NAR Genomics and Bioinformatics, 2025, **7**, lqaf048
https://doi.org/10.1093/nargab/lqaf048
Computational Methods for Transcriptome Analysis

OXFORD

Discovering governing equations of biological systems through representation learning and sparse model discovery

Mehrshad Sadria [1,*] **and Vasu Swaroop** [2]

[1]Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[2]Department of Computer Science Information Systems, BITS-Pilani, Pilani Campus, Pilani 333031, India
[*]To whom correspondence should be addressed. Email: msadria@uwaterloo.ca

Abstract

Understanding the governing rules of complex biological systems remains a significant challenge due to the nonlinear, high-dimensional nature of biological data. In this study, we present CLERA, a novel end-to-end computational framework designed to uncover parsimonious dynamical models and identify active gene programs from single-cell RNA sequencing data. By integrating a supervised autoencoder architecture with Sparse Identification of Nonlinear Dynamics, CLERA leverages prior knowledge to simultaneously extract related low-dimensional representation and uncover the underlying dynamical systems that drive the processes. Through the analysis of both synthetic and biological data, CLERA demonstrates robust performance in reconstructing gene expression dynamics, identifying key regulatory genes, and capturing temporal patterns across distinct cell types. CLERA's ability to generate dynamic interaction networks, combined with network rewiring using Personalized PageRank to highlight central genes and active gene programs, offers new insights into the complex regulatory mechanisms underlying cellular processes.

Introduction

Across many scientific disciplines, discovering governing equations has traditionally served as the cornerstone of understanding systems [1]. Derived from mathematical and physical laws, these equations provide interpretable and generalizable frameworks for explaining and predicting various phenomena. In areas such as biology [2], epidemiology [3], and finance [4], mathematical models are used to model signalling pathways, population dynamics and disease spread, and market fluctuations, respectively. However, for complex systems with high dimensionality and nonlinearity, including biological processes, traditional approaches often fall short [5]. Discovering the main equations governing these systems can be challenging, and even when partial knowledge exists, relying solely on first principles becomes impractical [6].

The modern era, with its abundance of data and computational power, has facilitated the emergence of data-driven model discovery as a powerful paradigm in scientific exploration [7]. This approach directly leverages data to uncover the hidden principles that govern complex systems. In the context of cellular biology, single-cell RNA sequencing (scRNA-seq) provides an unprecedented window into individual cells, which offers insights into gene expression variation across diverse cellular populations [8]. By analysing this data, researchers can investigate the molecular machinery underlying development, disease, and response to external perturbation [9, 10]. The noisy, nonlinear, and high-dimensional characteristics of scRNA-seq data and the biological processes it captures pose significant challenges for analysis and interpretation [11]. These complexities make it difficult to uncover the underlying principles of biological processes and pinpoint their key drivers [12]. Recent advances in modeling cell state transitions have sought to address these challenges. Methods such as Vector Field Reconstruction [13], DeepVelo [14], and PRESCIENT [15], along with approaches based on optimal transport [16] and transition path and dynamical systems theory [9, 17], have leveraged scRNA-seq data to infer cellular dynamics effectively. Some of these, including PRESCIENT [15] and Deep Lineage [18], incorporate generative AI frameworks to predict cell trajectories and simulate potential interventions. These methods are highly effective at reconstructing velocity fields, modeling cell dynamics, and predicting cellular responses across various scenarios. While they have achieved success in specific tasks, limitations remain. For instance, the correlative nature of most methods prevents them from capturing causal features and true representations, thus limiting their generalizability [19]. Furthermore, these models struggle to discover governing relationships among underlying variables in a parsimonious manner, similar to classical physics settings, which further hinders true interpretability. Therefore, a crucial step in understanding any biological process lies in developing models that not only can accurately predict but also reveal the underlying connections between features in an interpretable and parsimonious manner [20]. This ensures the models can be applied across diverse environments and

Received: October 8, 2024. **Revised:** March 19, 2025. **Editorial Decision:** April 2, 2025. **Accepted:** April 11, 2025

provides clearer insights into the mechanisms governing the process [1].

In the realm of high-dimensional biological data like scRNA-seq, the ability to capture causal representations of the data becomes particularly valuable. This approach goes beyond identifying correlations and allows us to understand the true relationships between variables [21]. In this context, identifiability, the ability to uniquely recover the underlying causal structure from observed data, becomes a crucial aspect of representation learning. Traditional methods like Independent Component Analysis (ICA) have achieved success in many areas of linear representation learning [22]. In fact, if all latent components are non-Gaussian and independent, ICA can be identifiable. However, ICA struggles with the inherent nonlinearities and complex interactions present in biological data [22]. While perfect identifiability, especially in nonlinear settings, remains a challenge, incorporating temporal structure, employing additional tasks, or using auxiliary information can facilitate the way to attain identifiability [23, 24]. Notably, autoencoders can offer a promising avenue for achieving identifiability in nonlinear settings [25]. By carefully designing their architecture and loss function, autoencoders can help extract meaningful representations from complex biological data [18, 26–28].

In this work, we present CLERA (Cellular Latent Equation and Representation Analysis), a novel end-to-end computational framework that combines the power of data-driven model discovery, specifically Sparse Identification of Nonlinear Dynamics (SINDy), and representation learning. Leveraging a supervised autoencoder architecture, CLERA simultaneously extracts a compact and relevant representation from high-dimensional data and uses it to discover the underlying low-dimensional, nonlinear dynamical model governing the system. This learned embedding further allows us to not only identify active gene programs and key genes but also track their transitions over time across cell types, providing insights into the complex dynamic mechanisms of biological systems. We validate CLERA's performance on both simulated data (with known active gene programs) with different sizes and real-world biological datasets.

Materials and methods

Autoencoders

An autoencoder is a neural network architecture consisting of an encoder and a decoder [29]. The encoder compresses high-dimensional input data into a lower-dimensional latent representation reducing dimensionality while preserving essential information. The decoder reconstructs the original data from this latent representation. Regularizations and architectural choices can be applied to constrain the latent embeddings, influencing their quality and utility. This flexibility allows autoencoders to be used for dimensionality reduction, feature learning, and data denoising.

Training involves minimizing the reconstruction error, typically measured by mean squared error, to align the input with its reconstruction. For an input X, the encoder E produces latent embeddings $z = E(X)$, which the decoder D uses to reconstruct the input as $\hat{X} = D(z)$. The reconstruction loss guides the optimization process so that the latent layer accurately captures the input data. The reconstruction loss for inputs is given by

$$\text{Reconstruction loss} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2$$

SINDy and SINDy autoencoder

SINDy is a regression technique used to discover a best fitting dynamical system from training data [7]. It takes input data $x_t \in R^n$ and predicts a system of ordinary differential equations (ODEs) as $\frac{dx_t}{dt} = f(x_t)$, where f is a library of candidate functions composed of basis functions that we want SINDy to model. SINDy selects a subset of active terms from these candidates, promoting parsimony in the model, resulting in a simpler and more interpretable model that captures the essential dynamics. For m input snapshots, we stack them to form $X = [x_1, x_2, \cdots, x_m]^T$ and compute the corresponding derivatives $\dot{X} = [\dot{x}_1, \dot{x}_2, \cdots, \dot{x}_m]^T$. The library is constructed as $\Theta(X) = [1, \theta_1(X), \theta_2(X), \cdots, \theta_p(X)] \in R^{m\times p}$, where p is the length of the candidate library, and θ_i are the candidate model terms. The coefficient matrix can be represented as $\Xi = |\xi_1, \xi_2, \cdots, \xi_n| \in R^{p\times n}$, where ξ_i contains the coefficients for the library functions of differential equation i.

To address the challenge of high-dimensional data, SINDy autoencoders leverage an autoencoder to find a low-dimensional state $z = \phi(x)$, where x is the time-series input, z is the latent embedding, and ϕ is the encoder [1]. We incorporate sparse regression into the autoencoder training to facilitate the simultaneous discovery of latent embeddings and the corresponding ODE. The SINDy Autoencoder finds a latent dimension $z_t = \phi(x_t) \in R^d$, where d is the size of the latent embedding, and an associated dynamic model $\frac{dz_t}{dt} = f(z_t) = \Theta(z_t)\Xi$, where f has few active terms.

