RECURRENT MEMORY AUGMENTATION OF GENA-LM IMPROVES PERFORMANCE ON LONG DNA SEQUENCE TASKS

Yuri Kuratov[1,2] **Aleksei Shmelev**[1,3] **Veniamin Fishman**[1,4] **Olga Kardymon**[1] **Mikhail Burtsev**[5]
[1]AIRI, Moscow, Russia [2]Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia
[3]HSE University, Moscow, Russia [4]Institute of Cytology and Genetics, Novosibirsk, Russia
[5]London Institute for Mathematical Sciences, London, UK
yurii.kuratov@phystech.edu, mb@lims.ac.uk

ABSTRACT

Utilizing DNA language models based on the transformer architecture represents a significant advancement in the field of computational genomics. However, these models face a critical challenge due to their inherent limitations in handling input lengths comparable to those of individual vertebrate genes (ranging from 10^4 to 10^5 nucleotides) and complete genomes (typically around 10^9 nucleotides). Currently, the architecture with the longest sequence input among publicly available transformer-based DNA language models, GENA-LM, is constrained to a maximum input length of merely $3 \cdot 10^4$ nucleotides. In this study, we investigate the efficacy of the Recurrent Memory Transformer (RMT) in enhancing GENA-LM for multiple genomic analysis tasks that require processing long DNA sequence inputs. Our results demonstrate that augmenting GENA-LMs with RMT leads to a substantial enhancement in performance, particularly in tasks such as species classification and prediction of epigenetic features. This underscores the significance of the recurrent memory approach in advancing the field of computational genomics and its potential for addressing critical challenges associated with processing long sequence inputs.

1 INTRODUCTION

Computational methods of genomics encounter several significant challenges. Firstly, the interconnection of various genomic characteristics is complex, with non-linear dependencies originating from multiple underlying mechanisms. For example, gene expression is often coordinated by multiple enhancers, with their effects depending on enhancer-promoter compatibility (Drew T. Bergman et al., 2022), spatial distance (Belokopytova et al., 2020), enhancer redundancy (Kvon et al., 2021), competition between promoters (Oudelaar et al., 2019), and many other factors, making gene expression prediction a challenging task. Machine learning methods, with their ability to resolve such complex dependencies, show promise in deducing unknown epigenetic properties either from measured characteristics or directly from DNA sequences. In the past decade, these computational approaches have evolved from ensemble trees (Belokopytova et al., 2020) to more sophisticated models like convolutional neural networks (Kelley et al., 2018; Zhou & Troyanskaya, 2015), recurrent neural networks (Quang & Xie, 2016), and, most notably, transformer-based architectures (Linder et al., 2023; Avsec et al., 2021). These advanced models have significantly improved the precision of inferring multiple genomic characteristics from DNA sequences (Linder et al., 2023; Avsec et al., 2021).

However, the expanding scale of datasets and the complexity of neural network architectures introduce a second challenge: the substantial computational resources required for training state-of-the-art models from scratch, which are often beyond the reach of many research groups. A promising solution to this has been the introduction of pretrained transformer models, which allow for achieving top-tier results through cost-effective fine-tuning of publicly available models (Zaheer et al., 2020; Dalla-Torre et al., 2023; Fishman et al., 2023; Ji et al., 2021; Nguyen et al., 2023; Zhou et al., 2023).

Transformer models generate high-quality predictions but face a limitation due to their quadratic increase in computation with the length of the input sequence. For instance, DNABERT, the pre-trained transformer model for DNA, was limited to processing sequences of only 512 base pairs (Ji et al., 2021). This limitation is particularly problematic in genomic data processing, where understanding long-range dependencies is crucial, as sequences located millions of base pairs apart can be spatially and functionally interconnected, presenting yet another challenge in the field of genomics (Fudenberg et al., 2020).

Architectural innovations like the sparse attention mechanism of BigBird (Zaheer et al., 2020), have expanded the capability of transformer models to handle sequences up to 32,000 base pairs (bp) or 32 kb in length. This extended capacity has allowed BigBird to outperform DNABERT in various downstream tasks due to its ability to process longer sequences, highlighting the critical role of processing extended DNA in genomic research.

Further advancements were made with the introduction of GENA DNA language models (GENA-LMs, Fishman et al. (2023)), which incorporate both the sparse attention mechanism and a Byte Pair Encoding (BPE) tokenization strategy. This combination allows handling inputs of 36 kb and demonstrated enhanced performance across a range of downstream tasks compared to both BigBird and DNABERT.

