Towards Hierarchical Causal Representation Learning for Nonstationary Multi-Omics Data

Haiyi Mao[1,2,6], Minxue Jia[1,2,6], Marissa Di[1,2], Kun Zhang[4,5], and Panayiotis V. Benos[1,2,3]

[1] Department of Computational and Systems Biology, University of Pittsburgh
[2] Joint Carnegie Mellon University – University of Pittsburgh PhD Program in Computational Biology
[3] Department of Epidemiology, University of Florida
[4] Department of Philosophy, Carnegie Mellon University
[5] Mohamed bin Zayed University of Artificial Intelligence
[6] These authors contributed equally to this work and share first authorship

Abstract. As the available sequencing data modalities increase, so does the potential biological insight that they are able to provide. Most existing methods to integrate co-profiled single-cell multi-omics data focus only on learning representations that capture stationary and shared information among these modalities. Current methods do not account for time-dependent and modality-specific information delineating cell states and subtypes, nor do they consider dynamics resulting from causal relations among modalities. For example, open chromatin may cause active transcription; however, it is also possible that gene expression responses lag behind changes in chromatin accessibility. To account for this time lag, the epigenome and transcriptome relationship can be characterized as "coupled" (changing dependently) or "decoupled" (changing independently). We propose the framework HALO (**H**ierarchical c**A**usal representation **L**earning for **O**mics data), which adopts a causal approach to model these nonstationary causal relations using independent changing mechanisms in co-profiled single-cell ATAC- and RNA-seq data. Our model factorizes these two modalities into both coupled and decoupled latent representations, allowing us to identify the dynamic interplay between chromatin accessibility and transcription through temporal modulations. In blood lineage and developing mouse brain data, where the balance between proliferation and differentiation is tightly regulated, HALO distinguishes between coupled and decoupled genes and links them with disparate processes that constitute these two complementary states.

Keywords: Multi-Omics · Single-cell · Casual Representation Learning

1 Introduction

Recent advances in single-cell sequencing technologies have enabled the probing of multiple "omics" layers, such as chromatin accessibility with single-cell ATAC-sequencing (scATAC-seq) and gene transcription with single-cell RNA-sequencing (scRNA-seq), which can deliver more comprehensive biological insights and uncover interactions between multiple omics layers at single-cell resolution [5,21]. However, our ability to integrate different omics modalities is still limited. Most integration methods hold onto a strict biological assumption that accessible chromatin and active transcription are consistently "coupled" (changing dependently). However, some of the true associations may be time lagged and incorporating complementary information from multiple modalities could be useful for extensive delineation of cell states [2,18]. Specifically, chromatin remodeling of a gene region and its transcription may become decoupled in situations such as: (1) when chromatin opens before transcription begins (also called "priming"), (2) RNA degradation and nuclear export occur, or (3) chromatin closes before transcription ends. We refer to this phenomenon as time lag or "decoupling" (changing independently) [17]. For example, Tle4 exhibits a delay between the induction of chromatin accessibility and the corresponding gene expression (Fig. 1a), while Ptprd demonstrates the inverse situation Fig. 1c). Decoupling between the transcription and chromatin accessibility of Epha5 demonstrates additional mechanisms contributing to the repression of gene expression (Fig. 1b). Utilizing information from one modality alone for clustering of single-cell genomics data may lead to identification of spurious associations between expressed genes and nearby open chromatin regions (peaks in ATAC-seq) [22].

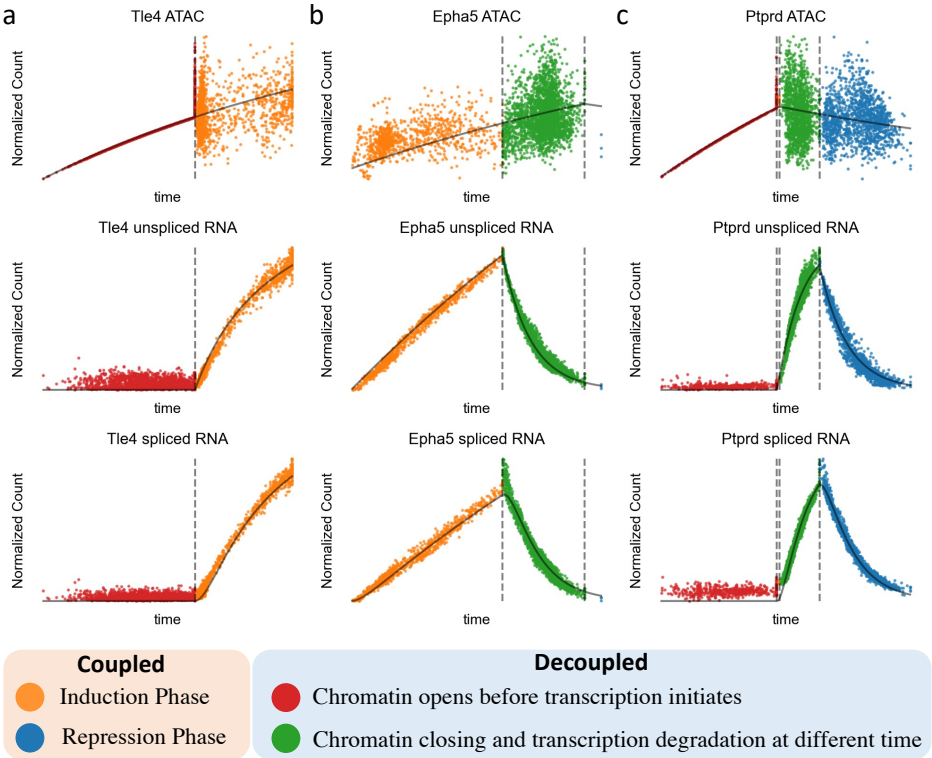

Fig. 1: Time lag results in decoupling between chromatin accessibility and transcription products. Normalized counts of chromatin accessibility (top), unspliced RNA (middle), and spliced RNA (bottom) plotted as a function of latent time and colored by coupled or decoupled states, for **a**. Tle4 (delay in increased transcription after chromatin has already opened), **b**. Epha5 (RNA levels decrease, although chromatin is open in the same time interval), and **c**. Ptprd (chromatin closes but RNA levels increase).

Inconsistent coupling of chromatin accessibility of genes and their transcription indicates that chromatin regulation and transcription mechanisms are non-stationary. Indeed, the underlying generating process of "omics" data tends to change over time at single-cell resolution. To model the data generating processes and

non-stationarity of multiple omics modalities, we can examine the connections between chromatin remodeling and transcription through the paradigm of causal relations. Causal models provide a succinct description of data generation processes. Recently, many endeavors have been made to enhance representation learning using causal inference and discovery by considering the data generating processes alongside interventions and changes [31,7,13]. Unlike traditional machine learning methods, causal machine learning attempts to model effects under intervention or distribution shift and have been proven successful in the analysis of omics and clinical data [1,28,3,29,30,32,33]. By utilizing the principles of independent causal mechanisms (modularity) and minimal changes or sparse mechanism shift [31,15,12], one can learn causally related representations instead of spurious associations.

One of the most important features of causal models is their aforementioned modularity, or the ability of mechanisms (data-generating processes) to vary independently of each other [27,13]. That is, if we assume that **cause** and **effect** are not influenced by confounders, then $P(\textbf{cause})$ and $P(\textbf{effect} \mid \textbf{cause})$ are fixed or change independently under local alterations. This particular form of independence helps the integration of distinct modalities based on both modality-specific (complementary) information and correlated information. For example, suppose chromatin remodeling of a gene is **cause**, and its transcription is **effect**. Here, $P(\textbf{effect} \mid \textbf{cause})$ represents the probability of transcription given a particular state of chromatin accessibility. This corresponds to the biological fact that transcription-related proteins require physical access to the appropriate DNA sites, and DNA accessibility depends on chromatin remodeling. However, $P(\textbf{cause})$, the probability of chromatin accessibility, is regulated by ATP-dependent nucleosome-remodelling complexes and histone modifications [6] and is independent of $P(\textbf{effect} \mid \textbf{cause})$. The time lag between chromatin accessibility and transcription may occur under the following situations: (1) chromatin region is open, but transcription decreases (e.g., due to disassembly of the transcription machinery), (2) chromatin region is open, while transcription levels stay constant because the necessary machinery and regulatory proteins are still in the process of recruitment, and (3) chromatin region is closed while the transcription machinery is still active (Fig. 1) [17] . In addition, the mechanisms of chromatin remodeling are distinct from those of transcription machinery assembly/disassembly [6]. For these decoupled time intervals, the causal relationship between chromatin accessibility and transcription is not constant and should not be modeled as such. Temporal information is needed to account for the decoupling phenomenon in a causal framework.

Recently, a causal discovery approach called Causal Discovery from Nonstationary/heterogeneous Data (CD-NOD) [13] has been developed for heterogeneous/non-stationary data by utilizing independent changes in the changing data distributions and estimating the "driving force" of the non-stationarity [13]. Hereby, we develop the framework HALO to learn low-dimensional representations of cross-profiled multi-omics data with causal regularization. This framework enables us to learn expressive and non-spurious representations of multiple modalities to further identify the underlying mechanisms of gene expression. Applications to various co-profiled scATAC and scRNA sequencing datasets demonstrate that HALO can effectively differentiate between coupled and decoupled omics information in latent space. HALO is also able to identify cell subtypes by using the combination of coupled or decoupled representations from two modalities. Moreover, HALO can identify epigenome and transcriptome states as well as distinguish between batch effects and biologically meaningful information by using independently changing causal modules.

2 Methods

In this section, we briefly introduce causal graphic models, structural causal models (SCMs), and non-stationary causal discovery conditions, then we give the formal definition of our problem.

