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

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.

Keywords: Single-Cell Multi-Omics, Transcriptional Regulatory Factor

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

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 an in silico computational method for identifying combinatorial regulatory redundancy at the single-cell level by connecting this problem with global feature importance explanations of black-box models.

Most existing computational and experimental methods for dissecting regulatory redundancy are designed to analyze individual CRE-to-gene effects at the tissue-level sequencing assays, rather than the combinatorial effects of multiple CREs using single-cell sequencing assays. For instance, several pioneer studies showed that seemingly redundant CREs (e.g., shadow enhancers) precise modulate gene expression during development and improve phenotypic robustness to physiological or genetic stress (Frankel et al., 2010). However, such studies mainly focused on CRE redundancy using single-locus approaches and reporter genes, lacking genome-wide insights (Perry et al., 2011). Technological advancements and efforts by consortia like ENCODE led to the identification of millions of CREs across various tissues (Luo et al., 2020), with large-scale transgenic reporter assays characterizing their in vivo activities (Manning et al., 2012). However, such tissue-level sequencing approaches overlooked regulatory heterogeneity across cell states and populations. Recently, the single-cell sequencing revolution, particularly the multi-modal genomic profiling, enables finer resolution analysis of transcriptional regulation at the cellular level. Computational methods developed for single-cell data have been successful in elucidating complex, cell type-specific regulatory mechanisms (Granja et al., 2021; Zhang et al., 2022). However, these approaches still focus on individual CRE-to-gene relationships, neglecting synergistic effects across multiple CREs.

To address this challenge, our initial step involves developing a black-box model capable of predicting gene expressions from a set of CREs. This leads us to naturally consider the task of identifying multi-CRE-to-gene relationships as a fundamental **feature importance explanation** task (Sood and Craven, 2022). In this context, various methods have been proposed, falling into two main categories: *local* feature importance and *global* feature importance. Most local feature importance explanation methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), focused on explaining individual predictions based on important features (instance-wise feature importance). Recent work also tried to address the black-box model explanations using a differentiable surrogate model

via generating explanations through learning minimal adversarial perturbations Chapman-Rounds et al. (2021). In addition, (Chen et al., 2018) developed a method to jointly train the surrogate model and generate explanations. However, identifying *generalizable* combinatorial effects of CREs is crucial for providing regulatory insights applicable across a broad spectrum of cells, a task best approached through global feature importance explanations (Doshi-Velez and Kim, 2017; Ibrahim et al., 2019). While numerous methods have been developed, they simplified the challenge of explaining combinatorial features as additive importance metrics, failing to account for complex, nonlinear interactions between CREs. Later, Schwab and Karlen (2019) and Lundberg and Lee (2017) approached global explanations through feature perturbation, masking feature values to zero. Covert et al. (2020) recently proposed a sampling-based approximated Shapley value method to consider subset feature effects in global explanations. However, this random sampling approximation struggled to converge and yield effective explanations in high-dimensional feature spaces (Sood and Craven, 2022), which is common in single-cell multi-modal data.

In this work, we propose GRIDS, a global feature explanation approach for efficient regulatory redundancy dissection using single-cell multi-modal data. Specifically, GRIDS comprises two components: a differentiable cross-modality surrogate mapping and a global explanation method for regulatory redundancy dissection via learnable subset perturbations. To elucidate the black-box regulatory model, the cross-modality surrogate mapping component initially learns modality-specific cell representations and then aligns them into a common semantic space through adversarial training. In the second step, GRIDS designs a learnable subset perturbation method with a state transition model, which dissects gene regulatory redundancy by generating a subset of globally important features to maximally modify a target gene's cell-type-specific expression. Our explanation approach belongs to the class of perturbation or removal-based model explanation methods. Unlike approaches that formulate combinatorial subset interactions as additive measures or use sampling approximation, our learnable subset perturbation method directly adds perturbation effects to the input CRE modality like a discrete replacement operation, while still using the standard auto-differentiation mechanism to update the subset elements as if they were continuous variables. This unique approach enables the generation of precise and efficient global feature importance explanations, crucial for analyzing large-scale biological datasets.

