Learnable Subset Perturbations for Understanding Transcriptional Regulatory Redundancy

Junhao Liu[1] **Siwei Xu**[1] **Dylan Riffle**[2] **Ziheng Duan**[1] **Jing Zhang**[1]
[1]University of California, Irvine [2]Cornell University
junhao.liu@uci.edu, zhang.jing@uci.edu

Abstract

Transcriptional regulation through cis-regulatory elements (CREs) is crucial for numerous biological functions, with its disruption potentially leading to various diseases. It is well-known that these CREs often exhibit redundancy, allowing them to compensate for each other in response to external disturbances, highlighting the need for methods to identify CRE sets that collaboratively regulate gene expression effectively. To address this, we introduce GRIDS, an in silico computational method that approaches the task as a global feature explanation challenge to dissect combinatorial CRE effects in two phases. First, GRIDS constructs a differentiable surrogate function to mirror the complex gene regulatory process, facilitating cross-translations in single-cell modalities. It then employs learnable perturbations within a state transition framework to offer global explanations, efficiently navigating the combinatorial feature landscape. Through comprehensive benchmarks, GRIDS demonstrates superior explanatory capabilities compared to other leading methods. Moreover, GRIDS's global explanations reveal intricate regulatory redundancy across cell types and states, underscoring its potential to advance our understanding of cellular regulation in biological research.[1]

1 Introduction

Transcriptional regulation via cis-regulatory elements (CREs) is essential to maintaining cell identity, responding to intra- and extra-cellular signals, and coordinating gene activities, whereas its dysregulation can cause a broad range of diseases [Hoch et al., 1990]. Unfortunately, after decades of CRE identification efforts, it is still challenging to directly validate single CREs' impacts at either intermediate (e.g., gene expression) or clinical phenotype level [Hong et al., 2008, Barolo, 2012]. Most recent research has found that multiple CREs may target a specific gene to drive overlapping spatiotemporal expression patterns, so if one CRE is damaged, another can step in to fulfill appropriate functions [Kassis, 1990]. Such combinatorial CRE effects, usually referred to as **regulatory redundancy**, widely exist in most genomes as a regulation buffer to provide phenotypic robustness [Kvon et al., 2021]. Existing research has primarily focused on using computational methods to uncover individual CRE-to-gene regulatory effects, while combinatorial regulatory redundancy remains largely unknown due to the complexities in accounting for interactions among CREs. In this work, We present a computational method to identify combinatorial regulatory redundancy at the single-cell level by linking it to global feature explanations of black-box models.

To address this challenge, our initial step involves developing a black-box model to predict gene expressions from CREs, framing the task of identifying multi-CRE-to-gene relationships as a feature importance explanation problem [Sood and Craven, 2022]. Methods for this task fall into two categories: local and global feature importance. Local methods, like LIME [Ribeiro et al., 2016] and

[1]The source code is available at `https://github.com/jhliu17/nnpert`.

D3S3: Data-driven and Differentiable Simulations, Surrogates, and Solvers @ NeurIPS 2024.

SHAP [Lundberg and Lee, 2017], explain individual predictions by identifying important features. Some approaches use differentiable surrogate models to generate explanations through minimal adversarial perturbations [Chapman-Rounds et al., 2021], or jointly train the surrogate model and generate explanations [Chen et al., 2018]. However, for generalizable regulatory insights across diverse cells, global feature importance explanations are crucial [Doshi-Velez and Kim, 2017, Ibrahim et al., 2019]. While various methods have been developed, many simplify combinatorial feature effects as additive, overlooking complex interactions between CREs. Approaches like CXPlain [Schwab and Karlen, 2019] and SHAP [Lundberg and Lee, 2017] use feature perturbation or sampling-based methods, but these struggle in high-dimensional feature spaces common in single-cell multi-modal data [Sood and Craven, 2022].