2.1 Causal Graphic Models and Structural Causal Models

Finding underlying causal relations is a fundamental task in many disciplines; see, e.g., [34]. Moreover, utilizing causal models to encode distributional and causal assumptions is becoming more and more popular [26,41,40]. Markovian causal models are one of the simplest model types and consist of a directed acyclic graph (DAG) over a set of vertices $V = \{V_1, \ldots, V_n\}$, representing variables of interest, as well as a set E of directed edges, or arrows, that connect these vertices. The interpretation of such a graph has two components, probabilistic and causal. The probabilistic interpretation treats the arrows as representations of probabilistic dependencies among the corresponding variables, and the missing arrows represent conditional independencies. Each variable is independent of all its nondescendants given its direct parents in the graph

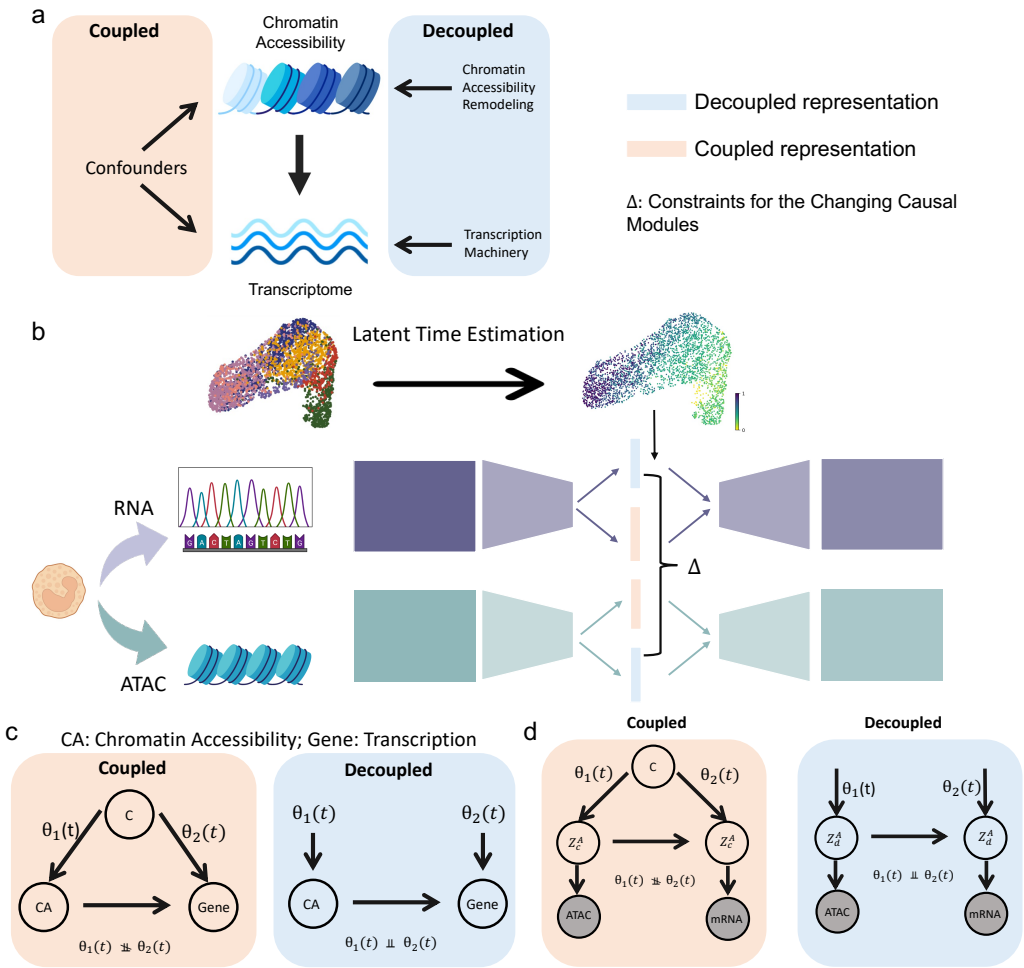

Fig. 2: HALO models coupling and decoupling with changing causal modules. a. Schematic causal diagram, which presents two categories of chromatin accessibility and transcription relations. Coupled chromatin accessibility and transcription are confounded by latent variables such that they change dependently. Meanwhile, decoupled chromatin accessibility and transcription result from independent biological mechanisms that induce independent changes. **b**. The general architecture of HALO. Given co-profiled scATAC X^A and scRNA X^R seqencing data, the latent time is estimated. Then, X^A and X^R are used as inputs to two encoders. The latent representations of scRNA-seq Z^R and scATAC-seq Z^A are factorized to $[Z_c^R, Z_d^R]$ and $[Z_c^A, Z_d^R]$, respectively. We denote $\{Z_c^R, Z_c^A\}$ as the coupled pair of latent representations, and $\{Z_d^R, Z_d^A\}$ as the decoupled pair; these pairs are regularized by causal constraints on the statistic Δ. **c**. Causal graphs summarizing the relationships between chromatin accessibility and transcription for coupled intervals and decoupled intervals at a time $T = t$. $\theta_1(t)$ and $\theta(t)$ are the parameters of causal functions in corresponding SCMs. **d**. Causal diagram of latent space of scATAC-seq and scRNA-seq data for both decoupled $\{Z_d^R, Z_d^A\}$ and coupled $\{Z_c^R, Z_c^A\}$ representations of scRNA-seq and scATAC-seq.

[34,27]. These assumptions are represented in the joint probability $P(V) = P(V_1, \ldots, V_n)$ factorization as the following product:

$$P(V) = \prod_{i=1}^{n} P(V_i | PA_i). \tag{1}$$

Here, $P(V_i|PA_i)$ is the conditional distribution of V_i given its parents.. The causal perspective of these models views the arrows as representing causal influences between the corresponding variables. In this case, the factorization in Eq. (1) still holds, but the factors are further assumed to represent autonomous data generation processes (here we refer to each $P(V_i|PA_i)$ as a causal module or causal mechanism [13]); that is, each conditional probability $P(V_i|PA_i)$ represents a stochastic process by which the values of V_i are generated in response to the values of PA_i, and the stochastic variation of this assignment is assumed to be

independent of the variations in all other assignments. In this sense, this representation is an alternative to an SCM. An SCM is defined [27] as a triple $M = \langle U, V, F \rangle$, where:

1. U is a set of exogenous (background) variables, that are determined by factors outside the model;
2. V is a set $\{V_1, V_2, \ldots, V_n\}$ of endogenous variables that are determined by variables in the model, $U \cup V$
3. F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and $PA_i \subseteq V \backslash V_i$. That is, $V_i = f_i(PA_i, U_i)$, $i = 1, \ldots, n$, assigns a value to V_i that is dependent on a set of variables in $V \cup U$.

2.2 Non-stationary Causal Discovery Conditions

In most biological processes, the biological mechanisms change with time, and this displacement is referred to as *distribution shift*. If there are distribution shifts (i.e. $P(V)$ changes across domains or over time), there must also be changes in the causal modules $P(V_i|PA_i), i \in N$; these causal models are called *changing causal modules*. In the setting of our framework, this corresponds to changes over time in the mechanism by which chromatin accessibility affects gene expression. It is usually assumed that the quantities that change over time can be written as functions of a time index, T [13]. We assume *pseudo causal sufficiency*, which means that any confounders can be written as smooth functions of time. Hence, in our multi-omics cases, the latent confounders can be presented as a function of pseudo-time. Denoting $\mathbf{g}(T)$ as the set of pseudo confounders (this set may also be empty), we can write the SCM as:

$$V_i = f_i(PA_i, g_i(T), \theta_i(T), U_i), \tag{2}$$

where $g_i(T) \subseteq \mathbf{g}(T)$ denotes the set of confounders related to T which influence V_i, $\theta_i(T)$ denotes the effective parameters in the model that are also assumed to be functions of T, and U_i is a disturbance term that is independent of T and PA_i, with a non-zero variance.

Causal Directions with Independent Changing Causal Modules. We model the decoupled scATAC-seq and scRNA-seq data with independent causal modules. Following CD-NOD, we construct a kernel embedding of the joint distribution conditioned on T and estimate the corresponding Hilbert Schmidt Independence Criterion (HSIC) to test the independence of changing causal modules. Before diving into the details, we generally introduce the principles behind identifying the corresponding causal directions. For the decoupled causal modules as in Figs. 2c and 2d, $\theta_1(T)$ and $\theta_2(T)$ are the independently changing parameters which represent chromatin remodeling mechanisms and transcription machinery. Note that, as a consequence, the distribution of chromatin accessibility $P(\text{CA}; \theta_1(T))$ which is parameterized by $\theta_1(T)$, and the conditional distribution of gene expression $P(\text{Gene}|\text{AC}; \theta_2(T))$ are changing independently for the decoupled genes-chromatin regions. Therefore, we can factorize the joint distribution as follows:

$$P(\text{CA}, \text{Gene}; \theta_1(T), \theta_2(T)) = P(\text{CA}; \theta_1(T)) P(\text{Gene}|\text{CA}; \theta_2(T)). \tag{3}$$

Furthermore, in Fig. 2d, the co-profiled scRNA and scATAC sequencing data are the proxies for chromatin accessibility and transcription. The latent representations entail the same causal relations between chromatin accessibility and transcription as shown in Fig 2d. Moreover, testing for the independence of changing $P(\textbf{cause})$ and $P(\textbf{effect}|\textbf{cause})$ can be achieved by HSIC with properly designed kernel embeddings of nonstationary conditional distributions [13].

Hilbert Schmidt Independence Criterion (HSIC). Given a set of observational samples $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ for random variables X and Y, HSIC is a statistic for testing statistical independence by evaluating the squared covariances C_{XY} between feature maps of X and feature maps of Y. Let us introduce the following theorem:

Theorem 1. *[10] Denote by $\mathcal{F}$, $\mathcal{G}$ Reproducing Kernel Hilbert Spaces (RKHS) with universal kernels k, l on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then $\|C_{xy}\|_{HS} = 0$ if and only if $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are independent.*

Specifically, let M_X and M_Y be the Gram matrices for X and Y estimated on the observed samples. We use following estimator of HSIC [11]:

$$\text{HSIC}_{XY} = \frac{1}{(N-1)^2} \operatorname{tr}(\mathbf{M}_X H \mathbf{M}_Y H), \tag{4}$$

where H is for centering the features, with entries $H_{ij} = \delta_{ij} - \frac{1}{N}$, and $\delta_{ij} = 1$ when $i = j$. Moreover, the normalized HSIC is given by:

$$\mathrm{HSIC}_{XY}^{\mathcal{N}} = \frac{\mathrm{tr}\left(\mathbf{M}_X H \mathbf{M}_Y H\right)}{\mathrm{tr}\left(\mathbf{M}_X H\right) \mathrm{tr}\left(\mathbf{M}_Y H\right)}. \tag{5}$$

HSIC can be tailored to serve as a powerful tool for evaluating the dependence score between changing causal modules. However, instead of directly applying it to the scATAC-seq and scRNA-seq data, we extend this approach to the latent space representations of these two noisy and sparse data modalities. In the next section, we introduce the formal definition of our problem and its solutions.

