CycleGRN: Inferring Gene Regulatory Networks from Cyclic Flow Dynamics in Single-Cell RNA-seq

Wenjun Zhao, Elana J. Fertig, Genevieve Stein-O'Brien

February 27, 2026

Abstract

Oscillatory processes such as the cell cycle play critical roles in cell fate determination and disease development, yet existing gene regulatory network (GRN) inference methods often fail to account for their dynamic nature. We propose *CycleGRN*, a novel framework that treats cell cycle gene expression observations as an invariant measure of a stochastic differential equation and learns from data a dynamical system that fits cycling biological processes. Using a directed graph constructed along the inferred flow field in the cell space, we estimate Lie derivatives for all genes, enabling velocity inference beyond the cell cycle subspace. To quantify regulatory interactions, we introduce a time-lagged correlation operator between any pair of genes supported on the flow-aligned directed graph, which respects the intrinsic geometry of the data manifold and allows temporal ordering consistent with the underlying oscillatory process. The method requires only raw gene expression data at single-cell resolution and a list of cycle genes, without temporal binning or splicing dynamics. We evaluate our method on four synthetic datasets generated from mechanistic models with known network structures with oscillatory subnetworks, and on a mouse retinal progenitor single-cell RNA-seq dataset spanning three cell types and a knockout condition. Across all settings, our method consistently ranks among the top-performing approaches and demonstrates strong recovery of oscillatory and directional interactions.

1 Introduction

Oscillatory processes such as the cell cycle play a central role in cell fate determination, tissue homeostasis, and disease progression, including cancer and developmental disorders [17, 27, 32]. Beyond cell cycle states, even individual pathways, such as *Notch* signaling [15], can demonstrate oscillatory dynamics. These processes induce structured temporal patterns in gene expression, often reflecting periodic or quasi-periodic dynamics. While such oscillations are readily detectable at the population level, they are typically only indirectly captured in single-cell RNA-sequencing experiments, where each cell is measured at a single time point and thus requires additional methods to infer its developmental or phase state [25].

Despite their biological importance, oscillatory processes have not been fully leveraged in current gene regulatory network (GRN) inference methods. For instance, a common practice in single-cell analysis is to treat cell cycle effects as unwanted confounders [6, 21], removing or regressing out cell cycle–associated genes prior to downstream analyses such as trajectory inference. While this can reduce technical variability, it also discards valuable information about intrinsic temporal structure that could inform regulatory relationships.

We hypothesize that by constraining GRN inference using an estimated invariant oscillatory process, we can better recover the temporal ordering of regulatory events and improve causal inference in oscillatory systems. Intuitively, the invariant process acts as a latent clock shared across cells, capturing the periodic structure

[0]Code with a simple notebook example available at https://github.com/WenjunZHAOwO/CycleGRN/tree/main

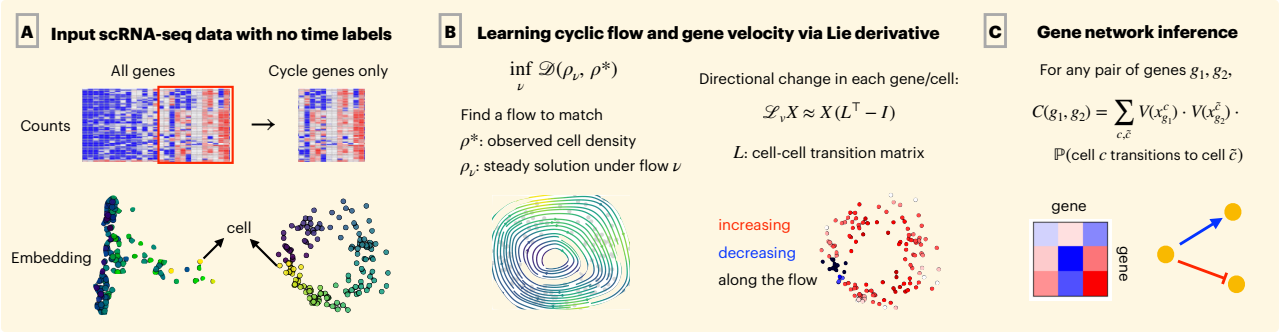

Figure 1: *Learning dynamic gene regulation with CycleGRN. (A) Input: scRNA-seq data and a list of cycle genes. (B) A continuous flow in the subspace of cycle genes is learned on this embedding to reproduce the observed cell density. Each gene's directional change along the flow is estimated. (C) Comparing these directional changes across cell transitions yields time-lagged correlations that reconstruct directed regulatory interactions.*

underlying gene expression variation. A natural approach is to leverage trajectory inference methods to learn a pseudotime for cell cycle [10]. For instance, Tricycle [31] identifies a template with clear cyclic structure, and uses transfer learning trained on large real datasets to map new datasets onto this template. Empirical evidence using such pseudotime showed improvement in accuracy [30], but required dividing data into discrete bins, which may fail to capture the continuous and periodic nature of the cell cycle.

To address this limitation, we propose to learn the latent cyclic dynamics directly from the data, without inferring explicit pseudotime. Intuitively, cycling genes trace out an orbit in expression space; we seek a vector field whose invariant distribution matches the observed point cloud of these genes [5, 28]. The learned flow defines a local velocity field that respects the cyclic structure of the underlying process. We then extend this invariant flow to all genes by computing cycle-aware directional derivatives, which estimate gene-wise rates of change along the learned flow using a flow-aligned transition operator.

Finally, we define a family of time-lagged gene-gene correlation operators built on the graph that follows the invariant flow of cells, enabling the estimation of directional dependencies between genes without explicit temporal binning or splicing information. This results in a signed, asymmetric gene–gene correlation matrix that reflects potential causal ordering along the oscillatory process. Figure 1 provides a schematic overview of the *CycleGRN* framework. We demonstrate the performance of our method on four simulated datasets with specified oscillatory subnetworks from HARISSA [26], and three real datasets of different cell types with active cell cycles [7], showing superior performance across all tasks, compared to traditional methods that do not leverage knowledge on the oscillatory process.

Our contributions. (1) We introduce a method that learns invariant cyclic dynamics directly from scRNA-seq data via PDE-constrained optimization, requiring neither time labels nor splicing information. (2) We develop a flow-aligned Lie derivative that extends the learned dynamics from cycle genes to the full transcriptome, yielding gene- and cell-specific velocity estimates. (3) We define a family of flow-aligned time-lagged correlation operators that exploit the learned dynamics to infer directional and signed gene regulatory interactions.

Scope and identifiability. CycleGRN is designed for settings in which the dominant latent dynamics form a one-dimensional cyclic manifold (topologically S^1), such as the cell cycle or other recurrent biological programs. In this regime, the invariant distribution determines the underlying flow up to a global orientation and time reparameterization. Our objective is therefore not to recover absolute biological time, but to infer flow-consistent directional influence within this identifiable equivalence class. Resolving the global orientation requires minimal external information (e.g., canonical phase markers), serving as an identifiability constraint rather than an imposed regulatory prior. We do not claim applicability to arbitrary non-cyclic or multi-scale dynamics.

2 Methodology

Given a single-cell gene expression matrix $X \in \mathbb{R}^{G \times C}$, where rows correspond to genes and columns to cells, we partition the gene set into two disjoint subsets, $G = G^{\text{inv}} \cup G^{\text{other}}$, where

- G^{inv}: genes associated with an invariant process (e.g., cell cycle or Notch signaling);
- G^{other}: all remaining genes reflecting other dynamics (e.g., differentiation or reprogramming).

Let $X^{\text{inv}} \in \mathbb{R}^{|G^{\text{inv}}| \times C}$ denote the submatrix of X restricted to G^{inv}. The ultimate goal is to use the invariant dynamics in the space of G^{inv} to constrain the temporal order of regulatory events, and infer causality that respects the cell cycle – for the full set of genes including both cell cycle genes G^{inv} and G^{other}. We describe the pipeline to achieve this in the following subsections.

