scMultiNODE : Integrative Model for Multi-Modal Temporal Single-Cell Data

Jiaqi Zhang[1], Manav Chakravarthy[1], and Ritambhara Singh[1,2,3]

[1] Department of Computer Science, Brown University
[2] Center for Computational Molecular Biology, Brown University
[3] Corresponding author
ritambhara@brown.edu

Abstract. Measuring single-cell genomic profiles at different timepoints enables our understanding of cell development. This understanding is more comprehensive when we perform an integrative analysis of multiple measurements (or modalities) across various developmental stages. However, obtaining such measurements from the same set of single cells is resource-intensive, restricting our ability to study them integratively. We propose an unsupervised integration model, scMultiNODE, that integrates gene expression and chromatin accessibility measurements in developing single cells while preserving cell type variations and cellular dynamics. scMultiNODE uses autoencoders to learn nonlinear low-dimensional cell representation and optimal transport to align cells across different measurements. Next, it utilizes neural ordinary differential equations to explicitly model cell development with a regularization term to learn a dynamic latent space. Our experiments on four real-world developmental single-cell datasets show that scMultiNODE can integrate temporally profiled multi-modal single-cell measurements better than existing methods that focus on cell type variations and tend to ignore cellular dynamics. We also show that scMultiNODE's joint latent space helps with the downstream analysis of single-cell development.

Availability: The data and code are publicly available at https://github.com/rsinghlab/scMultiNODE.

Keywords: single-cell development · multi-modal data integration · optimal transport · auto-encoders · neural ordinary differential equations · temporal single-cell analysis

1 Introduction

Biological systems are inherently dynamic, constantly changing and adapting over time, and operate at different levels, like organisms, cells, and molecules. Specifically, dynamic processes at the cell level reveal how cells grow, divide, and differentiate into various cell types [43,54]. So, understanding these cellular dynamics allows us to enhance our knowledge of cell development and diseases. Advances in single-cell measurements assist this study by capturing a high-resolution genomic snapshot of cell states. Recent research increasingly involves profiling multiple developmental single-cell measurements (or modalities), uncovering heterogeneity within the same tissue type and cellular dynamics across developmental stages [19,41,30,57]. Therefore, by examining these multiple single-cell modalities across different timepoints, we can comprehensively understand how biological processes evolve within cell populations.

While critical, integrative analysis of temporal multi-modal single-cell datasets is challenging. Cells are destroyed during single-cell sequencing, resulting in the measurement of separate cell populations (referred to here as "unaligned" datasets). New co-assay experimental protocols have been developed to simultaneously sequence different measurements in the same cells (like gene expression and chromatin accessibility) [34,26]. However, it is still costly to perform this joint measurement for developing single cells across multiple timepoints. Consequently, most temporal multi-modal single-cell datasets are unaligned across modalities and timepoints.

Many unsupervised computational methods have been proposed to integrate multi-modal single-cell datasets. These approaches integrate unaligned modalities without using prior cell correspondence information for supervision. For example, using canonical correlation analysis, the single-cell analysis platform Seurat [24] projects different feature spaces, such as genes and chromatin regions, into a common subspace. LIGER [31] uses non-negative matrix factorization to find shared factors for matching multiple single-cell modalities. However, they are based on linear operations and cannot deal with nonlinear associations across modalities [4,1]. Therefore, recently published methods adopt manifold alignment to capture complicated inter-modality relationships. For example, MMD-MA [32] uses maximum mean discrepancy for single-cell integration. Pamona [8], SCOT [18], SCOTv2 [17], and uniPort [7] use optimal transport to align modality-specific representations into a shared one. These methods have shown superior performance in integrating heterogeneous single-cell datasets by focusing on capturing cell type variations. Unfortunately, all the existing integration methods overlook the developmental dynamics of the temporal single-cell multi-modal dataset. That is, cellular dynamics in the developmental data are not naturally defined in their integrated spaces [13]. This is a crucial methodological gap because variations in cell types do not necessarily overlap with the variations of cell states during development [48,15]. Therefore, if integration focuses primarily on capturing cell type variation while neglecting cellular dynamics, the resulting integration for temporal data will be sub-optimal. This can obscure critical insights obtained from single-cell development data. Thus, explicitly incorporating cellular dynamics into multi-modal integration remains an important but unsolved problem.

Several previous works have considered modeling cellular developments for single-cell gene expression. For example, Cicero [39] characterizes cell trajectories during the myoblast differentiation. PRESCIENT [52], MIOFlow [27], and TrajectoryNet [46] apply neural ordinary differential equations (ODEs) [9] to model the cell trajectory from temporal gene expression data. Recently, our scNODE [55] method incorporated the dynamics learned from neural ODE to model cellular developments and predict gene expression. However, these methods are designed to model developmental trajectories for one specific modality and cannot be used for the integrative modeling of multiple single-cell measurements.

We propose single-cell Multi-modal Neural Ordinary Differential Equation (`scMultiNODE`), which integrates gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) profiles at multiple timepoints with optimal transport and explicitly models the cellular dynamics with neural ODE (Fig. 1). `scMultiNODE` is an unsupervised model that does not require prior cell correspondence information between the modalities for integration. First, due to the high dimensionality and sparsity of single-cell data, `scMultiNODE` learns low-dimensional latent representations of each modality with Auto-Encoders (AEs) [3]. Subsequently, it aligns modality-specific latent representations with Gromov-Wasserstein (GW) optimal transport [38], which facilitates the prediction of cell correspondence between the two modalities. This correspondence ensures that cells that exhibit similar biological profiles are aligned together. Finally, `scMultiNODE` constructs a joint latent space with the guidance of the predicted correspondence and explicitly incorporates the cellular dynamics using neural ODE. In this way, the joint latent space is able to retain both cell type variations and cellular dynamics.

