Reconstructing growth and dynamic trajectories from single-cell transcriptomics data

Nature Machine Intelligence(2024)

Yutong Sha, Yuchi Qiu, Peijie Zhou & Qing Nie
Reporter: Fengjiao Gong

October 8, 2024

Outline

1. Background
2. Trajectory Inference with Growth via Optimal transport and Neural network(TIGON)
3. Experiments

Background

Single-cell **RNA sequencing (scRNA-seq) methods** observing dynamics by sampling cells at different times.

- cells are killed during sequencing
- only provides unpaired snapshots

Thus,

- gene expression dynamics of individual cells are not traceable
- cell lineage relationship (cell trajectory) between different sequenced times is missing

Dynamic OT

Suppose **gene expression state** $x \in \mathbb{R}^d$ has **density** $\rho(x,t) \geq 0, t \in [0,T]$:

- smooth
- initial and final conditions:

$$\rho(\cdot,0) = \rho_0, \quad \rho(\cdot,T) = \rho_T \tag{1}$$

Then,

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{2}$$

with velocity

$$\mathbf{v}(x,t) = \frac{dx(t)}{dt}. \tag{3}$$

[ref] *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, 2000.*

Dynamic OT

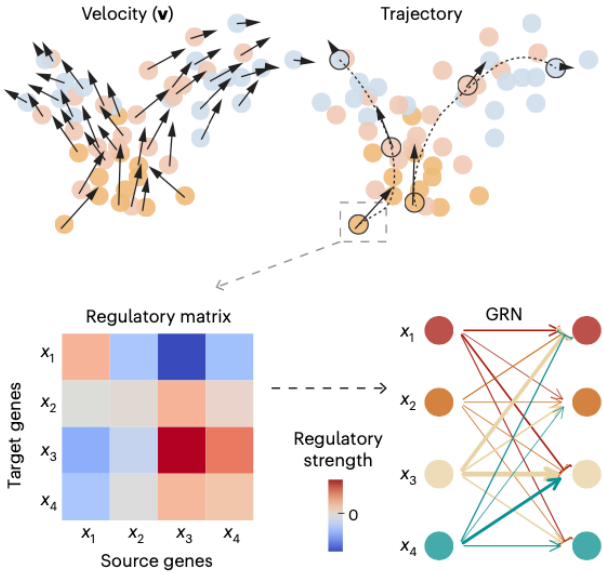
Cost function (kinetic energy):

$$\mathcal{L}_0 = T \int_0^T \int_{\mathbb{R}^d} |\mathbf{v}(x,t)|^2 \rho(x,t)\, dx\, dt. \tag{4}$$

$$\min \mathcal{L}_0 \quad \Leftrightarrow \quad \text{Wasserstein distance } (p=2)$$

[ref] *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, 2000.*

Illustration



- **Velocity**. Each dot represents a cell coloured by collection time and length of arrow denotes the magnitude of the velocity.
- **Trajectory** of each cell.
- **Gene Regulatory matrix** of a selected cell or cell type.
- **GRN**. The pointed arrows (blunt arrows) denote positive (negative) regulation from the source gene to the target gene and the width denotes regulatory strength.

Background

Cell populations may change in time due to *cell division* and *cell death*.

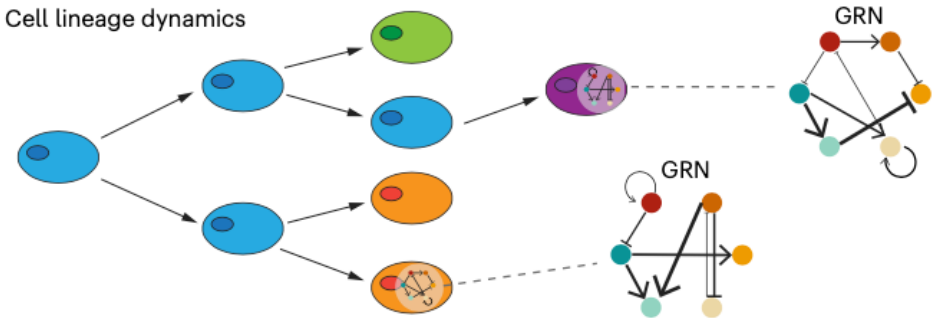


Figure 1: Illustrative graph of cell lineage dynamics which involves cell growth, transition and GRNs.

A growth term that captures such net change is needed. [ref] *Fundamental limits on dynamic inference from single-cell snapshots, 2018.*

Method — TIGON

1. Unbalanced Dynamic OT
2. Dimensionless Cost Function
3. Reconstruction Errors
4. Deep Learning-Based Dimensionless Solver

TIGON — Reconstruction of cell density

Suppose the **time-series discrete data** are given by

$$(t_1, C^1), (t_2, C^2), \ldots, (t_r, C^r)$$

where $C^i = \{c_{t_i}^{(j)}\}_{j=1}^{N_i} \in \mathbb{R}^{N_i \times d}$ is a set of N_i independent and identically distributed samples drawn from the distribution at a d-dimensional space at time t_i.

For each t_i, generate **ground truth density** ρ_{t_i} by

- Gaussian mixture model
- weighted by relative population ratio

TIGON — Reconstruction of cell density

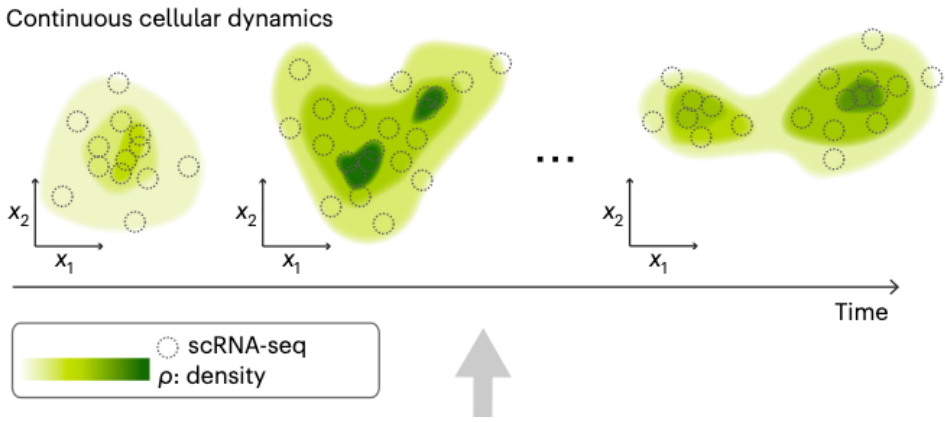


Figure 2: Time-series scRNA-seq data is used to generate density functions at the given discrete time points: $\rho_i = \rho(x, t), i = 1, 2, \ldots$, using a Gaussian mixture model.