2.1 Learning the dynamics of cell cycle genes

In single-cell datasets, it has been shown that sampling a population of unsynchronized cells allows the dataset to effectively capture temporal dynamics. Numerous trajectory inference methods and RNA-velocity methods have been developed to infer these dynamics [20], and shown to be successful at uncovering trajectories associated with cell fate decisions. However, these methods often rely on knowing a pre-defined initial state and are fundamentally manifold dependent, limiting the ability to infer more complex dynamics in cyclic processes that lack a unidirectional trajectory.

In the context of chaotic dynamical systems, it was first proposed by Yang et al [5, 28] that even without knowing the actual time label, one can still learn from data the dynamical system by matching the observed density and the invariant measure of a flow. Given a noisy point cloud representing the expression of the set of genes in G^{inv}, denoted as $x_c^{\text{inv}} \in \mathbb{R}^{|G^{\text{inv}}|}$ with $c = 1, ..., C$ for C cells in total, we represent the distribution of samples as a histogram over predefined bins of expression values. For each bin B, we count how many samples fall inside it and normalize by the total number of cells:

$$\rho^*(B) := \frac{1}{C} \sum_{c=1}^{C} \chi_B(x_c^{\text{inv}}), \quad \chi_B(x) = \mathbf{1}_{\{x \in B\}}. \tag{2.1}$$

We aim to construct a flow $\nu_\theta(x^{\text{inv}})$ where $\theta \in \Theta$ denotes the parameters in a simple neural network parameterizing the flow in the cell space. The optimal flow is solved through the optimization:

$$\inf_{\theta \in \Theta} \mathcal{D}(\rho_{\nu_\theta}, \rho^*), \tag{2.2}$$

where ρ^* is the density obtained from single cells representing cell cycle genes, ρ_{ν_θ} is the steady-state density induced by the flow ν_θ, and $\mathcal{D}$ is a metric that quantifies the discrepancy between two distributions, such as the quadratic Wasserstein distance and Kullback–Leibler divergence. At optima, the steady-state density ρ_{ν_θ} best matches the observed distribution of ρ^*.

We next describe how ρ_{ν_θ} is connected to our parameterized flow ν_θ. Derived from the Fokker–Planck equation of a stochastic differential equation $\mathrm{d}x_t = \nu(x_t)\mathrm{d}t + \sqrt{2D}\mathrm{d}W_t$, the steady-state solution $\rho_\nu(x)$ induced by the flow $\nu(x)$ satisfies:

$$\nabla \cdot (\rho_\nu(x)\nu(x)) = D\nabla^2 \rho_\nu(x), \tag{2.3}$$

where $\rho_\nu(x)$ represents the density that is invariant under flow $\nu(x)$, and D is a positive constant representing diffusion scale. Following the approach by [5], we solve the equation using a finite volume numerical solver, compute the gradient of the objective based on the adjoint-state method and backpropagation for the parameters in $\nu(x)$, and use gradient-based optimization algorithms to iteratively update them when solving (2.2).

Remark on existence and identifiability. The existence of a solution to (2.2) depends on the parameterization class of ν [5]. When ν is parameterized by a smooth (i.e. C^∞) neural network, a regular solution to the steady-state Fokker–Planck equation exists under mild conditions. In practice, we use a two-layer fully connected neural network with tanh activation throughout all experiments.

Beyond existence, the learned flow is not uniquely identifiable from the invariant measure alone. When the underlying dynamics lie on a one-dimensional cyclic manifold (topologically S^1), the invariant density determines the flow up to (i) global orientation reversal and (ii) time reparameterization along the cycle. Thus, additional information is required to fix the direction of traversal.

In our setting, this ambiguity is resolved using minimal biological orientation signals (e.g., canonical phase markers such as S/G2M genes [21]). This does not impose edge-level regulatory priors, but serves as an identifiability constraint analogous to selecting a temporal root in pseudotime methods. A detailed discussion of identifiability under cyclic dynamics is provided in Supplementary Section A.

2.2 Cycle-aware Lie derivatives of all genes

The learned flow $\nu(x)$ is defined only on the cycle subspace and therefore must be lifted to the full G-dimensional gene space. We approximate this lifting by constructing a cell-based operator that acts on expression profiles and captures directional changes along the flow field. Conceptually, this corresponds to computing a discrete Lie derivative of each gene's expression with respect to the flow ν.

We first construct a transition matrix $L \in \mathbb{R}_+^{C\times C}$ representing a directed K-nearest neighbor graph [3, 8]. Each entry $L(i,j)$ measures how likely cell i is to transition toward cell j following the local flow direction $\nu(x_i^{\text{inv}})$:

$$L_{ij} \;=\; \frac{\mathbf{1}\big\{\, j \in \mathcal{N}_K(i) \text{ and } \cos\!\big(x_j^{\text{inv}} - x_i^{\text{inv}},\ \nu(x_i^{\text{inv}})\big) > \delta\big\}}{\displaystyle\sum_{k\in\mathcal{N}_K(i)} \mathbf{1}\big\{\cos\!\big(x_k^{\text{inv}} - x_i^{\text{inv}},\ \nu(x_i^{\text{inv}})\big) > \delta\big\}},$$

where $\cos(a,b)$ denotes the cosine similarity, and $\delta \geq 0$ is a tolerance threshold. This construction ensures that L_{ij} is nonzero only for neighbors j whose displacement vectors are well aligned with the local flow direction $\nu(x_i^{\text{inv}})$. The denominator normalizes rows of L to sum up to 1.

Given this directed adjacency matrix, the velocity of other genes can be estimated by a discrete forward difference:

$$\mathcal{L}_\nu(X) \approx V(X) \;=\; X(L^\top - I), \quad V_{g,c} = \sum_{j=1}^{C} X_{g,j} L_{cj} - X_{g,c}, \quad g = 1,...,G, \quad c = 1,...,C, \tag{2.4}$$

where $X \in \mathbb{R}^{G\times C}$ is the full gene-by-cell matrix. Multiplying X by $L^\top$ propagates expression values one step forward along the directed cell-cell graph, analogous to a Lie derivative along the flow field, in the operator-theoretic view [8, 19]. The term $X(L^\top - I)$ thus represents the expected forward change of each gene's expression across cells. Each element $V_{g,c}$ quantifies the rate of change of gene g in cell c.

Remark on comparison to other velocity methods. Most existing velocity methods, such as RNA velocity [4, 12], rely on spliced and unspliced read counts, which may not be available when sequencing depth is limited. Other approaches incorporate additional information, for example, gene regulatory interactions [13], or rely on binning cells into experimental or pseudotime intervals [29]. In contrast, our method does not require splicing dynamics or explicit temporal binning. It operates solely on a predefined list of cell cycle genes to construct a flow field, and propagates this information to all genes while preserving the continuous granularity of cellular dynamics without discretizing time, under the assumption that the underlying dynamics are truly cyclic, a notoriously hard case for modern trajectory inference tools such as optimal transport (e.g., commented by Supplementary Material of Klein et al [11], "On the limitations of optimal transport for recovering cell cycles"). A comparison between flows found by optimal transport and our method can be found Figure 3, showing that optimal transport may still fail to identify a coherent cyclic flow even with suitable temporal binning.

2.3 Time-lagged correlation on the directed graph

The idea of approximating the manifold through a transition matrix can be extended beyond a single step to capture multi-step propagation along the velocity field ν. To propagate information along the local flow, we construct a lagged neighborhood operator built from L. Given a scalar $\alpha \in [0, 1)$, we define the propagation matrix

$$P_\alpha \;=\; \big(I - \alpha L\big)^{-1} L \;=\; \sum_{t=0}^{\infty} \alpha^t L^{t+1}, \tag{2.5}$$

where the latter expansion as a Neumann series corresponds to repeatedly applying the local flow-aligned transition matrix with exponentially decaying weights α raised to the power of t. When $\alpha = 0$, the propagation operator reduces to $P_0 = L$, which corresponds to a single-step update along the directed neighborhood graph. Because each row of L sums to one, this is equivalent to applying one step of a Markov chain whose transition probabilities are supported on the K-nearest neighbors aligned with the local flow direction.