We evaluated GRIDS against various baseline models using image classification benchmarks and single-cell multi-modal datasets. The findings indicate that GRIDS provides more semantically meaningful feature importance values, enabling effective analysis of regulatory redundancy across extensive genome regions. To our knowledge, this study is the first to integrate global feature explanations with regulatory redundancy analysis in the context of single-cell multi-modal data. The source code is available at `https://github.com/jhliu17/nnpert`.

2. Preliminary

2.1. Problem Definition of Regulatory Redundancy Dissection

According to the definition proposed by Wu et al. (2021), the CRE is typically represented by the ATAC-seq $\mathbf{x} \in \{0,1\}^{d_a}$, which is a binary vector. Each dimension of this vector indicates the peak state in chromosomes, with "1" denoting an open state and "0" indicating

a closed state. It is important to note that ATAC-seq data is usually high-dimensional, with $d_a > 10^5$. The gene expression values (i.e., RNA-seq) regulated by the CRE are denoted as a real value vector $\mathbf{y} \in \mathbb{R}^{d_r}$. d_a and d_r represent the number of peaks and genes. We provide a detailed discussion on the relationship between ATAC-seq and RNA-seq in **Appendix A**. For a single-cell multi-omics dataset, it is a collection of N single-cell multi-modal data $\mathcal{C} = \{\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \ldots, \boldsymbol{c}^{(N)}\}$, where each cell $\boldsymbol{c}^{(i)} = (\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ contains a ATAC-seq vector $\mathbf{x}^{(i)}$ and its corresponding RNA-seq vector $\mathbf{y}^{(i)}$. Besides, each cell $\boldsymbol{c}^{(i)}$ has a semantic label $\ell^{(i)} \in \{1, \ldots, T\}$ to indicate its cell type in T classes. 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 $\ell^{(i)} = k$.

As mentioned above, gene expression level is precisely controlled by transcriptional regulation via CREs, executed through complex biological processes within cells. This can be formulated as a regulatory function $\mathbf{y} = \mathcal{F}(\mathbf{x})$, where $\mathcal{F}(\mathbf{x}) : \mathbb{R}^{d_a} \to \mathbb{R}^{d_r}$. The regulatory function $\mathcal{F}$, a black-box model, is challenging to query frequently due to experimental costs. The problem of regulatory redundancy dissection aims to find a subset of L peak indices $\boldsymbol{r} = \{r_1, \ldots, r_L\}$ within the CRE (i.e., the subset feature in the ATAC-seq $\mathbf{x}_{\boldsymbol{r}} \equiv \{\mathbf{x}_j | j \in \boldsymbol{r}\}$) that are crucial for regulating the target gene's expression across a cell population. Therefore, the domain of $\boldsymbol{r}$ is the binomial combination subset $\binom{d_a}{L}$, which constitutes a large search space, especially considering that $\mathbf{x}$ typically contains more than 10^5 dimensions.

2.2. 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). For example, the cross-entropy can be adopted to measure probability degradation in binary classification, while in our task, we use a mean-squared loss to describe gene expression value degradation. The optimal subset $\boldsymbol{r}^*$ in Eq. 1 is the solution to the regulatory redundancy problem defined in Section 2.1.

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

$$\boldsymbol{r}^* = \underset{\boldsymbol{r}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbf{c} \sim \mathcal{C}}[\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\setminus \boldsymbol{r}}), \mathbf{y})]. \tag{2}$$

We discuss the surrogate modeling and training details in Section 3.1. However, even with a tractable objective, previous global explanation methods are still inefficient in large feature