We evaluate scMultiNODE on four developmental single-cell datasets, including two co-assay datasets and two unaligned datasets, with scRNA-seq and scATAC-seq assays measured at multiple timepoints from different species and tissues. Our qualitative and quantitative analyses demonstrate that scMultiNODE integrates two modalities well in both co-assay and unaligned datasets. Moreover, scMultiNODE significantly outperforms baseline models in capturing cellular dynamics while still retaining cell type variations. Additionally, we show that scMultiNODE generates an interpretable joint latent space, enabling the construction of cell transition paths for studying cell development. We envision that scMultiNODE will be helpful for integrative analyses of multi-modal temporal single-cell datasets, especially those with unaligned measurements.

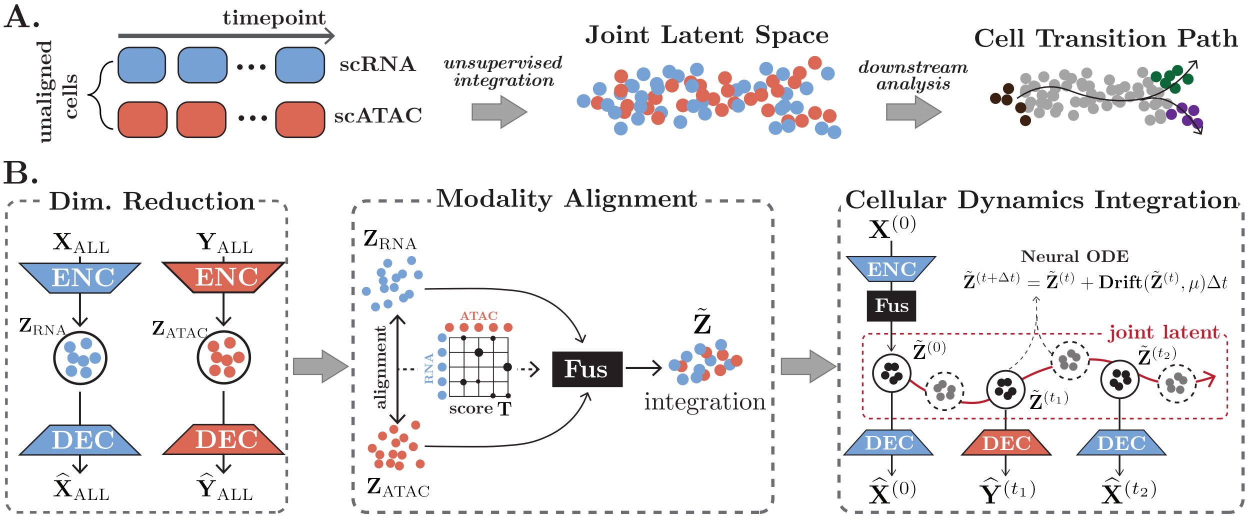
A.
timepoint
unaligned cells
scRNA
scATAC
unsupervised integration
Joint Latent Space
downstream analysis
Cell Transition Path
B.
Dim. Reduction
$\mathbf{X}_{\text{ALL}}$
$\mathbf{Y}_{\text{ALL}}$
ENC
ENC
$\mathbf{Z}_{\text{RNA}}$
$\mathbf{Z}_{\text{ATAC}}$
DEC
DEC
$\hat{\mathbf{X}}_{\text{ALL}}$
$\hat{\mathbf{Y}}_{\text{ALL}}$
Modality Alignment
$\mathbf{Z}_{\text{RNA}}$
alignment
ATAC
RNA
score $\mathbf{T}$
Fus
$\tilde{\mathbf{Z}}$
integration
$\mathbf{Z}_{\text{ATAC}}$
Cellular Dynamics Integration
$\mathbf{X}^{(0)}$
ENC
Fus
Neural ODE
$\tilde{\mathbf{Z}}^{(t+\Delta t)} = \tilde{\mathbf{Z}}^{(t)} + \text{Drift}(\tilde{\mathbf{Z}}^{(t)}, \mu)\Delta t$
joint latent
$\tilde{\mathbf{Z}}^{(0)}$
$\tilde{\mathbf{Z}}^{(t_1)}$
$\tilde{\mathbf{Z}}^{(t_2)}$
DEC
DEC
DEC
$\hat{\mathbf{X}}^{(0)}$
$\hat{\mathbf{Y}}^{(t_1)}$
$\hat{\mathbf{X}}^{(t_2)}$

Fig. 1: **Model overview.** (**A**) Our goal is to integrate scRNA-seq and scATAC-seq data measured at multiple timepoints. We assume there is no cell-to-cell correspondence across modalities and timepoints. With the integration, we can conduct various downstream analyses, such as cell transition path construction. (**B**) scMultiNODE first learns low-dimensional representations of each modality with AEs. Then, it aligns modality-specific latent representations with GW optimal transport, which predicts the cell correspondence between the two modalities. scMultiNODE constructs a joint latent space by mapping modality-specific latent representations to a shared space, with the guidance of the predicted correspondence where aligned cells stay together. scMultiNODE then uses neural ODEs to model cellular dynamics and embed dynamics into the joint latent space.

2 Method

scMultiNODE aligns scRNA-seq (gene expression) and scATAC-seq (chromatin accessibility) datasets measured across time. Therefore, we denote $\mathbf{X}^{(t)} \in \mathbb{R}^{m_t \times p}$ as the gene expression of m_t cells and p genes at timepoint t and $\mathbf{Y}^{(t)} \in \mathbb{R}^{n_t \times q}$ as the chromatin accessibility of n_t cells and q chromatin regions at timepoint t. Vectors are denoted in bold lowercase letters, and matrices are in bold capital letters. Given gene expression and chromatin accessibility matrices at observed timepoints $\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}} \subseteq \{0, 1, 2, \cdots\}$, our goal is finding a low-dimensional latent space that aligns the two modalities, captures cell type variations, and retains cellular dynamics to enable downstream developmental analysis. For scRNA-seq data, we assume the same set of genes is measured and for scATAC-seq data, we assume the same set of chromatin regions is measured at different timepoints. Also, we do not require any prior cell correspondence information, but we assume that there is some underlying shared biology information, such that cells from different measurements and timepoints have shared cell types and reflect a similar cell transition process. This assumption means that the data have a potentially meaningful integration.

scMultiNODE framework includes three major steps (Fig. 1B). (1) It first learns low-dimensional latent representations of each modality with separate Auto-Encoders (AEs). (2) Then, it aligns modality-specific latent representations with GW optimal transport that predicts cell correspondence between the two modalities in an unsupervised manner. scMultiNODE then maps modality-specific latent representations to a joint latent space while preserving local cell relationships and cell type variations. (3) Finally, `scMultiNODE` applies neural ODE to model the cellular dynamics and incorporate the learned dynamic into the joint latent space.