We then follow the approach of [29] to define a generalized notion of time-lagged correlation using the propagation matrix P_α, after normalizing each gene velocity V_g to have unit variance. For any pair of genes (g_1, g_2), we define

$$C_\alpha(g_1, g_2) \;=\; \langle V_{g_1}, P_\alpha\, V_{g_2} \rangle \;=\; \sum_{c_1=1}^{C} \sum_{c_2=1}^{C} V_{g_1,c_1}\, V_{g_2,c_2}\, P_{\alpha,c_1,c_2}, \tag{2.6}$$

where $V_{g,c}$ denotes the estimated velocity of gene g in cell c as in Equation (2.4). The resulting matrix $C_\alpha \in \mathbb{R}^{G\times G}$ thus quantifies the empirical correlation between gene-wise rates of change across cell pairs that are likely to be related as ancestor and descendant along the cycle.

Remark on the properties of C_α. The matrix C_α is signed: positive values indicate that two genes tend to change in the same direction along the flow, whereas negative values indicate opposing changes. Moreover, because P_α encodes time-lagged propagation, C_α is generally asymmetric, allowing one to infer potential directional relationships. Specifically, if changes in g_1 consistently precede those in g_2, then $C_\alpha(g_1, g_2)$ will differ from $C_\alpha(g_2, g_1)$, suggesting a source–target (causal) ordering between the two.

3 Results

We benchmark our algorithm on two types of datasets: (1) synthetic datasets simulated from mechanistic models [26], and (2) real mouse retinal progenitor cells undergoing cell cycle transitions during development [7] in three cell types and one knockout condition. We compare against the following representative algorithms:

- HARISSA and CARDAMOM [26]: mechanistic models for RNA count data that infer regulatory networks based on discretized temporal bins and expression matrices. CARDAMOM extends HARISSA by better exploring the energy landscape and metastable states.
- GENIE3 [1] and GRNBoost2 [14]: decision tree–based algorithms that do not use temporal information. These were ranked among the top-performing methods by BEELINE [18] for both curated and real datasets, with GRNBoost2 serving as a computationally efficient implementation of GENIE3.
- SINCERITIES [16]: a time-aware algorithm that quantifies Granger causality through a linear regression model on Kolmogorov–Smirnov statistic between consecutive time pairs.

For those methods that require time coordinates, we use either the Tricycle coordinates generated by [31], which use transfer learning to project cell cycle genes onto a common periodic template, or the simulation/experimental times when available. Performance is quantified by the area under precision curve (AUPR) and its variants.

Because the current implementation of the finite-volume solver for the Fokker–Planck equation is limited to two dimensions, in our proposed method *CycleGRN*, we learn the flow ν based on the first two principal components of cycle genes for synthetic data. For real datasets, we report results using both the first two principal components and the two-dimensional embedding produced by Tricycle [31]. Gene velocities are then computed using Equation (2.4) and normalized to unit variance across cells prior to computing time-lagged correlations. The lag parameter α in Equation (2.5) is set to 0. More details can be found in Supplementary Information Section 1.

3.1 Results on simulations

We first evaluate the performance of our method on simulated datasets where the ground truth regulatory network is known and can be directly compared with the inferred results. HARISSA [26] is a software framework based on a mechanistic model that simulates RNA counts given any user-specified network. We used three benchmark networks: (1) FN4, a 4-gene network with a feedback loop; (2) CN5, a 5-gene network exhibiting cyclic dynamics; and (3) FN8, an 8-gene network with a feedback loop. To further mimic the complexity of real signaling pathways, we also simulated additional datasets inspired by the *Notch* pathway, including the motif between *Notch1*, *Dll1*, and *Hes1*. For each network, we generated 10 replicates with different random seeds, each containing 1,000 cells. Ground-truth networks and example simulated datasets are shown in the top two rows of Figure 2, with additional visualizations in Supplementary Figure S1.

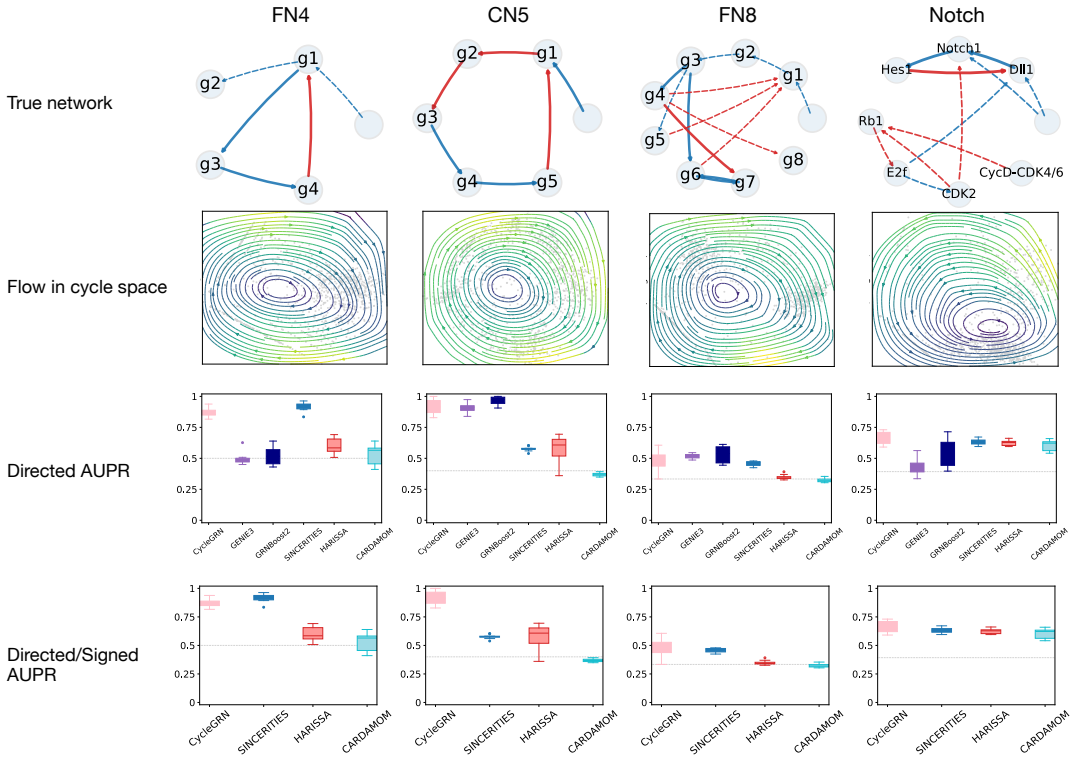

Figure 2: *Results on oscillatory networks simulated by HARISSA [26]. Top row: ground-truth gene regulatory networks, where solid lines indicate edges associated with cycle genes used to learn the invariant flow, and edge color represents activation (blue) or repression (red). Second row: invariant flow $\nu_\theta(x)$ learned by our algorithm on the oscillatory subspace. Third row: area under the precision–recall curve (AUPR) for directed regulation. Bottom row: signed AUPR that additionally accounts for the regulation type (activation vs. inhibition). GENIE3 and GRNBoost2 are omitted because they do not predict regulation type.*

Each dataset also records the simulated time point for every cell, visualized by color in Supplementary Figure S1. To mimic experimental conditions, we intentionally introduced long lags between consecutive time points so that

simulated "experimental time" does not align perfectly with the internal cell-cycle clock. These simulated times were provided to algorithms that require temporal information, including SINCERITIES [16], HARISSA, and CARDAMOM [26]. HARISSA and CARDAMOM both infer networks by directly fitting the mechanistic model used to generate the data, with CARDAMOM offering better exploration of the energy landscape, metastability, and computational efficiency. In contrast, our algorithm does not use simulation time but instead leverages prior knowledge of oscillatory genes (last row in Supplementary Figure S1). Orientation was fixed as described in Section 2.1 using minimal biological markers (the stimulus gene acting as an activator).

We first learn the flow fields on cycle genes and visualize them in the second row of Figure 2. The learned vector field shows smooth transitions in the space of cycle genes and aligns well with the data distribution. We set $K = 200$ and $\delta = 0.95$ to construct the transition matrix L, and infer the gene regulatory network using the time-lagged correlation operator described above.

