Advancing DNA Language Models: The Genomics Long-Range Benchmark

Chia-Hsiang Kao[*2], **Evan Trop** [*1], **McKinley Polen**[*1,3], **Yair Schiff**[*2], **Bernardo P. de Almeida**[1], **Aaron Gokaslan**[2], **Thomas Pierrot**[1], **Volodymyr Kuleshov**[2]

[1]InstaDeep
[2]Cornell University
[3]Massachusetts Institute of Technology
t.pierrot@instadeep.com, vk379@cornell.edu

Abstract

Building on the successes in other domains, there has been rapid development of language models (LMs) for genomics. Key to this development is the establishment of proper benchmarks and systematic evaluation approaches. The benchmarks that have been proposed so far have focused on tasks that depend on short-range sequence contexts, while the evaluation of models for long-range tasks that are integral to genomics, such as gene expression and genetic variant prediction, is lacking. In this work, we propose a benchmark that fills this need and introduce the genomics long-range benchmark – an evaluation tool that is designed to encompass tasks requiring long-range sequence dependencies, an aspect which we deem crucial to genomic applications of DNA language models. In addition to clearly defining and organizing relevant tasks into a cohesive benchmark, we provide preliminary results of several prominent and recent DNA LMs evaluated on the proposed benchmark. Finally, we probe the tasks in our benchmarks by exploring the effect of context length extension methods for one of the evaluated DNA LMs, the Nucleotide Transformer. By proposing this benchmark we hope to stimulate the ongoing development of DNA LMs and provide a fruitful testing ground for future developments that aim to capture long-range sequence modeling in genomics.

Introduction

Foundation models have emerged as a promising approach to tackle a broad array of problems across different domains (Bommasani et al. 2021), such as natural language processing and computer vision, and more recently in biology. An important driver for the development of such types of models has been the creation of proper benchmarks and systematic evaluation approaches. For example, it is difficult to separate the advent and success of AlphaFold in protein structure prediction (Jumper et al. 2021) from the Critical Assessment of Protein Structure Prediction competitions that spurred the development of this model (Kryshtafovych et al. 2021). More recent successes of foundation models include applications to the field of genomics and DNA sequences (Dalla-Torre et al. 2023; Ji et al. 2021; Nguyen et al. 2023; Zhou et al. 2023; Benegas et al. 2023). DNA language models (LMs) are showing promise in solving a diverse set of genomics tasks, and to foster their development and adoption it is important to establish benchmarks adapted to the tasks of core importance to the genomics community.

The Nucleotide Transformer (NT) work was the first attempt to establish a systematic study and benchmark of language models for DNA, testing models of varying sizes up to 2.5B parameters and pre-trained on diverse genomes from different individuals and species (Dalla-Torre et al. 2023). This benchmark included 18 diverse tasks of chromatin, splicing, and regulatory element predictions. Since then, other benchmark datasets have been published (Nguyen et al. 2023; Zhou et al. 2023). Although these datasets have provided important means by which we can compare existing models, they are limited to tasks that require short-range sequence context, typically up to 10k base pairs (bp). There remains a need for similar benchmarks for long-context tasks such as gene expression and genetic variant predictions, which are of higher relevance to the scientific community.

Tasks such as predicting the expression of every gene in different cell types from sequence alone and the impact of genetic variants on gene expression, remain challenging tasks in biology because they depend on long-range interactions between the regulatory elements that modulate the expression of the target gene, which can be as far as 1 million bp (Furlong and Levine 2018). The recently proposed architecture, known as Enformer, achieved state of the art performance on these tasks by combining convolutional and transformer layers to effectively integrate information from up to 100k bp away in the genome (Avsec et al. 2021). Enformer was trained in a supervised way to predict thousands of epigenetic and transcriptional profiles from hundreds of cell types using only the DNA sequence as input, thus also learning the intricate correlations between these diverse molecular entities. Towards the goal of building a standardized benchmark that encapsulates meaningful long-range genomic applications we make the following contributions:

1. We propose the genomics long-range benchmark (LRB) as an evaluation suite that tests the capabilities of models for solving long-range genomics tasks.

[*]These authors contributed equally.

2. We use this benchmark to compare the performance of different LMs with the Enformer as a baseline.
3. We study how to efficiently extend the context of current shorter-range LMs and demonstrate increased performance with increased context size.

