2025 IEEE International Conference on Image Processing (ICIP) | 979-8-3315-2379-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIP55913.2025.11084625

MULTIMODAL CELL CONTEXT INSTRUCTION TUNING FOR CONDITIONAL DNA REGULATORY SEQUENCE GENERATION WITH LARGE LANGUAGE MODELS

Junhao Liu, Pengpeng Zhang*, Siwei Xu, Shushrruth Sai Srinivasan, Yongxian Wu, Jing Zhang*†

University of California, Irvine
{junhao.liu, zhang.jing}@uci.edu

ABSTRACT

Designing biologically plausible regulatory sequences, such as enhancers, is a key challenge in synthetic biology. Current approaches either classify active enhancers or generate synthetic sequences independently but often fail to address the need for cell-type-specific recruitment of multiple transcription factors (TFs) for effective transcription activation. This oversight leads to suboptimal designs. To overcome these limitations, we propose LEONINE, a novel framework redefining enhancer sequence design as a multimodal question-answering task within a cellular context. LEONINE utilizes large language models (LLMs) trained on DNA sequences and integrates multimodal cellular data—including promoter sequences, gene expression profiles, and cell type information—to optimize enhancer sequences for specific regulatory environments. A large-scale dataset of over 1.5 million cell-type-specific promoter-enhancer pairs and a robust evaluation benchmark were developed to support this effort. Extensive evaluations across seven cell types demonstrate that LEONINE outperforms state-of-the-art LLMs, generating biologically and functionally coherent sequences. This work introduces a new paradigm for context-aware DNA sequence design, advancing research in synthetic biology and gene regulation.

Index Terms— Multimodal Large Language Models, Regulatory Sequence Generation, Discrete Signals

1. INTRODUCTION

In eukaryotic systems, distal regulatory elements, such as enhancers, spatially fold in three dimensions to interact with proximal sequences, including promoters, thereby orchestrating the precise spatiotemporal regulation of gene transcription. This intricate interplay between promoters and enhancers represents a fundamental mechanism that underpins gene expression and governs cellular behavior [1]–[3]. Consequently, the accurate modeling and the conditional generation of these interactions are essential for uncovering the principles of gene regulation and advancing applications in fields such as synthetic biology and gene therapy [4]–[6].

Previous research in this domain has predominantly concentrated on two main approaches. The first involves the development of classification models to predict promoter-enhancer interactions, aiming to determine whether specific promoter and enhancer sequences are functionally linked [7], [8]. While these models have enhanced the ability to identify regulatory pairs, they do not address the generation of novel sequences. The second approach applies various generative models to design DNA sequences, often training them from scratch, to produce synthetic DNA sequences with desired characteristics [9], [10]. However, these models typically generate enhancer or promoter sequences independently, without accounting for their specific interactions within a three-dimensional genomic context. Furthermore, it is well-known that distinct sets of regulatory genes, such as transcription factors (TF), are employed via short sequence patterns (e.g., motifs) in each cell type to guide enhancer-promoter interactions in a cell-type-specific manner. Current methodologies seldom incorporate this cell type specificity during the generative process, thereby failing to capture the influence of the cellular environment. Consequently, the generation of regulatory sequences that are *cell-type-* and *context*-specific remains a significant challenge.

To address these limitations, this study reconceptualizes the regulatory sequence design problem as a multimodal question-and-answer (Q&A) task within a defined cellular context. This framework facilitates: (1) the design of enhancer sequences (answers) that are optimally compatible with specific promoter sequences (questions), and (2) the optimization of potential enhancer-promoter interactions within a distinct cellular environment (e.g., cell type and gene expression levels). To achieve these objectives, we introduce LEONINE, a computational model leveraging DNA LLMs to design functional enhancer sequences. Unlike existing methodologies, LEONINE integrates diverse contextual information, including promoter sequences, gene expression levels, and cell type information, enabling the generation of biologically plausible enhancer sequences tailored to specific regulatory environments. To the best of our knowledge, this is the first framework to incorporate multimodal cellular context

*Both authors contributed equally to this work.