Recently, Recurrent Memory Transformers (RMT) have emerged as a novel solution to the challenge of extending the input length capacity of transformer models (Bulatov et al., 2022). Drawing inspiration from the principles of recurrent and memory networks, RMT sequentially processes lengthy inputs in chunks (or segments). To facilitate the transfer of information across these segments, RMT incorporates additional memory tokens into the standard BERT transformer architecture (Devlin et al., 2019), enabling the accumulation and retention of information over segments. RMT benchmarking (Bulatov et al., 2024) demonstrates the capability to handle extremely long inputs effectively (up to 2 million tokens). When integrated with GENA DNA language models (GENA-LMs), RMT has shown promise in various biological applications (Fishman et al., 2023). However, in genomic contexts, the performance of RMT-augmented GENA-LMs with sequence lengths surpassing the original limits of GENA-LMs (about 36kb) had not been thoroughly evaluated.

Here, we present the augmentation of GENA-LM with RMT, tested on three specific tasks: promoter and splice site prediction, prediction of epigenetic features and gene expression, and species classification. Our results indicate that augmenting GENA-LMs with RMT improves its performance in all three tasks examined and extends GENA-LMs to sequences up to 196,000bp. This underscores the potential of RMT in advancing the capabilities of genomic computational methods, particularly in handling and analyzing lengthy DNA sequences.

2 Recurrent Memory for GENA-LM

The GENA-LMs encompasses a variety of models, each distinguished by its architecture, training dataset, parameter count, and maximum input length (Fishman et al., 2023). We use GENA-LM transformer-based encoder-only models[1] trained on T2Tv2 human genome assembly of base (110M, bert-base-t2t) and large (350M, bert-large-t2t) sizes with 512 input sequence length, and sparse model (110M, bigbird-base-t2t, 4096 tokens) based on the BigBird architecture (Zaheer et al., 2020).

Recurrent Memory Transformers (Bulatov et al., 2022) is a plug-and-play approach for augmenting pre-trained transformers with memory to recurrently process segmented long sequences. Special memory tokens are added to each segment, and the corresponding outputs are passed as inputs to the next segment (Figure 1a).

Extending the input length of DNA language models with RMT can be approached through several strategies. One method involves integrating RMT at the initial pre-training phase of the foundational model (Figure 1a). Although this approach enables the model to capture long-range dependencies inherent in DNA sequences during pre-training, the extensive computational resources required for pre-training may render this option impractical in many scenarios. An alternative strategy entails utilizing a foundational model that has been pre-trained without RMT enhancements and incorpo-

[1] `https://huggingface.co/collections/AIRI-Institute/dna-language-models-65b229c1cd9b73cf8462ac96`

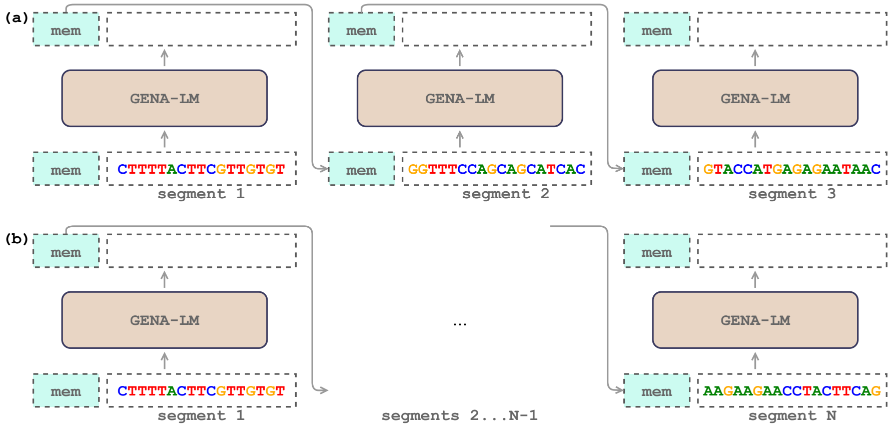

Figure 1: **Recurrent memory augmentation of GENA-LM models.** (a) Transformer-based GENA-LM model is augmented with recurrent memory ([mem]) and pre-trained on a long DNA sequence split on a fixed number of consecutive segments with MLM objective. During fine-tuning number of segments can be scaled up. (b) After training recurrent GENA-LM can be employed for predictions on a number of segments exceeding those encountered during training.

rating RMT during the fine-tuning stage. This allows the model to learn how to utilize memory tokens for information storage and retrieval throughout the fine-tuning process, offering a more resource-efficient solution.

