SKIP-ROPE: UNDERSTANDING LONG-RANGE GENOMIC INTERACTIONS USING SKIP-WISE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-read sequencing generates genomic data at scale through parallel processing of millions of DNA fragments, typically spanning hundreds of base pairs, whereas underlying regulatory interactions can span millions. Genomic language models (gLMs) often bridge the gap by training on long contiguous inputs assembled from short reads, but this incurs significant data generation and computational cost. We propose a modification of Rotary Position Embeddings (RoPE), Skip-RoPE, that encodes genomic distances between packed fragments by using their aligned genomic coordinates to skip missing bases. This biologically-informed decoupling of context window size and genomic distance lets a 100M-parameter model trained to an effective context length of 1 Mb using 60% fewer FLOPs than full-length training. Skip-RoPE also supports evaluation on discontinuous sequence inputs. We evaluate a 1 Mb model by predicting enhancer-gene interactions in a CRISPRi dataset through trained classifiers on frozen embeddings. Using Skip-RoPE, we achieve comparable results to full-context evaluation while requiring up to $2.1\times$ fewer input tokens. Apart from training and inference efficiency gains, we believe that this will enable a new class of gLMs that directly target raw short-read data.

1 INTRODUCTION

While much focus is put on protein-coding genomic regions, which in humans is less than 2% of the genome, roughly 93% of disease-associated variants are non-coding (Maurano et al., 2012), and are often located far from the genes they impact. Many fall in enhancers: regulatory elements that activate a target gene by looping to its promoter. The target is not always the nearest gene; only around 7% of looping interactions connect an element to its closest transcription start site (TSS) (Sanyal et al., 2012), and enhancer–promoter pairs can be separated by hundreds of kilobases to over a megabase (Smemo et al., 2014; Claussnitzer et al., 2015). In some cases, editing a single distal base can reverse diseased phenotype to a healthy state, and offer a pathway to therapeutic intervention (Claussnitzer et al., 2015).

In recent years, sequenced-based deep learning models have become a popular tool for understanding DNA variants. However, genomic language models (gLMs) are assumed to learn only those regulatory dependencies that fit within their training context windows. Supervised sequence-to-activity models have extended context accordingly, from Enformer's 200 kb (Avsec et al., 2021) to AlphaGenome's 1 Mb at base-pair resolution (Avsec et al., 2026). Autoregressive DNA models have done the same; HyenaDNA showed that single-nucleotide modeling at 1 Mb was feasible (Nguyen et al., 2023), and Evo 2 showed that data-scaling could allow sequence models to predict non-coding pathogenic variants without fine-tuning (Brixi et al., 2026).

However, the nature of genomics introduces two obstacles towards training for long context understanding. The first is data: most sequencing assays produce short reads (Illumina reads are typically only 50-300 bp (Goodwin et al., 2016)), and although these can be aligned to a reference and tiled into long windows, functional assays cover only a fraction of the genome, so many such windows are mostly empty. For example, ATAC-seq (Buenrostro et al., 2015) reads predominantly concentrate in candidate cis-regulatory elements, covering about 21% of the human genome (Moore et al., 2026), whereas RNA-seq (Mortazavi et al., 2008) reads concentrate on exons, at about 3% (ENCODE Project Consortium, 2012) of the whole genome. The second is simply computational; full

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

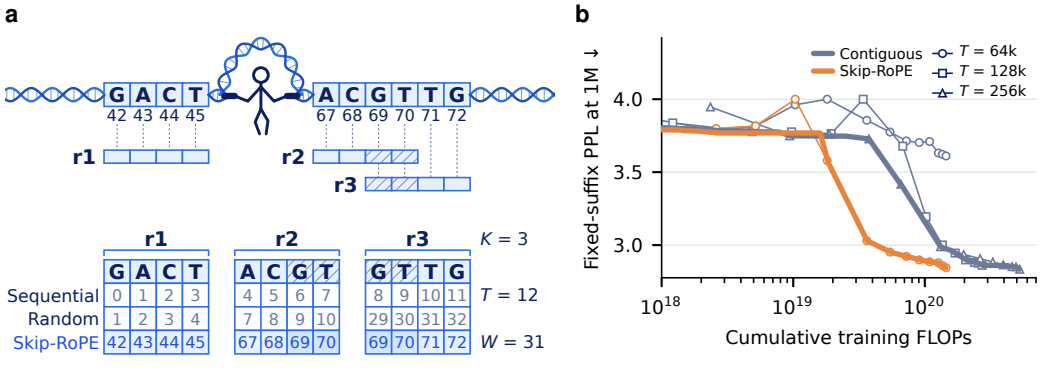


Figure 1: Overview of Skip-RoPE. **a**, Sequencing technologies generate data that resembles short fragments drawn from DNA. We simulate this by uniformly sampling K discontinuous reads from a window of size W of the human reference genome. A context window of length T is constructed from these packed reads. Existing options for assigning token positions include sequential indices and random offsets between contiguous chunks. Skip-RoPE assigns genomic coordinates. **b**, The Pareto frontier for fixed-suffix perplexity at 1 Mb of preceding context constructed against training FLOPs for Skip-RoPE (orange) at $T = 64$k, against a range of models trained on longer, contiguous sequences (gray).

attention scales quadratically with sequence length (Vaswani et al., 2017), incurring large pretraining costs.

In this paper, we propose a modification of rotary position embeddings (RoPE) (Su et al., 2024), Skip-RoPE, that lets a transformer train directly on short reads without requiring assembly into contiguous windows, while retaining the benefits of long-context training on long-range tasks.

