HALO: Hierarchical Causal Modeling for Single Cell Multi-Omics Data

Authors: Haiyi Mao†[1,2], Minxue Jia†[1,2], Marissa Di[1,2], Eleanor Valenzi[3], Xiaoyu Tracy Cai[4], Harinder Singh[5], Robert Lafyatis[6], Kun Zhang[7], *Panagiotis Benos[1,2,8]

Affiliations:

[1]Department of Computational and Systems Biology, University of Pittsburgh.
[2]Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology.
[3]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, University of Pittsburgh.
[4]Immunology Discovery and Regenerative Medicine, Genentech, Inc.
[5]Department of Immunology, University of Pittsburgh.
[6]Division of Rheumatology and Clinical Immunology, University of Pittsburgh.
[7]Department of Philosophy, Carnegie Mellon University.
[8]Department of Epidemiology, University of Florida.
†Equal Contributions
*Corresponding author. Email: pbenos@ufl.edu

Abstract: Though open chromatin may promote active transcription, gene expression responses may not be directly coordinated with changes in chromatin accessibility. Most existing methods for single-cell multi-omics data focus only on learning stationary and shared information among these modalities, overlooking modality-specific information delineating cellular states and dynamics resulting from causal relations among modalities. To account for this, the epigenome and transcriptome relationship can be characterized in relation to time as "coupled" (changing dependently) or "decoupled" (changing independently). We propose the framework HALO, which adopts a causal approach to model these temporal causal relations on two levels. On the representation level, HALO factorizes these two modalities into both coupled and decoupled latent representations, identifying the dynamic interplay between chromatin accessibility and transcription through temporal modulations in the latent space. On the individual gene level, HALO matches gene-peak pairs and characterizes changing dynamics between gene expression and local peaks with time. HALO reveals bipotency in a subset of AT2 cells that exhibit different decisions in lineage specification between systemic sclerosis (SSc) and normal conditions. We demonstrate that using coupled and decoupled information, HALO discovers analogous biological functions between modalities, distinguishes epigenetic factors for lineage specification, and identifies temporal *cis*-regulation interactions relevant to cellular differentiation and complex human diseases.

Introduction

Single-cell multi-omics technologies have revolutionized our understanding of cellular heterogeneity and complexity, enabling the simultaneous measurement of diverse molecular layers such as transcriptomics, epigenomics, and proteomics within individual cells. These technologies provide a unique opportunity to dissect the regulatory relationships underlying cellular functions and phenotypes. However, integration and analysis of such heterogeneous data types remain challenging, for example in distinguishing the effects of chromatin accessibility on expression across varying cellular states and over time. To address these challenges, we developed HALO (**H**ierarchical C**A**usal Modeling for Single Cell Mu**L**ti-**O**mics Data), a computational framework designed to model the causal relationships within single-cell co-profiled multi-omic data. The key hypothesis of HALO is that changes in chromatin accessibility causally influence gene expression. By leveraging measurements of transcriptomics and chromatin accessibility data from single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq), HALO enables a comprehensive analysis of the causal interactions that dictate cell state, function, and regulatory mechanisms.

Chromatin structure plays a crucial role in regulating the ability of transcriptional machinery to access DNA and activate gene expression. Open chromatin facilitates the binding of transcription factors, the recruitment of RNA polymerases, and the initiation of transcription. Consequently, gene expression and chromatin accessibility often correlate and exhibit similar dynamics across different cellular states or over time. However, they do not always display the same patterns due to several biological regulatory factors. For example, chromatin can be accessible, but the associated gene or regulatory region may not be immediately transcribed—a state known as chromatin priming. This primed state allows cells to respond more quickly to environmental changes or developmental signals. In this state, cells have the potential to differentiate into various cell types and adapt to different conditions[1]. Additionally, gene expression can be controlled post-transcriptionally through mechanisms such as mRNA stability, degradation, or translation efficiency. Meanwhile, active chromatin remodeling may occur independently of transcriptional changes, preparing the chromatin landscape for future gene activation or repression[2]. Previous representation learning methods for single-cell multi-omics data implicitly rely on the strict biological assumption that chromatin accessibility and active transcription are synchronized[3–6], though some recent methods consider the case where chromatin opening precedes transcription initiation[2,7]. In contrast to these works, HALO decomposes the relations between chromatin accessibility and gene expression into coupled and decoupled cases. In the coupled case, chromatin accessibility and gene expression exhibit correlated changes over time, while in the decoupled case, they change independently over time. Due to the high dimensionality and sparsity of single cell genomics data, HALO operates on two levels of analysis: the low-dimensional latent representation level and the individual gene level, incorporating real time points/estimated latent time to account for temporal dynamics in cellular development. The framework is equipped with an interpretable neural network to provide biological meaning to the latent representations. Moreover, HALO employs Granger causality to assess context-specific distal cis-regulation in cases where, despite associated chromatin regions becoming more accessible, gene transcription does not increase correspondingly[7]. We observe these situations frequently occur when the chromatin regions overlap with super enhancer regions.

Results

HALO: a causal machine learning framework to model the interactions between chromatin accessibility and gene expression. HALO models the co-assayed gene expression and chromatin accessibility in a low-dimensional latent space as well as individual gene and its linked peaks (open chromatin regions) level through a causal lens. In our study, we establish a framework to analyze the causal relationships between jointly profiled scRNA-seq and single-cell ATAC sequencing (scATAC-seq) data, incorporating temporal information. We hypothesize that scATAC-seq data, which is indicative of chromatin accessibility, causally precedes scRNA-seq data due to open chromatin regions influencing gene transcription. Specifically, we distinguish between two types of causal interactions: 1) In the **coupled** case, gene expression and chromatin accessibility exhibit *dependent* changes over time, indicating they are influenced by shared unknown confounders. 2) In the **decoupled** case, certain gene expressions and their local peak patterns change *independently* over time, suggesting the presence of distinct, unknown causal factors affecting gene expressions. This approach extends the concept of time-lagging observed in previous studies[2,8] to a broader framework, encompassing mechanisms that change independently or dependently over time; for example, chromatin open regions may become more accessible while gene expression level remains stable. This approach aims to elucidate causal relationships between gene expression and chromatin accessibility at both the levels of representations and individual genes, as depicted in **Figure 1A**.

On the representation level, we model the lower dimensional latent space interactions between ATAC and RNA modalities from a causal perspective. Specifically, HALO learns the representations that contain the "coupled" information between scATAC-seq $\mathbf{X}^{\mathrm{A}} \in \mathbb{Z}^{n_{\mathrm{cells}} \times n_{\mathrm{peaks}}}$ and scRNA-seq $\mathbf{X}^{\mathrm{R}} \in \mathbb{Z}^{n_{\mathrm{cells}} \times n_{\mathrm{genes}}}$ data as well as the independent "decoupled" information, such that the causal relations are preserved even across time (**Figure 1B**). Coupled representations encapsulate information where gene expression changes are dependent on chromatin accessibility over time, reflecting shared information across modalities. In contrast, decoupled representations extract information where gene expression changes independently of chromatin accessibility over time, emphasizing modality-specific information. The latent representations of scATAC-seq data can be decomposed to $\mathbf{Z}^{\mathrm{A}} = [\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}, \mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}]$. The latent representations of scRNA-seq data $\mathbf{Z}^{\mathrm{R}}$ may be similarly decomposed: $\mathbf{Z}^{\mathrm{R}} = [\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}, \mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}]$. $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ represent the decoupled latent representations derived from scATAC-seq and scRNA-seq data while $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ denote the coupled latent representations for scATAC-seq and scRNA-seq data, respectively. These representations capture distinct aspects of information from the two modalities, each varying over time or along a latent temporal variable. **Figure 1A** illustrates the underlying causal relationship, wherein $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ causes $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$, and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ causes $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$. However, $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ are influenced by independent causative factors, which are temporally related. This implies that, despite their causal linkage, $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ have independent causal factors, leading to independent temporal changes. In contrast to the decoupled representations, both $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ are influenced by common latent confounders, which also vary with time. This indicates that the changes in $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ are synchronized over time, reflecting their shared temporal dynamics.

To ensure the delineation of these relationships, HALO utilizes paired scRNA-seq and scATAC-seq data as its input. We employ two distinct encoders to derive the latent representations $\mathbf{Z}^{\mathrm{A}}$ and $\mathbf{Z}^{\mathrm{R}}$. Moreover, we have formulated specific decoupled and coupled causal constraints (**see**

Methods Section 3). A Multi-Layer Perceptron (MLP) is used to model the concept that $\mathbf{Z}_d^A$ causes $\mathbf{Z}_d^R$, while a decoupled constraint enforces the independent functional relations between them. Similarly, another MLP is used to align the coupled representations $\mathbf{Z}_c^A$ to $\mathbf{Z}_c^R$, but with a coupled constraint (**See Methods Section 4**). Additionally, we developed a nonlinear interpretable decoder (**Figure 1B**) that allows us to interpret the latent representations by decomposing the reconstruction of genes or peaks into additive contributions from individual representations **(See Methods Section 5**).

At the individual gene level, our approach begins with the use of denoised gene expressions and peaks (**Figure 1C**), applying negative binomial regression to correlate local peaks with gene expression. This method allows us to match local peaks to corresponding gene expression enabling the subsequent calculation of decouple and couple scores at the individual gene level. Specifically, the decouple score quantifies the independence of gene expression level changes in relation to local peaks over time. Conversely, the couple score evaluates the extent to which gene expression changes are dependent on local peaks throughout the temporal course (**See Methods Section 6.3**). Finally, through the application of Granger causality analysis, we explore the underlying mechanisms of distal peak-gene regulatory interactions. This analytical approach allows us to elucidate instances where local peaks increase, yet corresponding gene expression remains largely unchanged.

In summary, HALO offers several key contributions to the field of multi-omics analysis, which we enumerate as follows: (1) HALO learns latent representations that are causally informed, enhancing our understanding of the interactions across different omics modalities. (2) HALO enables learning of interpretable latent representations. (3) HALO causally characterizes the relations between gene expression and associated chromatin regions over time, unlocking insights about regulatory dynamics. (4) HALO identifies distal *cis*-regulation interactions between chromatin regions and nearby genes, specifically for chromatin regions overlapping with super enhancers.

HALO effectively separates coupled and decoupled representations, enhancing the analysis and interpretation of modality-shared and modality-specific information in mouse skin hair follicle data. Through the integration of information from both ATAC-seq and RNA-seq modalities, HALO is able to discern cell types and capture latent temporal dynamics. The ATAC and RNA representations from HALO are concatenated and visualized according to cell type (**Figure 2A, Figure S1A and S2**) and latent time (**Figure S1C**). **Figure 2B and F** display the UMAP projections constructed by the coupled RNA and ATAC representations, $\mathbf{Z}_c^A$ and $\mathbf{Z}_c^R$, respectively. In accordance with the concept of coupled representations, $\mathbf{Z}_c^A$ and $\mathbf{Z}_c^R$capture analogous information across the two modalities, which is reflected in the resemblance between their UMAP visualizations. However, the decoupled representations $\mathbf{Z}_d^A$ and $\mathbf{Z}_d^R$ convey different information. In particular, $\mathbf{Z}_d^R$ retains cell type information, while $\mathbf{Z}_d^A$ does not (**Figure 2C and G)**.

Examining the analysis of the decoupled ATAC representation $\mathbf{Z}_d^A$, we performed clustering based on it. Cluster 4 (**Figure 2D)** is notably characterized by the decoupled ATAC latent representation 14 (**Figure 2H)**. Additionally, we have mapped cluster 4 onto the original

SHARE-seq UMAP projection (**Figure 2E** and **Figure S1B**), revealing that cluster 4 corresponds to the novel root cells[8]. These novel roots cells were previously identified by chromatin potential, which is a quantitative measure of chromatin lineage-priming and used for cell fate prediction[8]. Moreover, it is also demonstrated that novel roots (cluster 4) express distinct marker genes in comparison to expected root cells (**Figure 2K).** To validate this novel cell state in HALO's framework, we evaluated the decoupled ATAC and RNA representations and found that, consistent with the definition of chromatin potential[8], the cosine distance between the decoupled representations is maximal within the novel roots (**Figure 2I - J)**. By utilizing the interpretable decoder, HALO can identify the enriched transcription factors (TFs) within specific latent ATAC representations. **Figure 2N** shows the significantly enriched TFs in coupled representation 3 and decoupled representation 14, including Wt1 and Pou5fl. Transcription factors Wt1 and Pou5f1 are important in the Wnt/β-Catenin signaling pathway, exerting regulatory control over Lef1 and Dach1[9,10], which serve as marker genes for novel root cells. In coupled ATAC representation 3, the transcription factor Runx1 is enriched, which directly promotes the proliferation of hair follicle stem cells and affects hair morphogenesis and differentiation. Coupled RNA representation 3, which is highly correlated with coupled ATAC representation 3, captures Eda, Wnt, and Sonic hedgehog (Shh) signaling pathways, which are important for hair follicle morphogenesis (**Figure 2P**)[11].

To validate the relations between the coupled and decoupled representations of the two modalities, we show the Pearson correlation matrix for the latent representations (**Figure 2L**), which reveals a strong correlation between the ATAC and RNA coupled representations. In contrast, the decoupled representations exhibit a weaker correlation. Additionally, we assessed the correlations of latent representations between modalities using various representation learning methods, including MIRA[4], GLUE[6], MultiVI[5], and scMVP[12]. The comparative analysis (**Figure 2M, S1D)** indicates that HALO's decoupled representations maintain the lowest correlations, whereas its coupled representations are highly correlated across modalities with other four distinct datasets: Mouse Brain from 10X genomics, NEAT-Seq[13], NeurIPS[14] and Systemic sclerosis (SSc) pulmonary epithelium. We further interrogate whether the decoupled representations identified by HALO represent modality-specific batch bias or truly modality-specific biological information. To address this, we compared HALO's performance in batch correction against that of other frameworks on the NeurIPS datasets[14], which are known to contain substantial batch effects (**Figure S1E**), by silhouette score[15] and Hilbert-Schmidt Independence Criterion (HSIC) (**see Methods Sections 8.2 & 8.3**). HALO is particularly effective at removing batch information, thereby confirming that its decoupled representations predominantly capture modality-specific biological information rather than mere batch effects.

To further examine our approach, we have developed a genomic matching score (**see Methods Section 8.1**) for coupled RNA and ATAC representation pairs that assesses the distance between their important genes and peaks by calculating the ratio of peaks that are located within the *cis*-regulation regions of the genes. **Figure 2O** presents genomic matching scores for the coupled ATAC-RNA representation pairs in the mouse skin hair follicle dataset, which demonstrates that HALO's coupled representations could capture regulatory interactions between peaks and gene expressions from the extent to which they are aligned with each other.

HALO characterizes gene-peak interactions in a temporal causal perspective. To further characterize the causal temporal relations between individual genes and peaks, we categorize

gene-peak pairs into coupled genes and decoupled genes (**Figure 1A, 3B**). Employing negative binomial regression from denoised nearby peaks to denoised gene expression penalized by genomic distance, we align gene-peak pairs (**See Methods Section 6.1**). Upon aggregating these matched peaks, HALO calculates decouple and couple scores to quantitatively assess the extent of "decoupledness" and "coupledness" in gene-peak relationships, with positive scores suggesting decoupled and coupled, respectively**. Figure S3A-C** display the simulation outcomes for the decouple scores of both simulated decoupled and coupled gene-peaks pairs (**See Methods Section 11**). The SHARE-seq hair follicle data represent several lineages, differentiated from transit-amplifying cells (TACs) to inner root sheath (IRS), medulla, and cuticle/cortex cells (**Figure 3A**). **Figure 3C & D** and **S5 A & B** highlight the decoupled/coupled gene-peaks across all branches**,** with the respective couple and decouple scores (**Figure 3H, S4A & C**). Additionally, HALO takes into account distinct developmental lineages: Cortex lineage (**Figure 3E, S5C & D**), IRS (**Figure 3F, S5E & F**), and Medulla (**Figure 3G, S5G & H**). By examining the decoupled genes (Abl2, Dnajb1, Dlx3 and Btg2) across different branches or within specific lineages, it becomes apparent that gene expression levels and corresponding peaks change independently over time. Specifically, Dlx3 exhibits decoupled behaviors on all three branches (**Figure 3H**); the gene expression remains relatively stable, but the aggregated peaks keep going up with time (**Figure 3E-G**). **Some genes (Notch1, Dlx3) exhibit coupled expression across all cells, but show decoupled behaviors within specific lineages, highlighting that gene regulation is highly context specific.**

We address the question of why these peaks continue to increase by examining the downstream effect of the accessibility of a local chromatin region on the gene expression of potential regulation targets other than the corresponding gene. Granger causality is well suited for the task of inferring these regulatory relationships due to the *time lag* between a local peak and target gene expression change. By utilizing the Granger causality test (**see Methods Section 10.1**), we are able to identify the distal regulation relations among Dlx3's corresponding peaks and nearby genes other than Dlx3**.** In **Figure 3I**, the pink loops are peak-gene interactions inferred by DORC[8], while the blue loops are uniquely detected by negative binomial regression. By predicting gene expression using the peaks of preceding cells with respect to ascending latent time, we find that nine local peaks within mouse skin hair follicle super enhancer region exhibit Granger causal relations with the expression of Itga3 (**Figure 3J-K** and **Figure S4B**)[16]. In summary, by analyzing gene-peak pairs at the individual gene level, HALO effectively distinguishes between coupled and decoupled interactions, as well as illustrates that local peaks of decoupled genes may contribute to expression of nearby genes.

HALO uncovers regulatory factors in primary human CD4+ effector T cells assayed by NEAT-seq. NEAT-seq profiles the intra-nuclear protein epitope abundance of a panel T cell master transcription factors (TFs), chromatin accessibility, and transcriptome in single cells[13]. Due to posttranscriptional regulatory mechanisms affecting GATA3 protein expression[17], we can leverage the protein quantification of the Th2 master TF GATA3 for more cell state information than can be gleaned from using RNA expression levels alone. HALO infers latent representations utilizing nuclear protein level of GATA3 as a proxy variable for latent time (**see Methods Section 7.2**) and constructs UMAP embedding to visualize distinct T cell subsets (**Figure 4A, S6 A-C, S7 & S8**). Notably, two coupled latent representations, RNA coupled 9 and ATAC coupled 9, show negative correlations with the nuclear protein level of GATA3 (**Figure 4 B-C**). RNA

coupled 9 captures mTORC1 signaling (**Figure 4D**), which negatively regulates Th2 differentiation[18]. Consistently, ATAC coupled 9 is enriched in ZFX/NR4A motifs (**Figure 4E**), where NR4A is known to suppresses Th2 genes[19,20]. Conversely, Th17-specific ATAC coupled representation 12 exhibits enrichment in NR2F6/ARID5A motifs (**Figure S8 & 4E**), which are essential for the differentiation and function of Th17 cells[21,22].