In this work, we introduce GRIDS, a global feature explanation method for efficient regulatory redundancy analysis using single-cell multi-modal data. GRIDS consists of two key components: a differentiable cross-modality surrogate mapping and a global explanation method using learnable subset perturbations. First, the surrogate mapping learns and aligns modality-specific cell representations in a common space through adversarial training. Next, GRIDS employs a learnable subset perturbation technique with a state transition model to identify globally important features that significantly affect gene expression. This approach, distinct from traditional additive or sampling-based methods, directly perturbs input CRE modality while leveraging auto-differentiation, enabling precise and efficient global feature importance explanations for large biological datasets.

2 Methodology

As defined in Wu et al. [2021], the CRE is represented by the ATAC-seq vector $\mathbf{x} \in \{0,1\}^{d_a}$, where each dimension indicates a peak state in chromosomes—"1" for open and "0" for closed. ATAC-seq data is typically high-dimensional, with $d_a > 10^5$. The gene expression values regulated by the CRE are denoted by a real-valued vector $\mathbf{y} \in \mathbb{R}^{d_r}$, where d_a and d_r represent the number of peaks and genes, respectively. A single-cell multi-omics dataset consists of N single-cell multi-modal data points $\mathcal{C} = \{\boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(N)}\}$, where each cell $\boldsymbol{c}^{(i)} = (\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ includes an ATAC-seq vector and its corresponding RNA-seq vector, along with a semantic label $\ell^{(i)}$ indicating its cell type among T classes. The gene expression level, controlled by CREs through complex biological processes, can be modeled as a regulatory function $\mathbf{y} = \mathcal{F}(\mathbf{x})$, where $\mathcal{F}(\mathbf{x}) : \mathbb{R}^{d_a} \to \mathbb{R}^{d_r}$. Due to experimental costs, querying this black-box regulatory function $\mathcal{F}$ is challenging. The task of regulatory redundancy dissection involves identifying a subset of L peak indices $\boldsymbol{r} = \{r_1, \ldots, r_L\}$ within the CRE (i.e., subset of features in the ATAC-seq $\mathbf{x}_{\boldsymbol{r}} \equiv \{\mathbf{x}_j | j \in \boldsymbol{r}\}$) that are crucial for regulating gene expression across a cell population. Given that $\mathbf{x}$ typically has over 10^5 dimensions, this creates a vast search space for the subset $\boldsymbol{r}$ within the binomial combination $\binom{d_a}{L}$.

Global Feature Explanations for Regulatory Redundancy Dissection To resolve the regulatory redundancy dissection problem, we propose an in silico computational method by modeling it within a global feature explanation framework. Conventionally, global explanation is defined by how much a model's performance degrades over an observed population of samples when features are removed [Chapman-Rounds et al., 2021]. In the context of regulatory redundancy, the global explanation objective can be expressed as

$$\boldsymbol{r}^* = \underset{\boldsymbol{r}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbf{c} \sim \mathcal{C}}[\mathcal{L}(\mathcal{F}(\mathbf{x}_{\setminus \boldsymbol{r}}), \mathbf{y})] \tag{1}$$

where $\mathcal{L}$ is a loss measurement for expected gene expression degradation. $\mathbf{x}_{\setminus \boldsymbol{r}}$ denotes the perturbed CREs induced by $\boldsymbol{r}$, replacing the original feature $\mathbf{x}_{\boldsymbol{r}}$ with preset perturbation values $\mathbf{p} \in \mathbb{R}^{d_a}$ at indices indicated by $\boldsymbol{r}$ (i.e., $\mathbf{x}_{\setminus \boldsymbol{r}, r_j} = \mathbf{p}_{r_j}$). If $\mathbf{p} = \mathbf{0}$, this equates to removal-based perturbation. The choice of loss function depends on the model and task [Covert et al., 2020]. The optimal subset $\boldsymbol{r}^*$ in Eq. 1 is the solution to the regulatory redundancy problem defined above.