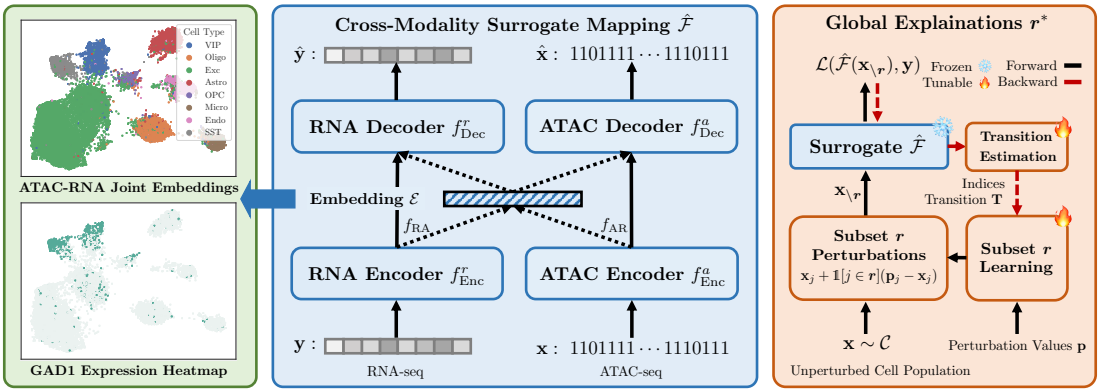

Figure 1: Overview of our proposed GRIDS method. It comprises two steps: training a cross-modality surrogate model and using a global explanation method to dissect regulatory redundancy. The left panel visualizes the learned ATAC-RNA joint embeddings and the heatmap of GAD1 gene expression.

spaces and infeasible in high-dimensional ATAC-seq. These methods either simplify the combinatorial subset effects as additive measures without considering interactions among different input features (Schwab and Karlen, 2019; Lundberg and Lee, 2017), or they use random sampling strategies to measure these effects (Covert et al., 2020). We describe our solution to this challenge in Section 3.2.

3. Methodology

We now describe the two key components of GRIDS: the cross-modality surrogate mapping and the global explanation method to efficiently dissect regulatory redundancy in the high-dimensional ATAC-seq space. The overview of our method is shown in **Figure 1**.

3.1. 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

$$\mathbf{h}_a^{(i)} = f^a_{\text{Enc}}(\mathbf{W}^a_{\text{Emb}}(\mathbf{x}^{(i)})),\ \mathbf{h}_r^{(i)} = f^r_{\text{Enc}}(\mathbf{W}^r_{\text{Emb}}(\mathbf{y}^{(i)})) \tag{3}$$

where $\mathbf{W}^a_{\text{Emb}} \in \mathbb{R}^{d_h \times d_a}$ is a category embedding module to accommodate the high-dimensional ATAC-seq data, $\mathbf{W}^r_{\text{Emb}} \in \mathbb{R}^{d_h \times d_r}$ is an embedding matrix for RNA-seq, f^a_{Enc} and f^r_{Enc} 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 $\hat{\mathbf{x}}^{(i)} = f^a_{\text{Dec}}(\mathbf{h}_a^{(i)})$, $\hat{\mathbf{y}}^{(i)} = f^r_{\text{Dec}}(\mathbf{h}_r^{(i)})$, where f^a_{Dec} and f^r_{Dec} are

two decoder networks for the two modalities, $\hat{\mathbf{x}}^{(i)}$ and $\hat{\mathbf{y}}^{(i)}$ represent the reconstructions with objective defined as

$$\mathcal{L}_{\text{Rec}} = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{C}}[\text{BCE}(\hat{\mathbf{x}}^{(i)}, \mathbf{x}^{(i)}) + \text{MSE}(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)})] \tag{4}$$

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

Alignment Embedding Adversarial Training To align the modality-specific embeddings and capture the regulatory regulations between them, two mapping layers are adopted to jointly align the two modalities

$$\tilde{\mathbf{h}}_r^{(i)} = f_{\text{AR}}(\mathbf{h}_a^{(i)}),\ \tilde{\mathbf{h}}_a^{(i)} = f_{\text{RA}}(\mathbf{h}_r^{(i)}) \tag{5}$$

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

$$\begin{aligned}\mathcal{L}_{\text{Dis}}^k =& \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[\log D_a^k(\mathbf{h}_a)] + \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[\log(1 - D_a^k(\tilde{\mathbf{h}}_a))] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[\log D_r^k(\mathbf{h}_r)] + \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[\log(1 - D_r^k(\tilde{\mathbf{h}}_r))].\end{aligned} \tag{6}$$