There are 3 aspects to this paper's contributions, illustrated in Figure 1:

1. We propose Skip-RoPE, which uses aligned genomic co-ordinates to encode distance between packed fragments, decoupling context window from genomic distance.
2. Skip-RoPE natively bridges the gap between the scale of short-read data (hundreds of base pairs) and regulatory interactions (millions).
3. We train a 100M-parameter model using Skip-RoPE to an effective 1 Mb context with 60% fewer FLOPs than full-length training, while achieving comparable benchmark results.

2 Related Work

In both large language models and gLMs, the need for long-range understanding has spurred new architectures and training regimes. Architecturally, many recent models replace some (Lenz et al., 2025; De et al., 2024) or all (Gu & Dao, 2024; Nguyen et al., 2023; Schiff et al., 2024; Nguyen et al., 2024) of their attention layers with sub-quadratic alternatives to full attention. However, across architectural choices, models trained purely on short sequences often do not generalize to long ones at inference time (Delétang et al., 2023), so long-range understanding has necessitated context window extension following pretraining (Xu et al., 2024; Shi et al., 2026).

One well-established class of context window extension methods is Position Interpolation (Chen et al., 2023), where a model that has been pretrained on a short context length is fine-tuned on longer contexts with positional indices interpolated down to the model's in-distribution range.

Other approaches instead artificially inflate the position indices that the model is exposed to. Ruoss et al. (2023) achieved this by sampling indices uniformly at random from a wider range; Zhu et al. (2024) improved on this by sampling random offsets between contiguous chunks. However, they claim that the augmented positional indices do not have to align with the corresponding sequence. Wu et al. (2024) demonstrates that combining positional offsets with structural biases, in this case instruction-tuning data, has positive effects. We extend this concept to genomic sequence model-

ing, which is rich in distal structure, and show that faithfully aligning position coordinates with corresponding content is important.

3 METHODOLOGY

3.1 SKIP-ROPE

A typical Transformer model (Vaswani et al., 2017) is trained on contiguous sequences of length T, with positional indices assigned to tokens in sequential order from 0 to $T-1$. We instead train on sets of discontinuous fragments and assign positional indices according to each token's genomic coordinate, and show in Section 4.2 that a match to the true genomic position is important. This requires the reads to be aligned to a reference genome (Langmead & Salzberg, 2012), which is already a standard step in sequencing pipelines, to obtain genomic coordinates. In the human genome, the value of this genomic coordinate can be up to 250 million (Genome Reference Consortium, 2013). RoPE (Su et al., 2024) encodes positional information by rotating two-dimensional subspaces of each query and key vector by an angle proportional to a positional coordinate, with a different frequency per subspace. Since rotations in the same plane compose, the query-key inner product depends on position only through the difference between the two tokens' coordinates. The essence of our method is that instead of passing sequential indices to RoPE, we pass genomic coordinates.

For genomic modeling, we prefer relative positional encodings (Shaw et al., 2018; Su et al., 2024) over absolute ones for several reasons. First, within non-reference genomes, insertions or deletions (indels) mean that variants shift bases away from their expected genomic coordinates, which themselves can differ between reference assemblies. Secondly, across multi-species corpora, absolute coordinates carry even less meaning when evolutionary differences prohibit multi-genome alignment. Finally, since a typical human genome differs from the reference at only 4–5 million sites (1000 Genomes Project Consortium, 2015), a sufficiently large gLM trained on next-token prediction could exploit absolute coordinates by indexing into a memorized reference rather than understanding the underlying genomic language. In total, we believe that not only relative distances are more robust to shifts within and across genomes, but also better reflect the underlying biology.

Putting this into practice requires only small changes to a standard decoder-only Transformer. We adopt the Qwen3 architecture (Yang et al., 2025), with the RoPE mechanism as above; and additionally promote coordinates to int64, RoPE angles to fp64, and the output head to fp32 precision.

For fair comparison, we control model size at 100M parameters and train on 30 B tokens sampled from the human genome assembly GRCh38 (hg38) (Genome Reference Consortium, 2013; Schneider et al., 2017). We use the AdamW optimizer with a peak learning rate of 10^{-3}, weight decay of 0.1, beta values $(0.9, 0.95)$, and gradient clipping of 1.0. The learning rate follows a warmup-stable-decay (WSD) learning rate schedule (Hu et al., 2024): warmup over the first 5% of steps, a constant phase, and decay to 10% of the peak over the final 10% of steps. We use a batch size of 2.1M tokens, 14.3k steps (30B tokens), 2.5h on 32 NVIDIA B300 GPUs. All models use single-nucleotide tokenization (Nguyen et al., 2023), apply reverse-complement augmentation with probability 0.5, and randomize fragment order.

3.2 DATASET PARAMETERIZATION

Theoretically, the context length T and the genomic window W from which those tokens are drawn are independent, subject to the constraint that they belong to the same chromosome or contig. T determines compute cost, while W determines genomic reach; contiguous training ties them together ($T = W$), whereas fragmented sampling decouples them.

Introducing a third variable, K, representing the number of fragments per sequence, we define a parameterization (T, W, K) of fragmented datasets. Fragments may be disjoint or overlap, as true sequenced data does at sufficient coverage. Concretely, paired-end reads from single-cell ATAC-seq correspond roughly to $(T, W, K) = (100, 1000, 2)$ (Buenrostro et al., 2015; Cusanovich et al., 2015) (see Figure 1a), whereas most gLMs train on data resembling $(8192, 8192, 1)$.

