Appendix for Multimodal Cell Context Instruction Tuning for Conditional DNA Regulatory Sequence Generation with Large Language Models

A1. DATASET PROCESSING DETAILS

As robust and biologically meaningful promoter-enhancer pairings are crucial for model interpretation and downstream analysis, we utilized the single-cell multiome dataset of the prefrontal cortex region from the PsychENCODE consortium [19]. Following the established data processing protocol, we began by removing batch effects from the curated single-cell ATAC-seq (scATAC-seq) dataset. We then call peaks using MACS2[2] [25] 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), specifying a peak width of $L = 500$ base pairs. The resulting cis-regulatory elements (CREs) was identified as suggested and subsequently divided into promoter and distal categories based on their distance to genes.

Given the importance of robust and biologically meaningful promoter-enhancer pairings for model interpretability and downstream analyses, we utilized the single-cell multiome dataset of the prefrontal cortex region provided by the PsychENCODE consortium [19]. Adhering to the established data processing protocol, we began by correcting batch effects within the curated single-cell ATAC-seq (scATAC-seq) dataset. Subsequently, we identified peaks using MACS2 [25] 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), specifying a peak width of $L = 500$ base pairs. The resulting cis-regulatory elements (CREs) were identified following recommended guidelines and were further categorized into promoter and distal elements based on their proximity to genes.

For the distal CREs, we integrated the log-normalized single-cell RNA-seq (scRNA-seq) data from the same cohort. We applied the addPeak2GeneLinks function of ArchR [20] to identify distal peak-gene pairs associated with specific genes, using a maximum search distance of $\pm 500,000$ base pairs and a Pearson correlation cutoff of 0.45. In this study, we considered these distal peaks as enhancers and extracted their genomic sequences from the hg38 reference genome based on their coordinates. For each associated gene, we extracted the first $L = 1024$ base pairs of the promoter region using the genomic annotation of hg38.

In summary, we curated a promoter-enhancer generation dataset with $N = 7$ cell types. In total, the proposed multimodal promoter-enhancer generation dataset compromises 1,536,274 pairs of promoter-enhancer. For each cell type, there are 109,734 promoter-enhancer pairs on average. Each promoter sequence is in $L = 1024$ base pairs and its corresponding enhancer sequence with $L = 500$ base pairs. To ensure a fair evaluation, we split the dataset by separating the corpus based on cell type and corresponding promoter sequences. The detailed dataset statistics are summarized in **Table 1**.

[2] version 2.2.6

A2. EVALUATION DETAILS

A2.1. JS Divergence

For two discrete probability distributions P and Q, the JS divergence is defined as:

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}\big(P \parallel M\big) + \frac{1}{2} D_{\mathrm{KL}}\big(Q \parallel M\big), \quad \text{(A1)}$$

where $M = \frac{1}{2}\big(P+Q\big)$, and $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler (KL) divergence. In our problem setting, the discrete probability distribution P is constructed by first computing TF motif distributions for each sequence in a given comparable group (e.g., either the generated or the training set), then averaging and normalizing these motif-hit counts to form a probability distribution. Analogously, Q is built in the same manner but from the endogenous ground-truth sequences. By comparing P and Q via the JS divergence, we obtain a measure of how closely the TF motif usage patterns in generated (or training) sequences mirror those in the real, biologically observed sequences.

A2.2. FID Classifier Training

Dataset The dataset used for FID classifier training is identical to that of LEONINE. Specifically, we took the training partition from the LEONINE dataset to train the classifier, then evaluated its performance on the same evaluation set used in LEONINE. Finally, we further tested the classifier on the testing set to ensure robust generalization.

Model Each input is an enhancer sequence of 500 tokens. We first embed the raw nucleotide tokens into a 128-dimensional latent space. The embedded sequences then pass through multiple layers of CNNs and max-pooling to capture local features, which are concatenated into a shared latent representation. This local feature representation is then fed into a multi-head self-attention module (with 4 attention heads) to extract global contextual information. After the attention layer, several CNN layers and a global average pooling operation are applied to further shuffle and distill the features, producing a 128-dimensional hidden representation. We use the

hidden representation to compute the FID features. Finally, a simple MLP classifier, built on top of the 128-dimensional hidden layer, outputs the final classification result.

Training The classifier was trained for 10,000 steps with a batch size of 2,560. We used the Adam optimizer with a learning rate of 8×10^{-5}. Model performance was monitored on the evaluation set at regular intervals to tune hyperparameters and prevent overfitting.

Results Our final FID classifier achieved an accuracy of 78.9% on the evaluation set and 78.1% on the testing set across seven cell types. These results confirm that the model captures meaningful features distinguishing real enhancer sequences from generated ones, making its hidden representations suitable for reliable FID computation.