Unbalanced Dynamic OT

Reconstructs $\rho(x,t)$ by a hyperbolic partial differential equation:

$$\partial_t \rho(x,t) + \nabla \cdot (\boldsymbol{v}(x,t)\rho(x,t)) = g(x,t)\rho(x,t). \tag{5}$$

- **convection term** $\nabla \cdot (\boldsymbol{v}(x,t)\rho(x,t))$ describes the transport of *cell density*.
- **growth term** $g(x,t)$ describes the instantaneous *population change* of cell density dynamics.

Unbalanced Dynamic OT

Equation (5) is solved using unbalanced OT by optimizing the **WFR cost**:

$$W_{0,T} = \int_0^T \int_{\mathbb{R}^d} \left(|\boldsymbol{v}(x,t)|^2 + \alpha g(x,t)^2 \right) \rho(x,t) dx dt. \tag{6}$$

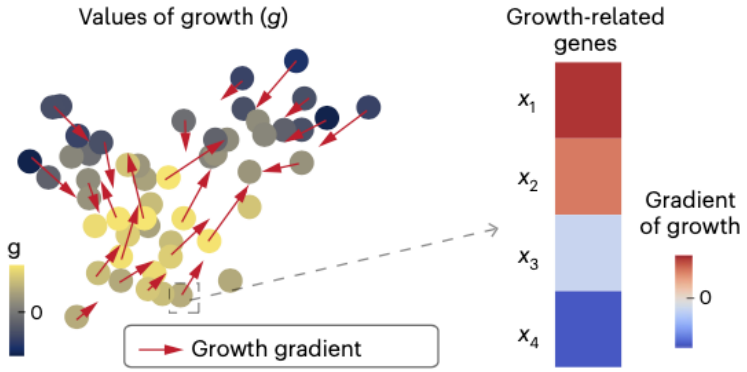
$\min W_{0,T}$, kinetic energy + growth energy

[ref] *An Interpolating Distance between Optimal Transport and Fisher-Rao, 2015.*

Growth-Related Genes

Assess contributions of each gene to growth by the **gradient**:

$$\nabla g = \left\{ \frac{\partial g}{\partial x_j} \right\}_{j=1}^{d} \tag{7}$$



Dimensionless Formulation

To deal with **high-dimensional** integrals in gene expression space:

$$\text{velocity } \boldsymbol{v}(x,t) \approx NN_1(x,t), \quad \text{growth } g(x,t) \approx NN_2(x,t)$$

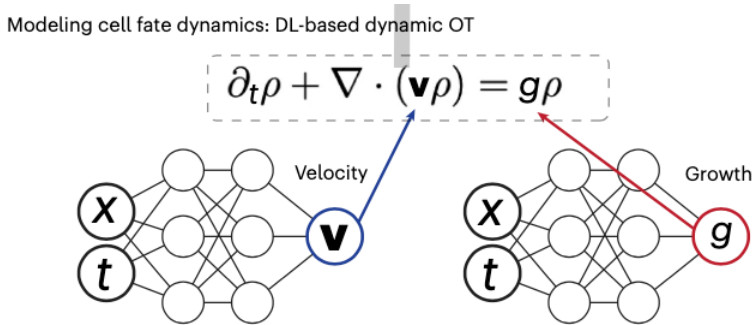


Figure 3: The density ρ is governed by a partial differential equation involving velocity $\boldsymbol{v}$ and growth g that are modelled by two neural networks. DL, deep learning.

Method - Dimensionless Formulation

Lemma: If density $\rho(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$, velocity field $\mathbf{v}(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ and growth $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ satisfy

$$\partial_t \rho(x,t) + \nabla \cdot (\mathbf{v}(x,t)\rho(x,t)) = g(x,t)\rho(x,t),$$

with initial condition

$$\rho(x,0) = \rho_0(x)$$

for all $0 \leq t \leq T$

$$\frac{dx(t)}{dt} = \mathbf{v}(x,t), \quad x(0) = x_0,$$

then we have

$$\frac{d(\ln \rho)}{dt} = g - \nabla \cdot \mathbf{v}. \tag{8}$$

Method - Dimensionless Formulation

Proof:

$$\frac{\partial \rho}{\partial t} = g\rho - \nabla \cdot (\mathbf{v}\rho) = g\rho - \nabla\rho \cdot \mathbf{v} - \rho\nabla \cdot \mathbf{v}$$

$$\begin{aligned}
\frac{d\rho}{dt} &= \nabla\rho \cdot \frac{dx}{dt} + \frac{\partial \rho}{\partial t} \\
&= \nabla\rho \cdot \mathbf{v} + \frac{\partial \rho}{\partial t} \\
&= \nabla\rho \cdot \mathbf{v} + g\rho - \nabla \cdot (\mathbf{v}\rho) \\
&= \nabla\rho \cdot \mathbf{v} + g\rho - \nabla\rho \cdot \mathbf{v} - \rho\nabla \cdot \mathbf{v} \\
&= g\rho - \rho\nabla \cdot \mathbf{v}
\end{aligned}$$

So that

$$\frac{d(\ln \rho)}{dt} = \frac{1}{\rho} \cdot \frac{d\rho}{dt} = g - \nabla \cdot \mathbf{v}$$

Method - Dimensionless Formulation

Theorem: If smooth density $\rho(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$, velocity field $\mathbf{v}(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ and growth rate $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ satisfy

$$\begin{cases} \partial_t \rho(x,t) + \nabla \cdot (\mathbf{v}(x,t)\rho(x,t)) = g(x,t)\rho(x,t) \\ \rho(x,0) = \rho_0(x) \end{cases}$$

for all $0 \leq t \leq T$ where

$$\begin{cases} \frac{dx(t)}{dt} = \mathbf{v}(x,t) \\ x(0) = x_0 \end{cases}$$

then for any measurable function $f(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, we have

$$\int_0^T \int_{\mathbb{R}^d} f(x,t)\rho(x,t)\,dx\,dt = \mathbb{E}_{x_0 \sim \rho_0}\left[\int_0^T f(x,t) e^{\int_0^t g(x,s)\,ds}\,dt\right]. \tag{9}$$