2.3 Problem Definition

In this section, we give the formal definition of our model. The goal of this paper is to learn the representations that contain the "coupled/overlapped" information between scATAC-seq X^A and scRNA-seq X^R data as well as the independent "decoupled/time lagging" information, such that the causal relations within information that changes with time are sustained. Specifically, the latent representation of scATAC-seq data Z^A can be decomposed to $Z^A = [Z_d^A, Z_c^A]$. The latent representation of scRNA-seq data Z^R may be similarly decomposed: $Z^R = [Z_d^R, Z_c^R]$. Z_d^R and Z_d^A denote the decoupled latent representations of scRNA- and scATAC-seq data, while Z_c^R and Z_c^A are their coupled counterparts. Moreover, following the general case [13], we denote parameters of the SCM for coupled and decoupled latent representations as the following: θ_{1c}, θ_{2c}, θ_{1d}, and θ_{2d} are the functional parameters of Z_c^A, Z_c^R, Z_d^A, and Z_d^R, respectively. Note that all these parameters are functions of time T, but we drop it for the convenience of notation. In this way, the latent representations contain both coupled and decoupled *causal* information. Fig. 2d shows the latent causal diagram of our model:

We postpone the topic of determining the time information T of the sequencing data to the Latent Time Estimation section. For now, let us assume we have time information and further develop statistical constraints for the coupled and decoupled causal cases. As presented in the previous section, the joint distribution in the decoupled case can be formulated as such:

$$P(Z_d^A, Z_d^R; \theta_{d1}, \theta_{d2}) = P(Z_d^A; \theta_{d1}) P(Z_d^R | Z_d^A; \theta_{d2}).$$

Based on this factorization, in light of CD-NOD [13], we formalize the specific joint distribution $\tilde{P}(\underline{Z_d^R}, Z_d^A | T = t_n)$ and its embeddings. $\tilde{P}(\underline{Z_d^R}, Z_d^A | T = t_n)$ is designed to capture the changes in only the conditional distribution $P(Z_d^R | Z_d^A, T = t_n)$. Formally, with $n = 1, \ldots, N$, it is defined as

$$\tilde{P}(\underline{Z_d^R}, Z_d^A | T = t_n) = P(Z_d^R | Z_d^A, T = t_n) P(Z_d^A). \tag{6}$$

Note that in Eq. (6), $P(Z_d^A)$ does not depend on T. As a result, $\tilde{P}(\underline{Z_d^R}, Z_d^A | T = t_n)$ encodes only the changes across T in $P(Z_d^R | Z_d^A, T = t_n)$. A corresponding joint distribution can be also applied to the coupled latent space. Let $Z^R \in \mathcal{Z}^R$ and $(\mathcal{H}, k)$ be an RKHS with a measurable kernel on $\mathcal{Z}^R$. Let $\phi(Z^R) \in \mathcal{H}$ be the feature map for $Z^R \in \mathcal{Z}^{\mathcal{R}}$, and assume $\mathbb{E}_{Z^R}[k(Z^R, Z^R)] \leq \infty$. Similar notations are used for Z^A, $Z_c^R, Z_d^R, Z_c^A, Z_d^A$ and T. Thus the general embedding of the joint distribution of coupled or decoupled latent representations can be formulated as the following:

Proposition 1. *[13] Let Z^A represent the direct causes of Z^R, and suppose that they have N observations. The kernel embedding of distribution $\tilde{P}\left(\underline{Z^R}, Z^A \mid T = t_n\right)$ can be represented as*

$$\hat{\tilde{\mu}}_{\underline{Z^R}, Z^A | T = t_n} = \frac{1}{n} \boldsymbol{\Phi}_{\mathbf{Z^R}} \left(\mathbf{K}_{\mathbf{Z^A}} \odot \mathbf{K}_{\mathbf{t}} + \lambda I\right)^{-1} \mathrm{diag}\left(\mathbf{k}_{\mathbf{t}, t_n}\right) \mathbf{K}_{\mathbf{Z^A}} \boldsymbol{\Phi}_{\mathbf{Z^A}}^{\top},$$

where $\boldsymbol{\Phi}_{\mathbf{Z^R}} := \left[\phi\left(Z_1^R\right), \ldots, \phi\left(Z_N^R\right)\right], \boldsymbol{\Phi}_{\mathbf{Z^A}} := \left[\phi\left(Z_1^A\right), \ldots, \phi\left(Z_N^A\right)\right], \mathbf{k}_{\mathbf{t}, t_n} := \left[k\left(t_1, t_n\right), \ldots, k\left(t_N, t_n\right)\right]^{\top}$, $\mathbf{K}_{\mathbf{Z^A}}\left(Z_t^A, Z_{t'}^A\right) = \left\langle \phi\left(Z_t^A\right), \phi\left(Z_{t'}^A\right) \right\rangle, \mathbf{K}_{\mathbf{t}}\left(t, t'\right) = \left\langle \phi\left(t\right), \phi\left(t'\right) \right\rangle$, and $\odot$ represents point-wise product.

The proof of this proposition can be found in the CD-NOD paper [13].

Independently Changing Causal Module Statistics. Our goal is to characterize the changing causal mechanisms between the latent space of scATAC- and scRNA-seq. In other words, for the decoupled case, Z_d^A and Z_d^R have independent changing mechanisms along T within the cell lifespan, but Z_c^A and Z_c^R change dependently. Thus, we aim to enforce that $P(Z_d^R | Z_d^A)$ and $P(Z_d^A)$ change independently, and that $P(Z_c^R | Z_c^A)$

and $P(Z_c^A)$ change dependently. CD-NOD [13] proposes a score function based on HSIC which measures the dependence of the changes in the two distributions $P(Y|X)$ and $P(X)$ over time. We translate the two-variable case to our Z^A and Z^R setting as follows:

$$\hat{\Delta}_{Z^A \to Z^R} = \frac{\mathrm{tr}(M_{Z^A} H M_{\underline{Z^R} Z^A} H)}{\mathrm{tr}(M_{Z^A} H)\mathrm{tr}(M_{\underline{Z^R} Z^A} H))}, \quad \hat{\Delta}_{Z^R \to Z^A} = \frac{\mathrm{tr}(M_{Z^R} H M_{\underline{Z^A} Z^R} H)}{\mathrm{tr}(M_{Z^R} H)\mathrm{tr}(M_{\underline{Z^A} Z^R} H))}. \tag{7}$$

where $\hat{\Delta}_{Z^A \to Z^R}$ measures the dependence between $P(Z^R|Z^A)$ and $P(Z^A)$. Similarly, $\hat{\Delta}_{Z^R \to Z^A}$ measures the dependence between $P(Z^A|Z^R)$ and $P(Z^R)$. Between these two dependence measures, if the value of $\hat{\Delta}_{Z^A \to Z^R}$ is smaller, we posit that $P(Z^R|Z^A)$ and $P(Z^A)$ change independently, and that the causal relation is $Z^A \to Z^R$. In our case, for the decoupled representations Z_d^R and Z_d^A, $\hat{\Delta}_{Z_d^A \to Z_d^R}$ should be smaller than $\hat{\Delta}_{Z_d^R \to Z_d^A}$, based on our knowledge of chromatin accessibility and transcription. On the other hand, for the coupled case, $\hat{\Delta}_{Z_c^A \to Z_c^R}$ and $\hat{\Delta}_{Z_c^R \to Z_c^A}$ should be similar and both larger than a threshold α [13]. Formally, the causal constraints of $[Z_d^R, Z_c^R]$ and $[Z_d^A, Z_c^A]$ are given by the constraints as follows,

$$\begin{aligned} &\text{Decoupled Case:} \quad \hat{\Delta}_d = \hat{\Delta}_{Z_d^A \to Z_d^R} - \hat{\Delta}_{Z_d^R \to Z_d^A} < 0, \quad \hat{\Delta}_{Z_d^A \to Z_d^R} < \alpha \\ &\text{Coupled Case:} \quad \hat{\Delta}_{Z_c^A \to Z_c^R} \geq \alpha, \quad \hat{\Delta}_{Z_c^R \to Z_c^A} \geq \alpha. \end{aligned} \tag{8}$$

Models. Finally, we formally introduce our representation learning model. We aim to learn scATAC-seq X^A representations $[Z_c^A, Z_d^A]$ and scRNA-seq X^R representations $[Z_c^R, Z_d^R]$ where coupled and decoupled representations as constrained as described in Eq. (8). In this paper, we adopt the variational autoencoder framework [14] to learn the latent representations of our data. Particularly, as Fig. 2c shows, we have two encoders and decoders each for scATAC-seq and scRNA-seq, in order to learn $Z^A = [Z_c^A, Z_d^A]$ and $Z^R = [Z_c^R, Z_d^R]$. Additionally, we enforce the decoupled representations Z_d^A, Z_d^R to have $\hat{\Delta}_d = \hat{\Delta}_{Z_d^A \to Z_d^R} - \hat{\Delta}_{Z_d^R \to Z_d^A} < 0$ and the coupled representations to have $\hat{\Delta}_{Z_c^A \to Z_c^R} \geq \alpha$, $\hat{\Delta}_{Z_c^R \to Z_c^A} \geq \alpha$. The Evidence Lower Bound (ELBO) of two marginal distribution can then be formulated as follows:

$$\begin{aligned} &\log P_\theta(X^A) \geq \mathbb{E}_{q_\phi(Z^A|X^A)}[\log p_\theta(X^A|Z^A)] - D_{KL}(q_\phi(Z^A|X^A)||p_\theta(Z^A)), \\ &\log P_\theta(X^R) \geq \mathbb{E}_{q_\phi(Z^R|X^R)}[\log p_\theta(X^R|Z^R)] - D_{KL}(q_\phi(Z^R|X^R)||p_\theta(Z^R)), \\ &\text{s.t.} \quad \hat{\Delta}_d = \hat{\Delta}_{Z_d^A \to Z_d^R} - \hat{\Delta}_{Z_d^R \to Z_d^A} < 0, \hat{\Delta}_{Z_d^R \to Z_d^A} < \alpha, \hat{\Delta}_{Z_c^A \to Z_c^R} \geq \alpha, \hat{\Delta}_{Z_c^R \to Z_c^A} \geq \alpha. \end{aligned} \tag{9}$$