An additional aspect that can be optimized in RMT applications is the number of segments processed during both training and inference phases. An important characteristic of RMT is its ability to generalize the use of memory tokens across a variable number of segments, beyond those seen during training (Figure 1b). For instance, recent experiments have demonstrated that a model trained with only 7 segments can effectively utilize its memory capabilities during inference on up to 4096 segments (Bulatov et al., 2024). This feature of RMT allows for reducing computation during the training or fine-tuning stages, as full-length samples may only need to be processed during the inference stage, offering a more resource-efficient approach to handling extensive sequences.

The training schedule plays an important role in the development of RMT models. Research has indicated that RMT training is particularly effective with curriculum learning (Bulatov et al., 2024). It involves initial RMT training on a small number of segments, and once the model reaches a point of convergence, the length is incrementally increased with additional segments. This makes possible for the RMT to adapt progressively to longer sequences, enhancing its ability to handle extended inputs effectively.

As backbones for memory augmentation, this study takes *gena-lm-bert-base-lastln-t2t* and *gena-lm-bert-large-t2t* models. In all RMT pre-training experiments, memory had 10 tokens and a segment length was 512 tokens. As a result, we pre-trained two models in the following setups:

- *gena-lm-bert-base-lastln-t2t*, 8 segments, no curriculum, 440k iterations
- *gena-lm-bert-large-t2t*, 2-4-8 segments, with curriculum, 200k-30k-60k iterations per curriculum step accordingly.

We used data pipeline from (Fishman et al., 2023) to pre-train RMT on Masked Language Modeling (MLM) task (Devlin et al., 2019). In this task, a fraction of input tokens is replaced by a special "[MASK]" token and should be predicted by the model. The MLM loss was computed on all segments without stop gradients and backprop truncation. We used batch size 256, AdamW optimizer, learning rate in {2e-05, 1e-05}, and constant linear rate schedule with 50k warm-up steps. The pre-training dataset comprises the human T2Tv2 genome along with SNP augmentations from gnomAD 1000-genomes[2]. To enrich the dataset, we used the same augmentations as in (Fishman et al., 2023): reverse-complementary sequences and random shifts. We employ the same tokenizer as used in GENA-LMs. The total size of the dataset is about 480×10^9 bp and 73×10^9 tokens. Our code is based on the original RMT implementation[3]. The results of RMT pre-training can be found in Appendix A.

[2] T2T: T2T-CHM13v2.0, NCBI: GCF_009914755.1; gnomAD: v3.1.2

[3] `https://github.com/booydar/recurrent-memory-transformer`

3 Results

3.1 Promoters and splice sites prediction

The accurate prediction of promoters and splice sites is a critical task in genomics, as these regions play key roles in the regulation of gene expression. Improved predictive models for these genomic features can significantly enhance our understanding of gene regulation mechanisms and contribute to advancements in areas such as disease research, personalized medicine, and biotechnology. In Table 1, we report results for the best configurations of each model, we choose from $\{5, 10, 20\}$ memory tokens for RMT. Results are averaged over 5 folds for promoters and over 3 runs with different seeds for splice sites. Detailed information regarding dataset construction can be found in Appendix B.1 and comparison with non-RMT GENA-LM models in Table 3.

Experiments using models pre-trained with RMT show superior results compared to those using fine-tuning with RMT alone on both tasks. There is no significant difference for large models on the promoter prediction task.

Table 1: **RMT pre-training of GENA-LM improves promoters and splice site prediction compared to fine-tuning only.** Models denoted with (+P) indicate those with RMT pre-training.

Model	Promoters (16 kb), F1	Splice sites (15 kb), PR-AUC
RMT+GENA-LM base	93.70 ± 0.46	0.9353 ± 0.001
RMT+GENA-LM base (+P)	$\mathbf{94.61} \pm 0.54$	$\mathbf{0.9429} \pm 0.001$
RMT+GENA-LM large	95.58 ± 0.43	0.9471 ± 0.001
RMT+GENA-LM large (+P)	95.36 ± 0.44	$\mathbf{0.9518} \pm 0.001$

3.2 Epigenetic features and gene expression prediction

Predicting epigenetic characteristics and gene expression from DNA sequences is one of the most challenging tasks in the field of computational genomics. The complexity of this task is heightened by the functional interconnections between genomic sequences that can be separated by substantial distances, ranging from tens to hundreds of kb. These long-range dependencies within the genome render the task particularly well-suited for applications of recurrent memory. RMT's ability to process and integrate information from long sequences makes promise for tackling the intricate relationships and interactions that underlie epigenetic features and gene expression patterns in the genome.