†Corresponding author. This work was supported by the National Institutes of Health [R01HG012572, R01NS128523].

into the conditional generation of DNA sequences.

To evaluate the performance of LEONINE, we conducted comparative analyses against **2** state-of-the-art LLMs across **7** cell types, yielding insights into its effectiveness and applicability from both biological and semantic perspectives. This framework establishes a new benchmark for promoter-enhancer sequence generation, representing a significant advancement in the application of DNA LLMs to genomics research. Our contributions can be summarized as follows:

- We introduce LEONINE, the first conditional enhancer sequence generation model utilizing a context instruction tuning strategy applied to DNA LLMs.
- Unlike existing approaches, LEONINE integrates multimodal cellular context (e.g., interacting DNA sequences and gene expression data) to generate enhancer sequences optimized for activating a specified promoter within a particular regulatory environment.
- We constructed a large-scale, cell-type-specific, multimodal dataset of promoter-enhancer sequences and conducted comprehensive evaluations of promoter-enhancer sequence generation.
- Extensive experimental results demonstrate that the proposed multimodal cell-context instruction tuning strategy produces enhancer sequences that are both biologically plausible and semantically coherent.

2. METHODOLOGY

2.1. Multimodal Promoter-Enhancer Generation

As shown in **Fig. 1**, our goal is to enable DNA LLMs to understand multimodal cellular contexts (e.g., cell types and gene expression profiles), which allows us to capture complex cellular information and generate biologically plausible enhancer sequences given a specific promoter sequence. We choose HyenaDNA [11] as the DNA LLM backbone of LEONINE $f_{\psi}(\cdot)$ parameterized by as it has demonstrated impressive unconditional generation capabilities for DNA sequences with publicly available model checkpoints [11], [12]. Additionally, we evaluated a transformer-based LLM backbone derived from the GPT-2 architecture [13], which had not been pre-trained on a large-scale DNA corpus. Our experiments revealed that this transformer-based backbone underperformed compared to HyenaDNA, as will be discussed in subsequent sections.

We utilized gene expression profiles and assigned cell types to represent the cellular context, both of which are widely recognized as influential factors in shaping promoter-enhancer interactions within the three-dimensional genomic space [14]. For a given cell type $\boldsymbol{x}_c \in [0 \quad N]$ and its corresponding gene expression profile $\boldsymbol{x}_g \in \mathbb{R}^{d_g}$, the promoter sequence $\boldsymbol{x}_p = [x_1 \quad x_L]$ that belongs to a cell type

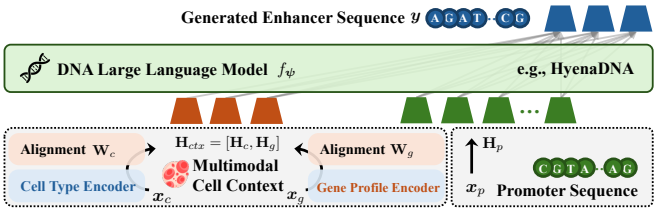

Fig. 1: The overview of LEONINE for promoter-enhancer generation using multimodal cell context instruction tuning.

$\boldsymbol{x}_c$ is regulated by an enhancer sequence $\boldsymbol{y} = [y_1 \quad y_L]$, where L is the DNA sequence length. Both x_i and y_i represent a vocabulary index to a token corresponding to one of the four unique nucleotide bases {A T C G}. Therefore, the multimodal promoter-enhancer generation is defined as a conditional generation problem:

$$p(\boldsymbol{y}|\boldsymbol{x}_c \ \boldsymbol{x}_g \ \boldsymbol{x}_p) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_c \ \boldsymbol{x}_g \ \boldsymbol{x}_p \ \boldsymbol{y}_{<i}) \tag{1}$$

where $\boldsymbol{\theta}$ represents the trainable parameters of the conditional probability function, and $\boldsymbol{y}_{<i}$ denotes all previous tokens preceding the current prediction token y_i.