GATA3 regulates the Th2 cell fate decision in CD4+ T cells, while Tbet orchestrates Th1 differentiation. Both T-bet and GATA3 are co-expressed in Th1 cells, rather than being exclusively present in their respective lineages (**Figure 4F**). Within Th1 cells, T-bet suppresses Th2 genes by redistributing GATA3 from Th2 gene loci to Tbet binding sites within Th1 gene regions[23]. Granzyme genes (GZMA, GZMB, GZMK) possess binding sites specifically targeted by both T-bet and GATA3 in Th1 cells[24]. To dissect the regulation mechanisms of different TFs, we calculate decouple scores for Th1 cells using protein levels of GATA3 and T-bet as temporal information. Intriguingly, expression of the GZMA gene and its local peaks exhibit decoupled dynamics in Th1 cells (**Figure 4G-H**). As GATA3 protein levels rise, local peaks of GZMA become less accessible, even though GZMA gene expression is increased at high GATA3 protein levels (**Figure 4H**). Consistently, CD4+ T cells co-expressing T-bet and GATA3 exhibit upregulation of GZMA gene expression, compared to cells expressing only T-bet (**Figure S6D**). Conversely, increases in nuclear T-bet levels enhance the accessibility of local GZMA peaks, while GZMA gene expression level remains constant (**Figure 4G**). Using nuclear protein levels as temporal information, Granger causality tests reveal that local peaks of GZMA mediate the gene expression of GZMK (**Figure 4I-L**). Strikingly, local peaks of GZMA within the Th1-specific T-bet super enhancer region are also crucial for regulating gene expression of GZMK[25]. Leveraging TF protein levels as proxies for temporal information, HALO coupled representations not only exhibit strong correlation between RNA and ATAC modalities, but also capture similar biological contexts (**Figure S6E**). HALO allows us to further investigate the regulation of a TF using Granger causality test to determine downstream enhancer and its gene targets.

HALO reveals epigenetic regulation of alveolar epithelial differentiation in systemic sclerosis-associated interstitial lung disease (SSc-ILD). Systemic sclerosis (SSc) is an autoimmune disease characterized by fibrosis in the lungs and other organs. In SSc-ILD, alveolar epithelial cells are decreased due to impaired regeneration function, while airway epithelial cells (ciliated, club, goblet, basal) increase in fibrotic lesions[26]. Recent scRNA-seq analysis of human distal lungs and alveoli has shown that alveolar type-2 (AT2) cells acquire a unique epithelial transition state, known as AT0, during primate lung regeneration and disease[27]. AT0 cells have the potential to differentiate into either alveolar type-1 (AT1) cells or terminal and respiratory bronchiole-specific secretory cells (TRB-SCs) when cultured *in vitro*. However, in idiopathic pulmonary fibrosis (IPF), AT0 cells predominantly transform into terminal secretory cells within severe fibrotic areas, referred to as 'bronchiolized regions' in IPF lungs[27].

To investigate the molecular underpinnings of impaired alveolar epithelial regeneration in SSc-ILD, we analyzed alveolar epithelial and terminal secretory cells using single-cell multiome sequencing data from six SSc-ILD and seven control lungs[28]. Due to its distinct genomic profile, airway basal cells were excluded from downstream analysis. We identified five clusters of alveolar epithelial and terminal secretory cells based on the expression of known marker genes

(**Figure 5A, S9A-D**). Two subpopulations of AT2 cells (clusters 0 and 1) exhibit similar transcriptomic profiles but differ in chromatin accessibility. Additionally, we identified AT1 cells (cluster 3) and TRB-SCs (cluster 4), which co-express SFTPB and SCGB3A2, along with club and goblet cells (cluster 2).

Our analysis revealed a substantial decrease in the AT2 cell population and an increase in secretory cells in SSc-ILD samples (**Figure 5B**). Additionally, AT2 cells in SSc-ILD lungs mirrored the transcriptional state of AT2 cells cultured in EGF-depleted organoids (**Figure S10A-E**), where AT2 cells convert to AT0 cells and subsequently progress to TRB-SCs following EGF depletion[27]. Consistently, trajectory inference algorithms (scVelo[29], CellRank 2[30], and Palantir[31]) applied to epithelial populations from both SSc-ILD and control samples reveal two terminal states in the AT2 differentiation trajectory. Cluster 1 progresses to secretory cells, passing through TRB-SCs, while cluster 0 differentiates into AT1 cells (**Figure 5C-D & S9E-F**).

To identify epigenetic regulations underlying AT2 lineage specification, we applied HALO to infer latent representations utilizing Palantir pseudotime. While RNA profiles alone are insufficient to distinguish the two AT2 subpopulations, decoupled ATAC representations are informative in separating these two subpopulations (**Figure S11, S12A-B**). Two decoupled ATAC representations (decoupled 14 and decoupled 15) characterize the transition from AT2 cells (cluster 1) to TRB-SCs (cluster 4) (**Figure 5E**). With an interpretable decoder, we find that top peaks contributing to the decoupled representations are enriched in airway epithelial-specific super enhancers and known transcription factors essential for proximal airway patterning during lung development[32,33](**Figure 5E-G**). In contrast, three coupled ATAC representations describe the alveolar epithelial populations (cluster 0, 1, 3 & 4), which are enriched in alveolar epithelium-specific super enhancers (**Figure 5E-F**). Consistently, ChromVAR motif activity analysis indicates that cluster 1 exhibits lower motif activity for alveolar lineage transcription factors (NKX2-1 and CEBPA) compared to cluster 0, the other AT2 cluster (**Figure S10F**). CEBPA restricts AT2 cell plasticity during development and injury-repair, while CEBPA-dependent regulation recruits alveolar epithelial lineage TF NKX2-1 to promote and maintain the AT2 program[34–37]. Additionally, there are increased motif activities of transcription factors for proximal airway patterning and decreased motif activity of CEBPA in AT2 cells (cluster 0 & 1) from SSc lungs (**Figure S10G**). Analysis of modality-specific decoupled representations using HALO's interpretable framework enables us to discover the epigenetic landscape and cell fates of alveolar epithelial cells.

We further evaluated transition probabilities among these clusters using optimal transport based on RNA and ATAC representations obtained from HALO (**see Methods Section 9**). In normal lungs, both AT2 clusters would progress to AT1. However, in SSc lungs, although cluster 0 still differentiated into AT1, cluster 1 predominantly transformed into TRB-SCs (**Figure 5N & S9G**), demonstrating that bipotent AT2 cells exhibit different cell fate decisions under SSc conditions. This suggests that HALO is able to depict a comprehensive portrait of alveolar differentiation, while defining epigenetic regulatory effectors of cell identity and pathology.

We then applied HALO to examine disease-associated decoupled genes. The epithelial-mesenchymal transition (EMT) master regulator SOX4 plays a crucial role in AT2-to-AT1 differentiation[38,39]. In SSc-ILD, both the expression and local peaks of SOX4 increase (**Figure 5H-I**). The dynamics of SOX4 gene expression and its local peaks become decoupled during

terminal differentiation to AT1 in SSc lungs (**Figure 5J**); the local peaks remain accessible while the gene expression decreases. Strikingly, these local peaks of SOX4 are significantly Granger causally related to the expression of the EMT-associated long non-coding RNA (lncRNA) CASC15 (**Figure 5K-M).** Additionally, local peaks of SOX4 within AT1-specific super enhancer regions contribute to the expression of CASC15. In SSc condition, regulatory events of SOX4's local peaks exhibit longer time lag compared to the normal condition. A previous study has shown that CASC15 is upregulated in aberrant basaloid cells, an epithelial cell type displaying a partial EMT phenotype in ILD lungs[40]. Using HALO, we uncovered disease-specific decoupled genes that may contribute to SSc-ILD pathogenesis.

To further characterize dysregulated gene regulatory network of AT2 cells under SSc conditions, we inferred TF-CRE-gene linkages (**see Methods Section 14**). Among these, the EGR, RFX, TCF, and NFI family transcription factors (TFs) exhibited the largest positive average differences in chromVAR motif activity (**Figure S12C, Supp Table 1**). Notably, RFX family TFs are crucial for airway epithelial differentiation, while TCF and NFI family TFs play essential roles in alveolar epithelial differentiation, survival and regeneration[41–43].

Methods

See the Supplementary Methods Section.

Discussion

The relationship between chromatin accessibility and gene expression is complex and often asynchronous. Previous methods assume that the two are synced with respect to time, missing nuances like time-lagging effects and independent regulatory mechanisms. HALO is the first to address these gaps by differentiating between coupled (dependent) and decoupled (independent) changes from a causal perspective. The framework operates at both representations and individual gene levels to learn interpretable modality-specific and shared information, as well as characterizes gene and associated peak interactions.

We conduct extensive benchmarks to evaluate the performance of HALO from different perspectives. Across several datasets, we demonstrate that HALO effectively learns highly coupled information across modalities while preserving decoupled and modality-specific information rather than batch bias. The highly coupled RNA and ATAC representation pairs, characterized by high Pearson correlation, capture functionally analogous biological contexts. In contrast, decoupled ATAC representations can inform our knowledge of chromatin accessibility-mediated cell fate potentials during the differentiation of mouse skin hair follicles and alveolar epithelium. Notably, we confirmed the bipotency of a subset of AT2 cells that acquire a previously reported transitional cellular state with the potential to differentiate into TRB-SCs or AT1 cells depending on niche signals[27,44]. We demonstrate that epigenetic information can shape this cell fate decision by conducting motif enrichment and cell type-specific super enhancer analyses on HALO's decoupled representations. Achieving a deeper understanding of the factors underlying the diverging differentiation paths of AT2 cells opens the door to therapeutic options for a variety of lung diseases.

At the individual gene level, HALO distinguishes between coupled and decoupled gene-peak pairs. For decoupled cases where associated chromatin regions become more accessible, but gene

expression remains stable or decreases, HALO employs Granger causality test to assess distal regulatory interactions between associated peaks and nearby genes. We find that peaks overlapping with super enhancer regions often exhibit such decoupled behavior and are involved in distal regulation during cellular transition or developmental trajectories. Additionally, HALO uncovers SSc-ILD-specific regulatory mechanisms, such as the decoupling of SOX4 expression and its local peaks during AT2-to-AT1 differentiation. We show that during this differentiation under SSc conditions, chromatin regions in the intersection of SOX4 local peaks and AT1-specific super enhancers are responsible for regulation of the lncRNA CASC15.

To enhance HALO's applicability, future work can extend the framework to include profiling of additional modalities, such as methylation and protein levels, to provide a more comprehensive understanding of gene regulation dynamics. Additionally, leveraging spatial epigenome-transcriptome co-profiling technologies could allow HALO to uncover spatiotemporal dynamics and genome-wide gene regulation mechanisms within tissue contexts[45]. These advancements would extend HALO’s utility as a valuable tool for understanding the functions and regulatory mechanisms of cell populations.

Bibliography

1. Burdziak, C. *et al.* Epigenetic plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. *Science (1979)* 380, (2023).
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. *Nat Biotechnol* 41, 387–398 (2023).
3. Liu, J. *et al.* Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat Protoc* 15, 3632–3662 (2020).
4. Lynch, A. W. *et al.* MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods* 19, 1097–1108 (2022).
5. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat Methods* 20, 1222–1231 (2023).
6. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 40, 1458–1466 (2022).
7. Wu, A. P., Singh, R. & Berger, B. Granger causal inference on DAGs identifies genomic loci regulating transcription. *ArXiv* abs/2210.10168, (2022).
8. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* 183, 1103-1116.e20 (2020).
9. Huggins, I. J. *et al.* The WNT target SP5 negatively regulates WNT transcriptional programs in human pluripotent stem cells. *Nat Commun* 8, 1034 (2017).
10. Paris, N. D., Coles, G. L. & Ackerman, K. G. Wt1 and β-catenin cooperatively regulate diaphragm development in the mouse. *Dev Biol* 407, 40–56 (2015).
11. Sulic, A.-M. *et al.* Transcriptomic landscape of early hair follicle and epidermal development. *Cell Rep* 42, 112643 (2023).
12. Li, G. *et al.* A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. *Genome Biol* 23, 20 (2022).
13. Chen, A. F. *et al.* NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* 19, 547–553 (2022).
14. Lance, C. *et al.* Multimodal single cell data integration challenge: Results and lessons learned. in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track* (eds. Kiela, D., Ciccone, M. & Caputo, B.) vol. 176 162–176 (PMLR, 2022).
15. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 19, 41–50 (2022).
16. Adam, R. C. *et al.* Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* 521, 366–370 (2015).
17. Cook, K. D. & Miller, J. TCR-Dependent Translational Control of GATA-3 Enhances Th2 Differentiation. *The Journal of Immunology* 185, 3209–3216 (2010).
18. Chi, H. Regulation and function of mTOR signalling in T cell fate decisions. *Nat Rev Immunol* 12, 325–338 (2012).
19. Sekiya, T. *et al.* Suppression of Th2 and Tfh immune reactions by Nr4a receptors in mature T reg cells. *Journal of Experimental Medicine* 212, 1623–1640 (2015).
20. Sekiya, T. *et al.* Regulation of peripheral Th/Treg differentiation and suppression of airway inflammation by Nr4a transcription factors. *iScience* 24, 102166 (2021).
21. Masuda, K. *et al.* Arid5a regulates naive CD4+ T cell fate through selective stabilization of Stat3 mRNA. *Journal of Experimental Medicine* 213, 605–619 (2016).
22. Hermann-Kleiter, N. & Baier, G. Orphan nuclear receptor NR2F6 acts as an essential gatekeeper of Th17 CD4+ T cell effector functions. *Cell Communication and Signaling* 12, 38 (2014).
23. Hertweck, A. *et al.* The TH1 cell lineage-determining transcription factor T-bet suppresses TH2 gene expression by redistributing GATA3 away from TH2 genes. *Nucleic Acids Res* 50, 4557–4573 (2022).
24. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat Commun* 3, 1268 (2012).
25. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep* 15, 2756–2770 (2016).
26. Valenzi, E. *et al.* Disparate Interferon Signaling and Shared Aberrant Basaloid Cells in Single-Cell Profiling of Idiopathic Pulmonary Fibrosis and Systemic Sclerosis-Associated Interstitial Lung Disease. *Front Immunol* 12, (2021).
27. Kadur Lakshminarasimha Murthy, P. *et al.* Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature* 604, 111–119 (2022).

28. Valenzi, E. *et al.* Altered AP-1, RUNX, and EGR chromatin dynamics drive fibrotic lung disease. Preprint at https://doi.org/10.1101/2024.10.23.619858 (2024).
29. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 38, 1408–1414 (2020).
30. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nat Methods* 21, 1196–1205 (2024).
31. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* 37, 451–460 (2019).
32. Cao, S. *et al.* Single-cell RNA sequencing reveals the developmental program underlying proximal–distal patterning of the human lung at the embryonic stage. *Cell Res* 33, 421–433 (2023).
33. Wang, Y. *et al.* SEdb 2.0: a comprehensive super-enhancer database of human and mouse. *Nucleic Acids Res* 51, D280–D290 (2023).
34. Toth, A. *et al.* Alveolar epithelial progenitor cells require Nkx2-1 to maintain progenitor-specific epigenomic state during lung homeostasis and regeneration. *Nat Commun* 14, 8452 (2023).
35. Little, D. R. *et al.* Differential chromatin binding of the lung lineage transcription factor NKX2-1 resolves opposing murine alveolar cell fates in vivo. *Nat Commun* 12, 2509 (2021).
36. Murray, C. W. *et al.* LKB1 drives stasis and C/EBP-mediated reprogramming to an alveolar type II fate in lung cancer. *Nat Commun* 13, 1090 (2022).
37. Hassan, D. & Chen, J. CEBPA restricts alveolar type 2 cell plasticity during development and injury-repair. *Nat Commun* 15, 4148 (2024).
38. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat Commun* 11, 3559 (2020).
39. Habermann, A. C. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 6, (2020).
40. Uthaya Kumar, D. B. *et al.* Bronchial epithelium epithelial-mesenchymal plasticity forms aberrant basaloid-like cells in vitro. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 322, L822–L841 (2022).
41. Hsu, Y.-C. *et al.* Mesenchymal Nuclear factor I B regulates cell proliferation and epithelial differentiation during lung maturation. *Dev Biol* 354, 242–252 (2011).
42. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat Commun* 11, 3559 (2020).
43. Wang, Z. *et al.* Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. *Nat Commun* 12, 6500 (2021).
44. Cai, X. T. *et al.* IL-4-induced SOX9 confers lineage plasticity to aged adult lung stem cells. *Cell Rep* 43, 114569 (2024).
45. Zhang, D. *et al.* Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature* 616, 113–122 (2023).
46. Mitra, S. *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat Genet* 56, 627–636 (2024).