To establish a baseline for this task, we fine-tuned GENA-LMs without the integration of RMT. Utilizing the *gena-lm-bert-large-t2t* model, which accommodates an input length of 512 tokens corresponding to 24 bins (equivalent to 3072 bp), we achieved a Pearson correlation coefficient (R) of 0.5899. However, *gena-lm-bigbird-base-t2t*, which has the same number of parameters as *gena-lm-bert-base-t2t*, but allows processing longer inputs (4096 tokens, equal to 192 bins or 24576 bp) due to sparse attention mechanism, results in substantially better performance: R=0.6146. Please refer to Appendix B.2 for details on data processing and models training.

We subsequently enhanced the *gena-lm-bert-large-t2t* model with RMT, enabling the processing of longer sequences with the GENA-LM that has the largest parameter count. As previously shown (Bulatov et al., 2024), the application of curriculum learning proved to be critical for the effective training of RMT models (Figure 2). The *rmt+gena-lm-bert-large-t2t* model, when trained on just 2 segments, does not generalize to a higher number of segments. Conversely, models trained on four or more segments demonstrated robust generalization when tested with an increased number of segments during inference. The best performance was achieved with the model trained on 24 segments and evaluated on 48 segments, achieving $R = 0.6151$. Eventually, RMT with segments of 512 tokens achieves slightly better results on 48 segments compared to a sparse model capable of processing 4096 tokens.

In RMT, samples are processed from left to right, which means that the latter segments benefit from a significantly richer contextual information accumulated in the memory tokens, as opposed to the

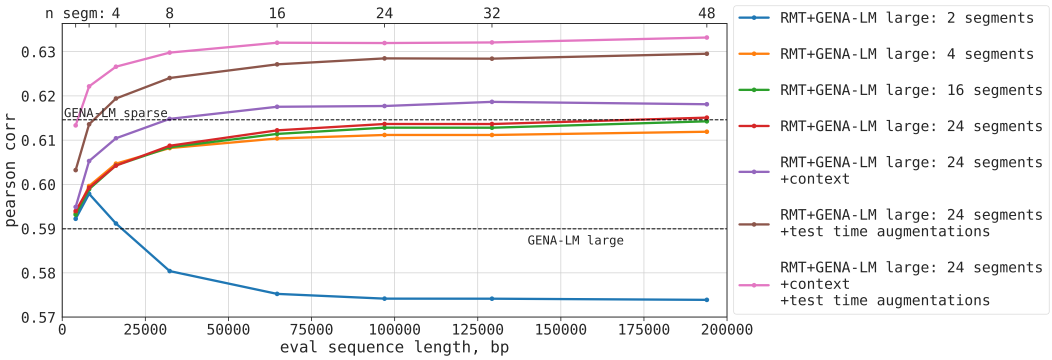

Figure 2: **RMT makes GENA-LM significantly better for epigenetic features and gene expression prediction.** RMT demonstrates the ability to generalize to lengths not seen during training, starting with training on 4 segments. As the number of training segments increases, RMT continues to improve. RMT significantly enhances GENA-LM results by utilizing a longer context, up to 24 segments in fine-tuning. While extending the input length at inference from 24 to 48 segments still brings some improvements, they are not as pronounced. The best results for RMT with GENA-LM large are obtained with both test time augmentations and with half of the segments used as additional context.

initial segments. In line with this, we adopted a strategy where half of the segments were designated as additional context (Figure 4). Specifically, for an input consisting of 24 segments, we updated memory states with the first 12 segments and performed predictions solely based on the subsequent 12 segments. This approach resulted in achieving an R-value of 0.619 for inputs with 48 segments. The RMT's ability to process sequences beyond its backbone model length of 512 tokens resulted in improvements over a sparse model, which handles 4096 tokens per segment, but lacks recurrence and generalization to longer inputs.

It was observed that augmenting the sequence during inference enhances the accuracy of predictions. By averaging the predictions over 8 augmentations: the original sequence, its reverse-complement, and sequences derived by shifting the original sequence by several base pairs, as suggested in (Avsec et al., 2021), we obtained our best result of $R = 0.6331$.