2.2. Multimodal Cell Context Instruction Tuning

Although a DNA LLM f_{ψ} can generate DNA sequence via large-scale pretraining on a genome corpus, it is challenging for the model to understand and follow the multimodal conditions defined by cell type $\boldsymbol{x}_c$, gene expression $\boldsymbol{x}_g$, and promoter sequence $\boldsymbol{x}_p$. Motivated by the recent progress in multimodal LLMs [15], [16], we propose to perform post-finetuning on f_{ψ} by leveraging the multimodal instruction tuning techniques.

Specifically, for a gene expression profile $\boldsymbol{x}_g$, we consider the pre-trained Geneformer single cell transcriptomes encoder [17], which provides the gene expression feature $\boldsymbol{z}_g = g(\boldsymbol{x}_g) \in \mathbb{R}^{d_g}$. The gene embedding features before the last Transformer layer are used in our experiments. For a discrete cell type index $\boldsymbol{x}_c$, we utilize a trainable cell type embedding module $e_{\boldsymbol{\phi}}(\cdot)$ that maps the discrete cell type index into an embedding vector $\boldsymbol{z}_c = e_{\boldsymbol{\phi}}(\boldsymbol{x}_c) \in \mathbb{R}^{d_e}$. Since both gene expression features and cell type embeddings are in different modalities compared to DNA tokens, we apply trainable modality alignment tensors $\mathbf{W}_g$ and $\mathbf{W}_e$ to convert $\boldsymbol{z}_g$ and $\boldsymbol{z}_c$ into language embedding tokens $\mathbf{H}_g \in \mathbb{R}^{l_g \times d_h}$ and $\mathbf{H}_c \in \mathbb{R}^{l_c \times d_h}$, which have the same dimensionality as the nucleotide token embedding space d_h in the DNA LLM:

$$\begin{aligned} \mathbf{H}_g &= \mathbf{W}_g \cdot \boldsymbol{z}_g \quad \text{with } \boldsymbol{z}_g = g(\boldsymbol{x}_g) \\ \mathbf{H}_c &= \mathbf{W}_c \cdot \boldsymbol{z}_c \quad \text{with } \boldsymbol{z}_c = e_{\boldsymbol{\phi}}(\boldsymbol{x}_c) \end{aligned} \tag{2}$$

where $\cdot$ represents tensor-vector product, l_g and l_c denote the token numbers of gene expression feature and cell type, receptively. Thus, we obtain a sequence of cell context tokens $\mathbf{H}_{ctx} = [\mathbf{H}_c \ \mathbf{H}_g]$. Similarly, the promoter sequence is encoded as $\mathbf{H}_p$ using the word embedding of f_{ψ}.

Consequently, given the cell context tokens $\mathbf{H}_{ctx}$ and a

pair of promoter-enhancer sequence ($\boldsymbol{x}_p$ $\boldsymbol{y}$), the loss function that maximizes the likelihood of conditional probability $p(\boldsymbol{y}|\boldsymbol{x}_c\ \boldsymbol{x}_g\ \boldsymbol{x}_p)$ is constructed as:

$$\mathcal{L} = -\sum_{i=1}^{L} \log p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_c\ \boldsymbol{x}_g\ \boldsymbol{x}_p\ \boldsymbol{y}_{<i}) \tag{3}$$

where $\boldsymbol{\theta} = \{\quad \phi\ \mathbf{W}_g\ \mathbf{W}_c\}$ is the trainable parameter set.

2.3. Training and Generation

We adopt a two-stage training strategy to help the DNA LLM better understand the multimodal input.

Stage 1: Feature Alignment Pre-training Since the multimodal context is initialized from scratch without any fine-tuning, it is not aligned with the embedding space of DNA LLM obtained during the pre-training stage on a large DNA sequence corpus. Therefore, we first focus on aligning the multimodal cell context with the DNA token embedding. In detail, we keep both the single cell transcriptomes encoder and the DNA LLM weights frozen, while minimizing the loss function in (3) with trainable parameters $\boldsymbol{\theta} = \{\phi\ \mathbf{W}_g\ \mathbf{W}_c\}$ only. This approach forces the multimodal context features $\mathbf{H}_{ctx}$ to align with the pre-trained LLM token embedding.