We evaluate accuracy on the directed network using two metrics: (1) the area under the precision–recall curve (AUPR) computed on the absolute values of predicted edges, and (2) a signed AUPR that additionally requires the correct regulation type (activation or inhibition). All metrics are scaled between 0 and 1, with higher values indicating better performance. Results are shown in the bottom two rows of Figure 2; results for undirected prediction are in Supplementary Figure S2. For the two smaller networks, FN4 and CN5, our method consistently ranks top in all metrics compared to any other method. This is not surprising, since the subnetwork used as cycle genes takes a significant portion of the original true network and almost dominates the dynamics. For the two larger and more complex networks, FN8 and Notch, we show comparable or marginally better performance even compared to HARISSA and CARDAMOM, which have access to the original mechanistic model used for simulating the data. We also note that all the metrics are compared directly against the ground truth network, which is directed (distinguishes source and target genes) and signed (distinguishes activation and inhibition), demonstrating that our method can infer the temporal information even not used as input of the algorithm. This is particularly advantageous in real single-cell datasets, which typically contain only one or a few asynchronous snapshots that do not align with biological time.

3.2 Results on mouse retinal progenitor cells

We further demonstrate the performance of our algorithm on real datasets of mouse retinal progenitor cells (RPCs) [7]. The cells were subset into 27,764 early RPCs, 16,422 late RPCs, and 6,929 neurogenic cells, respectively. We ran the algorithm using 205 cell cycle genes curated from the Gene Ontology term for mouse cell cycle (GO:0007049) [2], and then infer GRNs using $3,164$ highly variable genes. We benchmarked our method against GENIE3 and GRNBoost2 (which do not use temporal information), as well as SINCERITIES with experimental time and with Tricycle positions. Due to the extensive computational cost, we randomly subsampled 1,000 cells for GENIE3 and GRNBoost2 for each cell type. Following BEELINE [18], due to low absolute accuracy, we evaluated performance using the AUPR ratio (AUPR divided by network density) and AUROC.

The cell cycle embeddings are shown in Figure 4 and Supplementary Figure S3, where the point clouds appear as amorphous blobs without clear cyclic structure. Experimental time also misaligns with the cell cycle phase revealed by Tricycle coordinates as shown in Supplementary Figure S3. Though Tricycle coordinates successfully provide an ordering that agrees to known biology, existing method using optimal transport on that time coordinate still fails to recover a coherent cyclic flow as shown in Figure 3. Nevertheless, our method successfully recovers a dynamical system consistent with the Tricycle positions. The recovered flow fields of each cell type are shown in panel (a) of Figure 4. We then compute gene-wise velocities for all genes, where red indicates increasing and blue indicates decreasing expression. Velocities for *Top2a* are shown in panel (b) of Figure 4, peaking at the mid-cycle phase, consistent with rapid transcriptional dynamics during G2 [27], and is markedly reduced in neurogenic RPCs, reflecting their diminished proliferative activity and exit from the cell cycle.

Next, we apply the lagged correlation operator to 3,164 genes, selected as transcription factors and pattern

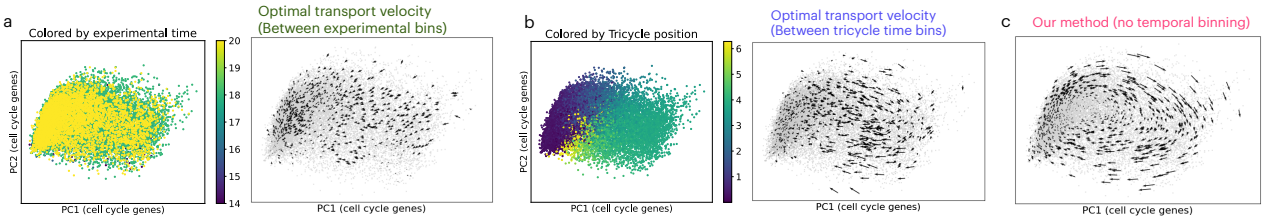

Figure 3: *Comparison of velocity fields on late RPCs learned using (a) OT velocity [29] using experimental time, (b) OT velocity [29] using Tricycle phase divided into 20 equally-sized bins, and (c) our invariant dynamics without any time labels. Optimal-transport–based velocity fields fail to recover a coherent cyclic flow under either temporal coordinate, whereas our method recovers a consistent circular dynamics.*

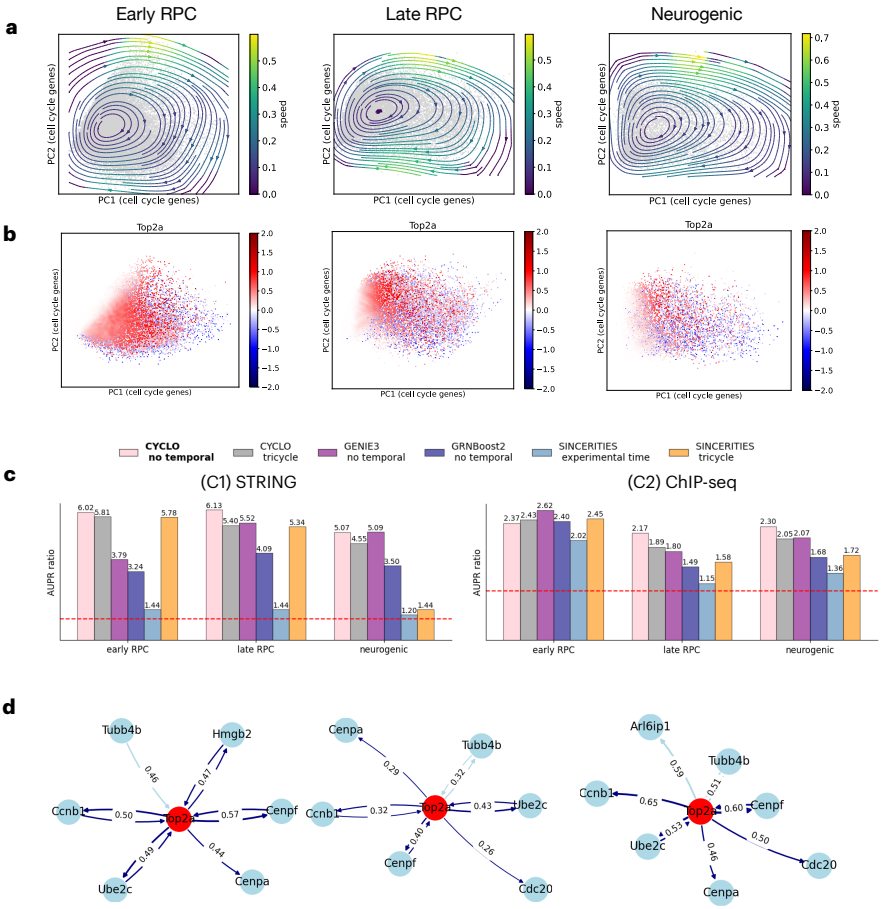

Figure 4: *Results on mouse retinal cells in 3 cell types: early retinal progenitor (first column), late retinal progenitor (second column), and neurogenic (third column) cells. Top row: invariant flow learned for each cell type, overlaid on the first two principal components of 205 cell cycle genes. Second row: gene velocity associated with Top2a, estimated by equation 2.4, consistent with the fact that the gene starts to peak in the middle of cycle rather than the beginning. Third row: accuracy of resulting network benchmarked against STRING (b1) and ChIP-seq (b2), quantified by AUPR ratio (AUPR divided by density of network). Our method consistently ranks among the top-performing approaches across all RPC populations, and remains competitive in neurogenic cells where cell-cycle activity is weakened. Bottom row: top 10 highest weighted edges associated with Top2a, with dark blue edges indicating evidence supported by STRING.*

markers, to infer directed gene–gene relationships. We benchmark the inferred networks against the STRING database [23], restricted to 80 pattern markers genes curated from non-negative matrix factorization analysis of the complete retinal dataset [7] with CoGAPS [9, 22] within each cell type performed previously on the same dataset. Only regulatory links with at least 80% confidence in STRING were retained for evaluation.