However, the regulatory function $\mathcal{F}$ is a black box and inefficient to query, which means that even model-agnostic explanation methods can be intractable in this setting. Therefore, we further define a surrogate $\hat{\mathcal{F}}(\mathbf{x}; \theta_f) : \mathbb{R}^{d_a} \to \mathbb{R}^{d_r}$, which is a neural network trained to be a differentiable approximation of $\mathcal{F}$ using the collected single-cell multi-modal data $\mathcal{C}$. Substituting $\hat{\mathcal{F}}(\mathbf{x}; \theta_f)$ for $\mathcal{F}$ in Eq. 1 yields a tractable objective. We elaborate the details of training differentiable cross-modality surrogate mapping $\hat{\mathcal{F}}$ in Appendix A. Given the differentiable surrogate $\hat{\mathcal{F}}$ and the removal

perturbation $\mathbf{p} = \mathbf{0}$, we define the loss measurement of expected gene expression degradation perturbed by a given perturbation subset $\boldsymbol{r}$ as:

$$\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}}), \mathbf{y}) = (\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}})_i / \mathbf{y}_i)^2 + \frac{\beta}{d_r - 1} \sum_{j=1, j \neq i}^{d_r} (\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}})_j / \mathbf{y}_j - 1)^2 \tag{2}$$

where i is the target gene index, β is a hyperparameter used to guide the learned perturbation $\boldsymbol{r}$ to be independent of non-target genes, and d_r is the total number of genes.

Learning Optimal Subset Perturbations $\boldsymbol{r}^*$ To work around the non-differentiable replacement operation in $\mathbf{x}_{\backslash \boldsymbol{r}, r_j} = \mathbf{p}_{r_j}$ with $r_j \in \boldsymbol{r}$, we observe that replacement with any perturbation values $\mathbf{p}$ can be unified as follows $\mathbf{x}_{\backslash \boldsymbol{r}, j} = \mathbf{x}_j + \mathbb{1}[j \in \boldsymbol{r}](\mathbf{p}_j - \mathbf{x}_j)$, where the jth dimensional feature is replaced by the perturbation value $\mathbf{p}_j$ if j is in the subset $\boldsymbol{r}$; otherwise, it retains its original value. Given a randomly initialized subset $\boldsymbol{r}$ and the global explanation objective in Eq. 1, the objective gradient with respect to the category embedding (or the perturbed feature if the input is continuous) can be easily computed through any automatic differentiation framework. This is represented as

$$\mathbf{G} = \partial \mathbb{E}_{\mathbf{c} \sim \mathcal{C}}[\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}}), \mathbf{y})] / \partial \mathbf{W}^a_{\text{Emb}}(\mathbf{x}_{\backslash \boldsymbol{r}}) \tag{3}$$

where $\mathbf{G} \in \mathbb{R}^{d_a \times d_h}$. Based on the gradient information of $\mathbf{G}$, we update the current global important subset $\boldsymbol{r}$ by constructing a state transition matrix of indices $\mathbf{T} \in \mathbb{R}^{L \times d_a}$, where each entry $\mathbf{T}_{i,j}$ in the matrix represents the advantage value of transitioning from replacing the previous index r_i with the new index j. The state transition matrix can be approximated by considering the objective gradient $\mathbf{G}$ and the replaced perturbations $\mathbf{W}^a_{\text{Emb}}(\mathbf{p}) - \mathbf{W}^a_{\text{Emb}}(\mathbf{x})$

$$\mathbf{d}_j = \mathbf{G}_j \cdot (\mathbf{W}^a_{\text{Emb}}(\mathbf{p})_j - \mathbf{W}^a_{\text{Emb}}(\mathbf{x})_j), \;\; \mathbf{T}_{i,j} = \mathbb{1}[j \notin \boldsymbol{r}]\mathbf{d}_j - \mathbb{1}[j \neq r_i]\mathbf{d}_{r_i} \tag{4}$$

where $\mathbf{d}_j \in \mathbb{R}^{d_a}$ represents the approximated objective descent value estimated for applying the potential perturbation $\mathbf{p}_j$. Given the estimated state transition matrix of indices $\mathbf{T}$, we use the coordinate descent method to update the global feature subset $\boldsymbol{r}$, which means we iteratively update each index in $\boldsymbol{r}$ by selecting the candidate indices with the top-k advantage values of the corresponding row in $\mathbf{T}$. In practice, we have found that the coordinate descent method achieves a good balance between convergence speed and explanation performance. As a result, the randomly initialized perturbation subset $\boldsymbol{r}$ can be effectively learned, leading to the optimal solution $\boldsymbol{r}^*$, through a batch iteration manner. The overall algorithm is summarized in Algorithm 1.

