Identifying Combinatorial Regulatory Genes for Cell Fate Decision via Reparameterizable Subset Explanations

Junhao Liu[1] Pengpeng Zhang[1] Martin Renqiang Min[2] Jing Zhang[1]*

[1]University of California, Irvine [2]NEC Laboratories America *Corresponding Author

Overview

Cell fate decisions are highly coordinated processes governed by complex interactions among numerous regulatory genes, while disruptions in these mechanisms can lead to developmental abnormalities and disease. Traditional methods often fail to capture such combinatorial interactions, limiting their ability to fully model cell fate dynamics.

Contributions

- We introduce MetaVelo, a global feature explanation framework for identifying key regulatory gene sets influencing cell fate transitions.
- MetaVelo models these transitions as a black-box function and employs a differentiable neural ordinary differential equation (ODE) surrogate to enable efficient optimization.
- By reparameterizing the problem as a controllable data generation process, MetaVelo overcomes the challenges posed by the non-differentiable nature of cell fate dynamics.

Results

- MetaVelo outperforms 12 baseline methods in identifying combinatorial regulatory gene sets across diverse stand-alone and longitudinal single-cell RNA-seq datasets and three black-box cell fate models.
- MetaVelo distinguishes: independent regulatory genes and synergistic regulatory genes.
- MetaVelo offers novel insights into gene interactions governing cell fate.

Blackbox of Cell Fate Decision Models

In scRNA-seq experiments, suppose we have a set of cells $\mathcal{E}$ with observed expression profile $\boldsymbol{x} \in \mathbb{R}^d$. Also, the cell type label for each cell is given as $y \in \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, C\}$. Here, d and C represent the number of genes and cell types/states.

Step 1: short-term RNA velocity prediction. Given the gene expression profile $\boldsymbol{x}(t) \in \mathbb{R}^d$ at time t, the RNA velocity $\dot{\boldsymbol{x}}(t)$ represents short term gene expression change predicted by a function v,

$$\dot{\boldsymbol{x}}(t) = v(\boldsymbol{x}(t), t), \tag{1}$$

where $t \in \{0, \ldots, T\}$ represents the latent time along the trajectory, and v is either a statistical model or a neural network.

Step 2: long-term cell trajectory inference. With the trained v, starting from a particular cell state $\boldsymbol{x}(t)$ at time t, the longer-term developmental trajectories ending at terminal cell state $\boldsymbol{x}(T)$ is evolved through an iterative cell transition process $\mathcal{T}$:

$$\begin{aligned} \boldsymbol{x}(t+1) &= \mathcal{T}_{t+1}(\boldsymbol{x}(t)) = \mathcal{Z}(\boldsymbol{x}(t), v(\boldsymbol{x}(t), t)), \\ \boldsymbol{x}(T) &= \mathcal{T}_{t:T}(\boldsymbol{x}(t)) = \mathcal{T}_T(\cdots \mathcal{T}_t(\boldsymbol{x}(t)) \cdots), \end{aligned} \tag{2}$$

where $\mathcal{Z}$ is the short term cell state transition function. $\mathcal{Z}$ involves discrete operations and is thus essentially non-differentiable, imposing computational challenges in the feature explanation step.

Identify Developmental Regulatory Genes via Global Feature Importance Explanation

Along the cell developmental trajectories, we assume that perturbing gene expression at any initial or intermediate state can disrupt the tightly regulated cell fate decision process $\mathcal{T}$, possibly leading to altered terminal states.

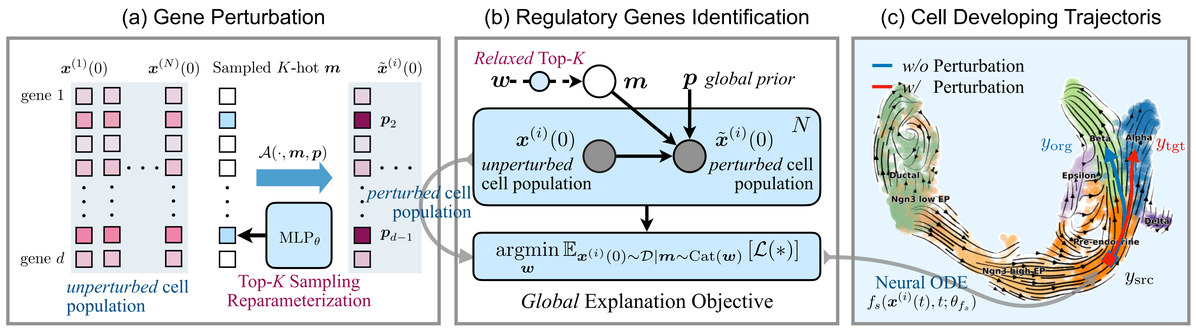

Figure 1. The overview of MetaVelo for combinatorial cell fate regulatory genes identification.

