Supplementary Material (Appendix) for "Understanding Transcriptional Regulatory Redundancy by Learnable Global Subset Perturbations"

Junhao Liu[*] JUNHAO.LIU@UCI.EDU
University of California, Irvine

Siwei Xu[*] S.XU@UCI.EDU
University of California, Irvine

Dylan Riffle DYR4005@MED.CORNELL.EDU
Weill Cornell Medicine, Cornell University

Ziheng Duan ZIHEND1@UCI.EDU
University of California, Irvine

Martin Renqiang Min RENQIANG@NEC-LABS.COM
NEC Laboratories America

Jing Zhang[†] ZHANG.JING@UCI.EDU
University of California, Irvine

Editors: Vu Nguyen and Hsuan-Tien Lin

Appendix A. Relationship between ATAC-seq and RNA-seq

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) and RNA-seq (RNA sequencing) are two complementary techniques employed to profile molecule quantity within cells at different layers.

- ATAC-seq measures the accessibility of DNA within the nucleus. Typically in higher eukaryotes, DNA is in a highly compact and inaccessible state. However, certain regions of the chromosome can open up in a cell-type-specific manner to perform various functions, such as regulating transcription. ATAC-seq is a technique used to assess chromatin accessibility of individual cells, generating a sparse, binary matrix $\mathbf{x}$, where a value of 1 indicates an accessible region, and 0 indicates an inaccessible region.

- RNA-seq measures RNA molecules being transcribed in a cell at a given time, providing direct information on gene expression levels. Thus, RNA-seq data is usually represented as a numerical array $\mathbf{y}$, with each value in the array corresponding to the expression level of a specific gene.

[*] Both authors contributed equally to this work.
[†] Corresponding author.

Relationship Open chromatin regions (represented by 1s in $\mathbf{x}$) identified by ATAC-seq are often associated with cis-regulatory elements (CREs) that activate gene transcription, leading to the production of multiple mRNA molecules (non-zero values in $\mathbf{y}$). However, the specific CREs in $\mathbf{x}$ that regulate particular genes in $\mathbf{y}$ are not always known. In single-cell studies, combining ATAC-seq and RNA-seq (sometimes within a single experiment, referred to as multiomics) enables researchers to capture these CRE-to-gene regulatory relationships. This relationship can be compared to the connection between an image and its corresponding textual description in the field of multimodal machine learning—where both modalities share similar semantics but have different structures.

Embedding Space As mentioned above, ATAC-seq and RNA-seq are complementary multiomics modalities, each profiled by distinct techniques that together provide a holistic characterization of cellular states. Developing a robust embedding space enables the seamless integration of ATAC-seq and RNA-seq data, facilitating a powerful cross-modal understanding and interaction.

Appendix B. GRIDS Algorithm

The algorithm is summarized in Algorithm 1.

Algorithm 1 GRIDS global feature importance explanation algorithm for regulatory redundancy dissection

```
**Input**: cross-modality surrogate mapping model $\hat{\mathcal{F}}$
**Parameter**: global feature explanation number $L$, explanation target gene $\mathbf{y}_j$, perturbation values $\mathbf{p}$
**Output**: explanation result $\boldsymbol{r}^*$
randomly initialize the subset $\boldsymbol{r}$
while not converged do
    sample a batch of data $(\mathbf{x}, \mathbf{y}) \sim \mathcal{C}$
    doing perturbation induced by $\boldsymbol{r}$ and $\mathbf{p}$ on the input data $\mathbf{x}$
    compute the global explanation objective with Eq. 2
    estimate the indices transition $\mathbf{T}$ using Eq. 14
    update the current $\boldsymbol{r}$ using the candidates in $\mathbf{T}$ using coordinate descent
end while
set optimal result $\boldsymbol{r}^* \leftarrow \boldsymbol{r}$
return $\boldsymbol{r}^*$
```

Appendix C. Datasets

C.1. Single-Cell Multimodal Dataset

Preprocessing We curated a set of deeply-sequenced post-mortem human pre-frontal cortex (PFC) cells of a healthy individual from the PsychENCODE consortium (Akbarian et al., 2015). In total, 10,266 cells were harvested and sequenced for both chromatin accessibility (ATAC-seq) and transcription activity (RNA-seq) after applying a series of quality control parameters (ATAC-seq sequencing depth greater than 1,000, RNA-seq number of mapped genes greater than 200, and TSS enrichment greater than 2.0) and initial processing using Cell Ranger ARC (Zheng et al., 2017). In the ATAC-seq dataset, we called 127,219 characteristic chromatin regions (peaks) using Macs2 (Zhang et al., 2008) with an average sequencing depth of 4811.34, resulting in a 2-dimensional matrix of $10,266 \times 127,219$ using the R package ArchR (Granja et al., 2021). Since each chromatin region must be either opened or closed, we binarized the matrix to obtain the ATAC-seq dataset used for model training. In the RNA-seq dataset, we mapped to a total of 19,607 genes or pseudogenes for each cell, generating a 2-dimensional matrix of $10,266 \times 19,607$ with raw reads (number of reads mapped to each gene for each cell). Since RNA-seq raw reads were heavily correlated by the total number of reads per each gene, we conducted a standard normalization process using the Pegasus package (Li et al., 2020) followed by a feature selection process in which we selected the top 3,000 most deferentially expressed genes. Finally, we obtained a matrix of $10,266 \times 3,000$ as the training RNA-seq dataset.