3 Experiments and Results

Single-Cell Multimodal Dataset and Baselines We curated a set of deeply-sequenced single-cell multi-modal data from postmortem human PFC [Akbarian et al., 2015]. In total, $N = 10,266$ cells with $T = 8$ different cell types were harvested and sequenced for both chromatin accessibility (ATAC-seq) and transcription activity (RNA-seq). On the ATAC-seq side, we called $d_a = 127,219$ peaks using Macs2 [Zhang et al., 2008] with an average sequencing depth (i.e. the number of open state) of 4811.34. For the RNA-seq, we conducted standard quality control and pre-processing using the default parameters recommended by Pegasus [Li et al., 2020]. The gene number d_r is 3000. We test GRIDS to generate different subset size of global important features $\boldsymbol{r}$ sequence lengths L on multiple target genes by do perturbation in the CRE input using masking $\mathbf{p} = \mathbf{0}$. In each experiment, the full dataset was randomly split into three subsets (training, validation, and test) with the ratio of 0.7, 0.1, and 0.2, respectively. The global explanations were learned in the training set and then evaluated its performance on the test set. We compared GRIDS against several feature importance explanation methods, including global and local. Details about dataset processing and implementation can be found in Appendix C and Appendix D.

The Surrogate Model Accurately Models the ATAC-to-RNA Relationship Since GRIDS relies on a differentiable surrogate model $\hat{\mathcal{F}}$ for ATAC-to-RNA translation, we first assessed the model's accuracy on a single-cell multimodal dataset. We selected marker genes from a previous study Lake et al. [2016] and compared the mean expressions between cell types and between the observed and translated cohorts (Fig. 1A). Marker genes, which are highly indicative of each cell type, served as category labels in this evaluation. Focusing on key marker genes, the UMAP showed consistent findings, with the translated expression highlighting these cell types and high correlations between observed and translated data (Fig. 1B).

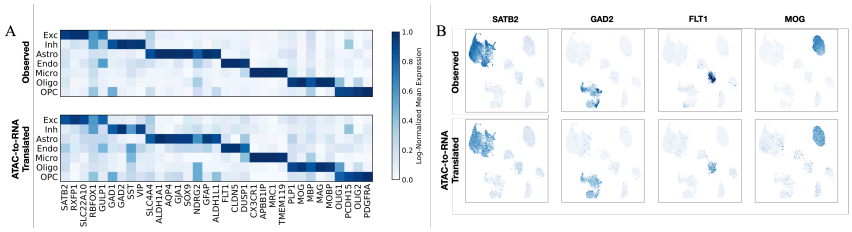

Figure 1: The trained differentiable surrogate model $\hat{\mathcal{F}}$ can accurately predict the RNA-seq modality from given single-cell ATAC-seq profiles. (A) The comparison of predicted marker gene expression with actual values across different cell types demonstrated high consistency and specificity to cell type (mean R^2=0.914). (B) The UMAP of real scRNA-seq data, colored according to both actual and predicted expression levels for marker genes, exhibited a strong similarity.

Table 1: Benchmark results by comparing expression drops of marker genes across all cell types (upper: $L = 10$, bottom: $L = 128$).