Genomics Long-Range Benchmark

The motivation of the genomics LRB is to compile a set of relevant genomic tasks requiring long-range dependencies which will act as a robust evaluation tool for genomic LMs. While serving as a strong basis of evaluation, the benchmark must also be efficient and user-friendly. To achieve this we strike a balance between task complexity and computational cost through strategic decisions, such as down-sampling or combining datasets.

Variant Effect Prediction

This task, derived from the Enformer paper (Avsec et al. 2021), involves predicting whether a single nucleotide polymorphism (SNP) directly perturbs gene expression. The input to the task is a sequence from the human reference genome whose center position corresponds to a SNP with a reference and alternative allele. The output is a binary label, where labels are assigned to the positive class if their causal probability, given by population-based fine mapping analysis tool SuSiE (Wang et al. 2020), is $> .9$ (Avsec et al. 2021). The dataset originally proposed in Enformer was composed of 48 sub-datasets each corresponding to a different tissue. For the sake of ease of evaluation, we combine the sequences across tissues into one dataset. To preserve tissue information, we combine model embeddings with one-hot encoded tissue-indicator vectors. Since no test set was specified in the original dataset, we designate chromosomes 9 and 10 as the held out test set and the remaining chromosomes are designated for training purposes.

CAGE Gene Expression Profiling

This task corresponds to a biological assay that allows for high-throughput expression profiling of genes at their transcription start sites (TSS). The Cap Analysis of Gene Expression (CAGE) task, originally proposed in the Basenji paper (Kelley et al. 2018) and subsequently used in Enformer (Avsec et al. 2021), involves predicting the measured CAGE levels across the positions of the gene in various tissues and cell types. Input is a sequence from the human reference genome centered around the TSS of an associated gene, and outputs are continuous values of CAGE signal across the sequence for several tissue types. It is important to note that the dataset from Enformer uses CAGE labels of size 896×638, which correspond to 896 bins of 128 bp each and 638 human tissue types or cell lines (Avsec et al. 2021). For the sake of ease of evaluation and to make the task less computationally intensive, we subset the labels by sampling a representative set of 50 tissues to a final tensor of 896×50. We maintain the original test dataset split from Basenji and Enformer, but we combine the train and validation sets into one training split. Finally, continuous labels are $\log(1 + x)$ transformed and subsequently processed with standard scaling.

Bulk RNA-seq Gene Expression

Bulk RNA-sequencing is another biological assay which measures average expression of genes from a population of cells in a given tissue. For this task, the input is also a sequence from the human reference genome that is centered around the TSS. The output here is a single vector of continuous values representing the bulk RNA levels of a gene across 218 different tissue types. This task differs from the CAGE gene expression task not only by the biological assay but also in that the predictions here have a single expression value per gene, while for CAGE, we predict sequence coverage at different bin positions across the gene. The original dataset presented in the training and evaluation of ExPecto (Zhou et al. 2018) assigned chromosome 8 to the test dataset and all remaining chromosomes to the train dataset, which we maintain as our dataset splits. Continuous labels for this task are also $\log(1 + x)$ transformed and standardized.

Evaluation Methodology

As an initial analysis of our proposed benchmark, we begin by evaluating recently developed genomic LMs, such as NT and HyenaDNA (Dalla-Torre et al. 2023; Nguyen et al. 2023), on the genomics LRB. We also evaluate Enformer with the same methodology described below to provide a baseline result against which we can compare the genomic LMs. In order to fairly and systematically evaluate these models, we employ several design decisions. Firstly, we carry out five-fold cross validation (CV), selecting the best performing model for each fold based on validation loss, and evaluating on the held out test set for which we report the mean performance across folds. Validation dataset sampling for CV is done as follows: for variant effect and bulk RNA 5 chromosomes from the train set are randomly selected to each serve as a single validation set in the CV process. For CAGE expression, given that the original test set was not split by chromosome and that sequences from the same chromosome appear in both train and test, we randomly select a distinct 10% portion of the train set to serve as the validation set for each fold. To train and evaluate models in a standardized manner, we first run inference of each model to obtain embeddings for input sequences. These embeddings are then used to train a multi-layer perceptron (MLP). MLP hidden dimensions are sized in an adaptive way such the hidden state size is equal to two times the base model's embedding dimension. Embeddings of shape $\text{sequence length} \times \text{embedding dimension}$ are processed uniquely for each task.