2.1 `scMultiNODE` compresses high dimensional and sparse single-cell data with AEs

Because single-cell measurements are high-dimensional and sparse, `scMultiNODE` first uses Auto-Encoder (AE) to reduce data dimensionality and remove noise. AE is a neural network-based model that maps high-dimensional data to a low-dimensional representation. It is widely used in many single-cell studies and shows a good dimensionality reduction performance [47,33,44]. The benefit of AEs for single-cell data is that they can effectively capture complex cell variations due to the non-linearity property of the neural networks.

`scMultiNODE` uses two separate AEs for the two modalities to perform multi-timepoint modeling with all cells $\mathbf{X}_{\text{ALL}} = \text{CONCAT}(\mathbf{X}^{(t)} | t \in \mathcal{T}_{\text{RNA}})$ and $\mathbf{Y}_{\text{ALL}} = \text{CONCAT}(\mathbf{Y}^{(t)} | t \in \mathcal{T}_{\text{ATAC}})$. AE consists of two neural networks: (1) the encoder network $\text{Enc}(\cdot, \phi)$ maps the input features to a low-dimensional space $\mathbb{R}^d$ ($d \ll \min\{p, q\}$), and (2) the decoder network $\text{Dec}(\cdot, \theta)$ maps latent variables back to the feature space to reconstruct the input. Specifically, given $\mathbf{X}_{\text{ALL}}$ and $\mathbf{Y}_{\text{ALL}}$, `scMultiNODE` learns modality-specific latent representations through

$$\begin{aligned} \mathbf{Z}_{\text{RNA}} = \text{Enc}_{\mathbf{X}}(\mathbf{X}_{\text{ALL}},\ \phi_{\mathbf{X}}), \quad \widehat{\mathbf{X}}_{\text{ALL}} = \text{Dec}_{\mathbf{X}}(\mathbf{Z}_{\text{RNA}},\ \theta_{\mathbf{X}}); \\ \mathbf{Z}_{\text{ATAC}} = \text{Enc}_{\mathbf{Y}}(\mathbf{Y}_{\text{ALL}},\ \phi_{\mathbf{Y}}), \quad \widehat{\mathbf{Y}}_{\text{ALL}} = \text{Dec}_{\mathbf{Y}}(\mathbf{Z}_{\text{ATAC}},\ \theta_{\mathbf{Y}}). \end{aligned} \tag{1}$$

The encoder and decoder networks are parameterized by ϕ and θ, correspondingly. AE minimizes the mean squared error (MSE) between input features and the reconstructions from the decoder as follows:

$$\mathcal{L}_{\text{RNA}} = \text{MSE}\left(\mathbf{X}_{\text{ALL}},\ \widehat{\mathbf{X}}_{\text{ALL}}\right) \qquad \text{and} \qquad \mathcal{L}_{\text{ATAC}} = \text{MSE}\left(\mathbf{Y}_{\text{ALL}},\ \widehat{\mathbf{Y}}_{\text{ALL}}\right). \tag{2}$$

Note that `scMultiNODE` trains these two AEs separately at this step to capture modality-specific cellular variations (like cell types).

2.2 `scMultiNODE` aligns single-cell modalities through GW Optimal Transport

Next, `scMultiNODE` maps latent variables $\mathbf{Z}_{\text{RNA}}$ and $\mathbf{Z}_{\text{ATAC}}$ into a joint latent space, and aligns the two modalities in this latent space. Specifically, `scMultiNODE` adds another neural network $\text{Fus}(\cdot,\ \omega) : \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by ω, named fusion layer, to map a modality-specific cell latent variable vector $\mathbf{z} \in \mathbb{R}^d$ of any modality to the d-dimensional joint latent space through

$$\tilde{\mathbf{z}} = \text{Fus}(\mathbf{z},\ \omega) \quad \text{with} \quad \begin{cases} \mathbf{z} = \text{Enc}_{\text{RNA}}(\mathbf{x},\ \phi_{\mathbf{X}}) \ \text{ and } \ \widehat{\mathbf{x}} = \text{Dec}_{\text{RNA}}(\tilde{\mathbf{z}},\ \theta_{\mathbf{X}}) & \text{for RNA data} \\ \mathbf{z} = \text{Enc}_{\text{ATAC}}(\mathbf{y},\ \phi_{\mathbf{Y}}) \ \text{ and } \ \widehat{\mathbf{y}} = \text{Dec}_{\text{ATAC}}(\tilde{\mathbf{z}},\ \theta_{\mathbf{Y}}) & \text{for ATAC data} \end{cases}. \tag{3}$$

With the fusion layer, `scMultiNODE` integrates latent representations of different feature spaces into the same space. We assume different modalities, though they profile different cells, should have some underlying shared biological information. Therefore, we hypothesize that if cells from different modalities are biologically similar, they should be aligned in the joint latent space and their latent representations should be close to each other. To this end, `scMultiNODE` accomplishes the alignment with the help of GW optimal transport.