Cell Type	Random		Saliency		SmoothGrad		FIMAP		GRIDS	
	Avg. Δ	Rel. Δ(%)	Avg. Δ	Rel. Δ(%)	Avg. Δ	Rel. Δ(%)	Avg. Δ	Rel. Δ(%)	Avg. Δ	Rel. Δ(%)
Astro	−0.085	−0.015	−2.163	−0.601	−2.155	−0.621	−13.502	−4.254	**−16.696**	**−5.837**
Endo	−1.073	−0.138	−4.974	−0.372	−9.726	−0.995	−38.997	−9.303	**−57.477**	**−11.816**
Micro	−0.012	−0.026	−23.757	−1.545	−32.944	−2.083	−73.752	−6.248	**−90.607**	**−7.671**
OPC	+0.823	−0.087	−54.645	−2.338	−48.438	−2.067	−77.167	−6.260	**−96.661**	**−8.256**
Oligo	−0.058	+0.026	−0.558	−0.173	−0.939	−0.220	−10.917	−4.252	**−16.760**	**−6.896**
SST	+0.159	+0.080	−5.201	−2.006	−5.201	−2.006	−16.453	−5.660	**−17.677**	**−6.365**
VIP	+0.012	+0.001	−0.654	−1.189	−0.634	−1.160	−2.732	−3.797	**−6.804**	**−7.195**
Avg.	+0.016	−0.021	−12.988	−1.209	−13.519	−1.290	−30.268	−5.367	**−39.103**	**−7.300**
Astro	−1.793	−0.533	−15.511	−4.853	−18.505	−6.217	−82.565	−24.766	**−100.556**	**−34.633**
Endo	+2.554	+0.468	−46.160	−6.217	−52.383	−7.893	−252.338	−41.790	**−259.920**	**−44.601**
Micro	−9.091	−0.490	−131.512	−9.122	−145.561	−10.116	−451.210	−39.695	**−470.430**	**−44.114**
OPC	−1.848	−0.165	−193.739	−10.260	−186.235	−9.891	**−415.231**	−35.687	−392.326	**−36.380**
Oligo	−1.134	−0.211	−19.809	−6.382	−21.136	−7.630	−69.460	−28.175	**−93.518**	**−38.982**
SST	−1.681	−0.615	−33.589	−11.675	−32.275	−11.115	−86.191	−29.198	**−93.772**	**−33.708**
VIP	+0.071	+0.002	−4.014	−4.876	−3.872	−4.782	−13.054	−16.757	**−19.703**	**−27.221**
Avg.	−1.843	−0.237	−68.620	−7.618	−70.292	−8.212	−202.368	−30.787	**−209.583**	**−36.893**

Evaluation of Regulatory Redundancy with Learnable Perturbations To verify the effectiveness and robustness of the explanation process, we benchmarked GRIDS with various baselines on focused marker genes of 7 cell types. The full list of marker gene of each cell type can be found in the Appendix C.2. We use two metrics to evaluate each method's effectiveness in masking L CRE features to suppress a target gene's expression, including the averaged expression change (Avg. Δ) [Han et al., 2020] and the ratio of expression change against the original value (Rel. Δ). We summarize our benchmarking results in **Table 1**. In our experiments, we observed that LIME, CXPlain, and SAGE failed to provide global explanations for the high-dimensional ATAC-seq data. Due to the curse of dimensionality, both LIME and CXPlain failed to generate reasonable explanations in ATAC-seq, which means using a simple model (K-Lasso in LIME) or a regular masking strategy (sliding window in CXPlain) to capture the additive effect of the important feature might be infeasible in the vast dimension space. Meanwhile, the SAGE method could not converge within a reasonable time frame, since it randomly samples the subset from the vast combinatorial feature space and then evaluates the expected performance degradation. This strategy is equivalent to the importance sampling method, which has the problem of high variance and weight degeneracy in high-dimensional spaces. GRIDS consistently outperforms all baselines across each cell type by introducing larger marker gene expression degradation.

4 Conclusions

In this paper, we present GRIDS, a global feature importance explanation method designed to dissect complex multi-CRE-to-gene regulatory redundancy using single-cell multi-modal data. GRIDS first employs cross-modality surrogate mapping to approximate the black-box regulatory function, unifying the regulatory redundancy problem with global feature importance explanations. It also introduces a subset perturbation learning framework to efficiently generate global feature importance subsets. Our method, applicable across various data modalities, outperforms state-of-the-art baselines on single-cell data. Moreover, cross-cell type and regional analysis demonstrate GRIDS's ability to characterize cell-type-specific regulatory redundancy, offering valuable insights for experimental validations in biological research.