The generators are trained to simultaneously fool the discriminator and keep the cycle consistency (Zhu et al., 2017)

$$\begin{aligned}\mathcal{L}_{\text{Gen}}^k =& \mathbb{E}_{\mathbf{x}\sim\mathcal{C}^k}[-\log D_r^k(\tilde{\mathbf{h}}_r) + \text{MSE}(f_{\text{RA}}(\tilde{\mathbf{h}}_r), \mathbf{h}_a)] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathcal{C}^k}[-\log D_a^k(\tilde{\mathbf{h}}_a) + \text{MSE}(f_{\text{AR}}(\tilde{\mathbf{h}}_a), \mathbf{h}_r)].\end{aligned} \tag{7}$$

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

$$\mathcal{L}_{\text{Adv}} = \min_{\theta_{\text{Gen}}} \max_{\theta_{\text{Dis}}} \mathbb{E}_{k\sim T}[\mathcal{L}_{\text{Gen}}^k + \mathcal{L}_{\text{Dis}}^k] \tag{8}$$

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

$$\mathcal{L}_{\text{Int}} = \mathcal{L}_{\text{Rec}} + \gamma\mathcal{L}_{\text{Adv}} \tag{9}$$

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

$$\hat{\mathcal{F}}(\mathbf{x}; \theta_f) = f_{\text{Dec}}^r(f_{\text{AR}}(f_{\text{Enc}}^a(\mathbf{W}_{\text{Emb}}^a(\mathbf{x})))). \tag{10}$$

3.2. Learnable Global Subset Explanations for Regulatory Redundancy

Our global explanation method aligns closely with the class of perturbation or removal-based methods. The objective in Eq. 2 illustrates how the replacement perturbation, induced by the selected subset $\boldsymbol{r}$, can affect the performance degradation of the surrogate model across a population of cells.

Global Explanation Objective 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:

$$\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}}), \mathbf{y}) = (\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}})_i / \mathbf{y}_i)^2 + \frac{\beta}{d_r - 1} \sum_{j=1, j\neq i}^{d_r} (\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}})_j / \mathbf{y}_j - 1)^2 \tag{11}$$

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. This definition aligns with our objective of multi-CRE-to-gene regulatory redundancy outlined in Section 2.1. To optimize this objective, the domain of potential perturbations $\boldsymbol{r}$, represented as $\binom{d_a}{L}$, is too large to allow for either a comprehensive or a sampling-based search, particularly in the extremely high-dimensional ATAC-seq space ($d_a > 10^5$). Given that the surrogate $\hat{\mathcal{F}}$ is a differentiable approximation, we keep the parameters in the surrogate frozen and propose that the gradient of the Eq. 2 can be leveraged to efficiently learn the CRE subset $\boldsymbol{r}$.

Learning Optimal Subset Perturbations $\boldsymbol{r}^*$ As we mentioned in Section 2.2, the features $\mathbf{x}_{\backslash \boldsymbol{r}}$ with replacement perturbations $\mathbf{p}$, induced by a subset $\boldsymbol{r}$, are given by $\mathbf{x}_{\backslash \boldsymbol{r}, r_j} = \mathbf{p}_{r_j}$, where $r_j \in \boldsymbol{r}$. Although the surrogate is fully differentiable, the replacement perturbation is a discrete operator, which means the subset cannot be directly optimized as a continuous variable through standard gradient descent methods. To work around the non-differentiable replacement operation, we observe that replacement with any perturbation values $\mathbf{p}$ can be unified as follows

$$\mathbf{x}_{\backslash \boldsymbol{r}, j} = \mathbf{x}_j + \mathbb{1}[j \in \boldsymbol{r}](\mathbf{p}_j - \mathbf{x}_j) \tag{12}$$

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. This unified replacement operation allows us to more easily analyze the feature-changing effects caused by any replacement perturbation strategies, including removal, mean value filling, etc. Given a randomly initialized subset $\boldsymbol{r}$ and the global explanation objective in Eq. 2, 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

$$\mathbf{G} = \partial \mathbb{E}_{\mathbf{c} \sim \mathcal{C}}[\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\backslash \boldsymbol{r}}), \mathbf{y})] / \partial \mathbf{W}^a_{\text{Emb}}(\mathbf{x}_{\backslash \boldsymbol{r}}) \tag{13}$$