So we can maximize the ELBO with following objective,

$$\begin{aligned} \arg\max_{\theta,\phi} \mathcal{L}(\theta, \phi, X^A, X^R) = &\mathbb{E}_{q_\phi}(\log P(X^A|Z^A)) + \mathbb{E}_{q_\phi}(\log P(X^R|Z^R)) \\ &- D_{KL}(q_\phi(Z^R|X^R)||p_\theta(Z^R)) - D_{KL}(q_\phi(Z^A|X^A)||p_\theta(Z^A)), \\ \text{s.t.} \quad \hat{\Delta}_d = \hat{\Delta}_{Z_d^A \to Z_d^R} - \hat{\Delta}_{Z_d^R \to Z_d^A} < 0, &\hat{\Delta}_{Z_d^R \to Z_d^A} < \alpha, \hat{\Delta}_{Z_c^A \to Z_c^R} \geq \alpha, \hat{\Delta}_{Z_c^R \to Z_c^A} \geq \alpha. \end{aligned} \tag{10}$$

In practice, we can maximize the ELBO with the following objective:

$$\begin{aligned} \arg\max_{\theta,\phi} \mathcal{L}(\theta, \phi, X^A, X^R) = &\mathbb{E}_{q_\phi}(\log P(X^A|Z^A)) + \mathbb{E}_{q_\phi}(\log P(X^R|Z^R)) \\ &- D_{KL}(q_\phi(Z^R|X^R)||p_\theta(Z^R)) - D_{KL}(q_\phi(Z^A|X^A)||p_\theta(Z^A)) \\ &- \mathbf{max}(0, \hat{\Delta}_d) - \mathbf{max}(0, \alpha - \hat{\Delta}_{Z_c^A \to Z_c^R}) - \mathbf{max}(0, \alpha - \hat{\Delta}_{Z_c^R \to Z_c^A}) - \mathbf{max}(0, \hat{\Delta}_{Z_d^R \to Z_d^A} - \alpha) \\ &- D_{KL}(q_\phi(Z^R|X^R)||q_\phi(Z^A|X^A)) - D_{KL}(q_\phi(Z^A|X^A)||q_\theta(Z^R|X^R)). \end{aligned} \tag{11}$$

Notice that $\mathbf{max}(0, \hat{\Delta}_d)$, $\mathbf{max}(0, \alpha - \hat{\Delta}_{Z_c^A \to Z_c^R})$, $\mathbf{max}(0, \alpha - \hat{\Delta}_{Z_c^R \to Z_c^A})$, $\mathbf{max}(0, \hat{\Delta}_{Z_d^R \to Z_d^A} - \alpha)$ and D_{KL} are all non-negative, so Eq. 11 is still a lower bound of $\log P(X^R), \log P(X^A)$.

Latent Time Estimation. Recall the need to address how to estimate the latent time and incorporate it into our framework. Generally speaking, we can use any latent time or pseudotime estimation methods to annotate cells, as long as they are consistent across samples. For the bone marrow dataset, we use pseudotime to order cells, and for the mouse brain dataset, we adopt MultiVelo to estimate the latent

time. MultiVelo, an extension of the dynamical RNA velocity model, uses a combination of scRNA-seq and scATAC-seq data to account for epigenomic regulation of gene expression [17]. It infers gene regulation information using a probabilistic latent variable model for switch time and rate parameters in a system of three ordinary differential equations. Thus, MultiVelo can be used to infer latent time of cells, which facilitates the enforcement of the constraints

3 Results

3.1 HALO Identifies Coupled or Decoupled Gene-Peak Pairs in Developing Mouse Brain Dataset

In this section, we show that independently changing causal module statistics are able to distinguish pairs of genes and aggregated nearby peaks with coupled or decoupled status from paired scATAC-seq and scRNA-seq normalized count data by computing the scores $\Delta_{X^A \to X^R}$ and $\Delta_{X^R \to X^A}$. Particularly, if $\Delta_{X^A \to X^R} - \Delta_{X^R \to X^A} < 0$, and $\Delta_{X^A \to X^R} < \alpha$ then we categorize the gene-peak pair as decoupled, while if $\Delta_{X^A \to X^R} > \alpha$ and $\Delta_{X^R \to X^A} > \alpha$, where $\alpha = 0.05$, the gene-peak pair is categorized as coupled Eq. (8). We evaluate this categorization on Tie4, Epha3, 1gfbpl1 and Csmd2 gene-peak pairs, whose normalized counts at different points in latent time are shown in Fig. 3a.The coupled and decoupled gene-peak pairs show different behaviors as latent time progresses. Decoupled pairs illustrate a "time-lagging" phenomenon, with trends that can be more heterogeneous. However, coupled pairs behave in a more consistent manner.

3.2 HALO Discriminates between Coupled and Decoupled Components Using Temporal Information

To assess whether HALO can accurately learn representation for coupled and decoupled events with temporal information, we apply our method to 10X Multiome embryonic (E18) mouse brain data. HALO infers coupled (c) and decoupled (d) latent representations of scATAC-seq (Z^A) and scRNA-seq Z^R data.

We infer Z^R from Z^A using a linear least squares regression fit, then evaluate the corresponding residual trends across latent time. As shown in Fig. 3b, the residuals of $Z^A \to Z^R$ change over time, even within individual cell types, illustrating the changing relationship between Z^A and Z^R. The temporally associated change is more drastic for the decoupled pair $Z^A_d \to Z^R_d$ residuals, while the coupled pair $Z^A_c \to Z^R_c$ residual curve exhibits relatively little change over time in Fig. 3b. Then, we make another inference from Z^A to the normalized RNA count matrix X^R to further evaluate the performance of our method in incorporating temporal information. The residual curve for this $Z^A \to Z^R$ regression fails to match up with the residual curve of inference from $Z^A \to$ scRNA because of dropout events in scRNA-seq data. We can still see that residual curve of $Z^A_c \to Z^R_c$ is similar to residual curve $Z^A_c \to$ RNA (Fig. 3c, top left), indicating that the representations reflect the presence of the confounder, denoted by C in Fig. 2d; $Z^A_c$and Z^R_cproperly represent the contemporaneous nature of chromatin accessibility and transcription in the coupled situation over latent time. In contrast, the residual curves using Z^A_d to infer $Z^R_d$and X^R demonstrate different slopes and are noticeably dissimilar over time (Fig. 3c, top right). The different trends of residuals in the decoupled case demonstrate the independently changing causal modules, as pictured in Fig. 2d. These results suggest that HALO can accurately distinguish decoupled and coupled information, incorporating both chromatin accessibility and transcription. When an open chromatin region and its active transcription are coupled, we can use the information from one modality to infer the other. On the other hand, when decoupling occurs, we observe modality-specific information.

We compare HALO with two other multi-omics integration methods, MultiVI and scMVP [2,18]. MultiVI and scMVP are deep generative models for integration of multi-view profiled scATAC and scRNA sequencing data. They both generate overlapping low-dimensional latent representations for downstream analysis. scMVP differs in that it utilizes a clustering consistency constraint to preserve the embedding and semantic information between the original and imputed data. Residual curves of the two inference tasks using the latent representations from scMVP or MultiVI also show slopes that differ between Z^R and X^R over latent time, showing that the latent representations from these tools do not account for time information to the same extent as HALO. Residual curves of the scMVP latent representation show very divergent trajectories, while the residual curves based on the MultiVI latent representation display divergent slopes during some time windows (Fig. 3c).