3.3 Species classification

The task of classifying mammalian species based on genomic sequence fragments was recently established as a benchmark for evaluating the capability of models to process long DNA sequences (Nguyen et al., 2023). The initial benchmarking conducted by Nguyen et al. (2023) highlighted a pronounced correlation between sequence length and classification accuracy: the accuracy surged from 61.1 with 1 kb sequences to an impressive 99.5 when the sequence length was extended to 1000 kb, as summarized in Table 2.

Without augmentation with RMT, the *gena-lm-bert-base-t2t* model is constrained to processing sequences with a maximum length of approximately 4 kb. To allow the processing of longer sequences, we augmented it with RMT. As a baseline, we used HyenaDNA scores reported recently (Nguyen et al., 2023).

For sequence inputs of 1 kb, the classification accuracy of both the *rmt+gena-lm-bert-base-t2t* model and HyenaDNA was relatively low, with the former achieving 61.45 ± 0.91 and the latter 61.1. A significant enhancement in classification accuracy was observed when the sequence length was increased to 32 kb, with *rmt+gena-lm-bert-base-t2t* achieving 99.24 ± 0.06, thereby surpassing the performance of HyenaDNA (93.4), as detailed in Table 2. Further extending the sequence length to 50 kb elevated the classification accuracy of *rmt+gena-lm-bert-base-t2t* to 99.67 ± 0.059, exceeding the accuracy HyenaDNA attained with 1000 kb sequences. This indicates that, within this experimental framework, RMT and the associated model architecture extract and leverage information from extended DNA sequences more efficiently than other technologies designed for processing long input sequences such as Hyena layers underlying HyenaDNA.

Table 2: **RMT with GENA-LM overperforms HyenaDNA on the species classification task.** RMT+GENA-LM achieves over 99% accuracy starting from 32kb, surpassing HyenaDNA, which needs 1000kb to reach 99.5% accuracy.

Model	Sequence length			
	1kb	32kb	50kb	1000kb
HyenaDNA (Nguyen et al., 2023)	61.1	93.4	-	99.5
RMT+gena-lm-bert-base-t2t	**61.45** ± 0.91	**99.24** ± 0.06	**99.67** ± 0.059	-

4 Discussion and Conclusions

In our study, we applied recurrent memory augmentation with RMT to GENA-LM DNA language models and assessed the resulting performance across several biological tasks. Initially, we explored the impact of incorporating RMT during the model's pre-training phase. Our findings indicate that the use of RMT does not markedly enhance the MLM accuracy. For promoters and splice site prediction, the inclusion of RMT during pre-training yielded improvement in results, but the advantage is not dramatic compared to those pre-trained without RMT. We speculate that RMT pre-training may be more beneficial for downstream tasks characterized by shorter input sequences, where the model has limited opportunity to learn how to utilize memory during fine-tuning. Identifying the specific attributes of datasets for which RMT pre-training is advantageous remains an area for further investigation. Based on our current results, we recommend utilizing models pre-trained with RMT when available. When such models are not accessible, starting experiments from augmentation with RMT during the fine-tuning phase represents a viable approach.

Our results suggest that RMT application during the fine-tuning phase is unequivocally beneficial for all the biological tasks examined in the study. Interestingly, the gene expression prediction task underscored the superiority of the model employing sparse attention mechanisms over larger models enhanced with RMT for processing inputs of similar lengths. This might suggest that models with sparse attention extract information from inputs more efficiently than RMT-based models. An alternative explanation could be that in an RMT setup, where the model processes sequences from left to right, the initial segments lack sufficient contextual information. Supporting this notion, our experiments indicated that the RMT-based model's predictions for the latter parts of a sequence were more accurate. Employing a bidirectional approach for processing samples, coupled with a sampling strategy that prioritizes segments with more extensive contextual information, could potentially mitigate this limitation. As input lengths exceed the capacity of models with sparse attention, RMT-enhanced models demonstrate significantly improved performance. This is likely due to the RMT-based model's ability to harness long-range regulatory connections prevalent among genomic elements, which becomes increasingly feasible with extended input lengths.

While our findings convincingly demonstrate the benefits of employing RMT for processing extended DNA sequences, the availability of suitable biological datasets for benchmarking this methodology remains limited. Most high-throughput genomic assays, including massive parallel reporter assays, CRISPR-based mutagenesis, and various biotechnological applications, typically focus on measuring the properties of relatively short DNA sequences, ranging from tens to hundreds of base pairs. However, with the ongoing advancements and wider adoption of long-read sequencing technologies, coupled with significant improvements in the quantity and quality of genome assemblies, genome editing techniques, and *in vitro* DNA synthesis systems, the importance of efficiently processing long DNA sequences in genomics is set to increase markedly.

