Identifiable Feedback-Controlled Latent Flow for Unpaired Single-cell Spatio-Temporal Dynamics

Jianle Sun[1], **Kun Zhang**[1,2]
[1] Carnegie Mellon University;
[2] Mohamed bin Zayed University of Artificial Intelligence

Abstract

Reconstructing the gene regulatory dynamics underlying tissue development from unpaired, cross-sectional spatial transcriptomic snapshots poses a significant challenge in single-cell biology. While recent optimal transport and flow-based methods have advanced trajectory interpolation, their unconstrained latent dynamics lack identifiability, hindering reliable discovery of the underlying regulatory mechanisms. To address this, we propose single-cell Identifiable Feedback-controlled Latent Flow (scIFLF), a generative framework that introduces a physics-inspired structural prior that decomposes the latent flow into a macroscopic developmental drift and a restorative feedback toward functional attractors. Crucially, the gene-regulatory Jacobian is shown to be uniquely identifiable, invariant to the latent ambiguity, enabling principled discovery of gene regulations directly from the learned flow. Implemented with a multimodal Neural ODE, scIFLF aligns latent trajectories using entropy-regularized optimal transport. Benchmark experiments demonstrate that scIFLF outperforms state-of-the-art methods in trajectory interpolation and spatial coherence, while successfully recovering key driver genes of tissue development and organogenesis, bridging the gap between deep generative flexibility and biological interpretability.

1 Introduction

Understanding how gene regulation drives continuous spatio-temporal cell fate transitions during tissue development is a central problem in biology. While spatial transcriptomics provides critical, high-resolution snapshots of cellular organization, the destructive nature of data acquisition yields inherently **unpaired cross-sectional samples**. This fundamental limitation obscures the recovery of continuous causal paths of cell fate, making reconstructing the underlying developmental dynamics intrinsically challenging [1].

Recent approches have explored Optimal Transport (OT) [2] and Flow-based frameworks [3] to interpolate trajectories between snapshots while suffering from underdetermined latent dynamics. Methods such as Waddington-OT [2], Spateo [4] and Moscot [5] leverage OT-coupling to align spatial snapshots across time, yet these pairwise matching approaches often struggle with complex, non-monotonic temporal dynamics and lack a mechanistic biophysical grounding. While deep generative models like Neural Ordinary Differential Equations (ODEs) [6] and OT-induced flows like STORIES [7] and NicheFlow [8] can flexibly interpolate between distributions, their latent spaces are notoriously unidentifiable (non-unique) [9]. In other words, they confronts the severe risk of generating artifactual vector fields that fail to reflect genuine underlying cell dynamics [10]. This lack of identifiability severely limits biological interpretability: different latent dynamics can generate

Source codes are available at `https://github.com/sjl-sjtu/scIFLF`.
Email Address: jianles@andrew.cmu.edu; kunz1@andrew.cmu.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

similar observations while implying drastically distinct gene regulatory mechanisms, preventing reliable causal interpretation in biology [11].

To bridge the gap between deep generative flexibility and biological interpretability, we propose the single-cell Identifiable Feedback-controlled Latent Flow (scIFLF), which integrates high-dimensional molecular profiles with their spatial microenvironmental contexts into a unified, continuous latent flow. We conceptualize cellular development not merely as statistical interpolation, but as a structured continuous dynamical system. Drawing inspiration from physics and control theory, scIFLF reformulates the latent trajectory as a superposition of two forces: a macroscopic developmental drift in the Waddington landscape [12], and an internal restorative feedback induced by gene interactions that steers the cell toward its functional attractor. Remarkably, by constraining the biologically motivated latent dynamics, we establish rigorous identifiability theory, showing that the crucial internal gene regulatory network (GRN) can be **uniquely recovered** regardless of the latent space ambiguity.

We summarize our main contributions as follows: (1) We introduce scIFLF, a physics-inspired generative model with a continuous-time multimodal Neural ODE and efficient mini-batch latent-space OT, to model the feedback-controlled latent flow that jointly drives the spatial-temporal gene expression dynamics; (2) We establish a rigorous identifiability theory, enabling the principled discovery of gene regulation relationships during the dynamic developmental process; (3) Extensive benchmarks demonstrate scIFLF's superior performance on spatiotemporal interpolation, and further biological inference successfully and robustly recovers key genes driving tissue development and organogenesis.

2 Related Work

Optimal Transport on Spatio-Temporal Trajectories. Reconstructing cellular dynamics from inherently unpaired cross-sectional snapshots is fundamentally a distribution alignment problem. OT has been highly successful, beginning with frameworks like Waddington-OT [2] for dissociated scRNA-seq. To accommodate the spatial modality, recent methods such as Moscot [5] utilize Fused Gromov-Wasserstein (FGW) OT [13] to compute discrete transition couplings between consecutive tissue slices, while tools like Spateo [4] estimate localized spatiotemporal vector fields. However, they generally lack a generative global, continuous dynamical structure, which fundamentally limits their capacity to model complex biological processes and precludes deeper mechanistic interpretations.

Generative Flows and Identifiability. To explicitly model continuous temporal evolution, continuous normalizing flows and flow matching paradigms [14] have emerged as state-of-the-art architectures. For unpaired single-cell data without observed trajectories, recent advances like STORIES [7] and NicheFlow [8] bridge OT and generative flow by introducing OT-induced flow [3, 15] or variational flow matching [16]. While these flow-based frameworks (along with standard Neural ODEs [6]) offer exceptional mathematical flexibility in interpolating cellular states, their unconstrained neural vector fields typically results in the lack of identifiability [9, 17]: the learned latent structure is inherently underdetermined, making the underlying biological mechanisms uninterpretable [11]. In contrast, recent identifiable temporal causal representation learning [18, 17, 19] largely focuses on observed time-series trajectories, leaving the more challenging unpaired cross-sectional setting underexplored. This directly motivates scIFLF, to leverage a physics-inspired feedback-control structure to achieve meaningful latent identifiability and reliable discovery of underlying gene regulations.

3 Model

Consider a biological developmental process where we observe **unpaired** cross-sectional samples drawn from marginal distributions $p_t(X, S)$ across discrete time points $t_0 < t_1 < \cdots < t_T$, where $\boldsymbol{X} \in \mathbb{R}^G$ denotes the gene expression vector across G genes, and $\boldsymbol{S} \in \mathbb{R}^2$ represents the cell coordinate. At a high level, we assume that the observed high-dimensional molecular profiles and spatial coordinates are both driven by a shared, lower-dimensional latent dynamical system, and we aim to uncover the continuous latent dynamics that generate these observed temporal profiles. We postulate that the true data generation proceeds through the following coupled mechanisms.

Latent Dynamics. We conceptualize the developmental trajectory of each cell i as a continuous evolution within a low-dimensional latent state space $\mathcal{Z} = \mathbb{R}^d$. Drawing intuition from physics and control theory, we postulate that the instantaneous velocity $\boldsymbol{v}(\boldsymbol{Z}_{it})$ of a cell i is governed by

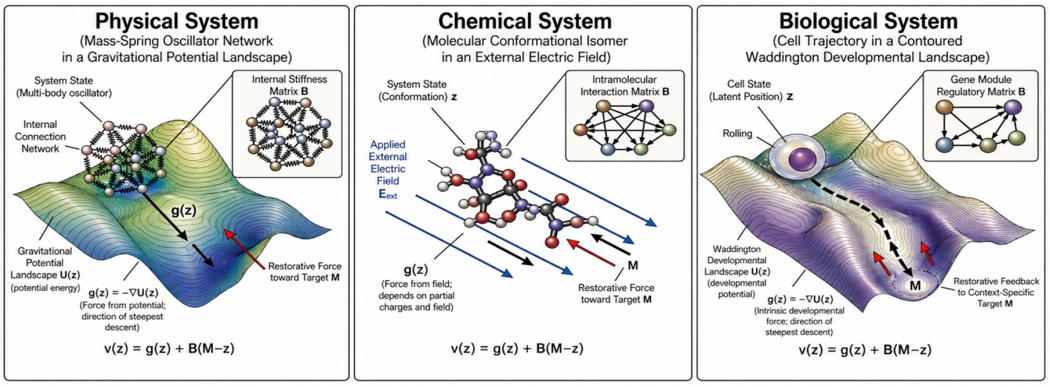


Figure 1: Physical analogy of latent dynamics. Left: a multi-body mass-spring oscillator in a gravitational field. Center: a molecular conformational isomer in an external electric field. Right: a cell evolving on the Waddington developmental landscape. The velocity field takes the form $\boldsymbol{v}(\boldsymbol{z}) = g(\boldsymbol{z}) + \mathbf{B}(\boldsymbol{M} - \boldsymbol{z})$: a non-linear drift induced by an external potential and a restorative force toward a specific target $\boldsymbol{M}$, mediated by the regulatory-interaction matrix $\mathbf{B}$.

a superposition of a macroscopic field shaping the overall developmental landscape (Waddington landscape) [12] and an innate local restorative feedback encoding the intrinsic regulation. Specifically, the latent state $\boldsymbol{Z}_{it}$ evolves as a continuous flow according to the following ODE

$$\boldsymbol{v}(\boldsymbol{Z}_{it}) := \frac{d\boldsymbol{Z}_{it}}{dt} = \underbrace{g(\boldsymbol{Z}_{it})}_{\text{developmental drift}} + \underbrace{\mathbf{B}(\boldsymbol{M}(\boldsymbol{Z}_{it}) - \boldsymbol{Z}_{it})}_{\text{restorative feedback}}. \tag{1}$$

Here, $g : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth nonlinear vector field representing the macroscopic developmental drift, while $\mathbf{B}(\boldsymbol{M}(\boldsymbol{Z}_{it}) - \boldsymbol{Z}_{it})$, introduces a internal restorative force, where $\mathbf{B} \in \mathbb{R}^{d\times d}$ acts as the controller gain matrix (innate interaction) continuously steering the cell toward its specific target state (attractor) $\boldsymbol{M}(\boldsymbol{Z}_{it}) \in \mathbb{R}^d$. This decomposition is unique up to a normalization, which Assumption 4 fixes. This formulation provides a highly interpretable biophysical grounding, which we formalize through two complementary lenses.

Remark 1 (**Physics Lens: Superposition of External Field and Internal Interactions**). *Analogous to classical dynamic systems (e.g., a multi-body mass-spring network in a gravitational landscape, or molecular conformation in an electric field, illustrated in Figure 1), the drift $g(\boldsymbol{Z})$ is driven by external potential field (e.g., the macroscopic Waddington landscape, $-\nabla U(\boldsymbol{Z})$). The feedback $\mathbf{B}(\boldsymbol{M}(\boldsymbol{Z}_{it}) - \boldsymbol{Z}_{it})$ serves as an internal restorative force pulling the cell toward a context-specific rest configuration $\boldsymbol{M}(\boldsymbol{Z}_{it})$, mediated by the regulatory-interaction matrix $\mathbf{B}$.*

Remark 2 (**Control-theoretical Lens: Closed-loop Feedback Control**). *Alternatively, Equation 1 implements a continuous-time proportional feedback controller, in contrast to conventional unconstrained Neural ODEs and flows [6, 3] whose open-loop vector fields $\boldsymbol{v}(Z, t)$ have no explicit reference target. Here, $g(\boldsymbol{Z})$ represents the plant's intrinsic dynamics, $\boldsymbol{M}(\boldsymbol{Z}_{it})$ the reference set-point, and $\mathbf{B}$ the controller gain matrix. The error signal $\boldsymbol{M}(\boldsymbol{Z}_{it}) - \boldsymbol{Z}_{it}$ closes the loop, continuously driving the system toward the cell-specific attractor. At the terminal steady state the error vanishes, yielding $\boldsymbol{Z}_{iT} \approx \boldsymbol{M}(\boldsymbol{Z}_{iT})$, enforced by the equilibrium loss $\mathcal{L}_{\text{eq}}$ during training.*

Observations Generation. The latent state $\boldsymbol{Z}$ jointly generates the gene expression profile and the canonical spatial morphology via smooth decoding maps $\phi_x : \mathbb{R}^d \to \mathbb{R}^G$ and $\phi_w : \mathbb{R}^d \to \mathbb{R}^2$. To account for the inherent dimensional imbalance between modalities, we define a weighted joint decoder $\Phi_\gamma(\boldsymbol{Z})$ parameterized by a fixed modality weight $\gamma > 0$,

$$(\boldsymbol{X}, \boldsymbol{S}_{\text{rel}}) = \Phi_\gamma(\boldsymbol{Z}) := \left[\phi_x(\boldsymbol{Z})^\top,\ \gamma\,\phi_w(\boldsymbol{Z})^\top\right]^\top \in \mathbb{R}^{G+2}. \tag{2}$$

To address the challenge of ambiguity in raw coordinates $\boldsymbol{S}$ due to rotation, reflection and translation in cross-sectional slices, we define a canonical relative coordinate via PCA-based normalization $\boldsymbol{S}_{\text{rel}} = (\boldsymbol{S} - \bar{\boldsymbol{S}})\,\mathbf{V}^\top$, where $\bar{\boldsymbol{S}}$ is the spatial centroid of the slice and $\mathbf{V}$ is the PCA rotation matrix. To eliminate reflection ambiguity, the sign of $\mathbf{V}$ is fixed such that the largest-absolute-value coordinate of each principal axis remains positive. This operation yield a cross-time aligned coordinate $\boldsymbol{S}_{\text{rel}}$, which is then passed into a relative position encoder to capture spatial representations [20].

4 Identifiability Theory

Having established the latent dynamic model, a central question is: given access only to the marginal distributions $\{p_t(\boldsymbol{X}, \boldsymbol{S})\}_t$, is the model uniquely determined (i.e., identified)? The short answer is **yes, up to a global similarity transformation**. Informally, any two parametrizations $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ that produce the same observations must be related by a similarity transformation $\tilde{\boldsymbol{Z}} = \lambda \mathbf{Q} \boldsymbol{Z} + \boldsymbol{m}$ for some orthogonal $\mathbf{Q}$, scale $\lambda > 0$, and translation $\boldsymbol{m}$. More crucially, the gene-expression velocity $\boldsymbol{v}_X = J_{\phi_x}(\boldsymbol{Z}) \boldsymbol{v}_Z(\boldsymbol{Z})$ and underlying GRN interactions are **absolutely identifiable** (invariant to $\mathbf{Q}, \lambda, \boldsymbol{m}$), making the whole system biologically interpretable.

4.1 Assumptions & Identifiability Theorem

Now we rigorously establish the identifiability theory. Let $\mathcal{Z} \subset \mathbb{R}^d$. We observe unpaired cross-sectional joint marginals $p_t(\boldsymbol{X}, \boldsymbol{S})$ at times $t \in [0, T]$. The weighted canonical observable is $\Phi_\gamma(\boldsymbol{Z})$ as in Equation (2), with a known modality weight $\gamma > 0$. Consider the following assumptions.

Assumption 1 (Injectivity & Isometry)**.** *Φ_γ is a smooth injective map and **scaled isometric embedding**, $\|\Phi_\gamma(\boldsymbol{Z}_i) - \Phi_\gamma(\boldsymbol{Z}_j)\|^2 \propto \|\boldsymbol{Z}_i - \boldsymbol{Z}_j\|^2$ for any cells i, j. Equivalently, the decoder Jacobian satisfies*

$$J_{\Phi_\gamma}(\boldsymbol{z})^\top J_{\Phi_\gamma}(\boldsymbol{z}) = J_{\phi_x}(\boldsymbol{z})^\top J_{\phi_x}(\boldsymbol{z}) + \gamma^2 J_{\phi_w}(\boldsymbol{z})^\top J_{\phi_w}(\boldsymbol{z}) = \eta^2 \mathbf{I}_d, \tag{3}$$

for some constant global scale $\eta > 0$.

Assumption 1 generalizes the isometric embedding [21, 22] $J_{\phi_x}^\top J_{\phi_x} = \lambda^2 \mathbf{I}_d$ to the multimodal setting with a fixed modality weight γ. Geometrically, it requires the joint decoder to be scaled isometric, ensuring the latent distances faithfully preserve the observed gene-spatial structure.

Assumption 2 (Benamou–Brenier Dynamics)**.** *The latent densities $\rho_t(\boldsymbol{z})$ are absolutely continuous, strictly positive on connected supports, and have finite second moments. The velocity field $\boldsymbol{v}(\boldsymbol{z}, t)$ is the unique minimizer of the Benamou–Brenier functional*

$$\mathcal{A}(\boldsymbol{v}) = \int_0^T \int_{\mathcal{Z}} \|v(\boldsymbol{z}, t)\|^2 \, \rho_t(\boldsymbol{z}) \, d\boldsymbol{z} \, dt, \tag{4}$$

subject to the continuity equation $\partial_t \rho_t + \nabla \cdot (\rho_t \boldsymbol{v}) = 0$.

Assumption 2 is the Benamou–Brenier characterization of dynamic OT [23], selecting the minimum kinetic energy velocity field consistent with the observed marginals, which reflects the biological principle of energetically efficient developmental paths and ensures that the velocity field is identifiable.

Assumption 3 (Regularity)**.** *Functions g and $\boldsymbol{M}$ are smooth, the required Jacobians are well-defined, and $J_{\boldsymbol{v}}$ is ρ-integrable.*

Assumption 4 (Canonical Normalization)**.** *Let ρ denote the latent density pooled over the observed time points. The feedback matrix satisfies*

$$\mathbf{B} = -\,\mathbb{E}_{\boldsymbol{z} \sim \rho}\big[J_{\boldsymbol{v}}(\boldsymbol{z})\big], \qquad \textit{equivalently} \qquad \mathbb{E}_{\boldsymbol{z} \sim \rho}\big[J_g(\boldsymbol{z}) + \mathbf{B}\, J_{\boldsymbol{M}}(\boldsymbol{z})\big] = \mathbf{0}. \tag{5}$$

Assumption 4 fixes the representative of the decomposition in Equation (1). For any $C \in \mathbb{R}^{d \times d}$, the pair $\big(g + C(\boldsymbol{M} - \boldsymbol{z}),\, \mathbf{B} - C\big)$ generates the same velocity field, so the velocity field determines $(g, \mathbf{B})$ up to this family of shifts. Equation (5) selects exactly one member of the family: starting from any parametrization $(g, \mathbf{B}, \boldsymbol{M})$, the shift $C^\star = \mathbb{E}_\rho[J_g + \mathbf{B} J_{\boldsymbol{M}}]$ is the unique one whose pair satisfies the assumption, and the resulting feedback matrix is $\mathbf{B} - C^\star = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$ (Lemma 2 in Appendix A). In this representative $\mathbf{B}$ carries the entire population-average linear part of the dynamics, leaving a residual $J_g + \mathbf{B} J_{\boldsymbol{M}}$ that averages to zero over ρ. Every velocity field of the form (1) therefore has exactly one normalized representative, and it can be computed in closed form from any fitted parametrization. The normalization admits the following interpretations.

- **Mathematical interpretation:** $\mathbf{B} = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$ is a functional of the velocity field $\boldsymbol{v}$ and the density ρ alone. Both are determined by the population marginals $\{p_t(\boldsymbol{X}, \boldsymbol{S})\}_t$, which is the level at which Theorem 1 operates, so the identifiability of $\boldsymbol{v}$ and ρ transfers directly to $\mathbf{B}$. The definition is global: it averages over all observed states and applies uniformly across the multiple attractor basins of $\boldsymbol{M}$.

- **Biological interpretation:** $\mathbf{B}$ is the average linear response of a cell's velocity to a perturbation of its state, taken over the observed cell population. It summarizes the regulation shared by the cells of the tissue across the sampled developmental window, which is the quantity we report as a dataset-level GRN. State-specific and non-linear regulation is carried by g and $\boldsymbol{M}$, and enters the observable GRN through the decoder Jacobian $J_{\phi_x}(\boldsymbol{Z})$.

This leads to our main identifiability theorem. The proof is deferred in Appendix A.1.

Theorem 1 (Identifiability up to Similarity)**.** *Suppose two parameter sets $\boldsymbol{\theta} = \{\phi_x, \phi_w, g, \mathbf{B}, \boldsymbol{M}\}$ and $\tilde{\boldsymbol{\theta}} = \{\tilde{\phi}_x, \tilde{\phi}_w, \tilde{g}, \tilde{\mathbf{B}}, \tilde{\boldsymbol{M}}\}$ with the same known modality weight γ yield identical marginal observation distributions $p_t(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\theta}) = p_t(\boldsymbol{X}, \boldsymbol{S}; \tilde{\boldsymbol{\theta}})$ for all t. Then under Assumptions 1, 2, 3, 4, there exist a time-independent orthogonal matrix $\mathbf{Q} \in O(d)$, a positive scalar $\lambda > 0$, and a vector $\boldsymbol{m} \in \mathbb{R}^d$ such that*

$$\tilde{\boldsymbol{Z}} = \lambda \mathbf{Q} \boldsymbol{Z} + \boldsymbol{m}, \quad \tilde{\mathbf{B}} = \mathbf{Q} \mathbf{B} \mathbf{Q}^\top, \quad \tilde{\Phi}_\gamma(\tilde{\boldsymbol{Z}}) = \Phi_\gamma(\boldsymbol{Z}). \tag{6}$$

*In other words, $\mathbf{B}$ and Φ are identified up to a **global similarity transformation** (a global scaled isometric affine transformation on $\boldsymbol{Z}$).*

Remark 3 (Full model identifiability)**.** *Under additional assumption that $\boldsymbol{M}$ is parameterized as a softmax prototype map with $K_{pro} \geq d + 1$ fixed-point attractors spanning all latent dimensions (Assumption 5 in Appendix A.2), the target map and intrinsic drift are also individually identified: $\tilde{\boldsymbol{M}}(\boldsymbol{z}) = \lambda \mathbf{Q} \boldsymbol{M}(\boldsymbol{z}) + \boldsymbol{m}$ and $\tilde{g}(\boldsymbol{z}) = \lambda \mathbf{Q} g(\boldsymbol{z})$ (Theorem 4 in Appendix A.2).*

Remark 4 (Generalization to cell proliferation and death)**.** *Assumption 2 (strict mass conservation) can be relaxed to more realistic scenario with cell proliferation and apoptosis [2], within the Wasserstein–Fisher–Rao (WFR) or Hellinger-Kantorovich metric framework, also known as unbalanced OT [24, 25]. The continuity equation becomes $\partial_t \rho_t + \nabla \cdot (\rho_t \boldsymbol{v}) = r \rho_t$, where $r(\boldsymbol{z}, t)$ is the cell growth/death rate. The Benamou–Brenier functional is generalized to*

$$\mathcal{A}(v, r) = \int_0^T \int_{\mathcal{Z}} \Big(\|\boldsymbol{v}(\boldsymbol{z}, t)\|^2 + \lambda_{\mathrm{UOT}} |r(z, t)|^2 \Big) \rho_t(\boldsymbol{z}) \, d\boldsymbol{z} \, dt. \tag{7}$$

Crucially, the reaction term $|r|^2$ is zeroth-order in spatial scale while $\|\boldsymbol{v}\|^2$ is second-order, so fixing the WFR hyperparameter λ_{UOT} as a known constant introduces an absolute scale reference for global isometry. The uniqueness of the WFR minimizer [24, 25] preserves the structural identification, guaranteeing that Theorem 1 holds in the presence of cell proliferation and apoptosis. Details are shown in Appendix A.3.