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 $\mathbf{W}^a_{\text{Emb}}(\mathbf{p}) - \mathbf{W}^a_{\text{Emb}}(\mathbf{x})$

$$\begin{aligned} \mathbf{d}_j &= \mathbf{G}_j \cdot (\mathbf{W}^a_{\text{Emb}}(\mathbf{p})_j - \mathbf{W}^a_{\text{Emb}}(\mathbf{x})_j) \\ \mathbf{T}_{i,j} &= \mathbb{1}[j \notin \boldsymbol{r}]\mathbf{d}_j - \mathbb{1}[j \neq r_i]\mathbf{d}_{r_i} \end{aligned} \tag{14}$$

where $\mathbf{d}_j \in \mathbb{R}^{d_a}$ represents the approximated objective descent value estimated for applying the potential perturbation $\mathbf{p}_j$. Though it may not be immediately obvious, our subset

replacement strategy can be directly extended to continuous features (such as images) without requiring an extra embedding module. In this situation, the objective gradient can be computed with respect to the perturbed input features $\mathbf{x}_{\setminus \boldsymbol{r}}$. Therefore, the approximated objective descent value $\mathbf{d}$ is simplified as $\mathbf{d}_j = \mathbf{G}_j \times (\mathbf{p}_j - \mathbf{x}_j)$. Meanwhile, the construction of the advantage value in the indices transition matrix remains the same as in Eq. 14.

Given the estimated state transition matrix of indices $\mathbf{T}$, there are two methods for updating the global feature subset $\boldsymbol{r}$. The first method involves coordinate descent, 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 $\mathbf{T}$. 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 $\mathcal{L}(\hat{\mathcal{F}}(\mathbf{x}_{\setminus \boldsymbol{r}'}), \mathbf{y})$. The other approach involves updating the entire subset $\boldsymbol{r}$ simultaneously as a sequence generation process using the beam search algorithm, which consistently maintains the best choices up to the beam size at each step. 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.

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 ($d_a > 10^4$). Our method can be extended to various kinds of perturbation tasks and can also be applied to other data modalities, including images and texts. The overall algorithm is summarized in **Algorithm 1**, located in **Appendix B**.

4. Related Work

Cicero (Pliner et al., 2018) was developed to link CREs to target genes using ATAC-seq data via a graphical lasso model. Then, ArchR relies on pair-wise correlations to link CREs to genes one at a time (Granja et al., 2021). Later on, DirectNet used a fitting model to predict gene expression using CRE status and then explore CRE-to-gene linkage via model selection. While promising, these methods still mainly focused on evaluating individual CRE's impact (Zhang et al., 2022). Therefore, it is still challenging to answer a key question - "upon modification, which set of CREs can jointly change a target genes' expression to the maximum degree?".

Feature importance explanation methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), focus on local explanations of individual predictions. Recent studies, like those by Chapman-Rounds et al. (2021) and (Chen et al., 2018), have shifted towards using differentiable surrogate models for explaining black-box models. Global explanation approaches, such as those by Schwab and Karlen (2019), Lundberg and Lee (2017), and Covert et al. (2020), involve perturbing features or using Shapley value methods to interpret complex feature interactions. However, these methods still struggle to yield effective explanations in high-dimensional feature spaces.

5. Experiments and Results

5.1. Expermental Setup

Single-Cell Multimodal Dataset We curated a set of deeply-sequenced single-cell multi-modal data from postmortem human PFC (Emani et al., 2024; 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. Details about the dataset can be found in **Appendix C**. 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. The detailed hyperparameter setting can be found in **Appendix D**.

MNIST Dataset To illustrate the combinatorial effects found by GRIDS, we conducted extensive experiments on the binary digit classification using MNIST. We compared both the global feature explanation estimation performance of GRIDS with strong baselines. We summarized the experiment results in **Appendix E**, where we observed that our method can effectively detect the combinatorial features in the feature space.

Baseline Comparisons We compared GRIDS 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.