Stage 2: End-to-End Generation Fine-tuning In this stage, we train all trainable parameters in $\boldsymbol{\theta} = \{\quad \phi\ \mathbf{W}_g\ \mathbf{W}_c\}$ to continue optimize the loss function in (3) by updating both the DNA LLM weights and the parameters of multimodal alignment modules. It is worth to note that we always keep the single cell transcriptomes encoder weights frozen. This training stage allows the DNA LLM $f_{\boldsymbol{\psi}}$ to better incorporate the multimodal instruction into modeling the promoter-enhancer generation. Once the promoter-enhancer generation fine-tuning is complete, new enhancer sequences can be generated by sampling from the trained conditional probability function $p_{\boldsymbol{\theta}}$.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

To effectively benchmark our model, we adopted the cell type-specific promoter-enhancer pairs from the PsychENCODE consortium [18]. Using their published scATAC-seq and scRNA-seq dataset, we identified distal peak-to-gene linkages for seven major cell types: Excitatory neurons (Ex), Inhibitory neurons (In), Astrocytes (Ast), Endothelial cells (End), Microglia (Mic), Oligodendrocytes (Oli), and Oligodendrocyte precursor cells (OPC) with ArchR [19] (search distance $\pm$500k, correlation > 0.45). Assuming the distal peaks to be the enhancers, we extracted genomic sequences of the genes (promoters) and peaks (enhancers) using the reference genome hg38. In total, we adopted **>1.5M** pairs with each promoter sequence of 1,024 base pairs and each enhancer sequence 500 base pairs, separated by the cell types and promoters (see **Table 1** and **Appendix A1**[1] for details).

Table 1: The statistics of the curated multimodal promoter-enhancer generation dataset.

Cell Type	# Promoter	# Enhancer	# Training	# Test
In	16,388	48,268	203,498	3,363
OPC	15,930	38,839	170,039	3,104
Oli	16,202	59,508	254,121	4,433
End	11,770	8,505	33,392	581
Mic	15,441	63,400	301,117	5,429
Ex	17,034	87,925	353,043	5,992
Ast	15,967	48,527	194,169	3,199

3.2. Implementation Details

To investigate the significance of utilizing pre-trained DNA LLMs for the promoter-enhancer generation task, we conducted experiments using two distinct LLM backbones $f_{\boldsymbol{\psi}}(\cdot)$: HyenaDNA [11] and GPT-2 [13]. HyenaDNA, a DNA-specific LLM pre-trained on the human genome [20], is capable of generating DNA sequences at single-nucleotide resolution. For our LEONINE model, we employed a pre-trained HyenaDNA checkpoint, trained on sequences of up to 160,000 nucleotides with 14.2 million parameters, available on the Hugging Face platform[2], as the backbone for completing the generation task.

To evaluate the performance of a model without leveraging a pre-trained language model on a DNA corpus, we employed the GPT-2 architecture for conditional enhancer generation from promoters. Specifically, we adopted the original GPT-2 architecture and trained it from scratch in our LEONINE-GPT-2 model. Notably, the LEONINE-GPT-2 model, with 137 million parameters, has a substantially larger parameter size compared to the LEONINE model (14.2 million parameters). This design choice was intended to assess the impact of pre-training on DNA sequence generation.

3.3. Evaluation Protocols

We assessed LEONINE's context-specific tuning from two perspectives: (1) its ability to generate cell-type-specific enhancer sequences reflective of the cellular environment, and (2) its capacity to produce enhancers that effectively interact with a given promoter in a specific cellular context.