Method - Dimensionless Cost Function

An equivalent dimensionless form of the cost function in WFR metric:

$$
\begin{aligned}
W_{0,T} &= T \int_0^T \int_{\mathbb{R}^d} \underbrace{\left(|\mathbf{v}(x,t)|^2 + \alpha g(x,t)^2\right)}_{f(x,t)} \rho(x,t)\, dx\, dt \\
&= T \mathbb{E}_{x_0 \sim \rho_0} \int_0^T \left(|\mathbf{v}(x,t)|^2 + \alpha g(x,t)^2\right) e^{\int_0^t g(x,s)\, ds}\, dt
\end{aligned}
\tag{10}
$$

where $\mathbb{E}_{x_0 \sim \rho_0}[\cdot]$ denotes the expectation for random variable x_0 following distribution ρ_0.

We assume the characteristic curves do not intersect.

Method - Dimensionless Cost Function

Compute (10) by summing up the cost between all pairs of consecutive time points:

$$W = \sum_{i=1}^{T-1} W_{t_i,t_{i+1}}. \tag{11}$$

where

$$W_{t_i,t_{i+1}} = (t_{i+1} - t_i) E_{x_i \sim \rho_{t_i}} \int_{t_i}^{t_{i+1}} \left(|\mathbf{v}(x,t)|^2 + \alpha g(x,t)^2 \right) e^{\int_{t_i}^{t} g(x,s)\, ds}\, dt \tag{12}$$

and trajectory $x = x(t)$ satisfying

$$\begin{cases} \frac{dx(t)}{dt} = \mathbf{v}(x,t) \\ x(t_i) = x_i \end{cases}$$

.

Method - Reconstruction Errors

Recall **Lemma**:

$$\frac{d(\ln \rho)}{dt} = g - \nabla \cdot \mathbf{v}$$

with condition

$$\begin{cases} \rho(x, 0) = \rho_0(x) \\ \dfrac{dx(t)}{dt} = \mathbf{v}(x, t), \forall t \in [0, T] \\ x(0) = x_0 \end{cases}$$

Compute density dynamics at each trajectory $x(t)$ with an initial value.

MSE(ground truth, estimation) w.r.t. density

Method - Reconstruction Errors

1. Pick a set of samples from the ground truth density at time $x_{t_j} \sim \rho_{t_j}$.
2. Integrate backward to the early time point t_i along the trajectory:

$$\hat{x}_{t_i} = x_{t_j} + \int_{t_j}^{t_i} \boldsymbol{v}(x, t)\, dt, \quad t_i < t_j. \tag{13}$$

3. Obtain ground truth density for these samples at t_i as $\rho_{t_i}(\hat{x}_{t_i})$.
4. Integrated the density from these sample points forward to x_{t_j}:

$$\ln \tilde{\rho}_{t_j}(x_{t_j}) = \ln \rho_{t_i}(\hat{x}_{t_i}) - \int_{t_j}^{t_i} \frac{d \ln p}{dt}\, dt \tag{14}$$

to calculate the estimated density at t_j as $\tilde{\rho}_{t_j}$.

Reconstruction Error

Suppose we have K samples, the **reconstruction error** is denoted as

$$R_{t_i,t_j} = \frac{1}{K}\sum_{k=1}^{K}\left[\tilde{\rho}_{t_j}\left(x_{t_j}^{(k)}\right) - \rho_{t_j}\left(x_{t_j}^{(k)}\right)\right]^2. \tag{15}$$

Combined reconstruction error:

$$R = \underbrace{\sum_{i=1}^{T-1} R_{t_i,t_{i+1}}}_{short-term} + \underbrace{\sum_{i=1}^{T-1} R_{t_1,t_{i+1}}}_{long-term}. \tag{16}$$

The combined reconstruction errors facilitate robust and accurate results by minimizing errors at different time scales.

Method - Overall Loss Function

Loss function:

$$L = W + \lambda_d R \tag{17}$$

with hyperparameter λ_d.

TIGON Algorithm

Require: A series of snapshots $\left(t_1, C^1\right), \left(t_2, C^2\right), \cdots, \left(t_T, C^T\right)$ where $C^i = \left\{c_{t_i}^{(j)}\right\}_{j=1}^{N^i} \in \mathbb{R}^{N^i \times d}$. If relative cell population $\tilde{N}^i$ is not provided, $\tilde{N}^i = \frac{N^i}{N^1}$

Ensure: Neural networks: $(x,t) \to NN_1 \to v(x,t)$ and $(x,t) \to NN_2 \to g(x,t)$

Preprocessing: Using Gaussian mixture model to generate density ρ_{t_i} from snapshot C^i

$$\rho_{t_i}(x) = \frac{\tilde{N}^i}{N^i} \sum_{j=1}^{N^i} \frac{\exp\left(-\frac{1}{2}\left(x - c_{t_i}^{(j)}\right)^T \Sigma^{-1} \left(x - c_{t_i}^{(j)}\right)\right)}{\sqrt{(2\pi)^d |\Sigma|}}, \Sigma = \sigma I \in \mathbb{R}^{d \times d}$$

for *epoch* from 1 to *Epochs* **do**

$Loss = 0$

for i from $T-1$ to 1 **do**

$x_{t_{i+1}} \sim \rho_{t_{i+1}}, x_{t_{i+1}} = \left(x_{t_{i+1}}^{(1)}, x_{t_{i+1}}^{(2)}, \cdots, x_{t_{i+1}}^{(\mathrm{K})}\right)$ ▷ i.i.d sampling

Integrating backward from t_{i+1} to t_i

$$\begin{cases} \dfrac{dx}{dt} = v(x,t) \\ \dfrac{d(z(x,t))}{dt} = g(x,t) - \nabla \cdot v(x,t) \end{cases}, \begin{cases} x\left(t_{i+1}\right) = x_{t_{i+1}} \\ z\left(t_{i+1}\right) = 0 \end{cases}$$