The overall performance, measured by AUPR ratio, is shown in panel (c) of Figure 4 under two complementary benchmarks: (C1) STRING-based evaluation and (C2) ChIP-seq–supported TF–target interactions. Under STRING (C1), our method outperforms all others in early and late RPCs, where cell-cycle activity is pronounced. Under ChIP-seq (C2), performance varies slightly across cell types: GENIE3 shows higher agreement in early RPCs, whereas our method achieves the best or near-best results in late RPC and neurogenic populations. Performance remains comparable to GENIE3 in neurogenic cells, while other methods approach random baseline. Results are robust to the choice of k and τ (Supplementary Figure S4). Learning the flow directly on the Tricycle embedding yields slightly reduced performance compared to learning on cell-cycle genes [31].

We further examine whether the method improves predictions for cell cycle genes, as hypothesized. For the representative cell cycle gene *Top2a*, we compare the inferred network edges to those predicted by GENIE3 (Supplementary Figure S5), focusing on the top ten edges by weight. GENIE3 correctly predicts 8, 8, and 6 edges across the three cell types, whereas our method predicts 9, 8, and 7, respectively. Notably, our method identified the edge pointing from *Top2a* to *Cenpa*, a G2/M gene that peaks after *Top2a* [27], while GENIE3 missed it. This is likely because this relationship is phase-dependent and directional within the cell-cycle trajectory, and can be blurred by static networks inferred from regression-based methods such as GENIE3. This result is consistent with our hypothesis and highlights the potential of our method for improving regulatory inference in cell types exhibiting strong oscillatory signals.

The methodology is time-lagged correlation–based; therefore, it is crucial to verify whether the inferred networks capture directional causal relationships rather than static associations. To this end, we performed differential GRN inference on (1) regular late retinal progenitor cells (RPCs) and (2) *Nfia/b/x* triple-knockout RPCs, in which the transcriptional regulators *Nfia*, *Nfib*, and *Nfix* are simultaneously disrupted [7]. This knockout is known to impair the specification of late-born retinal cell types and to cause a loss of Müller glia and bipolar differentiation. The results are summarized in Figure 5.

Figure 5 presents (a) the velocity field learned from cell-cycle genes, (b) the differential incoming and outgoing hub strengths, and (c) the subnetwork formed by the 50 most rewired edges. The inferred network differences recapitulate the experimental phenotype: in the knockout, a reinforced core of cell-cycle regulators (*Hmgb2*, *Top2a*, *Ube2c*, *Cdc20*) gains mutual connectivity and forms a self-sustaining proliferative module, while genes associated with differentiation and late mitotic execution (*Cenpf*, *Cenpa*, *Ccnb1*, *Tubb4b*) lose both incoming and outgoing regulatory edges, reflecting a collapse of coordinated lineage and mitotic control.

This asymmetric rewiring, where upstream proliferative drivers persist while downstream effectors and differentiation targets become decoupled, demonstrates that the inferred network structure reflects directional causality consistent with biological perturbation. The method therefore captures not only known regulatory phenotypes but also distinguishes causal persistence (sustained drivers) from causal disconnection (lost regulation) within perturbed developmental programs.

4 Discussion

We have presented *CycleGRN*, a framework that learns an invariant flow in the space of genes corresponding to oscillatory processes such as the cell cycle and leverages it for gene regulatory network inference. The method achieves strong performance on both simulated and real datasets, particularly enhancing inference for cell cycle genes. Although we focus on the cell cycle, the framework naturally extends to other recurrent processes like circadian rhythms and *Notch* oscillations.

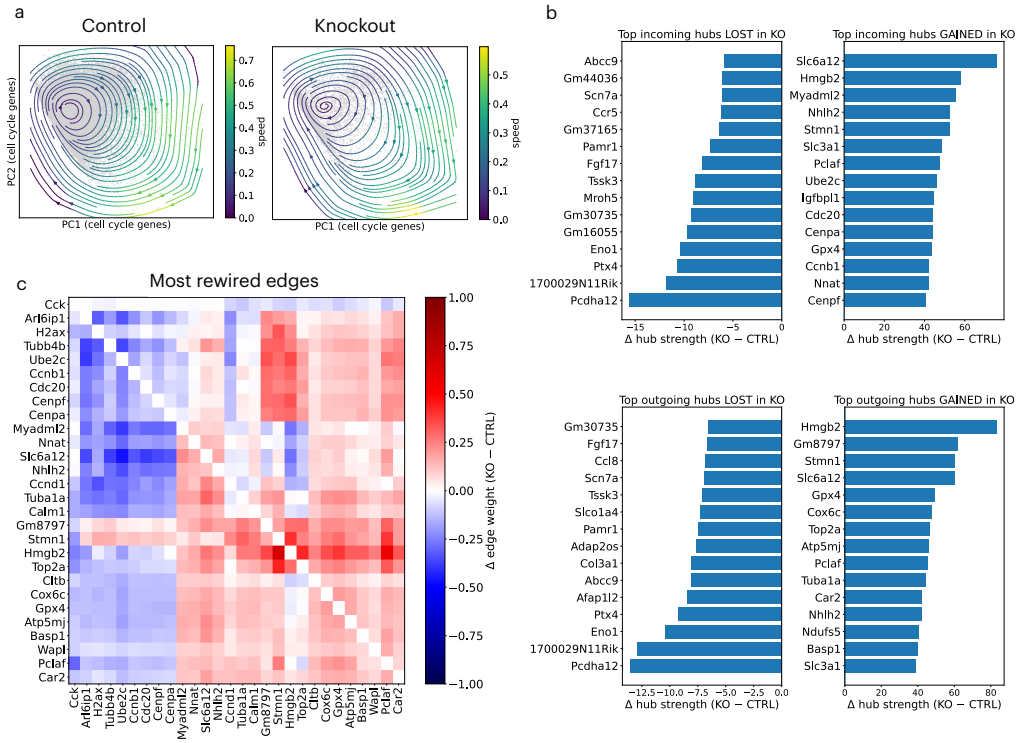

Figure 5: *Differential GRN inference on regular late RPCs versus Nfia/b/x triple-knockout RPCs. (a) Vector field learned from cell cycle genes, (b) incoming/outgoing hub genes, and (c) subnetwork for top 50 differential edges.*

There are several avenues for further improvement. First, the current implementation relies on a 2D embedding prior to learning the dynamics. This may fail in settings with substantial cell type heterogeneity, where a simple 2D projection does not reveal a clean cyclic structure. At the same time, the low-dimensional projection can act as a geometric denoising step, stabilizing local distances before flow estimation. Extending the flow learning procedure to higher-dimensional representations could improve robustness across diverse datasets while preserving this regularization effect. Second, as discussed in the methodology, the learned flow is not guaranteed to be unique. Prior work has leveraged time-delay coordinates and Takens' theorem [5, 24] to address this issue. Although true lagged measurements are not directly available in single-cell experiments, experimental time points or pseudotime estimates could potentially be used to construct suitable surrogate constraints. Third, our algorithm is naturally biased toward oscillatory genes, particularly cell cycle genes, and may not perform optimally for regulatory inference involving other processes. For example, we observed suboptimal recovery of interactions involving *Hes1* (Supplementary Figure S6).

More broadly, biological systems exhibit regulation across multiple temporal and functional scales, from fast cell-cycle oscillations to slower developmental and differentiation programs as well as pathway-specific signaling modules. Capturing these processes within a unified modeling framework remains a key challenge. We envision extending *CycleGRN* toward multi-scale inference, where dynamic regulatory flows corresponding to different timescales and pathways are jointly identified and integrated, bridging cyclic and hierarchical modes of regulation.

References

[1] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. Scenic: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, October 2017.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[3] Ralf Banisch and Péter Koltai. Understanding the geometry of transport: Diffusion maps for lagrangian trajectory data unravel coherent sets. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(3), February 2017.

[4] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, August 2020.

[5] Jonah Botvinick-Greenhouse, Robert Martin, and Yunan Yang. Learning dynamics on invariant measures using pde-constrained optimization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(6), June 2023.

[6] Mengjie Chen and Xiang Zhou. Controlling for confounding effects in single cell rna sequencing studies using both control and target genes. *Scientific Reports*, 7(1), October 2017.