We train the autoencoder with additional SINDy losses as regularizations to ensure that the latent embeddings have a parsimonious model representation. The reconstruction loss, $L_{recon} = ||x - \psi(\phi(x))||_2{}^2$ is used which ensures accurate embeddings, while the SINDy weight regularization, $L_{SINDyReg} = ||\Xi||_1$ promotes sparsity of the coefficients. We also define two SINDy losses: $L_{dz/dt} = ||\nabla_x\phi(x_t)\dot{x}_t - \Theta(z_t)\Xi||_2{}^2$, which enforces accurate prediction of dynamics, and $L_{dx/dt} = ||\dot{x}_t - \nabla_z \Psi(z_t)\Theta(z_t)\Xi||_2{}^2$, which reduces the error between the real-time derivatives of the input and the predicted derivatives using SINDy. We threshold the coefficients below a certain magnitude at regular intervals to promote parsimony. Finally, we conduct an additional refinement phase without L_1 loss and thresholding for a few epochs to refine the coefficients.

Overview of CLERA

CLERA is a deep learning-based model discovery method designed to uncover parsimonious and interpretable dynamical systems and active gene programs. It employs a specialized loss function to facilitate the joint discovery of a low-dimensional, interpretable latent space. The method's architecture integrates multiple key components to analyse gene expression data. At its core, CLERA processes gene expression data to generate a cell-by-latent embeddings matrix. This latent representation serves as a crucial intermediate step, enabling three primary functions:

1. Data reconstruction: A decoder network uses latent embeddings to reconstruct the original count matrix.

2. Cell-type classification: The latent space is used to classify distinct cell types within the dataset.
3. Dynamical system discovery: CLERA incorporates SINDy to uncover a system of ODEs. These ODEs describe the temporal dynamics and interactions of the latent embeddings over the underlying biological processes.

This integration allows CLERA to simultaneously capture cell identities, gene expression patterns, and temporal dynamics.

Unsupervised learning of identifiable nonlinear representations has long been recognized as theoretically impossible without incorporating inductive biases or suitable constraints on the model [22]. This fundamental challenge in representation learning has led researchers to explore various approaches to learn meaningful and causal representations of complex data. Autoencoders have emerged as a powerful tool for unsupervised representation learning, capable of discovering useful coordinate transformations and dimensionality reductions. However, while autoencoders can be trained in isolation, there is no guarantee that the learnt coordinates will be associated with sparse and physical dynamical models. To address this, CLERA integrates multiple components to learn interpretable and dynamically meaningful representations. By simultaneously training an autoencoder and a SINDy model, we constrain the network to learn coordinates associated with parsimonious dynamics. While CLERA includes a classifier, its purpose differs from traditional cell type annotation methods [10, 30, 31]. Instead, the classification task serves as an auxiliary objective, guiding the latent space to better reflect biologically relevant processes. By leveraging additional information during training, this approach helps constrain the space of possible latent variables, making the inferred representations more interpretable and biologically meaningful. This integration introduces inductive biases that enhance CLERA's ability to uncover governing equations describing the interactions between these latent variables.

Data preprocessing

To preprocess the cell-by-gene count matrix for pancreas and bone marrow, we applied a preprocessing pipeline that includes normalization, filtering, and transformation steps. Genes that were expressed in fewer than the 20 number of cells or genes with <20 total counts were filtered out to improve data quality. The counts for each cell were normalized by scaling the total gene expression counts for each cell to the median total counts across all cells. A log1p transformation was then applied to stabilize variance in the gene expression data which was then used to get the 2000 most highly variable genes based on their dispersion. The resulting matrix is analysed using pseudotime [32] to establish a temporal sequence in the data and enable the conversion of static gene expression profiles into a dynamic time series which is essential for SINDy modelling. Given the sensitivity of SINDy to noise and the inherently noisy nature of scRNA-seq data, the time series is constructed by averaging gene expression counts over a sliding window of four-time steps (pseudo cell) with a stride of two steps. We experimented with multiple sliding window lengths but found that smaller window sizes would result in less effective denoising, while larger window sizes would lead to insufficient training data. Ultimately, a window length of 4 was chosen as the optimal balance between effective denoising and maintaining sufficient training data (Supplementary Note 1). This averaging process helps to smooth the data, reduce noise, and increase power. During this process, the cell type within each window is determined and assigned to ensure that cell identity is preserved. Derivatives of gene expression are calculated under the assumption of unit time steps between cells, a reasonable assumption given the continuous nature of pseudotime. This allows for consistent derivative estimation across the trajectory. Constructing the time series requires careful selection of cell types based on biological relevance, as some cell types may not contribute meaningfully to the dynamics being studied. This selection ensures that the resulting time-by-gene matrix accurately reflects the biological processes of interest.

Components of CLERA objective function

The preprocessed time series results in a pseudo-cell by gene matrix. CLERA takes each step as input to the autoencoder and generates reconstructions of the time-averaged gene expression for that box. These steps form a batch, consisting of 1292 samples in the Pancreas dataset and 1172 samples in the hematopoietic differentiation dataset. During training, the derivatives calculated during preprocessing are used to calculate the SINDy losses while the cell type serves as the label for the classification network.

During classification, the latent representation is passed through a classification network, which outputs the probability of the embedding belonging to each cell type. The cell types are represented as labels $l = [l_1\ l_2 \cdots l_c]$ where c is the total number of cell types. The classification network C produces the probability of the input belonging to each cell type, $y = C(z)$ where $y = [\widehat{y_1}\widehat{y_2} \cdots \widehat{y_c}]$ and $\widehat{y_j}$ is the probability of the label being j. Multi-class cross-entropy loss $L_{class} = -\sum_{j=1}^{c} y_j \log(\widehat{y_j})$ is used for this task. To improve stability during training, L1 regularization $L_{NetReg} = ||\varphi||_1 + ||\psi||_1 + ||C||_1$ is applied to the network weights. This regularization encourages sparsity in the weights, reducing the risk of overfitting and improving the model's generalization capabilities.

The loss function for CLERA is a weighted sum of six losses: Reconstruction loss, Classification loss, SINDy loss over the latent embedding, SINDy loss over the reconstructed genes, SINDy weight regularization, and Network weight regularization which can be represented as

$$L = \lambda_{recon} L_{recon} + \lambda_{\frac{dz}{dt}} L_{\frac{dz}{dt}} + \lambda_{\frac{dx}{dt}} L_{\frac{dx}{dt}} + \lambda_{class} L_{class} + \lambda_{SINDyReg} L_{SINDyReg} + \lambda_{NetReg} L_{NetReg}$$

where different $\lambda_{loss\ type}$ represent the loss weights.

Modelling choices and training

The model hyperparameters significantly influence the nature of the resulting equations and require careful tuning to ensure effective training and accuracy of CLERA (refer Supplementary Table S3 for the sensitivity analysis and ablation of hyperparameters for the pancreas data). Several conditions are essential for this process. First, we monitor the loss curves and the number of active SINDy coefficients closely, as they provide key insights that guide refinement. Second, the functional forms used in the model are designed to align with established prior knowledge of the process. Additionally, we make assumptions to maintain the integrity of the system's dynamics: the right-hand side of every equation in the ODE

should be nonzero, as a zero value would indicate that SINDy was unable to model the temporal features for that latent variable. Moreover, the discovered equations should be parsimonious, balancing expressiveness and simplicity. The dimension of the autoencoder's latent layer plays a crucial role in this balance, which shows the trade-off between model complexity and parsimony. As the size of the latent dimensions increases, the SINDy library expands quadratically. Therefore, selecting the appropriate latent dimension is crucial, as a larger latent space can lead to more complex candidate functions, which complicates the model and increases the risk of overfitting. Experiments with higher-dimensional latent spaces (e.g. dimensions of 10 or 15) reveal that the model consistently produces ODEs where the right-hand side is zero for some of the latent variables. This indicates that higher latent dimensions produce variables that are less critical for capturing the dynamics of the system. Based on these findings, the final model uses a latent space of six dimensions, which provides a balance between capturing essential dynamics and avoiding unnecessary complexity (Supplementary Note 1). The autoencoder and classification weights are initialized with Xavier initialization which is known to promote stability and efficient learning. We train both the pancreas and bone marrow datasets 50 times each, selecting the best model and parameters based on the conditions and criteria outlined above, such as the lowest loss, parsimony, and model complexity. This approach ensures that, unlike studies relying on a single initialization, we systematically explore multiple initializations to infer the most accurate latent variables and their governing equations. Due to differences in initialization, certain training runs may yield slightly different differential equations or rankings of latent variables. However, we mitigate potential instability by enforcing parsimony and systematically filtering out degenerate solutions (e.g. equations containing either no functional terms or all possible terms). While minor variations across runs are possible, the core dynamical structure consistently remains biologically interpretable and reproducible.