Cell Types Furthermore, to guide the training process, we curated a set of cell type annotations using ATAC-seq and RNA-seq data separately. From the RNA-seq data, we conducted dimension reduction using PCA (number of components of 20) and clustering using LEIDEN (resolution of 1.0) (Traag et al., 2019). Using the gene expressions of the marker genes (Lake et al., 2016), we overlay the clustering and marker gene information to manually assign each cluster to a cell type. The annotation process for ATAC-seq dataset followed a similar pattern, with an extra step of transforming the ATAC-seq matrix into a gene activity matrix using ArchR. Finally, we assign all cells into one of the following cell types: excitatory neurons (Exc), inhibitory neurons (SST and VIP subtypes), astrocytes (Astro), endothelial cells (Endo), microglia cells (Micro), oligodendrocyte progenitor cells (OPC), and oligodendrocyte cells (Oligo). Note that co-assayed data is not necessarily required to train this model. As long as the ATAC-seq matrix (binarized), the RNA-seq matrix (normalized), and their corresponding cell type annotation were present, our model can be trained. The only requirement should be that the two modalities need to come from the same region (for example, the PFC region) so that the cell type annotation matches.

C.2. Marker Gene List

The full list of marker genes used in our experiments can be found in **Table 1**. For

Table 1: Marker gene list of each cell type used in the experiments.

Cell Type	Marker Genes
Astro	ALDH1A1, AQP4, GJA1
Endo	CLDN5, FLT1
Micro	APBB1IP, CX3CR1
OPC	NXPH1, OLIG1, OLIG2
Oligo	MOBP, MOG
SST	GAD1, GAD2
VIP	GAD1, GAD2

C.3. MNIST Dataset

We followed the experiment design first proposed by Shrikumar et al. (2017) to recognize two categories of digits (8 vs 3) on MNIST. At the preprocessing stage, each pixel value was normalized to be in the range of $[0, 1]$, and the full dataset was randomly split into three subsets (training, validation, and test) with a ratio of 0.7, 0.1, and 0.2, respectively. We employed a binary classifier multi-layer perceptron (MLP) model featuring rectified linear (ReLU) units, without using batch normalization. The model was trained using the Adam optimizer (Kingma, 2014), with a learning rate of $1e^{-4}$, and a batch size of 64. Training was conducted for a maximum of 50 epochs, incorporating an early stopping mechanism with a patience of 2000 steps, based on the validation set loss. This approach resulted in a test set accuracy of 97.9% in distinguishing between two-digit classes.

Appendix D. Implementation Details

Hyperparameters Our method is implemented using PyTorch. For the cross-modality surrogate mapping, we adopted four MLP layers with embedding dimension 32. The learned common latent dimension d_h is set to be 20. During the adversarial training, the weight of adversarial loss γ is set to be 0.3. The discriminator number T is set to be the number of cell types in the dataset. In the global explanation generation stage, we set the hyperparameter β to 0.1 and k to 32. We utilized the reference implementations for LIME[1], CXPlain[2], and SAGE[3], as provided by the original authors of these methods. For Saliency, SmoothGrad, and FIMAP, we developed our own implementations.

MNIST Benchmark For LIME, we followed the implementation in the paper by using random sampling to generate neighbor data for each sample. The neighbor number was set to 1024, and we used cosine distance to measure the neighbor distance. For CXPlain, we followed the setting in (Schwab and Karlen, 2019), we explained non-overlapping connected regions of 2×2 pixels for the MNIST benchmarks. Since the image dimensions were 28×28 for MNIST, the target attribution maps were of size 14×14. We used the CXPlain(U-net) model proposed by the author to learn the target attribution maps. The model was trained for 500 epochs with a batch size of 512 using Adam optimizer with a learning rate $5e^{-4}$. For SAGE, the model was trained for 300 epochs to converge with a batch size of 512 using Adam optimizer with a learning rate $1e^{-3}$. we set the permutation number to sample $10,000$ times with a batch size of 512. For GRIDS, we set the perturbation subset size to 64, and the candidate size to 32.