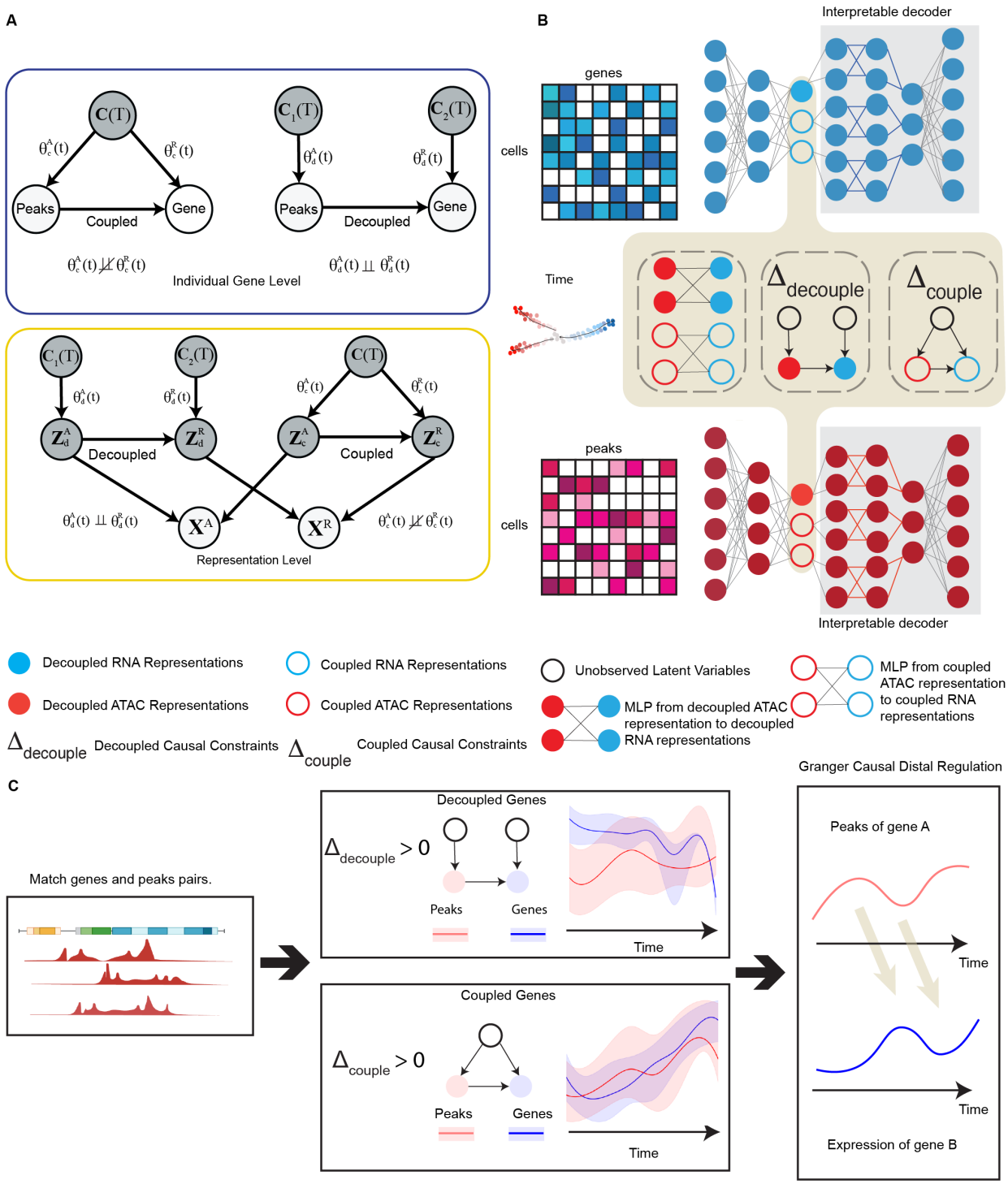

Figure 1. The Main Framework of HALO.

A. Top: The causal diagram of individual gene expression and its corresponding peaks; Bottom: the causal diagram of representation level of scRNA-seq and scATAC-seq.

B. Depicts the architecture for representation learning within a causal regularized variational autoencoder (VAE) framework. For the ATAC modality, the latent $\mathbf{Z}^{\mathbf{A}}$ is divided into

$\{\mathbf{Z_d^A}, \mathbf{Z_c^A}\}$, representing the decoupled and coupled latent representations, respectively. Similarly, the RNA modality's latent representations comprise $\mathbf{Z_d^R}$ (decoupled) and $\mathbf{Z_c^R}$ (coupled). The decoupled representations $\mathbf{Z_d^A}$ and $\mathbf{Z_d^R}$ adhere to the decoupled causal constraints $\Delta_{\mathbf{decouple}}$, whereas the coupled representations, $\mathbf{Z_c^A}$, $\mathbf{Z_c^R}$ conform to the coupled causal constraints $\Delta_{\mathbf{couple}}$.

C. Illustrates the gene-peak level analysis process. Initially, genes and ATAC peaks within gene proximities are matched using non-negative binomial regression, linking gene expression to neighboring ATAC peaks. Subsequently, this analysis computes decouple and couple scores to categorize gene-peaks into either decoupled or coupled genes. Finally, the analysis employs the Granger causality test to identify distal regulatory relationships between peaks and genes, uncovering potential mechanisms of genetic regulation.

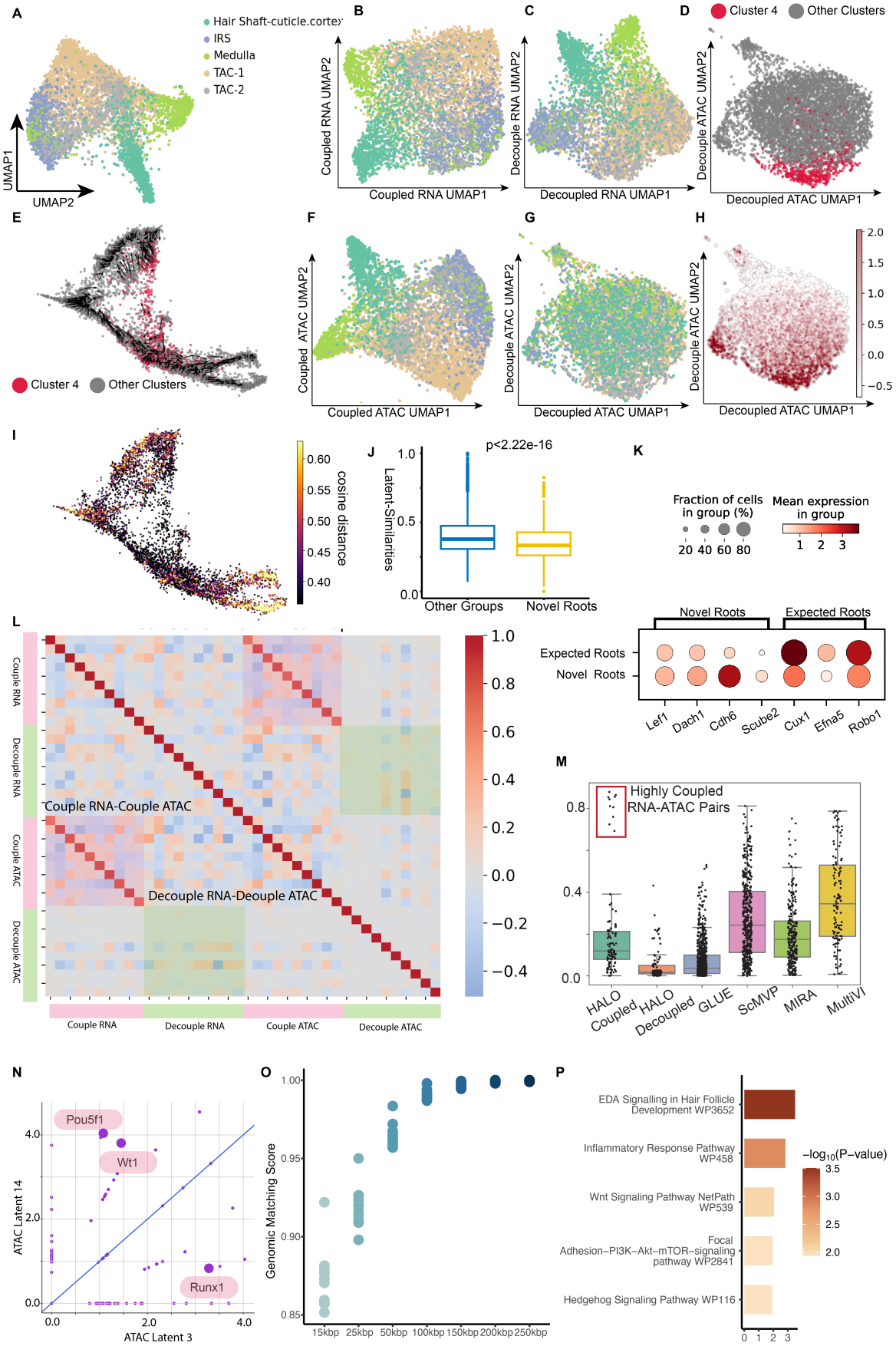
A
Hair Shaft-cuticle.cortex
IRS
Medulla
TAC-1
TAC-2
UMAP1
UMAP2
B
Coupled RNA UMAP2
Coupled RNA UMAP1
C
Decouple RNA UMAP2
Decoupled RNA UMAP1
D
Cluster 4
Other Clusters
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
E
Cluster 4
Other Clusters
F
Coupled ATAC UMAP2
Coupled ATAC UMAP1
G
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
H
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
I
cosine distance
J
p<2.22e-16
Latent-Similarities
Other Groups
Novel Roots
K
Fraction of cells in group (%)
Mean expression in group
Novel Roots
Expected Roots
Expected Roots
Novel Roots
Lef1
Dach1
Cdh6
Scube2
Cux1
Efna5
Robo1
L
Couple RNA
Decouple RNA
Couple ATAC
Decouple ATAC
Couple RNA-Couple ATAC
Decouple RNA-Deouple ATAC
M
Highly Coupled RNA-ATAC Pairs
HALO Coupled
HALO Decoupled
GLUE
ScMVP
MIRA
MultiVI
N
Pou5f1
Wt1
Runx1
ATAC Latent 14
ATAC Latent 3
O
Genomic Matching Score
15kbp
25kbp
50kbp
100kbp
150kbp
200kbp
250kbp
P
EDA Signalling in Hair Follicle Development WP3652
Inflammatory Response Pathway WP458
Wnt Signaling Pathway NetPath WP539
Focal Adhesion−PI3K−Akt−mTOR−signaling pathway WP2841
Hedgehog Signaling Pathway WP116
−log10(P−value)

Figure 2. The representation level results of SHARE-seq mouse skin hair follicle dataset. The decouple representations contain the chromatin potential information in the latent space.

- **A.** The UMAP constructed from concatenated RNA and ATAC representations, colored by cell type.
- **B.** The UMAP constructed from RNA coupled representations, colored by cell type.
- **C.** The UMAP of RNA decoupled representations, colored by cell type.
- **D.** The UMAP constructed from ATAC decoupled representations, colored by cluster membership (cluster 4 and the rest of the clusters).
- **E.** The distribution of cluster 4 and other clusters, along with chromatin potential vector field, shown on previously published UMAP[8,46].
- **F.** The UMAP constructed from ATAC coupled representations, colored by cell type.
- **G.** The UMAP constructed from ATAC decoupled representations, colored by cell type.
- **H.** The UMAP constructed from ATAC decoupled representations, colored by the value of ATAC decoupled latent representation 14, which mainly characterizes cluster 4.
- **I.** The original UMAP of SHARE-seq data[8] colored by the cosine distance of decoupled RNA and decoupled ATAC representations.
- **J.** The cosine similarity between decoupled RNA and decoupled ATAC representations of cluster 4 (Novel Roots) and rest of cells.
- **K.** The marker genes expression levels of the novel roots and expected roots.
- **L.** The Pearson correlations matrix of latent representations.
- **M.** The Pearson correlations for HALO (coupled representations), HALO (decoupled representations), GLUE, scMVP, MIRA, and MultiVI.
- **N.** The enriched transcription factors (TFs) for coupled ATAC latent representation 3 and decoupled ATAC latent representation 14.
- **O.** The Genomic Matching Score of coupled representations with different genomic distances from genes' TSS site.
- **P.** The gene ontology (GO) enrichment of top genes of RNA coupled representation 3.

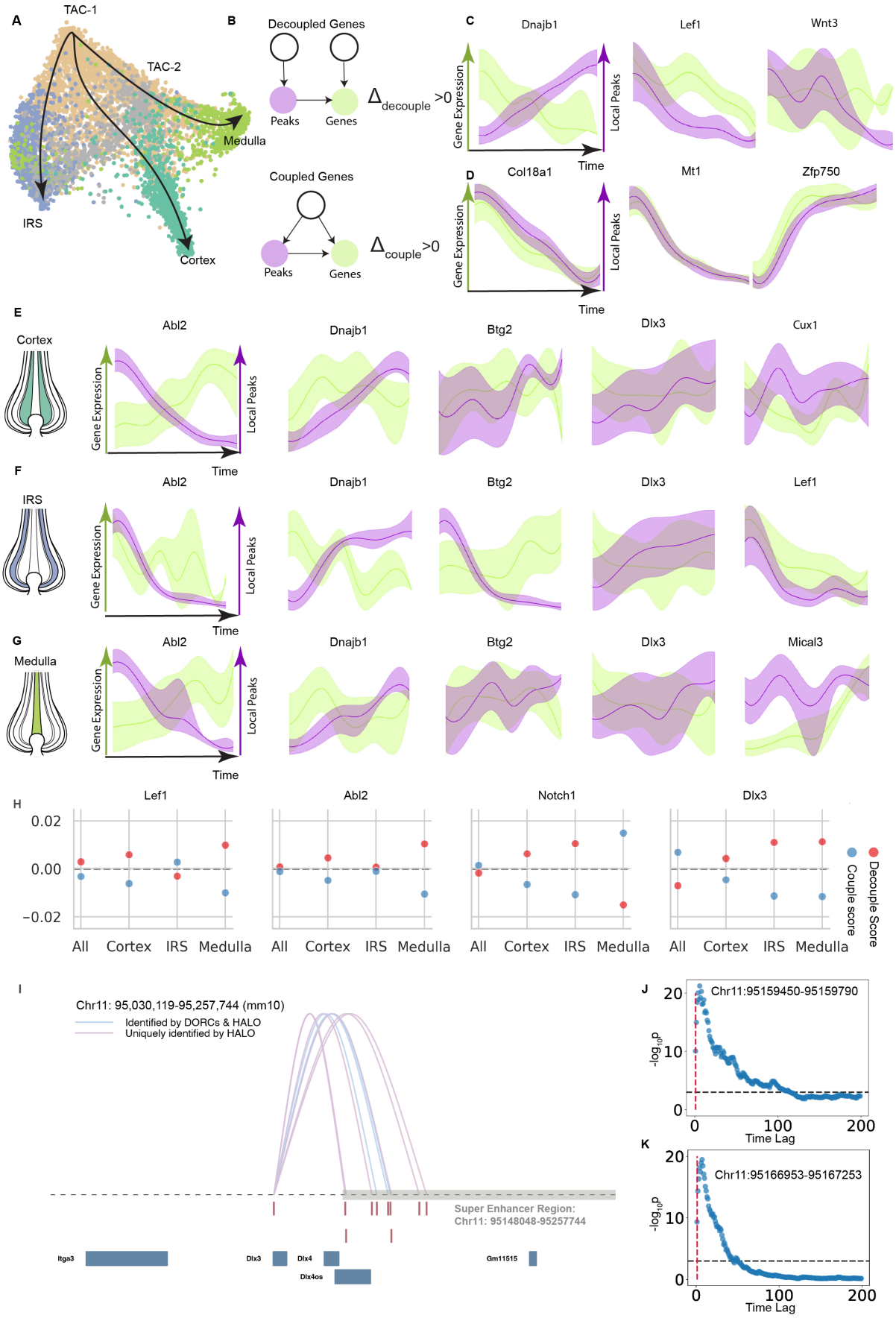
A
TAC-1
TAC-2
Medulla
IRS
Cortex
B
Decoupled Genes
Peaks
Genes
Δdecouple >0
Coupled Genes
Peaks
Genes
Δcouple >0
C
Dnajb1
Lef1
Wnt3
Gene Expression
Local Peaks
Time
D
Col18a1
Mt1
Zfp750
Gene Expression
Local Peaks
Time
E
Cortex
Abl2
Dnajb1
Btg2
Dlx3
Cux1
Gene Expression
Local Peaks
Time
F
IRS
Abl2
Dnajb1
Btg2
Dlx3
Lef1
Gene Expression
Local Peaks
Time
G
Medulla
Abl2
Dnajb1
Btg2
Dlx3
Mical3
Gene Expression
Local Peaks
Time
H
Lef1
Abl2
Notch1
Dlx3
0.02
0.00
−0.02
All
Cortex
IRS
Medulla
Couple score
Decouple Score
I
Chr11: 95,030,119-95,257,744 (mm10)
Identified by DORCs & HALO
Uniquely identified by HALO
Super Enhancer Region:
Chr11: 95148048-95257744
Itga3
Dlx3
Dlx4
Dlx4os
Gm11515
J
Chr11:95159450-95159790
-log10p
Time Lag
K
Chr11:95166953-95167253
-log10p
Time Lag

Figure 3. The individual gene level results of SHARE-seq mouse skin follicle hair dataset.

A. The development trajectory of mouse hair follicle cells.

B. The causal diagrams of decoupled and coupled individual gene-peak pairs. If the decouple score $\Delta_{decouple} > 0$, we classify the gene-peak pair as decoupled. If the couple score $\Delta_{couple} > 0$, we classify the gene-peak pair as coupled.

C. Examples of decoupled (top) and coupled (bottom) gene-peak pairs across all branches.

D. Examples of coupled gene-peak pairs across all branches.

E. Examples of decoupled gene-peak pairs in Cortex.

F. Examples of decoupled gene-peak pairs in inner root sheath (IRS).

G. Examples of decoupled gene-peak pairs in Medulla.

H. The decouple/couple score of genes on different developmental branches.

I. The loops denote significant connections between peaks within super enhancer region and RNA expression of Dlx3. The connections between the expression of Dlx3 and peaks are identified by a gene-peak-linking algorithm and reported by original study[8].

J. The scatter plot visualizes significance of Granger causal relations between Itga3 gene expression and peaks in Chr11: 95159490-95159790. The X-axis is the time lag (number of cells, sorted by latent time), Y-axis is -log(P-values). Granger causality significance is evaluated by using the peaks of all previous cells $[c_{t-n}, c_t)$ up to lag *n* (number of cells with index preceding time *t*) to determine the gene expression for cell c_t, where *t* is determined by latent time sequential ordering of all cells. We utilize the likelihood ratio test for the Granger causality-based regulation inference.

K. The scatter plot of Granger causal relations between Itga3 gene expression and peaks in Chr11: 95166953-95167253. The X-axis is the time lag (number of cells sorted by latent time), Y-axis is -log(P-values).