4.2 Biological Interpretability: Gene Regulation Identification

Remarkably, although by theorem 1 there is a residual ambiguity $(\mathbf{Q}, \lambda, \boldsymbol{m})$ in the latent space, it turns out that biological meaningful quantities are **absolutely** identifiable, invariant to this ambiguity.

GRN identifiability. The observable gene-expression velocity is $\boldsymbol{v}_X = J_{\phi_x}(\boldsymbol{Z}) \, \boldsymbol{v}_Z(\boldsymbol{Z})$. Since $\tilde{\phi}_x(\tilde{\boldsymbol{Z}}) = \phi_x(\boldsymbol{Z})$ and $\tilde{\boldsymbol{Z}} = \lambda \mathbf{Q} \boldsymbol{Z} + \boldsymbol{m}$, the chain rule gives $J_{\tilde{\phi}_x}(\tilde{\boldsymbol{Z}}) = J_{\phi_x}(\boldsymbol{Z})(\lambda^{-1} \mathbf{Q}^\top)$. Together with $\tilde{\boldsymbol{v}}_Z(\tilde{\boldsymbol{Z}}) = \lambda \mathbf{Q} \boldsymbol{v}_Z(\boldsymbol{Z})$, we have $\tilde{\boldsymbol{v}}_X = J_{\tilde{\phi}_x}(\tilde{\boldsymbol{Z}}) \, \tilde{\boldsymbol{v}}_Z(\tilde{\boldsymbol{Z}}) = J_{\phi_x}(\boldsymbol{Z})(\lambda^{-1} \mathbf{Q}^\top)(\lambda \mathbf{Q} \boldsymbol{v}_Z(\boldsymbol{Z})) = J_{\phi_x}(\boldsymbol{Z}) \, \boldsymbol{v}_Z(\boldsymbol{Z}) = \boldsymbol{v}_X$. Hence $\boldsymbol{v}_X$ is absolutely identifiable and invariant to $(\mathbf{Q}, \lambda, \boldsymbol{m})$. Since $\boldsymbol{v}_X$ is absolutely identifiable, the underlying gene regulation, i.e., GRN Jacobian $\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}}$, is also identifiable, invariant to latent ambiguity (Proposition 5). Here $\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}}$ is the full $G \times G$ **signed, directed** Jacobian, whose entry $[i, j]$ is the effect of gene j on the rate of change of gene i at a specific time in the developmental trajectory [26]. A positive value indicates that gene j promotes the expression of gene i, while a negative value indicates inhibition. Since $\boldsymbol{v}_X(t)$ is a time-dependent velocity, the inferred GRN is also temporally resolved, capturing the dynamic changes in gene regulation across different developmental stages.

Computational Approximation. By Assumption 4, $\mathbb{E}_\rho\big[\partial \boldsymbol{v}_Z / \partial \boldsymbol{Z}\big] = -\mathbf{B}$. Using this population value and neglecting decoder curvature, the GRN is

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}} \approx J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top, \tag{8}$$

indicating that we can directly evaluate the GRN from $\mathbf{B}$ and decoder Jacobian. By the isometry condition, Equation (8) is equal to $J_{\phi_x}(-\mathbf{B})(\eta^2 \mathbf{I}_d - \gamma^2 J_{\phi_w}^\top J_{\phi_w})^{-1} J_{\phi_x}^\top$. Compared to the vanilla

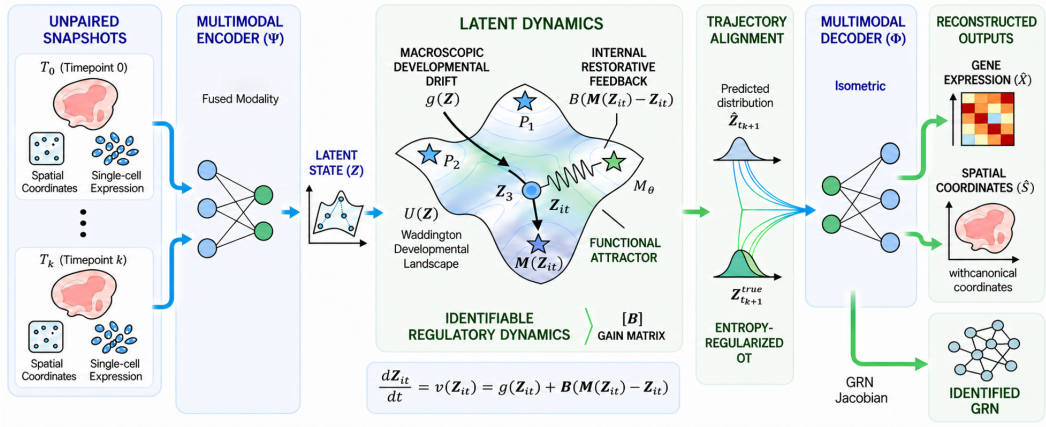


Figure 2: Model Overview. scIFLF uses a multimodal encoder to map expression and spatial information into an isometric latent space,an OT-induced latent flow to operationalize neural ODE and reconstruct underlying regulatory dynamics.

quadratic form $J_{\phi_x}(-\mathbf{B})J_{\phi_x}^\top$, it reveals that the spatial coordinates enter GRN as a **spatially adjusted precision matrix**, dynamically scaled by the isometric constraint η^2 and the inter-modality weight γ^2. In other words, incorporating spatial information not only improves trajectory prediction but also provides a principled correction to the inferred gene regulatory relationships. Details on the derivation are shown in Appendix A.4.

5 Architecture and Training

To operationalize our theoretical framework, we introduce a neural ODE [6] with multimodal encoder (Fig. 2). To ensure scalability and stability, the ODE flow operates within the low-dimensional latent space $\mathcal{Z} \subset \mathbb{R}^d$. Details about architecture and training strategy are deferred to Appendix B.

5.1 Architecture

The framework consists of three primary modules: a multimodal encoder, a continuous-time latent dynamics engine, and dual decoders.

Multimodal Encoder Ψ. The encoder maps the observed gene expression $\boldsymbol{X} \in \mathbb{R}^G$ and PCA-canonical spatial coordinates [20] $\boldsymbol{S}_{\text{rel}} \in \mathbb{R}^2$ into a latent state $\boldsymbol{Z} \in \mathbb{R}^d$. Specifically, we employ a Relative Spatial Encoder to capture local tissue microenvironment contexts. A natural way is to retrieve the k-nearest neighbors and aggregates their features [27, 28], yet graph construction is non-differentiable and sensitive to the neighbourhood radius and to cell density, which varies significantly (from 563 to 121,767 cells in our training data, Table 2) across different slices. Hence, we employ a continuous kernel, where a fixed grid of K anchor or inducing points $\{a_k\}$ [29, 30] is placed uniformly on the PCA-normalized coordinate space. Each cell's spatial context h_s is encoded as a Gaussian-kernel-weighted aggregation, and then fused with extracted gene expression encoding h_g by a per-cell continuous spatial fusion module (concatenation followed by an MLP) to produce the unified latent state. Beyond being an input and a reconstruction target, space is coupled to expression through the joint isometry (Assumption 1) and the precision adjustment in Equation (8).

Latent Dynamics and ODE Integration. The continuous flow of the latent state $\boldsymbol{Z}$ is governed by the parameterized neural ODE $\frac{d\boldsymbol{Z}}{dt} = g_\phi(\boldsymbol{Z}) + \mathbf{B}_\phi(\boldsymbol{M}_\theta(\boldsymbol{Z}) - \boldsymbol{Z})$, where g_ϕ is parameterized by a multi-layer neural network, $\mathbf{B}_\phi$ is a learnable linear map, and M_θ predicts the cell-specific target state. To endow the target predictor with an explicit multi-fate structure, $\boldsymbol{M}_\theta$ is implemented as a prototype attractor model [31], where K_{pro} learnable prototype vectors $\{\boldsymbol{P}_k\}_{k=1}^{K_{\text{pro}}} \subset \mathbb{R}^d$ represent distinct cell fates, and each cell's target is a softmax-weighted combination [32]

$$M_\theta(\boldsymbol{Z}_{it}) = \sum_{k=1}^{K_{\text{pro}}} \frac{\exp(-\|\boldsymbol{Z}_{it} - \boldsymbol{P}_k\|^2/\tau_{\text{pro}})}{\sum_{k'} \exp(-\|\boldsymbol{Z}_{it} - \boldsymbol{P}_{k'}\|^2/\tau_{\text{pro}})} \boldsymbol{P}_k, \tag{9}$$

where $\tau_{\text{pro}} > 0$ is a temperature parameter. It naturally guides different cells toward distinct attractors, and makes the equilibrium loss directly optimizable over the prototype positions.

Dual Decoders Φ_X, Φ_S. The latent representations are projected back to the observational spaces via the Gene Decoder ($\Phi_X : \mathbb{R}^d \to \mathbb{R}^G$) and the Spatial Decoder ($\Phi_S : \mathbb{R}^d \to \mathbb{R}^2$).

5.2 Training Objectives and Strategies

Given the inherently unpaired nature of cross-sectional spatial transcriptomics, we cannot leverage pointwise trajectory supervision. Instead, we rely on OT-induced distributional alignment of the latent flow, with reconstruction supervision on the latent manifold.

Latent Reconstruction Loss. To anchor the spatial and gene modalities and maintain consistency between the latent manifold and observational spaces, we enforce a denoising cyclic encoder-decoder constraint. Denoting by $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I})$ a noise-perturbed latent code, we supervise

$$\mathcal{L}_{\text{recon}} = \sum_{t \in \{t_k\}} \mathbb{E}_{(X,S)\sim p_t} \big[\|\Phi_X(\tilde{\Psi}(\boldsymbol{X}, \boldsymbol{S})) - \boldsymbol{X}\|^2 + w_{\text{recon_s}} \|\Phi_S(\tilde{\Psi}(\boldsymbol{X}, \boldsymbol{S})) - \boldsymbol{S}_{\text{rel}}\|^2 \big], \quad (10)$$

where $\tilde{\boldsymbol{Z}} = \tilde{\Psi}(\boldsymbol{X}, \boldsymbol{S}) = \Psi(\boldsymbol{X}, \boldsymbol{S}) + \boldsymbol{\epsilon}$ denotes the perturbed latent. The noise injection encourages the decoders to learn smoother, more generalizable mappings over the latent manifold.

Trajectory Alignment. To circumvent the lack of paired observations, we align the ODE-predicted latent distribution at t_{k+1} with the directly encoded target distribution using entropy-regularized OT in the latent space $\mathcal{Z}$ within each mini-batch [15], as OT with squared Euclidean cost coincides with the dynamic Benamou–Brenier minimizer [23]

$$\mathcal{L}_{\text{traj}} = \sum_k \min_\pi \mathbb{E}_{(\hat{\boldsymbol{Z}},\, \boldsymbol{Z}^{\text{true}})\sim\pi} \big[\|\hat{\boldsymbol{Z}} - \boldsymbol{Z}^{\text{true}}\|^2 \big] \approx \sum_k \text{OT}_{\varepsilon,\tau}\big(\hat{\boldsymbol{Z}}_{t_{k+1}},\ \boldsymbol{Z}^{\text{true}}_{t_{k+1}}\big), \quad (11)$$

where $\hat{\boldsymbol{Z}}_{t_{k+1}}$ is the ODE-predicted latent state by integrating from $\boldsymbol{Z}_{t_k}$, $\boldsymbol{Z}^{\text{true}}_{t_{k+1}} = \Psi(\boldsymbol{X}_{t_{k+1}}, \boldsymbol{S}_{t_{k+1}})$ is the directly encoded target, and $\text{OT}_{\varepsilon,\tau}$ denotes the Sinkhorn divergence with entropic regularization strength ε and unbalanced relaxation τ. Crucially, we operate latent OT flow in the mini-batch latent space rather than the global high-dimensional observation space, avoiding the inherent difficulty of jointly aligning feature and spatial distributions via computational intensive FGW OT [13, 5].

Isometric Regularization. Isometric regularization is adopted to preserve structural identifiability (Equation 3), i.e., $\mathcal{L}_{\text{iso}} = \mathbb{E}\big[\|J_{\phi_x}^\top J_{\phi_x} + \gamma^2 J_{\phi_w}^\top J_{\phi_w} - \lambda^2 \mathbf{I}_d\|_F^2\big]$. To bypass expensive explicit Jacobian computation, we adopt a randomized finite-difference Hutchinson approximation with $\delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ [33] and penalize the coefficient of variation of the per-sample Rayleigh quotient $r = \big[\|\Phi_\gamma(\boldsymbol{Z}_i + \delta) - \Phi_\gamma(\boldsymbol{Z}_i)\|^2\big] / \|\delta\|^2$ via $\mathcal{L}_{\text{iso}} = \text{Var}_{\boldsymbol{Z},\delta}[r] \,/\, \mathbb{E}_{\boldsymbol{Z},\delta}[r]^2$, which vanishes if and only if the encoder is isometric, and is scale-invariant by construction.

Latent diversity. We also introduce a latent diversity regularization $\mathcal{L}_{\text{div}} = -\text{std}(\boldsymbol{Z}_t)$ to further circumvent latent collapse, allowing cells to evolve toward diverse, cell-specific targets $\boldsymbol{M}_\theta(\boldsymbol{Z})$.

Equilibrium Loss. The equilibrium loss $\mathcal{L}_{\text{eq}} = \mathbb{E}_{\boldsymbol{Z}}[\|\boldsymbol{M}_\theta(\boldsymbol{Z}_T) - \boldsymbol{Z}_T\|^2]$ ensures that the restorative feedback successfully drives the latent state $\boldsymbol{Z}$ to its predicted target configuration.

Hence the final objective admits

$$\mathcal{L} = w_{\text{recon}}\, \mathcal{L}_{\text{recon}} + w_{\text{traj}}\, \mathcal{L}_{\text{traj}} + w_{\text{iso}}\, \mathcal{L}_{\text{iso}} + w_{\text{div}}\, \mathcal{L}_{\text{div}} + w_{\text{eq}}\, \mathcal{L}_{\text{eq}}. \quad (12)$$

Optimization proceeds in a two-phase curriculum. Phase 1 (T_1 epochs) pretrains the latent encoder-decoders to establish a stable structural embedding manifold. Phase 2 activates the ODE integration and full loss landscape, allowing the latent dynamics and the embedding manifold to jointly optimize and mutually refine each other. To prevent catastrophic divergence early before the meaningful target predictor $\boldsymbol{M}_\theta$ has learned, we scale the feedback term $\mathbf{B}_\phi(\boldsymbol{M}_\theta(\boldsymbol{Z}) - \boldsymbol{Z})$ by a scheduling multiplier β, which linearly warms up from 0 to 1 over the first T_2 training epochs of phase 2.

6 Experiments and Results

We would like to emphasize that scIFLF is primarily motivated by **biological interpretability** through identifiability guarantees, with the feedback-controlled latent flow and isometric regularization

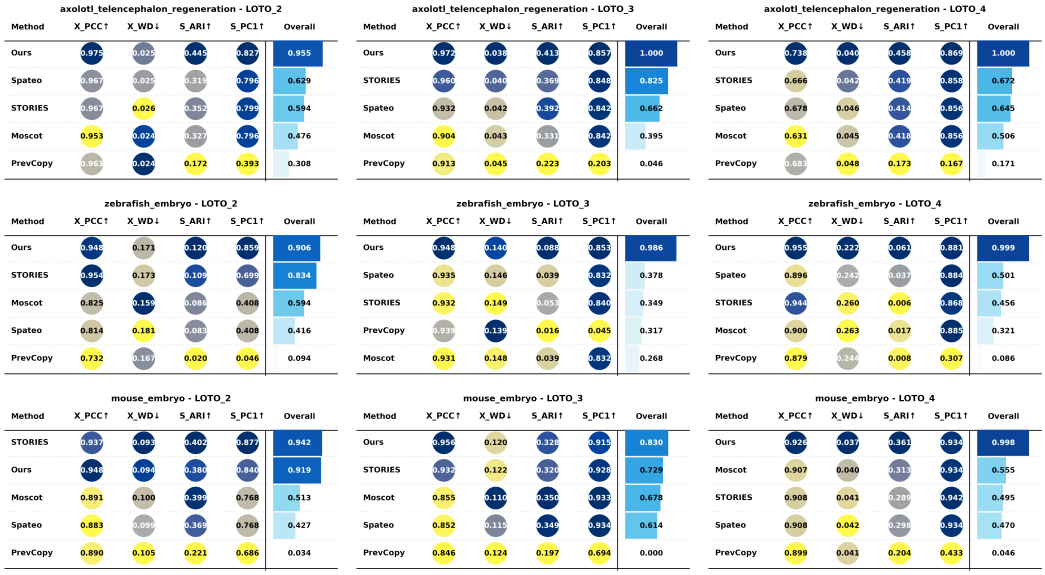


Figure 3: LOTO benchmark metrics across all datasets and held-out time points, with the color shade indicating the normalized value within each LOTO setting. X_PCC stands for gene Pearson correlation, X_WD stands for gene Wasserstein distance, S_PC1 stands for spatial coherence, and S_ARI stands for spatial ARI. Raw metrics are shown in Table 4.

introduced as structural priors enabling principled GRN recovery. To demonstrate that these priors do not compromise predictive fidelity, we first benchmark scIFLF on a leave-one-time-out (LOTO) protocol against state-of-the-art methods, before turning to the biological analysis on the GRN dynamics.

6.1 Leave-one-time-out Benchmark

We evaluate on three real-world spatiotemporal transcriptomics datasets drawn from the STOmics database, including Mouse Embryo (MOSTA) [34], Zebrafish Embryo (ZESTA) [35], and Axolotl Telencephalon Regeneration (ARTISTA) [36], which spans different organisms, tissue types, and biological processes (Appendix C.1). To access the model's ability in interpolation, we adopt a LOTO protocol for evaluation, i.e., slice at an specific intermediate time point t_k is held out and the model is trained on the remaining slices. We evaluate its ability to predict the cellular state at the held-out t_k, with metrics on both feature (Pearson correlation & Wasserstein distance) and spatial alignment (Spatial coherence & ARI), and aggregate min-max normalized metrics [37] into an overall score (Appendix C.3). We compare with several state-of-the-art methods for modeling spatiotemporal transcriptomics, including Moscot [5], Spateo [4] (based on Dynamo [26]), and STORIES [7], as well as a baseline that simply replicates the slice from the nearest past time point (PrevCopy). Since Moscot does not support prediction for unseen time points, we utilize its OT-induced velocity between the two closest time points to perform LOTO interpolation (Appendix C.2).

Fig. 3 summarizes the LOTO prediction results across three datasets and three held-out time points each (with unnormalized raw metrics in Table 4). Our model achieves the highest composite score in 8 out of 9 settings, with consistent advantages on the spatial alignment metrics. The gains are most pronounced on the ARTISTA dataset, which features non-monotonic injury-response dynamics, suggesting that scIFLF generalizes better than pairwise matching under complex temporal dynamics. On mouse embryo, STORIES slightly outperforms scIFLF at the earliest held-out time point but falls behind at later ones, showing it is sufficient under near-linear transcriptional change but less so at held-out points further along the trajectory. PrevCopy performs poorly on spatial metrics across all settings, confirming that non-trivial spatiotemporal prediction requires explicit modeling of dynamics.

In addition, scIFLF extends naturally to time-series scRNA-seq without spatial coordinates by setting modality weight $\gamma \to 0$, where Assumption 1 reduces to the unimodal isometric embedding and

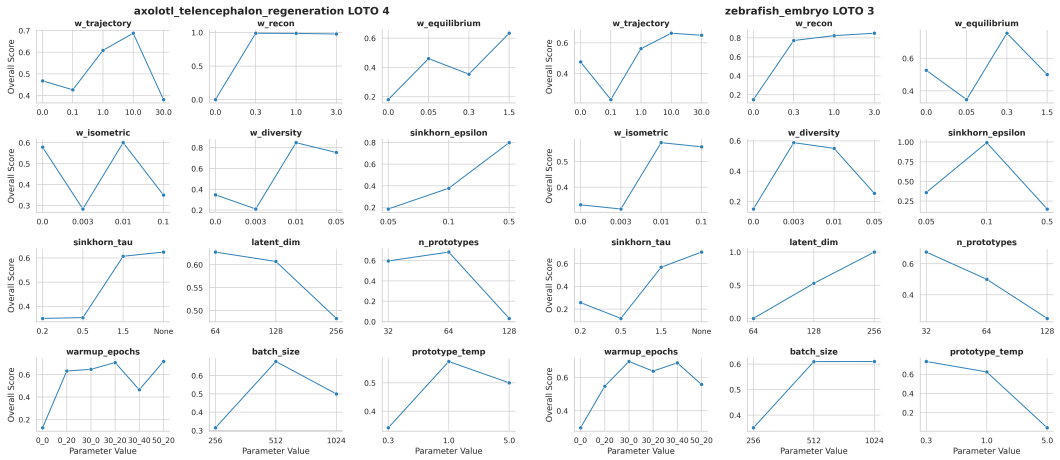


Figure 4: Ablation & hyperparameter sensitivity analysis results. Overall scores are min-max normalized across each hyperparameter group in each LOTO setting and then aggregated. Unnormalized raw metrics are shown in Table 5 and Table 6.

Equation (8) to $J_{\phi_x}(-\mathbf{B})J_{\phi_x}^{\top}/\eta^2$, and Theorems 1 and 4 carry over and the GRN remains identifiable. To verify this on gene-only scRNA data and to quantify the additional gain from spatial information, we ran the gene-only variant under the same LOTO protocol, with TrajectoryNet [3] as an external non-spatial reference. The gene-only variant is comparable to TrajectoryNet, and incorporating spatial information improves the overall performance on all metrics (Appendix D.3). Besides, to better assess spatial prediction ability, we further report two gene-level spatial metrics, the mean gene-wise spatial PCC between predicted and observed expression fields and the agreement of gene-wise Moran's I, which measures how well the spatial autocorrelation of each gene is recovered. Evaluating on the same LOTO settings in ARISTA and ZESTA datasets, scIFLF demonstrates superior performance among the spatial methods (Appendix D.4).