In modelling complex biological systems, it is crucial to leverage the prior knowledge we have of the system to carefully select a library of candidate functions. This selection should include functions capable of capturing a wide range of dynamics. Studies have considered various terms, such as polynomials up to degree 2, Michaelis–Menten kinetics, fractional terms, and exponential functions, to model nonlinear and dynamic processes [33–37]. Building on these approaches, we carefully select our SINDy library to include linear, second-order polynomial terms along with exponential, sinusoidal and fractional function forms. The sinusoidal terms are crucial for capturing oscillatory behaviours, while the exponential terms effectively model processes involving growth or decay. The fractional function helps with modelling nonlinear behaviours typically observed with biological processes, such as Michaelis–Menten kinetics (refer to the Supplementary Table S2 for a list of candidate terms used). The SINDy model coefficients are randomly initialized with 1 and −1 which ensures that there is no bias towards one particular direction (positive or negative coefficients). This makes the model better at discovering the true underlying relationships in the data. Refer to Supplementary Table S1 for different hyperparameters that are chosen while training CLERA on pancreas and bone marrow datasets.

In fields like NLP and computer vision, transfer learning has been demonstrated to result in better representation of underlying patterns, faster convergence during training, and higher accuracy compared to models initialized randomly [38, 39]. Accordingly, for the bone marrow dataset, the initialization of the autoencoder weights and SINDy coefficients is performed within a transfer learning framework. The SINDy library and autoencoder architecture are consistent across both the pancreas and bone marrow, ensuring that the trained weights and SINDy coefficients of the pancreas can be used as initialization for bone marrow. CLERA is initially trained on the Pancreas dataset using 15 different random initializations, resulting in 15 distinct instances of trained Pancreas weights. These trained instances are then evaluated using bone marrow data. The top six experiments that exhibit the lowest losses when tested on the bone marrow data are selected. From these six experiments, an average of the Autoencoder weights and SINDy coefficients is computed. This averaged set of weights and coefficients is used as the initialization for subsequent bone marrow training. By leveraging transfer learning in this manner, the model benefits from faster convergence and improved performance, as the pre-trained weights provide a more informed starting point (Supplementary Fig. S4).

Another key difference between the pancreas and bone marrow training lies in the SINDy thresholding interval. For the bone marrow dataset, the thresholding interval at 500 epochs is set higher than that for the pancreas dataset which is kept at 150. This adjustment is necessary as the SINDy coefficients derived from the pancreas experiments exhibit magnitudes much lower than 1, due to the application of L1 regularization. By setting a higher threshold for the bone marrow training, the network is better at adjusting and correcting the coefficient magnitudes, which ensures that key SINDy terms do not undergo early thresholding.

In the unsupervised training setting, where there is an absence of explicit target ODE terms and structure, an inductive bias is introduced to determine the appropriate number of training steps. A maximum acceptable term threshold is established, defining the maximum number of active SINDy coefficients permitted on the right-hand side of the ODE. During training, once the number of active coefficients falls below this threshold, the refinement phase is initiated, consisting of an additional 150 epochs. This threshold is kept at 30 to ensure that the model remains parsimonious.

Personalized Pagerank

PageRank is an algorithm that ranks nodes in a graph by evaluating the number and weight of links between them. Nodes that are connected to highly-ranked nodes are deemed more important and thus receive a higher PageRank score. The standard PageRank algorithm calculates a global importance score, treating all nodes equally in the process.

Personalized PageRank (PPR), on the other hand, is an extension of this algorithm that computes node rankings relative to a specific node or set of nodes within the graph. Unlike traditional PageRank, which provides a uniform importance score across the entire graph, PPR introduces a bias toward the selected nodes. This bias allows for the calculation of personalized importance scores, emphasizing the relevance of nodes in relation to the chosen node(s) of interest.

The interaction graph is constructed by including the gene–latent variable interaction using the top "K" SHAP for each latent variables and inter-latent-variable interaction via the ODE. As cells progress further in pseudotime and become

more dissimilar, the mean intersection percentage between top-k genes of different latent variables decreases for all latent variables (Supplementary Fig. S6 and Supplementary Note 2). To rewire the original interaction graph and capture direct gene interactions, we use PPR to quantify the probability of connectivity between genes. This approach models the likelihood that a random walk starting from one gene will visit another, thereby capturing both direct and indirect relationships in the network.

For each gene i, we initialize a personalized vector v_i where the ith entry is set to 1 and all other entries are 0. This biases the random walk to originate from gene i. We then compute the PPR vector π_i by recursively solving the equation:

$$\pi_i = \alpha v_i + (1-\alpha) A\pi_i$$

where A is the column-normalized adjacency matrix of the initial network and α is the teleportation probability (typically set to 0.85).

The resulting i vector contains PageRank scores for all genes relative to gene i. We interpret these scores as probabilities of connectivity, with higher values indicating stronger relationships. To construct the refined network, we select the top "E" edges for each gene based on these probabilities. In our implementation, we set E = 600 to balance network sparsity and connectivity (Supplementary Note 2). This process is repeated for all genes, resulting in a directed, weighted graph that emphasizes the most significant regulatory relationships in the network. Also, the PPR-based rewiring captures higher-order network topology beyond direct interactions, which can reveal important indirect connections. Unlike previous methods that focus on direct transcription factor-downstream gene interactions, CLERA uses PPR to capture higher-order network topology beyond direct connections. Most of the traditional GRN inference approaches produce static representations [40–42], while CLERA's PPR-based network rewiring identifies relationships, offering dynamic interaction networks that highlight central genes and active gene programs. However, a direct comparison between CLERA and traditional methods is not possible due to their fundamentally different approaches to network construction and the types of regulatory relationships they aim to capture.

Calculating various centrality metrics on the obtained connectivity graph helps us identify the most influential genes, determine hubs of activity, and uncover critical pathways. Each centrality metric offers a unique perspective on the role and significance of nodes in the network. The centrality metrics used include:

- **Degree centrality**: Degree centrality measures the number of direct connections a node has in a network. Nodes with a high degree of centrality are typically more active or prominent within the network.
- **Closeness centrality**: Closeness centrality assesses the average length of the shortest paths from a node to all other nodes in the network. Lower scores indicate that the node occupies a more central and important position in the network.
- **Eigenvector centrality**: Eigenvector centrality evaluates a node's influence based on both the quality and quantity of its connections. A node is more central if it is connected to other central nodes, highlighting its importance within the network.
- **PageRank**: PageRank assigns importance to a node based on the principle that connections from high-ranking nodes contribute more to a node's score, making it a more influential node in the network

By analysing these centrality metrics, we identify the key genes in the interaction network, gene markers and hubs that can play a key role in cell function and gene interaction.

Simulated data

To evaluate SINDy's ability to detect governing equations, we design a simple two-gene regulatory network. This system is governed by stochastic differential equations, incorporating both deterministic dynamics and stochastic fluctuations. The network features two mutually inhibiting genes, described by G_1 and G_2:

$$dG_1 = \left(\frac{m_1}{\left(1+{G_2}^2\right)} - k_D G_1\right) dt + \sigma_1 \sqrt{dt}\, dW_1$$

$$dG_2 = \left(\frac{m_2}{\left(1+{G_1}^2\right)} - k_D G_2\right) dt + \sigma_2 \sqrt{dt}\, dW_2$$

In this model k_D represents the degradation rate, m_1 and m_2 determine the synthesis rates of the genes, σ_1 and σ_2 are noise intensities, dt is the time step, dw_1 and dw_2 are increments of independent Wiener processes modelling Gaussian noise. By adjusting these parameters, the system can exhibit saddle-node and pitchfork bifurcations. To generate training data for SINDy, we set $m_1 = 3$, $m_2 = 1$, $k_D = 1$, and $\sigma_1 = \sigma_2 = 0.2$. The Euler–Maruyama method was used to simulate these stochastic differential equations, to capture the system's dynamics while accounting for noise that scales with time steps.

Result

Discovery of dynamical systems and gene programs from simulated data

We first investigate the performance of the SINDy part of CLERA in discovering the underlying governing equations of a simple simulated biological system with two driver genes (Fig. 1A). The dynamics of this system are described by a well-established set of differential equations commonly used in various biological contexts such as the lac operon, metabolic signalling pathways, and the cell cycle [43]. Synthetic data is generated using this system of equations with varying noise levels. We then apply SINDy, to recover the equations. Notably, the governing equations are discovered with high accuracy. Figure 1A shows this successful reconstruction, with the recovered parameters closely mirroring the original values (see the 'Materials and methods' section). Also, the results from solving the discovered differential equation closely match the generated data, further validating the accuracy of the equations discovered (Fig. 1B).