First, we examined the model's cell-type-specific capabilities from both biological and semantic aspects. Biologically, enhancers employ cell-type-specific TFs that recognize distinct DNA motifs to interact with promoters. Therefore, plausible generated enhancers should exhibit similar TF motifs to observed enhancers from matching cell types, differing significantly from those of unrelated cell types. To test this, we calculated the TF motif enrichment scores of the generated and observed enhancers and used these scores to compute the Jensen–Shannon (JS) divergence between matched and unmatched cell types. Next, on the semantic side, we evaluated the quality of the generated sequences using a metric inspired by the Fréchet Inception Distance (FID) [9], [21]. To this

[1]The appendix is available at https://doi.org/10.5281/zenodo.15492655.

[2]LongSafari/hyenadna-medium-160k-seqlen-hf

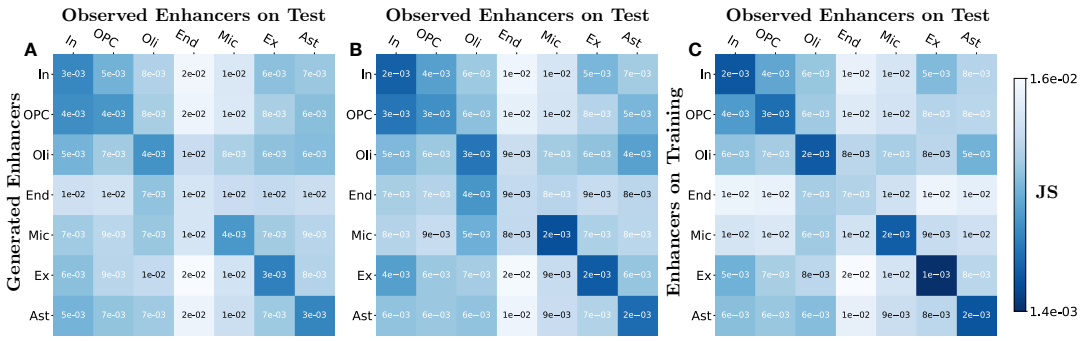

Fig. 2: Cell-type-specific heat maps generated by LEONINE-GPT-2 (**A**), LEONINE (**B**), and the observation (**C**) using the proposed multimodal cell context instruction tuning.

end, we trained seven binary classifiers, one for each cell type. For a given classifier, enhancer sequences from the target cell type in the training set were used as the positive class, while an equal number of enhancer sequences from the other six cell types were used as the negative class. After training, the hidden layer representations of the classifier served as embeddings for both generated and ground truth samples. The FID was calculated as the Wasserstein distance between two Gaussian distributions fitted to the embeddings of generated and ground truth enhancer sequences, providing a quantitative measure of generative quality. We provided details of evaluation process in **Appendix A3**.

Subsequently, we evaluated the sequence generation capabilities of LEONINE in producing enhancer sequences that interact with specified promoters. The promoters were categorized into two groups based on GC content: Low GC ($< 20\%$) and High GC ($> 80\%$). Given the substantial differences in nucleotide composition between these groups, we hypothesized that their corresponding enhancers would utilize distinct TFs with unique motifs, leading to markedly different distributions of TF enrichment scores. To assess whether LEONINE effectively captures the promoter-specific context, we computed pairwise motif enrichment score distances among the generated enhancers both within and across the groups.

3.4. Results

Multimodal Cellular Context Tuning Enables Biologically Plausible Enhancer Generation via DNA LLMs. We start with checking whether LEONINE can capture the impacts of the multimodal cellular context, including cell type information and gene expression levels. Specifically, we calculated the JS distance of generated and observed enhancers from both matched and unmatched cellular environments. As illustrated in **Fig.2**, enhancer sequences generated by both the LEONINE-GPT-2 and LEONINE models displayed pronounced cell-type specificity. For instance, the JS divergence scores for matched cell types (diagonal, $2e^{-3}$ to $9e^{-3}$) were significantly lower than those for mismatched cell types (off-diagonal, $4e^{-3}$ to $1e^{-2}$, **Fig.2 A-B**). Moreover, the LEONINE model demonstrated superior performance compared to LEONINE-GPT-2. For example, enhancer sequences generated by LEONINE exhibited notably higher similarity to observed sequences from matched cell types (**Fig. 2B-C**), resulting in more consistent patterns in the JS divergence heat map compared to LEONINE-GPT-2. These results emphasize the critical role of pretraining on extensive DNA text corpora in enhancing both the generative model's performance and its biological relevance.