5.2. The Surrogate Model Accurately Models the ATAC-to-RNA Relationship

As GRIDS depends on a differentiable approximating surrogate model $\hat{\mathcal{F}}$ for ATAC-to-RNA translation, we first evaluated the translation accuracy of the surrogate model in the

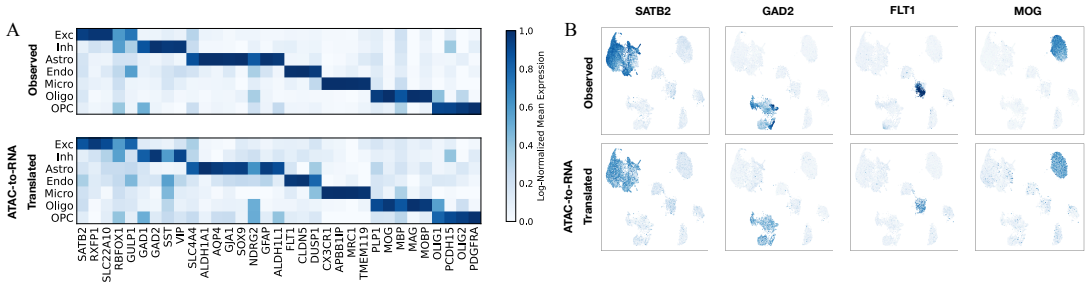

Figure 2: The trained differentiable surrogate model $\hat{\mathcal{F}}$ can accurately predict the RNA-seq modality from given single-cell ATAC-seq profiles. (A) The comparison of predicted marker gene expression with actual values across different cell types demonstrated high consistency and specificity to cell type. (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: Gene-focused 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**

single-cell multimodal dataset. Specifically, we selected a curated list of marker genes according to previous study Lake et al. (2016) and compared the mean expressions between cell types and between the observed and translated cohort (**Figure 2A**). The marker gene, representing the most expressively indicative gene for each cell type, is akin to the category label in classification tasks. We found that GRIDS managed to preserve similar expression patterns with the observed ground truth (mean R^2=0.914). We then focused on four key marker genes–SATB2 for excitatory cells, GAD2 for inhibitory cells, FLT1 for endothelial cells, and MOG for oligodendrocyte cells as they were the well-known marker genes commonly recognized. The UMAP of them agreed with our previous findings, with the translated expression highlighting the mentioned cell types, plus a high correlation between the observed and the translated ($R^2 = 0.633, 0.603, 0.460, 0.684$, **Figure 2B**). These results demonstrate that our surrogate model can accurately model the black-box ATAC-to-RNA translation process.

5.3. Global Subset Perturbations in Regulatory Redundancy Dissection

We then evaluated the performance of GRIDS to dissect multi-CRE-to-gene regulatory redundancy by generating global feature importance explanations in the high-throughput single-cell multi-omics data.

5.3.1. Evaluation of Regulatory Redundancy with Learnable Perturbations

One major challenge of single-cell data analysis is the data's high dimensionality. For instance, a typical ATAC-seq dataset generates hundreds of thousands of peaks, leaving it unrealistic for us to use traditional leave-one-out approaches (Li et al., 2016). We accelerate this process by using global feature importance explanations. To verify the effectiveness and robustness of the explanation process, we benchmarked GRIDS with various baselines in **cell-type-focused** and **gene-focused** settings, including (1) focused marker genes of 7 cell types, and (2) comprehensive highly-expressed gene sets from two representative cell types (VIP and Microglia). 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) and the ratio of expression change against the original value (Rel. Δ).

We summarize our benchmarking results of the cell-type-focused and gene-focused settings in **Table 1** and **Table 2**, respectively. In our experiments, we observed that although LIME, CXPlain, and SAGE are effective on the MNIST dataset (see **Appendix E**), they all 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.