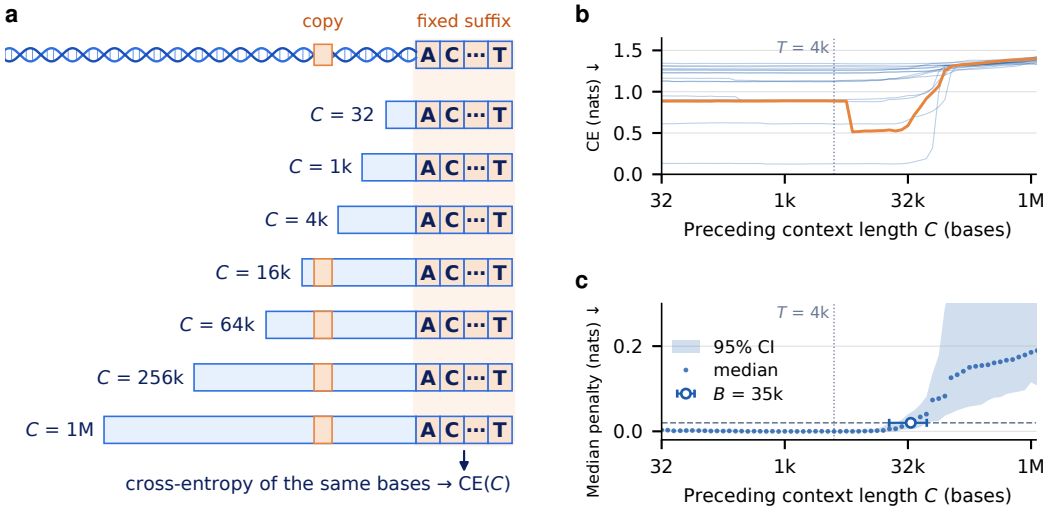


Figure 2: Measuring the onset of long-context breakdown, B, for an autoregressive model — in this case, a model trained on contiguous sequences of length $T = 4$k. **a**, Fixed-suffix evaluation: 512 bp sequences are scored in separate forward passes with C preceding bases of context (blue). Because the suffix remains constant, $\mathrm{CE}(C)$ is purely a function of added context. **b**, $\mathrm{CE}(C)$ across 16 sequences, depicting a wide range of entropies, due to some suffixes falling in repeat-rich regions. In one example (orange; hg38 chr1:149,081,056–149,081,568), CE falls by 0.37 nats around $C = 6$k, when an upstream copy of the suffix (429/512 bases identical) enters the preceding context. **c**, Median penalty with its pointwise 95% interval over resampled sequences (band). We define B as the first point beyond $C_b = 1$k for which the median penalty reaches 0.02 nats (dashed line) for two consecutive measurements.

We study the effects of training using Skip-RoPE under different data fragmentations by sweeping over a grid of (T, W, K) values. Each configuration defines a semi-synthetic dataset sampled from GRCh38, $\mathcal{G}$, according to Algorithm 1.

3.3 Break point analysis

Following Khandelwal et al. (2018), Sun et al. (2021), and Chevalier et al. (2023) we fix a 512 bp suffix and run separate forward passes with increasing preceding context lengths, C, recording the cross-entropy (CE) over the suffix tokens only (see Figure 2a).

We evaluate all models on 16 suffix sequences drawn from held-out validation chromosomes. So that sufficient upstream context is always available, we always exclude the first 4M bases of each chromosome. Additionally, we reject suffixes with more than 1% N content in the 4M upstream bases. We note that this introduces a selection bias against sequences that are subtelomeric, peri-centromeric, and next to gaps, which tend to be rich in repeats and segmental duplications (Bailey et al., 2002).

Figure 2b plots $\mathrm{CE}(C)$ against C for each locus. The high variance in absolute CE values largely reflects the underlying dynamic range of perplexities in genomic data; for example, repeat-rich regions are more predictable. We highlight a particular case (orange) where the suffix contains a motif that is repeated around 6 kb upstream. As soon as the preceding context C extends to include the upstream repeat, we see a sharp drop in CE loss.

Because absolute CE varies so widely across sequences, we measure each locus against its own baseline, defining a per-locus penalty $P_\ell(C_k) = \mathrm{CE}_\ell(C_k) - \mathrm{CE}_\ell(C_b)$, where ℓ denotes the locus, C_k the evaluated context length, and $C_b = 1{,}024$ bases. We take the median penalty across the 16 sequences and find the first $C_k > C_b$ where the median penalty reaches 0.02 nats for both C_k and C_{k+1}. The onset B is the context at which the median penalty reaches 0.02 nats, interpolated linearly in $\log C$ between C_{k-1} and C_k, and indicates breakdown at long context. For this particular model,

trained on contiguous sequences of length $T = 4$k, we measure $B = 35$ kb (95% CI 19-56 kb; Figure 2c).

3.4 Evaluating distal interactions using non-contiguous inputs

We use the results of the break-point analysis to extrapolate a parameterization that should efficiently deliver an effective context length beyond 1 Mb. We use this parameterization to pretrain a model with no held out data. We target 1 Mb because most validated enhancers act within a few hundred kilobases of their target promoters, and enhancer range is typically bounded by domains of roughly 1 Mb (Bower & Kvon, 2025; Dixon et al., 2012).

To test whether a model trained with Skip-RoPE can capture long-range regulatory relationships, we evaluate on a CRISPRi enhancer–gene (E2G) data (Gschwind et al., 2023). Each entry consists of a candidate enhancer, a paired gene, and a binary label: a pair is positive if CRISPRi knockdown of the enhancer significantly changes expression of the target gene, and negative otherwise (Figure 3). Performing well on this task requires combining the genomic distance between enhancer and promoter with sequence features that determine whether the enhancer is active and whether the promoter is responsive to it. Distance alone is already a strong baseline that many published gLMs struggle to beat.

