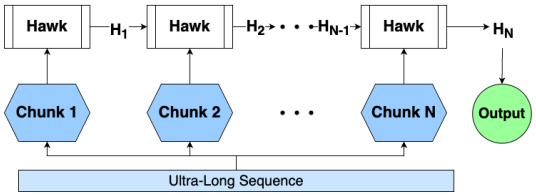

Figure 2: **Hidden-State Transfer in SSMs.** The ultralong sequence is split into chunks over which we do a linear scan. The final hidden state of a chunk initializes the next, preserving global context while limiting GPU memory usage.

Figure 3 demonstrates Hawk SSM's stable performance while tested on 1Mbp sequences on VEP ClinVar and eQTL tasks, while being pretrained on 12 kbp sequences. This demonstrates that our approach efficiently and reliably extrapolates to ultralong sequences on a single GPU.

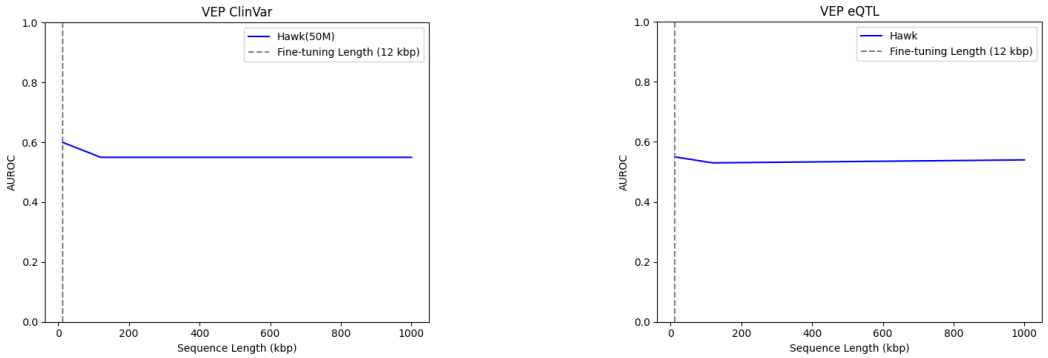

Figure 3: Hawk SSM zero-shot extrapolation to 1 Mbp on VEP ClinVar (left) and eQTL (right).

4 DISCUSSION AND CONCLUSION

Our findings demonstrate that SSMs are capable of reliably zero-shot extrapolating two orders of magnitude beyond their training window, maintaining stable performance on ultralong genomic sequences (up to 1Mbp on a single GPU), whereas standard transformers degrade without additional fine-tuning or complex inference modifications, and even advanced extensions like RoPE and YaRN fail to match this zero-shot scalability.

Looking ahead, SSMs hold promise for genomic analyses spanning hundreds of kilobases. An important observation from our work is that although SSMs can reliably extrapolate to longer genomic sequences, their performance on relevant tasks does not improve with the additional contextual information. Inspecting this phenomena could give us insights into utilizing longer contexts more thoroughly, thereby potentially improving their performance on all downstream tasks. Another direction is to expand these approaches to additional tasks like single-cell multi-omics or alternative splicing prediction. Our hidden-state transfer is implemented in a straightforward linear scanning manner for simplicity, future work could explore parallel and other more advanced scans to further enhance speed and scalability. This line of research could further establish SSMs as practical, resource-efficient architectures for capturing distant regulatory interactions across the genome.

URM STATEMENT

We affirm that at least one key author of this work meets the URM criteria of ICLR 2025 AI4NA Tiny Papers Track.

REFERENCES

Yu Bai, Xiyuan Zou, Heyan Huang, Sanxing Chen, Marc-Antoine Rondeau, Yang Gao, and Jackie Chi Kit Cheung. Citrus: Chunked instruction-aware state eviction for long sequence modeling, 2024. URL https://arxiv.org/abs/2406.12018.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024. Available at https://arxiv.org/abs/2402.19427.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. Recurrent chunking mechanisms for long-text machine reading comprehension, 2020. URL https://arxiv.org/abs/2005.08056.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

InstaDeep. Instadeepai/genomics-long-range-benchmark · datasets at hugging face. URL https://huggingface.co/datasets/InstaDeepAI/genomics-long-range-benchmark.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning, 2024. URL https://arxiv.org/abs/2401.01325.

Chia Hsiang Kao, Evan Trop, McKinley Polen, Yair Schiff, Bernardo P de Almeida, Aaron Gokaslan, Thomas Pierrot, and Volodymyr Kuleshov. Advancing dna language models: The genomics long-range benchmark. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution, 2023. URL https://arxiv.org/abs/2306.15794.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. URL https://arxiv.org/abs/2309.00071.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling, 2024. Available at https://arxiv.org/abs/2403.03234.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jiawen Xie, Pengyu Cheng, Xiao Liang, Yong Dai, and Nan Du. Chunk, align, select: A simple long-sequence processing method for transformers, 2024. URL https://arxiv.org/abs/2308.13191.