In conclusion, this study explores the effectiveness of Recurrent Memory Transformers (RMT) in augmenting GENA-LM, a transformer-based DNA language model, for processing long DNA sequences. The application of RMT both during the pre-training and fine-tuning phases across various genomic tasks showed significant improvements, particularly in species classification where it outperformed existing models. An essential aspect of this research is the recurrent nature of the RMT approach, which enables the processing of entire genomes with a single model in one pass, thus opening new avenues in genomic research. Another feature of RMT is its versatility, as it can extend any pre-trained transformer model to be integrated into various existing genomic computational frameworks, thereby advancing the field's capabilities in handling and analyzing extensive DNA sequences.

REFERENCES

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. *Nat. Methods*, 18:1196–1203, October 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x.

Polina S. Belokopytova, Miroslav A. Nuriddinov, Evgeniy A. Mozheiko, Daniil Fishman, and Veniamin Fishman. Quantitative prediction of enhancer-promoter interactions. *Genome Res.*, 30(1): 72–84, January 2020. ISSN 1549-5469. doi: 10.1101/gr.249367.119.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11079–11091. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/47e288629a6996a17ce50b90a056a0e1-Paper-Conference.pdf`.

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail Burtsev. Beyond attention: Breaking the limits of transformer context length with recurrent memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17700–17708, Mar. 2024. doi: 10.1609/aaai.v38i16.29722. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29722`.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, pp. 2023.01.11.523679v3, January 2023. doi: 10.1101/2023.01.11.523679.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL `https://aclweb.org/anthology/papers/N/N19/N19-1423/`.

# Drew T. Bergman, # Thouis R. Jones, Vincent Liu, Judhajeet Ray, Evelyn Jagoda, Layla Siraj, Helen Y. Kang, Joseph Nasser, Michael Kane, Antonio Rios, Tung H. Nguyen, Sharon R. Grossman, Charles P. Fulco, Eric S. Lander, and Jesse M. Engreitz. Compatibility rules of human enhancer and promoter sequences. *Nature*, 607(7917):176–184, July 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04877-w.

Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences. *bioRxiv*, pp. 2023.06.12.544594, November 2023. URL `https://doi.org/10.1101/2023.06.12.544594`.

Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods*, 17:1111–1117, November 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0958-x.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, January 2019.

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`.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, 28(5):739, May 2018. doi: 10.1101/gr.227819.117.

Evgeny Z. Kvon, Rachel Waymack, Mario Gad, and Zeba Wunderlich. Enhancer redundancy in development and disease. *Nat. Rev. Genet.*, 22:324–336, May 2021. ISSN 1471-0064. doi: 10.1038/s41576-020-00311-x.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *bioRxiv*, pp. 2023.08.30.555582, September 2023. URL https://doi.org/10.1101/2023.08.30.555582.

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. *arXiv*, June 2023. doi: 10.48550/arXiv.2306.15794.

A. Marieke Oudelaar, Caroline L. Harrold, Lars L. P. Hanssen, Jelena M. Telenius, Douglas R. Higgs, and Jim R. Hughes. A revised model for promoter competition based on multi-way chromatin interactions at the α-globin locus. *Nat. Commun.*, 10(5412):1–8, November 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13404-x.

Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, 44(11):e107, June 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw226.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, 12(10):931, October 2015. doi: 10.1038/nmeth.3547.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv*, June 2023. doi: 10.48550/arXiv.2306.15006.

A RMT PRE-TRAINING

We augmented GENA-LM with recurrent memory and pre-trained it on the MLM task. As shown in Figure 3, both base and large-sized models do not exhibit consistent improvements in MLM accuracy with increased length. This observation aligns with the results of the GENA-LM sparse model, which can handle 4096 tokens (equivalent to 8 segments for RMT models), but only marginally enhances the base model from 0.2297 to 0.2306. However, we find that pre-training with RMT enhances the base model's performance on sequences ranging from 2 to 10 segments, and from 8 to 10 segments for the large model. Unfortunately, models of both size exhibit a drop in MLM accuracy when processing beyond 8 segments they were trained on.

The MLM accuracy serves solely as a metric for pre-training and does not directly represent the final performance of the model on downstream tasks. We suppose that pre-training should be helpful for RMT to effectively learn memory operations. However, pre-training with memory for the large model appears to keep nearly the same results in MLM accuracy.