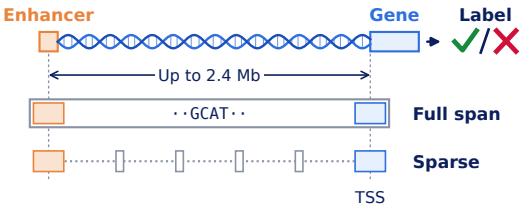


Figure 3: CRISPR interference data consists of enhancer-gene (E2G) pairs with binary labels indicating whether knocking down the enhancer has a significant effect on the gene's expression. Typically sequence models represent E2G pairs as the full span from enhancer to gene, which can reach several Mb. With Skip-RoPE, we explore sparse representations of the intervening sequence.

Conventional sequence-based models must approach this task by ingesting the entire contiguous sequence in between the enhancer and promoter, which can span up to 2.5 Mb. Skip-RoPE allows us to compress the input data by partially sampling the intervening sequence. We anchor each end of the sequence with 2,048 bp windows centered around the middle of the enhancer and the gene's TSS. We then infill the gap with N evenly-spaced 512 bp fragments, where if the gap size is less than $512 \times N$, we simply pass the full span. This allows us to measure the trade-off between compression and performance. We calculate the mean-pooled embedding across the whole input sequence at the final layer and use cross-fold validation to train linear classifiers on the binary labels.

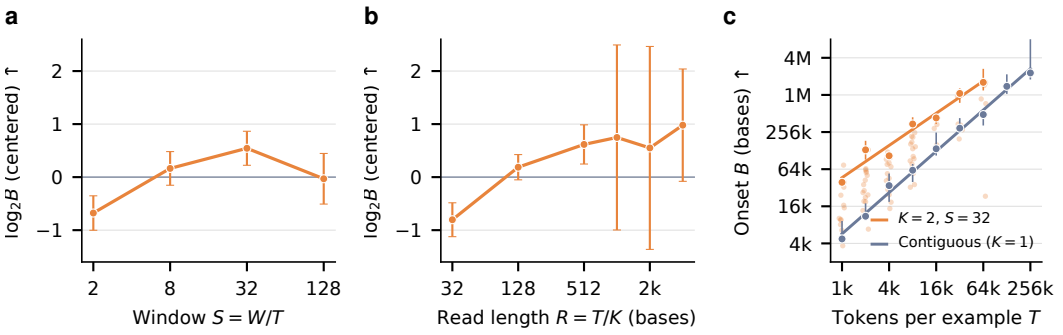


Figure 4: How training data parameterization (T, W, K) affects the long-context break point, B, across a sweep of values. **a**, Grouped by (K, T), increasing genomic window sparsity $S = W/T$ increases B up to a limit around $S = 32$. **b**, Grouped by (W, T), increasing read length $R = T/K$ increases B up to a plateau around 512-1,024 bp. **c**, Accordingly, we observe that $K = 2$, $S = 32$ parameterizations consistently lie near the frontier of long-context understanding across the sweep.

Table 1: Fine-tuning results on the revised Nucleotide Transformer benchmark. Matthews Correlation Coefficient (MCC) (%) is reported for benchmark tasks with supervised fine-tuning. Best in **bold**, second best underlined.

	NTv2				DNA-	En-	HyenaDNA		BP-	Skip-RoPE
Task	50M	100M	250M	500M	BERT-2	former	1 kb	32 kb	Net	100M
H2AFZ	50.1	49.2	51.3	49.3	49.0	**52.2**	45.5	46.7	47.3	50.2
H3K27ac	48.3	48.7	49.7	49.2	49.1	**52.0**	42.3	42.1	29.6	48.6
H3K27me3	59.5	59.5	60.0	**61.3**	59.9	55.2	54.1	55.0	54.3	60.4
H3K36me3	60.7	61.7	63.6	**64.2**	63.7	56.7	54.3	55.3	54.8	62.1
H3K4me1	47.8	48.5	49.1	49.7	49.0	**50.4**	43.0	42.3	43.6	49.2
H3K4me2	54.9	55.1	57.0	56.4	55.8	**62.6**	52.1	51.5	42.7	55.9
H3K4me3	62.7	63.3	64.0	63.6	**64.6**	63.5	59.6	60.3	44.5	62.8
H3K9ac	53.8	53.8	56.5	53.9	56.4	**59.3**	48.4	48.7	33.6	53.2
H3K9me3	43.0	44.5	46.7	47.5	44.3	45.3	37.5	41.9	29.8	**48.5**
H4K20me1	63.7	64.8	65.2	**65.8**	65.5	60.6	58.0	59.0	53.1	65.4
Enhancers	51.1	50.7	52.5	52.4	51.7	**61.4**	47.5	47.6	48.8	51.7
Enhancer types	45.9	46.5	49.2	49.8	47.6	**57.3**	44.1	44.5	44.9	47.9
Promoter all	73.1	75.3	77.4	**77.8**	75.4	74.5	69.3	69.8	69.6	72.3
Promoter no-TATA	75.6	76.6	**78.5**	77.4	76.9	76.3	72.3	72.9	71.7	74.3
Promoter TATA	74.7	82.6	87.0	87.8	78.4	79.3	64.8	66.6	84.8	**89.6**
Splice acceptors	92.2	94.7	**95.0**	94.4	83.7	74.9	81.5	80.8	85.9	93.7
Splice all	92.8	96.0	96.5	**96.7**	85.5	73.9	85.4	90.7	79.3	94.5
Splice donors	91.5	94.7	**96.7**	96.0	86.1	78.0	94.3	91.5	92.0	95.3
Mean	63.4	64.8	**66.4**	66.3	63.5	63.0	58.6	59.3	56.1	65.3

4 EXPERIMENTS

In this section, we now present results of experiments on our model; these start with training on fragmented sequences with Skip-RoPE, and conclude with evaluations of fully pretrained models.