In the variant effect task, mean embeddings are computed from a 1,500 bp window centered around the SNP for both the sequence with the reference and alternative allele. The mean embeddings for the reference and alternative alleles are subsequently concatenated together to create a vector that has dimension $2 \cdot \text{embedding dimension}$. Since sequences across all tissues were combined into a single dataset, we inject tissue information by additionally concatenating a one-hot encoded vector of tissue type with the concatenated mean embeddings of the previous step. This final representation is then passed as input to an MLP with two hidden layers, which is trained with a cross-entropy loss.

Given that CAGE expression is a task that requires making predictions for each position along the sequence, base model embeddings for each position in the sequence are input directly into an MLP with two hidden layers so that loss can be computed from every position in the sequence. It is important to note that models with a receptive field smaller than 114k bp, corresponding to the original 896 bins of 128 bp each, use a subset of the labels for computing the loss and model with receptive field larger than 114k bp use a sequence mask. We use mean squared error as the training objective.

For the bulk RNA task, mean embeddings are computed across a window of 384 bp upstream of the TSS and 256 bp downstream. The mean embeddings are subsequently fed into an MLP with a single hidden layer. As with CAGE, we use mean squared error as the training objective.

Benchmarking Results

The evaluation of the variant effect task highlights the Enformer's modest performance compared to genomic LMs. Despite variations in dataset partitioning between the original Enformer paper and our methodology, our results for Enformer yield an AUCROC value of 0.754, closely aligning with the reported 0.747 in the Enformer paper. We observe that all NT models exhibit slightly better performance than Hyena DNA models. Additionally, we note that performance appears to drop when model size reaches a certain point for both Hyena DNA and NT. We attribute this trend to the increased embedding dimension size in the larger models, leading to more trainable parameters in the MLP, which potentially causes over-fitting.

In CAGE expression, we observe a consistent ranking of model performance with Hyena DNA, NT, and Enformer in ascending order. We observe a Pearson R^{genes} across all positions of 0.701 for the Enformer model, closely aligning with the 0.712 reported value in the original paper. For Pearson $R^{tissues}$, we report 0.542 with the original paper reporting 0.532 for tissues with medium expression variance. Additionally, the significant difference in performance between Enformer and genomic LMs is likely attributable to Enformer's supervised training, which included CAGE as a subset of its original data. We observe a more prominent effect of model size on performance for NT for the CAGE task than for bulk RNA.

Bulk RNA results mirror the above rankings of model class by performance. A comparison of Enformer's performance between our methodology and the original paper shows similar Spearman R^{genes} values, with our methodology at 0.871 and the Enformer paper reporting 0.840. However, there is a larger difference between the Spearman $R^{tissues}$ metric, with values of 0.541 and 0.451 for our methodology and the original report, respectively. The increase in performance can likely be explained by the fact that we trained an MLP in comparison to only training a linear layer as done in the Enformer paper.

Within NT models, a positive correlation exists between model size and performance metrics such as Spearman R^{genes} and R^2. However, the same trend is not observed with Hyena DNA models.

Context Extension for DNA Language Models

In addition to the initial benchmarking of DNA LMs, we also conducted a more in-depth analysis of how context size affects performance on our proposed tasks, focusing on the variant effect task, where longer-range context is expected to improve performance (Avsec et al. 2021). For this analysis, we use the smallest NT model with 50M parameters due to computational considerations.

Context Extension for Transformer-based Models

Extending contexts for transformer-based models (Vaswani et al. 2017) is an open area of research. Owing to the quadratic cost of the attention mechanism in transformer-based models, training with long contexts is prohibitive in most scenarios. Therefore, much of the work in this realm has focused on either zero-shot or minimal-fine-tuning length extension. Several works have documented and explored the difficulties for length extension in attention-based models (Dubois et al. 2019; Press, Smith, and Lewis 2021; Anil et al. 2022; Kazemnejad et al. 2023). Given that the NT was trained using rotary positional embeddings (RoPE; Su et al. (2021)), we focus on recent approaches that have been proposed for extending contexts of RoPE models by converting the problem of length extrapolation into one of "interpolation". Specifically, we employ the method described in Peng et al. (2023), where the frequency used in RoPE embeddings is re-scaled to account for longer sequences. For more details on length interpolation for RoPE, see the Appendix.