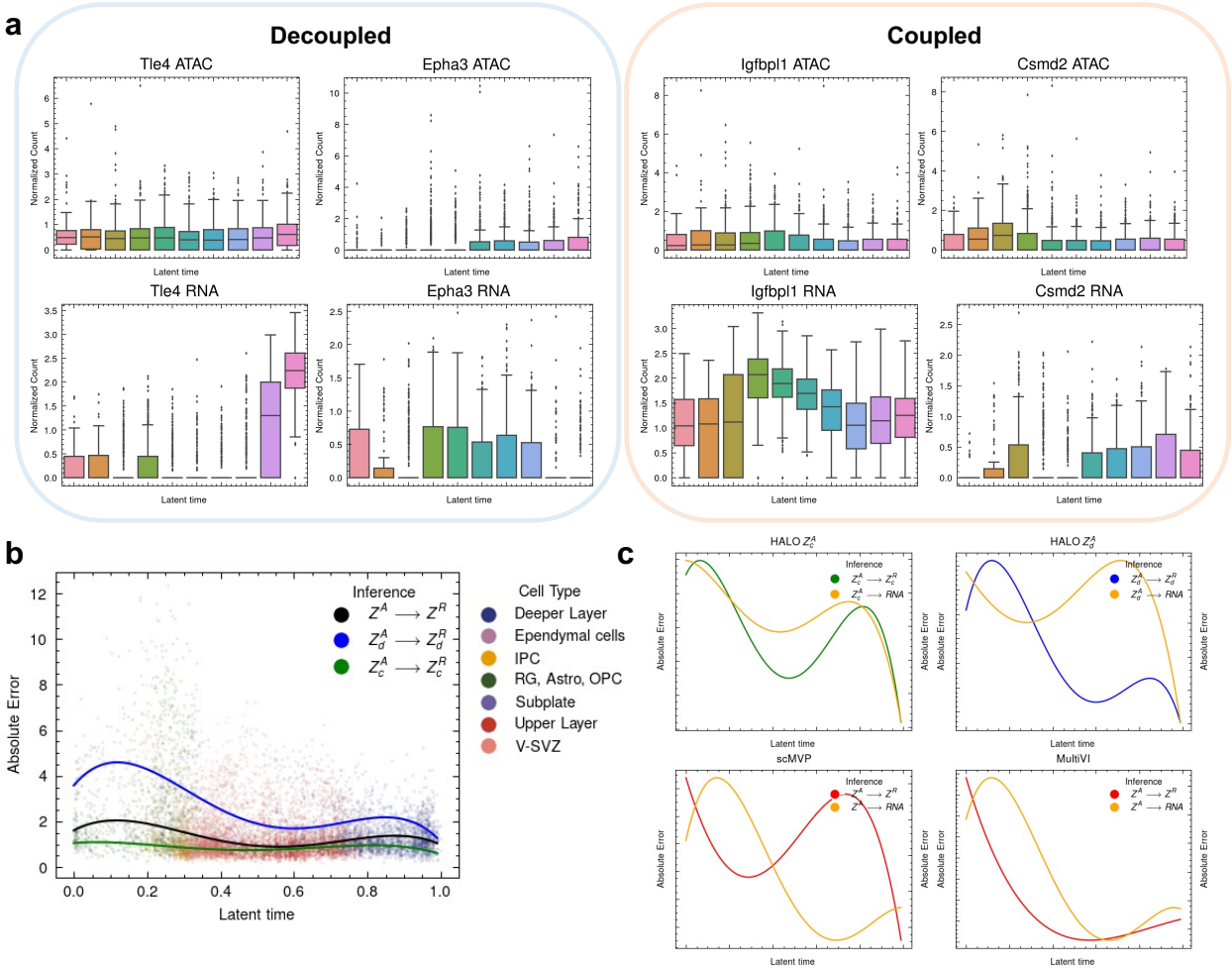

Fig. 3: HALO captures temporal information for transcriptome and epigenome states. a. Aggregated scATAC-seq peaks (top) and scRNA-seq levels (bottom) of selected decoupled (Tle4, Epha3) and coupled (Igfbpl1, Csmd2) genes. Independently changing causal module statistics were applied to individual gene-peak pairs to categorize them as decoupled or coupled. **b.** Inference of latent representations of scRNA-seq (Z^R, Z_d^R, Z_c^R) from various latent representations of scATAC-seq (Z^A, Z_d^A, Z_c^A) was performed using linear least squares regression. Residuals of these fits are plotted as a function of latent time. **c.** Comparison of HALO (top row) with other multi-omics integration methods, based on inference of Z^R and normalized scRNA-seq count matrix X^R from the latent representations Z^A of all methods. Residuals are plotted as a function of latent time as in **b**.

3.3 HALO Delineates Cell States in Embryonic Mouse Brain

We also apply HALO to developing mouse brain dataset with cell type annotation generated by [17]. Cells were clustered using the Scanpy pipeline [37] and manually annotated using canonical marker genes (Fig. 4a). Then, we perform re-clustering based on Z^R and Z^A (Fig. S1B), as a result of which deeper layer neurons are split into two clusters: cluster 3 & 4 (Fig. 4c). We postulate that these two clusters represent subtypes of deeper layer neurons. Although both cluster 3 and 4 highly express Tle4, which is a canonical marker of deep layer neurons [39], cluster 3 has significantly higher expression level of Tle4 than cluster 4 (Fig. 4d and Fig. S1A). We also observe that cluster 3 has higher expression of Robo1, Ldb2, Kirrel3 and Tox. These genes regulate neurons localized to specific areas of the brain, mostly in the lower back region: Robo1 affects the proliferation and generation of pyramidal neurons, [25,38], while Ldb2 and Kirrel3 are involved in corticospinal motor neuron and mossy fiber synapse differentiation [16,23]. Also, cluster 4 is earlier in latent time as inferred by multi-omic velocity (Fig. 4b), which could indicate our method's ability to further stratify cell types according to latent time. Our results suggest that HALO can characterize cell type at an even finer resolution and reveal development trajectory. To pinpoint which latent representation is most

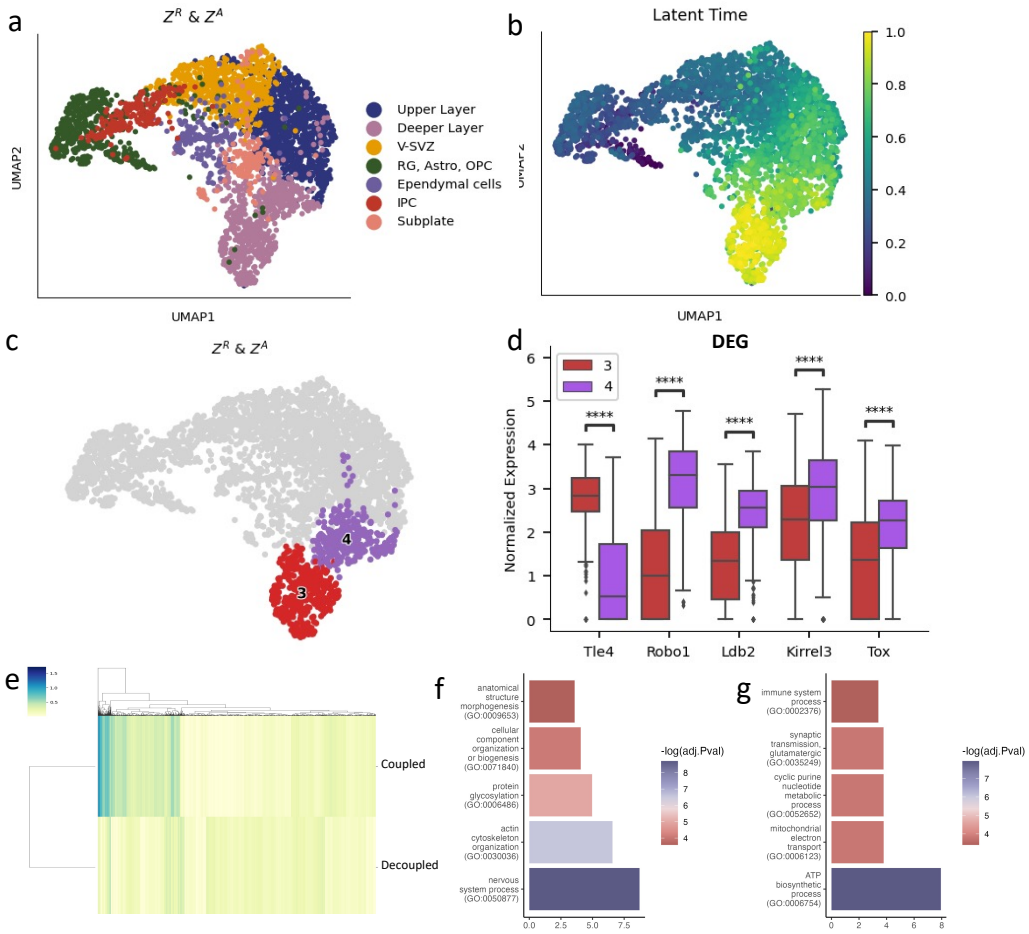

Fig. 4: HALO identifies coupled and decoupled genes, as well as subtypes of deeper layer neurons in the brains of embryonic mice. UMAP embedding visualization of neuron cells, colored by **a**. cell type annotation and **b**. latent time, where the UMAP embedding is constructed based on the latent representations of scATAC-seq and scRNA-seq, Z^A and Z^R. **c**. UMAP plot of two subtypes of deeper layer neuron cells. **d**. Boxplot of differentially expressed genes (DEGs) between the two subtypes in **c**., where four asterisks represents P values are less than 0.0001. **e**. Heatmap of coupled and decoupled scores inferred using a linear decoder [36]. **f, g**. Significant gene Ontology (GO) enrichment analysis results for **f**. decoupled genes and **g**. coupled genes.

informative for cell state depiction, we visualize clusters 3 & 4 with UMAP embeddings inferred from Z_c^A, Z_d^A, Z_c^R, Z_d^R (Figs. S1e-S1h). Cell members of cluster 3 & 4 are well-separated on the decoupled Z_d^R-based UMAP embedding (Fig. S1e), while the other latent representations were less informative in capturing more subtle cell-cell variation between deeper layer neuron cells. The decoupled information from the transcription modality (Z_d^R) also pictures undirected trajectory. In summary, we demonstrate HALO confers the advantage of utilizing both decoupled and coupled information for cell state delineation.

3.4 HALO Distinguishes Batch Effect and Biological Information

Large-scale single-cell omics datasets contain *batch effect*, which is a result of unavoidable technical or biologically irrelevant variations across batches and can further obscure downstream analysis and interpretation of the data. Moreover, batch effect can be seen as an independent exogenous variable in causal systems [40], contributing to independently changing causal modules. Accordingly, HALO is able to isolate the batch effect from entangled biological information. To evaluate the capacity of HALO to discern biological information and batch effect, we apply HALO to bone marrow mononuclear data, using annotations made by an expert committee as surrogate ground truth labels [19]. This competition benchmark dataset is generated with complex batch effects and human donor diversity. First, we train the HALO model with reconstruction

loss only, using visualizations of Z^A & Z^R to evaluate batch effect (Fig. S2). Cells on the Z^R-based UMAP embedding plot are able to be grouped by cell type (Fig. S2A), whereas Z^A is still highly affected by the (known) batch effect (Fig. S2B). During further training of the model, we include the causal constraint and evaluate the batch effect on each latent representation. The decoupled chromatin accessibility representation Z_d^A is affected by batch effect, while the coupled information representation Z_c^A instead captures biologically meaningful cell type separation between cells (Figs. S2e-i). Accordingly, using Z^R and Z_c^A only, we are able to discard batch effect on scATAC-seq data (Figs. S2a-b). When batch effect cannot be removed effectively from multi-omics data, the causal constraint enables HALO to pick out batch signal and semantic knowledge of the data for downstream analysis.