[7] Brian S. Clark, Genevieve L. Stein-O'Brien, Fion Shiau, Gabrielle H. Cannon, Emily Davis-Marcisak, Thomas Sherman, Clayton P. Santiago, Thanh V. Hoang, Fatemeh Rajaii, Rebecca E. James-Esposito, Richard M. Gronostajski, Elana J. Fertig, Loyal A. Goff, and Seth Blackshaw. Single-cell rna-seq analysis of retinal development identifies nfi factors as regulating mitotic exit and late-born cell specification. *Neuron*, 102(6):1111–1126.e5, June 2019.

[8] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. Special Issue: Diffusion Maps and Wavelets.

[9] Elana J. Fertig, Jie Ding, Alexander V. Favorov, Giovanni Parmigiani, and Michael F. Ochs. Cogaps: an r/c++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*, 26(21):2792–2793, September 2010.

[10] Xinyu Guo and Liang Chen. From g1 to m: a comparative study of methods for identifying cell cycle phases. *Briefings in Bioinformatics*, 25(2), January 2024.

[11] Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, Aimée Bastidas-Ponce, Perla Cota, Marta Tarquis-Medina, Shrey Parikh, Ilan Gold, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, and Fabian J. Theis. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, January 2025.

[12] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco,

Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. Rna velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[13] Jiachen Li, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. Tfvelo: gene regulation inspired rna velocity estimation. *Nature Communications*, 15(1), February 2024.

[14] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, November 2018.

[15] Yasutaka Niwa, Hiromi Shimojo, Akihiro Isomura, Aitor González, Hitoshi Miyachi, and Ryoichiro Kageyama. Different types of oscillations in notch and fgf signaling regulate the spatiotemporal periodicity of somitogenesis. *Genes &; Development*, 25(11):1115–1120, June 2011.

[16] Nan Papili Gao, S M Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, September 2017.

[17] Siim Pauklin and Ludovic Vallier. The cell-cycle state of stem cells determines cell fate propensity. *Cell*, 155(1):135–147, September 2013.

[18] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, January 2020.

[19] 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, Emanuelle I. Grody, Matthew J. Shurtleff, Ruoshi Yuan, Song Xu, Yian Ma, Joseph M. Replogle, Eric S. Lander, Spyros Darmanis, Ivet Bahar, Vijay G. Sankaran, Jianhua Xing, and Jonathan S. Weissman. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711.e45, February 2022.

[20] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, April 2019.

[21] Daniel Schwabe, Sara Formichetti, Jan Philipp Junker, Martin Falcke, and Nikolaus Rajewsky. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, 16(11), November 2020.

[22] Genevieve L Stein-O'Brien, Jacob L Carey, Wai Shing Lee, Michael Considine, Alexander V Favorov, Emily Flam, Theresa Guo, Sijia Li, Luigi Marchionni, Thomas Sherman, Shawn Sivy, Daria A Gaykalova, Ronald D McKay, Michael F Ochs, Carlo Colantuoni, and Elana J Fertig. Patternmarkers &; gwcogaps for novel data-driven biomarkers via whole transcriptome nmf. *Bioinformatics*, 33(12):1892–1894, February 2017.

[23] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, November 2022.

[24] Floris Takens. *Detecting strange attractors in turbulence*, page 366–381. Springer Berlin Heidelberg, 1981.

[25] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, March 2014.

[26] Elias Ventre, Ulysse Herbach, Thibault Espinasse, Gérard Benoit, and Olivier Gandrillon. One model fits all: Combining inference and simulation of gene regulatory networks. *PLOS Computational Biology*, 19(3):e1010962, March 2023.

[27] Michael L. Whitfield, Gavin Sherlock, Alok J. Saldanha, John I. Murray, Catherine A. Ball, Karen E. Alexander, John C. Matese, Charles M. Perou, Myra M. Hurt, Patrick O. Brown, and David Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13(6):1977–2000, June 2002.

[28] Yunan Yang, Levon Nurbekyan, Elisa Negrini, Robert Martin, and Mirjeta Pasha. Optimal transport for parameter identification of chaotic dynamics via invariant measures. *SIAM Journal on Applied Dynamical Systems*, 22(1):269–310, March 2023.

[29] Wenjun Zhao, Erica Larschan, Björn Sandstede, and Ritambhara Singh. Optimal transport reveals dynamic gene regulatory networks via gene velocity estimation. *PLOS Computational Biology*, 21(5):e1012476, May 2025.

[30] Wenjun Zhao, Alma Plaza-Rodriguez, Pichayathida Luanpaisanon, Elena Xinyi Wang, Linnea Gyllingberg, Elana Fertig, and Genevieve L Stein-O'Brien. Inferring the regulation dynamics of oscillatory networks from scrna-seq data. November 2025.

[31] Shijie C. Zheng, Genevieve Stein-O'Brien, Jonathan J. Augustin, Jared Slosberg, Giovanni A. Carosso, Briana Winer, Gloria Shin, Hans T. Bjornsson, Loyal A. Goff, and Kasper D. Hansen. Universal prediction of cell-cycle position using transfer learning. *Genome Biology*, 23(1), January 2022.

[32] Binghan Zhou, Wanling Lin, Yaling Long, Yunkai Yang, Huan Zhang, Kongming Wu, and Qian Chu. Notch signaling pathway: architecture, disease, and therapeutics. *Signal Transduction and Targeted Therapy*, 7(1), March 2022.

Data and code availability

The HARISSA simulations and scripts to reproduce the experiments are available at https://github.com/WenjunZHAOwO/CycleGRN/tree/main. The mouse retinal dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118614.

A Identifiability of cyclic flows from invariant measure

We characterize the identifiability of latent cyclic dynamics from an observed invariant measure in the presence of diffusion. This analysis clarifies the degree of non-uniqueness in the inferred flow and motivates the additional biological constraints used in the main text.

Setup. Let the latent cell-cycle coordinate be parameterized by $\theta \in S^1 \simeq \mathbb{R}/(2\pi\mathbb{Z})$. We consider a one-dimensional diffusion process

$$d\theta_t \;=\; b(\theta_t)\,dt \;+\; \sqrt{2D}\,dW_t, \tag{1.1}$$

where $b : S^1 \to \mathbb{R}$ is a 2π-periodic drift field and $D > 0$ is a constant diffusion coefficient. Suppose the process admits a strictly positive stationary density $\rho(\theta)$ satisfying

$$\int_0^{2\pi} \rho(\theta)\,d\theta = 1.$$

Stationary Fokker–Planck equation. The stationary density ρ satisfies the one-dimensional periodic Fokker–Planck equation

$$0 \;=\; -\partial_\theta\big(b(\theta)\rho(\theta)\big) \;+\; D\,\partial_{\theta\theta}\rho(\theta). \tag{1.2}$$

Define the stationary probability current

$$j(\theta) \;:=\; b(\theta)\rho(\theta) - D\,\partial_\theta\rho(\theta). \tag{1.3}$$

Equation (1.2) implies $\partial_\theta j(\theta) = 0$, hence

$$j(\theta) \equiv J$$

for some constant $J \in \mathbb{R}$ representing the net circulation around the cycle.

Drift decomposition. Solving (1.3) for the drift yields

$$b(\theta) \;=\; D\,\partial_\theta \log\rho(\theta) \;+\; \frac{J}{\rho(\theta)}. \tag{1.4}$$

This decomposition separates the drift into a density-determined component and a circulating component corresponding to persistent directional flow around the cycle.

Identifiability. Proposition. Let b_1 and b_2 be two drift fields on S^1 generating diffusion processes with the same diffusion coefficient $D > 0$ and the same strictly positive stationary density ρ. Then there exists a constant $c \in \mathbb{R}$ such that

$$b_2(\theta) - b_1(\theta) = \frac{c}{\rho(\theta)} \qquad \text{for all } \theta \in S^1.$$

Proof. Both drifts admit the representation (1.4) with constants J_1 and J_2, respectively. Subtracting the two expressions gives

$$b_2(\theta) - b_1(\theta) = \frac{J_2 - J_1}{\rho(\theta)},$$

which proves the claim. □

Interpretation. Thus, on a one-dimensional cyclic manifold, the invariant density ρ determines the spatial structure of the drift uniquely, leaving only a single scalar ambiguity corresponding to the magnitude and direction of circulation. This ambiguity cannot be resolved from the stationary distribution alone and reflects a fundamental non-identifiability of cyclic flows from steady-state data. In the main text, we resolve this remaining degree of freedom using biological prior information, specifically known cell-cycle marker genes, which provide temporal orientation along the cycle.