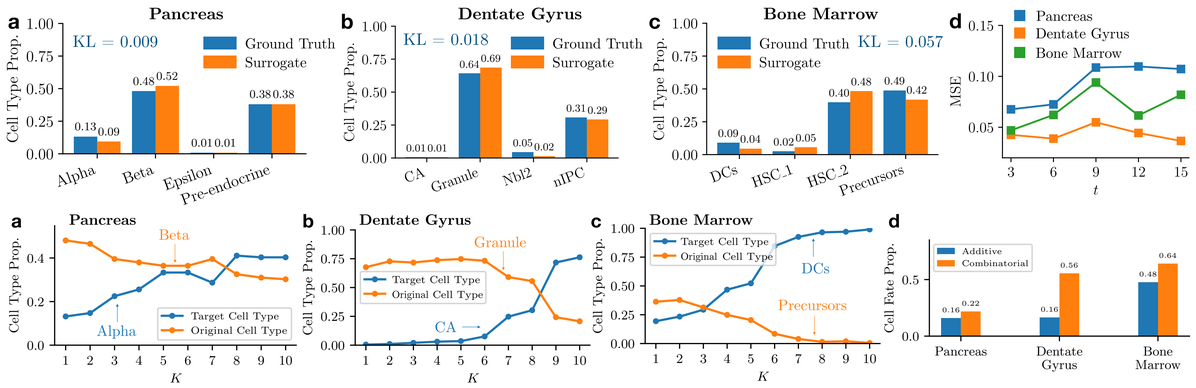

Figure 2. **Top:** MetaVelo's surrogate function accuracy evaluations. **Bottom:** (a-c) The cell type proportion changes with different K values. (d) Desired terminal cell state proportions under additive and combinatorial models.

Regulatory Genes Our objective is to identify a set of K genes, $\mathcal{X}^* = \{\mathcal{X}_1^*, \ldots, \mathcal{X}_K^* \mid \forall \mathcal{X}_i^* \in \{1, \ldots, d\}\}$, whose combinatorial perturbation induces most cells of state y_{src} to transition to the new terminal state y_{tgt} instead of the original terminal state y_{org}.

Global Feature Importance Explanation for Identifying Developmental Regulatory The developmental regulatory genes identification can be formulated as the following perturbation-based global feature explanation problem:

$$\underset{\boldsymbol{m}}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{x}^{(i)}(0)} \left[\mathcal{L}(g(\mathcal{T}_{1:T}(\mathcal{A}(\boldsymbol{x}^{(i)}(0), \boldsymbol{m}, \boldsymbol{p}))), y_{\mathrm{tgt}}) \right], \tag{3}$$

where $g(\cdot) : \mathbb{R}^d \to \mathbb{R}^C$ is a classifier that takes the gene profile $\boldsymbol{x}$ as input and predicts the cell type categorical distribution $\boldsymbol{y} \in \mathbb{R}^C$, $\mathcal{L}$ is the cross-entropy loss measuring the negative log-probability of the cell $\tilde{\boldsymbol{x}}^{(i)}(0)$ perturbed by the subset encoded in $\boldsymbol{m}$ finally evolved into the target cell type y_{tgt}.

Surrogate We use the cell development trajectory data $\mathcal{D}_T = \{\xi^{(i)} = (\boldsymbol{x}^{(i)}(0), \ldots, \boldsymbol{x}^{(i)}(T))\}_{i=1}^N$ as the supervised data, which are sampled from all cells in $\mathcal{D}$ belonging to the class y_{src}. To learn a good differentiable approximation, we adopt the neural ODE to model the sampled time-series data $\xi^{(i)}$, which is formulated as: $\hat{\boldsymbol{x}}^{(i)}(t) = \boldsymbol{x}^{(i)}(0) + \int_0^t f_s(\boldsymbol{x}^{(i)}(t), t; \theta_{f_s}) dt$, where $\hat{\boldsymbol{x}}^{(i)}(t)$ is the predicted gene expression profile at time t, θ_{f_s} is the trainable parameters in the surrogate model. The mean squared error loss is applied to train the surrogate model.