As shown in **Table 1**, in the gene-focused setting, GRIDS consistently outperforms all baselines across each cell type by introducing larger marker gene expression degradation. The average gene expression change by GRIDS is -7.30%, as compared to -0.02% to -5.36% in other methods (p-value < 0.01, one-sided t-test). We also evaluated our method by removing $L = 128$ global important features. Similar trends can be observed in the table, further verifying the searching performance of GRIDS. In the cell-type-focused setting, as shown in **Table 2**, we compared the top 100 highly expressed genes expression changes in two cell types by masking L important CRE inputs predicted by different explanation methods. We found that the Random, Saliency, and SmoothGrad can barely report effective solution.

Table 2: Cell-type-focused benchmark results in VIP and Microglia by comparing expression degradation of highly expressed genes after masking CRE features in the global explanation subset $\boldsymbol{r}$.

Type	L	Method	Avg. Δ	Rel. Δ (%)
VIP-100	10	Random	−0.448	−0.009
		Saliency	−18.822	−0.915
		SmoothGrad	−18.424	−0.927
		FIMAP	−56.469	−3.087
		GRIDS	**−64.016**	**−3.827**
Microglia-100	10	Random	−0.333	−0.008
		Saliency	−42.372	−1.941
		SmoothGrad	−44.125	−2.073
		FIMAP	−115.092	−5.863
		GRIDS	**−141.339**	**−7.466**

A possible reason for the poor performance of these methods is that the RNA-ATAC cross-mapping relationship is complex and requires multi-step dissections, so gradient information with only one step can be misleading without explicit CRE removals. In contrast, both the FIMAP method and GRIDS can significantly suppress gene expression by masking only 10 CRE features. Besides, the global future importance generated by GRIDS introduces noticeably larger expression change than the FIMAP method, demonstrating its effectiveness and robustness in dissecting the regulatory redundancy. We show the cell-type-specific marker gene expression changes by removing $L = 128$ CREs chosen from all cell types in **Figure 3**. It is not surprising to observe the largest marker gene expression drop by removing the explanation chosen from matched cell types (i.e., the diagonal entries in the heat map). On the other hand, masking important features derived from related cell types also introduced decent gene expression changes, reflecting the regulatory similarity in close cell types. For instance, SST and VIP are two sub-types of inhibitory neurons with a common pan-inhibitory marker gene GAD1. We observed a slightly smaller but still decent expression drop in SST cells by removing CRE sets chosen by VIP cells. Overall, GRIDS can identify optimal CRE set in a cell-type-specific manner, which is essential to characterize regulatory redundancy heterogeneity across diverse cell types.

5.3.2. Evaluation of Relevant Regulatory Effects

To evaluate whether the global explanations generated by several methods can be helpful for advancing biological discovery. We benchmarked the explanation results produced by different methods by measuring the CRE-to-gene distance. Since distance has the biggest impact on CRE-to-gene interaction, we defined a local neighborhood of 10 million base pairs (MB) to allow direct CRE-to-gene interaction via chromatin looping, which is widely used in genomics (Phanstiel et al., 2017; Swygert et al., 2021). Then, we adopted the Soft Hit Ratio (SHR) to measure how many of the reported L CREs are located in this neighborhood. We also use the Hit Ratio (HR) to calculate an exact match, which describes where the region found is exactly relevant with the target gene.

Figure 3: Normalized pair-wise expression changes of marker genes.

Table 3: The hit ratio of direct CRE-to-gene interactions.

Method	$L = 10$		$L = 128$	
	HR ↑	SHR ↑	HR ↑	SHR ↑
Saliency	0.00	0.00	6.25	18.75
SmoothGrad	0.00	0.00	0.00	18.75
FIMAP	12.50	12.50	18.75	56.25
GRIDS	**18.75**	**25.00**	**31.25**	**68.75**

As shown in **Table 3**, Saliency and SmoothGrad report very few directly interacting CREs in the local neighborhood for both $L = 10$ and $L = 128$ cases, which is not surprising given their limited expression drop by removing these CRE features. Our GRIDS model generated the global explanations with a substantially larger percent of directly interacting CREs as compared to the FIMAP baseline, demonstrating the effectiveness of our method. We also conducted an independent validation using cell-type-matched Hi-C experiments,

as reported in **Appendix F**. It is worth noting that it is usually difficult for expression-guided searching schemes to distinguish direct CRE-gene relations via physical contact from those via indirect mechanisms. For instance, some CREs may only physically regulate upstream transcription factors (TFs), which can pass the indirect effects to the target genes without requiring physical interactions. Modeling the regulatory function with additional side information would help to more accurately capture global feature interactions using model explanation methods.