References

M. Hoch, C. Schroder, E. Seifert, and H. Jackle. cis-acting control elements for kruppel expression in the drosophila embryo. *EMBO J*, 9(8):2587–95, 1990. ISSN 0261-4189 (Print) 1460-2075 (Electronic) 0261-4189 (Linking). doi: 10.1002/j.1460-2075.1990.tb07440.x. URL https://www.ncbi.nlm.nih.gov/pubmed/2114978.

J. W. Hong, D. A. Hendrix, and M. S. Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, 2008. ISSN 1095-9203 (Electronic) 0036-8075 (Print) 0036-8075 (Linking). doi: 10.1126/science.1160631. URL https://www.ncbi.nlm.nih.gov/pubmed/18772429.

S. Barolo. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays*, 34(2):135–41, 2012. ISSN 1521-1878 (Electronic) 0265-9247 (Print) 0265-9247 (Linking). doi: 10.1002/bies.201100121. URL https://www.ncbi.nlm.nih.gov/pubmed/22083793.

J. A. Kassis. Spatial and temporal control elements of the drosophila engrailed gene. *Genes Dev*, 4 (3):433–43, 1990. ISSN 0890-9369 (Print) 0890-9369 (Linking). doi: 10.1101/gad.4.3.433. URL https://www.ncbi.nlm.nih.gov/pubmed/2110923.

Evgeny Z. Kvon, Rachel Waymack, Mario Gad, and Zeba Wunderlich. Enhancer redundancy in development and disease. *Nature Reviews Genetics*, 22(5):324–336, 2021. ISSN 1471-0056. doi: 10.1038/s41576-020-00311-x.

Akshay Sood and Mark Craven. Feature Importance Explanations for Temporal Black-Box Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8351–8360, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i8.20810.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Matt Chapman-Rounds, Umang Bhatt, Erik Pazos, Marc-Andre Schulz, and Konstantinos Georgatzis. FIMAP: Feature Importance by Minimal Adversarial Perturbation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11433–11441, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i13.17362.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892. PMLR, July 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 279–287, 2019.

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.

Kevin E. Wu, Kathryn E. Yost, Howard Y. Chang, and James Zou. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021. ISSN 0027-8424. doi: 10.1073/pnas.2023070118.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding Global Feature Contributions With Additive Importance Measures. In *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223. Curran Associates, Inc., 2020.

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. URL `https://dx.doi.org/10.1038/nn.4156`.

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. URL `https://dx.doi.org/10.1186/gb-2008-9-9-r137`.

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.

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. URL `https://doi.org/10.1126/science.aaf1204`.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.492.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

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. URL `https://doi.org/10.1038/ncomms14049`.

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.

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. URL `https://doi.org/10.1038/s41598-019-41695-z`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Chanan, and et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise, June 2017.

A Cross-Modality Surrogate Mapping

Recalling the collection of single-cell multi-modal data $\mathcal{C}$, we can train the surrogate model $\hat{\mathcal{F}}$ by mapping RNA and ATAC modalities into the same embedding space $\mathcal{E}$. The advantage of using the embedding model is that it allows us to easily extend our surrogate model, which learns from paired RNA and ATAC data with known paired cell type labels, to unpaired data without prior knowledge of cell type class labels. We adopt two autoencoders to model the modality-specific feature. For ATAC-seq, each dimension in $\mathbf{x}$ is considered a binary categorical feature, with one low-dimensional embedding for each category. The encoder projects the raw input into semantics features as

$$\mathbf{h}_a^{(i)} = f_{\text{Enc}}^a(\mathbf{W}_{\text{Emb}}^a(\mathbf{x}^{(i)})),\ \mathbf{h}_r^{(i)} = f_{\text{Enc}}^r(\mathbf{W}_{\text{Emb}}^r(\mathbf{y}^{(i)})) \tag{5}$$