Ablation and Sensitivity Analysis. We conducted extensive ablation and sensitivity analyses on the optimization objectives and key hyperparameters. Fig. 4 reports a one-at-a-time sensitivity analysis on the ZESTA LOTO-3 and ARTISTA LOTO-4 settings. Overall, scIFLF is robust to moderate hyperparameter variation. The most critical hyperparameter is the reconstruction weight w_{recon}: the dramatic performance collapse when ablation of reconstruction loss confirms its critical rule in establishing a stable latent geometry. The trajectory weight w_{traj} improves performance up to moderate values but degrades when set too large, reflecting the need to balance the OT objective against reconstruction. Raw metrics are not sensitive to isometric and equilibrium weights (Table 5, 6), consistent with their role as auxiliary structural constraints. Balanced OT ($\tau = \text{None}$) with a moderate entropic regularization exhibit a quite robust performance across diverse data, though unbalanced marginal relaxation may perform better on the datasets with huge tissue deformation (e.g, ARTISTA LOTO-2, i.e., at time 10hpf, Fig. 9). More results and raw scores are shown in Appendix D, with additional discussion on the role of feedback structure in Appendix E.

6.2 Biological Interpretation

A key advantage of our framework is that the learned latent dynamics admit a theoretically grounded biological interpretation. Specifically, the identifiable gene velocity field $\boldsymbol{v}_X = J_{\phi_x} \cdot \boldsymbol{v}_Z$ can be lifted back to the observational space, and the approximate GRN Jacobian $\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}}$ (Equation 8) governing tissue development at the evaluation time point can be quantified directly from the learned feedback matrix $\mathbf{B}_\phi$ and the decoder Jacobian, without requiring any additional training. We characterize genes by their **outgoing influence** (GRN column norm, i.e., how strongly a gene's expression drives others) and **incoming influence** (GRN row norm, i.e., how strongly a gene's velocity is regulated by the global expression state).

The inferred regulatory landscape of mouse embryogenesis is dominated by three coherent biological programs (Fig. 5): **embryonic and fetal hemoglobins** (*Hba-a1/a2/x*, *Hbb-bs/bt/y*) reflecting the active erythropoietic and hematopoietic program; classical **imprinted growth genes** (*H19*, *Igf2*,

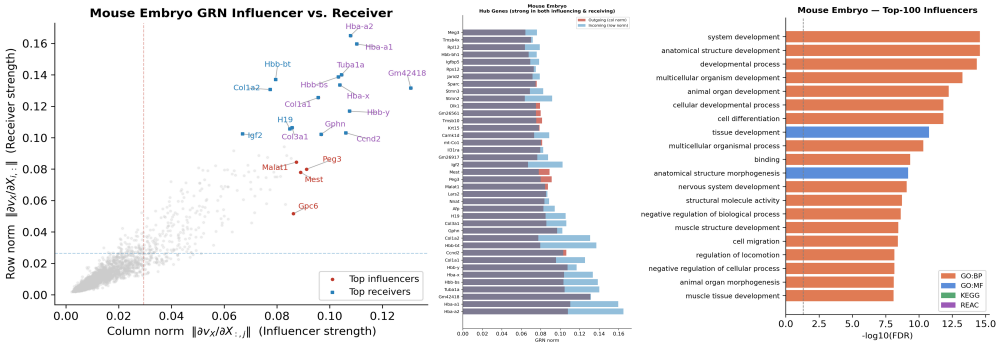


Figure 5: Biological analysis on Mouse Embryo Development. (a) Scatter plot of all genes by outgoing (column norm) vs. incoming (row norm) regulatory strength, with top influencers (red), receivers (blue), and hub genes (purple) labeled. (b) Hub genes ranked by the product of outgoing and incoming norms. (c) GO/KEGG enrichment (FDR < 0.05) for top-100 influencer genes.

Peg3, *Mest*, *Dlk1*, *Meg3*, *Nnat*) central to mammalian patterning and organogenesis; and **ECM/cell-cycle components** (*Col1a1/a2*, *Col3a1*, *Ccnd2*) reflecting the coordinated tissue morphogenesis and proliferation underlying development. The leading hub genes also include *Jarid2* (a Polycomb complex recruiter essential for developmental gene silencing) [38] and *Gphn* (an inhibitory-synapse scaffold central to GABAergic circuit assembly) [39]. GO enrichment of top influencers recovers broad development and morphogenesis terms, while top receivers are additionally enriched for nervous system development, neurogenesis, and neuron differentiation, suggesting that the broad developmental machinery further feeds into neural lineage specification. This data-driven GRN inference aligns with known mammalian organogenesis biology, demonstrating that scIFLF encodes biologically meaningful regulatory structure and yields interpretable mechanistic hypotheses on developmental drivers. Crucially, the inferred GRN is highly robust: scIFLF achieves substantially higher cross-seed Jaccard similarities for top influencer, receiver, and hub genes than the no-feedback ablation, empirically confirming that the proposed latent structure yields a robust, identifiable GRN (Appendix E, Fig. 10). More results for axolotl and zebrafish data, and the temporal-resolved dynamics of GRNs are reported in Appendix F.

7 Discussion

Conclusion. In summary, scIFLF introduces a biophysics-inspired generative framework that successfully bridges the gap between deep latent flow flexibility and biological interpretability. By providing rigorous identifiability guarantees for the feedback-controlled latent flow, scIFLF not only achieves the state-of-the-art spatiotemporal interpolation on complex tissue conformation process, but also enables the principled discovery of gene regulation dynamics. With advances in omics technologies, scIFLF will provide a powerful tool for elucidating the dynamic regulatory mechanisms underlying tissue development and organogenesis.

Limitations and Prospects. Despite its theoretical and empirical advantages, scIFLF still has several limitations. Although the GRN Jacobian $\frac{\partial \boldsymbol{v}_{\boldsymbol{X}}}{\partial \boldsymbol{X}}$ is time- and state-dependent, hence capable to capture the dynamic changes during different developmental stages, it assumes a globally constant regulatory feedback matrix $\mathbf{B}$, which provides a population-level linear baseline but may simplify the highly non-linear, cell-specific, and dynamically rewiring nature of localized interactions over prolonged morphogenesis. State dependence of the effective regulation currently enters only through $\boldsymbol{M}(\boldsymbol{z})$ and $J_{\phi_x}(\boldsymbol{Z})$. Additionally, while the continuous spatial kernel ensures differentiability, it may inadvertently smooth out ultra-fine-grained, single-cell level microenvironmental stochasticity. Additionally, as the rapid advancement of sequencing technology, extending this identifiable flow paradigm to integrate unpaired spatial multi-omics time-series with richer omics modalities [40], could provide a much more comprehensive, multimodal understanding of the epigenetic mechanisms driving these continuous spatiotemporal trajectories. Finally, scIFLF is limited to trajectory reconstruction and the identification of gene regulations within the observed temporal period. How to extrapolate these to make reliable predictions for future time points and perturbations remains an essential topic for further exploration.

Acknowledgments

We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from the MBZUAI-WIS Joint Program and the Al Deira Causal Education project.

References

[1] Longqi Liu, Ao Chen, Yuxiang Li, Jan Mulder, Holger Heyn, and Xun Xu. Spatiotemporal omics for biology and medicine. *Cell*, 187(17):4488–4519, 2024.

[2] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[3] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International Conference on Machine Learning*, pages 9526–9536. PMLR, 2020.

[4] Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu Zuo, Samuel King, et al. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26):7351–7373, 2024.

[5] Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025.

[6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Geert-Jan Huizing, Jules Samaran, Daniele Capocefalo, Anna Audit, Gabriel Peyré, and Laura Cantini. Stories: learning cell fate landscapes from spatial transcriptomics using optimal transport. *Nature Methods*, pages 1–10, 2025.

[8] Kristiyan Sakalyan, Alessandro Palma, Filippo Guerranti, Fabian J Theis, and Stephan Günnemann. Modeling microenvironment trajectories on spatial transcriptomics with nicheflow. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[10] Meichen Fang, Gennady Gorin, and Lior Pachter. Trajectory inference from single-cell genomics data with a process time model. *PLoS computational biology*, 21(1):e1012752, 2025.

[11] Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, 57(4):797–808, 2025.

[12] Ben D MacArthur, Avi Ma'ayan, and Ihor R Lemischka. Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10):672–681, 2009.

[13] Titouan Vayer, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR*, 2023.

[15] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, pages 1–34, 2024.

[16] Floor Eijkelboom, Grigory Bartosh, Christian A Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024.

[17] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

[18] Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

[19] Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

[20] Ruixuan Yu, Xin Wei, Federico Tombari, and Jian Sun. Deep positional and relational feature learning for rotation-invariant point cloud analysis. In *European Conference on Computer Vision*, pages 217–233. Springer, 2020.

[21] Yonghyeon Lee, Sangwoong Yoon, Minjun Son, and Frank C Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022.

[22] Jianle Sun, Chaoqi Liang, Ran Wei, Peng Zheng, LEI BAI, Wanli Ouyang, Hongliang Yan, and Peng Ye. scmrdr: A scalable and flexible framework for unpaired single-cell multi-omics data integration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[23] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[24] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[25] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

[26] Xiaojie Qiu, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, Marco Y Hein, Kyung Hoi Joseph Min, Li Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[27] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, 2022.

[28] Xiang Zhou, Kangning Dong, and Shihua Zhang. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nature Computational Science*, 3(10):894–906, 2023.

[29] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18, 2005.

[30] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[32] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

[33] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[34] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

[35] Chang Liu, Rui Li, Young Li, Xiumei Lin, Kaichen Zhao, Qun Liu, Shuowen Wang, Xueqian Yang, Xuyang Shi, Yuting Ma, et al. Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Developmental Cell*, 57(10):1284–1298, 2022.

[36] Xiaoyu Wei, Sulei Fu, Hanbo Li, Yang Liu, Shuai Wang, Weimin Feng, Yunzhi Yang, Xiawei Liu, Yan-Yun Zeng, Mengnan Cheng, et al. Single-cell stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444, 2022.

[37] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[38] David Landeira, Hakan Bagci, Andrzej R Malinowski, Karen E Brown, Jorge Soza-Ried, Amelie Feytout, Zoe Webster, Elodie Ndjetehe, Irene Cantone, Helena G Asenjo, et al. Jarid2 coordinates nanog expression and pcp/wnt signaling required for efficient esc differentiation and early embryo development. *Cell Reports*, 12(4):573–586, 2015.

[39] Anath C Lionel, Andrea K Vaags, Daisuke Sato, Matthew J Gazzellone, Elyse B Mitchell, Hong Yang Chen, Gregory Costain, Susan Walker, Gerald Egger, Bhooma Thiruvahindrapuram, et al. Rare exonic deletions implicate the synaptic organizer gephyrin (gphn) in risk for autism, schizophrenia and seizures. *Human Molecular Genetics*, 22(10):2055–2066, 2013.

[40] Xuhua Yan, Zhaoyu Fang, Kok Siong Ang, Lynn van Olst, Alex Edwards, Thomas Watson, Ruiqing Zheng, Di Zhang, Rong Fan, Min Li, et al. Mosaic integration of spatial multi-omics with spamosaic. *Nature Genetics*, pages 1–12, 2026.

[41] Vladimir G Ivancevic and Tijana T Ivancevic. *Applied differential geometry: a modern introduction*. World Scientific, 2007.

[42] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[43] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[44] Jiayi Ma, Ji Zhao, Jinwen Tian, Xiang Bai, and Zhuowen Tu. Regularized vector field learning with sparse approximation for mismatch removal. *Pattern Recognition*, 46(12):3519–3532, 2013.

[45] Huimin Cao, Shen Y Heazlewood, Brenda Williams, Daniela Cardozo, Julie Nigro, Ana Oteiza, and Susan K Nilsson. The role of cd44 in fetal and adult hematopoietic stem cell regulation. *Haematologica*, 101(1):26, 2016.

[46] Yui Imaizumi, Shohei Furutachi, Tomoyuki Watanabe, Hiroaki Miya, Daichi Kawaguchi, and Yukiko Gotoh. Role of the imprinted allele of the cdkn1c gene in mouse neocortical development. *Scientific Reports*, 10(1):1884, 2020.

[47] Fang Ye, Guodong Zhang, Weigao E, Haide Chen, Chengxuan Yu, Lei Yang, Yuting Fu, Jiaqi Li, Sulei Fu, Zhongyi Sun, et al. Construction of the axolotl cell landscape using combinatorial hybridization sequencing at single-cell resolution. *Nature Communications*, 13(1):4228, 2022.

[48] Lauren M Saunders, Sanjay R Srivatsan, Madeleine Duran, Michael W Dorrity, Brent Ewing, Tor H Linbo, Jay Shendure, David W Raible, Cecilia B Moens, David Kimelman, et al. Embryo-scale reverse genetics at single-cell resolution. *Nature*, 623(7988):782–791, 2023.

[49] James Palis. Primitive and definitive erythropoiesis in mammals. *Frontiers in Physiology*, 5:3, 2014.

Appendix

A Details on Identifiability Theory

We present the complete proof of the identifiability theorem. We first invoke the basic assumptions and the main identifiability results given in Section 4.

A1. Injectivity & Isometry. Decoder Φ_γ is a smooth injective map that defines a diffeomorphism between $\mathcal{Z}$ and its image $\mathcal{M}_\gamma := \Phi_\gamma(\mathcal{Z})$, and a **joint scaled isometric embedding** such that

$$J_{\Phi_\gamma}(\boldsymbol{z})^\top J_{\Phi_\gamma}(\boldsymbol{z}) = J_{\phi_x}(\boldsymbol{z})^\top J_{\phi_x}(\boldsymbol{z}) + \gamma^2 J_{\phi_w}(\boldsymbol{z})^\top J_{\phi_w}(\boldsymbol{z}) = \eta^2 \mathbf{I}_d,$$

for some constant global scale $\eta > 0$.

A2. Benamou–Brenier Dynamics. The latent densities $\rho_t(\boldsymbol{z})$ are absolutely continuous, strictly positive on connected supports, and have finite second moments. The velocity field $\boldsymbol{v}(\boldsymbol{z}, t)$ is the unique minimizer of the Benamou–Brenier functional

$$\mathcal{A}(\boldsymbol{v}) = \int_0^T \int_{\mathcal{Z}} \|\boldsymbol{v}(\boldsymbol{z}, t)\|^2 \ \rho_t(\boldsymbol{z})\, d\boldsymbol{z}\, dt,$$

subject to the continuity equation $\partial_t \rho_t + \nabla \cdot (\rho_t \boldsymbol{v}) = 0$.

A3. Regularity. The functions g and $\boldsymbol{M}$, and the decoders ϕ_x, ϕ_w are smooth, and $J_{\boldsymbol{v}}$ is ρ-integrable.

A4. Canonical Normalization. With ρ the latent density pooled over the observed time points, the feedback matrix satisfies

$$\mathbf{B} = -\,\mathbb{E}_{\boldsymbol{z}\sim\rho}\big[J_{\boldsymbol{v}}(\boldsymbol{z})\big], \qquad \text{equivalently} \qquad \mathbb{E}_{\boldsymbol{z}\sim\rho}\big[J_g(\boldsymbol{z}) + \mathbf{B}\, J_{\boldsymbol{M}}(\boldsymbol{z})\big] = \mathbf{0}.$$

Every velocity field admits exactly one such representative:

Lemma 2 (Normalized representative)**.** *Let $\boldsymbol{v} = g + \mathbf{B}(\boldsymbol{M} - \boldsymbol{z})$ with $J_{\boldsymbol{v}}$ ρ-integrable. Among the pairs $\big(g + C(\boldsymbol{M} - \boldsymbol{z}),\ \mathbf{B} - C\big)$, $C \in \mathbb{R}^{d\times d}$, exactly one satisfies A4, namely $C^\star = \mathbb{E}_\rho\big[J_g + \mathbf{B} J_{\boldsymbol{M}}\big]$, for which $\mathbf{B} - C^\star = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$.*

Proof. Write $g_C := g + C(\boldsymbol{M} - \boldsymbol{z})$ and $\mathbf{B}_C := \mathbf{B} - C$, so that $J_{g_C} = J_g + C(J_{\boldsymbol{M}} - \mathbf{I}_d)$. Then

$$\mathbb{E}_\rho\big[J_{g_C} + \mathbf{B}_C J_{\boldsymbol{M}}\big] = \mathbb{E}_\rho\big[J_g + \mathbf{B} J_{\boldsymbol{M}}\big] - C,$$

which vanishes if and only if $C = C^\star$. Using $J_{\boldsymbol{v}} = J_g + \mathbf{B}(J_{\boldsymbol{M}} - \mathbf{I}_d)$ gives $\mathbf{B}_{C^\star} = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$. □

Under Assumptions 1, 2, 3, 4, suppose there exist two sets of generative model parameters $\boldsymbol{\theta} = \{\phi_x, \phi_w, g, \mathbf{B}, \boldsymbol{M}\}$ and $\tilde{\boldsymbol{\theta}} = \{\tilde{\phi}_x, \tilde{\phi}_w, \tilde{g}, \tilde{\mathbf{B}}, \tilde{\boldsymbol{M}}\}$ with the same known modality weight γ that yield identical marginal observation distributions: $\forall t, p_t(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\theta}) = p_t(\boldsymbol{X}, \boldsymbol{S}; \tilde{\boldsymbol{\theta}})$. Then there exists a time-independent orthogonal matrix $\mathbf{Q}$, a positive scalar λ, and a vector $\boldsymbol{m}$ such that

$$\tilde{\boldsymbol{z}} = \lambda \mathbf{Q}\boldsymbol{z} + \boldsymbol{m}, \quad \tilde{\mathbf{B}} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top, \quad \tilde{\Phi}_\gamma(\tilde{\boldsymbol{Z}}) = \Phi_\gamma(\boldsymbol{Z}).$$

Hence, $\mathbf{B}$ and Φ are identified up to a global similarity transformation (a scaled isometric affine transformation on $\boldsymbol{z}$).

A.1 Proof of Main Identifiability Theorem 1

We first state the following lemma on Euclidean conformal rigidity, which is a standard conclusion from Liouville-type theorem (Cartan-Killing rigidity of Euclidean similarities) [41].

Lemma 3 (Euclidean conformal rigidity)**.** *Let $\mathcal{Z} \subset \mathbb{R}^d$ be a connected open set with $d \geq 2$, and let $T : \mathcal{Z} \to \mathbb{R}^d$ be a C^2 map satisfying $J_T(\boldsymbol{z})^\top J_T(\boldsymbol{z}) = c^2 \mathbf{I}_d$ for all $\boldsymbol{z} \in \mathcal{Z}$ and some constant $c > 0$. Then $T(\boldsymbol{z}) = c\,\mathbf{Q}\boldsymbol{z} + \boldsymbol{m}$ for some $\mathbf{Q} \in O(d)$ and $\boldsymbol{m} \in \mathbb{R}^d$.*

Proof. Differentiating $J_T^\top J_T = c^2 \mathbf{I}_d$ along any direction gives $(\partial_k J_T)^\top J_T + J_T^\top (\partial_k J_T) = 0$. Combined with the symmetry of mixed partial derivatives $\partial_k J_T^{ij} = \partial_j J_T^{ik}$ (Hessian symmetry of T), a standard cyclic-index manipulation yields $\partial_k J_T \equiv 0$ on $\mathcal{Z}$. Hence J_T is constant; writing $J_T = c\mathbf{Q}$ and using $c^2 \mathbf{Q}^\top \mathbf{Q} = c^2 \mathbf{I}_d$ gives $\mathbf{Q} \in O(d)$. Integrating recovers $T(\boldsymbol{z}) = c\mathbf{Q}\boldsymbol{z} + \boldsymbol{m}$. □

Now we prove the main identifiability theorem.

Proof

Step 1: Observable Manifold Alignment, Temporal Coherence, and Global Isometry Define the weighted observable mapping:

$$\Phi_\gamma(\boldsymbol{z}) := [\phi_x(\boldsymbol{z})^T, \gamma\phi_w(\boldsymbol{z})^T]^T.$$

Let $\mathcal{M}_\gamma := \Phi_\gamma(\mathcal{Z})$ be the image. Under A1, Φ_γ is a smooth embedding and hence a diffeomorphism between $\mathcal{Z}$ and $\mathcal{M}_\gamma$. Because $p_t = \tilde{p}_t$, the supports coincide $\mathcal{M}_\gamma = \tilde{\mathcal{M}}_\gamma$. Thus we define a coordinate transformation on the latent space

$$T := \tilde{\Phi}_\gamma^{-1} \circ \Phi_\gamma : \mathcal{Z} \to \mathcal{Z},$$

which is a well-defined smooth bijective map.

The pullback metric satisfies

$$J_{\Phi_\gamma}^T J_{\Phi_\gamma} = J_{\phi_x}^T J_{\phi_x} + \gamma^2 J_{\phi_w}^T J_{\phi_w} = \eta^2 \mathbf{I}_d.$$

The two parametrizations $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$ each satisfy A1 with their respective global scales $\eta, \tilde{\eta} > 0$. By the chain rule,

$$\begin{aligned} J_T(\boldsymbol{z})^\top J_T(\boldsymbol{z}) &= J_{\Phi_\gamma}(\boldsymbol{z})^\top \left[J_{\tilde{\Phi}_\gamma}(T(\boldsymbol{z}))^\top J_{\tilde{\Phi}_\gamma}(T(\boldsymbol{z}))\right]^{-1} J_{\Phi_\gamma}(\boldsymbol{z}) \\ &= J_{\Phi_\gamma}(\boldsymbol{z})^\top \left(\tilde{\eta}^{-2}\mathbf{I}_d\right) J_{\Phi_\gamma}(\boldsymbol{z}) = \frac{\eta^2}{\tilde{\eta}^2}\,\mathbf{I}_d. \end{aligned}$$

Define $\lambda := \eta/\tilde{\eta} > 0$, so $J_T^\top J_T \equiv \lambda^2 \mathbf{I}_d$ on the connected domain $\mathcal{Z}$. By Lemma 3,

$$T(\boldsymbol{z}) = \lambda \mathbf{Q}\boldsymbol{z} + \boldsymbol{m}, \quad \mathbf{Q} \in O(d),\ \boldsymbol{m} \in \mathbb{R}^d.$$

Since this holds globally on a connected domain $\mathcal{Z}$, by Liouville-type theorem (Cartan-Killing rigidity of Euclidean similarities) [41], we have

$$T(\boldsymbol{z}) = \lambda \mathbf{Q}\boldsymbol{z} + \boldsymbol{m}, \quad \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_d, \quad \lambda > 0,$$

for some time-independent global scalar constant λ, orthogonal matrix $\mathbf{Q}$, and vector $\boldsymbol{m}$. Note that λ is determined by the ratio $\eta/\tilde{\eta}$ rather than fixed a priori: the residual scale ambiguity arises because A1 does not pin η to a specific value across equivalent parametrizations.