4.1 UNDERSTANDING FRAGMENTED DATASETS

We sweep configurations of our fragment sampler (see Appendix B) and train models with Skip-RoPE on each. For analysis, we define window sparsity $S = W/T$ and read length $R = T/K$, and observe how varying each of these parameters in isolation affects the onset of long-context breakdown, B, in log-space.

Figure 4a shows that increasing sparsity directly increases effective context length, up to a peak ≈ 32 times the token budget, T. Beyond this point, further increasing S *degrades* model performance at long context. The size of fragments, $R = T/K$, does matter, but in Figure 4b we see that gains beyond $R = 512$ are diminishing.

We fit a power law of the form $B = nT^{\gamma}$ to the sweep results in Figure 4c (a straight line in log-log coordinates). We find $\gamma = 0.90(8)$ allowing us in principle to extrapolate to even greater target context lengths than evaluated in this paper.

4.2 GENOMIC BENCHMARK EVALUATION

We compare our 100M-parameter Skip-RoPE model, pretrained on 30B tokens sampled from hg38, against other gLMs on two common genomic benchmarks: (i) the revised Nucleotide Transformer (NT) benchmark (Dalla-Torre et al., 2025), which uses short sequences, and (ii) the Genomics Long-Range Benchmark (LRB) (Trop et al., 2024), designed to test long-range understanding.

The NT benchmark consists of 18 classification tasks that test a model's ability to identify regulatory features, including promoters, enhancers, and splice sites, as well as histone modifications. For model evaluation and comparison to results in Dalla-Torre et al. (2025), we fine-tune our model on

Table 2: Results on the Genomics Long-Range Benchmark. Baselines are from Trop et al. (2024), fine-tuned on each task, whereas our models (Skip-RoPE shaded) were evaluated on frozen embeddings. Average rank is taken over the tasks each model reports. Best in **bold**, second best underlined.

	Fine-tuned							Frozen (100M)		
	DNABERT		NTv2 500M		HyenaDNA		Caduceus	Skip-	Random	Sequential
Task (metric)	2	S	12 kb	96 kb	32 kb	160 kb	131 kb	RoPE	RoPE	RoPE
Causal eQTL (AUROC)	0.72	0.73	0.72	**0.74**	0.72	0.71	0.68	0.723	0.727	0.728
ClinVar (AUROC)	0.74	0.73	**0.78**	0.75	0.66	0.56	–	0.719	0.701	0.711
Bulk RNA (R^2)	0.51	0.52	**0.60**	0.56	0.41	0.46	0.52	0.561	0.546	0.550
CAGE (R^2)	–	–	**0.39**	0.36	0.22	0.19	–	0.223	0.234	0.210
Promoter (AUPRC)	0.71	0.75	**0.79**	0.78	0.56	0.67	0.75	0.781	0.776	0.775
Enhancer (AUROC)	0.81	**0.83**	0.82	0.82	0.73	0.74	–	0.827	0.827	0.825
Histone marks (AUPRC)	0.24	0.33	**0.38**	**0.38**	0.22	0.25	–	0.333	0.325	0.322
DNA accessibility (AUPRC)	0.15	0.16	**0.30**	**0.30**	0.084	0.11	–	0.162	0.163	0.158
Average rank (↓)	6.86	4.14	**2.44**	**2.44**	8.38	8.25	7.67	3.44	4.19	5.00

each task with ten-fold cross-validation, reporting the median test Matthews correlation coefficient (MCC) across folds[1]. Despite its small pretraining dataset and size, our Skip-RoPE model ranks third by mean MCC, behind only the 250M- and 500M-parameter NT-v2 models (Table 1).

The Long-Range Benchmark comprises 9 tasks across 4 categories: variant effect prediction, gene expression prediction, chromatin feature identification, and regulatory element detection. We drop OMIM, as it only supports zero-shot evaluation and models perform at chance. The LRB dataloader allows for arbitrary length input sequences around the locus of interest. To evaluate our model[2], we use the dataloader-defined defaults: 100 kb for all tasks except CAGE prediction, which uses 114,688 bp. Results are shown in Table 2, and the Skip-RoPE model has an average rank of 3.44; whereas models not trained with true genomic coordinates (random and sequential) lag behind.

4.3 Enhancer-gene pair classification

We also tested on the enhancer-gene pair prediction benchmarks. In our results, we found that our model outperforms TSS-enhancer distance-based baselines in every group, stratified by enhancer-gene distance (Table 4). AlphaGenome remains the strongest, which is expected: it is trained end-to-end on multi-omic tracks (e.g., gene expression, chromatin accessibility, and 3D genome contacts) over 1 Mb windows, whereas our model is trained on sequences alone. For a fairer comparison against other sequence-only models, we evaluate on the DNALongBench enhancer–gene benchmark (Cheng et al., 2025), where our model outperforms most existing sequence models (Table 3).

Table 3: AUPRC on DNALongBench enhancer–gene prediction.

Model	AUPRC
Distance	0.431
ABC	0.407
DNAChunker	0.468
Caduceus-Ph	0.441
HyenaDNA	0.249
NTv3 100M	0.440
GeneZip 70M	0.446
Skip-RoPE 100M	**0.597**

[1]We fine-tune using low-rank adaptation (Hu et al., 2022) of rank 8 on the attention QKV and output projections of all blocks, alongside training a linear classifier on mean-pooled final-layer embeddings. We use a learning rate of 3×10^{-4} for the adapters and 10^{-3} for the classifier on all tasks.