Table 2: The cell-type-specific FID score generated by LEONINE and LEONINE without using multimodal cell context $\mathbf{H}_{ctx}$ instruction.

	In	OPC	Oli	End	Mic	Ex	Ast	Avg. ↓
LEONINE	**34.177**	**38.718**	**44.701**	45.970	**33.936**	**32.848**	**40.134**	**38.641**
w/o $\mathbf{H}_{ctx}$	38.840	43.325	52.227	**45.867**	48.109	44.826	48.905	46.014

To further illustrate the impact of the cellular environment, we examined the scatter plot of enrichment scores for 879 motifs derived from JASPAR 2024 databases [22]. Ideally, plausible generated enhancers should exhibit similar patterns to those observed in matched cell types. As anticipated, six of the seven cell types—excluding endothelial cells (End)—demonstrated a high concordance between the TF motif proportions of the generated and observed enhancers (R squared value > 0.95, **Fig.3A-G**). Endothelial cells, however, exhibited a slightly lower R-squared value (0.91, **Fig.3D**) compared to other cell types (0.97 to 0.98, **Fig.3A-C, E-G**). This discrepancy is likely attributable to the limited representation of this cell type in the dataset, resulting in insufficient information for accurate modeling. These findings further validate that our LEONINE model effectively captures diverse cellular contexts to generate biologically meaningful enhancer sequences.

Pre-training Is a Crucial Step for High-Quality DNA Sequence Generation. To understand the effect of pretraining on the DNA sequence generation process, we compared the cell type-specific heat maps generated by LEONINE-GPT-2 and LEONINE (i.e., **Fig.2A** vs. **Fig.2B**). It is evident that LEONINE generates better cell type-specific enhancer sequences compared to LEONINE-GPT-2, as reflected by the normalized Frobenius norms of 0.5072 (LEONINE-GPT-2 vs. Observation) and 0.062 (LEONINE vs. Observation). This

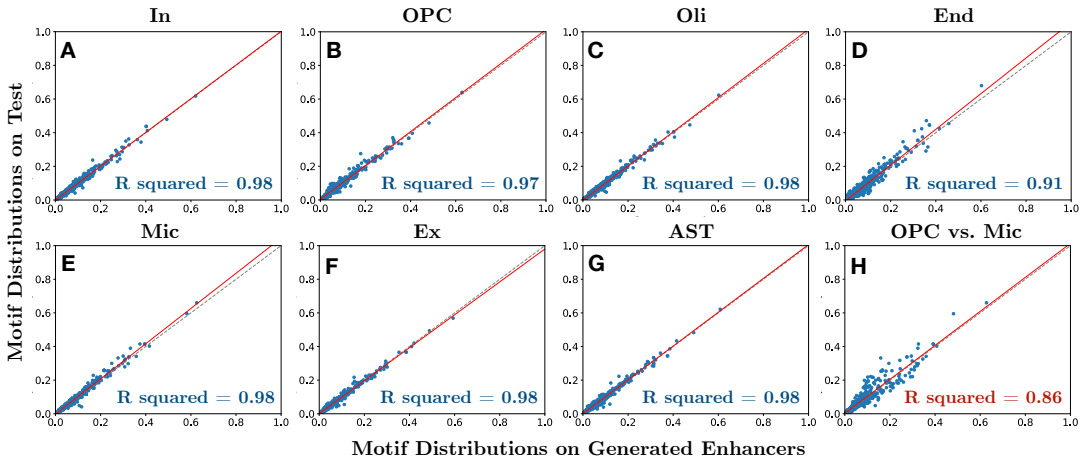

Fig. 3: (**A**-**G**) Scatter plots illustrating the distribution of TF motifs in enhancer sequences from both the observed and generated enhancer sequences of LEONINE across seven cell types. (**H**) A scatter plot illustrating the distribution of TF motifs observed in OPC compared to those generated in Mic.