3.5 HALO Reveals Coupled and Decoupled Genes Using a Linear Decoder

In order to evaluate how chromatin accessibility/transcription coupling status relates to biological function, we identified which genes are associated with either coupled or decoupled intervals, using the same embryonic mouse brain dataset as above. We label genes as coupled or decoupled by using the respective latent representations Z_c^R and Z_d^R of scRNA-seq data as input for LDVAE [35], which uses a linear function $ZW^\top$ on the latent representation, instead of the neural network $f_W(Z)$ in the scVI generative model. This ensures that the expression level of the reconstructed data is affected by the weights in the loading matrix W through whether it is in the coupled or decoupled row. Intuitively, we determine whether (the latent vectors of) the coupled or the decoupled representations contribute more to the reconstructed scRNA-seq output, thereby assigning the label of coupled or decoupled to each gene. This assignment is done by thresholding the scores (Fig. 4e), i.e., genes that had a coupled score > 0.5 and decoupled score < 0.1 were labeled as coupled, and vice versa for the decoupled genes.

For the biological significance of these genes, specifically pertaining to the coupled and decoupled contingents, we used gene ontology (GO) enrichment analysis [9] to obtain annotations for the coupled and decoupled gene sets. Significant terms for the decoupled genes are related to lineage development (Fig. 4f), such as anatomical structure morphogenesis, actin cytoskeleton organization, which has been shown to have a role in most aspects of neuronal differentiation and function [4], as well as glycosylation, which is crucial for proper mammalian brain development [24]. While a cell is undergoing differentiation, it has more dynamic chromatin access and transcription globally with larger or more frequent fluctuations in transcriptional programs, hence it has a higher association with time lag events (decoupled intervals). The coupled genes are related to various aspects of metabolism (Fig. 4g), e.g., ATP biosynthesis and metabolic processes associated with purine, which is highly produced during proliferation and whose regulation affects the cell cycle during brain development [8]. During the cell cycle, the cell's chromatin remodeling and transcription are more stable, so any transcription is more likely to be paired with specific chromatin opening events (coupled intervals). These associations of coupled and decoupled gene sets with biological functions capture the dynamic developmental attributes of the dataset, and they are consistent with results from another study modeling regulatory potential of individual gene loci using the same dataset [20]. Since the decoupled representation in our bone marrow analysis displayed batch effect or noise unrelated to cell state (Fig. S2H), we used only the coupled genes (Fig. S3A) from erythroblast lineage cells to perform GO enrichment analysis. Strikingly, we find further evidence supporting the results from the embryonic mouse brain data: the five most significant GO terms are all related to cell cycle regulation (Fig. S3B). In CD8 lymphocytes from peripheral blood mononuclear cell (PBMC) data, we identified the coupled and decoupled genes using HALO as above. Again, we find that coupled genes are enriched for metabolic and cell cycle processes (Fig. S4A), including DNA replication.

Genetic redundancy, or the presence of multiple genes performing the same function, is an evolutionary strategy to ensure the robustness of an organism's vital functions. Given the key role of the global, cell cycle terms associated with coupled genes in cell survival, it is reasonable that we find the number of coupled genes (329) to be severalfold higher than the number of decoupled genes (89) (Fig. 4e). Although there are fewer decoupled genes, they could be individually more important for explaining the mechanisms of decoupling, especially in particular differentiation lineages, and may warrant more detailed experimental study. Together, these results indicate that our method is able to distinguish genes associated with either coupling or decoupling of chromatin accessibility and transcription, as well as demonstrate that coupling is associated with metabolic reactions and the cell cycle, whereas decoupled genes are linked to cell structural changes and differentiation.

4 Discussion

In this paper, we propose HALO, a novel method for representation learning of jointly profiled scATAC-seq and scRNA-seq data, which advances the field by considering both coupled and uncoupled peak-gene causal relations in the two modalities over time. In particular, HALO enforces hierarchical causal relations between the chromatin accessibility and transcription in a latent representation learning scenario. The latent spaces are further decomposed into coupled and decoupled components, which we find to be significantly associated with cell cycle and cell fate dynamic processes, respectively. Results using experimental datasets show that HALO can effectively capture the temporal information, biological mechanisms, and cell subtypes. The major limitation of HALO is that it only considers the ATAC $\rightarrow$ mRNA causal direction. It is also possible that transcription may itself cause chromatin accessibility changes, because the components of chromatin remodeling complexes are also translated from mRNA. Although this seems to introduce a cycle in the corresponding causal graph, in reality these "feedback" effects occur at different time points. Also, the collected data represent a snapshot of epigenome and transcriptome, and the time scale is too short to observe such a phenomenon.

References

1. Abecassis, I., Sedgewick, A.J., Romkes, M., Buch, S., Nukui, T., Kapetanaki, M.G., Vogt, A., Kirkwood, J.M., Benos, P.V., Tawbi, H.: PARP1 rs1805407 increases sensitivity to PARP1 inhibitors in cancer cells suggesting an improved therapeutic strategy. Sci. Rep. **9**(1), 3309 (Mar 2019)
2. Ashuach, T., Gabitto, M.I., Jordan, M.I., Yosef, N.: Multivi: deep generative model for the integration of multimodal data. bioRxiv (2021)
3. Caparosa, E.M., Sedgewick, A.J., Zenonos, G., Zhao, Y., Carlisle, D.L., Stefaneanu, L., Jankowitz, B.T., Gardner, P., Chang, Y.F., Lariviere, W.R., LaFramboise, W.A., Benos, P.V., Friedlander, R.M.: Regional molecular signature of the symptomatic atherosclerotic carotid plaque. Neurosurgery **85**(2), E284–E293 (Aug 2019)
4. Cheever, T.R., Ervasti, J.M.: Actin isoforms in neuronal development and function. International Review of Cell and Molecular Biology **301**, 157–213 (2013)
5. Chen, S., Lake, B.B., Zhang, K.: High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature Biotechnology **37**(12), 1452–1457
6. Clapier, C.R., Iwasa, J., Cairns, B.R., Peterson, C.L.: Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nature Reviews Molecular Cell Biology **18**(7), 407–422 (May 2017)
7. Feng, F., Huang, B., Zhang, K., Magliacane, S.: Factored adaptation for non-stationary reinforcement learning. arXiv preprint arXiv:2203.16582 (2022)
8. Garcia-Gil, M., Camici, M., Allegrini, S., Pesi, R., Petrotto, E., Tozzi, M.G.: Emerging role of purine metabolizing enzymes in brain function and tumors. International Journal of Molecular Sciences **19**(11), 3598 (2018)
9. Gene Ontology Consortium: The gene ontology resource: enriching a GOld mine. Nucleic Acids Res. **49**(D1), D325–D334 (Jan 2021)
10. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International conference on algorithmic learning theory. pp. 63–77. Springer (2005)
11. Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. Advances in neural information processing systems **20** (2007)
12. Huang, B., Lu, C., Leqi, L., Hernandez-Lobato, J.M., Glymour, C., Schölkopf, B., Zhang, K.: Action-sufficient state representation learning for control with structural constraints. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 9260–9279. PMLR (17–23 Jul 2022)
13. Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., Schölkopf, B.: Causal discovery from heterogeneous/nonstationary data. J. Mach. Learn. Res. **21**(1) (jan 2020)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., Zhang, K.: Partial disentanglement for domain adaptation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 11455–11472. PMLR (17–23 Jul 2022)
16. Leone, D.P., Panagiotakos, G., Heavner, W.E., Joshi, P., Zhao, Y., Westphal, H., McConnell, S.K.: Compensatory actions of ldb adaptor proteins during corticospinal motor neuron differentiation. Cereb. Cortex **27**(2), 1686–1699 (Feb 2017)
17. Li, C., Virgilio, M.C., Collins, K.L., Welch, J.D.: Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. Nature Biotechnology (Oct 2022)
18. Li, G., Fu, S., Wang, S., Zhu, C., Duan, B., Tang, C., Chen, X., Chuai, G., Wang, P., Liu, Q.: A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. Genome Biol. **23**(1), 20 (Jan 2022)
19. Luecken, M., Burkhardt, D., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A., Deconinck, L., Detweiler, A., Granados, A., Huynh, S., Isacco, L., Kim, Y., Klein, D., DE KUMAR, B., Kuppasani, S., Lickert, H., McGeever, A., Melgarejo, J., Mekonen, H., Morri, M., Müller, M., Neff, N., Paul, S., Rieck, B., Schneider, K., Steelman, S., Sterr, M., Treacy, D., Tong, A., Villani, A.C., Wang, G., Yan, J., Zhang, C., Pisco, A., Krishnaswamy, S., Theis, F., Bloom, J.M.: A sandbox for prediction and integration of dna, rna, and proteins in single cells. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
20. Lynch, A.W., Theodoris, C.V., Long, H.W., Brown, M., Liu, X.S., Meyer, C.A.: Mira: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. Nature Methods **19**(9), 1097–1108 (2022)
21. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., Law, T., Lareau, C., Hsu, Y.C., Regev, A., Buenrostro, J.D.: Chromatin potential identified by shared single-cell profiling of rna and chromatin. Cell **183**(4), 1103–1116.e20 (2020)
22. Mao, H., Jia, M., Dou, J.X., Zhang, H., Benos, P.V.: Coem: Cross-modal embedding for metacell identification (2022). https://doi.org/10.48550/ARXIV.2207.07734
23. Martin, E.A., Muralidhar, S., Wang, Z., Cervantes, D.C., Basu, R., Taylor, M.R., Hunter, J., Cutforth, T., Wilke, S.A., Ghosh, A., Williams, M.E.: The intellectual disability gene kirrel3 regulates target-specific mossy fiber synapse development in the hippocampus. Elife **4**, e09395 (Nov 2015)
24. Medina-Cano, D., Ucuncu, E., Nguyen, L.S., Nicouleau, M., Lipecka, J., Bizot, J.C., Thiel, C., Foulquier, F., Lefort, N., Faivre-Sarrailh, C., et al.: High n-glycan multiplicity is critical for neuronal adhesion and sensitizes the developing cerebellum to n-glycosylation defect. Elife **7**, e38309 (2018)