[2]Due to computational constraints, we do not fine-tune our model on the LRB; instead, we train probes on frozen mean-pooled embeddings.

Table 4: AUPRC on the ENCODE-rE2G K562 CRISPRi benchmark, stratified by enhancer–TSS distance. Our models were evaluated with linear probes on final-layer frozen embeddings, fitted on five folds. Sparse inputs consist of two 2,048 bp anchors and 448 512 bp fragments, 2.1× fewer tokens than the full span. Random is the true positive rate. Best in **bold**, second best underlined.

			Zero-shot		Skip-RoPE 100M	
	Random	Distance	AlphaGenome	Borzoi	Full span	Sparse
All	0.045	0.412	**0.749**	0.67	0.536	0.517
<10 kb	0.593	0.706	**0.909**	0.88	0.776	0.751
10–100 kb	0.161	0.271	**0.779**	0.69	0.508	0.496
0.1–2.5 Mb	0.010	0.035	**0.347**	0.24	0.096	0.084

We explore the effects of passing sparse inputs instead of the full span from enhancer to gene. Figure 5 plots the change in performance compared to the full span input as we increase sparsity, up to the extreme where the input is only the two 2,048 bp anchors centered around the enhancer and TSS. We see that the model is able to maintain comparable performance, within noise bands, on up to 2.1× shorter inputs.

5 Discussion

In this section, we discuss the results of our experiments and evaluate the relative advantages and disadvantages of Skip-RoPE. Our contributions divide naturally into two parts: firstly, an efficient recipe for pretraining long-context gLMs; and how this transfers to biological understanding. We also assess the limitations of this paper, and finish by discussing future work.

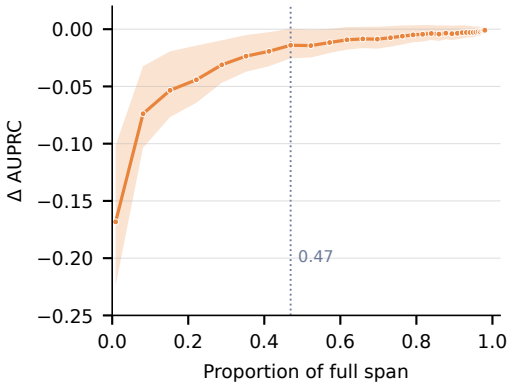


Figure 5: Comparison of sparse and full span input evaluation on the rE2G task. We plot $\Delta\text{AUPRC} = \text{AUPRC}(\text{sparse}) - \text{AUPRC}(\text{full})$ against the mean compression ratio, defined as the reciprocal of the mean $T(\text{full})/T(\text{sparse})$. Our results show sparse matches the full span performance at a compression of 0.47.

5.1 Pretraining models with Skip-RoPE & Compute Efficiency

To investigate the ability of Skip-RoPE to train models with long context understanding, we created semi-synthetic, fragmented datasets subsampled from the human genome. We demonstrated the ability to train models with 1 Mb understanding using a training context length of only 64 kb, but found that performance degrades at the extremes of either high data sparsity over the genome, or very low fragment size.

This is an informative result for constructing large-scale corpora: while performance remains robust in the $50 - 300$ bp range of Illumina sequencing read lengths, we observe small improvements up to the $512 - 1024$ bp range, suggesting that it may be worth investing resources into read stitching or sourcing data from alternative sources which provide longer reads; for example, Oxford Nanopore sequencing routinely produces reads of in the $10 - 100$ kb (Jain et al., 2018).

Crucially, we found that Skip-RoPE consistently enables model understanding at context lengths 32× greater than those trained on. Compared to full-context training, this allows us to achieve equivalent context length models at 2.5× fewer FLOPs (see Figure 6).

5.2 Performance on Genomics Benchmarks & Enhancer-Gene Classification

We further demonstrate that training on short-read-like data using Skip-RoPE is sufficient

for the model to understand distal relationships between regulatory elements, as evidenced by its performance on the CRISPR enhancer–gene dataset. Our 100M-parameter model consistently outperforms the distance-only baseline across all enhancer–gene pairs when stratified by enhancer–gene distance (Table 4), and also outperforms several sequence-based models (Table 3).

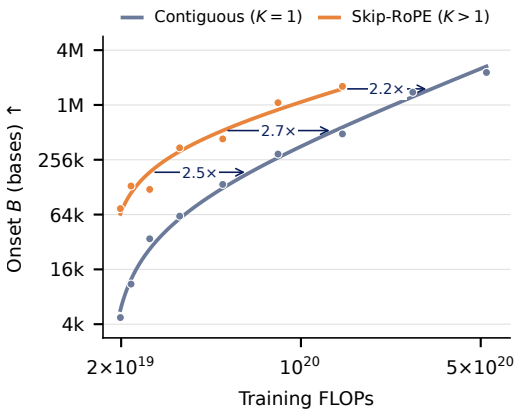


Figure 6: Skip-RoPE attains equivalent long-range understanding compared to a model trained on long contiguous sequences using $2.5\times$ fewer training FLOPs.

5.3 Limitations & Future Work

Skip-RoPE's strong performance on long contiguous sequences did not transfer seamlessly to long-range inputs represented as fragmented sequences: replacing full span inputs with enhancer–gene-anchored windows in the ENCODE rE2G task sharply degraded performance in long-range bins. We attribute this to pretraining, where the observed gaps between fragments were on average far shorter than the maximum span, so the model rarely saw the longest distances it was later asked to handle. Sparse infilling of the gap recovers much of this loss, but a mismatch between pretraining and downstream input structure remains. Finally, context extrapolation is bounded by chromosome size, since genomic coordinates do not extend across chromosomes; the model; and it cannot capture the small number of important inter-chromosomal interactions.