Experiment Setup

We fine-tuned the NT 50M model, initially pre-trained on a context size of 12k bp, on the Human Genome Dataset Dalla-Torre et al. (2023), employing context sizes of 24k and 48k bp and utilizing the NTK-aware method described in Peng et al. (2023). The model extended to 24k bp was trained for approximately 80k steps and the model extended to 40k bps steps was trained for about 48k steps. We also vary the length of the sequence provided to the model during the downstream variant effect task fine-tuning, with lengths ranging from 12k to 192k bp; the latter being nearly on par with Enformer's context window (Avsec et al. 2021). For both the pre-trained and the context extended models, we apply appropriate NTK-aware rescaling given the input sequence length for the downstream task. We use 5-fold CV and report average performance on the test set.

Results on Variant Effect Prediction

The results are shown in Table 2. We observe that the pre-trained NT 50M model improves performance when the input token length is extended from the original 12k bp with which the model was pre-trained. This indicates that we can potentially see gains even with 'zero-shot' context extension on this model when employing the NTK-aware rescaling. On the other hand, we also asses how additional training of the base model on an extended context length of 24k and 48k affects how the model performs on longer context inputs. We observe a steady increase in performance from the

		HyenaDNA				Nucleotide Transformer				Enformer
# Params / Context Length (bp)		1.6M / 1K	.4M / 16K	3.3M / 32K	6.6M / 160K	50M / 12K	100M / 12K	250M / 12K	500M / 12K	251M / 196K
Variant Effect Prediction	AUC-ROC	0.705	0.704	0.713	0.706	0.714	0.722	0.721	0.719	**0.755**
	Accuracy	0.648	0.649	0.658	0.647	0.661	0.664	0.661	0.657	**0.668**
Bulk RNA Expression	Spearman R^{genes}	0.701	0.622	0.605	0.726	0.750	0.761	0.776	0.780	**0.871**
	Spearman $R^{tissues}$	0.182	0.166	0.167	0.230	0.225	0.219	0.237	0.306	**0.554**
	R^2	0.377	0.264	0.225	0.399	0.397	0.458	0.470	0.478	**0.802**
CAGE Profile Expression	Pearson R^{genes}	0.278	0.273	0.328	0.297	0.446	0.483	0.508	0.524	**0.701**
	Pearson $R^{tissues}$	0.082	0.106	0.139	0.177	0.157	0.162	0.171	0.170	**0.541**
	R^2	0.071	0.073	0.110	0.089	0.201	0.235	0.260	0.276	**0.492**

Table 1: Baseline Results of Contemporary DNA Language Models on the Genomics Long-Range Benchmark. Best values in each row are bolded.

pre-trained model to models that undergo additional training with extended context lengths of 24k and 48k bp, especially as length of input sequences increases . Of note, the evaluation when using 192k bp inputs for the 48k context-extended model brings the unsupervised pre-trained NT model performance close to that of the state-of-the-art Enformer model on this task. These results underscore the critical role of context length in model performance for this type of task and the benefits of fine-tuning with extended context lengths. In Figure 1, we show that the benefits of context-extended models with longer input sequences are even more pronounced as the distance between SNP and nearest gene TSS grows.

	Model / Context Length (bp)		
Input Context Length	Pre-trained / 12k	Extended / 24k	Extended / 48k
12K	0.718	0.725	0.729
24K	0.720	0.725	0.730
48K	0.726	0.728	0.732
96K	0.734	0.732	0.738
192K	0.730	0.735	**0.742**

Table 2: Context extension improves variant effect prediction. AUC-ROC values shown. Best value is bolded.

We present initial assessment findings regarding genomics LRB, highlighting that while genomic language models such as NT and Hyena DNA exhibit promising performance, Enformer outperforms them. It's worth noting that Enformer, unlike these language models, underwent supervised training on data closely aligned with the specifics of our proposed tasks.