Then to evaluate the performance of CLERA in a more realistic scenario with a larger dataset, we apply it to simulated data generated by SERGIO, which incorporates various types of noise for realistic data generation [44]. SERGIO (Single-cell ExpRession of Genes In silicO) is a simulator that generates realistic scRNA-seq data by incorporating gene regulatory networks and stochastic transcriptional

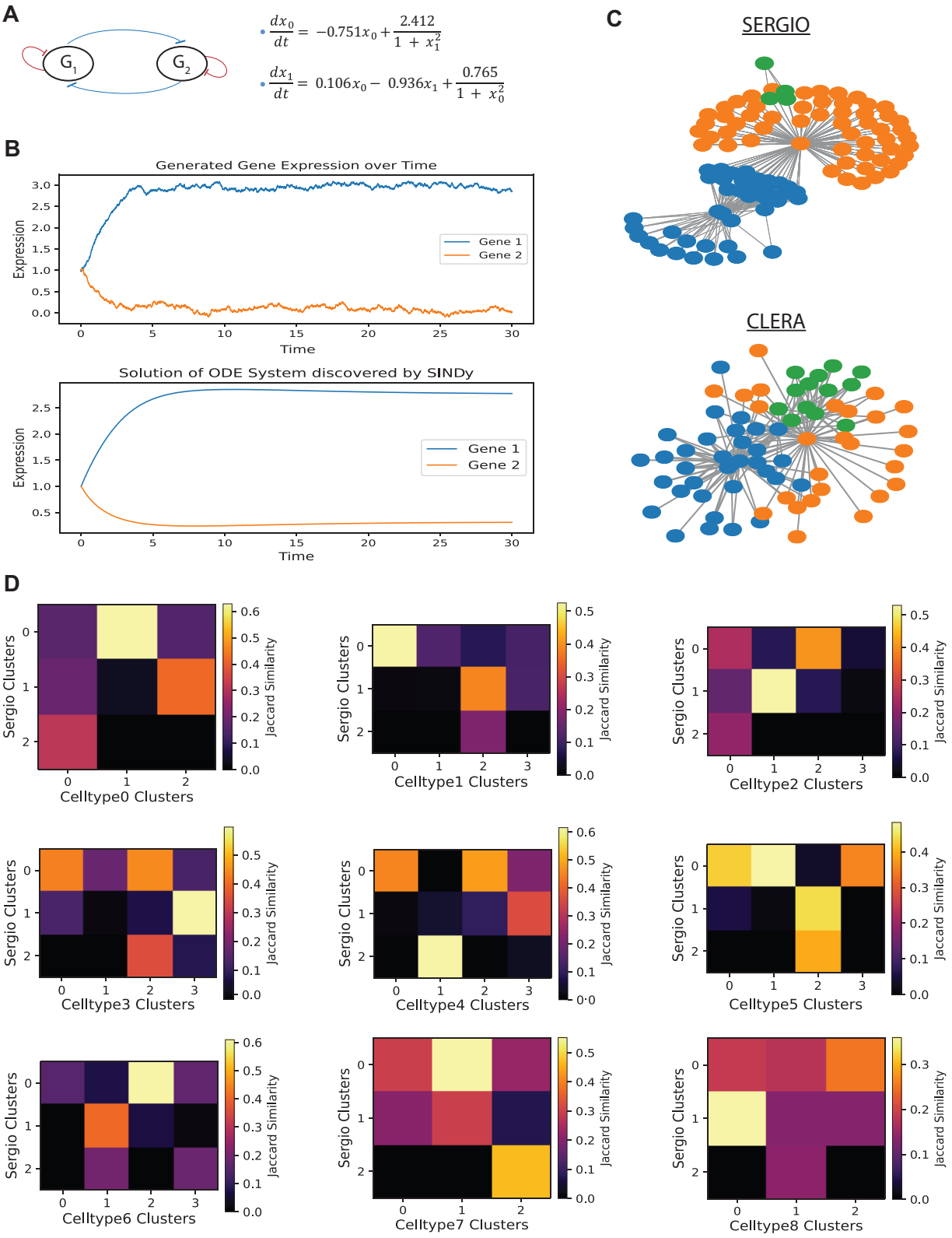

Figure 1. CLERA discovers dynamical systems and gene programs from simulated data. (**A**) Schematic of a two-gene regulatory network (G_1 and G_2) with discovered governing equations and parameters shown. (**B**) Comparison of generated gene expression data (top) and solutions from equations discovered by SINDy (bottom) for the two-gene system over time. Gene 1 and Gene 2 expression levels are plotted against time. (**C**) Gene interaction networks for cell type 7 derived from SERGIO ground truth (top) and CLERA (bottom). Nodes represent genes, coloured by gene programs identified through clustering. (**D**) Heatmaps showing Jaccard similarity between SERGIO and CLERA-derived gene program clusters across nine cell types (CellType0 to CellType8). Colour intensity indicates the degree of similarity, with lighter colours representing higher similarity and darker colours lower similarity.

processes. It models the interaction between transcription factors and their target genes to generate synthetic datasets that accurately replicate the statistical properties and biological noise of experimental data, providing a robust and flexible framework for benchmarking computational methods in single-cell transcriptomics. We train CLERA on simulated data, with 6300 cells across 100 genes and nine distinct cell types. We hypothesize that an optimal representation learned by the model should not only achieve high accuracy in data reconstruction (autoencoder loss) but also should perform well in tasks such as cell type classification and sparse dynamical model discovery.

To address stochasticity and ensure robustness in finding the optimal latent embedding, we run CLERA multiple times (50 for this data) using various initial conditions. We select the model with the lowest combined loss (see the 'Materials and methods' section) while also prioritizing parsimony in the discovered model. In our analysis, we observe that CLERA successfully identified a latent embedding with high accuracy in both reconstruction and cell type classification (Supplementary Figs S1 and S5A).

We then leverage the representation learned by CLERA to identify active gene programs and their dynamics over time. To uncover the connection between latent nodes and genes we compute SHAP [45] values between each node in the autoencoder's latent layer and genes and rank the identified genes based on their SHAP values (choosing the top 30 genes for each node). Using the results from the SHAP method and the discovered differential equations, we construct a network of interactions between latent nodes and genes. We then apply Personalized PageRank (PPR) to this network, starting from each gene, to identify the most relevant genes for the selected gene [46]. This approach enables us to refine the network by selecting only the top connected genes with the highest PPR scores while filtering out the latent nodes. A clustering algorithm is applied to this graph to detect the gene programs. Given that CLERA can uniquely incorporate a time component, this process can be done for different stages of the trajectory and cell types. To assess CLERA's performance in capturing active gene programs, we perform the same clustering analysis on the SERGIO ground truth network and compare the resulting clusters obtained (Fig. 1C for celltype7). We also observed a high degree of similarity between the gene programs of the SERGIO predefined network and the identified gene interaction networks for each cell type, as measured by the Jaccard similarity. This suggests that the latent embedding learned by CLERA can effectively capture active gene programs (Fig. 1D).

CLERA uncovers dynamics and gene programs in pancreatic development

We further evaluate CLERA on biological scRNA-seq data from mouse pancreas during embryonic development. This dataset comprises 3696 cells clustered into eight distinct cell types [47]. Following hyperparameter optimization and preprocessing, we trained CLERA several times with varying initializations (see the 'Materials and methods' section), using the gene expression and computed pseudotime (cell ordering) information. CLERA successfully identifies a set of sparse and interpretable differential equations with all individual loss terms in our total loss function decreasing (Fig 2A and Supplementary Fig. S2). Also, we observe the temporal dynamic of different latent variables captures distinct patterns for each cell type (Fig. 2B and Supplementary Fig. S9A). CLERA also achieves a high classification accuracy using the latent variables where certain latent variables emerge as dominant predictors for individual cell types (Supplementary Figs S5B and S7).