Discrete Combinatorial Optimization as A Controllable Generative Process This process is formulated as $\tilde{\boldsymbol{x}}(0) \sim \mathcal{A}(\boldsymbol{x}(0), \boldsymbol{m}, \boldsymbol{p})$. We let the K-hot mask $\boldsymbol{m}$ to be a random variable that follows a latent categorical distribution $\boldsymbol{m} \sim \mathrm{Cat}(\boldsymbol{w})$ with where $\boldsymbol{w} = \{w_1, \ldots, w_d\}$. This parameterization enables us to sample K-hot mask $\boldsymbol{m}$ from a categorical distribution $\mathrm{Cat}(\boldsymbol{w})$ via sampling without replacement (i.e., the top-K categories are selected). Then the generative model of the perturbed data is trained to maximize the likelihood of cells developing into terminal state y_{tgt}:

$$\mathbb{E}\left[\mathcal{L}(g(\mathcal{A}(\boldsymbol{x}^{(i)}(0), \boldsymbol{m}, \boldsymbol{p}) + \int_0^T f_s(\tilde{\boldsymbol{x}}^{(i)}(t), t; \theta_{f_s}) dt), y_{\mathrm{tgt}}) \right], \tag{4}$$

where the expectation is computed over $\boldsymbol{x}^{(i)}(0) \sim \mathcal{D} | \boldsymbol{m} \sim \mathrm{Cat}(\boldsymbol{w})$.

Differentiable top-K Sampling For continuous relaxation of K-subset sampling, the Gumbel-max trick is adopted to reparameterize the sampling of the categorical distribution. Given the categorical weight $\boldsymbol{w}$, the Gumbel noise is added to the categorical weight as follows:$G_i = -\log(-\log u_i)$, $u_i \sim \mathrm{U}(0, 1)$, $r_i = \log w_i + G_i$, where $\boldsymbol{r} = \{r_1, \ldots, r_d\}$ is the unnormalized weight of the relaxed categorical distribution, and G_i is the Gumbel noise. The top-K relaxation proposed by Plötz T, Roth S provides a differentiable top-K sampling procedure with respect to the unnormalized weight $\boldsymbol{r}$. We propose to reparameterize the distribution parameter $\boldsymbol{w} = \mathrm{MLP}_\theta(\mathbf{E}_\theta)$ by a latent embedding matrix ($\mathbf{E}_\theta$) paired with a MLP_θ to avoid unstable optimization and failed convergence.

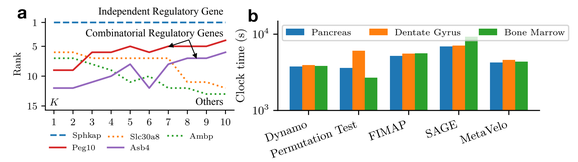

Figure 3. (**a**) The learned gene rank in the order of the categorical weight $\boldsymbol{w}$ on the Pancreas dataset. (**b**) The clock time (in log scale) of searching $K = 10$ genes for each method.

Experiment Results

- MetaVelo's Surrogate Function Accurately Predicts Cell Trajectories in Both Settings (Figure 1 Top)
- Multiple Combinatorial Genes Are Required for Effective Modulation of Cell Fates (Figure 1 Bottom)
- MetaVelo Outperforms Existing Methods in Modifying Cell Development Trajectories (Table 3 in the paper)
- MetaVelo Highlights Developmental Regulatory Genes with Biology Insights (Figure 3(a) and Table 4 in the paper)
- MetaVelo Improves Computational Efficiency in Considering Combinatorial Effects (Figure 3(b))

Key Takeaway Messages

- **Challenge Addressed:** Identifying key regulatory gene sets that determine cell fate is difficult due to inefficiencies in exhaustive perturbation experiments and the complexity of gene interactions.
- **Core Innovation:** MetaVelo reframes the problem as a global feature explanation task, enabling identification of combinatorial gene interactions rather than individual markers.
- **Methodology:** Uses an adversarial perturbation model to simulate gene knockouts by sampling top-K genes from a learnable distribution and applying global perturbations.
- **Optimization Strategy:** Leverages stochastic backpropagation to solve a combinatorial, discrete objective efficiently and differentiably.
- **Evaluation:** Validated on diverse stand-alone and longitudinal scRNA-seq datasets and multiple cell fate decision models.
- **Outcome:** Achieves both computational efficiency and biological interpretability, making it suitable for guiding wet-lab validation and advancing developmental biology.

Links

Paper GitHub Zhang Lab Job Openings

Acknowledgments

This work was supported by the National Institutes of Health [R01HG012572, R01NS128523].

Email: {junhao.liu, zhang.jing}@uci.edu

GitHub: aicb-ZhangLabs/MetaVelo