$z_{t_i} = \int_{t_{i+1}}^{t_i} (g(x,t) - \nabla \cdot v(x,t))dt = \int_{t_{i+1}}^{t_i} \frac{d(\ln \rho(x,t))}{dt} dt$

▷ Intermediate variable z_{t_i}

$\ln \tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}\right) = \ln \rho_{t_i} \left(\hat{x}_{t_i}\right) - z_{t_i}$

▷ Estimate $\tilde{\rho}_{t_{i+1}}$

$W_{t_i,t_{i+1}} = (t_{i+1} - t_i)\, \mathbb{E}_{x_i \sim \rho_{t_i}} \int_{t_i}^{t_{i+1}} \left(|v(x,t)|^2 + \alpha |g(x,t)|^2 \right) e^{\int_{t_i}^{t} g(x,s) ds} dt$

▷ Compute transport cost

$R_{t_i,t_{i+1}} = \frac{1}{K} \sum_{j=1}^{K} \left[\tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right) - \rho_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right) \right]^2$

▷ Compute short-term reconstruction error

Integrating backward from t_{i+1} to t_1

▷ Estimate $x\left(t_1\right) = \hat{x}_{t_1}$

$z_{t_1} = \int_{t_{i+1}}^{t_1} (g(x,t) - \nabla \cdot v(x,t))dt = \int_{t_{i+1}}^{t_1} \frac{d(\ln \rho(x,t))}{dt} dt$

▷ Intermediate variable z_{t_1}

$\ln \tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}\right) = \ln \rho_{t_1} \left(\hat{x}_{t_1}\right) - z_{t_1}$

▷ Estimate $\tilde{\rho}_{t_{i+1}}$

$R_{t_1,t_{i+1}} = \frac{1}{K} \sum_{j=1}^{K} \left[\tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right) - \rho_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right) \right]^2$

▷ Compute long-term reconstruction error

$Loss = W_{t_i,t_{i+1}} + \lambda_d R_{t_i,t_{i+1}} + \lambda_d R_{t_1,t_{i+1}}$

end for

Update NN_1 and NN_2 using the Adam optimizer by minimizing the $Loss$

end for

Experiments

1. Benchmark on a three-gene model
2. Model predictions align with lineage tracing experiments
3. Reconstructing cellular dynamics in EMT
4. Identifying bifurcation of directed differentiation in iPSCs

Three-gene Simulation Model

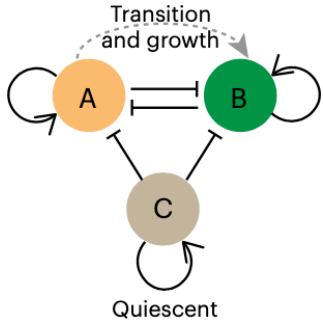


Figure 4: Illustrative diagram of the GRN of the three-gene model.

An in-silico stochastic model based on a three-gene GRN, which consists of three cell states.

Benchmark on a three-gene model

The simulation generates two groups of cells with distinct cell dynamics.

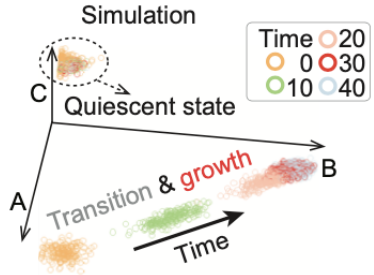


Figure 5: Simulation results at five time points.

- Quiescent state — maintaining tissue balance
- Transition A to B — enhance population growth

Benchmark on a three-gene model

TIGON identifies two groups of cells and growth that are consistent with the ground truth.

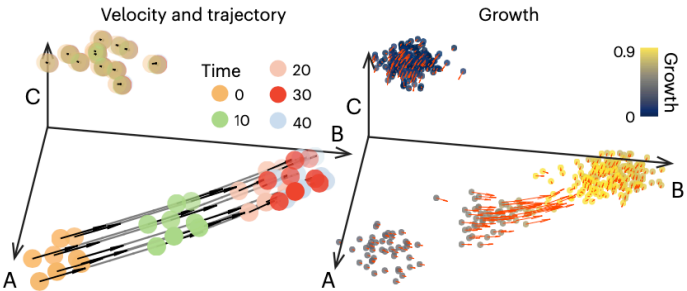


Figure 6: Cellular dynamics for cells sampled at time $t = 0$ by TIGON.

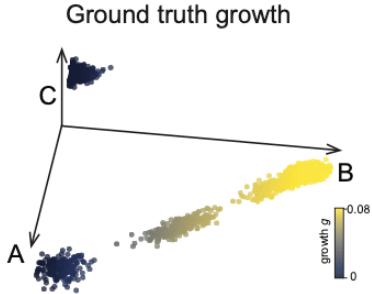


Figure 7: Ground truth values of growth.

Gene analysis for transition cells at time $t = 5$

The gene analysis identifies gene B as the only gene that upregulates growth

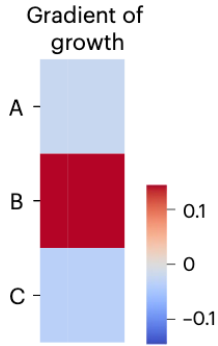


Figure 8: Gradient of growth.

Gene analysis for transition cells at time $t = 5$

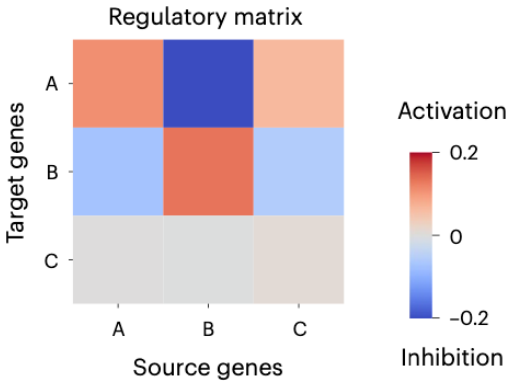


Figure 9: Regulatory matrix.