To explore the connection between latent nodes and genes, we calculate the SHAP values [45] for each gene–latent node pair and identify the top "K" genes (K = 300 for this data) connected to each latent node, ranked by their absolute values. Using the discovered equations (latent node–latent node interaction) and SHAP values (gene–latent node interaction), we generate a series of interaction graphs for various stages of pancreas development (Fig. 2C). Unlike traditional network methods, which only produce a static graph for the whole process, our approach captures dynamic graphs over time. Next, we apply a clustering algorithm to the interaction graphs to identify groups of interconnected and potentially co-regulated genes. These graphs, representing different stages of pancreatic development, allowed us to observe changes in gene interactions over time. To understand the similarity of active gene programs across different cell types, we analyse the clustering results for a specific cell type and transfer the identified gene colours to the analysis of other cell types (Fig. 2D). We observe a high degree of similarity in shared genes for cluster 1 among Ductal, Ngn3 low EP, and Ngn3 high EP cell types. Analysing these shared genes reveals several previously known key genes, such as Sox9, Neurog3, Hes1, Foxa3, and Nfib, as well as important signalling pathways like Wnt, Notch, and TGF-β [48]. Furthermore, Gene Set Enrichment Analysis reveals several pathways related to pancreatic development, demonstrating the biological relevance of the gene programs discovered by CLERA (Fig. 2E). Interestingly, CLERA also captures pathways involved in neurogenesis and neural development, which aligns with previous studies and highlights the molecular and cellular similarities between pancreatic and neural cell differentiation [47, 49].

To identify key and central genes for each cell type, we restructure the interaction network using the PPR technique to remove latent nodes. This network rewiring allows us to focus directly on gene interactions. We then apply centrality measures to the restructured network to identify the most influential genes for each cell type (Fig. 2F and Supplementary Fig. S10A). As a result of the centrality analysis, several key genes are identified, including Spp1 [50], a regulator of the epithelial-mesenchymal transitory axis and duct cell dedifferentiation; Chgb [51], a neuroendocrine cell marker; and Neurog3 [52, 53], crucial for endocrine cell differentiation. The analysis also confirms the central roles of Ins1 and Ins2 in beta cell function, along with Clu [54] and Sox9 [55], both critical for progenitor cell maintenance and differentiation. CLERA correctly captures these key genes, aligning with prior studies that emphasize their importance in pancreatic development.

CLERA reveals central genes and dynamics in hematopoietic differentiation

Next, we investigate bone marrow development, examining the differentiation of hematopoietic stem cells (HSPCs) into erythroids, monocytes, and dendritic cells (DCs). This dataset

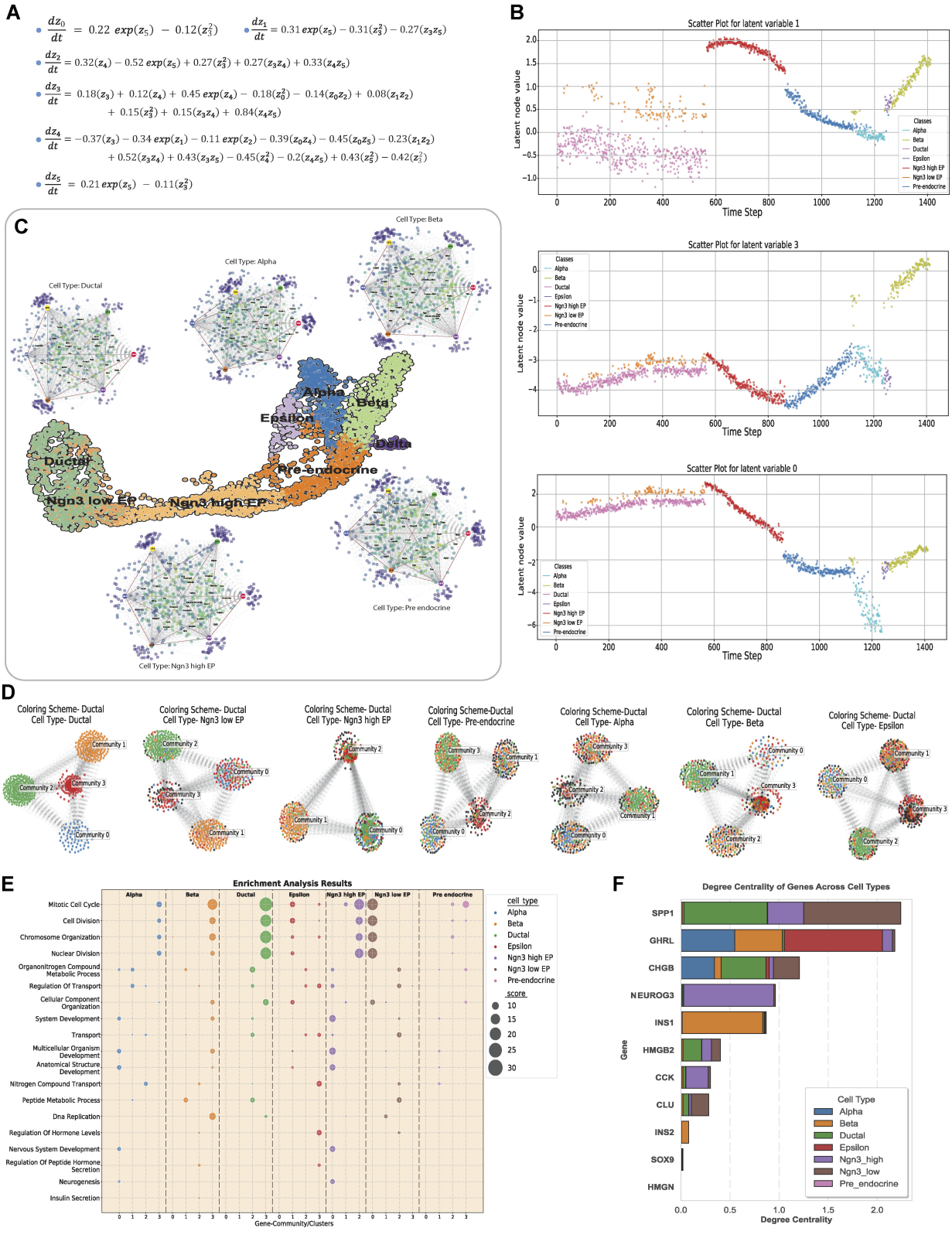

Figure 2. CLERA uncovers dynamics and gene programs in pancreatic development. (**A**) Discovered differential equations governing mouse pancreas development data from scRNA-seq, showing sparse and interpretable models and connections between latent variables. (**B**) Temporal dynamics of latent variables, which illustrate distinct patterns across cell types. (**C**) Interaction graphs for various stages of pancreatic development, show dynamic gene interactions over time for different cell types. (**D**) Clustering results showing gene program similarities across cell types, with shared genes in cluster 1 among Ductal, Ngn3 low EP, and Ngn3 high EP cell types. (**E**) Gene Set Enrichment Analysis of identified genes indicating pathways related to pancreatic development and neurogenesis. (**F**) Degree centrality analysis identifying key genes for each cell type, including Spp1, Chgb, Neurog3, Ins1, Ins2, Clu, and Sox9.

comprises 5780 cells and 14 319 genes clustered into 10 distinct cell types [32].

To enhance CLERA's performance on this data, we apply transfer learning from our previous pancreas study (see the 'Materials and methods' section). By initializing CLERA with pre-trained weights, we leverage the knowledge and relationships obtained from the previous part, which results in a faster optimization and more accurate representation of the data. Also, we observe a decrease across all components of the loss function, showing that all loss terms were effectively optimized and that the discovered equations are also parsimonious (Fig. 3A and Supplementary Fig. S3). Furthermore, investigating the latent space shows that the temporal dynamics of different latent variables capture distinct patterns for each cell type, which shows that the embedding learnt by CLERA can identify and characterize unique behavioural signatures for each cell type (Fig. 3B and Supplementary Fig. S9B). We observe that the latent embedding discovered by CLERA achieves high classification accuracy and also shows distinctive cell type-level differentiation, where specific latent variables drive the classification of particular cell types (Supplementary Figs S5C and S8).

Then, by identifying top genes using the SHAP method and leveraging the discovered equations, we generate a series of graphs representing different stages of bone marrow development (Fig. 3C). Through clustering analysis on these graphs, we identify groups of co-regulated genes at each developmental stage (Fig. 3D). Moreover, using label transfer techniques, we identify a significant similarity in co-regulated genes between precursors (cluster 0), monocytes (cluster 3 in Mono_1 and cluster 2 in Mono_2) and DCs (cluster 0). Some of the key genes discovered have been shown to be crucial for monocyte development, including ID2, TYROBP, FLT3, PDE4B, and GLIPR1 [56–58]. We also observe a strong similarity between the two erythroid subpopulations Ery_1 and Ery_2, particularly between clusters 1 and 3, and between clusters 2 and 1. Similarly, the monocyte subpopulations Mono_1 and Mono_2 show considerable overlap, with cluster 3 in Mono_1 closely aligning with cluster 1 in Mono_2, cluster 2 in Mono_1 resembling cluster 0 in Mono_2, and cluster 0 in Mono_1 closely matching cluster 2 in Mono_2. This suggests that these subpopulations have many common genes, which shows similarities in their developmental pathways and active gene programs.