Remarks. If detailed balance is assumed, then $J = 0$ and the drift is uniquely determined by ρ via $b(\theta) = D\,\partial_\theta \log \rho(\theta)$. If the diffusion coefficient D is unknown, additional non-identifiability arises and cannot be resolved from the invariant density alone.

B Implementation note on learning the flow

The flow field ν_θ was trained following the implementation of Botvinick-Greenhouse et al [5], without much modification to the original code available at `https://github.com/jrbotvinick/Learning-Dynamics-on-Invariant-Measures`. The invariant-measure solver and adjoint gradient are implemented using a finite–volume discretization of the Fokker–Planck equation coupled to a neural-network parameterization of the drift field.

Domain and discretization. We restricted the invariant-gene embedding to a two–dimensional square domain $[-4, 4] \times [-4, 4]$ with spatial step $\Delta x = 0.1$ and time step $\Delta t = 0.01$, yielding $n_x = n_y = 81$ grid points. The diffusion coefficient was fixed at $D = 10^{-3}$ and a small teleportation parameter $\alpha = 10^{-8}$ was added to ensure ergodicity. The empirical steady–state density ρ^* was estimated from the cell embeddings via a two–dimensional histogram, normalized to integrate to one, and smoothed by a Gaussian kernel with standard deviation $\sigma = 2$ grid cells. We also note that because the space is discretized and we only use the relative counts of cells within each bin to learn the dynamics, computationally, the method scales very well with number of cells.

Velocity parameterization. The drift field $\nu_\theta(x) = (u_\theta, v_\theta)$ was represented by a fully connected neural network with two hidden layers of 100 units each and tanh activation, which guarantees smoothness and consequently existence of a valid solution.

The network was implemented in PyTorch and evaluated on a staggered grid aligned with the finite–volume stencil, enforcing no–flux boundary conditions.

Optimization and loss. We minimized the Kullback–Leibler divergence between the empirical and model-implied stationary densities,

$$\mathcal{D}(\rho^*, \rho_{\nu_\theta}) = \sum_i \rho_i^* \log\left(\frac{\rho_i^*}{\rho_{\nu_\theta,i}}\right),$$

where ρ_{ν_θ} is obtained by solving the steady–state Fokker–Planck equation $\nabla\cdot(\rho_{\nu_\theta}\nu_\theta) = D\nabla^2\rho_{\nu_\theta}$. The corresponding adjoint equation was solved analytically in discrete form to compute exact gradients of the loss with respect to ν_θ, which were back-propagated through the neural network using a custom `torch.autograd.Function`. Parameters were optimized using Adam with learning rate 10^{-3} for $1,000$ iterations, and random seeds were fixed for reproducibility.

C Supplementary figures

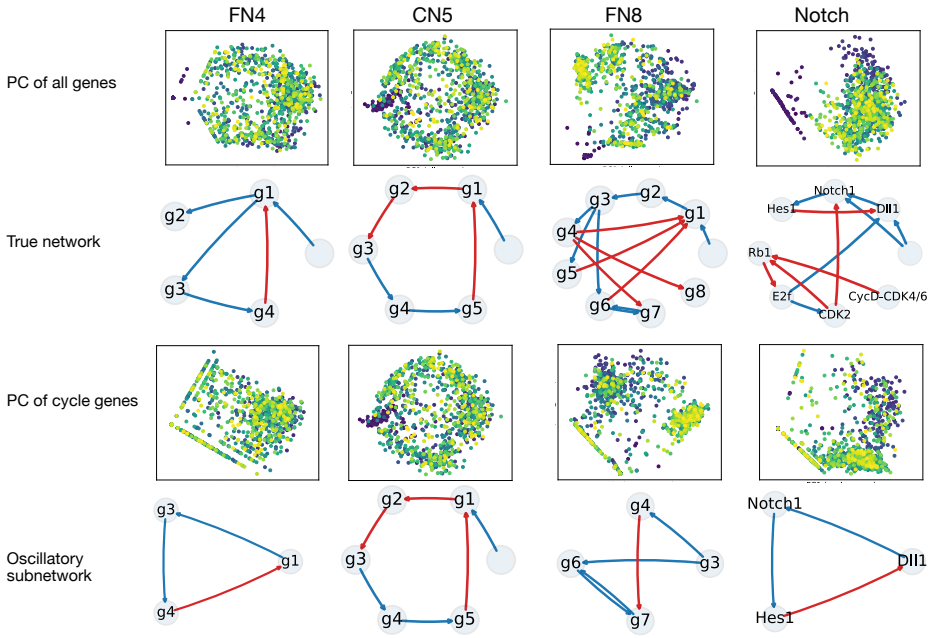

Figure S1: *Datasets with oscillatory networks simulated by HARISSA [26]. This figure shows low dimensional visualization and ground truth networks either with all genes (second row) or just a subset that corresponds to oscillatory process (last row). In the principal components, cells are colored according to the time points at which they were simulated.*

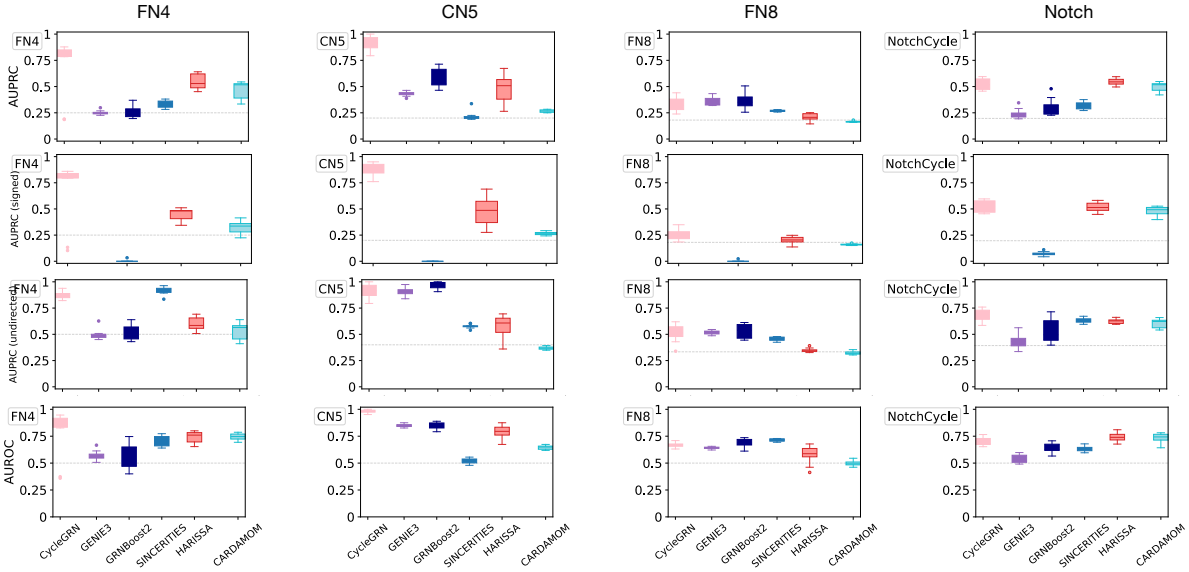

Figure S2: *Accuracy metrics on simulated data, resulting from a wide range of algorithms. GENIE3 and GRNBoost2 do not produce signed prediction and is omitted from the third row on signed AUPRC.*