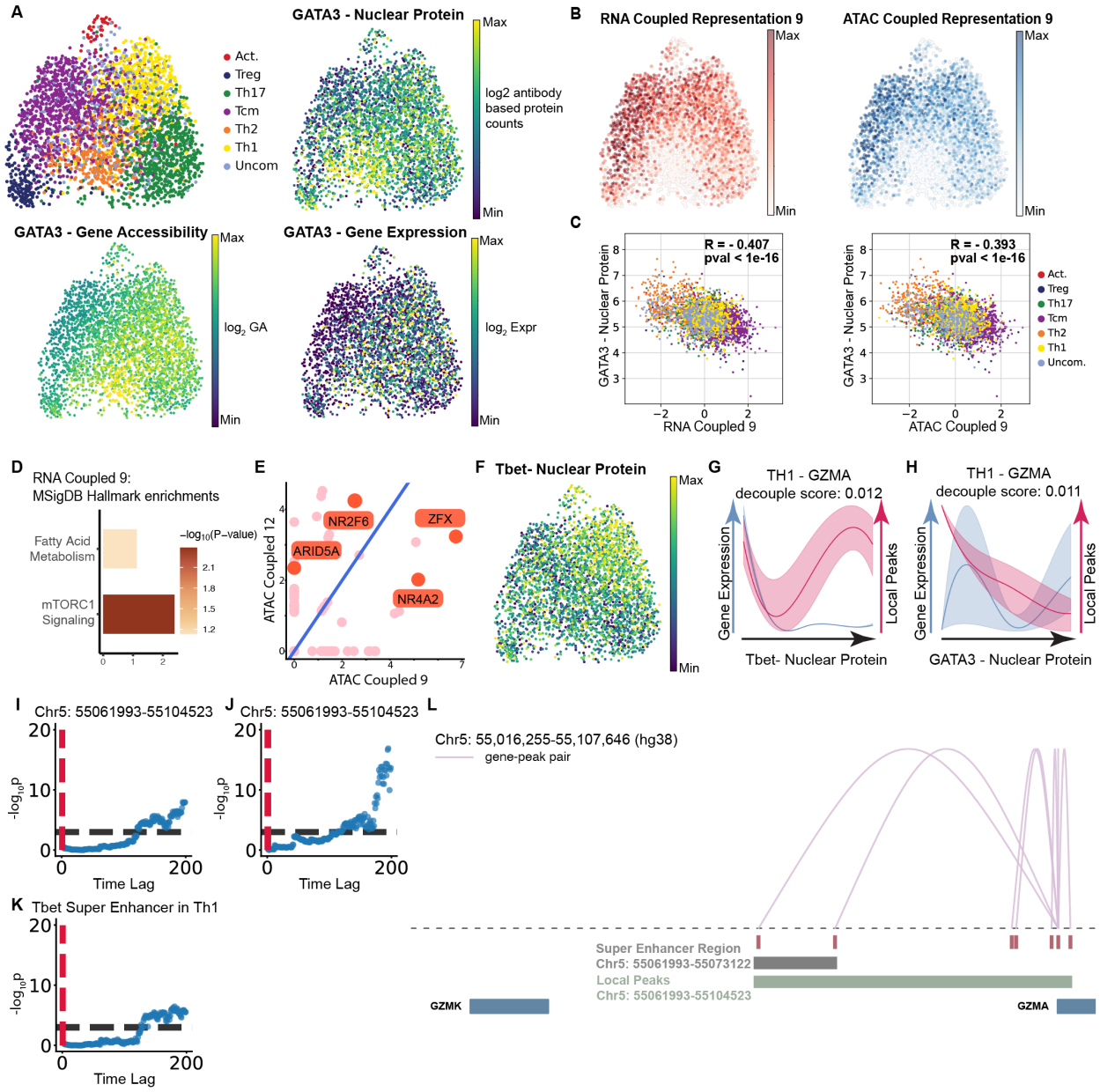

Figure 4. HALO reveals regulatory factors of human CD4+ effector T cell dataset

A. UMAP visualization of human CD4+ effector T cells constructed from concatenated RNA and ATAC representations, colored by original cell type, as well as nuclear protein, gene accessibility, and gene expression of GATA3.
B. RNA coupled 9 and ATAC coupled 9 representations on UMAP embedding of RNA and ATAC representations.
C. Scatter plots visualization GATA3 nuclear protein level and coupled latent representations (RNA coupled 9 and ATAC coupled 9)
D. Gene set enrichment for RNA coupled 9.
E. Motif enrichment for ATAC coupled 9 and ATAC coupled 12.
F. Nuclear protein level of Tbet on UMAP embedding of RNA and ATAC representations.
G. Gene-peak pairs of GZMA in Th1 cells sorted by Tbet nuclear protein level.
H. Gene-peak pairs of GZMA in Th1 cells sorted by GATA3 nuclear protein level.

I. The scatter plot of Granger causal relations between GZMK gene expression and peaks in Chr5: 55061993-55104523. The X-axis is the time lag (number of cells, sorted by Tbet nuclear protein level), Y-axis is -log(P-values). We utilized the F-test for the Granger causality-based regulation inference.
J. The scatter plot of Granger causal relations between GZMK gene expression and peaks in Chr5: 55061993-55104523. The X-axis is the time lag (number of cells, sorted by GATA3 nuclear protein level), Y-axis is -log(P-values).
K. The scatter plot of Granger causal relations between GZMK gene expression and peaks in Th1-specific T-bet super enhancer region (Chr5: 55061993-55073122). The X-axis is the time lag (number of cells, sorted by Tbet nuclear protein level), Y-axis is -log(P-values).
L. The loops denote significant connections between local peaks and RNA expression of GZMA. The connections between the expression of GZMA and peaks are identified by a gene-peak-linking algorithm.

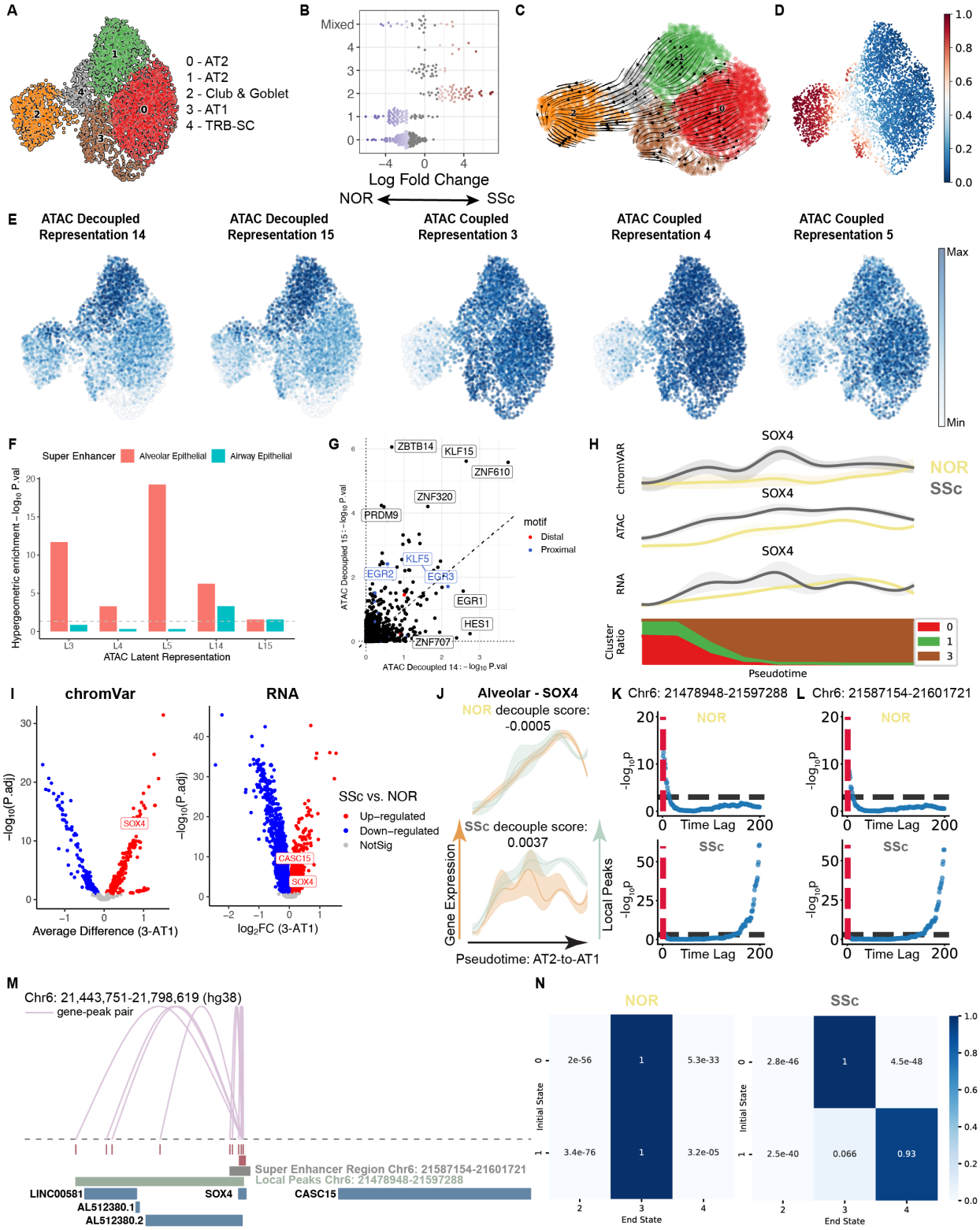

Figure 5. HALO uncovers chromatin lineage-priming in AT2 cells collected from SSc-ILD lungs.

A. UMAP visualization of pulmonary epithelial cells constructed from concatenated RNA and ATAC representations, colored by identified cell types.

B. Differential abundance test using Milo to compare control and SSc-ILD samples.

C. RNA velocity analysis showing differentiation trajectories of pulmonary epithelial cells.
D. UMAP visualization of pulmonary epithelial cells, colored by pseudotime.
E. ATAC latent representations (3, 4, 5, 14, 15) on UMAP embedding.
F. Hypergeometric enrichment of cell type specific super enhancer for ATAC latent representations. The dashed line indicates a p-value of 0.05.
G. Motif enrichment for ATAC decoupled representations 14 and 15, where TFs essential for proximal and distal airway patterning during lung development are colored blue and red, respectively
H. ChromVAR motif activity, local peaks, and gene expression of SOX4 in alveolar epithelial cells (clusters 0, 1, 3) from SSc and normal (NOR) lung, sorted by pseudotime.
I. Volcano plots visualize SSc-associated changes in gene expression and ChromVAR motif activity in cluster 3 (AT1).
J. Gene-peak pairs of SOX4 during AT2-to-AT1 differentiation (clusters 0, 1, 3) in alveolar epithelium, sorted by pseudotime.
K. The scatter plot showing Granger causal relations of CASC15 gene expression and local peaks of SOX4 in Chr6: 21478948-21597288. The X-axis is the time lag (number of cells sorted by pseudotime), Y-axis is -log(P-values). We utilized the likelihood ratio test for the Granger causality-based regulation inference.
L. The scatter plot showing Granger causal relations of CASC15 gene expression and AT1-specific super enhancer region (Chr6: 21587154-21601721). The X-axis is the time lag (number of cells sorted by pseudotime), Y-axis is -log(P-values).
M. The loops denote significant connections between local peaks and RNA expression of SOX4. Connections between the expression of SOX4 and peaks are identified by a gene-peak-linking algorithm.
N. Transition probability matrix of AT2 cell (clusters 0, 1) differentiation trajectories under different conditions, using optimal transport.

HALO Methods

Haiyi Mao[1,2†], Minxue Jia[1,2†], Marissa Di[1,2], Eleanor Valenzi[3], Xiaoyu Tracy Cai[4], Harinder Singh[5], Robert Lafyatis[6], Kun Zhang[7], Panayiotis V. Benos[1,2,8]

[1]Department of Computational and System Biology, University of Pittsburgh, Pittsburgh, PA, USA.
[2]CMU-Pitt Joint Program in Computational Biology, Pittsburgh, PA, USA.
[3]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, University of Pittsburgh, Pittsburgh, PA, USA.
[4]Immunology Discovery and Regenerative Medicine, Genentech, Inc., South San Francisco, CA, USA.
[5]Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA.
[6]Division of Rheumatology and Clinical Immunology, University of Pittsburgh, Pittsburgh, PA, USA.
[7]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA.
[8]Department of Epidemiology, University of Florida, Gainesville, FL, USA.

Contributing authors: ham112@pitt.edu; minxue.jia@pitt.edu; dim@pitt.edu; valenzie@upmc.edu; xiaoyu.t.cai@gmail.com; harinder@pitt.edu; rlafyatis@gmail.com; kunz1@cmu.edu; pbenos@ufl.edu;
[†]These authors contributed equally to this work.

Contents

1 Table of Symbols and Variables **4**

2 Problem Definition **5**

3 Causal Constraints **6**
3.1 Causal Graphic Models and Structural Causal Models . . . 6
3.2 Coupled and Decoupled Latent Representations . . . 7
3.3 Hilbert Schmidt Independence Criterion (HSIC) . . . 7
3.4 Independently Changing Constraints on Representations . . . 8

4 Generative Models **9**

5 Neural Interpretable Decoder **10**
5.1 Important Genes and Peaks for Representations . . . 10

6 Individual Gene Level Analysis **11**
6.1 Gene-Peak Matching . . . 11
6.2 Individual Gene Causal Conditions . . . 11
6.3 Gene Level Couple Score and Decouple Score . . . 12

7 Time Estimation **13**
7.1 Latent Time Estimation . . . 13
7.2 Protein-based Latent Time Estimation . . . 13
7.3 METAcell Construction . . . 13

8 Evaluation Metrics **13**
8.1 Genomic Matching Score . . . 13
8.2 Silhouette Score . . . 14
8.3 HSIC on Batches . . . 14
8.4 Normalized Mutual Information (NMI) . . . 14
8.5 Definition of Adjusted Rand Index (ARI) . . . 15

9 Optimal Transport for Lineage Prediction **15**
9.1 Optimal Transport . . . 15
9.2 Lineage Prediction with Optimal Transport . . . 16

10 Granger Causality-based Inference for Gene Regulatory Interactions **16**
10.1 Granger Causality and Tests . . . 16
10.2 Granger Causal Regulatory Interaction Inference . . . 17
10.2.1 Given Latent Time T . . . 17
10.2.2 Given Intra-nuclear Protein Information . . . 17

11 Individual Gene Level Simulation **17**

12 Dataset Preprocessing and Settings **18**
12.1 SHARE-seq Mouse Skin (Hair Follicle) Dataset . . . 18
12.2 NEAT-Seq Dataset . . . 19
12.3 SSc Human Lung Epithelium Dataset . . . 19
12.4 NeurIPS Dataset . . . 20
12.5 Human Brain Dataset . . . 20
12.6 10x Embryonic E18 Mouse Brain Dataset . . . 20

13 Enrichment Analysis **20**
13.1 Motifs . . . 20
13.2 Genes . . . 20

14 Gene Regulatory Network **20**
14.1 METAcell . . . 20
14.2 Identification of Cis-Regulatory Elements . . . 20
14.3 Construction of gene regulatory network . . . 21

15 Training and Testing Setups **21**

16 Ablation Study on Hyperparameters **21**
16.1 Hyperparameters Tuning . . . 21
16.2 α . . . 21
16.3 ω_1 . . . 21
16.4 ω_2 . . . 21
16.5 Number of Latent Representation Dimensions `dim` . . . 22

1 Table of Symbols and Variables

Variables	Meaning
n_{cells}	The number of cells
n_{genes}	The number of genes
n_{peaks}	The number of peaks
$\mathbf{X}^A$	The ATAC matrix, with dimension $\mathbb{Z}^{n_{cells} \times n_{peaks}}$
$\mathbf{X}^R$	The RNA matrix, with dimension $\mathbb{Z}^{n_{cells} \times n_{genes}}$
$\tilde{\mathbf{X}}^A$	The estimated ATAC matrix
$\tilde{\mathbf{X}}^R$	The estimated RNA matrix
$X^R_{i,g}$	The random variable of gene g value in cell i
$X^A_{i,j}$	The random variable of peaks j value in cell i
$X^R_{\cdot,g}$	The random variable of gene g across all cells
X^R_t	The expression of a given gene at time $T = t$
X^A_t	A local peak value associated with X^R_t at time $T = t$
$\tilde{X}^A_t$	An averaged local peak value of a given gene across cells at time t
$\tilde{X}^R_t$	An averaged expression value of a given gene across cells at time t
$\mathbf{X}^A_{\cdot,j}$	The vector of peak j value across all cells
$\mathbf{X}^R_{\cdot,g}$	The vector of gene expression value g across all cells
$\mathbf{X}^A_{i,\cdot}$	The vector of peaks values of cell i
$\mathbf{X}^R_{i,\cdot}$	The vector of gene expressions of cell i
n_A	The number of dimensions for ATAC representations
n_R	The number of dimensions for RNA representations
n_c	The number of dimensions for coupled representations
n_d	The number of dimensions for decoupled representations
$\mathbf{Z}^A$	The latent representation of ATAC, and $\mathbf{Z}^A \in \mathbb{R}^{n_A}$
$\mathbf{Z}^R$	The latent representation of RNA
$\mathbf{Z}^A_c$	The coupled latent representations of ATAC
$\mathbf{Z}^A_d$	The decoupled latent representations of ATAC
$\mathbf{Z}^R_c$	The coupled latent representations of RNA
$\mathbf{Z}^R_d$	The decoupled latent representations of RNA
T	Time T

Table 1: Table of Symbols and Variables

Variables	Meaning
θ_c^A	The functional parameters of $\mathbf{Z}_c^A$ or coupled peaks
θ_d^A	The functional parameters of $\mathbf{Z}_d^A$ or decoupled peaks
θ_c^R	The functional parameters of $\mathbf{Z}_c^R$ or coupled genes
θ_d^R	The functional parameters of $\mathbf{Z}_d^R$ or decoupled genes
V	Generic (causal) random endogenous variables
PA	The parent variables of a random variable in graphs
E	Exogenous (background) variables
f	The generating function of endogenous variables
$C(T)$	The latent confounder which can be presented as a function of time T
M_X	The Gram matrix of X
M_Y	The Gram matrix of Y
$\mathbf{B}_i^R$	The scRNA-seq batch variables for cell i
$\mathbf{B}_i^A$	The scATAC-seq batch variables for cell i
l_i^A	The library size of cell i's scATAC-seq
l_i^R	The library size of cell i's scRNA-seq
ρ_{ig}^R	The parameter of probability of success in the Negative Binomial (NB) distribution
θ_g^R	The Negative Binomial (NB) distribution dispersion parameter in cell i of gene g
p_{ij}^A	The normalized frequency of peaks in region j across all regions in cell i
$\hat{\rho}_{ig}^R$	The estimated normalized frequency of gene g in cell i by the local peaks
β_j	The scaling factor of influence of upstream, downstream and proximal regions to a gene
N_g	The set of peaks in the regions upstream, downstream, and proximal to the gene g TSS site
η_{gj}	The genomic distance from the region j to the TSS sight of gene g
δ_{gj}	The decay parameter for gene-peak matching
ω_1	The training weight of the modality aligning loss
ω_2	The training weight of causal constraint loss
$\phi(\cdot)$	The feature map of the RKHS
$\hat{\Delta}_{\tilde{X}^A \to \tilde{X}^R}$	The measure of dependence between $P(X^R \mid X^A)$ and $P(X^A)$
$\hat{\Delta}_{\tilde{X}^R \to \tilde{X}^A}$	The measure of dependence between $P(X^A \mid X^R)$ and $P(X^R)$
α	The threshold for couple and decouple constraint/score

Table 2: Table of Symbols and Variables, continued

2 Problem Definition

In this section, we give the formal definition of our model. The input is co-assayed scATAC-seq and scRNA-seq data. Specifically, scATAC-seq $\mathbf{X}^A \in \mathbb{Z}^{n_{cells} \times n_{peaks}}$ is a matrix with n_{cells} cells and n_{peaks} peaks, scRNA-seq $\mathbf{X}^R \in \mathbb{Z}^{n_{cells} \times n_{genes}}$ is a matrix with n_{cells} cells and n_{genes} genes. Moreover, $X_{\cdot,j}^A$ represents the value of peak j across all cells, $X_{i,j}^A$ represents the value of peak j in cell i. $X_{\cdot,g}^R$ is the count for gene g across all cells, $X_{i,g}^R$ is the expression of gene g in cell i.