GW optimal transport aims at moving data points from one metric space to another while preserving the original local geometry that is captured using intra-domain distances [38]. The central concept of GW is finding the best data correspondence matrix that denotes the probability of alignment between each data point across metric spaces. In our cases, it is finding the cell correspondence matrix between latent representations $\mathbf{Z}_{\text{RNA}}$ and $\mathbf{Z}_{\text{ATAC}}$, which `scMultiNODE` will use as the guidance of alignments. Specifically, for each modality, we compute intra-modality distance matrices $\mathbf{D}^{\text{RNA}} = \text{kNN}(\mathbf{Z}_{\text{RNA}}, \mathbf{Z}_{\text{RNA}})$ and $\mathbf{D}^{\text{ATAC}} = \text{kNN}(\mathbf{Z}_{\text{ATAC}}, \mathbf{Z}_{\text{ATAC}})$ based on k-nearest neighbor (kNN), as done previously in [6,18,17]. Then, GW finds the optimal cell correspondence matrix $\mathbf{T}$, where $\mathbf{T}_{ij}$ indicates the alignment probability of RNA cell i and ATAC cell j, by minimizing the GW distance

$$\text{GW distance} = \sum_{i,i',j,j'} \parallel \mathbf{D}^{\text{RNA}}_{ii'} - \mathbf{D}^{\text{ATAC}}_{jj'} \parallel^2 \mathbf{T}_{ij}\mathbf{T}_{i'j'} \tag{4}$$

However, the exact computation of GW distance is NP-hard and requires expensive computational costs for large-scale single-cell datasets. Therefore, scMultiNODE utilizes a recently proposed approximation algorithm, Quantized Gromov-Wasserstein (QGW) [11], to break the GW problem into smaller subproblems and significantly speed up computations. We binarize predicted $\mathbf{T}$ such that for a RNA cell i, $\mathbf{T}_{ij} = 1$ if ATAC cell j has the highest alignment probability with i; otherwise $\mathbf{T}_{ij} = 0$. Notice that scMultiNODE does not require cell correspondence information as a prior and predicts them in an unsupervised manner. If cell correspondence is known (even partially) ahead, we can encode the correspondence information into $\mathbf{T}$ and skip the alignment procedure. Our experiments assume no cell correspondence is given and conduct fully unsupervised integration.

With the cell correspondence matrix $\mathbf{T}$, scMultiNODE aligns modalities in the joint latent space by minimizing the loss function

$$\begin{gathered}
\mathcal{L}_{\text{fusion}} = \alpha \sum_{i,j} \mathbf{1}\left(\mathbf{T}_{ij} \neq 0\right) \parallel \tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j \parallel_2^2 + \text{MSE}\left(\mathbf{X}_{\text{ALL}},\ \widehat{\mathbf{X}}_{\text{ALL}}\right) + \text{MSE}\left(\mathbf{Y}_{\text{ALL}},\ \widehat{\mathbf{Y}}_{\text{ALL}}\right) \quad \text{where} \\
\tilde{\mathbf{z}}_i = \text{Fus}(\mathbf{z}_i, \omega) \text{ and } \tilde{\mathbf{z}}_j = \text{Fus}(\mathbf{z}_j, \omega) \text{ for RNA cell } i \text{ and ATAC cell } j \quad \text{(from Eq. 3)} \\
\widehat{\mathbf{X}}_{\text{ALL}} = \text{Dec}_{\mathbf{X}}(\tilde{\mathbf{Z}}_{\text{RNA}},\ \theta_{\mathbf{X}}) \text{ with } \tilde{\mathbf{Z}}_{\text{RNA}} = \{\tilde{\mathbf{z}}_i \mid \text{all RNA cells } i\}, \\
\widehat{\mathbf{Y}}_{\text{ALL}} = \text{Dec}_{\mathbf{Y}}(\tilde{\mathbf{Z}}_{\text{ATAC}},\ \theta_{\mathbf{Y}}) \text{ with } \tilde{\mathbf{Z}}_{\text{ATAC}} = \{\tilde{\mathbf{z}}_j \mid \text{all ATAC cells } j\}.
\end{gathered} \tag{5}$$

Here, $\mathbf{1}(s)$ is an indication function defined as $\mathbf{1}(s) = 1$ if statement s is true, and $\mathbf{1}(s) = 0$ otherwise. Therefore, when training the fusion layer we minimize the distance between the latent representations of the cells that are mapped by the GW optimal transport alignment. Hyperparameter α is a loss term coefficient. scMultiNODE freezes the parameter of two encoders ($\phi_{\mathbf{X}}$ and $\phi_{\mathbf{Y}}$) and updates fusion layer (ω) and decoders ($\theta_{\mathbf{X}}$ and $\theta_{\mathbf{Y}}$). Because decoders now operate on the joint latent space, we add the MSE reconstruction loss in the fusion loss Eq. 5 and update decoders accordingly. After this step, scMultiNODE has learned a joint latent space that aligns two modalities in an unsupervised manner and captures the cell type variations. In the next step, scMultiNODE further integrates cellular dynamics in the joint latent space.

2.3 scMultiNODE integrates cellular dynamics with neural ODE

scMultiNODE uses neural ODEs to explicitly model cell developmental dynamics in the joint latent space. ODE describes how a quantity a changes with respect to an independent variable b, such that $\mathrm{d}a = f(a; b)\mathrm{d}b$ where function f represents the derivative. Therefore, we can use differential equations to model how cell states change with respect to time in the joint latent space. But finding the solution of the derivative function f through numerical methods is intractable and computationally expensive [29]. Therefore, recent studies have adopted neural networks to approximate the derivative function and proposed neural ODEs [9]. Previous studies [41,10,55] have used neural ODEs (with respect to time) to construct continuous-time trajectories and model single-cell development for gene expression data. Here, scMultiNODE quantifies changes of cell latent representation $\tilde{\mathbf{Z}}^{(t)}$ in the joint latent space at time t through a neural ODE

$$\mathrm{d}\tilde{\mathbf{Z}}^{(t)} = \text{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right) \cdot \mathrm{d}t. \tag{6}$$