Discussion

In this work, we introduce the genomics LRB, a tool to robustly evaluate genomic LMs. We provide preliminary evaluation results on the genomics LRB for genomic LM's Hyena DNA and NT and show comparisons with state of the art supervised model Enformer. Additionally, we explored

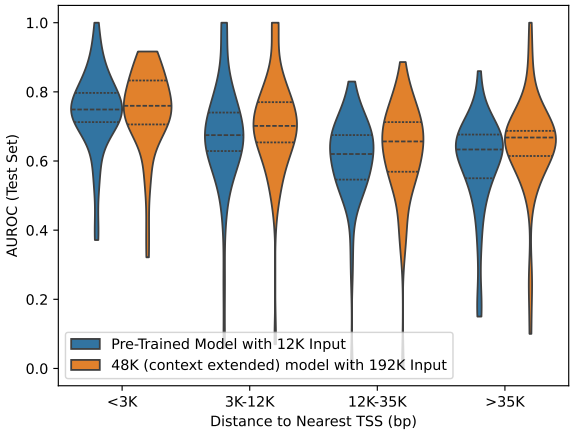

Figure 1: Performance on the variant effect prediction task broken down by distance to nearest gene TSS. Violin plots represent measures for the 49 tissues. The context-extended model with a longer input context improves AUC-ROC when compared with the pre-trained 50M NT.

context length extension for NT, ultimately displaying that increasing context window, either 'zero-shot' or with additional training, improves performance on variant effect prediction. We propose to further develop the genomics LRB by adding new tasks such as genome annotation and regulatory activity prediction. We believe these tasks are not only integral to genomics but would also permit for a fairer comparison with Enformer. To further build on genomic LM evaluations we plan to assess other models, such as DNABERT (Zhou et al. 2023) and GPN-MSA (Benegas et al. 2023). Finally, we hope to create an evaluation pipeline that the community can integrate into their own workflows and a leader board to show current standings on the genomics LRB. It is our hope that the work presented here can continue to spur development and meaningful advancement of the field of genomics LMs.

References

Anil, C.; Wu, Y.; Andreassen, A.; Lewkowycz, A.; Misra, V.; Ramasesh, V.; Slone, A.; Gur-Ari, G.; Dyer, E.; and Neyshabur, B. 2022. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35: 38546–38556.

Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.

Benegas, G.; Albors, C.; Aw, A. J.; Ye, C.; and Song, Y. S. 2023. GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction. *bioRxiv*.

bloc97. 2023. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023–01.

Dubois, Y.; Dagan, G.; Hupkes, D.; and Bruni, E. 2019. Location attention for extrapolation to longer sequences. *arXiv preprint arXiv:1911.03872*.

Furlong, E. E. M.; and Levine, M. 2018. Developmental enhancers and chromosome topology. *Science*, 361(6409): 1341–1345.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

kaiokendev. 2023. Things I'm learning while training superhot.

Kazemnejad, A.; Padhi, I.; Ramamurthy, K. N.; Das, P.; and Reddy, S. 2023. The Impact of Positional Encoding on Length Generalization in Transformers. *arXiv preprint arXiv:2305.19466*.

Kelley, D. R.; Reshef, Y. A.; Bileschi, M.; Belanger, D.; McLean, C. Y.; and Snoek, J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5): 739–750.

Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; and Moult, J. 2021. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12): 1607–1617.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Press, O.; Smith, N. A.; and Lewis, M. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; and Liu, Y. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33: 7537–7547.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, G.; Sarkar, A.; Carbonetto, P.; and Stephens, M. 2020. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5): 1273–1300.

Zhou, J.; Theesfeld, C. L.; Yao, K.; Chen, K. M.; Wong, A. K.; and Troyanskaya, O. G. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8): 1171–1179.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