We see considerable promise for Skip-RoPE in future applications. First, because its position encoding is relative, Skip-RoPE extends naturally to multi-genome and multi-species datasets. It can be combined with context-extension fine-tuning to further reduce the context length required during pretraining, and because it is agnostic to the underlying Transformer architecture, it can be applied in hybrid-attention or masked language modeling settings. Second, it opens up a broad range of biological applications. The most compelling opportunity is training directly on raw sequence data paired with readouts from functional assays, such as ATAC-seq for chromatin accessibility or RNA-seq for gene expression. In summary, we expect Skip-RoPE to slot flexibly into a wide range of gLMs across these settings.

6 Conclusion

Although Skip-RoPE is a simple adjustment to RoPE, it solves a fundamental problem in genomic language modeling with an attention-native solution. We demonstrate its efficacy in the simplest case — training a small model on data sampled from the human reference genome. From our experiments, we derive recommendations for constructing data corpora, demonstrate $2.5\times$ savings in training compute, explore a performance-matched 52% compaction of test-time sequences, and achieve competitive benchmark results against purpose-trained models, despite only intending for our models to be trained as proof-of-concept.

AI use statement

In this work, we used generative AI tools to provide feedback on research methodology and experiments, refine hypotheses, to implement parts of our method, and to support qualitative data analysis. We have not used generative AI tools to generate synthetic datasets, develop theoretical models, assist in any mathematical formulation or proof construction, or provide translation.

Additionally, we used generative AI tools to draft parts of the paper, create scientific figures, identify relevant literature and format references, and create code.

All AI-assisted work has been reviewed: references have been checked against original records, AI-drafted text greatly revised by authors, and all generated code was checked by at least two authors. The research ideas themselves are our own and were not generated by AI, and all final result inter-

pretation was carried out by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

An ethics statement will be added after the double-blind review process is complete.

Reproducibility statement

All datasets used for pretraining, fine-tuning, and evaluation are publicly available. Details for downloading, processing, and sampling our data are provided in Appendix A. We intend to open source the code and weights after the double-blind review process is complete.

Author Contributions

Author contributions will be added after the double-blind review process is complete.

Acknowledgments

Acknowledgments will be added after the double-blind review process is complete.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015. doi: 10.1038/nature15393.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649:1206–1218, 2026. doi: 10.1038/s41586-025-10014-0. URL https://www.nature.com/articles/s41586-025-10014-0.

Jeffrey A. Bailey, Zhiping Gu, Royden A. Clark, Knut Reinert, Rhea V. Samonte, Stuart Schwartz, Mark D. Adams, Eugene W. Myers, Peter W. Li, and Evan E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002. doi: 10.1126/science.1072047.

Grace Bower and Evgeny Z. Kvon. Genetic factors mediating long-range enhancer–promoter communication in mammalian development. *Current Opinion in Genetics & Development*, 90: 102282, 2025. doi: 10.1016/j.gde.2024.102282.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9, 2015. doi: 10.1002/0471142727.mb2129s109. URL https://doi.org/10.1002/0471142727.mb2129s109.

Jonathan Casper, Matthew L. Speir, Brian J. Raney, Gerardo Perez, Luis R. Nassar, Christopher M. Lee, Angie S. Hinrichs, Jairo Navarro Gonzalez, Clay Fischer, Mark Diekhans, Hiram Clawson,

Anna Benet-Pages, Galt P. Barber, Charles J. Vaske, Marijke J. van Baren, Karen Wang, Yesenia Joanna Puga Rodriguez, Jaidan Ashlyn Jenkins-Kiefer, Megna Chalamala, David Haussler, William James Kent, and Maximilian Haeussler. The UCSC Genome Browser database: 2026 update. *Nucleic Acids Research*, 54(D1):D1331–D1335, 2026. doi: 10.1093/nar/gkaf1250.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Wenduo Cheng, Zhenqiao Song, Yang Zhang, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, and Jian Ma. Dnalongbench: a benchmark suite for long-range dna prediction tasks. *Nature communications*, 16(1):10108, 2025.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=kp1U6wBPXq.

Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015. doi: 10.1056/NEJMoa1502214.

Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015. doi: 10.1126/science.aab1601.

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, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *International Conference on Learning Representations*, 2023.

Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. doi: 10.1038/nature11082.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

Genome Reference Consortium. Homo sapiens (human) genome assembly GRCh38 (hg38). NCBI Assembly database, 2013. URL https://www.ncbi.nlm.nih.gov/assembly/88331. Accession: GCA_000001405.15 / GCF_000001405.26.

Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. doi: 10.1038/nrg.2016.49.

Andreas R. Gschwind, Kristy S. Mualim, Alessandra Alessandrini, Rebecca Bergstrom, Ziyi Wang, Dirk Vandebergh, et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv*, 2023. doi: 10.1101/2023.11.09.563812.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *Conference on Language Modeling*, 2024.

Miten Jain, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, Andrew D. Beggs, Alexander T. Dilthey, Ian T. Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, 2018. doi: 10.1038/nbt.4060.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 284–294, 2018. doi: 10.18653/v1/P18-1027.

Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012. doi: 10.1038/nmeth.1923.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JFPaD7lpBD.

Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012. doi: 10.1126/science.1222794.