where $\mathbf{W}_{\text{Emb}}^a \in \mathbb{R}^{d_h \times d_a}$ is a category embedding module to accommodate the high-dimensional ATAC-seq data, $\mathbf{W}_{\text{Emb}}^r \in \mathbb{R}^{d_h \times d_r}$ is an embedding matrix for RNA-seq, f_{Enc}^a and f_{Enc}^r are encoder networks to generate embeddings $\mathbf{h}_a, \mathbf{h}_r \in \mathbb{R}^{d_h}$ in $\mathcal{E}$ of dimension d_h. The decoder generates reconstructions via $\hat{\mathbf{x}}^{(i)} = f_{\text{Dec}}^a(\mathbf{h}_a^{(i)})$, $\hat{\mathbf{y}}^{(i)} = f_{\text{Dec}}^r(\mathbf{h}_r^{(i)})$, where f_{Dec}^a and f_{Dec}^r are two decoder networks for the two modalities, $\hat{\mathbf{x}}^{(i)}$ and $\hat{\mathbf{y}}^{(i)}$ represent the reconstructions with objective defined as

$$\mathcal{L}_{\text{Rec}} = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{C}}[\text{BCE}(\hat{\mathbf{x}}^{(i)}, \mathbf{x}^{(i)}) + \text{MSE}(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)})] \tag{6}$$

where BCE is the binary cross-entropy loss, and MSE is the mean-squared error.

Alignment Embedding Adversarial Training Given a cell type $T = k$, we define $\mathcal{C}^k$ as a subset of $\mathcal{C}$, where each cell $\boldsymbol{c}^{(i)} \in \mathcal{C}^k$ has the same label $\ell^{(i)} = k$. To align the modality-specific embeddings and capture the regulatory regulations between them, two mapping layers are adopted to jointly align the two modalities

$$\tilde{\mathbf{h}}_r^{(i)} = f_{\text{AR}}(\mathbf{h}_a^{(i)}),\ \tilde{\mathbf{h}}_a^{(i)} = f_{\text{RA}}(\mathbf{h}_r^{(i)}) \tag{7}$$

where f_{AR} aims to map the ATAC embeddings to the RNA embeddings and f_{RA} does the opposite. We use a generative adversarial training mechanism [Arjovsky et al., 2017, Goodfellow et al., 2014] to let both encoders and mapping layers act as two generators to learn the modality-agnostic latent space $\mathcal{E}$. And then we apply the discriminator D_a^k in each cell type k for binary classification, aiming to differentiate whether $\mathbf{h}_a$ and $\tilde{\mathbf{h}}_a$ of the ATAC embedding belongs to the cell type k or not. The D_r^k does the similar operation for the RNA embeddings $\mathbf{h}_r$ and $\tilde{\mathbf{h}}_r$. Then, the discrimination loss can be formulated as

$$\begin{aligned}\mathcal{L}_{\text{Dis}}^k =& \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[\log D_a^k(\mathbf{h}_a)] + \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[\log(1 - D_a^k(\tilde{\mathbf{h}}_a))] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[\log D_r^k(\mathbf{h}_r)] + \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[\log(1 - D_r^k(\tilde{\mathbf{h}}_r))].\end{aligned} \tag{8}$$

The generators are trained to simultaneously fool the discriminator and keep the cycle consistency [Zhu et al., 2020]

$$\begin{aligned}\mathcal{L}_{\text{Gen}}^k =& \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[-\log D_r^k(\tilde{\mathbf{h}}_r) + \text{MSE}(f_{\text{RA}}(\tilde{\mathbf{h}}_r), \mathbf{h}_a)] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[-\log D_a^k(\tilde{\mathbf{h}}_a) + \text{MSE}(f_{\text{AR}}(\tilde{\mathbf{h}}_a), \mathbf{h}_r)].\end{aligned} \tag{9}$$