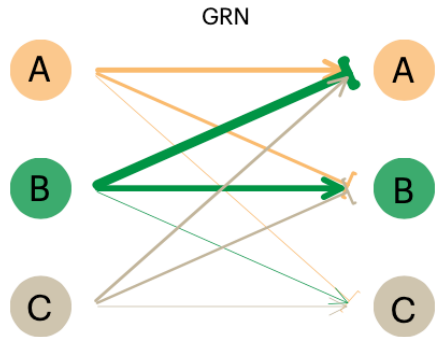
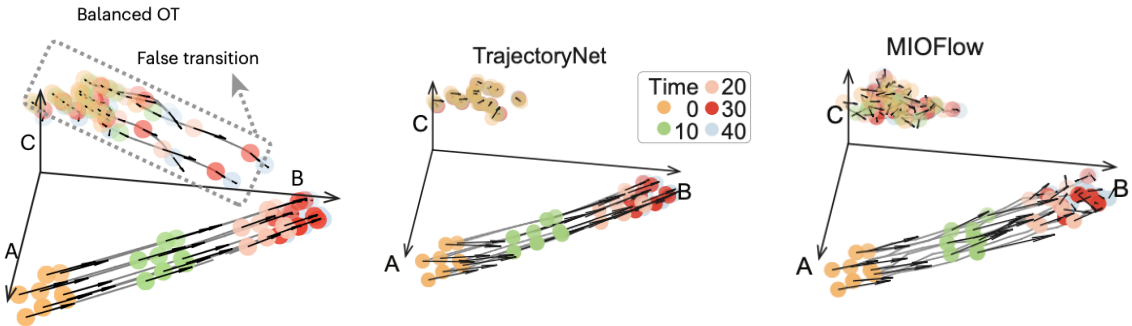


Figure 10: GRN displayed in a form of weighted directed graph.

Compared with three OT-based trajectory inference methods



Benchmark on a three-gene model

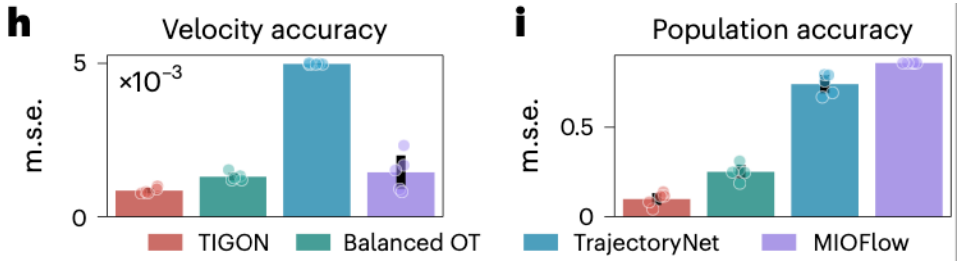


Figure 11: Comparisons between TIGON and OT-based trajectory inference methods measured by accuracy in velocity predictions (h), and accuracy in predicting ratio of cell population (i) between transition cells and quiescent cells .

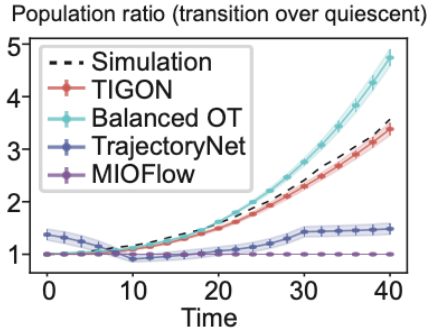


Figure 12: Comparisons of predicted cell population ratio between transition cells and quiescent cells as inferred by TIGON, balanced OT, TrajectoryNet, and MIOFlow.

Compared with other GRN inference methods

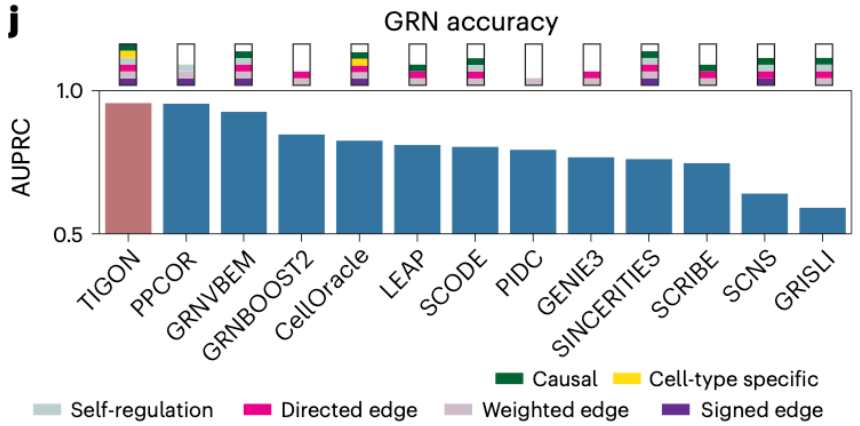


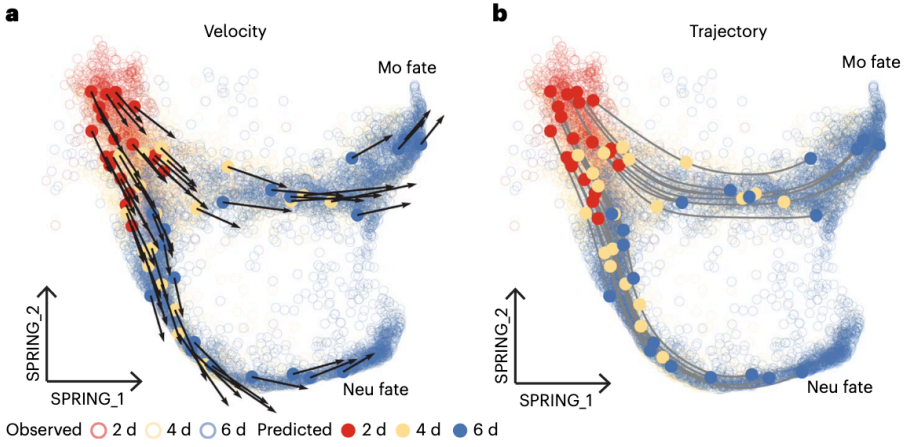
Figure 13: Comparison of GRN inference methods. GRNs are calculated for transition cells at time $t = 0, 10, 20, \ldots, 40$.