To identify the critical genes within these networks, we apply PPR for network rewiring, which allows us to remove latent nodes and focus on direct gene interactions. Centrality measures then pinpoint key genes driving cellular differentiation during hematopoiesis (Fig. 3E and Supplementary Fig. S10B). MPO [59], crucial for neutrophil differentiation, is identified as a key myeloid marker, while HOPX [60] emerges as a regulator of primitive hematopoiesis, guiding early progenitor cell fate. Malat1 [61], known for regulating gene expression in HSPCs, and FOS [62, 63], linked to cell proliferation and differentiation under cytokine signalling, are also highlighted. Also, CD52 [64], a marker of mature lymphocytes, FAm30A, which has shown links to immune response regulation and other hematopoietic lineages, and CD74 [65], essential for antigen presentation in immune cells, are captured. CLERA effectively identified these genes, which align with their known roles in hematopoiesis.

Discussion

Here we introduce a novel method, CLERA which represents a significant advancement in our ability to uncover the underlying principles governing complex biological systems. CLERA integrates data-driven model discovery and representation learning, to provide a robust way for uncovering interpretable and parsimonious dynamical models from high-dimensional scRNA-seq data. The ability to simultaneously learn coordinates and governing equations is perhaps the most complementary approach to traditional physics modelling. While classical methods rely on predefined equations based on known physical laws, CLERA also allows the data to guide the discovery of both the relevant coordinates (latent representation) and the equations that govern their dynamics. This approach is particularly valuable in biological systems where the underlying mechanisms are often complex and not fully understood.

Our results demonstrate CLERA's ability to accurately recover governing equations from both simulated and real-world biological data. In the simple simulated biological system, CLERA precisely predicts the governing equations, even under varying noise levels. This performance is further validated with a larger simulated dataset generated by SERGIO, which incorporates realistic noise patterns, showing CLERA's accuracy in modelling complex, high-dimensional data. When applied to real-world datasets from the pancreas and bone marrow development, CLERA leads to biologically relevant insights. The model effectively identifies key genes through network rewiring and centrality measures, as well as detecting active gene programs involved in these processes aligned with previous experimental studies. Another key strength of CLERA lies in its ability to capture the temporal dynamics of gene programs. Unlike traditional network inference methods that produce static graphs, CLERA generates dynamic networks that evolve over time across cell types. This temporal aspect is crucial for understanding differences in complex regulatory mechanisms of cellular processes such as differentiation and development. The identification of potentially novel key genes and gene programs, along with their temporal progression, can present new opportunities for further experimental exploration of dynamic cellular processes.

CLERA integrates multiple components in its loss function to address the challenges of unsupervised learning of identifiable nonlinear representations. The combination of reconstruction loss, classification loss, and SINDy losses introduces inductive biases that guide the model toward learning more meaningful and interpretable embeddings. This comprehensive loss function imposes constraints on latent dynamics, to ensure consistency with known biological principles while maintaining the flexibility to discover new relationships from the data. By balancing these different loss components, CLERA achieves a robust and parsimonious framework for uncovering interpretable dynamical models. Also, the use of transfer learning from the pancreas to the bone marrow dataset demonstrates the model's adaptability and potential for generalization across different biological contexts. This approach not only improves performance but also suggests that certain learned representations may be conserved across different developmental processes.

In CLERA, we intentionally avoid assuming independence between latent variables, diverging from some recent methods

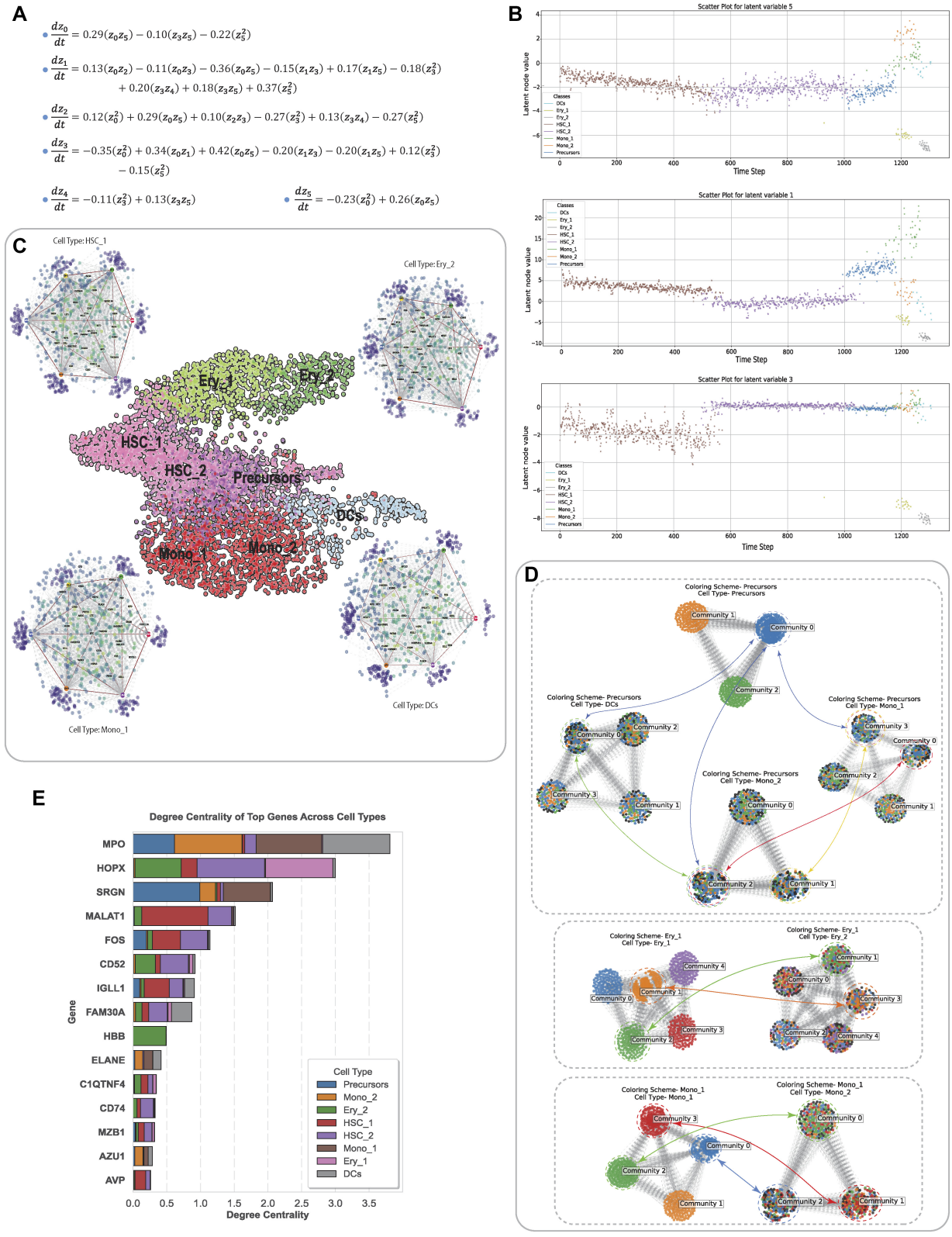

Figure 3. CLERA reveals central genes and dynamics in hematopoietic differentiation. (**A**) Differential equations discovered for bone marrow development data, showing connections between latent variables. (**B**) Temporal dynamics of latent variables, with distinct patterns across cell types. (**C**) Interaction graphs for different stages of bone marrow development, which capture dynamic gene interactions for each cell type. (**D**) Clustering analysis of co-regulated gene groups at each developmental stage, with significant similarities between precursors, monocytes, DCs, and among Ery-1 and Ery-2, Mono-1, and Mono-2 subpopulations. (**E**) Degree centrality analysis identifies key genes driving cellular differentiation, including MPO, HOPX, Malat1, FOS, CD52, FAM30A, and CD74.