The low-dimensional latent causal representations of scATAC-seq data $\mathbf{Z}^A \in \mathbb{R}^{n_A}, n_A \ll n_{peaks}$ can be decomposed to $\mathbf{Z}^A = [\mathbf{Z}_d^A, \mathbf{Z}_c^A]$, where $\mathbf{Z}_d^A \in \mathbb{R}^{n_d}, \mathbf{Z}_c^A \in \mathbb{R}^{n_c}$, and $n_c + n_d = n_A$. The latent representation of scRNA-seq data $\mathbf{Z}^R \in \mathbb{R}^{n_R}$ may be similarly decomposed: $\mathbf{Z}^R = [\mathbf{Z}_d^R, \mathbf{Z}_c^R]$, where $\mathbf{Z}_d^R \in \mathbb{R}^{n_d}, \mathbf{Z}_c^R \in \mathbb{R}^{n_c}$, and $n_c + n_d = n_R$. For the sake of simplicity, we set $n_c = n_d$ and $n_R = n_A$. $\mathbf{Z}_d^R$ and $\mathbf{Z}_d^A$ denote the decoupled latent representations of scRNA- and scATAC-seq data, while $\mathbf{Z}_c^R$ and $\mathbf{Z}_c^A$ are their coupled counterparts. Moreover, following the general case [1], we denote parameters of the structural causal model (SCM) for coupled and decoupled latent representations as the following: θ_c^A, θ_c^R, θ_d^A, and θ_d^R are the functional parameters of $\mathbf{Z}_c^A$, $\mathbf{Z}_c^R$, $\mathbf{Z}_d^A$, and $\mathbf{Z}_d^R$, respectively. To model the dynamics/changes of the representations with time, we further

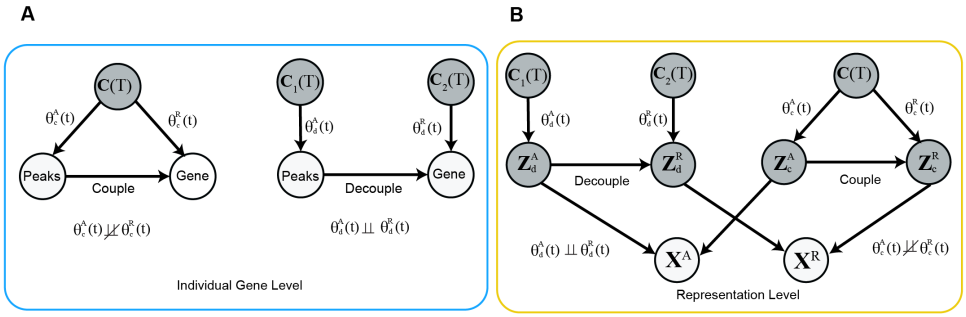

Fig. 1: The causal diagram of the scATAC-seq and scRNA-seq. A. The causal diagram at individual gene level. The decoupled gene-peak pairs have independently changing parameters while coupled gene-peak pairs have dependently changing parameters. B. The causal diagram at representation level. Coupled and decoupled latent representations are denoted as scATAC-seq $\mathbf{Z}_c^A$, $\mathbf{Z}_d^A$, and scRNA-seq $\mathbf{Z}_c^R$, $\mathbf{Z}_d^R$. Similarly, coupled representations have dependently changing parameters, while decoupled representations have independently changing mechanisms.

assume that these functional parameters can be written as functions of time T, that is $\theta_c^A(T)$, $\theta_c^R(T)$, $\theta_d^A(T)$, and $\theta_d^R(T)$. For simplicity of notation, we omit the T in the remainder of this paper.

3 Causal Constraints

In this section, we briefly introduce causal graphic models, structural causal models (SCMs), and non-stationary causal discovery conditions, then we introduce our causal constraints on representations.

3.1 Causal Graphic Models and Structural Causal Models

Finding underlying causal relations is a fundamental task in many disciplines [2]. Moreover, utilizing causal models to encode distributional and causal assumptions is becoming increasingly popular [3–5]. 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 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 [2, 6]. 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 PA_i. 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 [1]); 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 [6] as a triple $M = \langle E, V, F \rangle$, where:

1. E is a set $\{E_1, E_2, \ldots, E_n\}$ 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, $E \cup V$; and
3. F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from $E_i \cup PA_i$ to V_i, where $E_i \subseteq E$ and $PA_i \subseteq V \backslash V_i$. That is, $V_i = f_i(PA_i, E_i)$, for $i = 1, \ldots, n$, assigns a value to V_i that is dependent on a set of variables in $V \cup E$.

3.2 Coupled and Decoupled Latent Representations

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 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 [1]. We assume *pseudo causal sufficiency*, which means that any confounders can be written as smooth functions of time. Hence, in our multi-omic cases, the latent confounders can be presented as a function of pseudo-time. Denoting $\mathbf{C}(T)$ as the set of pseudo confounders (this set may also be empty), we can write the SCM as:

$$V_i = f_i(PA_i, C_i(T), \theta_i(T), E_i), \tag{2}$$

where $C_i(T) \subseteq \mathbf{C}(T)$ 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 E_i is the exogenous variables that is independent of T and PA_i, with a non-zero variance.

We construct a multi-modal representation learning model through a causal lens, incorporating temporal dynamics. Drawing from prior knowledge, we recognize that certain parts of scATAC-seq and scRNA-seq data exhibit strong correlations over time, whereas others do not (Fig. 1A). Our causal modeling framework incorporates this temporal information and the core hypothesis that open chromatin causes active transcription. Specifically, we pose the causal relations as $\mathbf{Z}_c^A \rightarrow \mathbf{Z}_c^R$ and $\mathbf{Z}_d^A \rightarrow \mathbf{Z}_d^R$.

We incorporate low dimensional representations in this biological process. Moreover, $\mathbf{Z}_c^R$ and $\mathbf{Z}_d^R$ altogether generate scRNA-seq, while decoupled and coupled representations $\mathbf{Z}_c^A$ and $\mathbf{Z}_d^A$ altogether generate scATAC-seq (Fig. 1B). Additionally, the independently changing mechanisms can be modeled as independently changing parameters of the generating functions. The correlated changing mechanisms can be formulated as confounded parameters of the generating functions. To be specific, it can be shown to be the following:

$$\mathbf{Z}_d^A = f_1(E_1, \theta_d^A(T)), \quad \mathbf{Z}_d^R = f_2(Z_d^A, E_2, \theta_d^R(T)), \tag{3}$$

$$\mathbf{Z}_c^A = f_3(E_3, \theta_c^A(T)), \mathbf{Z}_c^R = f_4(\mathbf{Z}_c^A, E_4, \theta_c^R(T)) \tag{4}$$

$$X_A = f_A(\mathbf{Z}_d^A, \mathbf{Z}_c^A, E_5), \quad X_R = f_R(\mathbf{Z}_d^R, \mathbf{Z}_c^R, E_6) \tag{5}$$

$$\theta_d^A(T) \perp\!\!\!\perp \theta_d^R(T), \quad \theta_c^A(T) \not\perp\!\!\!\perp \theta_c^R(T). \tag{6}$$

Further, we have following joint distributions,

$$P(\mathbf{Z}_d^A, \mathbf{Z}_d^R; \theta_d^A(T), \theta_d^R(T)) = P(\mathbf{Z}_d^A; \theta_d^A(T)) P(\mathbf{Z}_d^R | \mathbf{Z}_d^A; \theta_d^R(T)). \tag{7}$$

3.3 Hilbert Schmidt Independence Criterion (HSIC)

Testing for the independence of changing $P(\textbf{cause})$ and $P(\textbf{effect}|\textbf{cause})$ can be achieved by using Hilbert Schmidt Independence Criterion with properly designed kernel embeddings of nonstationary conditional distributions [1]. 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. *[7] 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 $\mathbf{M}_X$ and $\mathbf{M}_Y$ be the Gram matrices for X and Y estimated on the observed samples. We use following estimator of HSIC [8]:

$$\text{HSIC}_{XY} = \frac{1}{(N-1)^2} \operatorname{tr}(\mathbf{M}_X H \mathbf{M}_Y H), \tag{8}$$

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:

$$\text{HSIC}_{XY}^{\mathcal{N}} = \frac{\operatorname{tr}(\mathbf{M}_X H \mathbf{M}_Y H)}{\operatorname{tr}(\mathbf{M}_X H) \operatorname{tr}(\mathbf{M}_Y H)}. \tag{9}$$

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, which tend to be noisy and sparse, we extend this approach to the latent space representations of these two modalities. In the next section, we introduce the formal definition of our problem and its solutions.

3.4 Independently Changing Constraints on Representations

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

$$P(\mathbf{Z}_d^A, \mathbf{Z}_d^R; \theta_d^A, \theta_d^R) = P(\mathbf{Z}_d^A; \theta_d^A) P(\mathbf{Z}_d^R | \mathbf{Z}_d^A; \theta_d^R).$$

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

$$\tilde{P}(\underline{\mathbf{Z}_d^R}, \mathbf{Z}_d^A | T = t) = P(\mathbf{Z}_d^R | \mathbf{Z}_d^A, T = t) P(\mathbf{Z}_d^A). \tag{10}$$

Note that in Eq. (10), $P(\mathbf{Z}_d^A)$ does not depend on T. As a result, $\tilde{P}(\underline{\mathbf{Z}_d^R}, \mathbf{Z}_d^A | T = t)$ encodes only the changes across T in $P(\mathbf{Z}_d^R | \mathbf{Z}_d^A, T = t)$. A corresponding joint distribution can be also applied to the coupled latent space. Let $\mathbf{Z}^R \in \mathcal{Z}^R$ and $(\mathcal{H}, k)$ be a Reproducing Kernel Hilbert Space (RKHS) with a measurable kernel on $\mathcal{Z}^R$. Let $\phi(Z_i^R) \in \mathcal{H}, \mathbf{Z}^R = [Z_1^R, \dots Z_i^R, \dots Z_{n_R}^R]$ be the feature map for $\mathbf{Z}^R \in \mathcal{Z}^{\mathcal{R}}$, and assume $\mathbb{E}_{\mathbf{Z}^R}[k(\mathbf{Z}_i^R, \mathbf{Z}_j^R)] \leq \infty$. Similar notations are used for $\mathbf{Z}^A$, $\mathbf{Z}_c^R, \mathbf{Z}_d^R, \mathbf{Z}_c^A, \mathbf{Z}_d^A$, and T. Thus the general embedding of the joint distribution of coupled or decoupled latent representations can be formulated as the following:

Proposition 2. *[1] Let $\mathbf{Z}^A$ represent the direct causes of $\mathbf{Z}^R$, and suppose that they have N observations. The kernel embedding of distribution $\tilde{P}\left(\underline{\mathbf{Z}^R}, \mathbf{Z}^A \mid T = t_n\right)$ can be represented as*

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

where $\mathbf{\Phi}_{\mathbf{Z}^R} := \left[\phi\left(\mathbf{Z}_1^R\right), \dots, \phi\left(\mathbf{Z}_N^R\right)\right], \mathbf{\Phi}_{\mathbf{Z}^A} := \left[\phi\left(\mathbf{Z}_1^A\right), \dots, \phi\left(\mathbf{Z}_N^A\right)\right], \mathbf{k}_{\mathbf{t}, t_n} := \left[k\left(t_1, t_n\right), \dots, k\left(t_N, t_n\right)\right]^{\top}$, $\mathbf{K}_{\mathbf{Z}^A}\left(\mathbf{Z}_i^A, \mathbf{Z}_j^A\right) = \left\langle \phi\left(\mathbf{Z}_i^A\right), \phi\left(\mathbf{Z}_j^A\right) \right\rangle, \mathbf{K}_{\mathbf{t}}\left(t, t'\right) = \left\langle \phi\left(t\right), \phi\left(t'\right) \right\rangle$, and $\odot$ represents point-wise product.

The proof of this proposition can be found in the CD-NOD paper [1].

Our goal is to characterize the changing causal mechanisms between the latent spaces of scATAC- and scRNA-seq. Here we focus on the causal constraint for the representation level; for the gene-level analysis, please refer to Section Gene-Peak Matching. So for the decoupled case, $\mathbf{Z}_d^A$ and $\mathbf{Z}_d^R$ have independently changing mechanisms along T within the cell lifespan, but $\mathbf{Z}_c^A$ and $\mathbf{Z}_c^R$ change dependently. Thus, we aim to enforce that $P(\mathbf{Z}_d^R | \mathbf{Z}_d^A)$ and $P(\mathbf{Z}_d^A)$ change independently, and that $P(\mathbf{Z}_c^R | \mathbf{Z}_c^A)$ and $P(\mathbf{Z}_c^A)$ change dependently. CD-NOD [1] 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 $\mathbf{Z}^A$ and $\mathbf{Z}^R$ setting as follows:

$$\hat{\Delta}_{\mathbf{Z}^A \to \mathbf{Z}^R} = \frac{\operatorname{tr}(M_{\mathbf{Z}^A} H M_{\underline{\mathbf{Z}^R} \mathbf{Z}^A} H)}{\operatorname{tr}(M_{\mathbf{Z}^A} H) \operatorname{tr}(M_{\underline{\mathbf{Z}^R} \mathbf{Z}^A} H))}, \quad \hat{\Delta}_{\mathbf{Z}^R \to \mathbf{Z}^A} = \frac{\operatorname{tr}(M_{\mathbf{Z}^R} H M_{\underline{\mathbf{Z}^A} \mathbf{Z}^R} H)}{\operatorname{tr}(M_{\mathbf{Z}^R} H) \operatorname{tr}(M_{\underline{\mathbf{Z}^A} \mathbf{Z}^R} H))}. \tag{11}$$

where $\hat{\Delta}_{\mathbf{Z}^A \to \mathbf{Z}^R}$ measures the dependence between $P(\mathbf{Z}^R | \mathbf{Z}^A)$ and $P(\mathbf{Z}^A)$. Similarly, $\hat{\Delta}_{\mathbf{Z}^R \to \mathbf{Z}^A}$ measures the dependence between $P(\mathbf{Z}^A | \mathbf{Z}^R)$ and $P(\mathbf{Z}^R)$. Between these two dependence measures, if the value of $\hat{\Delta}_{\mathbf{Z}^A \to \mathbf{Z}^R}$ is smaller, we posit that $P(\mathbf{Z}^R | \mathbf{Z}^A)$ and $P(\mathbf{Z}^A)$ change independently, and that the causal relation is $\mathbf{Z}^A \to \mathbf{Z}^R$. In our case, for the decoupled representations $\mathbf{Z}_d^R$ and $\mathbf{Z}_d^A$, $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R}$ should be smaller than $\hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$, based on our knowledge of chromatin accessibility and transcription. On the other hand, for the coupled case, $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}$ and $\hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$ should be similar in value and both larger than a threshold α [1]. Formally, the causal constraints of $[\mathbf{Z}_d^R, \mathbf{Z}_c^R]$ and $[\mathbf{Z}_d^A, \mathbf{Z}_c^A]$ are given by the constraints as follows,

$$\begin{aligned} &\text{Decoupled Conditions:} \quad \hat{\Delta}_d = \hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} - \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A} < 0, \quad \hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} < \alpha \\ &\text{Coupled Conditions:} \quad \hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R} \geq \alpha, \quad \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A} \geq \alpha. \end{aligned} \tag{12}$$

4 Generative Models

We introduce our generative model designed to learn the lower dimensional representations $\mathbf{Z}_c^A$,$\mathbf{Z}_d^A$, $\mathbf{Z}_c^R$, $\mathbf{Z}_d^R$. Utilizing inputs from scRNA-seq $\mathbf{X}^R$ and scATAC-seq $\mathbf{X}^A$, along with modality-specific batch information $\mathbf{B}^R$ and $\mathbf{B}^A$, we use mean field variational inference to estimate $\mathbf{Z}_c^A$, $\mathbf{Z}_d^A$, $\mathbf{Z}_c^R$, $\mathbf{Z}_d^R$ [9, 10]. To model the expression value of gene g in cell i, denoted $X_{i,g}^R$, we specify the following distributions.