25. Nural, H.F., Todd Farmer, W., Mastick, G.S.: The slit receptor robo1 is predominantly expressed via the dutt1 alternative promoter in pioneer neurons in the embryonic mouse brain and spinal cord. Gene Expr. Patterns **7**(8), 837–845 (Oct 2007)
26. Pearl, J., Bareinboim, E.: Transportability of causal and statistical relations: A formal approach. In: Proc. AAAI 2011. pp. 247–254 (2011)
27. Pearl, J.: Causality. Cambridge University Press, 2 edn. (2009)
28. Raghu, V.K., Beckwitt, C.H., Warita, K., Wells, A., Benos, P.V., Oltvai, Z.N.: Biomarker identification for statin sensitivity of cancer cell lines. Biochem. Biophys. Res. Commun. **495**(1), 659–665 (Jan 2018)
29. Raghu, V.K., Ramsey, J.D., Morris, A., Manatakis, D.V., Sprites, P., Chrysanthis, P.K., Glymour, C., Benos, P.V.: Comparison of strategies for scalable causal discovery of latent variable models from mixed data. Int. J. Data Sci. Anal. **6**(1), 33–45 (Feb 2018)
30. Raghu, V.K., Zhao, W., Pu, J., Leader, J.K., Wang, R., Herman, J., Yuan, J.M., Benos, P.V., Wilson, D.O.: Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. Thorax **74**(7), 643–649 (Jul 2019)
31. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
32. Sedgewick, A.J., Buschur, K., Shi, I., Ramsey, J.D., Raghu, V.K., Manatakis, D.V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., Sciurba, F.C., Spirtes, P., Glymour, C., Benos, P.V.: Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics **35**(7), 1204–1212 (Apr 2019)
33. Sedgewick, A.J., Buschur, K., Shi, I., Ramsey, J.D., Raghu, V.K., Manatakis, D.V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., Sciurba, F.C., Spirtes, P., Glymour, C., Benos, P.V.: Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics **35**(7), 1204–1212 (Apr 2019)
34. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
35. Svensson, V., Gayoso, A., Yosef, N., Pachter, L.: Interpretable factor models of single-cell rna-seq via variational autoencoders. Bioinformatics **36**(11), 3418–3421 (2020)
36. Svensson, V., Pachter, L.: Interpretable factor models of single-cell rna-seq via variational autoencoders. bioRxiv (2019)
37. Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome Biology **19**(1), 15 (Feb 2018)
38. Yeh, M.L., Gonda, Y., Mommersteeg, M.T.M., Barber, M., Ypsilanti, A.R., Hanashima, C., Parnavelas, J.G., Andrews, W.D.: Robo1 modulates proliferation and neurogenesis in the developing neocortex. J. Neurosci. **34**(16), 5717–5731 (Apr 2014)
39. Zahr, S.K., Yang, G., Kazan, H., Borrett, M.J., Yuzwa, S.A., Voronova, A., Kaplan, D.R., Miller, F.D.: A translational repression complex in developing mammalian neural stem cells that regulates neuronal specification. Neuron **97**(3), 520–537.e6 (Feb 2018)
40. Zhang, K., Gong, M., Stojanov, P., Huang, B., Liu, Q., Glymour, C.: Domain adaptation as a problem of inference on graphical models. In: Neural Information Processing Systems (NeurIPS) (2020)
41. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: ICML-13 (2013)

1 Appendix I: Supplementary Figures

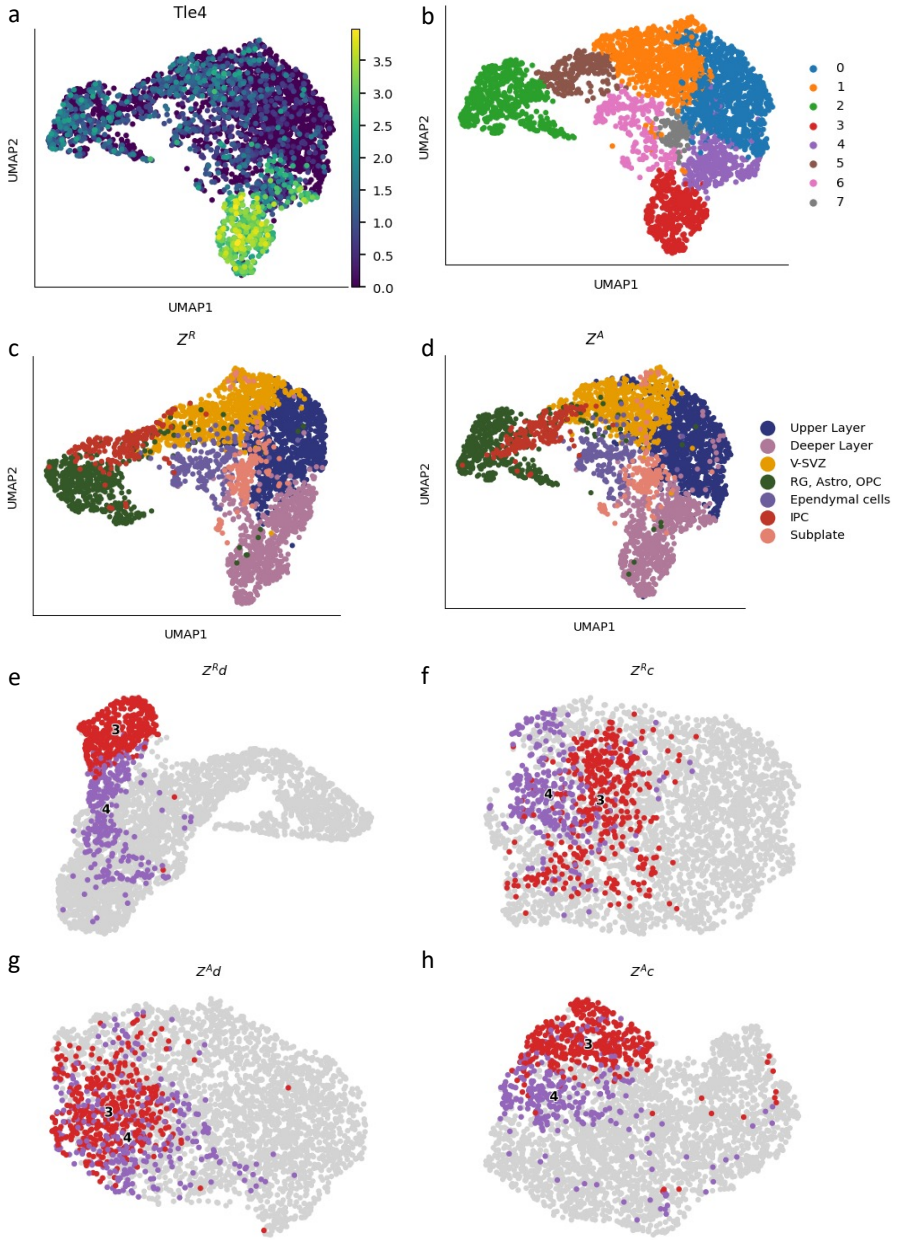

Fig. S1: **HALO identifies two subtypes from deep layer neuron cells. a-b**. UMAP plot of neuron cells colored by gene expression of Tle4 (canonical marker of deep layer neurons) and cluster assignment, where the UMAP embedding construction and clustering are performed based on the latent representations Z^R and Z^A of the two omics modalities. **c-d**. Neuron cells are visualized with UMAP embedding constructed from Z^R and Z^A. **e-h**. Two subtypes of deeper layer neuron cells are visualized using UMAP embeddings constructed from Z_c^A, Z_d^A, Z_c^R, Z_d^R.

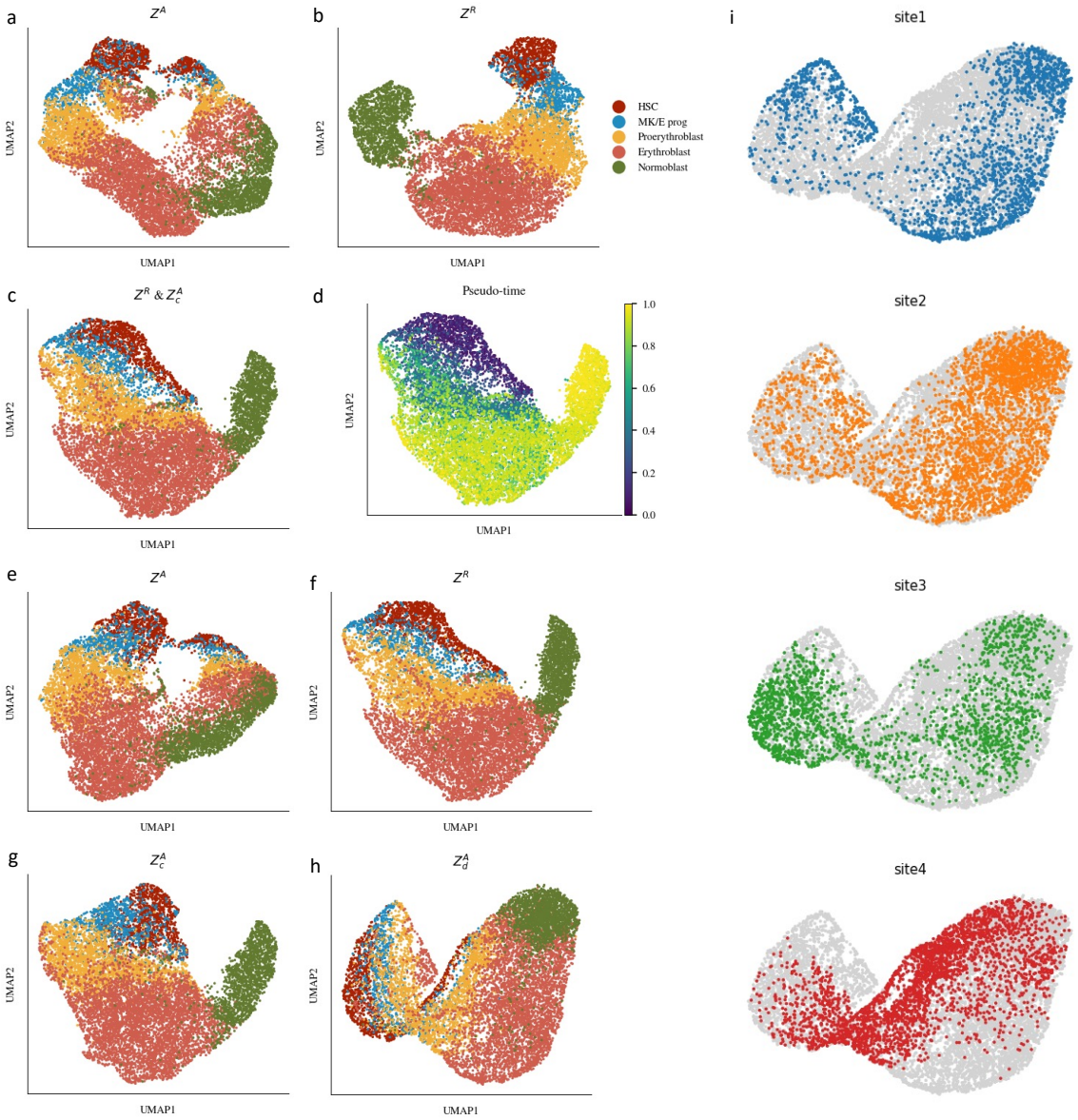

Fig. S2: **HALO can distinguish biological information and batch effect in bone marrow mononuclear cells with causal constraints. a-b**. UMAP embedding of latent representations inferred with reconstruction loss only. After including causal constraints into our model: **c-d**. UMAP embedding is generated from Z^R and Z_c^A to visualize cells colored by cell type annotation and pseudotime, **e-h**. latent representations Z_c^A, Z_d^A, Z_c^R, and Z_d^R, are used for UMAP visualization, **i**. Z_d^A-based UMAP plot of cells is colored by batch information.