Step 2: Velocity Field Equivariance Under the change of variables $\tilde{\boldsymbol{z}} = \lambda \mathbf{Q}\boldsymbol{z} + \boldsymbol{m}$, the latent density transforms as $\tilde{\rho}_t(\tilde{\boldsymbol{z}}) = \lambda^{-d}\rho_t(\boldsymbol{z})$. Define the candidate velocity field

$$\hat{\boldsymbol{v}}(\tilde{\boldsymbol{z}}, t) := \nabla T \cdot \boldsymbol{v}(\boldsymbol{z}, t) = \lambda \mathbf{Q}\boldsymbol{v}(\boldsymbol{z}, t).$$

This induces the same density evolution under change of variables $\tilde{\rho}_t(\tilde{\boldsymbol{z}}) = \rho_t(\boldsymbol{z})|\det(\lambda\mathbf{Q})|^{-1} = \lambda^{-d}\rho_t(\boldsymbol{z})$, and hence one can verify that $(\hat{\boldsymbol{v}}, \tilde{\rho}_t)$ satisfies the continuity equation $\partial_t\tilde{\rho}_t + \nabla_{\tilde{\boldsymbol{z}}} \cdot (\tilde{\rho}_t \hat{\boldsymbol{v}}) = 0$. Moreover we have $\|\lambda\mathbf{Q}\boldsymbol{v}\|^2 = \lambda^2\|\boldsymbol{v}\|^2$ and hence

$$\mathcal{A}(\hat{\boldsymbol{v}}) = \int \|\lambda\mathbf{Q}\boldsymbol{v}(\boldsymbol{z})\|^2 \lambda^{-d}\rho_t(\boldsymbol{z})\lambda^d d\boldsymbol{z}dt = \int \lambda^2\|\boldsymbol{v}(\boldsymbol{z})\|^2\rho_t(\boldsymbol{z})d\boldsymbol{z}dt = \lambda^2\mathcal{A}(\boldsymbol{v}).$$

By the standard characterization of the Benamou–Brenier minimizer [42], under regularity conditions including absolute continuity and finite second moments, the minimum-kinetic-energy velocity field consistent with $\{\rho_t\}$ is ρ_t-a.e. unique and belongs to the $L^2(\rho_t)$-closure of gradient fields [42]. Scaling the functional by a constant λ^2 preserves its unique Benamou–Brenier minimizer, $\hat{\boldsymbol{v}}$ is the Benamou–Brenier minimizer associated with the marginals $\{\tilde{\rho}_t\}$ if and only if $\boldsymbol{v}$ is the Benamou–Brenier minimizer associated with $\{\rho_t\}$.

By A2, both parametrizations satisfy the Benamou–Brenier minimality condition, and the Benamou–Brenier minimizer is *unique* in $L^2(\tilde{\rho}_t\, dt)$ ([42]). Both $\tilde{\boldsymbol{v}}$ (the velocity field of $\tilde{\boldsymbol{\theta}}$) and $\hat{\boldsymbol{v}}$ are Benamou–Brenier minimizers for $\{\tilde{\rho}_t\}$, and therefore must coincide $\tilde{\rho}_t$-a.e., i.e.,

$$\tilde{\boldsymbol{v}}(\lambda\mathbf{Q}\boldsymbol{z} + \boldsymbol{m}, t) = \lambda\mathbf{Q}\boldsymbol{v}(\boldsymbol{z}, t), \quad \rho_t\text{-a.e..}, \forall t. \tag{13}$$

Step 3: Identification of $\mathbf{B}$ Differentiating the velocity equivariance (13) with respect to $\boldsymbol{z}$ gives

$$J_{\tilde{\boldsymbol{v}}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m})\,\lambda\mathbf{Q} = \lambda\mathbf{Q}\,J_{\boldsymbol{v}}(\boldsymbol{z}),$$

and cancelling λ yields the pointwise relation

$$J_{\tilde{\boldsymbol{v}}}(\tilde{\boldsymbol{z}}) = \mathbf{Q}\,J_{\boldsymbol{v}}(\boldsymbol{z})\,\mathbf{Q}^\top, \qquad \tilde{\boldsymbol{z}} = \lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}. \tag{14}$$

The change of variables $\boldsymbol{z} \mapsto \lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}$ pushes ρ forward to $\tilde{\rho}$, so under A3 the relation survives integration:

$$\mathbb{E}_{\tilde{\rho}}\big[J_{\tilde{\boldsymbol{v}}}\big] = \mathbb{E}_{\rho}\big[J_{\tilde{\boldsymbol{v}}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m})\big] = \mathbf{Q}\,\mathbb{E}_{\rho}\big[J_{\boldsymbol{v}}(\boldsymbol{z})\big]\,\mathbf{Q}^\top.$$

By A4, $\mathbf{B} = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$ and $\tilde{\mathbf{B}} = -\mathbb{E}_{\tilde{\rho}}[J_{\tilde{\boldsymbol{v}}}]$, hence

$$\tilde{\mathbf{B}} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top.$$

This completes the proof of Theorem 1. □

A.2 Full Model Identification under Additional Attractor Assumption

Beyond Theorem 1, since both $\boldsymbol{M}$ and g are flexible non-linear functions, identifying them individually requires additional structural and parametric assumptions on the target map. As an example, we establish the full model identification theory with softmax attractor parametrization, and formally articulate the structural assumption now. It is also worth noting that the full identification of $\boldsymbol{M}$ and g turns out to be not necessary, since we can guarantee the identifiability of GRN dynamics $\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}}$ only based on theorem 1.

Assumption 5 (Attractor Structure)**.** *The target map $\boldsymbol{M} : \mathcal{Z} \to \mathcal{Z}$ possesses $K_{pro} \geq d+1$ distinct fixed-point attractors $\{P_k\}_{k=0}^{K_{pro}-1} \subset \mathcal{Z}$ satisfying:* (i) $\boldsymbol{M}(P_k) = P_k$*;* (ii) $g(P_k) = \mathbf{0}$*;* (iii) $\{P_k - P_0\}_{k=1}^{d}$ *are linearly independent. Furthermore, $\boldsymbol{M}$ is parameterized as $\boldsymbol{M}(\boldsymbol{z}) = \sum_k \alpha_k(\boldsymbol{z})P_k$ with $\alpha_k(\boldsymbol{z}) = \mathrm{softmax}(-\|\boldsymbol{z}-P_k\|^2/\tau_{pro})_k$, so that $\{P_k\}$ are its unique attractors corresponding to K_{pro} distinct cell fates.*

Theorem 4 (Full Model Identifiability)**.** *Under Assumptions 1, 2, 3, 4, and 5, there exist a permutation $\sigma \in S_{K_{pro}}$ such that the prototype attractors satisfy $\tilde{P}_{\sigma(k)} = \lambda\mathbf{Q}P_k + \boldsymbol{m}$ (with the temperature reparameterized as $\tilde{\tau}_{pro} = \lambda^2\tau_{pro}$), and consequently*

$$\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}\boldsymbol{M}(\boldsymbol{z})+\boldsymbol{m}, \qquad \tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}g(\boldsymbol{z}).$$

Proof. Without loss of generality we relabel the prototypes $\{\tilde{P}_k\}$ via the permutation σ so that $\tilde{P}_{\sigma(k)}$ corresponds to P_k; the prototype attractor model is invariant under permutation of $\{P_k\}$ by construction.

We continue from Theorem 1: $\tilde{\boldsymbol{v}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}\boldsymbol{v}(\boldsymbol{z})$ and $\tilde{\mathbf{B}} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top$. By substituting the specific forms of the velocity field into both sides of the equivariance equation, we have

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \tilde{\mathbf{B}}\big(\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) - (\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m})\big) = \lambda\mathbf{Q}\big(g(\boldsymbol{z}) + \mathbf{B}(\boldsymbol{M}(\boldsymbol{z})-\boldsymbol{z})\big).$$

Expanding the feedback terms on both sides yields

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \tilde{\mathbf{B}}\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) - \tilde{\mathbf{B}}\lambda\mathbf{Q}\boldsymbol{z} - \tilde{\mathbf{B}}\boldsymbol{m} = \lambda\mathbf{Q}g(\boldsymbol{z}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}) - \lambda\mathbf{Q}\mathbf{B}\boldsymbol{z}.$$

Using the relation $\tilde{\mathbf{B}} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top$, we have $\tilde{\mathbf{B}}\mathbf{Q} = \mathbf{Q}\mathbf{B}$, which implies $\tilde{\mathbf{B}}\lambda\mathbf{Q} = \lambda\mathbf{Q}\mathbf{B}$. Therefore, the term $-\tilde{\mathbf{B}}\lambda\mathbf{Q}\boldsymbol{z}$ on the left-hand side exactly cancels the term $-\lambda\mathbf{Q}\mathbf{B}\boldsymbol{z}$ on the right-hand side. This simplifies the equation to

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \tilde{\mathbf{B}}\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) - \tilde{\mathbf{B}}\boldsymbol{m} = \lambda\mathbf{Q}g(\boldsymbol{z}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}). \tag{15}$$

Identification of $\boldsymbol{M}$. At each attractor P_k (A5: $g(P_k) = \mathbf{0}$, $\boldsymbol{M}(P_k) = P_k$), the velocity equivariance gives $\tilde{\boldsymbol{v}}(\tilde{P}_k) = \mathbf{0}$ where $\tilde{P}_k := \lambda\mathbf{Q}P_k + \boldsymbol{m}$. Hence $\tilde{P}_k$ is a fixed-point attractor of $\tilde{\boldsymbol{M}}$ with $\tilde{g}(\tilde{P}_k) = \mathbf{0}$.

Since $\boldsymbol{M}$ (resp. $\tilde{\boldsymbol{M}}$) is parameterized by softmax prototypes and its attractors are exactly $\{P_k\}$ (resp. $\{\tilde{P}_k\}$), and since $\|\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}-\tilde{P}_k\|^2 = \lambda^2\|\boldsymbol{z}-P_k\|^2$ implies $\tilde{\alpha}_k(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \alpha_k(\boldsymbol{z})$ (up to temperature reparameterization $\tau_{\text{pro}} \mapsto \lambda^2\tau_{\text{pro}}$), we obtain

$$\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \sum_k \alpha_k(\boldsymbol{z})\,\tilde{P}_k = \sum_k \alpha_k(\boldsymbol{z})(\lambda\mathbf{Q}P_k+\boldsymbol{m}) = \lambda\mathbf{Q}\boldsymbol{M}(\boldsymbol{z})+\boldsymbol{m}.$$

Identification of g. Substituting the identified target map $\tilde{\boldsymbol{M}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}\boldsymbol{M}(\boldsymbol{z})+\boldsymbol{m}$ into Eq. 15, we expand the left-hand side

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \tilde{\mathbf{B}}\big(\lambda\mathbf{Q}\boldsymbol{M}(\boldsymbol{z})+\boldsymbol{m}\big) - \tilde{\mathbf{B}}\boldsymbol{m} = \lambda\mathbf{Q}g(\boldsymbol{z}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}).$$

Notice that the translation term $+\tilde{\mathbf{B}}\boldsymbol{m}$ exactly cancels with $-\tilde{\mathbf{B}}\boldsymbol{m}$, reducing the equation to

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \tilde{\mathbf{B}}\lambda\mathbf{Q}\boldsymbol{M}(\boldsymbol{z}) = \lambda\mathbf{Q}g(\boldsymbol{z}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}).$$

Applying $\tilde{\mathbf{B}}\lambda\mathbf{Q} = \lambda\mathbf{Q}\mathbf{B}$ once again, the second term on the left-hand side becomes $\lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z})$:

$$\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}) = \lambda\mathbf{Q}g(\boldsymbol{z}) + \lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z}).$$

Canceling $\lambda\mathbf{Q}\mathbf{B}\boldsymbol{M}(\boldsymbol{z})$ from both sides naturally yields $\tilde{g}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}g(\boldsymbol{z})$. □

A.3 Relaxation to Unbalanced OT

The strict mass conservation in A2 can be relaxed to accommodate cell proliferation and apoptosis within the Wasserstein-Fisher-Rao (WFR) framework [24, 25], without compromising identifiability (Remark 4). We formally establish the argument now.

Suppose we extend the latent dynamics to the Unbalanced OT (UOT) framework equipped with the WFR metric. Incorporating a scalar growth/death rate $r(\boldsymbol{z},t)$, the unbalanced continuity equation becomes $\partial_t\rho_t + \nabla\cdot(\rho_t\boldsymbol{v}) = r\rho_t$, and the generalized Benamou–Brenier functional is

$$\mathcal{A}_{\lambda_{\mathrm{UOT}}}(\boldsymbol{v},r) = \int_0^T \int_{\mathcal{Z}} \Big(\|\boldsymbol{v}\|^2 + \lambda_{\mathrm{UOT}}|r|^2\Big)\rho_t\, d\boldsymbol{z}\, dt, \tag{16}$$

where $\lambda_{\mathrm{UOT}} > 0$ balances transport and reaction costs. Under the change of variables to momentum $m = \rho\boldsymbol{v}$ and source $\mu = \rho r$, $\mathcal{A}$ becomes strictly convex, and subject to regular boundary marginals the minimizing pair $(\boldsymbol{v}^*, r^*)$ is unique [24].

Scale consistency under similarity transformation. Under the similarity $\tilde{\boldsymbol{z}} = \lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}$, the density transforms as $\tilde{\rho}_t(\tilde{\boldsymbol{z}}) = \lambda^{-d}\rho_t(\boldsymbol{z})$. The reaction rate r is intrinsically defined as the relative growth rate $r(\boldsymbol{z},t) = \partial_t \log\rho_t(\boldsymbol{z})$ along Lagrangian trajectories, hence transforms as a zeroth-order scalar: $\hat{r}(\tilde{\boldsymbol{z}},t) := r(\boldsymbol{z},t)$. The velocity, being a first-order tensor, pulls back as $\hat{\boldsymbol{v}}(\tilde{\boldsymbol{z}},t) := \lambda\mathbf{Q}\,\boldsymbol{v}(\boldsymbol{z},t)$. A direct calculation verifies that $(\hat{\boldsymbol{v}},\hat{r},\tilde{\rho}_t)$ satisfies the unbalanced continuity equation in the new coordinates

$$\partial_t\tilde{\rho}_t = \lambda^{-d}\big[-\nabla_{\boldsymbol{z}}\cdot(\rho_t\boldsymbol{v}) + r\rho_t\big], \qquad \nabla_{\tilde{\boldsymbol{z}}}\cdot(\tilde{\rho}_t\hat{\boldsymbol{v}}) = \lambda^{-d}\nabla_{\boldsymbol{z}}\cdot(\rho_t\boldsymbol{v}),$$

giving $\partial_t\tilde{\rho}_t + \nabla_{\tilde{\boldsymbol{z}}}\cdot(\tilde{\rho}_t\hat{\boldsymbol{v}}) = \lambda^{-d}r\rho_t = \hat{r}\,\tilde{\rho}_t$.

However, evaluating the functional gives

$$\mathcal{A}_{\tilde{\lambda}_{\mathrm{UOT}}}(\hat{\boldsymbol{v}},\hat{r}) = \iint \Big(\lambda^2\|\boldsymbol{v}\|^2 + \tilde{\lambda}_{\mathrm{UOT}}|r|^2\Big)\rho_t\, d\boldsymbol{z}\, dt. \tag{17}$$

Under this condition, $\mathcal{A}_{\tilde{\lambda}}(\hat{\boldsymbol{v}},\hat{r}) = \lambda^2\,\mathcal{A}_{\lambda_{\mathrm{UOT}}}(\boldsymbol{v},r)$, and the uniqueness of the WFR minimizer transfers to give $\tilde{\boldsymbol{v}}(\lambda\mathbf{Q}\boldsymbol{z}+\boldsymbol{m}) = \lambda\mathbf{Q}\boldsymbol{v}(\boldsymbol{z})$ ρ_t-a.e.

Identifiability is preserved. We now show that the residual scale ambiguity λ is removed under the WFR framework. Consider two equivalent parametrizations $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$ producing identical observation marginals, both modeled with the *same* WFR action functional with shared hyperparameter λ_{UOT}. For the Benamou—Brenier-equivalence in Step 2 of Theorem 1's proof to extend, we require $\hat{\boldsymbol{v}}$ to be the minimizer of the *same* functional $\mathcal{A}_{\lambda_{\mathrm{UOT}}}$, not a rescaled version. The homogeneity calculation gives

$$\mathcal{A}_{\lambda_{\mathrm{UOT}}}(\hat{\boldsymbol{v}},\hat{r}) = \int_0^T \int_{\mathcal{Z}} \Big(\lambda^2\|\boldsymbol{v}\|^2 + \lambda_{\mathrm{UOT}}|r|^2\Big)\rho_t\, d\boldsymbol{z}\, dt. \tag{18}$$

For this to equal $\lambda^2\,\mathcal{A}_{\lambda_{\mathrm{UOT}}}(\boldsymbol{v},r)$ (i.e., a uniform rescaling that preserves minimality), the reaction term must scale as $\lambda^2|r|^2$, which is incompatible with the zeroth-order $|r|^2$ unless $\lambda = 1$. In practice, λ_{UOT} is a fixed hyperparameter of the model. Under fixed λ_{UOT}, the only consistent scale is $\lambda = 1$: the second-order kinetic term and the zeroth-order reaction term cannot co-scale under a nontrivial

similarity, with λ_{UOT} serving as an absolute scale anchor analogous to fixing η in A1. We emphasize that this is not a strengthening of identifiability through weakening assumptions, but rather the consequence of the WFR functional carrying an external scale reference. If λ_{UOT} were instead allowed to vary across parametrizations, the ambiguity would become $(\lambda, \mathbf{Q}, \boldsymbol{m})$ together with a coupled rescaling of λ_{UOT} by λ^2, recovering the balanced ambiguity. Combined with the uniqueness of the WFR minimizer [24, 25], the structural decoupling of $\{\mathbf{B}, \boldsymbol{M}, g\}$ in Step 3 carries over to the unbalanced setting.

A.4 Derivation of the Gene Regulatory Network Jacobian

This section provides the detailed derivation for evaluating the gene-regulatory Jacobian $\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}}$ presented in Section 4.2 from the learned model parameters. We emphasize that $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$ is absolutely identifiable as a relation between observable quantities (since $\boldsymbol{v}_X$ is invariant to the latent ambiguity and $\boldsymbol{X}$ is observed). The question addressed here is how to express it in closed form using the learned $\mathbf{B}$ and decoder Jacobians, and which approximations this involves. We first summarize the main findings as Proposition 5.

Proposition 5 (Biological invariants)**.** *Under the conditions of Theorem 1:*

(i) The gene-expression velocity $\boldsymbol{v}_X = J_{\phi_x}(\boldsymbol{Z})\, \boldsymbol{v}_Z(\boldsymbol{Z})$ is absolutely identifiable, i.e., $\tilde{\boldsymbol{v}}_X = \boldsymbol{v}_X$, invariant to latent ambiguity $(\mathbf{Q}, \lambda, \boldsymbol{m})$.

(ii) The gene-regulatory Jacobian $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$ is also absolutely identifiable, invariant to latent ambiguity.

(iii) Equation (8)*, i.e.*

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}} \approx J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top,$$

gives a closed-form expression in terms of the learned $\mathbf{B}$ and decoder Jacobian, obtained by neglecting decoder curvature and replacing the pointwise latent Jacobian by its population average $-\mathbf{B}$ (Assumption 4). The underlying $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$ itself remains exactly identifiable.

Proposition (i) and (ii) have been established in Section 4.2, as $\boldsymbol{v}_X$ is absolutely identifiable, invariant to the latent ambiguity and $\boldsymbol{X}$ is observed. We now establish the computational expression (Proposition (iii)).

Step 1: Chain Rule Decomposition & Projection onto the Tangent Space. We aim to compute the Jacobian of the gene expression velocity $\boldsymbol{v}_X$ with respect to the gene expression state $\boldsymbol{X}$. Using the multivariate chain rule via the latent state $\boldsymbol{Z}$, we decompose this as

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{Z}} \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}}. \tag{19}$$

Since the generative map $\boldsymbol{X} = \phi_x(\boldsymbol{Z})$ is a smooth injective embedding from the low-dimensional latent space $\mathbb{R}^d$ to the high-dimensional observation space $\mathbb{R}^G$ (where $G \gg d$), its Jacobian $J_{\phi_x} \in \mathbb{R}^{G \times d}$ has full column rank. Taking the differential on both sides yields $d\boldsymbol{X} = J_{\phi_x} d\boldsymbol{Z}$. To solve for $d\boldsymbol{Z}$, we left-multiply by $J_{\phi_x}^\top$, and then get

$$J_{\phi_x}^\top d\boldsymbol{X} = J_{\phi_x}^\top J_{\phi_x} d\boldsymbol{Z}.$$

Because J_{ϕ_x} has full column rank, the metric tensor $J_{\phi_x}^\top J_{\phi_x} \in \mathbb{R}^{d \times d}$ is invertible. Left-multiplying by its inverse gives the left pseudo-inverse $d\boldsymbol{Z} = (J_{\phi_x}^\top J_{\phi_x})^{-1} J_{\phi_x}^\top d\boldsymbol{X}$, and hence $\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} = (J_{\phi_x}^\top J_{\phi_x})^{-1} J_{\phi_x}^\top$.

By invoking the joint scaled isometric embedding constraint (Assumption 1), we have $J_{\phi_x}^\top J_{\phi_x} = \eta^2 \mathbf{I}_d - \gamma^2 J_{\phi_w}^\top J_{\phi_w}$. Substituting this into the pseudo-inverse yields

$$\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} = (J_{\phi_x}^\top J_{\phi_x})^{-1} J_{\phi_x}^\top = \left(\eta^2 \mathbf{I}_d - \gamma^2 J_{\phi_w}^\top J_{\phi_w}\right)^{-1} J_{\phi_x}^\top. \tag{20}$$

Strictly speaking, since $\boldsymbol{X} = \phi_x(\boldsymbol{Z})$ defines a d-dimensional submanifold in $\mathbb{R}^G$, tangent vectors $d\boldsymbol{X}$ live in the tangent bundle of this submanifold. The pseudoinverse $(J_{\phi_x}^\top J_{\phi_x})^{-1} J_{\phi_x}^\top$ realizes the canonical pull-back from tangent vectors of the gene manifold to latent tangent vectors. With this caveat, we use $\partial \boldsymbol{Z} / \partial \boldsymbol{X}$ as a shorthand for this pull-back operation.