Here, Drift is a non-linear neural network parameterized by μ, modeling the developmental cell velocities in the joint latent space, such that $\text{Drift}(\tilde{\mathbf{Z}}^{(t)}, \mu)$ represents the direction and strength of cellular transitions. scMultiNODE calculates the initial condition $\tilde{\mathbf{Z}}^{(0)}$ of cells at the first time $t = 0$, defined as the earliest observed timepoint for both modalities. The encoder corresponding to the modality with the earliest timepoint, along with the fusion layer (pre-trained in the previous step), is used to determine the initial condition. scMultiNODE then predicts the subsequent cell states step-wise at any timepoint t through (assume RNA modality has the first timepoint as $0 \in \mathcal{T}_{\text{RNA}}$)

$$\begin{gathered}
\tilde{\mathbf{Z}}^{(t+\Delta t)} = \tilde{\mathbf{Z}}^{(t)} + \text{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right) \Delta t \\
\text{with} \qquad \tilde{\mathbf{Z}}^{(0)} = \text{Fus}\left(\mathbf{Z}_{\text{RNA}}^{(0)},\ \omega\right) \quad \text{and} \quad \mathbf{Z}_{\text{RNA}}^{(0)} = \text{Enc}_{\mathbf{X}}\left(\mathbf{X}^{(0)},\ \phi_{\mathbf{X}}\right),
\end{gathered} \tag{7}$$

Here, hyperparameter Δt denotes step size and drift term $\text{Drift}\left(\tilde{\mathbf{Z}}^{(t)}, \mu\right) \Delta t$ represents the forward steps taken in the joint latent space. We use the first-order Euler method (in Eq. 7) for convenience of explanation and one can specify any ODE solver in our implementations.

To fit the continuous trajectory (controlled by Drift neural network) to the observations, `scMultiNODE` minimizes the difference between the input and the reconstructed data. Specifically, at each measured timepoint $t \in \mathcal{T}_{\text{RNA}}$ or $t \in \mathcal{T}_{\text{ATAC}}$, `scMultiNODE` uses the decoder $\text{Dec}_{\mathbf{X}}/\text{Dec}_{\mathbf{Y}}$ to convert latent variables $\tilde{\mathbf{Z}}^{(t)}$ generated from Eq. 7 back to the high-dimensional feature space. Because we have no correspondence between true cells and cells generated from the ODE model, `scMultiNODE` utilizes the Wasserstein metric [14] to measure the distance between distributions defined by ground truth $\mathbf{X}$ and predictions $\widehat{\mathbf{X}}$ as

$$\text{Wass}(\mathbf{X}, \widehat{\mathbf{X}}) = \left(\min_{\Gamma \sim \Pi(\mathbf{X}, \widehat{\mathbf{X}})} \sum_{i,j} D_{ij}^2 \Gamma_{ij} \right)^{1/2} \qquad \text{with} \qquad D_{ij} = \|\mathbf{X}_i - \widehat{\mathbf{X}}_j\|_2. \tag{8}$$

Wasserstein distance measures the distribution distance in the same metric space, different from GW version that is applied for aligning different metric spaces. Here, $\Pi(\mathbf{X}, \widehat{\mathbf{X}})$ denotes the set of all transport plans between each cell of $\mathbf{X}$ and $\widehat{\mathbf{X}}$ and D_{ij} represents the ℓ_2 distance, such that the Wasserstein metric adopts the minimal-cost transport plan Γ to measure the data dissimilarity. `scMultiNODE` utilizes Wasserstein distance as reconstruction loss when training the neural ODE.

Furthermore, to integrate the cellular dynamics captured by neural ODE into the joint latent space, `scMultiNODE` uses a dynamic regularization term to update the joint latent space and capture both local cellular variations and the global developmental dynamics. The dynamic regularization is proposed in our previous work [55] for modeling temporal scRNA-seq data, which incorporates cellular dynamics into the latent space to make it more robust to distribution shifts in the measurements across time. Here, we extend dynamic regularization in the multi-modal integration setting. Specifically, the dynamic regularization minimizes the difference between the joint latent representations generated by the fusion layer (i.e., $\text{Fus}\left(\text{Enc}_{\text{RNA}}(\mathbf{X}, \phi_{\mathbf{X}})\right)$ or $\text{Fus}\left(\text{Enc}_{\text{ATAC}}(\mathbf{Y}, \phi_{\mathbf{Y}})\right)$) and the dynamics learned by the ODE (i.e., $\tilde{\mathbf{Z}}$). Because we have no correspondence between them, `scMultiNODE` again uses Wasserstein distance to evaluate their difference at each timepoint and defines the dynamic regularization as

$$\begin{aligned}
\mathcal{R}\left(\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}}\right) = \sum_{t \in \mathcal{T}_{\text{RNA}}} \text{Wass}\left(\tilde{\mathbf{Z}}^{(t)}_{\text{fus}},\ \tilde{\mathbf{Z}}^{(t)}\right) + \sum_{t \in \mathcal{T}_{\text{ATAC}}} \text{Wass}\left(\tilde{\mathbf{Z}}^{(t)}_{\text{fus}},\ \tilde{\mathbf{Z}}^{(t)}\right) \text{ with} \\
\begin{cases} \tilde{\mathbf{Z}}^{(t)}_{\text{fus}} = \text{Fus}\left(\text{Enc}_{\text{RNA}}(\mathbf{X}^{(t)}, \phi_{\mathbf{X}}), \omega\right) & \text{for } t \in \mathcal{T}_{\text{RNA}} \\ \tilde{\mathbf{Z}}^{(t)}_{\text{fus}} = \text{Fus}\left(\text{Enc}_{\text{ATAC}}(\mathbf{Y}^{(t)}, \phi_{\mathbf{Y}}), \omega\right) & \text{for } t \in \mathcal{T}_{\text{ATAC}} \end{cases}, \\
\text{and} \quad \tilde{\mathbf{Z}}^{(t)} \text{ comes from neural ODE} \quad (\text{Eq. } 7)
\end{aligned} \tag{9}$$

Therefore, `scMultiNODE` jointly optimizes AEs, fusion layer, and neural ODE components by minimizing the regularized loss function

$$\mathcal{L}_{\text{dyn}} = \sum_{t \in \mathcal{T}_{\text{RNA}}} \text{Wass}\left(\mathbf{X}^{(t)},\ \widehat{\mathbf{X}}^{(t)}\right) + \sum_{t \in \mathcal{T}_{\text{ATAC}}} \text{Wass}\left(\mathbf{Y}^{(t)},\ \widehat{\mathbf{Y}}^{(t)}\right) + \beta \mathcal{R}\left(\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}}\right), \tag{10}$$

so that the overall dynamics update the final latent space of `scMultiNODE` through dynamic regularization and corresponding hyperparameter β. The embedding of cellular dynamics improves upon previous integration models, whose integration focuses solely on cell type variations. This improvement allows `scMultiNODE` to fit the data more effectively, resulting in a joint latent space that is both more robust and interpretable, as it captures cellular dynamics alongside variations in cell types. Pseudocodes for `scMultiNODE` are provided in Supplementary Sec. S2.