Therefore, the adversarial training process can be summarized in the following objective function

$$\mathcal{L}_{\text{Adv}} = \min_{\theta_{\text{Gen}}} \max_{\theta_{\text{Dis}}} \mathbb{E}_{k\sim T}[\mathcal{L}_{\text{Gen}}^k + \mathcal{L}_{\text{Dis}}^k] \tag{10}$$

where θ_{Gen} is the trainable parameters of encoders $f_{\text{Enc}}^r, f_{\text{Enc}}^a$ and the cross-mapping layers $f_{\text{AR}}, f_{\text{RA}}$, θ_{Dis} collects parameters of all T pairs of discriminators D_a^k, D_r^k. The overall objective of the surrogate $\hat{\mathcal{F}}$ is

$$\mathcal{L}_{\text{Int}} = \mathcal{L}_{\text{Rec}} + \gamma\mathcal{L}_{\text{Adv}} \tag{11}$$

where γ is a hyperparameter to weigh the adversarial loss. After the training, the surrogate $\hat{\mathcal{F}}(\mathbf{x}; \theta_f)$ is defined as

$$\hat{\mathcal{F}}(\mathbf{x}; \theta_f) = f_{\text{Dec}}^r(f_{\text{AR}}(f_{\text{Enc}}^a(\mathbf{W}_{\text{Emb}}^a(\mathbf{x})))). \tag{12}$$

B GRIDS Algorithm

The algorithm is summarized in Algorithm 1. Given the estimated state transition matrix of indices $\mathbf{T}$, we iteratively update each index in $\boldsymbol{r}$ by selecting the candidate indices with the top-k advantage values of the corresponding row in $\mathbf{T}$. We further evaluate the best index among these k candidates by assessing which index can make the updated global feature subset $\boldsymbol{r}'$ most significantly decrease the global explanation objective $\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\setminus \boldsymbol{r}'}), \mathbf{y})$. Our global explanation method enables the learning of the global feature combinatorial subset $\boldsymbol{r}^*$ using gradient guidance, rather than relying on random sampling [Covert et al., 2020]. Our experiments prove it to be more efficient and converges more quickly to find the optimal $\boldsymbol{r}^*$ in a high-dimensional space. It facilitates the efficient generation of global explanations in high-throughput biological data, such as the ATAC-seq ($d_a > 10^4$).

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. 1
  estimate the indices transition $\mathbf{T}$ using Eq. 4
  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 $\boldsymbol{r}^*$

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 2. For

Table 2: 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

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. We utilized the reference implementations for LIME[2], CXPlain[3], and SAGE[4], as provided by the original authors of these methods. For Saliency, SmoothGrad, and FIMAP, we developed our own implementations.

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 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 [Paszke et al., 2019]. We use slurm as the job scheduler. For each experiment, we allocate 4 CPU cores, 1 GPU, and 90 GB memory.

[2]https://github.com/marcotcr/lime

[3]https://github.com/d909b/cxplain

[4]https://github.com/iancovert/sage

Baseline Comparisons We compared GRIDS with against several feature importance explanation methods, including global and local: (1) **Random**, a naive baseline that randomly selects global important features to perturb the model input. (2) **Saliency** [Simonyan et al., 2014], a widely used model interpretation method utilizing the gradient information w.r.t the input feature to select the most effective ones. We aggregate local feature importance scores to generate global ones. (3) **LIME** [Ribeiro et al., 2016], a local explanation method. It uses the submodular pick algorithm to convert local feature importance scores into global ones. (4) **SmoothGrad** [Smilkov et al., 2017], a method commonly used in computer vision which samples noise to generate neighbor samples and evaluate global feature importance via the average gradient saliency map, (5) **FIMAP** [Chapman-Rounds et al., 2021], a neural network based approach that learns the feature importance through finding minimal adversarial perturbation. (6) **CXPlain** [Schwab and Karlen, 2019], a global approach that involves training a surrogate model for explanations. This method perturbs features with perturbation values to determine their importance scores. (7) **SAGE** [Covert et al., 2020], extends the SHAP method [Lundberg and Lee, 2017] to offer global explanations based on approximated Shapley values by sampling important subsets.