Step 2: Population-Level First-Order Approximation. The observable gene-expression velocity is generated by $\boldsymbol{v}_X(\boldsymbol{Z}) = J_{\phi_x}(\boldsymbol{Z})\boldsymbol{v}_Z(\boldsymbol{Z})$. Taking the derivative with respect to $\boldsymbol{Z}$ requires the matrix-vector product rule:

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{Z}} = J_{\phi_x}(\boldsymbol{Z})\frac{\partial \boldsymbol{v}_Z}{\partial \boldsymbol{Z}} + \left(\nabla_{\boldsymbol{Z}} J_{\phi_x}(\boldsymbol{Z})\right)\bar{\times}\boldsymbol{v}_Z(\boldsymbol{Z}), \tag{21}$$

where the second term involves a third-order tensor $\nabla_{\boldsymbol{Z}} J_{\phi_x}$ contracted with the velocity vector $\boldsymbol{v}_Z$, representing the non-linear high-order deformations caused by the local curvature of the decoder manifold $\mathcal{M}_\gamma$.

To extract the dominant global regulatory network driving the developmental trajectory, we introduce two physically motivated first-order approximations:

1. **Low curvature:** We assume the local decoder manifold is sufficiently smooth, such that the curvature term $\nabla_{\boldsymbol{Z}} J_{\phi_x}$ is negligible compared to the primary linear variations. Thus, the second term vanishes.
2. **Population-average latent Jacobian:** The latent Jacobian expands as $\frac{\partial \boldsymbol{v}_Z}{\partial \boldsymbol{Z}} = J_g(\boldsymbol{Z}) + \mathbf{B}J_{\boldsymbol{M}}(\boldsymbol{Z}) - \mathbf{B}$, whose population average is exactly $-\mathbf{B}$ by Assumption 4. We use this value in place of the pointwise Jacobian. The pointwise residual is $J_g(\boldsymbol{Z}) + \mathbf{B}J_{\boldsymbol{M}}(\boldsymbol{Z})$.

Applying these two approximations, we obtain the dominant first-order linear component:

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{Z}} \approx J_{\phi_x}(\boldsymbol{Z}) \cdot (-\mathbf{B}). \tag{22}$$

This replacement is exact in expectation under Assumption 4.

Step 3: Final Assembly. Substituting the partial derivatives from Equation 20 and Equation 22 back into the chain rule Eq 19, we arrive at the final approximation for the spatially adjusted GRN Jacobian:

$$\frac{\partial \boldsymbol{v}_X}{\partial \boldsymbol{X}} \approx J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top = J_{\phi_x} \cdot (-\mathbf{B}) \cdot \left(\eta^2 \mathbf{I}_d - \gamma^2 J_{\phi_w}^\top J_{\phi_w}\right)^{-1} \cdot J_{\phi_x}^\top. \tag{23}$$

In practice, we can directly use $J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top$ to evaluate the GRN from $\mathbf{B}$ and decoder Jacobian. To better demonstrate the biological meaning, referring the derivation $J_{\phi_x} \cdot (-\mathbf{B}) \cdot \left(\eta^2 \mathbf{I}_d - \gamma^2 J_{\phi_w}^\top J_{\phi_w}\right)^{-1} \cdot J_{\phi_x}^\top$, which, compared to the vanilla quadratic form $J_{\phi_x}(-\mathbf{B})J_{\phi_x}^\top$, formally justifies how the innate latent regulatory interactions ($\mathbf{B}$) are lifted to the gene observation space via the structural maps (J_{ϕ_x}), while being dynamically re-weighted by the tissue's spatial microenvironment via the precision adjustment term $J_{\phi_w}^\top J_{\phi_w}$. Empirically, the learned feedback is predominantly restorative: on MOSTA, 71.1% of the eigenvalues of $\mathbf{B}$ have positive real part, with spectral radius 1.35.

A.5 Realization of Assumption 4 in the Trained Model

By Lemma 2, the normalized representative $\mathbf{B}^\star = -\mathbb{E}_\rho[J_{\boldsymbol{v}}]$ is obtained in closed form from the fitted model. On MOSTA, the relative residual $\|\mathbb{E}_\rho[J_g + \mathbf{B}J_{\boldsymbol{M}}]\|_F / \|\mathbf{B}\|_F$ at the fitted parametrization is 0.385, and the fitted and normalized matrices have matrix cosine 0.961. Propagating both through the same decoder Jacobian yields nearly identical GRNs (Table 1), in contrast to a null of 100 norm-matched random operators. The signed-edge agreement between the two is 1.000 / 0.978 / 0.959 on MOSTA / ARTISTA / ZESTA.

Table 1: GRN-level agreement between the fitted and the normalized representative on MOSTA. The null replaces the normalized operator by 100 norm-matched random operators (mean ± std). Sign agreement and $P@10^4$ are computed on the 10^4 strongest edges.

Quantity	Fitted vs. normalized	Random-operator null
Operator cosine (GRN-effective subspace)	0.968	0.00 ± 0.02
GRN matrix correlation	0.977	-0.01 ± 0.04
Sign agreement (10^4 strongest edges)	1.000	0.49 ± 0.04
Top-edge retrieval, $P@10^4$	0.855	0.21 ± 0.01

B Detailed Architecture and Training Specifications

In this section, we provide the complete architectural details, hyperparameters, and explicit formulations for the auxiliary loss functions omitted from the main text (Section 5).

B.1 Network Architecture

Multimodal Encoder Ψ. The encoder maps each cell's raw observation $(\boldsymbol{X}, \boldsymbol{S})$ to a latent pattern $\boldsymbol{Z} \in \mathbb{R}^d$ in four stages.

(1) Spatial normalization. The raw coordinate S is converted to $\boldsymbol{S}_{\text{rel}}$ via global z-score normalization and PCA canonicalization ($\boldsymbol{S}_{\text{rel}} = (\boldsymbol{S} - \bar{\boldsymbol{S}})\,\mathbf{V}^\top$).

(2) Gene feature extraction. The raw gene expression vector $\boldsymbol{X} \in \mathbb{R}^G$ is processed by a three-layer MLP with dimensions $G \to 512 \to 256 \to 128$. The first hidden layer is followed by ReLU and Dropout, and the second hidden layer is followed by ReLU only to output molecular encoding vector h_g.

(3) Relative spatial encoding. A deterministic uniform grid of K anchor points $\{a_k\}_{k=1}^K$ is placed on the PCA-normalized coordinate space. Gaussian-kernel weights $w_k \propto \exp(-\|\boldsymbol{S}_{\text{rel}} - a_k\|^2/2\sigma^2)$ (with learnable bandwidth σ) aggregate relative offsets and RBF distance features into a 64-dimensional spatial context vector h_s.

(4) Spatial fusion and projection. The extracted gene features h_g and spatial context representations h_s are fused by L per-cell fusion layers. Each layer concatenates the cell's current feature with its own spatial context vector and passes them through an update MLP. The fusion is per-cell: spatial information enters through the anchor-kernel context h_s. After the fusion layers, a two-layer projection MLP (fusion-dim $\to 2d \to d$) followed by a single LayerNorm produces the final $\boldsymbol{Z} \in \mathbb{R}^d$.

Algorithm 1 Encoder $\Psi(\boldsymbol{X}, \boldsymbol{S}) \to \boldsymbol{Z}$

Require: Gene expression $\boldsymbol{X} \in \mathbb{R}^G$, spatial coordinate $\boldsymbol{S} \in \mathbb{R}^2$
$\boldsymbol{S}_{\text{rel}}, \bar{\boldsymbol{S}}, \mathbf{V} \leftarrow \text{PCANorm}(\boldsymbol{S})$ {PCA Canonicalization}
$h_g \leftarrow \text{Linear}_{256\to128}(\text{ReLU}(\text{Dropout}(\text{Linear}_{G\to512}(\boldsymbol{X}))))$ {gene MLP, 128-dim}
$w_k \propto \exp(-\|\boldsymbol{S}_{\text{rel}} - a_k\|^2/2\sigma^2)$ {anchor weights}
$h_s \leftarrow \text{MLP}_s(\sum_k w_k(a_k - \boldsymbol{S}_{\text{rel}}),\ \sum_k w_k\,\phi_{\text{RBF}}(\|\boldsymbol{S}_{\text{rel}} - a_k\|))$ {spatial context, 64-dim}
for $\ell = 1, \ldots, L$ **do**
  $h_g \leftarrow \text{ReLU}\big(\text{Fuse}_\ell(h_g,\, h_s)\big)$ {per-cell fusion}
end for
$\boldsymbol{Z} \leftarrow \text{LayerNorm}\big(\text{Linear}_{2d\to d}(\text{ReLU}(\text{Linear}_{\cdot\to 2d}(h_g)))\big)$
return $\boldsymbol{Z} \in \mathbb{R}^d$

Latent Dynamics: g_ϕ, B_ϕ, M_θ. The nonlinear drift $g_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is parameterized by a multi-layer MLP with an initial Linear($d \to h$) + LayerNorm + ReLU, followed by residual blocks (i.e., $x \mapsto \text{ReLU}(\text{LayerNorm}(x + \text{Linear}(\text{ReLU}(\text{Linear}(x)))))$ for each block), and a final linear map ($h \to d$). The feedback matrix $B_\phi \in \mathbb{R}^{d\times d}$ is a bias-free linear layer initialized with $\mathcal{N}(0, 0.01^2)$ weights to ensure weak initial feedback.

The target predictor $M_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is implemented as a prototype attractor model with K_{pro} learnable prototypes $\{P_k\}_{k=1}^{K_{\text{pro}}} \subset \mathbb{R}^d$. Each cell's predicted target is the softmax-weighted combination

$$M_\theta(\boldsymbol{Z}_i) = \sum_{k=1}^{K_{\text{pro}}} a_k(\boldsymbol{Z}_i)\, P_k, \qquad a_k(\boldsymbol{Z}_i) = \frac{\exp(-\|\boldsymbol{Z}_i - P_k\|^2/\tau_{\text{pro}})}{\sum_{k'} \exp(-\|\boldsymbol{Z}_i - P_{k'}\|^2/\tau_{\text{pro}})}, \tag{24}$$

where $\tau_{\text{pro}} > 0$ is a learnable temperature (parameterized in log-space for positivity). The K prototypes correspond to distinct cell-fate attractors; the equilibrium loss directly optimizes their positions, yielding denser and more stable gradients than a free MLP. ODE integration in experiments uses a single-step RK4 solver.

Decoders Φ_X and Φ_S. The gene decoder $\Phi_X : \mathbb{R}^d \to \mathbb{R}^G$ is a multi-layer MLP with hidden dimension ReLU activations. The spatial decoder $\Phi_S : \mathbb{R}^d \to \mathbb{R}^2$ is a three-layer MLP with linear output, since canonical coordinates $\boldsymbol{S}_{\text{rel}}$ lie in an unbounded range after PCA normalization. During evaluation, predicted canonical coordinates are mapped back to physical space via the inverse PCA transformation $\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}}_{\text{rel}}\,\mathbf{V} + \bar{\boldsymbol{S}}$.

B.2 Loss Functions

Given the inherently unpaired nature of cross-sectional spatial transcriptomics, we cannot rely on pointwise trajectory supervision. Instead, we train the model using a composite objective

$$\mathcal{L} = w_{\text{recon}}\,\mathcal{L}_{\text{recon}} + w_{\text{traj}}\,\mathcal{L}_{\text{traj}} + w_{\text{iso}}\,\mathcal{L}_{\text{iso}} + w_{\text{div}}\,\mathcal{L}_{\text{div}} + w_{\text{eq}}\,\mathcal{L}_{\text{eq}}. \tag{25}$$

Reconstruction $\mathcal{L}_{\text{recon}}$. Within each forward pass, the decoders are supervised using noise-perturbed latent codes $\tilde{\boldsymbol{Z}}_t = \boldsymbol{Z}_t + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$, to encourage smoother decoder mappings (denoising autoencoder). For each time point, $\boldsymbol{Z}_{t_k} = \Psi(\boldsymbol{X}_{t_k}, \boldsymbol{S}_{t_k})$, and we have $\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{recon-x}} + w_{\text{recon_s}}\mathcal{L}_{\text{recon-s}}$, where

$$\mathcal{L}_{\text{recon-x}} = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{S})\sim p_{t_k}}\big[\|\Phi_X(\tilde{\boldsymbol{Z}}_{t_k}) - \boldsymbol{X}_{t_k}\|^2\big], \tag{26}$$

$$\mathcal{L}_{\text{recon-s}} = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{S})\sim p_{t_k}}\big[\|\Phi_S(\tilde{\boldsymbol{Z}}_{t_k}) - \boldsymbol{S}_{\text{rel},t_k}\|^2\big]. \tag{27}$$

Trajectory alignment $\mathcal{L}_{\text{traj}}$. For every consecutive time pair (t_k, t_{k+1}), the ODE-predicted latent distribution is aligned with the directly encoded target distribution via entropy-regularized (potential unbalanced) OT

$$\mathcal{L}_{\text{traj}} = \sum_k \min_\pi \mathbb{E}_{(\hat{\boldsymbol{Z}},\,\boldsymbol{Z}^{\text{true}})\sim\pi}\big[\|\hat{\boldsymbol{Z}} - \boldsymbol{Z}^{\text{true}}\|^2\big] \approx \sum_k \text{OT}_{\varepsilon,\tau}\big(\hat{\boldsymbol{Z}}_{t_{k+1}},\ \boldsymbol{Z}^{\text{true}}_{t_{k+1}}\big), \tag{28}$$

where $\hat{Z}_{t_{k+1}}$ is obtained by integrating the ODE from the encoding $\boldsymbol{Z}_{t_k}$, and $\boldsymbol{Z}^{\text{true}}_{t_{k+1}} = \Psi(\boldsymbol{X}_{t_{k+1}}, \boldsymbol{S}_{t_{k+1}})$ is treated as a fixed target. The transport cost is the squared Euclidean distance in $\mathcal{Z}$. We use stop-gradient on $\boldsymbol{Z}_{t_k}$ to ensure that the trajectory loss trains only the ODE parameters: without it, the encoder could trivially minimize the transport cost by mapping $\boldsymbol{Z}_{t_k} \approx \boldsymbol{Z}^{\text{true}}_{t_{k+1}}$ (i.e., zero velocity), which would conflict with the diversity loss and prevent the ODE from learning any dynamics. Despite that, in each mini-batch, we optimize with both reconstruction and trajectory loss so that we still achieve an EM-style of jointly optimization and mutually refinement of encoder-decoder and latent flow parameters.

Concretely, $\text{OT}_{\varepsilon,\tau}$ is the Sinkhorn divergence implemented via Python library `geomloss` [43] with entropic regularization ε, unbalanced marginal relaxation τ, and debiasing. The unbalanced formulation allows mass creation and annihilation, which is appropriate for cell differentiation where the cellular composition differs between time points.

Assumption 2 posits that the true velocity field is the Benamou–Brenier minimizer. Our trajectory loss does not enforce this directly—instead, it aligns ODE-forwarded distributions to encoded targets via Sinkhorn divergence. By the Benamou–Brenier theorem, static OT with squared Euclidean cost coincides with the Benamou–Brenier minimizer along displacement interpolation [23]; combined with the smoothness inductive bias of neural ODE parametrization, the learned velocity field is implicitly biased toward low-kinetic-energy solutions. Empirically, adding an explicit kinetic regularizer $\mathbb{E}\|\boldsymbol{v}_Z\|^2$ yields negligible improvement and can even harm performance in some cases (Appendix D, Table 7).

Isometric regularization $\mathcal{L}_{\text{iso}}$. Isometric regularization enforces the scaled isometry condition (Assumption 1), computed on Z_{t_k}. Exact Jacobian computation is computationally intensive, and hence we instead use a random finite-difference approximation with $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$, $\sigma = 0.05$. By a first-order Taylor expansion, the per-sample ratio

$$r_i = \frac{\|\Phi_X(\boldsymbol{Z}_i{+}\varepsilon_i) - \Phi_X(\boldsymbol{Z}_i)\|^2 + \gamma^2\|\Phi_S(\boldsymbol{Z}_i{+}\varepsilon_i) - \Phi_S(\boldsymbol{Z}_i)\|^2}{\|\varepsilon_i\|^2} \approx \frac{\varepsilon_i^\top(J_{\phi_x}^\top J_{\phi_x} + \gamma^2 J_{\phi_w}^\top J_{\phi_w})\varepsilon_i}{\|\varepsilon_i\|^2} \tag{29}$$

approximates the directional scaling of the encoder map at Z_i, and equals λ^2 for all $(\boldsymbol{Z}_i, \varepsilon_i)$ if and only if the isometry condition $J_{\phi_x}^\top J_{\phi_x} + \gamma^2 J_{\phi_w}^\top J_{\phi_w} = \lambda^2 \mathbf{I}_d$ holds everywhere. We therefore set

$$\mathcal{L}_{\text{iso}} = \frac{\text{Var}(r)}{\mathbb{E}[r]^2}. \tag{30}$$

This coefficient-of-variation formulation is scale-invariant: it vanishes if and only if r_i is constant across all cells and directions, and cannot be trivially minimized by rescaling decoder weights (unlike fixing a batch-estimated λ^2 with a stop-gradient). This requires only one additional forward pass and no explicit Jacobian.

Latent diversity $\mathcal{L}_{\text{div}}$. Latent diversity regularization prevents latent mode collapse

$$\mathcal{L}_{\text{div}} = -\text{std}(\boldsymbol{Z}_{t_k}). \tag{31}$$

Equilibrium $\mathcal{L}_{\text{eq}}$. Applied when t_{k+1} is the final normalized time point ($t_{k+1} = 1.0$), the equilibrium loss enforces the control-theoretic steady-state condition and directly supervises the prototype positions in M_θ

$$\mathcal{L}_{\text{eq}} = \mathbb{E}_{\boldsymbol{Z} \sim p_T} \left[\| M_\theta(\boldsymbol{Z}) - \boldsymbol{Z} \|^2 \right], \tag{32}$$

where Z is encoded directly from the true observations at the terminal time T.

B.3 Optimization and Training Schedule

All parameters are optimized with AdamW (learning rate 10^{-4}, default weight decay), with gradient clipping at maximum norm 1.0. Raw time points are linearly normalized to $[0, 1]$ before ODE integration. Training runs for 200 epochs with a two-phase curriculum.

- **Phase 1** (epoch $< T_1$). We introduce an encoder-decoder warm-up stage with T_1 epochs, where `PatternDynamics` and `TargetPredictor` are frozen and the feedback gain is fixed at $\beta = 0$. In this phase, each training step processes a **single** time point (rather than a time pair), so each slice is visited exactly once per epoch. Only the encoder Ψ and decoders Φ_X, Φ_S are updated, under $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{iso}}$. This establishes a stable latent geometry before any dynamical constraints are introduced.
- **Phase 2** (epoch $\geq T_1$). The dynamics module is unfrozen and $\mathcal{L}_{\text{traj}}$ and $\mathcal{L}_{\text{eq}}$ are activated alongside the Phase 1 terms. Training now iterates over consecutive time pairs. To prevent catastrophic divergence early before the target predictor M_θ has learned to yield meaningful set-points, we scale the restorative feedback term $B_\phi(M_\theta(\boldsymbol{Z}) - \boldsymbol{Z})$ by a scheduling multiplier β, where β is linearly annealed across a warm-up interval $[T_1, T_1 + T_2]$

  $$\beta(\text{epoch}) = \min\left(1,\ \frac{\text{epoch} - T_1 + 1}{T_2}\right), \tag{33}$$

  reaching $\beta = 1$ exactly at T_2, ensuring the restorative feedback $B_\phi(M_\theta(\boldsymbol{Z}) - \boldsymbol{Z})$ is fully engaged only once the ODE direction has stabilized.

C Details on Benchmark Experiments

C.1 Datasets

Mouse Embryo (MOSTA). Available at `https://db.cngb.org/stomics/mosta/`, it provides spatially resolved transcriptomic profiling of mouse embryos across eight developmental stages from embryonic day E9.5 to E16.5 (9.5, 10.5, 11.5, 12.5, 13.5, 14.5, 15.5, 16.5 days post-fertilization) [34].

Zebrafish Embryo (ZESTA). From the ZESTA atlas at `https://db.cngb.org/stomics/zesta/`, this dataset profiles zebrafish embryo development across six time points (3.3, 5.25, 10, 12, 18, and 24 hours post-fertilization, hpf), spanning early embryogenesis through the pharyngula stage [35].

Axolotl Telencephalon Regeneration (ARTISTA). Available at `https://db.cngb.org/stomics/artista/`, this dataset tracks brain regeneration in the axolotl (*Ambystoma mexicanum*) telencephalon following injury, with spatial transcriptomic snapshots collected at 2, 5, 10, 15, 20, 30, and 60 days post-injury (DPI), plus an uninjured control (we treat it as the regeneration target at 100 days) [36].

Spatial visualizations of these spatial transcriptomic slices are shown in Fig. 6, 7, 8. Detailed information about the datasets are shown in Table 2.

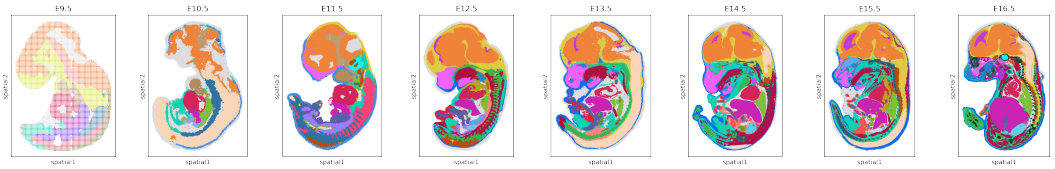


Figure 6: Spatial visualization of MOSTA data, with cell colored by its cell-type annotations.