$$X_{i,g}^R \sim \text{NegativeBinomial}(l_i^R, \rho_{ig}^R, \theta_g^R), \tag{13}$$

where l_i^R is the scaling factor in cell i, ρ_{ig}^R is the normalized composition of gene g across all genes in cell i, and θ_g^R is the gene dispersion. Similarly, for chromatin region j of cell i, we consider $X_{i,j}^A$ as a Multinomial distribution,

$$X_{i,j}^A \sim \text{MultiNomial}(l_i^A, p_{ij}^A), \tag{14}$$

where p_{ij}^A is the normalized frequency of accessibility across all regions in each cell, l_i^A is the library size (scaling factor) for cell i. Moreover, we use $X_{i,\cdot}^R, \mathbf{B}_i^R$ and $X_{i,\cdot}^A, \mathbf{B}_i^A$ to infer library sizes l_i^A, l_i^R of each cell by variational inference, where l_i^A, l_i^R follow log-norm distributions, $\mathbf{L}^A = [l_1^A, \ldots, l_{n_{cells}}^A]$, $\mathbf{L}^R = [l_1^R, \ldots, l_{n_{cells}}^R]$ are the vectors of the library size of gene expression and peak of all cells. The mean field variational is used to estimate the latent representations with the following factorized posterior:

$$\begin{aligned}
&q(\mathbf{Z}^A, \mathbf{Z}^R, l^A, l^R | \mathbf{X}^R, \mathbf{X}^A, \mathbf{B}^R, \mathbf{B}^A) \\
&= q(\mathbf{Z}_c^A, \mathbf{Z}_d^A, \mathbf{Z}_c^R, \mathbf{Z}_c^R, l^A, l^R | \mathbf{X}^R, \mathbf{X}^A, \mathbf{B}^R, \mathbf{B}^A) \\
&= q(\mathbf{Z}_c^A, \mathbf{Z}_d^A | \mathbf{X}^A, \mathbf{B}^A) q(\mathbf{Z}_c^R, \mathbf{Z}_d^R | \mathbf{X}^R, \mathbf{B}^R) q(l^A | \mathbf{X}^A, \mathbf{B}^A) q(l^R | \mathbf{X}^R, \mathbf{B}^R) \\
&= q(\mathbf{Z}^A | \mathbf{X}^A, \mathbf{B}^A) q(\mathbf{Z}^R | \mathbf{X}^R, \mathbf{B}^R) q(\mathbf{L}^A | \mathbf{X}^A, \mathbf{B}^A) q(\mathbf{L}^R | \mathbf{X}^R, \mathbf{B}^R).
\end{aligned} \tag{15}$$

Then we use the following Evidence Lower Bound (ELBO),

$$\begin{aligned}
\log p(\mathbf{X}^A, \mathbf{X}^R | \mathbf{B}^A, \mathbf{B}^R) = \mathcal{L} &= \log p(\mathbf{X}^A | \mathbf{B}^A) + \log p(\mathbf{X}^R | \mathbf{B}^R) \\
&\geq \mathbb{E}_{q(\mathbf{Z}^A | \mathbf{X}^A, \mathbf{B}^A)} \log p(\mathbf{X}^A | \mathbf{Z}^A, l^A, \mathbf{B}^A) \\
&\quad + \mathbb{E}_{q(\mathbf{Z}^R | \mathbf{X}^R, \mathbf{B}^R)} \log p(\mathbf{X}^R | \mathbf{Z}^R, l^R, \mathbf{B}^R) \\
&\quad + D_{kl}(q(\mathbf{Z}^R | \mathbf{X}^R, \mathbf{B}^R) || p(\mathbf{Z}^R)) \\
&\quad + D_{kl}(q(\mathbf{Z}^A | \mathbf{X}^A, \mathbf{B}^A) || p(\mathbf{Z}^A)).
\end{aligned} \tag{16}$$

Specifically, we have the following posterior distributions,

$$q(\mathbf{Z}^A | \mathbf{X}^A, \mathbf{B}^A) \sim \mathcal{N}(\mu^A, \mathbf{\Sigma}^A), \tag{17}$$

$$q(\mathbf{Z}^R | \mathbf{X}^R, \mathbf{B}^R) \sim \mathcal{N}(\mu^R, \mathbf{\Sigma}^R), \tag{18}$$

$$q(l_i^A | \mathbf{X}_{i,\cdot}^A, \mathbf{B}_i^A) \sim \log \mathcal{N}(\mu_{li}^A, (\sigma_{li}^A)^2), \tag{19}$$

$$q(l_i^R | \mathbf{X}_{i,\cdot}^R, \mathbf{B}_i^R) \sim \log \mathcal{N}(\mu_{li}^R, (\sigma_{li}^R)^2). \tag{20}$$

where μ_A, μ_R are the means and Σ_A, Σ_R are the covariance matrices of $\mathbf{Z}^A, \mathbf{Z}^R$, respectively. We use deep neural networks to estimate the following:

$$\mu_A = f_A^\mu(\mathbf{X}^A, \mathbf{B}^A), \mathbf{\Sigma}_A = f_A^\Sigma(\mathbf{X}^A, \mathbf{B}^A); \tag{21}$$

$$\mu_R = f_R^\mu(\mathbf{X}^R, \mathbf{B}^R), \mathbf{\Sigma}_R = f_R^\Sigma(\mathbf{X}^R, \mathbf{B}^R); \tag{22}$$

$$\mu_{li}^A = f_{l^A}^\mu(\mathbf{X}_{i,\cdot}^A, \mathbf{B}_i^A), \sigma_{li}^A = f_{l^A}^\sigma(\mathbf{X}_{i,\cdot}^A, \mathbf{B}_i^A); \tag{23}$$

$$\mu_{li}^R = f_{l^R}^\mu(\mathbf{X}_{i,\cdot}^R, \mathbf{B}_i^R), \sigma_{lj}^R = f_{l^R}^\sigma(\mathbf{X}_{i,\cdot}^R, \mathbf{B}_i^R). \tag{24}$$

The prior distribution is specified as standard Gaussian distribution,

$$p(\mathbf{Z}^A) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{25}$$

$$p(\mathbf{Z}^A) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{26}$$

$$p(l_i^R) \sim \log \mathcal{N}(0,1), \tag{27}$$
$$p(l_i^A) \sim \log \mathcal{N}(0,1). \tag{28}$$

Next, we integrate the causal constraints outlined in the previous section into our final optimization objective. Both the coupled and decoupled representations are subject to the constraints specified in Equation (12). Thus, we define the causal constraint $\mathcal{C}$ with a hyperparameter α,

$$\begin{aligned}\mathcal{C} = \mathbf{max}(0, \hat{\Delta}_d) + \mathbf{max}(0, \alpha - \hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}) \\ + \mathbf{max}(0, \alpha - \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}) + \mathbf{max}(0, \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A} - \alpha).\end{aligned} \tag{29}$$

Moreover, we want to align the paired decoupled and coupled representations so that under the coupled or decoupled assumptions, $\mathbf{Z}_c^A$ causes $\mathbf{Z}_c^R$ and $\mathbf{Z}_d^A$ causes $\mathbf{Z}_d^R$. To realize these assumptions, we add the following layers,

$$\hat{\mathbf{Z}}_d^R = f_d(\mathbf{Z}_d^A),\ \hat{\mathbf{Z}}_c^R = f_c(\mathbf{Z}_c^A). \tag{30}$$

We want to minimize the $L2$ norm, $\mathcal{M}$:

$$\mathcal{M} = \mathbb{E}(||\hat{\mathbf{Z}}_d^R - \mathbf{Z}_d^R||_2^2) + \mathbb{E}(||\hat{\mathbf{Z}}_c^R - \mathbf{Z}_c^R||_2^2). \tag{31}$$

The main loss can be defined with parameters ω_1, ω_2 as follows:

$$\mathcal{L}_{total} = -\mathcal{L} + \omega_1 \mathcal{C} + \omega_2 \mathcal{M}. \tag{32}$$

The ablation study of hyperparameters $\alpha, \omega_1, \omega_2$ can be found in section Hyperparameters Tuning. Thus the optimization objective is as follows,

$$\arg\min_{\phi,\psi} \mathcal{L}_{total}. \tag{33}$$

5 Neural Interpretable Decoder

We introduce the neural interpretable decoder in the HALO framework. In light of Neural Additive Models [11], for each latent variable, we utilize a sub-neural network with non-linear transformations to capture information specific to each individual latent variable. A logistic regression layer is added to the end of the neural network (Fig. 2). Thus, by analyzing the logistic regression weights (loading matrix), we can conclude how much one latent variable Z_i for $i \in [1, \ldots, n]$ contributes to the reconstruction of a specific gene or peak (X). The non-linear transformation of latent representations can be formulated as follows,

$$\mathbb{E}(X) = \text{softmax}(\sum_i^n w_i f_i(Z_i) + \beta), \tag{34}$$

where f_i for $i \in [1, \ldots, n]$ are the non-linear functions which can be presented as neural networks, w_i are the weights of the the last linear layer, and β is the bias of the last layer. Moreover, f_i is constructed as follows:

$$f_i = \text{ReLU}(\text{MLP}(\text{ReLU}(\text{MLP}(\ldots \text{ReLU}(\text{MLP}(Z_i))). \tag{35}$$

Since all f_i for $i \in [1, \ldots, n]$ are non-negative, the weights w_i can interpreted as the importance/contribution of Z_i for the reconstruction of X.

5.1 Important Genes and Peaks for Representations

For each representation, we are able extract the important genes and peaks to further interpret the latent representations. Specifically, for Z_i^R, the ith element of the latent representations, we sort genes by the weights of last linear layer of the neural network, and pick the top 100 genes as the most "significant" genes for Z_i^R. Similarly, we are able to retrieve the top 100 peaks as the most "significant" peaks for Z_j^A.

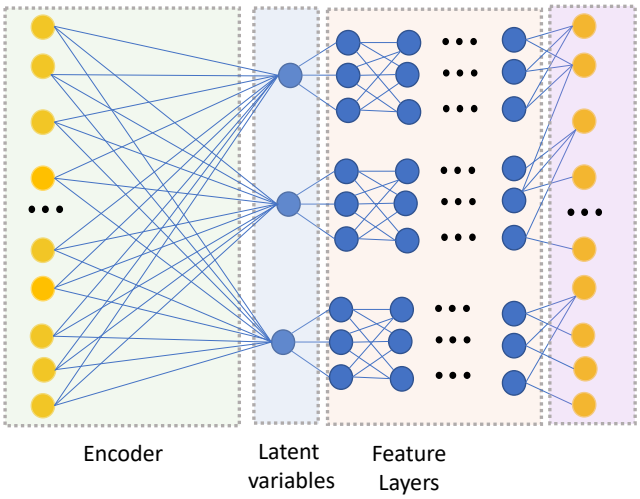

Fig. 2: Structure of the interpretable decoder

6 Individual Gene Level Analysis

6.1 Gene-Peak Matching

In light of MIRA [12], HALO correlates local scATAC-seq peaks to gene expression values, by searching within a certain distance upstream and downstream of local peaks to maximize the probability of observing gene expression values. Specifically, for gene $g \in \{1, n_{genes}\}$ in cell i, as well as estimated chromatin accessibility probability at chromatin region j and its probability p^A_{ij} (see Section Generative Models), we have the following definitions:

$$\hat{\rho}^R_{ig} = \sum_{j \in N_g} \beta_j p^A_{ij} 2^{-\eta_{gj}\delta_{gj}}, \tag{36}$$

where N_g is the set of peaks j in the regions upstream, downstream, and proximal to the TSS of gene g; $\beta_j \sim \text{HalfNormal}(0, 1)$ is the scaling factor for the influence of peaks in upstream, downstream, and proximal regions to the gene g; η_{gj} is the TSS distance; and δ_{gj} is the decay parameter. HALO learns these parameters $\{\beta_j, \eta_{gj}, \delta_{gj}\}$ by maximizing the likelihood of estimated ρ^R_{ig} in the previous representation learning phase. The upstream and downstream region sets encompass regions between 1.5 and 600 kbp from the TSS, while the proximal region is within 1.5 kbp. HALO estimates the parameters by maximizing the scaled $X^R_{i,g}$ likelihood:

$$\begin{aligned} &\underset{\{\beta_j, \eta_{gj}, \delta_{gj}\}}{\arg\max} \ \log p(\hat{\rho}^R_{ig} | p^A_{ij}); \\ &\text{where } X^R_{i,g} \sim \text{NegativeBinomial}(l^R_i, \rho^R_{ig}, \theta^R_g), \quad \forall \text{region j} \in N_g. \end{aligned} \tag{37}$$

The optimization procedures follow MIRA [12].

6.2 Individual Gene Causal Conditions

In this section, we apply the causal constraint to the individual gene level. After gene-peak pair matching, we can calculate the gene-level decouple score, further depicting the gene-peak behavior by characterizing how individual gene and corresponding matched peaks change with time. Recall the definition of coupled and decoupled causal relations: given time T, coupled gene-peak pairs have shared latent confounders that cause them to change in a correlated manner; in contrast, the decoupled gene-peak pairs have independently changing mechanisms with time, but are still associated in a causal manner. Fig 1A illustrates the coupled and decoupled gene-peak relations. As described in previous sections, we can calculate the independence of the parameters of peaks and genes to characterize the degree of their "decoupledness."

Similarly, given a gene g, its expression value X^R, and associated peaks X^A (for simplicity of notation, we drop the gene and cell indices here). we define the individual gene level changing joint distribution as follows:

$$\tilde{P}(\underline{X}^R, X^A | T = t) = P(X^R | X^A, T = t) P(X^A), \tag{38}$$

where $\tilde{P}(\underline{X}^R, X^A|T=t)$ captures the changes of X^R with T that are not introduced by X^A. Then we can calculate its kernel embeddings according to the proposition below. Let $X^R \in \mathcal{X}^R$ and $(\mathcal{H}, k)$ be a Reproducing Kernel Hilbert Space (RKHS) with a measurable kernel on $\mathcal{X}^R$. Moreover, $\mathbf{X}^R = [X^R_{t_1}, X^R_{t_2}, \dots X^R_{t_N}]$, and $\mathbf{X}^A = [X^A_{t_1}, X^A_{t_2}, \dots X^A_{t_N}]$ are samples we have for gene expression and associated peaks, where $t_1, \dots, t_N$ are the time labels for each sample and N is the number of samples. Let $\phi(X^R) \in \mathcal{H}$ be the feature map and assume $\mathbb{E}_{X^R}[k(X^R_{t_i}, X^R_{t_j})] \leq \infty$ and $\mathbb{E}_{X^A}[k(X^A_{t_i}, X^A_{t_j})] \leq \infty$. Thus the general embedding of the joint distribution of coupled or decoupled latent representations can be formulated as the following:

Proposition 3. *Let X^A represent the gene g's corresponding peaks which are the direct causes of the expression value X^R of gene g, and suppose that they have N observations. The kernel embedding of distribution $\tilde{P}\left(\underline{X}^R, X^A \mid T=t_n\right)$ can be represented as*

$$\hat{\tilde{\mu}}_{\underline{X}^R, X^A|T=t_n} = \frac{1}{n} \mathbf{\Phi_{X^R}} \left(\mathbf{K_{X^A}} \odot \mathbf{K_t} + \lambda I\right)^{-1} \operatorname{diag}\left(\mathbf{k}_{\mathbf{t}, t_n}\right) \mathbf{K_{X^A}} \mathbf{\Phi}^{\top}_{\mathbf{X^A}},$$

where $\mathbf{\Phi_{X^R}} := \left[\phi\left(X^R_{t_1}\right), \dots, \phi\left(X^R_{t_N}\right)\right]$, $\mathbf{\Phi_{X^A}} := \left[\phi\left(X^A_{t_1}\right), \dots, \phi\left(X^A_{t_N}\right)\right]$, $\mathbf{k}_{\mathbf{t}, t_n} := \left[k\left(t_1, t_n\right), \dots, k\left(t_N, t_n\right)\right]^{\top}$, $\mathbf{K_{X^A}}\left(X^A_t, X^A_{t'}\right) = \left\langle \phi\left(X^A_t\right), \phi\left(X^A_{t'}\right)\right\rangle$, $\mathbf{K_t}\left(t, t'\right) = \left\langle \phi\left(t\right), \phi\left(t'\right)\right\rangle$, *and* $\odot$ *represents point-wise product.*

With this proposition, we define $\hat{\Delta}_{X^A \to X^R}$ to measure the dependence between $P(X^R|X^A)$ and $P(X^A)$, $\hat{\Delta}_{X^R \to X^A}$ to measure the dependence between $P(X^A|X^R)$ and $P(X^R)$,

$$\begin{aligned} \hat{\Delta}_{X^A \to X^R} &= \frac{\operatorname{tr}(M_{X^A} H M_{\underline{X}^R X^A} H)}{\operatorname{tr}(M_{X^A} H) \operatorname{tr}(M_{\underline{X}^R X^A} H))}, \\ \hat{\Delta}_{X^R \to X^A} &= \frac{\operatorname{tr}(M_{X^R} H M_{\underline{X}^A X^R} H)}{\operatorname{tr}(M_{X^R} H) \operatorname{tr}(M_{\underline{X}^A X^R} H))}, \end{aligned} \tag{39}$$

where M_{X^A}, M_{X^R} are the Gram Matrices [13] for X^A and X^R, H is used to center the feature with entries $H_{ij} = \delta_{ij} - \frac{1}{N}$, and $\delta_{ij} = 1$ when $i = j$.

Calculating $\hat{\Delta}_{X^A \to X^R}$ and $\hat{\Delta}_{X^R \to X^A}$ is challenging because single cell sequencing data suffer from sparsity problems. We utilize the aggregated estimated peaks $\tilde{X}^A_t$ as X^A_t and averaged gene expression $\tilde{X}^R_t$ as X^R_t for all cells with time label t. Specifically,

$$\tilde{X}^A_t = \frac{1}{[T=t]} \sum_{t_i = t} \sum_{j \in N_g} l^A_i \hat{p}^A_{ij}, \tag{40}$$

$$\tilde{X}^R_t = \frac{1}{[T=t]} \sum_{t_i = t} l^R_i \hat{\rho}^R_{ig}, \tag{41}$$

where N_g is all matched peaks for gene g in the upstream, downstream, and proximal areas of the TSS (See section 6.1), $[T=t]$ is the number of samples/cells where the time label is equal to t. $\hat{p}^A_{ij}$ is the estimated parameter p^A_{ij} for peak j's Multinomial distribution. l^A_i is the estimated library size for peaks (scaling factor), $\hat{\rho}^R_{ig}$ is the shape variable in cell i of gene g's NB distribution, and l^R_i is the estimated library size for gene expression. Given this estimated $\tilde{X}^R_t$ and $\tilde{X}^A_t$, we can further calculate gene level couple and decouple scores.