Single-Cell Multimodal Benchmark For LIME, we employed random sampling to generate neighboring data for each sample. We set the number of neighbors at 1024 and utilized cosine distance to measure the proximity between neighbors. In the case of CXPlain, we tried to explain non-overlapping sliding windows measuring the size of $w = 4$ peaks for the ATAC-seq (we also tried different window sizes including $w = 16, 32, 64$). Given that the ATAC sequence length is $127, 219$, the resulting target attribution maps were consequently sized at $127, 219//w$. We adopted the CXPlain (U-net) model using the one-dimensional

[1] https://github.com/marcotcr/lime
[2] https://github.com/d909b/cxplain
[3] https://github.com/iancovert/sage

convolutional neural network to learn these target attribution maps. This model underwent training for 2000 epochs, using a batch size of 512 and the Adam optimizer with a learning rate of 5×10^{-4}. For SAGE, we configured the permutation number to sample $256,000$ times, maintaining a batch size of 512. For SAGE, the model was trained for 300 epochs to converge with a batch size of 512 using Adam optimizer with a learning rate $1e^{-3}$. We tried to run SAGE with more sampling times but it still cannot coverage. For FIMAP, For GRIDS, we set the perturbation subset size to 10 and 128 depending on the experiment setting, and the candidate size to 32.

Computing Infrastructure All model training and experiments are conducted on a server equipped with an AMD EPYC 7662 64-Core Processor with 1 TB memory, 32 CPU cores, and eight NVIDIA RTX A6000 GPUs. The code is implemented in PyTorch. We use slurm as the job scheduler. For each experiment, we allocate 4 CPU cores, 1 GPU, and 90 GB memory.

Time Analysis Our GRIDS can finish global explanation generation $L = 64$ on the MNIST benchmark within 3 minutes. On the single-cell benchmark, GRIDS can finish global explanation generation for each marker-gen and cell type pair of $L = 10$ within 3 minutes, while 24 minutes for $L = 128$.

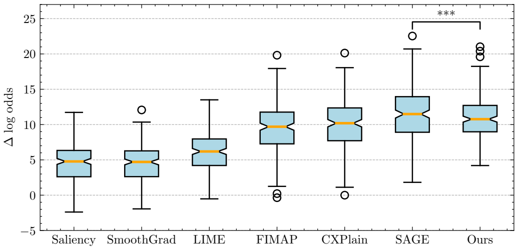

Figure 1: Benchmarks of the distributions of the log odds changes on the test dataset after masking $L = 64$ most important pixels according to different explanations learned across the training dataset. Our method is better than CXPlain with p-value < 0.004.

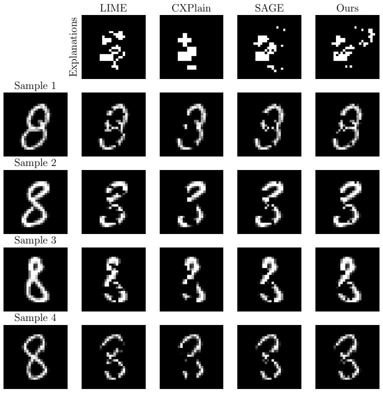

Figure 2: An examination of the most significant $L = 64$ pixels identified by various methods, is presented through their masked results on four distinct images in the test set. Our subset perturbation learning method can find a similar combinatorial pattern as SAGE.

Appendix E. Global Subset Perturbations in MNIST

To compare the global feature importance estimation performance of GRIDS to existing state-of-the-art methods, we evaluated its ability to identify the global important features in MNIST (LeCun et al., 2010) images. We followed the experiment design first proposed by Shrikumar et al. (2017), and trained binary classification models to recognize two categories of digits (8 vs 3) on MNIST. At the preprocessing stage, each pixel value was normalized to be in the range of $[0, 1]$, and the full dataset was randomly split into three subsets (training, validation, and test). Experiment details can be found in **Appendix C**

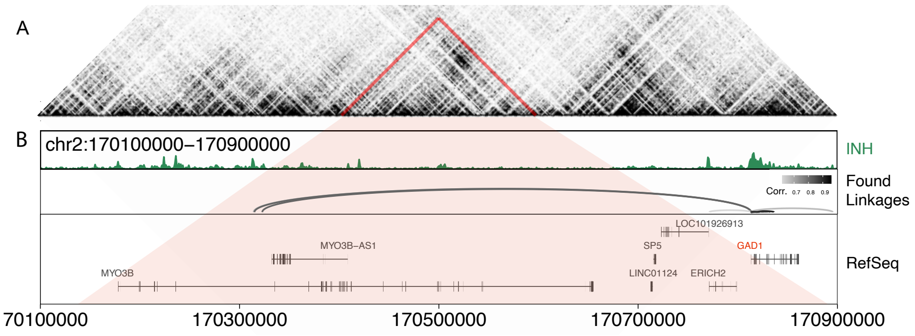

Figure 3: GRIDS effectively finds cis-regulatory elements known to associate with cell type-specific marker genes. (A) Hi-C topological domain (TAD) visualization of chromosome 2, with the specific TAD that contains GAD1 marked in red. (B) Visualization of: (top) inhibitory-specific chromatin accessibility; (center) the found enhancer-promoter linkages by GRIDS for GAD1; (bottom) the reference genome with GAD1 gene marked in red.