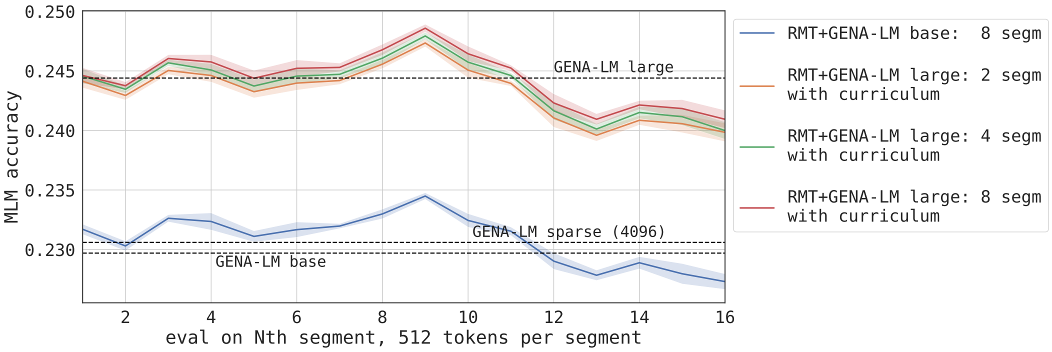

Figure 3: Pre-training RMT with GENA-LM on the MLM task yields improved results on 2–10 segments for the base-size model. However, the large model shows no clear benefit from pre-training in terms of MLM accuracy. Similarly, GENA-LM sparse, with an input length of 4096 tokens (equivalent to 8 segments), also doesn't show a noticeable advantage from increased context length on MLM accuracy. Results are averaged over 5 runs on valid set with different seeds for masking.

B DATASETS AND TASKS

B.1 PROMOTERS AND SPLICE SITES PREDICTION

We followed GENA-LM (Fishman et al., 2023) setup to run experiments on promoter activity prediction and splice site annotation tasks. For promoters, we used 16kb sequences from the EPDnew database[4] and followed instructions from GENA-LM repository[5] to get train/valid/test splits. For splice sites, we used data from (Jaganathan et al., 2019) and followed GENA-LM instructions[6] to get 15,000bp sequences. Comparison with non-RMT GENA-LM models is in Table 3.

B.2 EPIGENETIC FEATURES AND GENE EXPRESSION

For the prediction of epigenetic features and gene expression, a subset of human data from the Enformer (Avsec et al., 2021) dataset was utilized. This subset comprises processed signals from 5,313 experimental measurements. The dataset's structure is elaborated upon in detail in the cited reference. In summary, each sample in the dataset encompasses a target region spanning 114,688 bp, which is divided into 896 consecutive genomic bins, each 128 bp in length. This central target

[4] https://epd.epfl.ch/EPDnew_select.php

[5] https://github.com/AIRI-Institute/GENA_LM/tree/main/downstream_tasks/promoter_prediction

[6] https://github.com/AIRI-Institute/GENA_LM/tree/main/downstream_tasks/SpliceAI

Table 3: **Augmenting GENA-LMs with RMT improves results on promoter and splice site prediction tasks for both base and large size models.** In addition, RMT with GENA-LM large outperforms the GENA-LM model with sparse attention. For base size models, RMT shows competitive results compared to the sparse attention model. Models denoted with (+P) indicate those with RMT pre-training.

Model	Len, bp	Promoters, F1	Len, bp	Splice sites, PR-AUC
GENA-LM base (Fishman et al., 2023)	2000	91.18 ±0.62	4500	0.9263 ±0.001
GENA-LM large (Fishman et al., 2023)	2000	93.70 ±0.44	4500	0.9360 ±0.001
GENA-LM base sparse (Fishman et al., 2023)	16000	94.64 ±0.30	15000	0.9478 ±0.001
RMT+GENA-LM base	16000	93.70 ±0.46	15000	0.9353 ±0.001
RMT+GENA-LM base (+P)	16000	94.61 ±0.54	15000	0.9429 ±0.001
RMT+GENA-LM large	16000	**95.58** ±0.43	15000	0.9471 ±0.001
RMT+GENA-LM large (+P)	16000	95.36 ±0.44	15000	**0.9518** ±0.001