6.3 Gene Level Couple Score and Decouple Score

By utilizing the individual gene level causal constraints, we develop couple and decouple scores to measure the "coupledness" and "decoupledness" of genes and their associated peaks. Given the estimated aggregated gene g expression $\tilde{X}^R_g$, and its corresponding aggregated peaks j value $\tilde{X}^A_j$. Note "average" gene and peaks mean that the gene or peaks are averaged across all cells or cells in a specific branch. To simplify notation, we drop the i and j in the rest of this section. Given a threshold value α, we define the following scores,

$$\begin{aligned} \text{Decouple Score} &= \operatorname{sign}(\alpha - \hat{\Delta}_{\tilde{X}^A \to \tilde{X}^R})(\hat{\Delta}_{\tilde{X}^A \to \tilde{X}^R} - \hat{\Delta}_{\tilde{X}^R \to \tilde{X}^A}) \\ \text{Couple Score} &= -\text{Decouple Score}. \end{aligned} \tag{42}$$

From the definition of decouple score and couple score, given a threshold value α, if the decouple score is positive, we can identify gene g and peaks j as decoupled; otherwise if the couple score is positive, g and peaks j are coupled.

7 Time Estimation

7.1 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. For the SSc human lung epithelium dataset, we use Palantir pseudotime to order cells [14], and for the mouse brain, mouse skin datasets, we utilize latent time inferred using MultiVelo [15].

7.2 Protein-based Latent Time Estimation

We can also utilize the normalized levels of important intra-nuclear proteins as a proxy variable for latent time. In the NEAT-seq dataset of CD4 memory T cells, the transcription factor (TF) antibody-derived tag (ADT) count data were normalized using NPC antibody counts and hashtag oligos (HTOs), as described in the original study [16]. The NPC normalization can be formulated as,

$$\text{Normalized counts} = \log_2(250 \times \frac{\text{TF ADT counts}}{\text{anti-NPC HTO counts}} + 1) \tag{43}$$

7.3 METAcell Construction

If temporal information is unavailable, we can infer representations of paired scRNA-seq and scATAC-seq data based on reconstruction loss. Since cells at a similar developmental stage should have similar omics profiles, individual cells exhibiting similar states are identified then aggregated into METAcells [17, 18]. These METAcells are constructed using cell-cell similarities inferred from the latent representations, which account for potential decoupling phenomena between scRNA-seq and scATAC-seq data. We then assess distribution shifts by utilizing METAcell membership information C.

Here, we summarize the averaged peaks $X^A_{c,j}$ and averaged gene expression $X^R_{c,g}$ for all cells with METAcell membership label c. Specifically,

$$X^A_{c,j} = \frac{1}{[C=c]} \sum_{c_i=c} X^A_{i,j}, \tag{44}$$

$$X^R_{c,g} = \frac{1}{[C=c]} \sum_{c_i=c} X^R_{i,g}, \tag{45}$$

$$X^A_{c,g} = \frac{1}{[C=c]} \sum_{c_i=c} \sum_{j \in N_g} X^A_{i,j} \tag{46}$$

$$\tag{47}$$

c_i represents the membership of cell i. $[C=c]$ is the number of samples/cells assigned to a METAcell label c. Given these summarized values $X^A_{c,g}$ and $X^R_{c,g}$, we can further calculate gene-level couple and decouple scores.

8 Evaluation Metrics

8.1 Genomic Matching Score

To evaluate the distance between important genes and peaks from coupled RNA and ATAC representation pairs, we calculate the genomic matching score. This score assesses the fraction of important peaks within the ATAC representation that lie within the distal gene regulation distance (up to 250kbp) from the transcription start site (TSS) of significant genes in the RNA representation.

First, we select pairs of coupled RNA and ATAC representations with the highest Pearson correlation coefficients. Next, we identify significant RNA features and peak features that contribute to the latent representations. Specifically, for the SHARE-seq dataset, we extracted the top 100 RNA features and the top 2000 peak features from the corresponding coupled representation pairs.

8.2 Silhouette Score

The Silhouette width [19] is used to evaluate the extent to which sample dissimilarity is minimized within a batch and maximized across batches. For a sample x, $d_a(x)$ is the average dissimilarity between x and all other data points of the batch A that x belongs to, while $d_c(x)$ is the average dissimilarity between x and the samples in batch C where $C \neq A$, meaning that $\min(d_c(x))$ is the average dissimilarity for the batch that is most similar to A. Using these definitions, we have the following silhouette width $s(x)$ of x,

$$s(x) = \frac{\min(d_c(x)) - d_a(x)}{\max(d_a(x), \min(d_c(x)))}. \tag{48}$$

Then the batch silhouette score for the cell label i is defined as,

$$\text{silhouette_score}_i = \frac{1}{|L_i|} \sum_{x \in L_i} 1 - |s(x)|, \tag{49}$$

where L_i is the set of cells with the cell label i and $|L_i|$ denotes the number of cells in that set. Then the average silhouette score is defined as

$$\text{silhouette score} = \frac{1}{M} \sum_{i \in M} \text{silhouette_score}_i, \tag{50}$$

where M is the number of unique cell labels.

8.3 HSIC on Batches

We utilize HSIC (see Section Hilbert Schmidt Independence Criterion (HSIC)) to evaluate how independent the latent representations are from batch variables. The lower the values are, the more independent the latent representations are from batch variables.

8.4 Normalized Mutual Information (NMI)

Normalized Mutual Information (NMI) is a metric used to evaluate the similarity between two clustering assignments. It scales the Mutual Information (MI) score to a value between 0 and 1, where 1 indicates perfect agreement and 0 indicates no mutual information between the clusterings.

The NMI between two clusterings $\mathcal{C}$ and $\mathcal{C}'$ is defined as:

$$\text{NMI}(\mathcal{C}, \mathcal{C}') = \frac{2 \times I(\mathcal{C}; \mathcal{C}')}{H(\mathcal{C}) + H(\mathcal{C}')}$$

where:

- $I(\mathcal{C}; \mathcal{C}')$ is the Mutual Information between $\mathcal{C}$ and $\mathcal{C}'$.
- $H(\mathcal{C})$ and $H(\mathcal{C}')$ are the entropies of $\mathcal{C}$ and $\mathcal{C}'$, respectively.

The Mutual Information $I(\mathcal{C}; \mathcal{C}')$ is calculated as:

$$I(\mathcal{C}; \mathcal{C}') = \sum_{i=1}^{k} \sum_{j=1}^{k'} p_{ij} \log \left(\frac{p_{ij}}{p_i p'_j} \right)$$

and the entropy $H(\mathcal{C})$ is:

$$H(\mathcal{C}) = -\sum_{i=1}^{k} p_i \log p_i$$

Similarly for $H(\mathcal{C}')$:

$$H(\mathcal{C}') = -\sum_{j=1}^{k'} p'_j \log p'_j$$

where:

- $p_{ij} = \frac{n_{ij}}{n}$ is the probability that a randomly chosen data point belongs to cluster i in $\mathcal{C}$ and cluster j in $\mathcal{C}'$.
- $p_i = \frac{a_i}{n}$ is the probability of cluster i in $\mathcal{C}$.
- $p'_j = \frac{b_j}{n}$ is the probability of cluster j in $\mathcal{C}'$.
- n_{ij} is the number of data points in both cluster i and cluster j.
- $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$ are the sums over rows and columns of the contingency table, respectively.
- n is the total number of data points.

8.5 Definition of Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) measures the similarity between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. It adjusts the Rand Index (RI) for the chance grouping of elements.

The ARI is defined as:

$$\text{ARI} = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}\right]}$$

where:

- n_{ij} is the number of data points in both cluster i in $\mathcal{C}$ and cluster j in $\mathcal{C}'$.
- $a_i = \sum_j n_{ij}$ is the number of data points in cluster i in $\mathcal{C}$.
- $b_j = \sum_i n_{ij}$ is the number of data points in cluster j in $\mathcal{C}'$.
- n is the total number of data points.
- $\binom{n}{2}$ denotes the binomial coefficient, representing the number of ways to choose 2 items from n.

9 Optimal Transport for Lineage Prediction

9.1 Optimal Transport

Given two probability vectors r and c, we consider the set of transport plans $U(r, c)$:

$$U(r, c) = \left\{P \in \mathbb{R}_+^{d \times d} \mid P\mathbf{1} = r, \quad P^\top \mathbf{1} = c\right\}$$

Here, $\mathbf{1}$ is a vector of ones. The set $U(r, c)$ includes all non-negative $d \times d$ matrices P where the sums of the rows equal r and the sums of the columns equal c.

The optimal transport distance between r and c using a cost matrix M and transport matrix P is defined as:

$$d_M(r, c) = \min_{P \in U(r,c)} \langle P, M \rangle$$

where $\langle P, M \rangle$ denotes the sum of the element-wise products of P and M. This represents the minimal total cost of transporting the distribution r to c under the cost M.

The Sinkhorn distance, introduced by Cuturi [20], is a widely used method to make optimal transport computations more efficient by incorporating entropic regularization, which simplifies the optimization process. The dual problem of the entropic regularized optimal transport becomes:

$$d_M^\lambda := \langle P^\lambda, M \rangle, \quad \text{where } P^\lambda = \underset{P \in U(r,c)}{\arg\min} \langle P, M \rangle - \frac{1}{\lambda} h(P). \tag{51}$$

Here, $\lambda > 0$ is the regularization parameter, and $h(P)$ denotes the entropy of P. Solving this regularized problem is computationally much cheaper than the classical optimal transport problem for suitable values of λ [20]. In the following sections, we use `SinkhornOT`() to represent the function to compute P given r, c, and M.

9.2 Lineage Prediction with Optimal Transport

We utilize optimal transport to further validate that the latent coupled/decoupled representations are able to predict lineage trajectory of cells. For each cell i, with $i = 1, 2, \ldots, n$, we have the (decoupled or coupled or both) latent representations $\mathbf{Z}_i$. Then we construct the cost matrix, where each element is the distance d between each cells. Formally, the cost matrix can be formulated as follows,

$$M = \begin{bmatrix} d(\mathbf{Z}_1, \mathbf{Z}_1), d(\mathbf{Z}_1, \mathbf{Z}_2), d(\mathbf{Z}_1, \mathbf{Z}_3), \ldots, d(\mathbf{Z}_1, \mathbf{Z}_n) \\ d(\mathbf{Z}_2, \mathbf{Z}_1), d(\mathbf{Z}_2, \mathbf{Z}_2), d(\mathbf{Z}_2, \mathbf{Z}_3), \ldots, d(\mathbf{Z}_2, \mathbf{Z}_n) \\ \vdots \\ d(\mathbf{Z}_n, \mathbf{Z}_1), d(\mathbf{Z}_n, \mathbf{Z}_2), d(\mathbf{Z}_n, \mathbf{Z}_3), \ldots, d(\mathbf{Z}_n, \mathbf{Z}_n) \end{bmatrix} \tag{52}$$

Then we solve the Optimal Transport problem using Sinkhorn Distance with the defined cost matrix M, initial cell states r, and the end cell states c. The goal is to compute the transportation matrix P. Specifically,

1. Compute M with Eq. 52.
2. Compute c, where $c_i = \frac{\#\texttt{initial cell type i}}{\#\texttt{initial state cells}}$ for all initial state cell types i.
3. Compute r, where $r_j = \frac{\#\texttt{end cell type j}}{\#\texttt{end state cells}}$ for all the end state cell types j.
4. Compute the transportation matrix $P = \texttt{SinkhornOT}(r, c, M)$.
5. Aggregate the matrix P to get cell type transition matrix T, where $T_{ij} = \frac{\sum_{lk} p_{lk}}{\sum_l p_{l\cdot}}$, l is the cell index in cell type i, k is cell index in cell type j.

Finally we utilize the POT v0.9.3 package to compute optimal transport.

10 Granger Causality-based Inference for Gene Regulatory Interactions

In this section, we introduce Granger causality and the Granger causal test, then we present the application of the Granger causal test to identify the distal, time-lagged regulation relations between the open chromatin regions and expression of nearby genes.

10.1 Granger Causality and Tests

Granger [21] proposed a notion of causality based on how well past values of a time series y_t could predict future values of another series x_t. Let $H_{<t}$ be the history of all relevant information up to time $t-1$ and $P(x_t|H_{<t})$ be the optimal prediction of x_t given $H_{<t}$. Granger defined y to be causal for x if

$$\text{var}(x_t - P(x_t|H_{<t})) < \text{var}(x_t - P(x_t|H_{<t/y<t})), \tag{53}$$

where $H_{<t/y<t}$ is the optimal prediction that does not include the information of y up to time t. Granger causality assumes causal effects are ordered in time (i.e., cause before effect), under some assumptions: if y can predict x, then there must be a mechanistic (i.e., causal) effect; that is, predictability implies causality. Granger causality corresponds to nonzero entries in the autoregressive coefficients. In particular, for a bivariate model:

$$\begin{aligned} \text{Full Model: } a_x^0 X_t &= \sum_{k=1}^{d} a_{xx}^k x_{t-k} + \sum_{k=1}^{d} a_{xy}^k y_{t-k} + e_{t,x}, \\ \text{Reduced Model: } a_x^0 X_t &= \sum_{k=1}^{d} a_{xx}^k x_{t-k} + e_{t,x}. \end{aligned} \tag{54}$$

Series y is Granger causal for series x if and only if $a_{xy}^k \neq 0$ for some $1 \leq k \leq d$. However, Granger causality based on only two variables severely limits the interpretation of the findings: Without adjusting for all relevant covariates, a key assumption of Granger causality is violated. In another words, Granger causality cannot resolve any latent confounders between two variables. Despite this limitation, Granger causal tests

are still capable of revealing meaningful correlations between variables with time lagging information. Here we briefly introduce the Granger causal tests.

$$H_0 : y \text{ does not Granger-cause } x. \tag{55}$$

$$H_1 : y \text{ Granger-causes } x. \tag{56}$$

We can then construct the **F-test** test statistic

$$F = \frac{(RSS_R - RSS_F)/d}{RSS_F/(N - 2d - 1)}, \tag{57}$$

where RSS_R, RSS_F are the residual sums of squares from the reduced and full models, respectively; N is the total number of samples; and d is the number of time lags [21]. If F exceeds the critical value from the F-distribution under the null hypothesis, reject H_0. We can also utilize the **likelihood ratio test**,

$$LR = -2\left(\log \frac{L_R}{L_F}\right), \tag{58}$$

where L_R, L_F are the likelihoods of the restricted and unrestricted models, respectively. Compute the LR statistic and compare it against a p-value from the chi-squared distribution with degrees of freedom equal to the number of time lags. If the LR statistic is greater than the p-value, reject H_0.

10.2 Granger Causal Regulatory Interaction Inference

In this section, we show how we utilize the Granger Causality test to reveal distal regulation with time lags. Specifically, we test whether peaks Granger-cause expression of nearby genes with time lags (Algorithm 1).

10.2.1 Given Latent Time T

We introduce the algorithm to identify the Granger causal relations between a peak or some aggregated peak count $\tilde{X}^A$ and gene expression of nearby gene $\tilde{X}^R$ (for simplicity of notation, we drop the gene and cell indices here). In this test, we determine whether $\tilde{X}^A$ Granger-causes $\tilde{X}^R$.

Algorithm 1 Granger Causal Regulation Inference

Require: peaks $\tilde{X}^A_T$, expression of nearby gene $\tilde{X}^R_T$, and time lag range $[1, \ldots, d]$, time label on each cell.
Ensure: Granger causal p-value given different time lags.
  Sort the cells by their time label t.
  $k = 1$
  while $k \leq d$ **do**
    Conduct Granger causality test with $\tilde{X}^A_{T=t-m}$, $\forall m \in [1, \ldots k]$, and $\tilde{X}^R_{T=t}$;
    record the p-value;
    $k \mathrel{+}= 1$;
  end while

10.2.2 Given Intra-nuclear Protein Information

The Granger causal regulatory interaction inference can be extended to use the protein levels of intra-nuclear TFs. By utilizing TF protein information as a proxy variable for latent time to order cells, the Granger causality test can be used to uncover TF-associated regulation by downstream enhancers of their gene targets.

11 Individual Gene Level Simulation

The simulation process consists of two main steps:

1. **Simulating scRNA-seq Gene Expression Data**: For each gene g in cell i, we generate gene expression data using the Negative Binomial distribution:

$$X_{i,g}^R \sim \text{NegativeBinomial}(\rho_{ig}^R, \theta_g^R)$$

2. **Simulating Corresponding Peaks**: For each peak j in cell i, we generate data using the Bernoulli distribution:

$$X_{i,j}^A \sim \text{Bernoulli}(p)$$

While a Multinomial distribution could also be used, we opt for the Bernoulli distribution here for simplicity, as suggested in [10].

Based on the previous sections, we establish the following causal relationships:

$$\begin{aligned}
&X_{i,j}^A = f_A(T, E_A) \\
&X_{i,g}^R = f_R(T, X_{i,j}^A, E_R) \\
&X_{i,j}^A \sim \text{Bernoulli}(p_{ij}^A(T)) \\
&X_{i,g}^R \sim \text{NegativeBinomial}(\rho_{ig}^R(X_{i,j}^A, T), \theta_g^R(X_{i,j}^A, T)),
\end{aligned}$$

where E_A and E_R are the exogenous variables for gene expression and peaks respectively. The procedures of simulating time T and peaks $X_{i,j}^A$ can be formulated as follows,

$$\begin{aligned}
&T \sim \text{Uniform}(0, 1) \\
&p_{ij}^A = T \\
&X_{i,j}^A \sim \text{Bernoulli}(p_{ij}^A).
\end{aligned} \tag{59}$$

Note that here the parameter p_{ij}^A is a function of T (see section Problem Definition). Next, we simulate the gene expression $X_{i,g}^R$ for the coupled and decoupled cases. For both coupled and decoupled cases,

$$\begin{aligned}
&\text{Coupled case:} \quad p_{ij}^A \not\perp\!\!\!\perp \rho_{ig}^R \\
&\text{Decoupled case:} \quad p_{ij}^A \perp\!\!\!\perp \rho_{ig}^R
\end{aligned} \tag{60}$$

Specifically, for the coupled case, we simulate the following distributions,

$$\rho_{ig}^R = T + X_{i,j}^A/100, \quad \theta_g^R = T/10. \tag{61}$$

For the decoupled case, we simulate the following three distributions,

$$\begin{aligned}
&1. \quad \rho_{ig}^R = 1 - \text{Uniform}(0, T + X_{i,j}^A/100), \quad \theta_g^R = \exp(10T), \\
&2. \quad \rho_{ig}^R = |\cos(100T + 100X_{i,j}^A)|, \quad \theta_g^R = \exp(10T), \\
&3. \quad \rho_{ig}^R = \frac{1}{\exp(100T + X_{i,j}^A)}, \quad \theta_g^R = \exp(10T).
\end{aligned} \tag{62}$$

We repeat the simulation 100 times, with a sample size of 300 for each simulation.