is because LEONINE-GPT-2 was trained from scratch on the promoter-enhancer generation task without pre-training on an existing DNA sequence corpus. In contrast, LEONINE was pre-trained on human reference genomes before being adapted to the multimodal promoter-enhancer generation task. Although the LEONINE-GPT-2 architecture contains significantly more trainable parameters than LEONINE (i.e., 137M vs. 14.2M parameters), its generation capability is still inferior to that of the smaller model with pre-training. These findings highlight the critical importance of pre-training for DNA sequence generation.

LEONINE Allows for Semantically Informed Enhancer Generation Conditioned on Cellular Context. Next, we evaluated the cell-type specificity of the generated enhancers to further assess their ability to capture cellular context from a semantic perspective. Specifically, we calculated the FID scores of the generated enhancers after training a neural network for a cell type classification task using observed data (details provided in **Section 3.3**), as summarized in **Table 2**. Our analysis revealed that, across various conditions, the incorporation of multimodal cellular context and instruction tuning significantly reduced the FID score from 46.014 to 38.641, indicating a notable improvement in generation quality (Avg. column in **Table 2**). This trend was consistent across six of the seven cell types in the human brain dataset, with the sole exception being endothelial cells, likely due to their very limited representation in the dataset. These findings highlight the efficacy of the proposed multimodal cellular context instruction tuning strategy.

LEONINE Enables Targeted Enhancer Generation for Optimal Enhancer-Promoter Interaction. To evaluate the model's capacity to generate enhancer sequences that interact appropriately with a given promoter, we compared the results of enhancer generation under two distinct promoter conditions for each cell type. Promoter sequences within each cell type were categorized into Low GC and High GC groups, as outlined in the evaluation protocols. We then calculated the JS distance of TF motif enrichment scores both within and across these groups for each cell type. As summarized in **Table 3**, the model successfully generated enhancers with distinct TF motif patterns corresponding to the input promoter sequences for six out of seven cell types. For instance, the within-group diagonal JS distance scores were consistently smaller than the off-diagonal scores, confirming that the proposed multimodal cell context instruction tuning effectively equips DNA LLMs to model promoter-enhancer interactions at the cell-type level.

Table 3: Cell-type-specific JS distances illustrating the promoter-enhancer regulation pattern generated by LEONINE.

	In		OPC		Oli	
Group	Low	High	Low	High	Low	High
Low	**$1\ 0e^{-2}$**	$2\ 3e^{-2}$	**$6\ 1e^{-3}$**	$2\ 4e^{-2}$	**$7\ 9e^{-3}$**	$3\ 1e^{-2}$
High	$3\ 3e^{-2}$	**$9\ 3e^{-3}$**	$1\ 6e^{-2}$	**$9\ 8e^{-3}$**	$3\ 1e^{-2}$	**$7\ 3e^{-3}$**

	Mic		Ex		AST	
Group	Low	High	Low	High	Low	High
Low	**$3\ 6e^{-3}$**	$2\ 0e^{-2}$	**$5\ 4e^{-3}$**	$2\ 0e^{-2}$	**$8\ 7e^{-3}$**	$3\ 5e^{-2}$
High	$1\ 4e^{-2}$	**$5\ 7e^{-3}$**	$1\ 8e^{-2}$	**$4\ 9e^{-3}$**	$2\ 6e^{-2}$	**$1\ 1e^{-2}$**

4. CONCLUSION

In this study, we introduced a multimodal cell context instruction tuning strategy for DNA LLMs to generate enhancer sequences conditioned on promoter sequences and regulatory environments. By integrating multimodal cell context, such as cell type and gene expression, our approach addresses key limitations of prior models. Extensive experiments showed that fine-tuned DNA LLMs outperform models trained from scratch in generating promoter-enhancer pairs. This framework sets a new benchmark for cell-type-specific sequence generation, advancing genomics research and synthetic regulatory design. We hope this work inspires further exploration of DNA LLMs in multimodal contexts, bridging AI and genomics for precision medicine and synthetic biology.