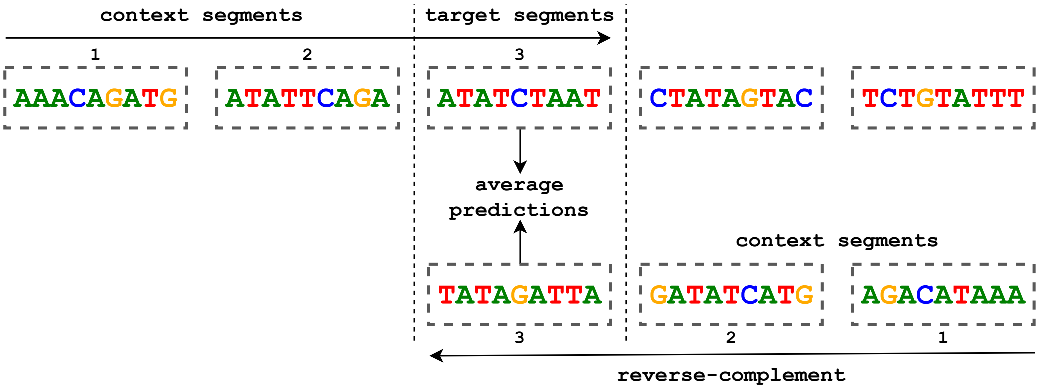

Figure 4: **Augmenting RMT with context segments and reverse-complement.** Context segments allow the use of longer context before making predictions for target segments. Since RMT processes sequences from left to right, reverse-complement adds information from the right context to target segments predictions. As a result, information from both directions is taken into account for target segments.

region is further surrounded by 40,960 bp of contextual information on either side, bringing the total length of each sample to 196,608 bp (calculated as $128 \cdot 896 + 40960 \cdot 2$ bp). For every one of the 896 target bins, there are 5,313 measurements available. The train-test split was maintained as per the original configuration in the Enformer dataset.

To process the samples, we use the following strategy. Initially, the target region is segmented into bins, each comprising 128 base pairs (bp), and each bin is tokenized independently. Subsequently, these tokenized bins are concatenated, with SEP tokens added to separate each bin. The sequence is then divided into segments, with each segment accommodating a number of tokens that corresponds to the input capacity of the GENA-LM being utilized (either 512 or 4096 tokens). If the model's input exceeds the resulting sample length, we add a tokenized context sequence. The formatted input for processing thus adopts the following structure: [CLS] left_context [SEP] bin1 [SEP] bin2 [SEP] ... [SEP] rigth_context [SEP].

We started RMT curriculum learning from *gena-lm-bert-large-t2t* fine-tuned on the Enformer dataset and gradually increased the number of segments: 2-4-16-24, making up to 12k tokens per sample. Each segment consists of 512 tokens, including 5 memory tokens and special tokens. We followed (Avsec et al., 2021) and augmented the training data with reverse-complement and random shifts (±0–3bp). To evaluate model quality, we compute Persons's correlation between targets and predictions across samples for each feature independently. Next, we averaged correlations obtained for each of the 5313 features to derive a single score.

B.3 SPECIES CLASSIFICATION

We constructed a dataset for species classification as described in (Nguyen et al., 2023). Genomes from 5 species (human, lemur, mouse, pig, hippo) were downloaded from NCBI (RefSeq assemblies GCF_000001405.40, GCF_020740605.2, GCF_000001635.27, GCF_000003025.6, GCF_030028045.1 respectively). Four chromosomes (chromosomes 1, 3, 12, and 13) were used for models evaluation, other chromosomes were utilized during training. We sampled sequences from chromosomes randomly, using the uniform distribution. We used 5-way classification and reported top-1 accuracy. For each task length, we collected a total of 50000 DNA subsequences from each species, ensuring a comprehensive dataset for our analysis.

We used *gena-lm-bert-base-t2t* model that has been augmented with RMT (8 segments) during the pre-training phase. The processes of fine-tuning were enhanced through the application of a curriculum learning strategy. This meant that our initial step includes fine-tuning the model on DNA subsequences of 1000 bp in length (single segment). Following this initial phase, we proceeded to extend the fine-tuning process to handle longer DNA subsequences while using the model weights from the 1000 bp fine-tuned model as an initial weights, increasing the task length to 32 kb (8 segments). Continuing with this progressive training methodology, we further advanced our model's capabilities by eventually fine-tuning it to efficiently process and analyze DNA subsequences extending up to 50 kb in length (12 segments). This gradual fine-tuning approach, in line with the principles of curriculum learning, facilitated the model in sequentially learning tasks of increasing complexity, thereby enhancing its analytical precision and performance on genetic classification tasks.