12 Dataset Preprocessing and Settings

12.1 SHARE-seq Mouse Skin (Hair Follicle) Dataset

We obtained this processed dataset directly from MultiVelo website[15]. There are total of 6,436 cells and 962 genes in the MultiVelo processed data. For the latent representations, the dimensions of latent coupled and decoupled representations are both 20. The Granger causality test was performed using likelihood ratio test to assess cis-regulation.

Algorithm 2 Gene-level Simulation

Require: Sample Size m and Repeated Times n.

Ensure: Simulated time T, peaks value $X^A_{i,j}$ and corresponding coupled gene expression value $X^R_{i,c}$, and the decoupled gene expression value $X^R_{i,d}$.

i = 0

while $i < n$ **do**

1. Sample the time T following Eq. 59.
2. Sample the peaks $X^A_{i,j}$ with time T, following Eq. 59.
3. Sample the coupled gene expression $X^R_{i,c}$ with Eq. 61 and decoupled gene expression $X^R_{i,d}$ with Eq. 62.
4. $n \leftarrow n + 1$

end while

12.2 NEAT-Seq Dataset

The NEAT-seq dataset profiles primary human CD4 memory T cells using a panel of master TFs that drive T cell subsets, including Tbet, GATA3, RORγT, FOXP3, and Helios [16]. We obtained the processed dataset from GSE178707. After filtering out low-quality droplets, the dataset comprised 3370 cells with 13,380 genes and 78,203 peaks. The dimensions of the latent coupled and decoupled representations are both set to 30. The Granger causality test was performed using F-test explore gene regulation.

Binding sites of Tbet and GATA3 for Type 1 T helper (Th1) and type 2 T helper (Th2) cells were downloaded from Data S1 of https://doi.org/10.1038/ncomms2260 [22]. The set of Th1-specific T-bet super enhancers was downloaded from https://doi.org/10.1016/j.celrep.2016.05.054 [23].

To validate the effect of coexpressing T-bet and GATA3 on T cells compared to T cells expressing T-bet only, bulk RNA-seq data of EL4-Tbet+GATA3 and EL4-Tbet+Plum cells were downloaded from GSE171410.

12.3 SSc Human Lung Epithelium Dataset

We used the 10X Genomics Multiome technology for paired snRNA-seq/snATAC-seq on nuclei from 6 SSc-ILD and 7 control human lung explants. For more details, please see [24].

4708 alveolar epithelial and terminal secretory cells, with 3000 highly variable genes and 178316 peaks, were used as input for the HALO model. Initially, the HALO model was trained based on ELBO loss only, and then the inferred latent representations were used for clustering and UMAP embedding construction using Scanpy [25]. Identified clusters were annotated using previously described cell markers [26]. Subsequently, the HALO model was trained using Palantir pseudotime for latent coupled and decoupled representations, where the dimensions of both are set to 20. The Granger causality test was performed using likelihood ratio test for distal regulation identification.

The spliced and unspliced RNA count matrices were generated using Cell Ranger output BAM files with Velocyto CLI (v0.17.17) [27]. These matrices were then processed with scVelo (v0.3.1) to compute RNA velocity using the "Dynamical" mode [28]. The RNA velocity-based cell-cell transition matrix was combined with cellular similarity to infer the initial and terminal states of cellular dynamics using CellRank (v2.0.2) [29]. Cells from the initial state were used as the root to compute pseudotime with Palantir (v1.3.2) [14].

To evaluate the information captured by HALO coupled and decoupled representations, a Multi-layer Perceptron classifier (scikit-learn v1.1.3) was trained to predict pulmonary epithelial cell type annotation.

Based on HALO representations, a cell-cell transition matrix was inferred using entropic-regularized optimal transport with a Sinkhorn solver (package POT v0.9.3) [20] and a cost matrix inferred based on squared Euclidean distance. This cell-cell transition matrix was then further aggregated into a cluster-cluster transition matrix. Additionally, we performed differential abundance testing to evaluate differences in cell abundances associated with disease using Milopy (v0.1.1) [30].

TF binding sites in the top important peaks were identified using FindMotifs function from the Signac package [31]. Association between top peaks and super enhancer sets was evaluated using the hypergeometric test, where super enhancer sets were downloaded from Data S1 of https://doi.org/10.1016/j.cell.2013.09.053 [32] and SEdb 2.0 [33]. We estimated DNA motif activity score using the Signac wrapper for chromVAR with motif profiles from the JASPER 2022 database [34, 35].

The scRNA-seq data of human lung organoids from GSE178360 were re-analyzed using Seurat and Harmony for normalization, batch correction, dimensionality reduction, and clustering [36, 37]. Marker genes

and module scores for AT0, AT1, AT2, and TRB-SC were used to annotate cell types [26]. AT2 (Cluster 0) was isolated for differential gene expression analysis by comparing the EGF depletion group to the control group (Fig. S10). Differentially expressed genes (with absolute logFC > 0.5, min.pct > 0.2, and p.adj < 0.05) were used to calculate the EGF depletion module score for SSc lung epithelium data.

12.4 NeurIPS Dataset

The NeurIPS single-cell multiomics dataset was collected from mobilized peripheral CD34+ hematopoietic stem and progenitor cells (HSPCs) isolated from four human donors at five time points. Samples were prepared using a standard protocol at four sites. The dataset was designed with a nested batch layout, with some donor samples measured at multiple sites and some donors measured at a single site. The processed data was downloaded from GSE194122. In our paper, we subset 10952 cells (Erythroblast, HSC, MK/E progenitor, Normoblast, and Proerythroblast) with pseudotime information for downstream analysis.The latent representation dimensions for scRNA-seq and scATAC-seq are both 10.

12.5 Human Brain Dataset

We obtained this processed dataset directly from GSE162170. We followed MultiVelo's protocol to process the dataset. The dimensions of the latent coupled and decoupled representations are both set to 20.

12.6 10x Embryonic E18 Mouse Brain Dataset

We obtained this processed dataset directly from the MultiVelo website [15]. In total, it includes 3,365 cells, 936 highly variable genes, and 112,656 peaks; these matrices were then used for downstream analysis. The latent coupled and decoupled representation dimensions are both 10.

13 Enrichment Analysis

With the interpretable decoder of HALO, we can understand how a latent representation contributes to the reconstruction of gene or peak features. This allows us to obtain a list of gene or peak features contributing to a latent representation for enrichment analysis.

13.1 Motifs

Using the JASPAR CORE collection, we compute motif profiles within scATAC-seq peaks using the MOODS algorithm [38], setting the adjusted p-value threshold to $p < 1 \times 10^{-5}$. Then, we identify TFs with predicted binding sites that are enriched in top peaks of ATAC latent representations $\mathbf{Z}^A$ (versus the remaining peaks) using the Fisher exact test.

13.2 Genes

HALO uses Enrichr to find overlaps between the top genes from RNA latent representations $\mathbf{Z}^R$ and precompiled ontologies [39].

14 Gene Regulatory Network

We utilize the HALO representation to construct TF-CRE-gene linkages from scRNA/scATAC-seq data. For the regulatory network of AT2 cells, we use our larger independent scRNA-seq dataset ($n = 17$ SSc, 13 control lungs) to identify differentially expressed genes (DEGs) for the comparisons of interest [40].

14.1 METAcell

Due to the sparsity inherent in single-cell genomics data, we aggregate individual cells exhibiting similar states into METAcells (see METAcell Construction for details).

14.2 Identification of Cis-Regulatory Elements

To identify cis-regulatory elements (CREs), we adopt the nonlinear regression-based framework, DIRECT-NET [41]. Specifically, we use the non-linear predictive model XGBoost to regress the expression levels of

genes against the accessibility values of candidate distal CREs, defined as peaks located within 250 kbp upstream or downstream of the transcription start site. CREs are selected based on the importance scores derived from the XGBoost model.

14.3 Construction of gene regulatory network

To construct a disease-associated regulatory network, we first identify DEGs (logFC > 0.5 and p.adj < 0.05) across conditions. Next, we identify differentially accessible CREs (logFC > 0.1 and p.adj < 0.05) corresponding to these DEGs. TFs within the differentially accessible CREs and promoter regions are identified using the JASPAR motif database and the motifmatchr function in the ChromVAR package [34, 35].

15 Training and Testing Setups

We extract 20% of the data from each dataset for testing. The remaining data is split into training and validation sets with a 4:1 ratio. For every dataset, we conducted an exhaustive hyperparameter grid search with `n_epoch`=100 on the total loss with training and validation datasets. We denote `dim` as the dimension of the decoupled or coupled representations and we further set $\texttt{dim}(\mathbf{Z}_d^R) = \texttt{dim}(\mathbf{Z}_d^A) = \texttt{dim}(\mathbf{Z}_c^R) = \texttt{dim}(\mathbf{Z}_c^A)$. The search spaces for the hyperparameters $[\omega_1, \omega_2, \alpha, \texttt{dim}]$ are shown as follows:

$$\omega_1 = [1, 10, 50, 100, 500], \tag{63}$$

$$\omega_2 = [0, 0.001, 0.01, 0.1, 1, 10, 100] \tag{64}$$

$$\alpha = [0.001, 0.005, 0.01, 0.065, 0.1] \tag{65}$$

$$\texttt{dim} = [5, 10, 15, 20, 30] \tag{66}$$

HALO is optimized using the Adam optimizer with a learning rate of 0.01, a weight decay of 0.001, and a minibatch size of 1024.

16 Ablation Study on Hyperparameters

16.1 Hyperparameters Tuning

Here we examine the following hyperparameters: α in the coupled and decoupled constraints; ω_1, the weight of causal constraints; and ω_2, the weight of the modality aligner. In the ablation studies for α, ω_1, and ω_2, we fix the latent representation dimensions as 10 for both scRNA-seq and scATAC-seq.

16.2 α

The ablation study for different values of α was performed on two datasets, the mouse brain and human brain datasets, with $\alpha = [0.1, 0.065, 0.02]$ For this ablation study, we fixed $\omega_1 = 10$ and $\omega_2 = 10$. We report the values of $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}$, $\hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$, $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R}$, $\hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$, as well as $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R} - \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$ and $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} - \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$. We also report the clustering results for cell type annotation in terms of NMI and ARI in Fig. **??**. From the results, we can see that generally, the results are most robust when $\alpha = 0.065$.

16.3 ω_1

We conducted the ablation study of the causal constraints weights ω_1 on the mouse brain dataset. ω_1 are in the range of $[1, 10, 50, 100, 500]$, while we fixed $\alpha = 0.065$ and $\omega_2 = 10$. We report the values of $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}$, $\hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$, $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R}$, $\hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$, as well as $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R} - \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$ and $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} - \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$. We also report the clustering results for cell type annotation in terms of NMI and ARI in Fig. **??**. From the results, we find that $\omega_1 = 10$ generally achieves the most stable performance.

16.4 ω_2

We conducted the ablation study of the modality aligner weights ω_2 on the mouse brain dataset. ω_2 are in the range of $[0, 0.001, 0.01, 0.1, 1, 10, 100]$, while we fixed $\alpha = 0.065$ and $\omega_1 = 10$. We report the values of $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}$, $\hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$, $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R}$, $\hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$, as well as $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R} - \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$ and $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} - \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$. We also

report the clustering results for cell type annotation in terms of NMI and ARI in Fig. **??**. From the results, we find that $\omega_2 = 10$ generally achieves the most stable performance.

16.5 Number of Latent Representation Dimensions `dim`

We conducted the ablation study of the latent representation dimensions n on the mouse brain dataset. n are in the range of $[6, 10, 18, 30]$, while we fixed $\alpha = 0.065$, $\omega_1 = 10$, and $\omega_2 = 10$. We report the values of $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R}$, $\hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$, $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R}$, $\hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$, as well as $\hat{\Delta}_{\mathbf{Z}_c^A \to \mathbf{Z}_c^R} - \hat{\Delta}_{\mathbf{Z}_c^R \to \mathbf{Z}_c^A}$ and $\hat{\Delta}_{\mathbf{Z}_d^A \to \mathbf{Z}_d^R} - \hat{\Delta}_{\mathbf{Z}_d^R \to \mathbf{Z}_d^A}$. We also report the clustering results for cell type annotation in terms of NMI and ARI in Fig. **??**. From the results, we find that $n = 10$ generally achieves the most stable performance.

References

[1] 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) (2020)

[2] Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, Prediction, and Search. MIT press, USA (2000)

[3] Pearl, J., Bareinboim, E.: Transportability of causal and statistical relations: A formal approach. In: Proc. AAAI 2011, pp. 247–254 (2011)

[4] Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: ICML-13 (2013)

[5] 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)

[6] Pearl, J.: Causality, 2nd edn. Cambridge University Press, Cambridge (2009)

[7] 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 (2005). Springer

[8] 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)

[9] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018)

[10] Ashuach, T., Gabitto, M.I., Jordan, M.I., Yosef, N.: Multivi: deep generative model for the integration of multi-modal data. bioRxiv (2021)

[11] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., Hinton, G.E.: Neural additive models: Interpretable machine learning with neural nets. Advances in neural information processing systems **34**, 4699–4711 (2021)

[12] 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)

[13] Schölkopf, B.: The kernel trick for distances. Advances in neural information processing systems **13** (2000)

[14] Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., Pe'Er, D.: Characterization of cell fate probabilities in single-cell data with palantir. Nature biotechnology **37**(4), 451–460 (2019)

[15] 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 (2022)

[16] Chen, A.F., Parks, B., Kathiria, A.S., Ober-Reynolds, B., Goronzy, J.J., Greenleaf, W.J.: Neat-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. Nature Methods **19**(5), 547–553 (2022)

[17] Mao, H., Jia, M., Dou, J.X., Zhang, H., Benos, P.V.: Coem: cross-modal embedding for metacell identification. arXiv preprint arXiv:2207.07734 (2022)

[18] Persad, S., Choo, Z.-N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C.C., Sharma, R., Pe'er, I., *et al.*: Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nature Biotechnology **41**(12), 1746–1757 (2023)

[19] Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., Theis, F.J.: A test metric for assessing single-cell rna-seq batch correction. Nature methods **16**(1), 43–49 (2019)

[20] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)

[21] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424–438 (1969)

[22] Kanhere, A., Hertweck, A., Bhatia, U., Gökmen, M.R., Perucha, E., Jackson, I., Lord, G.M., Jenner, R.G.: T-bet and gata3 orchestrate th1 and th2 differentiation through lineage-specific targeting of distal regulatory elements. Nature communications **3**(1), 1268 (2012)

[23] Hertweck, A., Evans, C.M., Eskandarpour, M., Lau, J.C., Oleinika, K., Jackson, I., Kelly, A., Ambrose, J., Adamson, P., Cousins, D.J., *et al.*: T-bet activates th1 genes through mediator and the super elongation complex. Cell reports **15**(12), 2756–2770 (2016)

[24] Valenzi, E., Jia, M., Gerges, P., Fan, J., Tabib, T., Behera, R., Zhou, Y., Sembrat, J., Das, J., Benos, P.V., Singh, H., Lafyatis, R.: Altered ap-1, runx, and egr chromatin dynamics drive fibrotic lung disease. bioRxiv (2024)

[25] Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)

[26] Kadur Lakshminarasimha Murthy, P., Sontake, V., Tata, A., Kobayashi, Y., Macadlo, L., Okuda, K., Conchola, A.S., Nakano, S., Gregory, S., Miller, L.A., *et al.*: Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. Nature **604**(7904), 111–119 (2022)

[27] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., *et al.*: Rna velocity of single cells. Nature **560**(7719), 494–498 (2018)

[28] 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 (2020)

[29] Weiler, P., Lange, M., Klein, M., Pe'er, D., Theis, F.: Cellrank 2: unified fate mapping in multiview single-cell data. Nature Methods, 1–10 (2024)

[30] Dann, E., Henderson, N.C., Teichmann, S.A., Morgan, M.D., Marioni, J.C.: Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nature biotechnology **40**(2), 245–253 (2022)

[31] Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., Satija, R.: Single-cell chromatin state analysis with signac. Nature methods **18**(11), 1333–1341 (2021)

[32] Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., Young, R.A.: Super-enhancers in the control of cell identity and disease. Cell **155**(4), 934–947 (2013)

[33] Wang, Y., Song, C., Zhao, J., Zhang, Y., Zhao, X., Feng, C., Zhang, G., Zhu, J., Wang, F., Qian, F., *et al.*: Sedb 2.0: a comprehensive super-enhancer database of human and mouse. Nucleic Acids Research

51(D1), 280–290 (2023)

[34] Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., *et al.*: Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic acids research **50**(D1), 165–173 (2022)

[35] Schep, A.N., Wu, B., Buenrostro, J.D., Greenleaf, W.J.: chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods **14**(10), 975–978 (2017)

[36] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., Raychaudhuri, S.: Fast, sensitive and accurate integration of single-cell data with harmony. Nature methods **16**(12), 1289–1296 (2019)

[37] Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., *et al.*: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature biotechnology **42**(2), 293–304 (2024)

[38] Korhonen, J.H., Palin, K., Taipale, J., Ukkonen, E.: Fast motif matching revisited: high-order pwms, snps and indels. Bioinformatics **33**(4), 514–521 (2017)

[39] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., *et al.*: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research **44**(W1), 90–97 (2016)

[40] Papazoglou, A., Huang, M., Bulik, M., Lafyatis, A., Tabib, T., Morse, C., Sembrat, J., Rojas, M., Valenzi, E., Lafyatis, R.: Epigenetic regulation of profibrotic macrophages in systemic sclerosis–associated interstitial lung disease. Arthritis & rheumatology **74**(12), 2003–2014 (2022)

[41] Zhang, L., Zhang, J., Nie, Q.: Direct-net: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. Science advances **8**(22), 7393 (2022)