3 Experiment Setup

Datasets We use four publicly available developmental single-cell datasets with scRNA-seq and scATAC-seq assays to demonstrate the capabilities of `scMultiNODE` in integrating modalities in an unsupervised manner. These datasets are summarized in Table 1. They have multiple timepoints and cover different species and tissues. To make computations tractable, we relabel timepoints with consecutive natural numbers starting from 0. In each experiment, we select the top 2000 most highly variable genes (HVGs) for the scRNA-seq assay and normalize the unique molecular identifier (UMI) count expression through a log transformation with pseudo-count. For scATAC-seq measurements, we select the top 2000 most variable features. We use the data after removing batch effects among different timepoints (see details in Supplementary Sec. S1).

Table 1: Data descriptions of the four real-world single-cell datasets used in experiments.

ID	Dataset	Species	# RNA cells	# RNA timepoints	# ATAC cells	# ATAC timepoints	Co-assay Data	Source
HC	human cortex	*Homo sapiens*	2277	10	2277	10	Yes	[56]
HO	human organoid	*Homo sapiens*	10000	11	10000	10	Yes	[21]
DR	drosophila	*Drosophila melanogaster*	2738	11	4246	11	No	[5]
MN	mouse neoctex	*Mus musculus*	6098	3	1914	3	No	[53]

Baselines We compare `scMultiNODE` with six state-of-the-art unsupervised single-cell integration methods that are capable of aligning different single-cell measurements and computing a joint multi-modal latent space. The single-cell analysis platform `Seurat` [24] projects two datasets into a common space with linear canonical correlation analysis that maximizes cross-dataset correlation. `UnionCom` [6] matches two datasets based on geometrical matrix matching. `uniPort` [7] computes latent space with coupled variation autoencoder and aligns cells with optimal transport. `Pamona` [8], `SCOTv1` [18], and `SCOTv2` [17] adopt GW optimal transport to integrate different modalities. The details of baseline models are included in Supplementary Sec. S3.

Evaluation metrics We evaluate each model's integrated latent representations from three perspectives: modality integration, capturing cell type variation, and preserving cellular dynamics. Therefore, we adopt the following evaluation metrics (detailed descriptions and computations of these metrics are included in Supplementary Sec. S4):

- **Modality integration:** We use `batch entropy` to evaluate the integration of unaligned datasets (DR and MN). `Batch entropy` is originally introduced in Xiong et al. [51] and previously adopted by Cao et al.[7]. It evaluates the sum of regional mixing entropies between different datasets where a high score indicates cells from different modalities are mixed well. For the co-assay datasets (HC and HO) where one-to-one cell correspondence information is available, we additionally use the fraction of samples closer than the true match (`FOSCTTM`) [32,18], `neighborhood overlap` [6], and Spearman correlation coefficient (`SCC`). Specifically, for each data point in the joint latent space, `FOSCTTM` computes the fraction of data points that are closer than its true nearest neighbor (i.e., the matched cell). We average these fraction values for all the cells in both modalities. A perfect integration implies that all cells should be closest to their true match, resulting in a `FOSCTTM` of zero. Therefore, a lower `FOSCTTM` value denotes a better integration performance. Furthermore, the `neighborhood overlap` is defined similarly and computes the ratio of cells that can find their correspondence cells from the other dataset in their neighborhood. We use the averaged ratio of `neighborhood overlap` of the two modalities. The `neighborhood overlap` ranges from 0 to 1, and a higher value implies a better recovery of cell-to-cell correspondence between the two modalities. Lastly, based on the intuitive assumption that matched cells should have similar latent representations in the joint latent space, we use `SCC` to evaluate representation similarities between matched cells, such that a better integration leads to a higher `SCC` value.
- **Cell type variation:** We also evaluate integration using cell type labels through label transfer accuracy (`LTA-type`) as in previous studies [6,7,17]. This metric assesses the clustering of cell types after integration by training a k-nearest neighbor (kNN) classifier on joint latent representations of one modality and then evaluates its predictive accuracy on another modality. It ranges from 0 to 1, and a higher metric value indicates better integration performance as cells that belong to the same cell type are aligned close together.
- **Cellular dynamics:** As the main objective of our research, we evaluate how well the integration captures the cellular variations across different timepoints. Therefore, we compute label transfer accuracy using timepoint labels (named as `LTA-time`), such that a higher `LTA-time` indicates better integration performance as cells that belong to the same timepoint are aligned close together. Furthermore, we hypothesize that good latent representations, if they retain the developmental dynamics, should highly correlate with the timepoint label. Therefore, we define the `time correlation`, which computes the distance correlation [45] between cell representations in the joint latent space and their corresponding timepoint labels. The distance correlation measures linear and nonlinear association between two datasets of arbitrary dimensions. Hence, the `time correlation` ranges from 0 to 1, where a higher value implies a better integration, which is highly associated with cellular dynamics.

Hyperparameter tuning On co-assay datasets (HC and HO), we select corresponding hyperparameters for all methods (scMultiNODE and baselines) that yield the minimum FOSCTTM value; on unaligned datasets (DR and MN), we select hyperparameters that yield the maximum LTA-type based on common cell type labels. We use Optuna [2] to automatically determine the optimal hyperparameters and use sufficiently large hyperparameter ranges for search and evaluation. The hyperparameter search ranges of scMultiNODE and baselines are listed in Supplementary Table S1. We set the joint latent space dimension as 50 for all methods. We use the first-order Euler ODE solver and set ODE step size $\Delta t = 0.1$ in scMultiNODE . We run each method for sufficient iterations to ensure they converge. Moreover, we evaluate scMultiNODE performance using different hyperparameter settings, conduct ablation studies, and give heuristic guidance on how to set hyperparameters in real-world scenarios in Sec. 4.3.