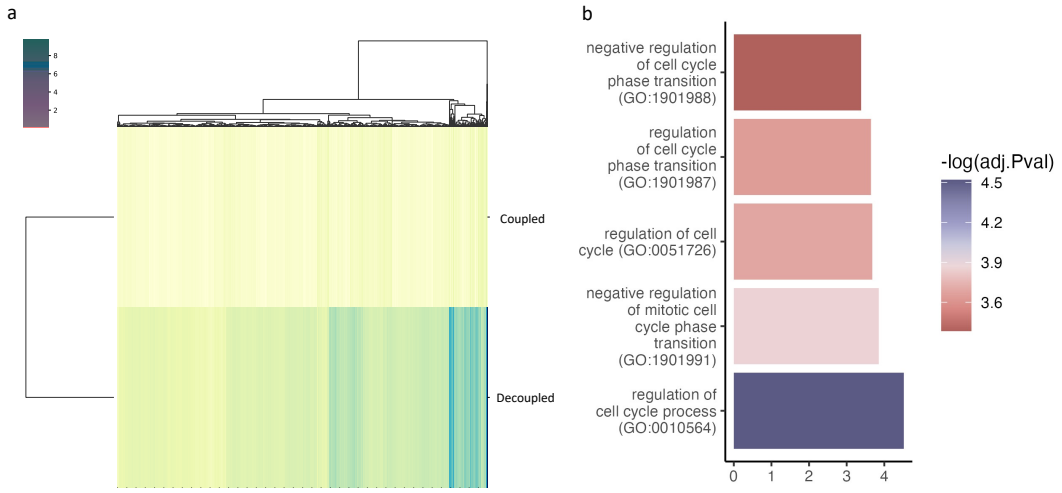

Fig. S3: **Identification of coupled and decoupled genes in erythroblast lineage cells from bone marrow. a.** Decoupled and coupled scores inferred from a linear decoder. Decoupled genes suffer from batch effects and were excluded from GO analysis. **b.** Top significant terms from gene ontology enrichment analysis of coupled genes.

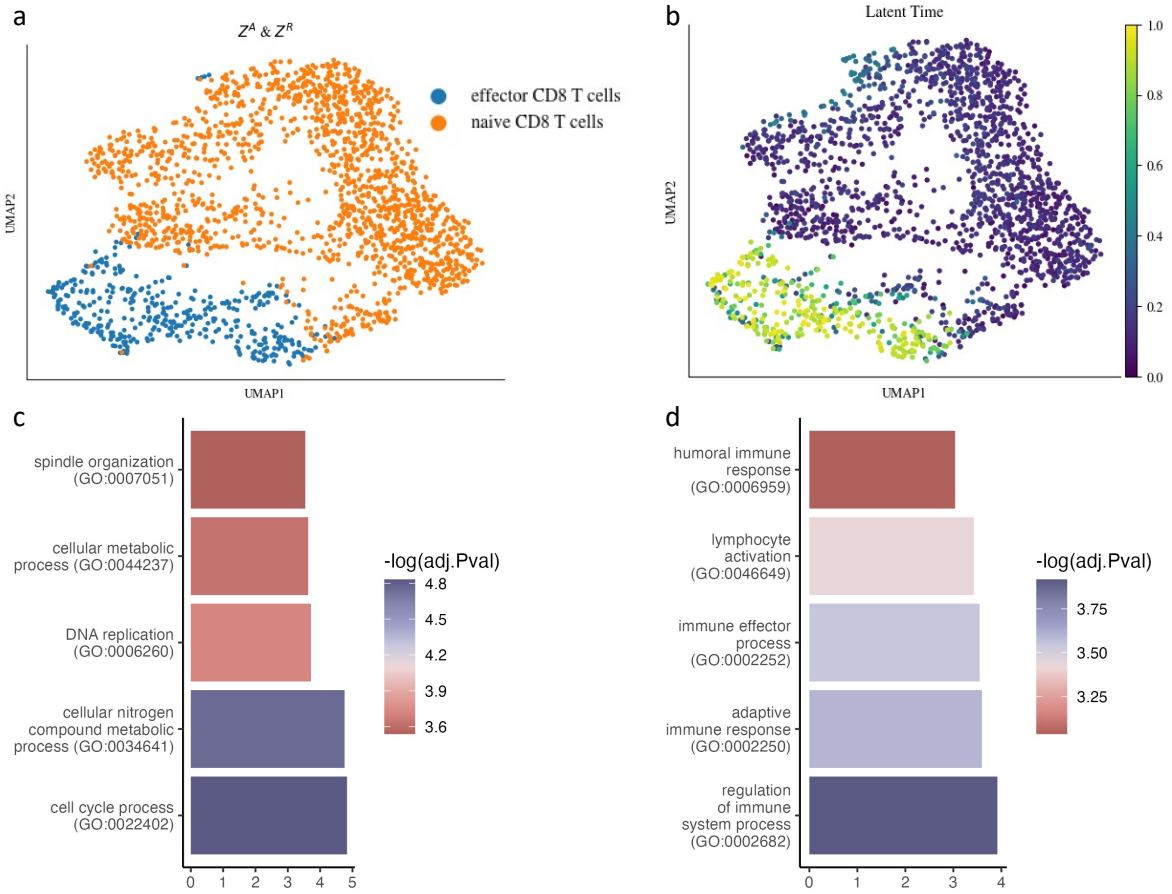

Fig. S4: **Identification of coupled and decoupled genes of CD8 T lymphocytes from 10X PBMC multiome data.** UMAP visualization of CD8 T lymphocytes, colored by **a.** cell label and **b.** latent time. Gene ontology enrichment analysis of **c.** coupled genes and **d.** decoupled genes.

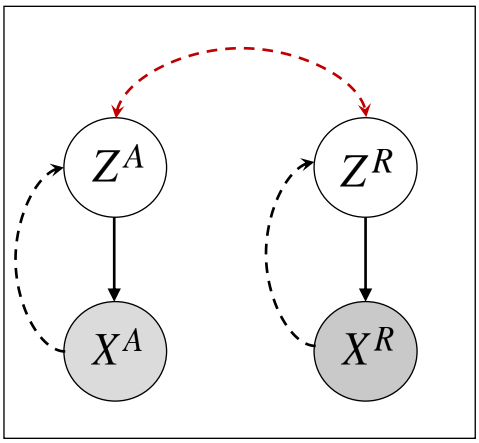

Fig. S5: **Graphical model of our modified auto-encoder.** Solid lines are used to denote the generative model, black dashed lines denote the variational approximation, and the red dashed line denotes our causal constraints.

2 Appendix II: Datasets and Data Preprocessing

We use both human and mouse multiome datasets from different tissues to evaluate our method: (1) human peripheral blood mononuclear cells (PBMCs) multiome data from 10X Genomics, (2) embryonic (E18) mouse brain multiome data from 10X Genomics, and (3) human bone marrow mononuclear cells from the NeurIPS competition training data [4]. For the PBMC dataset, we use the cell type annotation generated by the 10X Genomics R&D team as surrogates for the ground truth [3]. Similarly, we used cell type annotations generated by [2] and [4] for the mouse brain dataset and bone marrow dataset, respectively. For temporal information, we utilize diffusion pseudotime annotation from the bone marrow dataset, while we infer latent time through RNA velocity analysis [1,2] for the mouse brain and PBMC datasets. To visualize and determine cell type, we use the Python package Scanpy to generate UMAP embeddings, perform clustering, and rank genes for clusters of interest [5].

References

1. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., Theis, F.J.: Generalizing rna velocity to transient cell states through dynamical modeling. Nature Biotechnology **38**(12), 1408–1414 (Dec 2020)
2. Li, C., Virgilio, M.C., Collins, K.L., Welch, J.D.: Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. Nature Biotechnology (Oct 2022)
3. Liu, Q., Chen, S., Jiang, R., Wong, W.H.: Simultaneous deep generative modelling and clustering of single-cell genomic data. Nature Machine Intelligence **3**(6), 536–544 (Jun 2021)
4. Luecken, M., Burkhardt, D., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A., Deconinck, L., Detweiler, A., Granados, A., Huynh, S., Isacco, L., Kim, Y., Klein, D., DE KUMAR, B., Kuppasani, S., Lickert, H., McGeever, A., Melgarejo, J., Mekonen, H., Morri, M., Müller, M., Neff, N., Paul, S., Rieck, B., Schneider, K., Steelman, S., Sterr, M., Treacy, D., Tong, A., Villani, A.C., Wang, G., Yan, J., Zhang, C., Pisco, A., Krishnaswamy, S., Theis, F., Bloom, J.M.: A sandbox for prediction and integration of dna, rna, and proteins in single cells. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
5. Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome Biology **19**(1), 15 (Feb 2018)