[66]. In biological systems, the strict assumption of independence can be problematic, as many genes participate in multiple programs simultaneously. This interconnectedness makes it challenging to capture the true nature of these systems with models that enforce strict independence. By allowing for overlapping gene programs and complex interactions, CLERA's approach is better suited to reflect the complexity and reality we have in biology. We believe by not imposing independence between latent variables, CLERA can more accurately model the temporal dynamics and complex interactions captured by ODEs, leading to a better understanding of the underlying biological processes.

A critical aspect of using SINDy for ordinary differential equation (ODE) discovery in biological systems is the question of whether there is a single unique way to represent (model) a given process or if multiple valid representations exist. While CLERA effectively recovers governing equations from data, it is important to recognize that the method may yield different but equally valid solutions due to its consideration of linear transformations and rotations of the underlying dynamical systems. This inherent flexibility reflects the possibility that processes might not have a singular unique representation, but rather several that are mathematically equivalent yet distinct in form. This raises an important challenge in interpreting the biological significance of the discovered ODEs, as it suggests that multiple models could potentially describe the same underlying process. Especially for systems with limited prior knowledge, further investigation is needed to determine whether this flexibility captures biological variability or whether it points to the need for additional constraints to narrow down the solution space to the most biologically relevant model.

Despite its strengths, CLERA has some limitations that should be addressed in future work. First, the model's performance is sensitive to hyperparameter choices, particularly the dimension of the latent space and library of candidate functions. Developing more robust methods for automatic hyperparameter tuning could improve the model's consistency. Second, while CLERA can handle noise in the data to some extent, high noise levels can still impact its performance. Further improvements in denoising and data preprocessing techniques could enhance the model's robustness. Third, the interpretation of the learned latent space and governing equations requires careful consideration. While we have shown that the model captures biologically relevant information, additional work is needed to develop standardized methods for interpreting and benchmarking these results in the context of specific biological questions. Lastly, CLERA's performance may vary depending on the biological process being studied and the effectiveness of methods for pseudotime analysis or tools such as MIOFlow [67], which are designed to capture continuous population dynamics. Consequently, it is essential to ensure that the latent variables are time-differentiable to accurately model these processes.

In conclusion, CLERA represents a significant step forward in our ability to uncover the governing principles of complex biological systems from high-dimensional scRNA-seq data in a parsimonious and generalizable manner. By bridging the gap between data-driven discovery and mechanistic understanding, this approach has the potential to deepen our understanding of cellular processes and the interactions between active gene programs.

Acknowledgements

Author contributions: M.S. conceived the project. V.S and M.S. implemented the method and performed the analyses. V.S. and M.S. interpreted the results. M.S. and V.S. drafted the first manuscript. M.S. supervised the study. All authors read and approved the final manuscript.

Supplementary data

Supplementary data are available at NAR Genomics & Bioinformatics Online.

Conflict of interest

None declared.

Funding

This work was not supported by a dedicated financial support.

Data availability

The datasets used in this study are publicly available in public repositories. Simulated scRNA data was generated using SERGIO, which is accessible at https://github.com/PayamDiba/SERGIO. Preprocessed versions of both the pancreas development data (accession number GSE132188) and the human hematopoiesis data (accessed through the Human Cell Atlas data portal under the Human Hematopoietic Profiling project) can be downloaded from https://scvelo.readthedocs.io/en/stable/.

Code availability

Our Python implementation of CLERA can be found at: https://github.com/vasu-swaroop/CLERA and https://doi.org/10.5281/zenodo.15185597.

References

1. Champion K, Lusch B, Kutz JN *et al.* Data-driven discovery of coordinates and governing equations. *PLoS Biol* 2019;**116**:S131–2. https://doi.org/10.1073/pnas.1906995116
2. Sadria M, Seo D, Layton AT. The mixed blessing of AMPK signaling in cancer treatments. *BMC Cancer* 2022;**22**:105. https://doi.org/10.1186/s12885-022-09211-1
3. Layton AT, Sadria M. Understanding the dynamics of SARS-CoV-2 variants of concern in Ontario, Canada: a modeling study. *Sci Rep* 2022;**12**:2114. https://doi.org/10.1038/s41598-022-06159-x
4. Hirsa A, Neftci SN. *An Introduction to the Mathematics of Financial Derivatives*. books.google.com (18 December 2013, date last accessed).
5. Camps-Valls G, Gerhardus A, Ninad U *et al.* Discovering causal relations and equations from data. *Phys Rep* 2023;**1044**:1–68. https://doi.org/10.1016/j.physrep.2023.10.005
6. Lejarza F, Baldea M. Data-driven discovery of the governing equations of dynamical systems via moving horizon optimization. *Sci Rep* 2022;**12**:11836. https://doi.org/10.1038/s41598-022-13644-w
7. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc Natl Acad Sci USA* 2016;**113**:3932–7. https://doi.org/10.1073/pnas.1517384113

8. Li P-H, Kong X-Y, He Y-Z *et al*. Recent developments in application of single-cell RNA sequencing in the tumour immune microenvironment and cancer therapy. *Mil Med Res* 2022;**9**:52.
9. Sadria M, Layton A. scVAEDer: integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis. *Genome Biol* 2025;**26**:64. https://doi.org/10.1186/s13059-025-03519-4
10. Sadria M, Layton A, Bader GD. Adversarial training improves model interpretability in single-cell RNA-seq analysis. *Bioinform Adv* 2023;**3**:vbad166. https://doi.org/10.1093/bioadv/vbad166
11. Saliba A-E, Westermann AJ, Gorski SA *et al*. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**:8845–60. https://doi.org/10.1093/nar/gku555
12. Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. *Nat Rev Genet* 2019;**20**:536–48. https://doi.org/10.1038/s41576-019-0130-6
13. Qiu X, Zhang Y, Martin-Rufino JD *et al*. Mapping transcriptomic vector fields of single cells. *Cell* 2022;**185**:690–711. https://doi.org/10.1016/j.cell.2021.12.045
14. Chen Z, King WC, Hwang A *et al*. *DeepVelo*: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Sci Adv* 2022;**8**:eabq3745. https://doi.org/10.1126/sciadv.abq3745
15. Yeo GHT, Saksena SD, Gifford DK. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat Commun* 2021;**12**:3222. https://doi.org/10.1038/s41467-021-23518-w
16. Sha Y, Qiu Y, Zhou P *et al*. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nat Mach Intell* 2024;**6**:25–39. https://doi.org/10.1038/s42256-023-00763-w
17. Wang W, Ni K, Poe D *et al*. Transiently increased coordination in gene regulation during cell phenotypic transitions. *PRX Life* 2024;**2**:43009. https://doi.org/10.1103/PRXLife.2.043009
18. Sadria M, Zhang A, Bader GD. Deep Lineage: Single-Cell Lineage Tracing and Fate Inference Using Deep Learning. bioRxiv, https://doi.org/10.1101/2024.04.25.591126, 26 April 2024, preprint: not peer reviewed.
19. Wen Y, Huang J, Guo S *et al*. Applying causal discovery to single-cell analyses using CausalCell. *Elife* 2023;**12**:e81464. https://doi.org/10.7554/eLife.81464
20. Kutz JN, Brunton SL. Parsimony as the ultimate regularizer for physics-informed machine learning. *Nonlinear Dyn* 2022;**107**:1801–17. https://doi.org/10.1007/s11071-021-07118-3
21. Schölkopf B. Causality for machine learning. In: Geffner H, Dechter R, Halpern JY (eds.), *Probabilistic and Causal Inference: The Works of Judea Pearl*. New York, NY, USA: ACM; 2022, 765–804. https://doi.org/10.1145/3501714.3501755
22. Hyvärinen A, Khemakhem I, Morioka H. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns (NY)* 2023;**4**:100844. https://doi.org/10.1016/j.patter.2023.100844
23. Hyvarinen A, Morioka H. Nonlinear ICA of temporally dependent stationary sources. In: *Artificial intelligence and statistics*. PMLR, 2017, **54**, 460–9.
24. Hyvarinen A, Sasaki H, Turner R. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, 859–68.
25. Khemakhem I, Kingma D, Monti R *et al*. Variational autoencoders and nonlinear ica: A unifying framework. In: *International conference on artificial intelligence and statistics*. PMLR, 2020, 2207–17.
26. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**:715–21. https://doi.org/10.1038/s41592-019-0494-8
27. Sadria M, Layton A, Goyal S *et al*. Fatecode enables cell fate regulator prediction using classification-supervised autoencoder perturbation. *Cell Rep Methods* 2024;**4**:100819. https://doi.org/10.1016/j.crmeth.2024.100819
28. Eraslan G, Simon LM, Mircea M *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390. https://doi.org/10.1038/s41467-018-07931-2
29. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;**323**:533–6. https://doi.org/10.1038/323533a0
30. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246. https://doi.org/10.1038/s41467-022-28803-w
31. Domínguez Conde C, Xu C, Jarvis LB *et al*. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**:eabl5197. https://doi.org/10.1126/science.abl5197
32. Setty M, Kiseliovas V, Levine J *et al*. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* 2019;**37**:451–60. https://doi.org/10.1038/s41587-019-0068-4
33. Ma L, Zheng J. A polynomial based model for cell fate prediction in human diseases. *BMC Syst Biol* 2017;**11**:126. https://doi.org/10.1186/s12918-017-0502-5
34. Sadria M, Layton AT. Use of angiotensin-converting enzyme inhibitors and Angiotensin II receptor blockers during the COVID-19 pandemic: a modeling analysis. *PLoS Comput Biol* 2020;**16**:e1008235. https://doi.org/10.1371/journal.pcbi.1008235
35. Sadria M, Layton AT. Aging affects circadian clock and metabolism and modulates timing of medication. *iScience* 2021;**24**:102245. https://doi.org/10.1016/j.isci.2021.102245
36. Sadria M, Layton AT. Interactions among mTORC, AMPK and SIRT: a computational model for cell energy balance and metabolism. *Cell Commun Signal* 2021;**19**:57. https://doi.org/10.1186/s12964-021-00706-1
37. Chow S-M. Practical tools and guidelines for exploring and fitting linear and nonlinear dynamical systems models. *Multivariate Behav Res* 2019;**54**:690–718. https://doi.org/10.1080/00273171.2019.1566050
38. Zhuang F, Qi Z, Duan K *et al*. A comprehensive survey on transfer learning. *Proc IEEE* 2020;**109**:43–76. https://doi.org/10.1109/JPROC.2020.3004555
39. Iman M, Arabnia HR, Rasheed K. A review of deep transfer learning and recent advancements. *Technologies* 2023;**11**:40. https://doi.org/10.3390/technologies11020040
40. Yuan Q, Duren Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nat Biotechnol* 2025;**43**:247–57. https://doi.org/10.1038/s41587-024-02182-7
41. Matsumoto H, Kiryu H, Furusawa C *et al*. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* 2017;**33**:2314–21. https://doi.org/10.1093/bioinformatics/btx194
42. Huynh-Thu VA, Irrthum A, Wehenkel L *et al*. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;**5**:e12776. https://doi.org/10.1371/journal.pone.0012776
43. Freedman SL, Xu B, Goyal S *et al*. A dynamical systems treatment of transcriptomic trajectories in hematopoiesis. *Development* 2023;**150**:dev201280. https://doi.org/10.1242/dev.201280
44. Dibaeinia P, Sinha S. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell Syst* 2020;**11**:252–71. https://doi.org/10.1016/j.cels.2020.08.003
45. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., 2017, 4768–77.
46. Jeh G, Widom J. Scaling personalized web search. *Proceedings of the Twelfth International Conference on World Wide Web - WWW '03*. New York, New York, USA: ACM Press; 2003, 271.
47. Bastidas-Ponce A, Tritschler S, Dony L *et al*. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* 2019;**146**:dev173849. https://doi.org/10.1242/dev.173849