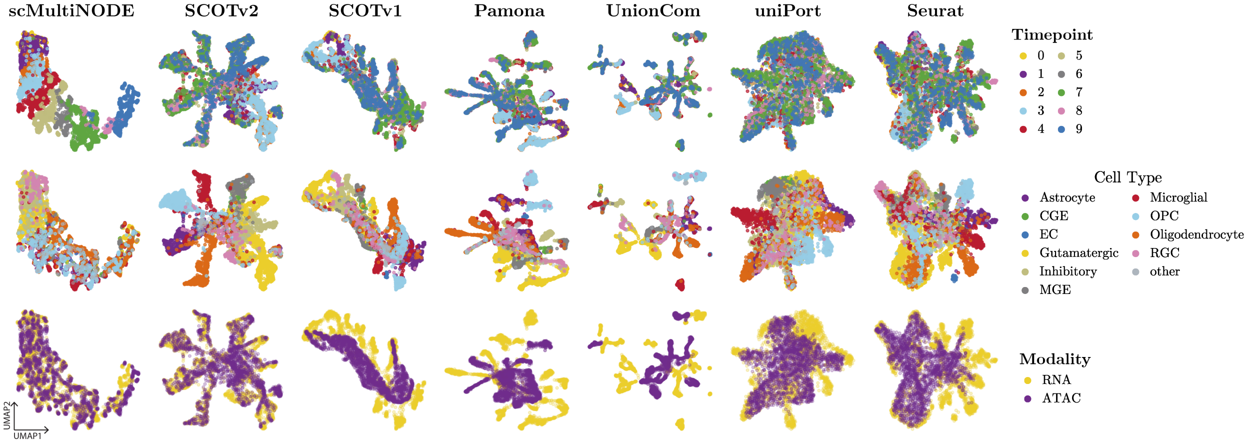

Fig. 2: 2D UMAP visualization of joint latent representations on the HC dataset. The representations are colored by timepoint labels, cell types, and modalities.

4 Experiment Results

4.1 scMultiNODE captures cellular developmental dynamics during multi-modal integration

We compare scMultiNODE 's performance with the baseline methods for aligning scRNA-seq and scATAC-seq measurements across multiple timepoints. Fig. 2 visualizes the joint latent representations of the HC dataset in 2D Uniform Manifold Approximation and Projection (UMAP) [35] space. Integration results for the other three datasets are shown in Supplementary Fig. S1-S3. The visual results indicate that scMultiNODE effectively aligns both modalities for all the timepoints, while capturing the cellular dynamics. The baseline models, on the other hand, either cannot align two modalities across the timepoints well (e.g., Pamona and UnionCom) or ignore the temporal structure (e.g., SCOTv2) with the cell type signal dominating the joint space. Furthermore, for unaligned datasets (DR and MN) as well, scMultiNODE effectively integrates the two modalities (Supplementary Fig. S2 and S3). Qualitative evaluation of all datasets indicates that scMultiNODE can effectively preserve cell type variations and cellular dynamics during integration, even in unaligned datasets.

Next, we quantitatively evaluate the integration performance of all the methods using the evaluation metrics introduced in Sec. 3. Fig. 3 shows the performance, evaluating integration from the three different perspectives. Detailed lists of all these metrics' values are included in Supplementary Tables S2 and S3.

For batch entropy and time correlation (Fig. 3A-B), scMultiNODE clearly outperforms the baselines for all datasets, meaning it can align the two modalities well while preserving the underlying dynamics. In Fig. 3C, we plot the LTA-type and LTA-time scores on the X and Y axes for all the methods, respectively. These scores measure how well the variations in the dataset are preserved upon integration. LTA-type captures this for cell type variations and LTA-time for temporal variations. A high position in the scatter plot for scMultiNODE indicates that we retain better overall performance in preserving both variations. This result highlights scMultiNODE 's superior capability to learn the heterogeneity in the developmental scRNA-seq and scATAC-seq datasets when integrating them. For example, on the HO and DR datasets, scMultiNODE has the highest LTA-type and LTA-time scores (Fig. 3C), denoting its good performance in capturing variations of both cell type and timepoints. On the HC and MN datasets, scMultiNODE balances the trade-off between the cell type and timepoint variations, where it reports a median level of LTA-type score across all models and the highest LTA-time. Despite this trade-off, scMultiNODE obtains the best integration (Fig. 3A) for these two methods with batch entropy=0.67 (HC) and 0.50 (MN). Additionally, Fig. 3D shows SCC, neighborhood overlap, and FOSCTTM, the three metrics only possible to calculate for co-assay datasets. scMultiNODE consistently has good performance with respect to these evaluations. Even when scMultiNODE is not the best model in some cases, its integration performs similarly to the best. For example, on HC dataset, scMultiNODE has SCC=0.88 while the best baseline SCOTv1 has SCC=0.89.

Therefore, these experiments and analyses demonstrate that scMultiNODE outperforms existing methods in integrating temporal multi-modal single-cell data, effectively handling both co-assay and unaligned datasets while capturing variations in cell type and cellular dynamics. Additionally, our findings emphasize the importance of evaluating multi-modal integration from multiple perspectives, as demonstrated in our experiments (Sec. 3), to ensure a thorough and comprehensive assessment.