Experiments

1. Benchmark on a three-gene model
2. **Model predictions align with lineage tracing experiments**

Model predictions align with lineage tracing experiments

Apply TIGON to a temporal scRNA-seq dataset in **mouse hematopoiesis** using a lineage tracing technique.



Model predictions align with lineage tracing experiments

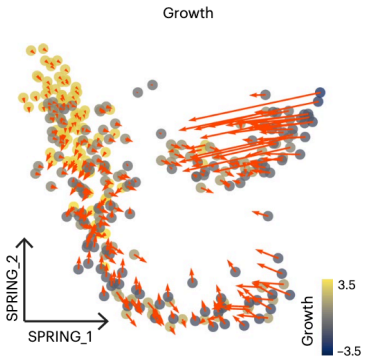
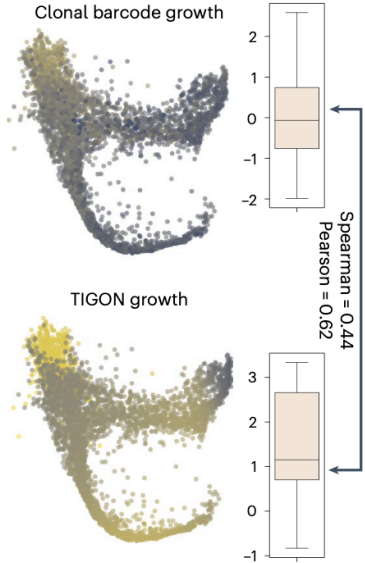
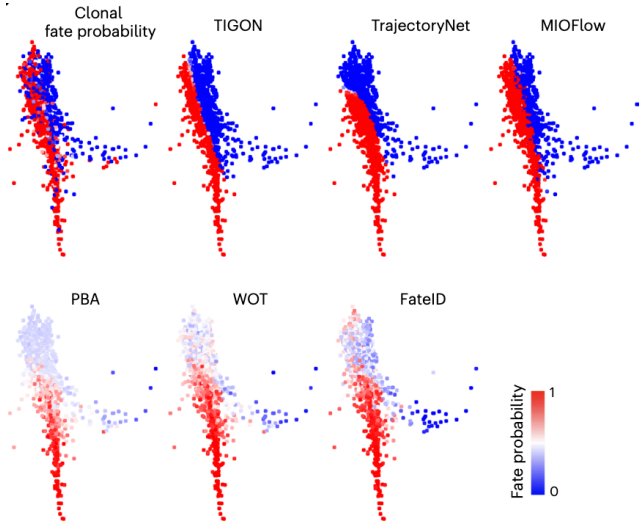


Figure 14: Cellular dynamics inference for growth and gradient of growth

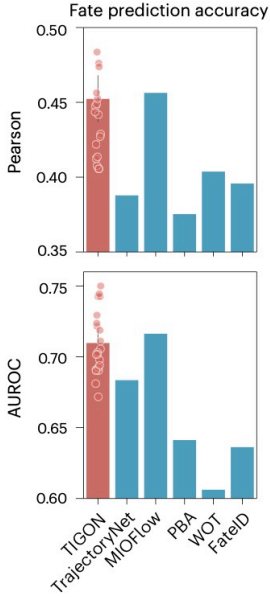


Compared with other trajectory inference methods

Compare the **fate probabilities** for each cell .



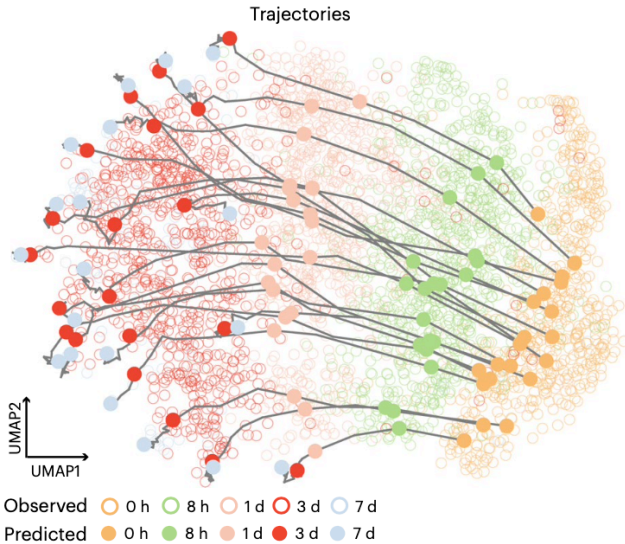
Model predictions align with lineage tracing experiments



Experiments

1. Benchmark on a three-gene model
2. Model predictions align with lineage tracing experiments
3. **Reconstructing cellular dynamics in EMT**

Reconstructing cellular dynamics in EMT



Apply TIGON to a time-series scRNA-seq dataset from an **A549 cancer cell line**, in which cells were exposed to TGFB1 to induce **EMT**(Epithelial-to-Mesenchymal Transition) at the first five time points.

Reconstructing cellular dynamics in EMT

- **decreasing expression level** for two epithelial (E) markers (CDH1 and CLDN1)
- **increasing values** for four mesenchymal (M) markers (VIM, CDH2, FN1 and MMP2) over time

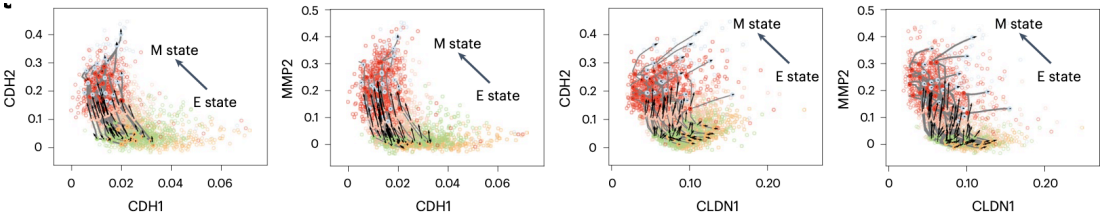


Figure 15: Trajectories and velocity for cells at gene expression space.

Reconstructing cellular dynamics in EMT

Moreover, the patterns of TIGON-inferred growth exhibit higher values at the intermediate stage compared to the epithelial (E) or mesenchymal (M) stage.