48. Lee J-H, Lee J-H, Rane SG. TGF-β signaling in pancreatic islet β cell development and function. *Endocrinology* 2021;**162**:bqaa233. https://doi.org/10.1210/endocr/bqaa233
49. Szlachcic WJ, Ziojla N, Kizewska DK *et al.* Endocrine pancreas development and dysfunction through the lens of single-cell RNA-sequencing. *Front Cell Dev Biol* 2021;**9**:629212. https://doi.org/10.3389/fcell.2021.629212
50. Kilic G, Wang J, Sosa-Pineda B. Osteopontin is a novel marker of pancreatic ductal tissues and of undifferentiated pancreatic precursors in mice. *Dev Dyn* 2006;**235**:1659–67. https://doi.org/10.1002/dvdy.20729
51. Dugnani E, Sordi V, Pellegrini S *et al.* Gene expression analysis of embryonic pancreas development master regulators and terminal cell fate markers in resected pancreatic cancer: a correlation with clinical outcome. *Pancreatology* 2018;**18**:945–53. https://doi.org/10.1016/j.pan.2018.09.006
52. Zhu Y, Liu Q, Zhou Z *et al.* PDX1, Neurogenin-3, and MAFA: critical transcription regulators for beta cell development and regeneration. *Stem Cell Res Ther* 2017;**8**:240. https://doi.org/10.1186/s13287-017-0694-z
53. Gradwohl G, Dierich A, LeMeur M *et al.* neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci USA* 2000;**97**:1607–11. https://doi.org/10.1073/pnas.97.4.1607
54. Kim BM, Kim SY, Lee S *et al.* Clusterin induces differentiation of pancreatic duct cells into insulin-secreting cells. *Diabetologia* 2006;**49**:311–20. https://doi.org/10.1007/s00125-005-0106-2
55. Puri S, Maachi H, Nair G *et al.* Sox9 regulates alternative splicing and pancreatic beta cell function. *Nat Commun* 2024;**15**:588. https://doi.org/10.1038/s41467-023-44384-8
56. Ishiguro A, Spirin KS, Shiohara M *et al.* Id2 expression increases with differentiation of human myeloid cells. *Blood* 1996;**87**:5225–31. https://doi.org/10.1182/blood.V87.12.5225.bloodjournal87125225
57. Böiers C, Buza-Vidas N, Jensen CT *et al.* Expression and role of FLT3 in regulation of the earliest stage of normal granulocyte-monocyte progenitor development. *Blood* 2010;**115**:5061–8. https://doi.org/10.1182/blood-2009-12-258756
58. Rochford I, Joshi JC, Rayees S *et al.* Evidence for reprogramming of monocytes into reparative alveolar macrophages in vivo by targeting PDE4b. *Am J Physiol Lung Cell Mol Physiol* 2021;**321**:L686–702. https://doi.org/10.1152/ajplung.00145.2021
59. Scholz W, Platzer B, Schumich A *et al.* Initial human myeloid/dendritic cell progenitors identified by absence of myeloperoxidase protein expression. *Exp Hematol* 2004;**32**:270–6. https://doi.org/10.1016/j.exphem.2003.12.007
60. Lin CC, Yao CY, Hsu YC *et al.* Knock-out of Hopx disrupts stemness and quiescence of hematopoietic stem cells in mice. *Oncogene* 2020;**39**:5112–23. https://doi.org/10.1038/s41388-020-1340-2
61. Cremer S, Michalik KM, Fischer A *et al.* Hematopoietic deficiency of the long noncoding RNA MALAT1 promotes atherosclerosis and plaque inflammation. *Circulation* 2019;**139**:1320–34. https://doi.org/10.1161/CIRCULATIONAHA.117.029015
62. Wang ZQ, Ovitt C, Grigoriadis AE *et al.* Bone and haematopoietic defects in mice lacking c-fos. *Nature* 1992;**360**:741–5. https://doi.org/10.1038/360741a0
63. Konturek-Ciesla A, Olofzon R, Kharazi S *et al.* Implications of stress-induced gene expression for hematopoietic stem cell aging studies. *Nat Aging* 2024;**4**:177–84. https://doi.org/10.1038/s43587-023-00558-z
64. Morisot S, Georgantas RW, Civin CI. 345. Hematopoietic stem-progenitor cells express CD52 mRNA and membrane protein. *Mol Ther* 2006;**13**:S131–2. https://doi.org/10.1016/j.ymthe.2006.08.403
65. Becker-Herman S, Rozenberg M, Hillel-Karniel C *et al.* CD74 is a regulator of hematopoietic stem cell maintenance. *PLoS Biol* 2021;**19**:e3001121. https://doi.org/10.1371/journal.pbio.3001121
66. Yang M, Liu F, Chen Z *et al.* Causalvae: disentangled representation learning via neural structural causal models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, 9593–602.
67. Huguet G, Magruder DS, Tong A *et al.* Manifold interpolating optimal-transport flows for trajectory inference. *Adv Neural Inf Process Syst* 2022;**35**:29705–18.

Received: October 8, 2024. **Revised:** March 19, 2025. **Editorial Decision:** April 2, 2025. **Accepted:** April 11, 2025