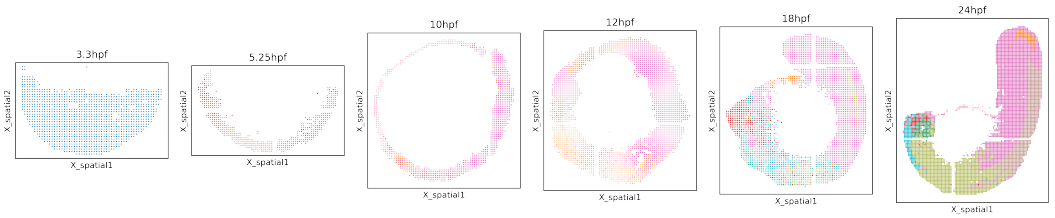


Figure 7: Spatial visualization of ZESTA data, with cell colored by its cell-type annotations.

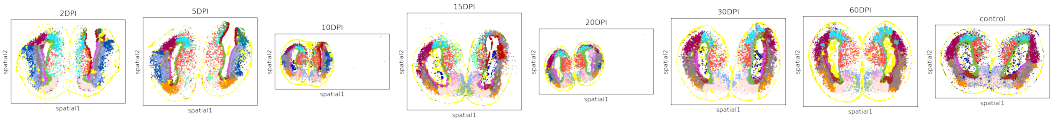


Figure 8: Spatial visualization of ARTISTA data, with cell colored by its cell-type annotations.

C.2 Benchmark Methods

We compare against three representative baselines that span the major paradigms in spatiotemporal trajectory inference. Top 3,000 highly variable genes (HVGs) are selected on the training slices and gene expression is preprocessed with standard library-size normalization and log1p transformation. Spatial coordinates are global z-score across all training time points jointly to preserve inter-slice tissue growth, and PCA-canonical coordinates are used throughout training and evaluation. During the LOTO evaluation, to prevent data leakage, we normalized the global gene expression and spatial coordinates across all training slices, and used the normalization parameters from the training set to normalize the hold-out evaluation slice. For each dataset, we conduct LOTO experiments for slice 2, 3, and 4 (the index is established with 0 as the start, i.e., index 2 means the slice at the third timepoint).

Moscot [5]. Moscot formulates spatiotemporal alignment as an OT problem that jointly transports cells in gene-expression and physical space across time. However, Moscot in its original form does not support leave-one-out inference, as it requires direct access to both endpoints of each time transition. To adapt it to the LOTO setting, we design a velocity-field extension: we solve the OT problem on PCA-reduced features ($d = 50$) for each pair of consecutive training slices, using the recommended settings: entropic regularization $\varepsilon = 0.1$, spatial-feature trade-off $\alpha = 0.5$, unbalanced marginals $\tau_a = \tau_b = 0.95$, and low-rank approximation (rank $= 1000$) to handle large cell counts. For each training pair, barycentric projection of the row-normalized transport plan yields a per-cell displacement, which divided by the time interval dt gives a velocity estimate. These velocity samples

Table 2: Detailed information on real-world spatiotemporal transcriptomics dataset in experiments.

Dataset	Time	Slice	Cells	Cell types
MOSTA	9.5	Mouse_embryo_E9.5	5913	12
MOSTA	10.5	Mouse_embryo_E10.5	18408	12
MOSTA	11.5	Mouse_embryo_E11.5	30124	19
MOSTA	12.5	Mouse_embryo_E12.5	51365	23
MOSTA	13.5	Mouse_embryo_E13.5	77369	19
MOSTA	14.5	Mouse_embryo_E14.5	102519	26
MOSTA	15.5	Mouse_embryo_E15.5	113350	26
MOSTA	16.5	Mouse_embryo_E16.5	121767	25
ZESTA	3.3	Zebrafish_embryo_3.3hpf	1167	1
ZESTA	5.25	Zebrafish_embryo_5.25hpf	563	7
ZESTA	10	Zebrafish_embryo_10hpf	1036	4
ZESTA	12	Zebrafish_embryo_12hpf	2081	5
ZESTA	18	Zebrafish_embryo_18hpf	3048	8
ZESTA	24	Zebrafish_embryo_24hpf	5271	9
ARTISTA	2	Axolotl_telencephalon_2DPI	7668	18
ARTISTA	5	Axolotl_telencephalon_5DPI	8106	19
ARTISTA	10	Axolotl_telencephalon_10DPI	9440	18
ARTISTA	15	Axolotl_telencephalon_15DPI	9676	21
ARTISTA	20	Axolotl_telencephalon_20DPI	11319	20
ARTISTA	30	Axolotl_telencephalon_30DPI	9252	17
ARTISTA	60	Axolotl_telencephalon_60DPI	10964	16
ARTISTA	100	Axolotl_telencephalon_control	11698	17

across all training pairs are then interpolated into a global RBF velocity field. At inference, source cells at t_{k-1} are propagated forward to the held-out t_k by integrating this field with an ODE solver (RK45), and gene expression is recovered by inverting the PCA transform.

Spateo [4]/Dynamo [26]. This approach estimates a pseudo-velocity field from the observed PCA-space distributions across training time points: for each training time point, the per-cell velocity is approximated by the population-mean shift toward the next time point (scaled by $1/dt$) plus a deviation term proportional to the cell's displacement from the current mean. A global vector field is then learned via Dynamo's SparseVFC (sparse approximation of vector field consensus) [44] using these estimated velocities, and source cells at t_{k-1} are propagated forward by integrating this learned field with an ODE solver (RK45) over the interval $\Delta t = t_k - t_{k-1}$. Gene expression is recovered by inverting the PCA transform.

STORIES [7]. STORIES learns a flow between consecutive spatial transcriptomics slices under an entropic regularization that jointly penalizes gene-expression transport cost and spatial displacement. The model is trained on the $T-1$ training slices as a sequence of pairwise flows, and inference at the held-out time is performed by applying the learned flow from the nearest preceding source slice in PCA space, with gene expression recovered by inverting the PCA transform. We train for up to 2,000 iterations with learning rate 10^{-2}, batch size 1,000, and quadratic regularization weight 5×10^{-3}, with early stopping patience of 150 iterations.

PrevCopy. As a naive baseline, PrevCopy predicts the held-out time point by replicating the mean gene expression profile of the source slice at t_{k-1} to every cell in the target slice. This baseline captures how much predictive information is already contained in the marginal distribution of the nearest preceding time point, without any modeling of dynamics.

C.3 Evaluation Metrics

All methods are evaluated on the unseen held-out timepoint with the following metrics.

HVG Pearson Correlation (X_PCC, ↑). We select the top-100 HVGs from $X_{t_k}^{\text{true}}$ and compute the Pearson correlation coefficient (PCC) between the predicted and true mean expression vectors

over these genes. It measures how accurately the predicted population-level gene expression profile matches the true profile on informative genes. Higher is better.

Wasserstein Distance (X_WD, ↓). We compute the per-gene 1D Wasserstein distance between the predicted and true expression distributions $\mathcal{W}_1\left(\hat{X}_{t_k}^{(g)},\ X_{t_k}^{\mathrm{true},(g)}\right)$, and take the median across all HVGs. It captures the distributional discrepancy in gene expression marginals. Lower is better.

Spatial Coherence (S_PC1, ↑). Following the definition in STORIES [7], spatial coherence measures whether the predicted gene expression pattern, when mapped onto the true spatial coordinates of the held-out slice, exhibits biologically realistic spatial autocorrelation. We project the true cells $X_{t_k}^{\mathrm{true}}$ onto a direction extracted from the predicted expression $\hat{X}_{t_k}$, and compute Moran's I on the resulting scalar field using $S_{t_k}^{\mathrm{true}}$ with a $k = 10$ nearest-neighbor spatial graph. We use the first principal component (PC1) direction of $\hat{X}_{t_k}$ (fit on the predicted cells) as the projection direction, which captures the dominant mode of predicted expression variation. Higher Moran's I indicates that the predicted expression programs are spatially organized in a manner consistent with the true tissue architecture.

Spatial Domain ARI (S_ARI, ↑). We perform Leiden clustering on the predicted gene expression projected into the true spatial coordinate space and compute the Adjusted Rand Index (ARI) against the provided cell-type annotations, which captures whether the predicted expression landscape reproduces the spatial organization of distinct cell-type domains.

Overall Score. To facilitate a comprehensive comparison, we apply min-max normalization (commonly used in single-cell biology bechmark) [37] to each metric across all methods (after negating the Wasserstein distance). We treat X_PCC as the main metric for gene expression prediction, as it robustly captures biologically critical, scale-invariant relative trends of gene regulation, while utilizing X_WD secondarily to verify macroscopic distributional fidelity. Similarly, S_ARI serves as the primary metric for spatial coherence with S_PC1 as an auxiliary, because S_ARI more comprehensively evaluates the structural preservation of distinct cell-type spatial distributions. Hence, an aggregate score is calculated by combining these four min-max normalized metrics (X_PCC, X_WD, S_ARI, S_PC1) with weights of 0.35, 0.15, 0.35, and 0.15, respectively.

C.4 Implementation Details

All experiments are run on a compute cluster with a single GPU (SLURM partition, 8 CPU cores, 100 GB RAM). The running time with 200 epochs and batch size 512 on each config is shown in Table 3.

Table 3: Running time of 200 epochs with batch size 512

Data	LOTO	Elapsed Time
axolotl_telencephalon_regeneration	2	38 min 58.70 s
axolotl_telencephalon_regeneration	3	31 min 28.54 s
axolotl_telencephalon_regeneration	4	28 min 24.39 s
zebrafish_embryo	2	6 min 01.94 s
zebrafish_embryo	3	5 min 46.79 s
zebrafish_embryo	4	5 min 34.96 s
mouse_embryo	2	4 h 18 min 31 s
mouse_embryo	3	4 h 10 min 22 s
mouse_embryo	4	2 h 20 min 55 s

D Detailed Experiments Results

We provided the detailed metrics values in benchmark and ablation experiments.

D.1 Benchmark Results

Table 4 shows detailed benchmark results.

Table 4: Details on benchmark results.

dataset	loto_idx	method	X_PCC↑	X_WD↓	S_ARI↑	S_PC1↑	X_PCC_norm	X_WD_norm	S_ARI_norm	S_PC1_norm	overall_score↑	rank
ARTISTA	2	Ours	0.975	0.025	0.445	0.827	1.000	0.699	1.000	1.000	0.955	1
ARTISTA	2	Spateo	0.967	0.025	0.319	0.796	0.615	0.573	0.538	0.929	0.629	2
ARTISTA	2	STORIES	0.967	0.026	0.352	0.799	0.639	0.000	0.658	0.936	0.594	3
ARTISTA	2	Moscot	0.953	0.024	0.327	0.796	0.000	0.922	0.565	0.929	0.476	4
ARTISTA	2	PrevCopy	0.963	0.024	0.172	0.393	0.451	1.000	0.000	0.000	0.308	5
ARTISTA	3	Ours	0.972	0.038	0.413	0.857	1.000	1.000	1.000	1.000	1.000	1
ARTISTA	3	STORIES	0.960	0.040	0.369	0.848	0.831	0.788	0.766	0.986	0.825	2
ARTISTA	3	Spateo	0.932	0.042	0.392	0.842	0.416	0.396	0.888	0.977	0.662	3
ARTISTA	3	Moscot	0.904	0.043	0.331	0.842	0.000	0.334	0.568	0.978	0.395	4
ARTISTA	3	PrevCopy	0.913	0.045	0.223	0.203	0.130	0.000	0.000	0.000	0.046	5
ARTISTA	4	Ours	0.738	0.040	0.458	0.869	1.000	1.000	1.000	1.000	1.000	1
ARTISTA	4	STORIES	0.666	0.042	0.419	0.858	0.325	0.724	0.863	0.985	0.672	2
ARTISTA	4	Spateo	0.678	0.046	0.414	0.856	0.439	0.321	0.846	0.982	0.645	3
ARTISTA	4	Moscot	0.631	0.045	0.418	0.856	0.000	0.382	0.862	0.981	0.506	4
ARTISTA	4	PrevCopy	0.683	0.048	0.173	0.167	0.490	0.000	0.000	0.000	0.171	5
MOSTA	2	STORIES	0.937	0.093	0.402	0.877	0.833	1.000	1.000	1.000	0.942	1
MOSTA	2	Ours	0.948	0.094	0.380	0.840	1.000	0.936	0.880	0.807	0.919	2
MOSTA	2	Moscot	0.891	0.100	0.399	0.768	0.117	0.414	0.987	0.430	0.513	3
MOSTA	2	Spateo	0.883	0.099	0.369	0.768	0.000	0.507	0.818	0.429	0.427	4
MOSTA	2	PrevCopy	0.890	0.105	0.221	0.686	0.098	0.000	0.000	0.000	0.034	5
MOSTA	3	Ours	0.956	0.120	0.328	0.915	1.000	0.276	0.859	0.920	0.830	1
MOSTA	3	STORIES	0.932	0.122	0.320	0.928	0.789	0.167	0.805	0.976	0.729	2
MOSTA	3	Moscot	0.855	0.110	0.350	0.933	0.082	1.000	1.000	0.997	0.678	3
MOSTA	3	Spateo	0.852	0.115	0.349	0.934	0.053	0.658	0.991	1.000	0.614	4
MOSTA	3	PrevCopy	0.846	0.124	0.197	0.694	0.000	0.000	0.000	0.000	0.000	5
MOSTA	4	Ours	0.926	0.037	0.361	0.934	1.000	1.000	1.000	0.984	0.998	1
MOSTA	4	Moscot	0.907	0.040	0.313	0.934	0.289	0.424	0.692	0.984	0.555	2
MOSTA	4	STORIES	0.908	0.041	0.289	0.942	0.324	0.274	0.545	1.000	0.495	3
MOSTA	4	Spateo	0.908	0.042	0.298	0.934	0.323	0.000	0.597	0.985	0.470	4
MOSTA	4	PrevCopy	0.899	0.041	0.204	0.433	0.000	0.303	0.000	0.000	0.046	5
ZESTA	2	Ours	0.948	0.171	0.120	0.859	0.972	0.441	1.000	1.000	0.906	1
ZESTA	2	STORIES	0.954	0.173	0.109	0.699	1.000	0.359	0.886	0.803	0.834	2
ZESTA	2	Moscot	0.825	0.159	0.086	0.408	0.419	1.000	0.658	0.445	0.594	3
ZESTA	2	Spateo	0.814	0.181	0.083	0.408	0.369	0.000	0.629	0.445	0.416	4
ZESTA	2	PrevCopy	0.732	0.167	0.020	0.046	0.000	0.625	0.000	0.000	0.094	5
ZESTA	3	Ours	0.955	0.139	0.069	0.854	1.000	1.000	1.000	1.000	1.000	1
ZESTA	3	STORIES	0.932	0.149	0.053	0.840	0.040	0.000	0.693	0.983	0.404	2
ZESTA	3	Spateo	0.935	0.146	0.039	0.832	0.135	0.315	0.432	0.973	0.392	3
ZESTA	3	Moscot	0.931	0.148	0.039	0.832	0.000	0.075	0.422	0.974	0.305	4
ZESTA	3	PrevCopy	0.939	0.139	0.016	0.045	0.325	0.950	0.000	0.000	0.256	5
ZESTA	4	Ours	0.934	0.210	0.064	0.869	0.847	1.000	1.000	0.973	0.943	1
ZESTA	4	STORIES	0.944	0.260	0.006	0.868	1.000	0.051	0.000	0.971	0.503	2
ZESTA	4	Spateo	0.896	0.242	0.037	0.884	0.256	0.385	0.542	0.999	0.487	3
ZESTA	4	Moscot	0.900	0.263	0.017	0.885	0.328	0.000	0.197	1.000	0.334	4
ZESTA	4	PrevCopy	0.879	0.244	0.008	0.307	0.000	0.352	0.046	0.000	0.069	5

D.2 Ablation and Hyperparameter Sensitivity

Results of ablation and hyperparameter sweep experiments are shown in Table 5 (for ARTISTA-LOTO4 setting) and Table 6 (for ZESTA-LOTO3 setting). The overall score in a sweep experiment is calculated by weighting the min-max normalized results from all parameter combinations. The baseline setting is w_trajectory=1.0, w_recon=1.0, w_equilibrium=0.01, w_diversity=0.01, w_isometric=0.01, decoder_warmup_epochs=30, feedback_warmup_epochs=20, sinkhorn_epsilon=0.1, sinkhorn_tau=0.5, latent_dim=128, n_prototypes=64, batch_size=512, prototype_temp=1. We have repeated the baseline results in the corresponding rows of each hyperparameter tuning group to make the comparison more intuitive. All experiments are conducted with 200 epochs and AdamW learning rate 0.0001. Note that although to fully identify g and $\boldsymbol{M}$ we need $K_{\text{pro}} \geq d+1$ (Theorem 4), by Theorem 1, the identification of biological meaningful gene interaction structure does not depend on that. Empirical results also show that choosing a relative larger latent dimension compared to the prototype number sometimes yields better interpolation performance.

Ablation and sensitivity analysis on more datasets. We conduct loss component ablation and hyperparameter sweep experiments on more datasets (ARTISTA LOTO-2, ARTISTA LOTO-3, ARTISTA LOTO-4, ZESTA LOTO-2, ZESTA LOTO-3, ZESTA LOTO-4), shown in Fig. 9. The basic observations consist with the main results in Section 6, where the reconstruction loss is the crucial part. The sinkhorn parameter is somewhat nuanced: for time points with significant deformation, a more relaxed non-balanced OT is required, whereas for periods when the tissue structure is relatively stable, an balanced OT often performs better.

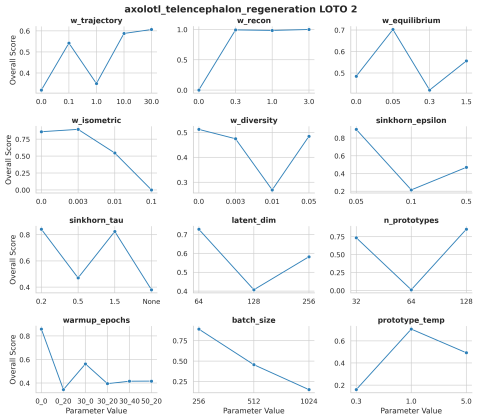
Why we do not include an explicit kinetic regularizer. One thing that we want to emphasize is that we operationalize Benamou–Brenier via sinkhorn OT, as by the Benamou–Brenier theorem, static OT with squared Euclidean cost coincides with the Benamou–Brenier minimizer along displacement

Table 5: Details on ablation and hyperparameter sweep results on ARTISTA-LOTO4. The value of warm-up epochs is denoted as $T_1_T_2$, where T_1 is the decoder warm-up epochs and T_2 is the feedback warm-up epochs.

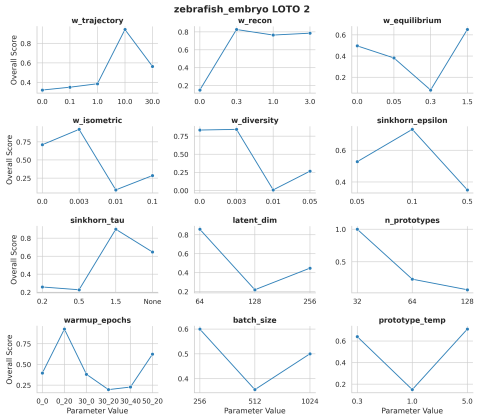
param_name	value	X_PCC↑	X_WD↓	S_ARI↑	S_PC1↑	overall score↑
batch_size	256	0.683	0.040	0.450	0.869	0.315
batch_size	512	0.698	0.040	0.468	0.869	0.677
batch_size	1024	0.707	0.041	0.437	0.871	0.500
w_diversity	0	0.690	0.040	0.448	0.869	0.348
w_diversity	0.003	0.684	0.040	0.446	0.869	0.211
w_diversity	0.01	0.698	0.040	0.468	0.869	0.848
w_diversity	0.05	0.698	0.041	0.462	0.870	0.754
w_equilibrium	0	0.698	0.040	0.466	0.869	0.181
w_equilibrium	0.05	0.698	0.040	0.471	0.869	0.461
w_equilibrium	0.3	0.698	0.040	0.468	0.869	0.354
w_equilibrium	1.5	0.701	0.041	0.468	0.869	0.634
w_isometric	0	0.687	0.039	0.464	0.870	0.578
w_isometric	0.003	0.688	0.041	0.455	0.869	0.284
w_isometric	0.01	0.698	0.040	0.468	0.869	0.600
w_isometric	0.1	0.731	0.042	0.451	0.868	0.350
w_recon	0	0.061	0.119	0.117	0.078	0.000
w_recon	0.3	0.723	0.041	0.458	0.871	0.989
w_recon	1	0.698	0.040	0.468	0.869	0.986
w_recon	3	0.684	0.040	0.466	0.869	0.977
w_trajectory	0	0.696	0.041	0.464	0.869	0.468
w_trajectory	0.1	0.699	0.041	0.454	0.869	0.426
w_trajectory	1	0.698	0.040	0.468	0.869	0.609
w_trajectory	3	0.698	0.041	0.457	0.869	0.441
w_trajectory	10	0.699	0.041	0.467	0.869	0.688
w_trajectory	30	0.695	0.041	0.477	0.869	0.381
latent_dim	64	0.699	0.040	0.449	0.870	0.628
latent_dim	128	0.698	0.040	0.468	0.869	0.607
latent_dim	256	0.697	0.041	0.473	0.870	0.483
n_prototypes	32	0.708	0.041	0.445	0.868	0.596
n_prototypes	64	0.698	0.040	0.468	0.869	0.683
n_prototypes	128	0.697	0.041	0.431	0.867	0.031
prototype_temp	0.3	0.700	0.040	0.466	0.869	0.341
prototype_temp	1	0.698	0.040	0.468	0.869	0.575
prototype_temp	5	0.702	0.042	0.466	0.869	0.500
sinkhorn_epsilon	0.05	0.698	0.040	0.453	0.869	0.185
sinkhorn_epsilon	0.1	0.698	0.040	0.468	0.869	0.377
sinkhorn_epsilon	0.5	0.701	0.040	0.466	0.869	0.798
sinkhorn_tau	0.2	0.697	0.041	0.473	0.869	0.350
sinkhorn_tau	0.5	0.698	0.040	0.468	0.869	0.353
sinkhorn_tau	1.5	0.701	0.039	0.472	0.872	0.607
sinkhorn_tau	None	0.733	0.038	0.434	0.872	0.624
warmup_epochs	0_0	0.683	0.041	0.445	0.870	0.126
warmup_epochs	0_20	0.688	0.040	0.465	0.872	0.632
warmup_epochs	30_0	0.698	0.041	0.463	0.869	0.647
warmup_epochs	30_20	0.698	0.040	0.468	0.869	0.707
warmup_epochs	30_40	0.698	0.040	0.445	0.869	0.465
warmup_epochs	50_20	0.697	0.041	0.478	0.870	0.716

Table 6: Details on ablation and hyperparameter sweep results on ZESTA-LOTO3. The value of warm-up epochs is denoted as $T_1_T_2$, where T_1 is the decoder warm-up epochs and T_2 is the feedback warm-up epochs.