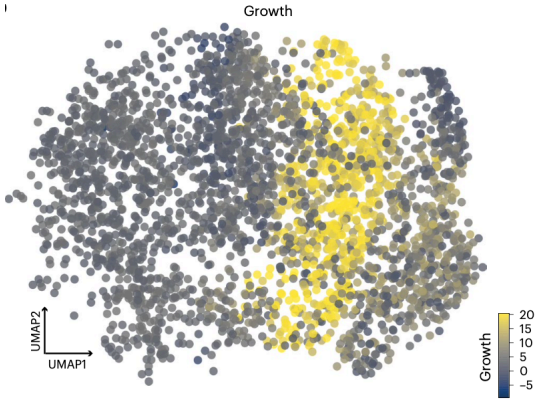
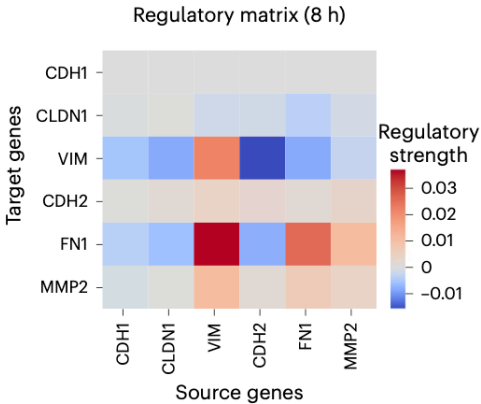
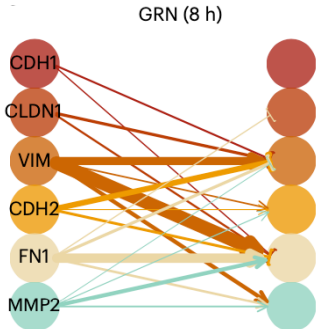


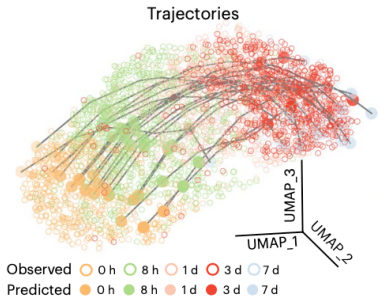
Figure 16: Values of growth for all observed cells.

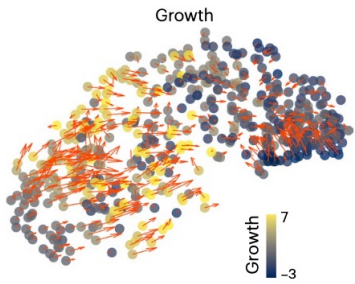
Reconstructing cellular dynamics in EMT

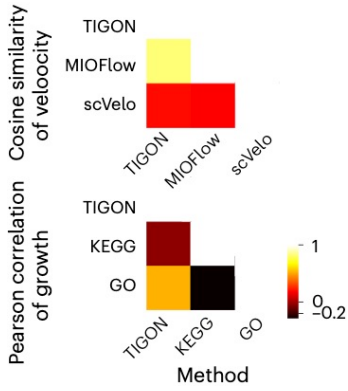




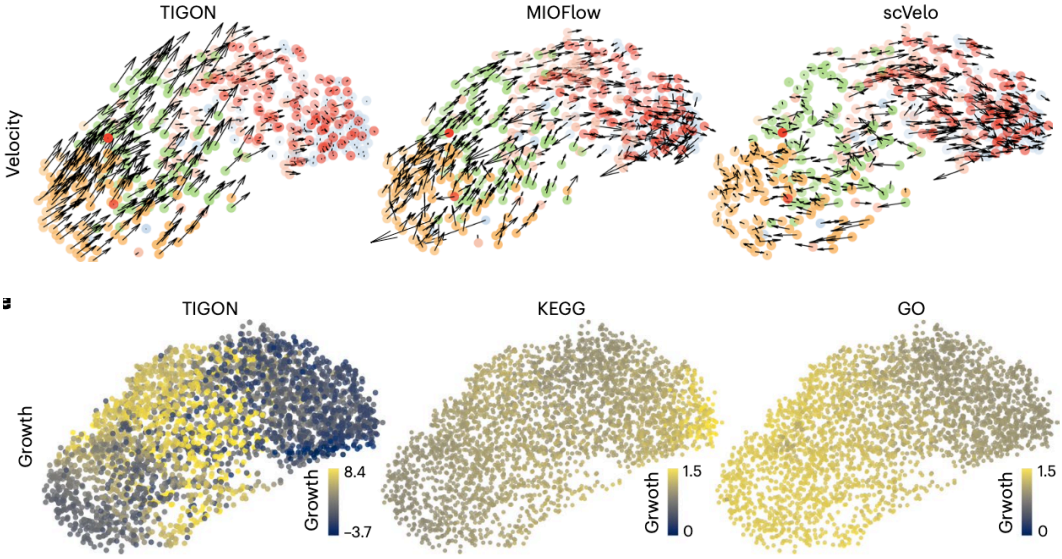
Compare TIGON with other trajectory inference methods







Reconstructing cellular dynamics in EMT



Experiments

1. Benchmark on a three-gene model
2. Model predictions align with lineage tracing experiments
3. Reconstructing cellular dynamics in EMT
4. **Identifying bifurcation of directed differentiation in iPSCs**

Identifying bifurcation of directed differentiation in iPSCs

Study single-cell qPCR datasets at eight time points, showing a bifurcation process for differentiation of iPSCs in cardiomyocytes.

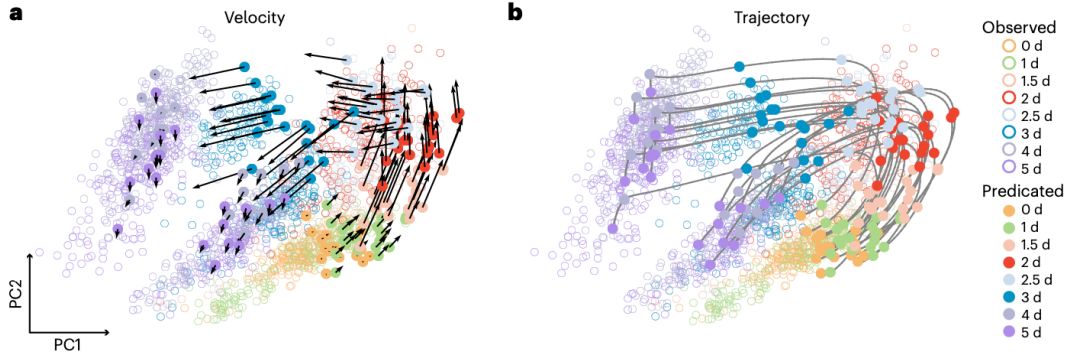


Figure 17: Velocity (a) and trajectories (b) of 20 cells initially sampled from the density at day 0.