5. REFERENCES

[1] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: From properties to genome-wide predictions," *Nature Reviews Genetics*, 2014.

[2] H. K. Long, S. L. Prescott, and J. Wysocka, "Ever-changing landscapes: Transcriptional enhancers in development and evolution," *Cell*, 2016.

[3] S. Schoenfelder and P. Fraser, "Long-range enhancer promoter contacts in gene expression control," *Nature Reviews Genetics*, 2019.

[4] Z. Chen, V. Snetkova, G. Bower, S. Jacinto, B. Clock, A. Dizehchi, I. Barozzi, B. J. Mannion, A. Alcaina-Caro, J. Lopez-Rios, D. E. Dickel, A. Visel, L. A. Pennacchio, and E. Z. Kvon, "Increased enhancer–promoter interactions during developmental enhancer activation in mammals," *Nature Genetics*, 2024.

[5] N. Kubo, P. B. Chen, R. Hu, Z. Ye, H. Sasaki, and B. Ren, "H3k4me1 facilitates promoter-enhancer interactions and gene activation during embryonic stem cell differentiation," *Molecular Cell*, 2024.

[6] D. Murphy, E. Salataj, D. C. Di Giammartino, J. Rodriguez-Hernaez, A. Kloetgen, V. Garg, E. Char, C. M. Uyehara, L.-s. Ee, U. Lee, M. Stadtfeld, A.-K. Hadjantonakis, A. Tsirigos, A. Polyzos, and E. Apostolou, "3d enhancer–promoter networks provide predictive features for gene expression and coregulation in early embryonic lineages," *Nature Structural & Molecular Biology*, 2023.

[7] W. Zeng, M. Wu, and R. Jiang, "Prediction of enhancer-promoter interactions via natural language processing," *BMC genomics*, 2018.

[8] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer–promoter interactions with neural network based on pre-trained dna vectors and attention mechanism," *Bioinformatics*, 2019.

[9] H. Stark, B. Jing, C. Wang, G. Corso, B. Berger, R. Barzilay, and T. Jaakkola, "Dirichlet flow matching with applications to dna sequence design," in *ICML*, 2024.

[10] K. Chen, H. Zhao, and Y. Yang, "Capturing large genomic contexts for accurately predicting enhancer-promoter interactions," *Briefings in Bioinformatics*, 2022.

[11] E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. Rabideau, and Y. Bengio, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *NeurIPS*, 2024.

[12] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie, "Sequence modeling and design from molecular to genome scale with Evo," *Science*, 2024.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[14] I. I. Taskiran, K. I. Spanier, H. Dickmänken, N. Kempynck, A. Pančíková, E. C. Ekşi, G. Hulselmans, J. N. Ismail, K. Theunis, R. Vandepoel, V. Christiaens, D. Mauduit, and S. Aerts, "Cell-type-directed design of synthetic enhancers," *Nature*, 2023.

[15] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, 2024.

[16] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv:2306.13549*, 2023.

[17] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, and X. S. Liu, "Transfer learning enables predictions in network biology," *Nature*, 2023.

[18] P. S. Emani, J. J. Liu, and D. Clarke, "Single-cell genomics and regulatory networks for 388 human brains," *Science*, 2024.

[19] J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf, "Archr is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nature Genetics*, 2021.

[20] V. Schneider and D. Church, "Genome reference consortium," *The NCBI Handbook [Internet], 2nd edn. National Center for Biotechnology Information (US), Bethesda, MD*, 2013.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, 2017.

[22] I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu, J. A. Castro-Mondragon, K. Ferenc, V. Kumar, R. B. Lemma, J. Lucas, J. Chèneby, and D. Baranasic, "Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles," *Nucleic Acids Research*, 2024.

[23] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of chip-seq (macs)," *Genome Biology*, 2008.