and **D**. After the training, the binary classification model can achieve 97.9% accuracy over the test set. We then used several global feature importance explanation methods to estimate which input pixels were most important for the classification models' predictions on the training set. The explanation methods learn to mask $L = 64$ the most important pixels. To ensure that the explanations of all methods are in the form of a subset, we select L pixels with the top L important scores as the subset. Those pixels were masked to zero (i.e. $\mathbf{p} = \mathbf{0}$) if they were selected as the global important features. To evaluate the generalization ability of the estimated globally important feature, we apply the generated explanation to the remaining test set and measured the resulting change in the confidence of classification models by checking the difference in log odds $\Delta\phi = \mathbb{E}[\phi(p(\mathbf{x})) - \phi(p(\mathbf{x}_{\setminus \boldsymbol{r}}))]$, where $\phi(p) = \log(p) - \log(1-p)$ represents the log odds function, $p(\mathbf{x})$ and $p(\mathbf{x}_{\setminus \boldsymbol{r}})$ are the classification models' category probability prediction from the original image and the masked image with the most important pixels removed, respectively. The log odds changes distribution were summarized in **Figure 1**. We created visual representations of the generated global feature importance explanations and the corresponding masked images in **Figure 2**. This was done to qualitatively evaluate the effectiveness of each method in identifying the key global features in the training images. If the estimations are precise, the resulting masked image should bear a closer resemblance to the number 3 rather than 8. This outcome is expected because the pixels that primarily contribute to the digit being recognized as an 8 ought to have been eliminated.

Appendix F. Cell-Type-Matched Hi-C Experiments

We conducted an independent validation using cell-type-matched Hi-C experiments from PsychENCODE Emani et al. (2024) to confirm the activity of our predicted CREs on the target gene GAD1 within the VIP cell type. As shown in **Figure 3**, the Hi-C experiment results encompass three key points:

- Visualization of Hi-C topological domains (TADs) on chromosome 2, with the specific TAD containing GAD1 highlighted in red.
- Chromatin accessibility data (ATAC-seq) for inhibitory-specific cell types, and
- The enhancer-promoter linkages for GAD1 identified by GRIDS. These results demonstrate that GRIDS effectively identifies cis-regulatory elements (e.g., enhancer-promoter linkages) associated with cell-type-specific marker genes, such as the GAD1 gene within the VIP cell type.

References

Schahram Akbarian, Chunyu Liu, James A Knowles, Flora M Vaccarino, Peggy J Farnham, Gregory E Crawford, Andrew E Jaffe, Dalila Pinto, Stella Dracheva, Daniel H Geschwind, and et al. The psychencode project. *Nature Neuroscience*, 18(12):1707–1712, 2015. ISSN 1097-6256. doi: 10.1038/nn.4156.

Prashant S Emani, Jason J Liu, Declan Clarke, Matthew Jensen, Jonathan Warrell, Chirag Gupta, Ran Meng, Che Yu Lee, Siwei Xu, Cagatay Dursun, et al. Single-cell genomics and regulatory networks for 388 human brains. *Science*, 384(6698):eadi5199, 2024.

Jeffrey M. Granja, M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411, 2021. ISSN 1061-4036. doi: 10.1038/s41588-021-00790-6.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Blue B. Lake, Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293): 1586–1590, June 2016. doi: 10.1126/science.aaf1204.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Bo Li, Joshua Gould, Yiming Yang, Siranush Sarkizova, Marcin Tabaka, Orr Ashenberg, Yanay Rosen, Michal Slyper, Monika S. Kowalczyk, Alexandra-Chloé Villani, and et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus rna-seq. *Nature Methods*, 17(8):793–798, 2020. ISSN 1548-7091. doi: 10.1038/s41592-020-0905-x.

Patrick Schwab and Walter Karlen. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153. PMLR, July 2017.

V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019. doi: 10.1038/s41598-019-41695-z.

Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and et al. Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

Grace X. Y. Zheng, Jessica M. Terry, and et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), January 2017. doi: 10.1038/ncomms14049.