6. Conclusions

In this paper, we propose GRIDS, a global feature importance explanation method designed to dissect complex multi-CRE-to-gene regulatory redundancy using single-cell multi-modal data. To achieve this goal, GRIDS first facilitates cross-modality surrogate mapping to create a differentiable approximation of the black-box regulatory function. This surrogate enables us to unify the regulatory redundancy problem with global feature importance explanations. Furthermore, GRIDS introduces a subset perturbation learning framework for efficiently generating subsets of global feature importance explanations. Our explanation method can be efficiently applied across various data modalities, including high-throughput biological sequence data. Experimental results on the image benchmark and single-cell data demonstrate the superiority of the GRIDS method over other state-of-the-art baselines. Additionally, cross-cell type and regional analysis reveal that GRIDS can effectively and efficiently characterize cell-type-specific regulatory redundancy mechanisms by generating global explanations using single-cell multi-modal data. These results offer significant potential for directing experimental validations in wet labs, highlighting the practical relevance of our method in biological research.

Acknowledgments

This work was supported by the National Institutes of Health [R01HG012572, R01NS128523]. We thank the UCI ICS Computing Support for guidance and use of the research computing infrastructure.

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.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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.

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.

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.

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

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.

Nicolás Frankel, Gregory K. Davis, Diego Vargas, Shu Wang, François Payre, and David L. Stern. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493, 2010. ISSN 0028-0836. doi: 10.1038/nature09158.

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

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.

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.

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.

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.

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.

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.

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.

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.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

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.

Y. Luo, B. C. Hitz, I. Gabdank, J. A. Hilton, M. S. Kagda, B. Lam, Z. Myers, P. Sud, J. Jou, K. Lin, U. K. Baymuradov, K. Graham, C. Litton, S. R. Miyasato, J. S. Strattan, O. Jolanki, J. W. Lee, F. Y. Tanaka, P. Adenekan, E. O'Neill, and J. M. Cherry. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic Acids Res*, 48(D1):D882–d889, 2020. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz1062.

Laurina Manning, Ellie, Maria, Jourdain Roberts, Alysha, Jason, Jill, Marie, Josh, Anna, and et al. A resource for manipulating gene expression and analyzing cis-regulatory modules in the drosophila cns. *Cell Reports*, 2(4):1002–1013, 2012. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.09.009.

Michael W. Perry, Alistair N. Boettiger, and Michael Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the drosophila embryo. *Proceedings of the National Academy of Sciences*, 108(33):13570–13575, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1109873108.

D. H. Phanstiel, K. Van Bortle, D. Spacek, G. T. Hess, M. S. Shamim, I. Machol, M. I. Love, E. L. Aiden, M. C. Bassik, and M. P. Snyder. Static and dynamic dna loops form ap-1-bound activation hubs during macrophage development. *Mol Cell*, 67(6):1037–1048.e6, 2017. ISSN 1097-2765 (Print) 1097-2765. doi: 10.1016/j.molcel.2017.08.006.

Hannah A. Pliner, Jonathan S. Packer, and et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular Cell*, 71(5):858–871.e8, 2018. ISSN 1097-2765. doi: 10.1016/j.molcel.2018.06.044.

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.

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.

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. *arXiv preprint arXiv:1706.03825*, 2017.

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.

S. G. Swygert, D. Lin, S. Portillo-Ledesma, P. Y. Lin, D. R. Hunt, C. F. Kao, T. Schlick, W. S. Noble, and T. Tsukiyama. Local chromatin fiber folding represses transcription and loop extrusion in quiescent cells. *Elife*, 10, 2021. ISSN 2050-084x. doi: 10.7554/eLife.72062.

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.

Lihua Zhang, Jing Zhang, and Qing Nie. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science Advances*, 8(22), June 2022. doi: 10.1126/sciadv.abl7393.

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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.