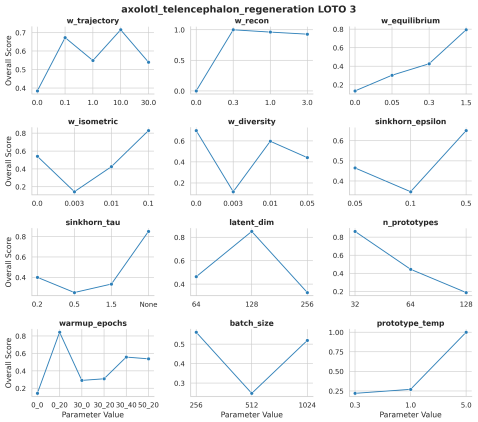
param_name	value	X_PCC↑	X_WD↓	S_ARI↑	S_PC1↑	overall score↑
batch_size	256	0.937	0.146	0.096	0.850	0.350
batch_size	512	0.947	0.142	0.068	0.853	0.610
batch_size	1024	0.946	0.141	0.069	0.854	0.610
w_diversity	0	0.938	0.143	0.062	0.855	0.150
w_diversity	0.003	0.942	0.143	0.090	0.854	0.588
w_diversity	0.01	0.947	0.142	0.068	0.853	0.550
w_diversity	0.05	0.941	0.141	0.063	0.853	0.253
w_equilibrium	0	0.946	0.143	0.067	0.853	0.526
w_equilibrium	0.05	0.946	0.144	0.059	0.853	0.346
w_equilibrium	0.3	0.947	0.142	0.068	0.853	0.754
w_equilibrium	1.5	0.944	0.147	0.073	0.854	0.500
w_isometric	0	0.941	0.143	0.060	0.855	0.331
w_isometric	0.003	0.944	0.148	0.065	0.854	0.315
w_isometric	0.01	0.947	0.142	0.068	0.853	0.574
w_isometric	0.1	0.939	0.141	0.090	0.854	0.557
w_recon	0	0.009	0.131	0.007	0.046	0.150
w_recon	0.3	0.942	0.143	0.059	0.854	0.772
w_recon	1	0.947	0.142	0.068	0.853	0.824
w_recon	3	0.944	0.145	0.080	0.854	0.849
w_trajectory	0	0.949	0.142	0.061	0.850	0.476
w_trajectory	0.1	0.945	0.145	0.057	0.853	0.230
w_trajectory	1	0.947	0.142	0.068	0.853	0.562
w_trajectory	3	0.946	0.143	0.067	0.853	0.433
w_trajectory	10	0.948	0.141	0.070	0.854	0.663
w_trajectory	30	0.951	0.136	0.045	0.856	0.650
latent_dim	64	0.932	0.146	0.063	0.853	0.000
latent_dim	128	0.947	0.142	0.068	0.853	0.529
latent_dim	256	0.948	0.140	0.088	0.853	1.000
n_prototypes	32	0.941	0.144	0.082	0.855	0.676
n_prototypes	64	0.947	0.142	0.068	0.853	0.500
n_prototypes	128	0.936	0.146	0.072	0.856	0.247
prototype_temp	0.3	0.947	0.146	0.074	0.854	0.735
prototype_temp	1	0.947	0.142	0.068	0.853	0.625
prototype_temp	5	0.945	0.147	0.064	0.853	0.040
sinkhorn_epsilon	0.05	0.946	0.145	0.067	0.853	0.359
sinkhorn_epsilon	0.1	0.947	0.142	0.068	0.853	0.992
sinkhorn_epsilon	0.5	0.945	0.142	0.064	0.853	0.150
sinkhorn_tau	0.2	0.950	0.140	0.067	0.851	0.258
sinkhorn_tau	0.5	0.947	0.142	0.068	0.853	0.119
sinkhorn_tau	1.5	0.952	0.141	0.069	0.856	0.568
sinkhorn_tau	None	0.956	0.142	0.073	0.851	0.700
warmup_epochs	0_0	0.938	0.144	0.065	0.856	0.297
warmup_epochs	0_20	0.940	0.143	0.074	0.854	0.547
warmup_epochs	30_0	0.946	0.144	0.073	0.853	0.695
warmup_epochs	30_20	0.947	0.142	0.068	0.853	0.637
warmup_epochs	30_40	0.947	0.147	0.074	0.854	0.687
warmup_epochs	50_20	0.948	0.141	0.062	0.854	0.558



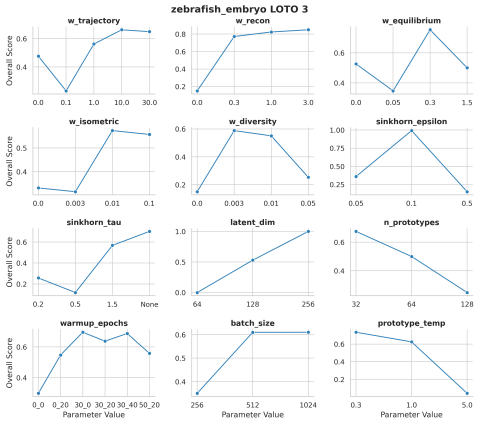
(a) ARTISTA LOTO-2



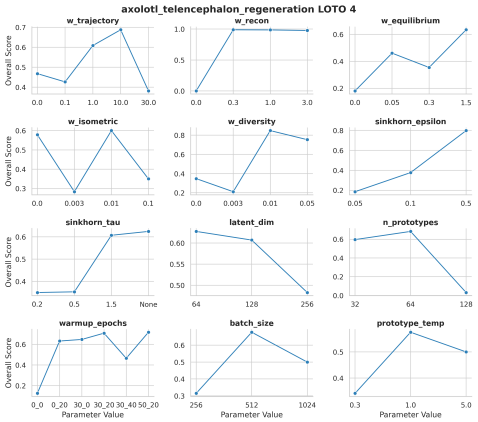
(b) ZESTA LOTO-2



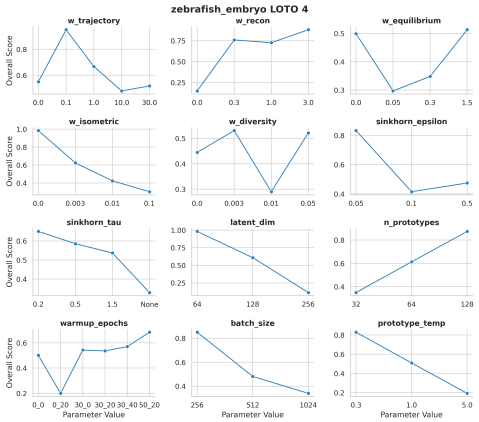
(c) ARTISTA LOTO-3



(d) ZESTA LOTO-3



(e) ARTISTA LOTO-4



(f) ZESTA LOTO-4

Figure 9: Hyperparameters ablation and sensitivity analysis on more settings. The overall scores are min-max normalized across each hyperparameter group in each LOTO setting and then aggregated, to better show how each hyperparameter affect the performance.

interpolation [23]; combined with the smoothness inductive bias of neural ODE parametrization, the learned velocity field is implicitly biased toward low-kinetic-energy solutions. Adding an explicit kinetic regularizer on the L2 norm $\|\boldsymbol{v}\|^2$ yields negligible improvement and even undermine the performance, and hence is unnecessary (Table 7).

Table 7: Effects of additional regularization on squared L2 kinetics. Column Strength represents the weight of the regularization. All the remaining hyperparameters follow the baseline setting.

dataset	loto_idx	w_velocity	X_PCC	X_WD	S_ARI	S_PC1
ZESTA	2	0	0.916	0.171	0.110	0.696
ZESTA	2	0.001	0.906	0.171	0.109	0.675
ZESTA	2	0.01	0.781	0.174	0.113	0.355
ZESTA	2	0.1	0.753	0.174	0.090	0.466
ZESTA	3	0	0.947	0.142	0.068	0.853
ZESTA	3	0.001	0.946	0.145	0.071	0.853
ZESTA	3	0.01	0.948	0.139	0.084	0.852
ZESTA	3	0.1	0.948	0.140	0.084	0.851
ZESTA	4	0	0.952	0.221	0.035	0.881
ZESTA	4	0.001	0.947	0.220	0.076	0.879
ZESTA	4	0.01	0.893	0.211	0.054	0.880
ZESTA	4	0.1	0.884	0.213	0.074	0.878
ARTISTA	2	0	0.975	0.025	0.423	0.828
ARTISTA	2	0.001	0.975	0.025	0.372	0.828
ARTISTA	2	0.01	0.975	0.025	0.416	0.828
ARTISTA	2	0.1	0.975	0.025	0.375	0.828
ARTISTA	3	0	0.944	0.039	0.392	0.851
ARTISTA	3	0.001	0.941	0.038	0.403	0.850
ARTISTA	3	0.01	0.943	0.039	0.382	0.850
ARTISTA	3	0.1	0.942	0.039	0.386	0.850
ARTISTA	4	0	0.698	0.040	0.466	0.869
ARTISTA	4	0.001	0.699	0.041	0.457	0.869
ARTISTA	4	0.01	0.700	0.041	0.467	0.868
ARTISTA	4	0.1	0.701	0.042	0.458	0.869

D.3 Performance on time-series single-cell transcriptomics without spatial information

For single-cell time series data without spatial locations available, we can remove the spatial decoder from the joint observable Φ_γ in Equation (2), i.e., setting $\gamma \to 0$. Assumption 1 then reduces to the unimodal isometric embedding $J_{\phi_x}^\top J_{\phi_x} = \eta^2 \mathbf{I}_d$, and the GRN of Equation (8) reduces to $J_{\phi_x}(-\mathbf{B})J_{\phi_x}^\top/\eta^2$, without the spatial precision adjustment. Theorems 1 and 4 apply verbatim. scIFLF then applies directly to time-series scRNA-seq data.

We ran the identical model without the spatial modality on the same LOTO protocol (LOTO-2/3/4), and compare it against the full model, the PrevCopy baseline, and TrajectoryNet [3] as an external non-spatial reference. TrajectoryNet was run at its recommended settings, and PrevCopy was re-run inside the gene-only pipeline. Table 8 reports per-dataset means over the three held-out time points. S_ARI and S_PC1 are computed by projecting the predicted expression onto the true spatial coordinates, as in Appendix C.3.

Averaged over all 9 settings, adding spatial information improves all four metrics, so the spatial modality also improves gene-level inference. The full model also outperforms TrajectoryNet on all four metrics, while the gene-only variant is comparable to it (better X_WD, S_ARI and S_PC1, lower X_PCC).

D.4 Additional Gene-Wise Spatial Metrics

To assess gene-wise and spatial-pattern recovery more directly than the metrics of Appendix C.3, we re-ran the experiments on the same LOTO settings (hence slightly different values from Table 4) and

Table 8: Gene-only ($\gamma \to 0$) variant against the full model and non-spatial references: per-dataset means over the three held-out time points, and the overall mean over all 9 settings. Best value per dataset in bold.

dataset	model	X_PCC↑	X_WD↓	S_ARI↑	S_PC1↑
MOSTA	full (with space)	**0.943**	0.084	**0.356**	**0.896**
MOSTA	gene-only	0.930	**0.073**	0.332	0.881
MOSTA	TrajectoryNet	0.927	0.088	0.342	0.883
MOSTA	PrevCopy	0.878	0.090	0.218	0.604
ZESTA	full (with space)	**0.946**	**0.173**	**0.084**	**0.861**
ZESTA	gene-only	0.884	0.188	0.070	0.727
ZESTA	TrajectoryNet	0.938	0.194	0.063	0.706
ZESTA	PrevCopy	0.850	0.184	0.010	0.133
ARTISTA	full (with space)	**0.895**	**0.034**	**0.439**	**0.851**
ARTISTA	gene-only	0.854	0.035	0.373	0.833
ARTISTA	TrajectoryNet	0.872	0.038	0.359	0.831
ARTISTA	PrevCopy	0.853	0.039	0.216	0.255
Overall (9 settings)	full (with space)	**0.928**	**0.097**	**0.293**	**0.869**
Overall (9 settings)	gene-only	0.889	0.099	0.259	0.813
Overall (9 settings)	TrajectoryNet	0.912	0.107	0.254	0.807
Overall (9 settings)	PrevCopy	0.860	0.104	0.148	0.330

report two additional gene-wise spatial metrics together with the original gene expression prediction metrics.

Mean gene-wise spatial PCC (↑). Let (s_i^p, x_i^p) denote the predicted expression profiles with their spatial support, and (s_j^t, x_j^t) the observed target cells. Since the predicted and observed time points contain different numbers of cells, at each observed target coordinate s_j^t we estimate the predicted expression by inverse-distance-weighted interpolation from its $k = 20$ nearest predicted-support coordinates,

$$\hat{x}_{jg} = \sum_{i \in N_k(j)} w_{ji}\, x_{ig}^p, \qquad w_{ji} = \frac{(d_{ji} + \epsilon)^{-1}}{\sum_{\ell \in N_k(j)} (d_{j\ell} + \epsilon)^{-1}}, \tag{34}$$

where $d_{ji} = \|s_j^t - s_i^p\|$. We then compute $\mathrm{PCC}_g = \mathrm{corr}_j(\hat{x}_{jg}, x_{jg}^t)$ across target locations for each highly variable gene and report the mean across genes. Only coordinates are used for interpolation, which accommodates unequal numbers of predicted and observed cells.

Gene-wise spatial autocorrelation agreement (Moran's I agreement, ↑). We compute Moran's I for each highly variable gene on the predicted expression with its spatial support, independently compute the true Moran's I values from the observed expression and coordinates at the target time point, and report the Spearman correlation between the predicted and true gene-wise Moran's I values.

For scIFLF, the spatial support is the predicted target-time coordinates. For Moscot, Spateo, and STORIES, whose predicted profiles remain attached to the source particles, it is the source-particle coordinates. All spatial methods share the same interpolation rule, coordinate preprocessing, highly variable genes, and k.

Among the spatial methods (Table 9), scIFLF achieves the best average mean gene-wise spatial PCC, spatial-autocorrelation agreement, and expression WD on both datasets, with comparable mean-profile PCC. The ranking is unchanged for $k = 10$ and $k = 50$. The absolute gene-wise PCC values are low for all methods, as expected given dropout and the absence of one-to-one cell correspondence.

E Cross-Seed GRN Consistency: Feedback dynamics and GRN identifiablity

Identifiable feedback structure ensures robust and reliable GRN inference. A central claim of scIFLF is that the inferred gene-regulatory Jacobian $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$ is absolutely identifiable as a relation

Table 9: Additional metrics on ZESTA and ARTISTA, as means over the three LOTO settings (LOTO-2/3/4). The experiments were re-run for this evaluation, so mean-profile PCC and expression WD differ slightly from Table 4. Best value per dataset in bold. "–" marks spatial metrics that are not defined for expression-only methods.

Dataset	Method	Mean-profile PCC↑	Expression WD↓	Mean gene-wise spatial PCC↑	Moran's I agreement↑
ZESTA	Ours	0.9323	0.1897	**0.0876**	**0.3707**
ZESTA	Moscot	0.8856	0.1900	0.0685	0.1803
ZESTA	Spateo	0.8814	0.1899	0.0698	0.1784
ZESTA	STORIES	**0.9447**	0.1950	0.0725	0.1156
ZESTA	Ours (gene-only)	0.8836	**0.1875**	–	–
ZESTA	TrajectoryNet	0.9380	0.1937	–	–
ARTISTA	Ours	**0.8710**	**0.0350**	**0.0066**	**0.5697**
ARTISTA	Moscot	0.8292	0.0372	0.0010	0.3786
ARTISTA	Spateo	0.8590	0.0375	0.0010	0.3813
ARTISTA	STORIES	0.8660	0.0357	0.0019	0.3790
ARTISTA	Ours (gene-only)	0.8498	0.0367	–	–
ARTISTA	TrajectoryNet	0.8702	0.0382	–	–

between observable quantities (Section 4.2), invariant to the residual latent ambiguity $(\mathbf{Q}, \lambda, \boldsymbol{m})$. We empirically validate this claim by assessing the **cross-seed consistency** of the inferred GRN: if the framework is truly identifiable, independent training runs with different random seeds, which explore different latent gauges, should converge to the same gene-regulatory structure up to the invariant quantities.

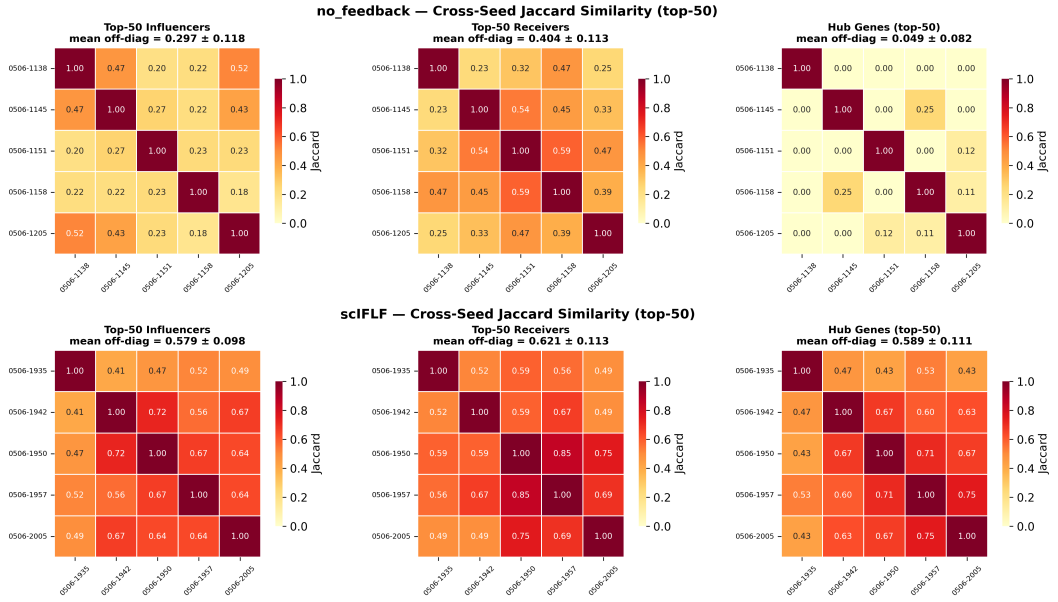


Figure 10: Cross-seed Jaccard similarity of top-50 influencer, receiver, and hub genes (those genes that are both top-50 influencers and receivers) across five independent training runs on the ZESTA dataset. Upper row: no-feedback ablation. Lower row: scIFLF. Mean ± std of off-diagonal entries are shown above each heatmap. scIFLF achieves substantially higher cross-seed consistency across all three gene categories, with the most pronounced improvement on hub genes ($0.049 \rightarrow 0.589$), empirically validating the identifiability guarantees of Theorem 1.

We train both variants with five different random seeds on the ZESTA dataset and measure the cross-seed Jaccard similarity (the ratio of the intersection size of sets obtained in experiments with different seeds to the union size) of the inferred top-50 influencer, receiver, and hub gene sets (Fig. 10). For scIFLF, the GRN is evaluated via Equation (8) using the closed-form $J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top$. In practice, we compute $(J_{\phi_x}^\top J_{\phi_x})^{-1}$ via truncated eigendecomposition, retaining only eigenvectors with eigenvalues above $\tau_{\text{rel}} = 0.1$ of the maximum, to suppress numerically unreliable low-energy directions. For the no-feedback ablation ($\boldsymbol{v} = g_\phi(\boldsymbol{Z})$), since there is no longer a closed-form expression of $\mathbf{B}$ available, the GRN Jacobian is evaluated directly via $\partial \boldsymbol{v}_X / \partial \boldsymbol{X} \approx J_{\phi_x} \cdot J_{g_\phi}(\bar{\boldsymbol{Z}}) \cdot$

$J_\Psi(\bar{\boldsymbol{X}})$, where $J_{g_\phi}(\bar{\boldsymbol{Z}})$ is the finite-difference Jacobian of the drift g_ϕ at the population mean latent state $\bar{\boldsymbol{Z}}$, and $J_\Psi(\bar{\boldsymbol{X}})$ is the autograd encoder Jacobian at the mean gene expression $\bar{\boldsymbol{X}}$. scIFLF achieves mean cross-seed Jaccard of 0.579, 0.621, and 0.589 for influencer, receiver, and hub genes respectively, versus 0.297, 0.404, and 0.049 for the no-feedback ablation. The gap is most pronounced for hub genes ($+0.540$), which require simultaneous consistency in both outgoing and incoming regulatory strength and thus constitute the strictest test of GRN reproducibility. These results confirm that $\mathbf{B}$ encodes a canonical gene-interaction structure that is substantially more stable across random initializations than the unconstrained drift $g_\phi(\boldsymbol{Z})$, providing concrete empirical grounding for the identifiability guarantees of Theorem 1, illustrating that the latent feedback structure effectively encodes the modular interaction information and enables reliable and robust GRN inference.

The structure prior achieves biological interpretibility without compromising performance. Since the feedback structure $\mathbf{B}(\boldsymbol{M}(\boldsymbol{Z}) - \boldsymbol{Z})$ in the latent dynamics is introduced primarily for biophysical interpretability, we are curious about whether this prior structure will weaken predictive performance. To assess this, we compare scIFLF against a stripped-down variant in which the feedback term is removed (i.e., a pure neural ODE drift $\boldsymbol{v} = g_\phi(\boldsymbol{Z})$ within the same isometric backbone, Table 10). The two variants achieve comparable interpolation performance, with feedback occasionally bringing moderate improvements (notably on ZESTA). This indicates that the feedback structural prior **does not trade predictive fidelity for interpretability**.

Integrating all the observations, it turns out that what feedback contributes is a **closed-form readout of the gene-regulatory structure**. With feedback, the population-average latent Jacobian is exactly $-\mathbf{B}$ (Assumption 4), which gives the GRN readout $J_{\phi_x} \cdot (-\mathbf{B}) \cdot (J_{\phi_x}^\top J_{\phi_x})^{-1} \cdot J_{\phi_x}^\top$, where $\mathbf{B}$ is a learnable $d \times d$ matrix with a clear biophysical meaning: the average linear response of the gene programs over the observed cell population (Section 4.2, Equation 8). Without feedback, the latent Jacobian is $J_g(\boldsymbol{Z})$, which varies continuously across cell states, requires pointwise autograd evaluation, and lacks a canonical baseline summary. Both formulations admit identifiable observable Jacobians $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$, but only the feedback formulation exposes the gene-regulatory structure as a fixed, low-dimensional, biologically interpretable object. This complements the identifiability theory: isometry and Benamou–Brenier dynamics secure identifiability of the observable quantities, while the feedback structure secures their interpretability as canonical biological invariants, ensuring reliable and robust GRN inference.