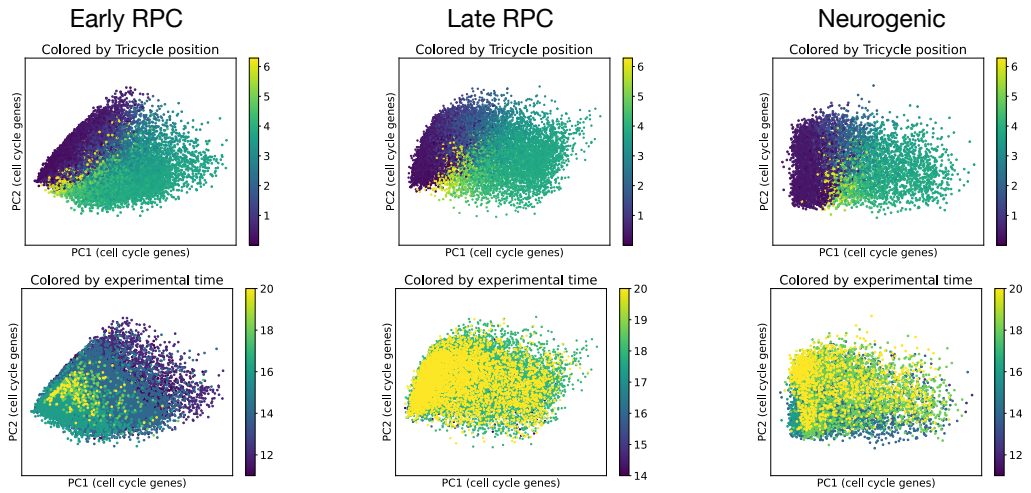

Figure S3: *Visualization of mouse retinal cells colored by experimental time (first row) and tricycle coordinates (second row)*

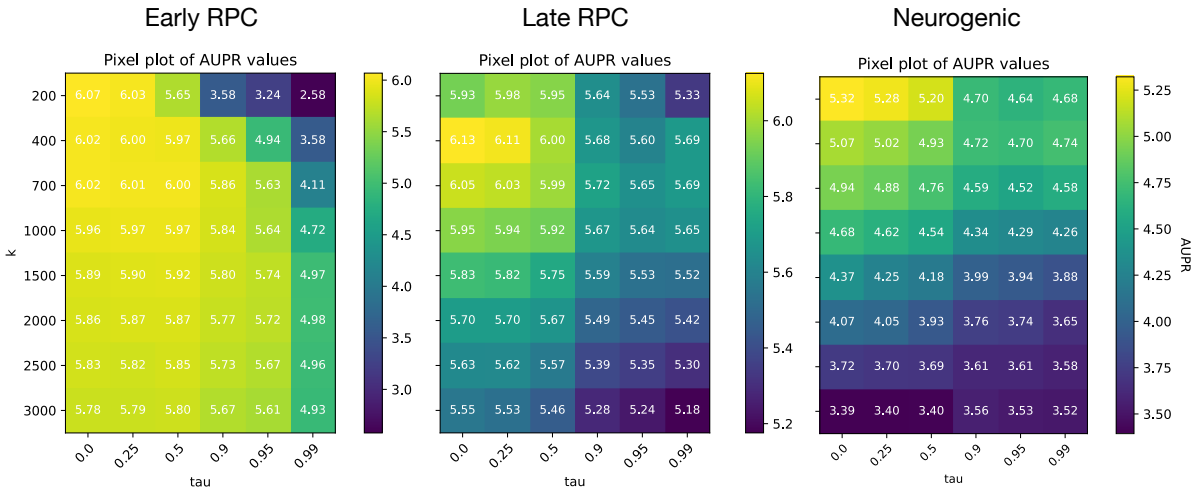

Figure S4: *Sensitivity analysis on the hyperparameters τ and k used in constructing the transition matrix on real data. Overall, the performance quantified by AUPRC ratio is robust for moderate τ and k. We have used $\tau = 0$ and $k = 400$ in all real data studies.*

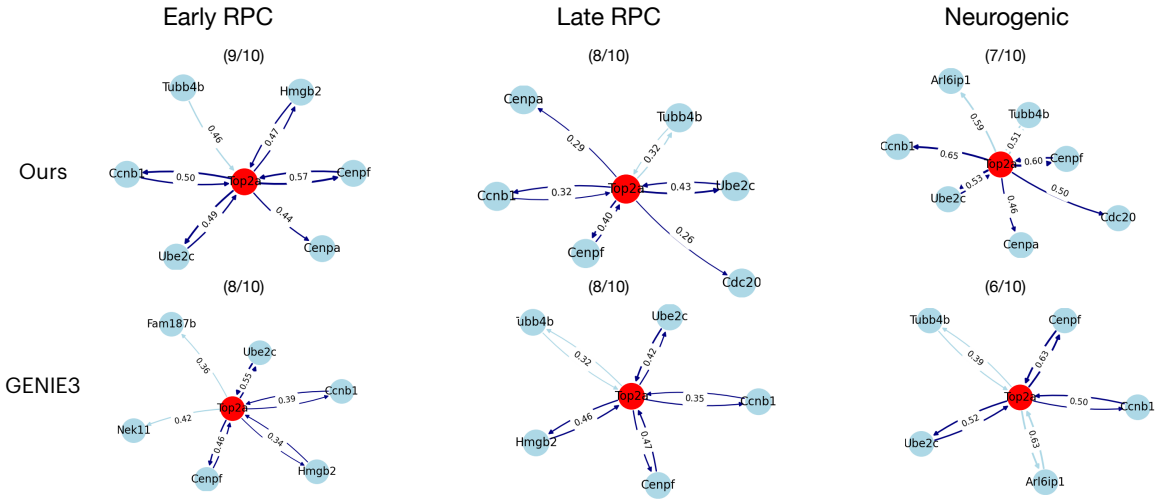

Figure S5: *Comparison of networks between GENIE3 and ours on Top2a for top 10 edges. Our method is able to identify one more edge connecting Cenpa, a cell cycle gene that peaks after Top2a, for early RPC and neurogenic cells.*

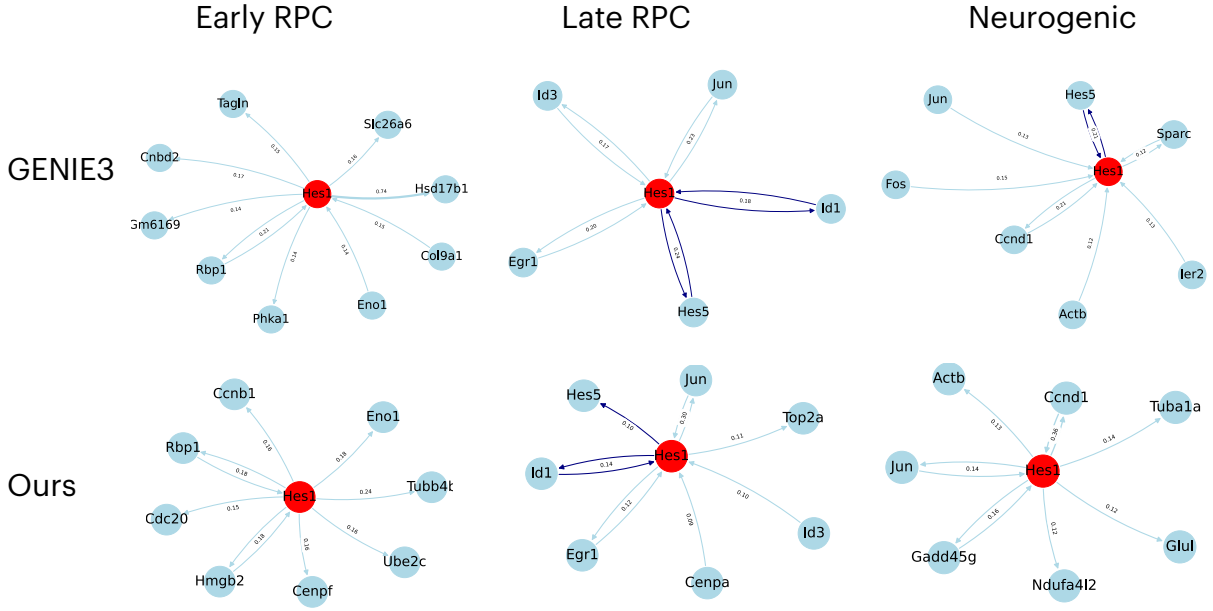

Figure S6: *Comparison of networks between GENIE3 and ours on Hes1. As our method prioritized cell cycle genes that do not include Hes1, it can fail to identify edges if its dynamics do not respect the same temporal ordering.*