Identifying bifurcation of directed differentiation in iPSCs

Large values of growth were observed near the branching time from day 2 to day 3, suggesting a strong dividing potential.

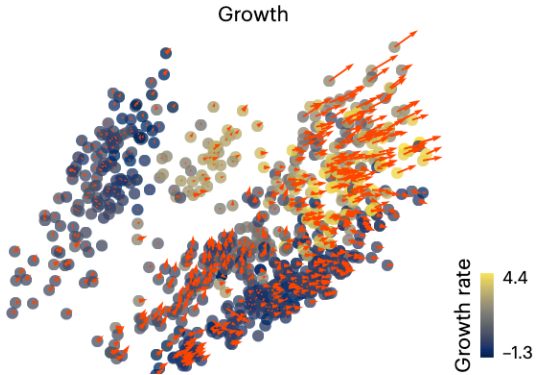


Figure 18: Values of growth and gradient of growth.

Identifying bifurcation of directed differentiation in iPSCs

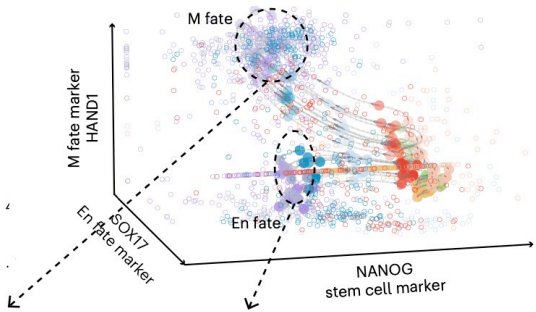


Figure 19: Trajectories of cells on gene expression space of three bifurcation marker genes.

Identifying bifurcation of directed differentiation in iPSCs

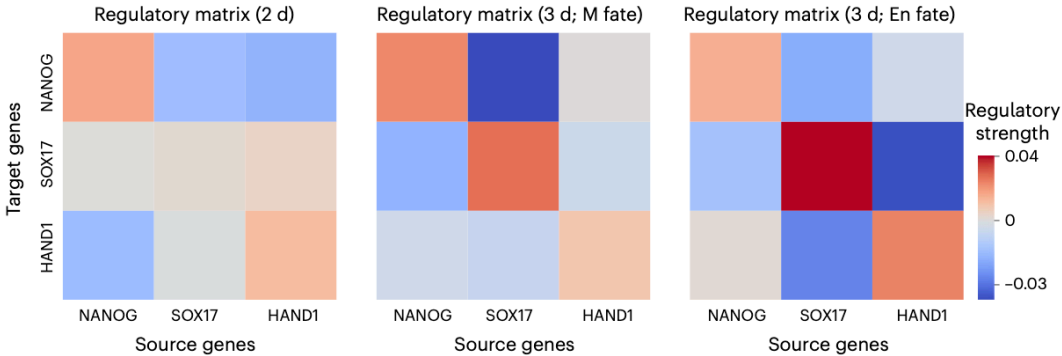


Figure 20: Regulatory metrics for the three marker genes for cells at (left) day 2, (middle) day 3 with M fate and (right) day 3 with En fate.

Identifying bifurcation of directed differentiation in iPSCs

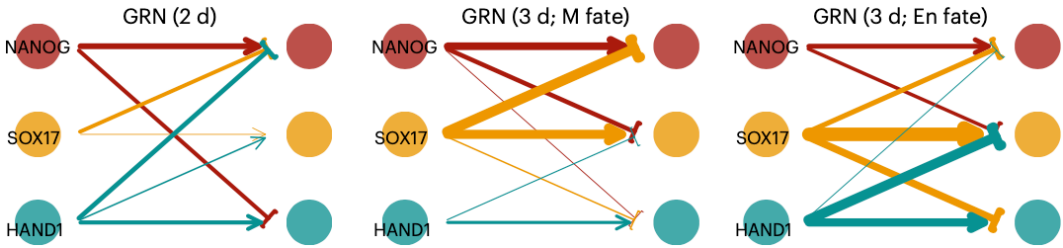


Figure 21: GRNs for the three marker genes for cells at (left) day 2, (middle) day 3 with M fate and (right) day 3 with En fate.

Interestingly, the toggle-switch interaction between HAND1 and SOX17, self-activation and mutual inhibition between two genes, was previously reported. [ref] *Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells, 2017.*

Analyse the contribution of genes to the growth

The top five candidates at day 2 are all previously reported as growth-related genes in the UniProtKB database.

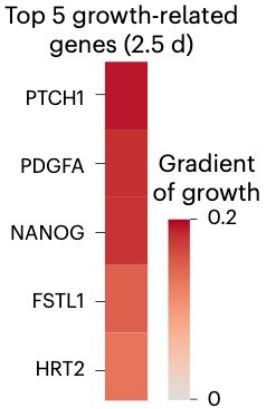


Figure 22: Gradient of growth for top five growth-related genes for cells at day 2.5.

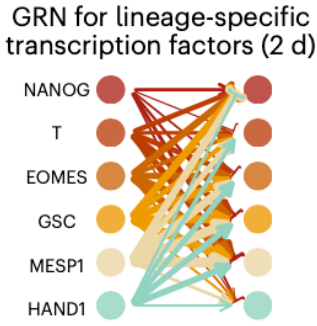


Figure 23: GRN for lineage-specific transcription factors for cells at day 2.

Conclusion

TIGON infers cell dynamics for unpaired time-series scRNA-seq data.

- a dynamic unbalanced OT model
- a mesh-free, dimensionless formulation
- inference of temporal, causal gene regulatory networks (GRNs) and growth-related genes

Thanks!