Jill Moore, Henry Pratt, Kaili Fan, Nishigandha Phalke, Jonathan Fisher, Shaimae Elhajjajy, Gregory Andrews, Mingshi Gao, Nicole Shedd, Yu Fu, Matthew Lacadie, Jair Meza, Mansi Khandpekar, Mohit Ganna, Eva Choudhury, Ross Swofford, Huong Phan, Christian Ramirez, Maxwell Campbell, and Zhiping Weng. An expanded registry of candidate cis-regulatory elements. *Nature*, pp. 1–10, 01 2026. doi: 10.1038/s41586-025-09909-9.

Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008. doi: 10.1038/nmeth.1226.

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 Christopher Ré. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers, 2023. URL https://arxiv.org/abs/2305.16843.

Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012. doi: 10.1038/nature11279.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning*, 2024.

Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017. doi: 10.1101/gr.213611.116.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018. doi: 10.18653/v1/N18-2074.

Jingzhe Shi, Qinwei Ma, Hongyi Liu, Hang Zhao, Jenq-Neng Hwang, and Lei Li. Intrinsic entropy of context length scaling in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vnipyA8c9V.

Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L. Credidio, Débora R. Sobreira, Nora F. Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D. Acemel, Miguel Manzanares, Andras Nagy, Nancy J. Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, and Marcelo A. Nóbrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375, 2014. doi: 10.1038/nature13138.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 807–822, 2021. doi: 10.18653/v1/2021.emnlp-main.62.

Evan Trop, Chia-Hsiang Kao, McKinley Polen, Yair Schiff, Bernardo P. de Almeida, Aaron Gokaslan, Thomas Pierrot, and Volodymyr Kuleshov. The genomics long-range benchmark: Advancing DNA language models. OpenReview, 2024. URL https://openreview.net/forum?id=8O9HLDrmtq.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wenhao Wu, Yizhong Wang, Yao Fu, Xiang Yue, Dawei Zhu, and Sujian Li. Long context alignment with short instructions and synthesized positions, 2024. URL https://arxiv.org/abs/2405.03939.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and weipeng chen. Base of roPE bounds context length. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=EiIelh2t7S.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dawei Zhu, Yutao Zeng, Li Dong, Zhixhao Zhang, Yuan Xu, Furu Wei, and Chang Su. Pose: Efficient context window extension of llms via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Z1gxuAQrA.

A Data Processing

Pretraining. GRCh38 chromosomes 1-22 (Schneider et al., 2017) from UCSC (Casper et al., 2026), converted from FASTA files and stored as per-chromosome uint8 memory maps. Sweep models hold out chr1, chr3 and chr9 for validation and chr11 as a sealed test set. The released model trains on all 22.

ENCODE-rE2G. K562 CRISPRi element–gene pairs from Gschwind et al. (2023) (`EngreitzLab/CRISPR_comparison` v1.0.0). All 10,356 pairs match the rE2G feature table (`karbalayghareh/ENCODE-E2G`, `7b4a653`) with no label disagreements. TSS coordinates are the rE2G curated promoter midpoints. To match AlphaGenome's release (Avsec et al., 2026), we drop three negatives whose genes are absent from GENCODE v46, leaving 10,353 pairs (471 positives). We keep the original five folds and purge each fold's training set of test genes and of sequences overlapping test windows. Sequences come from the pretraining hg38 release plus chrX (MD5-checked), used only for evaluation. This is a standard approach in genomic language modeling.

DNALongBench ETGP. The benchmark's `CRISPRi_EPI_K562_hg19.data.tab`, blacklist and bundled `hg19.fa.gz` from Harvard Dataverse (Cheng et al., 2025). The `.tab` must be fetched with `format=original` to get the published checksum. We use the published train/validation/test split unchanged.

Nucleotide Transformer (revised). All 18 tasks of `nucleotide_transformer_downstream_tasks_revised` (Hugging Face revision `851f994`). The published test split (chr20, chr21) is unchanged. Training rows form ten chromosome- and sequence-grouped folds. We report the median test MCC over the ten fold models.

Genomics Long-Range Benchmark. 8 tasks, excluding zero-shot-only OMIM, of `genomics-long-range-benchmark`, built with InstaDeep's loader at revision `b56f68e`. We use its data, sequence lengths (100 kb; 114,688 bp for CAGE) and filtering unchanged, and the UCSC genomes it downloads match UCSC's MD5 checksums. The test split is unchanged. For validation, chr6 and chr7 are held out of training, except for CAGE, which uses the loader's validation split.

B SEMI-SYNTHETIC DATA SAMPLING

We simulate the DNA cut-sequencing process by defining the following Algorithm for sampling fragments of length $R = T/K$ uniformly at random from a genome $\mathcal{G}$. By default, the fragments are left unordered and may overlap, since they are drawn with replacement.

Algorithm 1 (T, W, K)-parameterized dataset sampling

Require: Genome $\mathcal{G}$; token budget T; window W; fragment count K s.t. $K \mid T, 1 \leq T/K \leq W$
Ensure: Tokens x and positions π, both of length T

$R \leftarrow T/K$ ▷ fixed read length
Draw a window $[w, w + W)$ uniformly from all placements inside $\mathcal{G}$
for $k = 1, \ldots, K$ **do**
  $o_k \sim \mathcal{U}\{0, \ldots, W - R\}$ ▷ independent; fragments may overlap
  $F_k \leftarrow [w + o_k,\ w + o_k + R)$
end for
$x \leftarrow \mathrm{seq}(F_1) \,\|\, \cdots \,\|\, \mathrm{seq}(F_K)$ ▷ tokens, unsorted
$\pi \leftarrow \mathrm{coords}(F_1) \,\|\, \cdots \,\|\, \mathrm{coords}(F_K)$ ▷ genomic coordinates
return (x, π)