Table 10: The effects of feedback term in interpolation performance

dataset	loto_idx	method	X_PCC	X_WD	S_ARI	S_PC1
ZESTA	2	scIFLF	0.916	0.171	0.110	0.696
ZESTA	2	no_feedback	0.910	0.174	0.099	0.644
ZESTA	3	scIFLF	0.947	0.142	0.068	0.853
ZESTA	3	no_feedback	0.947	0.146	0.068	0.853
ZESTA	4	scIFLF	0.952	0.221	0.035	0.881
ZESTA	4	no_feedback	0.948	0.217	0.038	0.881
ARTISTA	2	scIFLF	0.975	0.025	0.423	0.828
ARTISTA	2	no_feedback	0.975	0.025	0.420	0.828
ARTISTA	3	scIFLF	0.944	0.039	0.392	0.851
ARTISTA	3	no_feedback	0.944	0.038	0.402	0.850
ARTISTA	4	scIFLF	0.698	0.040	0.466	0.869
ARTISTA	4	no_feedback	0.698	0.040	0.458	0.869

F Extended Biological Analysis

F.1 Gene Regulatory Network Extraction

For each dataset we extract the approximate GRN Jacobian $\partial \boldsymbol{v}_X / \partial \boldsymbol{X}$ by Equation 8 at the median time point, where the inverse $(J_{\phi_x}^\top J_{\phi_x})^{-1}$ is computed via truncated eigendecomposition retaining eigenvectors with eigenvalues above $\tau_{\text{rel}} = 0.1$ of the maximum, suppressing numerically unreliable low-energy directions. We characterize each gene along two axes: outgoing influence (column norm, measuring regulatory output) and incoming influence (row norm, measuring regulatory input). Genes

in the high-outgoing / low-incoming quadrant act as master regulators; genes in the high-incoming / low-outgoing quadrant act as downstream targets; genes prominent in both (as the top 50 influencers and receivers simultaneously) are hub genes. The Gene Ontology (GO) enrichment analysis is based on Fisher exact hypergeometric test, with a threshold 0.05 on the Benjamini-Hochberg False Discovery Rate (FDR) adjusted P-values, implemented with Python library `gprofiler`.

F.2 Mouse Embryo

Shown in Fig. 11, the inferred regulatory landscape of mouse embryogenesis is organized around three coherent programs that span the E9.5–E16.5 organogenesis window covered by MOSTA. The top influencer and receiver sets are dominated by **embryonic and fetal hemoglobins** (*Hba-a1/a2/x*, *Hbb-bs/bt/y/bh1*), reflecting the active erythropoietic program that proceeds through yolk-sac and fetal-liver stages over this window. Alongside these, classical **imprinted growth and patterning genes** (*H19*, *Igf2*, *Peg3*, *Mest*, *Dlk1*, *Meg3*, *Nnat*) and **ECM/cytoskeletal components** (*Col1a1*, *Col1a2*, *Col3a1*, *Tuba1a*, *Tmsb4x/Tmsb10*) feature prominently as hub genes, capturing the imprinted regulatory machinery and tissue-remodeling apparatus central to mammalian organogenesis [38, 45, 46]. *Ccnd2*, a key cell-cycle driver, appears as a top hub, consistent with the high proliferative rate of the developing embryo. The chromatin regulator *Jarid2* and the inhibitory-synapse scaffold *Gphn* are also recovered as strong hubs [38, 39].

A modest but informative **influencer–receiver asymmetry** distinguishes mouse embryogenesis from the zebrafish and axolotl datasets: while top influencers and receivers overlap substantially in the hub-gene set, several genes are predominantly receivers (*Hbb-bt*, *Col1a2*, *H19*, *Igf2*, *Col3a1*, *Col1a1*) versus predominantly influencers (*Peg3*, *Mest*, *Malat1*, *Gpc6*). Correspondingly, GO enrichment of top-100 influencers recovers broad development and morphogenesis terms (system/tissue/anatomical structure development, cell migration, muscle development), whereas top-100 receivers are additionally enriched for **nervous system development, neurogenesis, neuron differentiation, and neuron projection development**. This directional signal indicates that the broad developmental machinery drives neural lineage specification during this window. Most top influencers exhibit a balanced activation/repression profile, with embryonic β-globins (*Hbb-y*, *Hbb-bs*) and *Afp* biased toward activation and the lncRNAs (*Gm42418*, *Malat1*) biased toward repression.

F.3 Axolotl Telencephalon Regeneration

The inferred regulatory landscape of axolotl telencephalon regeneration is dominated by **radial glia / ependymoglia markers**, which are the cellular substrate of the regenerative program [36]. The top influencers include *FABP7* (a fatty-acid binding protein selectively expressed in radial glia and a canonical neural stem-cell marker), *SLC1A3* (GLAST/EAAT1, the glial glutamate transporter expressed by radial glia and reactive astrocytes, crucial for brain development and neurogenesis), *GLUL.L* (glutamine synthetase, another radial glia marker), and *VIM* (vimentin, a hallmark intermediate filament of progenitor and de-differentiated cells), together with *SCGN* (secretagogin, a calcium sensor expressed in interneurons) and *HB-B* (hemoglobin-β, reflecting the active vascular remodeling of the regenerating tissue). Notably, *FABP7* acts as a strong **net repressor** (activation fraction < 0.3, Fig. 12h), consistent with its role in maintaining the progenitor state by suppressing downstream differentiation programs [47].

A key feature of the regenerative regulatory network is the **strong overlap between top influencers and top receivers**: hub genes such as *FABP7*, *SLC1A3*, *SCGN*, *SFRP1* (a Wnt antagonist implicated in stem-cell maintenance and neuroprotection), and *HB-B* simultaneously rank in both top sets, indicating that the regenerative program is highly self-reinforcing rather than feed-forward. This is corroborated by GO enrichment: both top-100 influencers and top-100 receivers are dominated by the **translation machinery**, such as eukaryotic translation initiation, elongation and termination, 40S/60S ribosomal subunit assembly, NMD, SRP-dependent protein targeting, and cellular response to amino-acid starvation (*EIF2AK4*/GCN2 axis), reflecting the massive translational reprogramming that radial glia undergo during de-differentiation and re-entry into the cell cycle. Additionally, **ROBO/SLIT signaling** appears prominently in both enrichment panels, consistent with the axon guidance and progenitor-positioning cues required to rebuild telencephalic architecture (Fig. 12). Together, these results paint a coherent picture of axolotl brain regeneration as a radial-glia-driven, translationally hyper-active, self-reinforcing program, closely matching the mechanistic model of telencephalon regeneration established in the original ARTISTA study [36].

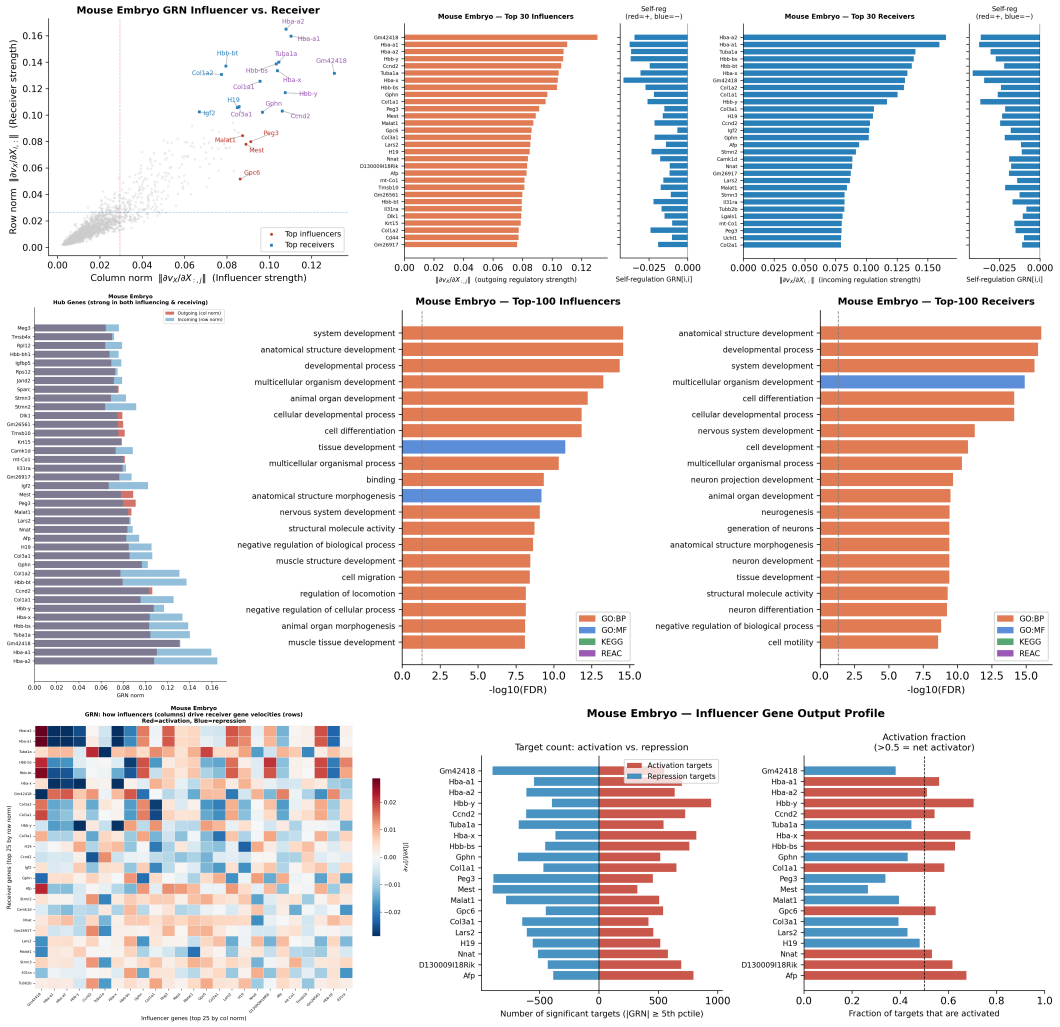


Figure 11: Biological analysis on Mouse Embryo Development. (a) Scatter plot of all genes by outgoing (column norm) vs. incoming (row norm) regulatory strength, with top influencers (red), receivers (blue), and hub genes (purple) labeled. (b) Top-30 influencer genes ranked by column norm, with self-regulation scores. (c) Top-30 receiver genes ranked by row norm, with self-regulation scores. (d) Hub genes ranked by the product of outgoing and incoming norms. (e) GO/KEGG enrichment (FDR < 0.05) for top-100 influencer genes. (f) GO/KEGG enrichment for top-100 receiver genes. (g) GRN submatrix heatmap of top-25 influencers (columns) × top-25 receivers (rows); red = activation, blue = repression. (h) Activation and repression target counts and fractions for each top influencer gene.

F.4 Zebrafish Embryo Development

The inferred regulatory landscape of zebrafish embryogenesis is dominated by the **biosynthetic and bioenergetic machinery** that supports rapid embryonic proliferation, together with epithelial structural components and yolk-lipid metabolism. Top influencers include keratins (*cyt1*, *krt4*, *krt5*, *krt18a.1*, the major epithelial intermediate filaments forming the enveloping layer of the early embryo), the stress-response chaperone *hspb1*, ribosomal proteins (*rplp0*, *rpsa*, *rpl15*, *rps3/5*), and apolipoproteins (*apoa1a*, *apoa2*, *apoc1*, *apoeb*, *apobb.1*) reflecting the active mobilization of yolk lipids that fuels early embryogenesis. Notably, the canonical mesoderm transcription factors *tbxta* (T-box, notochord/mesoderm specification) and *msgn1* (presomitic mesoderm segmentation) [48] also appear among the top influencers and hub genes, indicating that the framework recovers known devel-

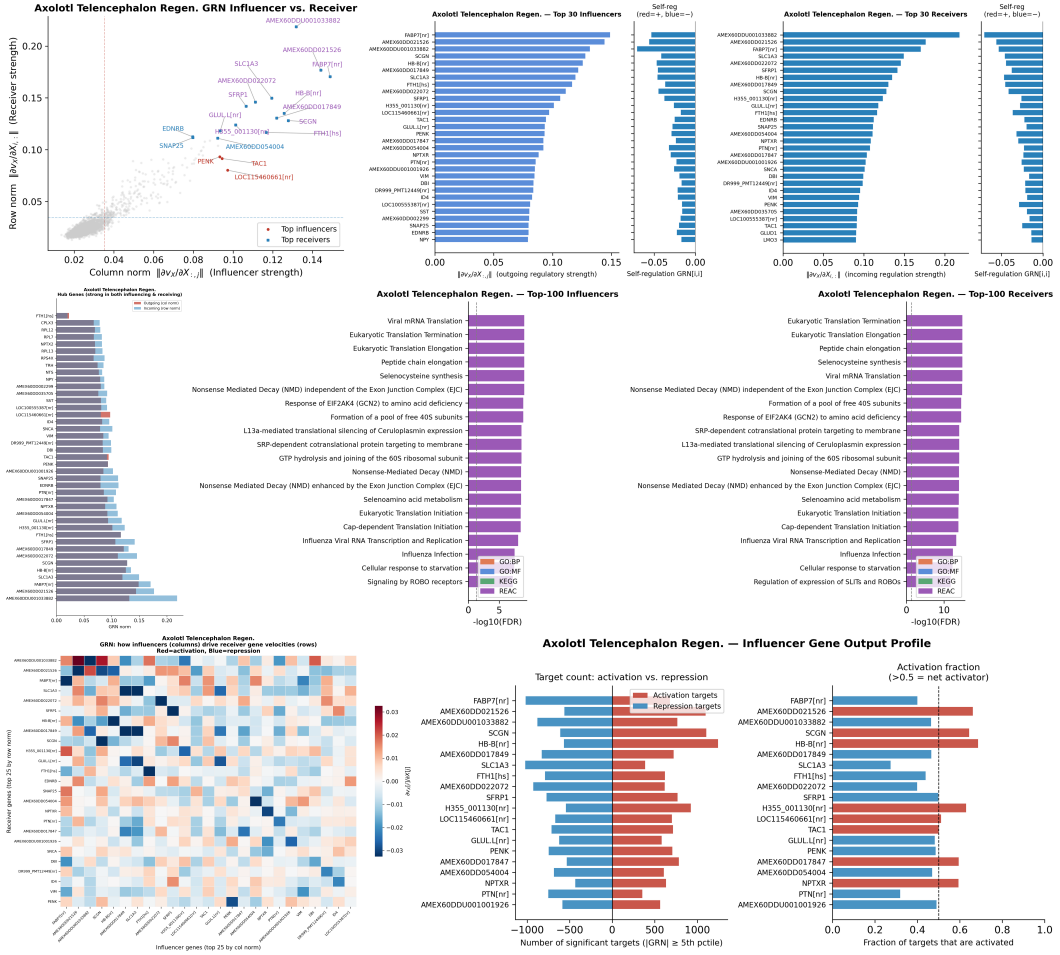


Figure 12: Biological analysis on Axolotl Telencephalon Regeneration. (a) Scatter plot of all genes by outgoing (column norm) vs. incoming (row norm) regulatory strength, with top influencers (red), receivers (blue), and hub genes (purple) labeled. (b) Top-30 influencer genes ranked by column norm, with self-regulation scores. (c) Top-30 receiver genes ranked by row norm, with self-regulation scores. (d) Hub genes ranked by the product of outgoing and incoming norms. (e) GO/KEGG enrichment (FDR < 0.05) for top-100 influencer genes. (f) GO/KEGG enrichment for top-100 receiver genes. (g) GRN submatrix heatmap of top-25 influencers (columns) × top-25 receivers (rows); red = activation, blue = repression. (h) Activation and repression target counts and fractions for each top influencer gene.

opmental drivers despite their lower magnitude in the GRN compared to the dominant biosynthetic program.

We can similarly observe the strong overlap between top influencers and top receivers: hub genes such as *hspb1*, *krt4*, *cyt1*, *rplp0*, *rpsa*, ribosomal proteins, and apolipoproteins simultaneously rank in both top sets, indicating a self-reinforcing rather than feed-forward regulatory program—similar to what we observe in axolotl regeneration but distinct from the asymmetric cascade in mouse embryogenesis. This is corroborated by GO enrichment: both top-100 influencers and top-100 receivers are dominated by **ribosome biogenesis, translation, ATP biosynthesis, oxidative phosphorylation, and proton motive force-driven ATP synthesis**, reflecting the high translational and bioenergetic demand of rapid cell proliferation during early somitogenesis. Influencer enrichment additionally surfaces **lipid transport and lipoprotein metabolic processes**, consistent with the dependence of early zebrafish embryos on yolk-derived lipid substrates. Most top influencers exhibit a balanced or mildly biased activation profile (Fig. 13h), with the keratins (*cyt1*, *krt4*, *krt5*) acting as predominant repressors and

several ribosomal/translation-associated factors as net activators, suggesting differential roles for structural maintenance versus biosynthetic ramp-up.



Figure 13: Biological analysis on Zebrafish Embryo Development. (a) Scatter plot of all genes by outgoing (column norm) vs. incoming (row norm) regulatory strength, with top influencers (red), receivers (blue), and hub genes (purple) labeled. (b) Top-30 influencer genes ranked by column norm, with self-regulation scores. (c) Top-30 receiver genes ranked by row norm, with self-regulation scores. (d) Hub genes ranked by the product of outgoing and incoming norms. (e) GO/KEGG enrichment (FDR < 0.05) for top-100 influencer genes. (f) GO/KEGG enrichment for top-100 receiver genes. (g) GRN submatrix heatmap of top-25 influencers (columns) × top-25 receivers (rows); red = activation, blue = repression. (h) Activation and repression target counts and fractions for each top influencer gene.

F.5 Time-Resolved GRN

The GRN reported above is evaluated at the median time point, and the inferred GRN evolves with time: $\mathbf{B}$ is constant, while the observable GRN depends on the latent state through the decoder Jacobian $J_{\phi_x}(\boldsymbol{Z})$, so evaluating Equation (8) on the cells of each observed time point gives a time-resolved GRN. On MOSTA, the top regulators drift smoothly along development (Table 11): the lower first-vs-last than adjacent-time Jaccard similarity indicates a gradual regulatory transition. The leading regulators shift in a biologically coherent way, from an early extracellular-matrix signal (*Postn*), through a dominant erythropoietic phase in mid-gestation (*Hba-a1*/*Hba-a2*), to a late connective-tissue/collagen signal (*Col1a1*). The mid-gestation erythropoietic phase coincides with the known

peak of fetal-liver erythropoiesis around E12–E14.5 [49], so the ECM → erythropoiesis → collagen progression tracks the broad course of organogenesis.

Table 11: Time-resolved GRN on MOSTA: Jaccard similarity of the top influencer and receiver gene sets between adjacent time points and between the first and last time point.

MOSTA time-resolved GRN	Influencers	Receivers
Adjacent-time Jaccard	0.389	0.425
First-vs-last Jaccard	0.235	0.299

F.6 Static-Expression Controls for the GO Analysis

Genes with large GRN norms could in principle be simply the highly expressed or highly variable genes. To separate the regulatory readout from static expression, we rerun the identical enrichment pipeline on control gene sets of the same size ranked by expression variance, by mean expression, and at random.

Broad developmental terms are shared, as expected (e.g., system development is the top term for every set), but the overlap is partial and the difference is systematic. On MOSTA, the top influencers yield 903 significant terms, against 690 for the variance-ranked, 729 for the mean-ranked, and only 35 for the random control (Table 12). The term-level Jaccard similarity between the influencer set and the variance-ranked control is 0.395. Moreover, 408 influencer terms appear in no control set. They are led by positive regulation of DNA-templated transcription (FDR 5.3×10^{-7}), DNA-binding transcription activator activity, and cis-regulatory region sequence-specific DNA binding, whereas the 130 receiver-unique terms are led by embryonic skeletal system morphogenesis, neuron projection regeneration, and postsynapse organization.

The GRN readout also carries a directional split that a static ranking lacks: the influencer and receiver sets differ (gene-level Jaccard 0.575, term-level Jaccard 0.472), and their unique terms separate transcriptional activation from neuronal and skeletal morphogenesis, where a variance ranking gives one undirected list.

Table 12: Number of significant enriched terms (FDR < 0.05) on MOSTA for the top influencer set and for static-expression control gene sets under the identical enrichment pipeline.

Dataset	Influencers	Variance-ranked	Mean-ranked	Random
MOSTA	903	690	729	35

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in the abstract and introduction are supported by the theoretical and empirical results in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical 4 and empirical results 6, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the assumptions are provided and discussed in Section 4. The complete proof and detailed remarks are shown in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We summarize the main architecture and training process in Section 5. And we provide detailed model architecture, training loss, training strategies, and hyperparameter settings in Appendix B. The source code is provided as a supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the url link for all the real spatiotranscriptomics data in Appendix C.1. The data processing is shown in Section 6. The source code is provided as a supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The training hyperparameters and optimization details are shown in Appendix B and Appendix C.3. The hyperparameters for compared methods follow the default setting and are shown in Appendix C.2. This is an unsupervised learning task. We conducted leave-one-out benchmark that we take one slice at an intermediate point as evaluation data and train the model on the remaining slices and interpolate the leave-out slice via learned dynamics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: This is an unsupervised learning task, but we conducted leave-one-out benchmark that we take one slice at an intermediate point as evaluation data and train the model on the remaining slices and interpolate the leave-out slice via learned dynamics. The downstream biological gene enrichment analysis is based on Fisher exact hypergeometric test, with a threshold 0.05 on the Benjamini-Hochberg False Discovery Rate (FDR) adjusted P-values. We also conducted cross-seed agreement analysis on the inferred GRN matrices.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on computer resources and running time are shown in Appendix C.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We make sure to preserve anonymity and follow other NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the broad impacts and potential limitations of our methods in biomedical and other nature sciences fields research in the introduction and discussion section, also in the remarks of our feedback-controlled dynamics model.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No such risks for our method and data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All the codes for comparing methods are open sourced, public available and have been cited in the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A]

    Justification: No new datasets or pre-trained models are released with this submission. The source codes are included in the anonymous supplement where licensing permits.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs only used for polishing writing and help writing some auxiliary codes like data-processing, with human review. It does not affect the core methodology scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.