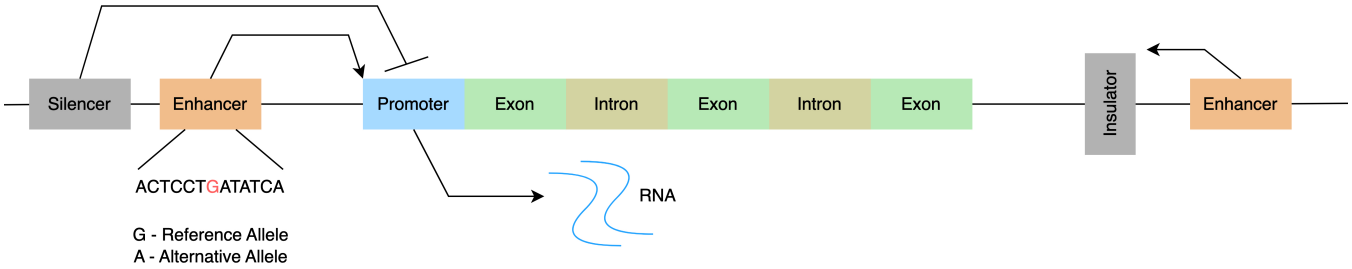

Figure 2: Shown in the figure is a model of gene transcription regulation. Distal regulatory elements such as enhancers, silencers, and insulators can effect gene transcription and subsequently expression. Enhancers recruit transcription factors and increase the target gene's transcription while elements, such as silencers, repress it. An insulator's function is to block the activity of both enhancers and silencers. These elements can act on genes over long distances through co-location in 3D space. Single nucleotide polymorphism's (SNPs) are single base pair changes located in both coding and non-coding regions. SNPs can alter the function of the elements in which they lie, for example by changing motifs which are critical for the binding of regulatory proteins. The CAGE expression task involves predicting the RNA expression quantitative profile at the transcription start sites of human genes as measured by CAGE, while the bulk RNA-seq tasks involve predicting the global expression value of the whole gene in a given cell type. In variant effect prediction, the task is to predict whether a given SNP and it's alleles affect gene expression.

Appendix

Length Interpolation for Rotary Embeddings

Rotary Embeddings

In attention-based modules, such as those used in Transformer models (Vaswani et al. 2017), for a sequence of length L, the model takes embeddings in $\{\mathbf{x}\}_{j=1}^{L}, \mathbf{x}_j \in \mathbb{R}^d$, where d is the dimension of the embeddings, and computes query, key, and value vectors at every m^{th} and n^{th} position in the sequence:

$$\begin{aligned}\mathbf{q}_m &= f_q(\mathbf{x}_m, m)\\ \mathbf{k}_n &= f_k(\mathbf{x}_n, n)\\ \mathbf{v}_n &= f_v(\mathbf{x}_n, n).\end{aligned}$$

f_q, f_k, f_v are query, key, and value transformations, respectively. For rotary embeddings (RoPE; Su et al. (2021)), we can think of $\mathbb{R}^d$ as equivalent to the complex field $\mathbb{C}^{d/2}$ and define f_q and f_k as:

$$\begin{aligned}f_q(\mathbf{x}_m, m) &= e^{im\Theta}\mathbf{W}_q\mathbf{x}_m\\ f_k(\mathbf{x}_n, n) &= e^{in\Theta}\mathbf{W}_k\mathbf{x}_n,\end{aligned}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are linear transformations and $\Theta = \text{diag}(\theta_1, \ldots, \theta_{d/2})$ is a diagonal matrix, with $\theta_j = b^{-2j/d}$ and $b = 10000$.

RoPE Position Interpolation

In the concurrent works of Chen et al. (2023) and kaiokendev (2023), the method of position interpolation was introduced, whereby longer sequences of length $L' > L$ are accommodated by simply rescaling the position input to f_q and f_k, e.g., $f_q(\mathbf{x}_m, m\frac{L}{L'})$.

NTK-aware RoPE Interpolation

An alternative interpolation scheme, attributed to bloc97 (2023), is motivated by the claim that Position Interpolation may lead to the loss of high frequency information. The approach that purportedly resolves this issue is related to the theory of Neural Tangent Kernels (NTK) by means of an analogy between RoPE and Fourier Features (Tancik et al. 2020), and is thus named "NTK-aware" interpolation. This scheme is characterized by a rescaling applied not to the position but rather to the basis of rotation, as follows:

$$\begin{aligned}\theta_j &= b'^{-2j/d}\\ b' &= b \cdot \left(\frac{L}{L'}\right)^{\frac{d}{d-2}}\end{aligned}$$

In the experiments on context extension presented in the main text, we adopt this interpolation scheme, both for zero-shot and additional training context extension.

We note that the authors in Peng et al. (2023) further tweak and build on NTK-aware interpolation to create their proposed interpolation scheme, which they title YaRN. However, the full YaRN approach, as presented in Peng et al. (2023) requires several manually tuned hyperparameters, which were carefully selected for the decoder-only generative Llama-2 7 billion parameter model (Touvron et al. 2023a,b). We therefore adopted the simpler NTK-aware approach in our experiments. Future work will explore the effect of the interpolation scheme on the results of our context extension analysis.