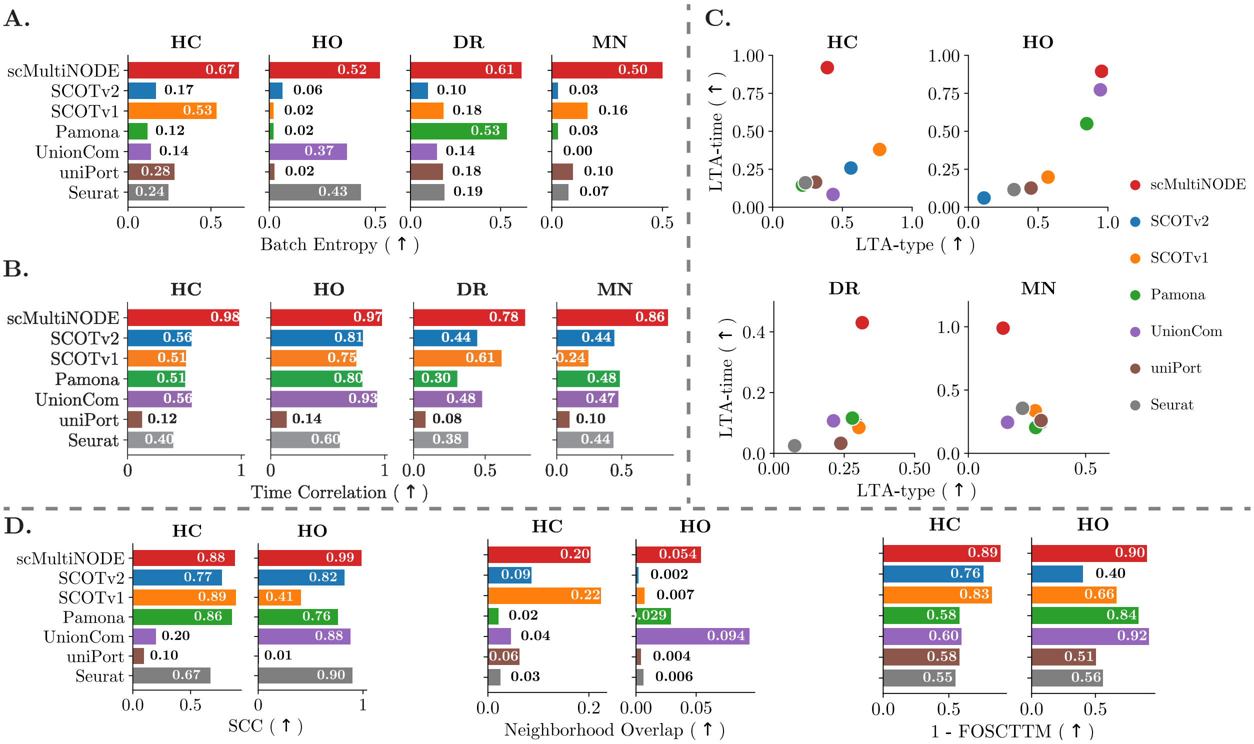

Fig. 3: Integration evaluation on four datasets, including two co-assay datasets (HC and HO) and two unaligned datasets (DR and MN). (**A** and **B**) The batch entropy and time correlation. (**C**) We plot the LTA-type and LTA-time scores on the X and Y axes for all the methods, respectively.LTA-type measures cell type variations and LTA-time for temporal variations. (**D**) Three metrics that are only available on co-assay datasets: SCC, neighborhood overlap, and FOSCTTM. We show 1-FOSCTTM instead of FOSCTTM here to unify figure plotting for all metrics, where a higher metric value implies better integration performance.

4.2 scMultiNODE 's latent space assists with understanding cell state transition

Here, we demonstrate that scMultiNODE can learn an interpretable joint latent space, which allows us to study the cell developmental transitions for multi-modal single-cell datasets. We use the joint latent space learned for the HC dataset for this task. We pick this dataset as it contains detailed cell annotations of human cerebral cortex development, allowing us to validate our analysis.

First, we train scMultiNODE on the HC dataset and map all cells to the joint latent space. In this joint space, we construct a most probable path between two cell populations through the Least Action Path (LAP) method [41,40,49]. LAP finds the optimal path between any two cell states while minimizing its action and transition time (see details in Supplementary Sec. S6). In Fig. 4A, we construct these LAP paths from cells at the starting point (i.e., $t = 0$) to oligodendrocyte (OL) and glutamatergic neuron (GN) populations. These paths are colored in purple and green, respectively.

After the optimal path is constructed, we use the widely used Wilcoxon rank-sum test to find differentially expressed (DE) genes along each path. These genes represent the potential key driver genes for the calculated cell development paths. Fig. 4B-C show the expression of top-rank DE genes for the GN path (SV2B and ANO3) and OL path (SOX6 and SLC1A3) across the timepoints. We plot their normalized values (calculated using z-scores) per timepoint averaged for the cells on the path (colored solid line). We also plot their z-score values for cells out of the path (colored dotted line) to show the distinct variations of these genes. Moreover, we plot the average z-score of five randomly selected genes for the cells on the path (black solid line) and out of the path (black dotted line). These arbitrarily chosen genes have relatively stable expression levels across all timepoints. In comparison, we see that the four DE genes (SV2B, ANO3, SOX6, and SLC1A3) have a higher variance associated with the corresponding cell development paths.

Next, we corroborate these findings with existing literature. Previous studies [37,12] have found that SV2B transcript is expressed in glutamatergic neurons. The Human Protein Atlas [28] shows that ANO3 is enriched in excitatory neurons, majorly consisting of glutamatergic neurons in the central nervous system. On the other hand, SOX6 plays an important role in the central nervous system by regulating oligodendrocyte proliferation [23]. Also, as The Human Protein Atlas shows, SLC1A3 is enhanced in oligodendrocytes.

Finally, we compare the genes obtained from our joint latent space analysis with the cell type marker genes obtained from scRNA-seq/scATAC-seq datasets using the traditional single-cell analysis pipeline (Supplementary Table S4). The RNA-derived marker genes distinguishing glutamatergic neurons lack the top DE genes (SV2B and ANO3) found in our joint latent space analysis, despite their known roles in cell development. The oligodendrocyte marker genes also exclude OL path driver genes SOX6 and SLC1A3. Therefore, these findings suggest that explicitly modeling cell dynamics aids in identifying development-related DE genes better than marker gene analysis of cell types. Interestingly, SV2B appears as a marker gene in chromatin accessibility data, further highlighting the need to perform joint analysis of these two modalities.

Overall, `scMultiNODE` effectively captures an interpretable joint latent space, enabling the detection of cell development-related genes by leveraging information from both modalities.

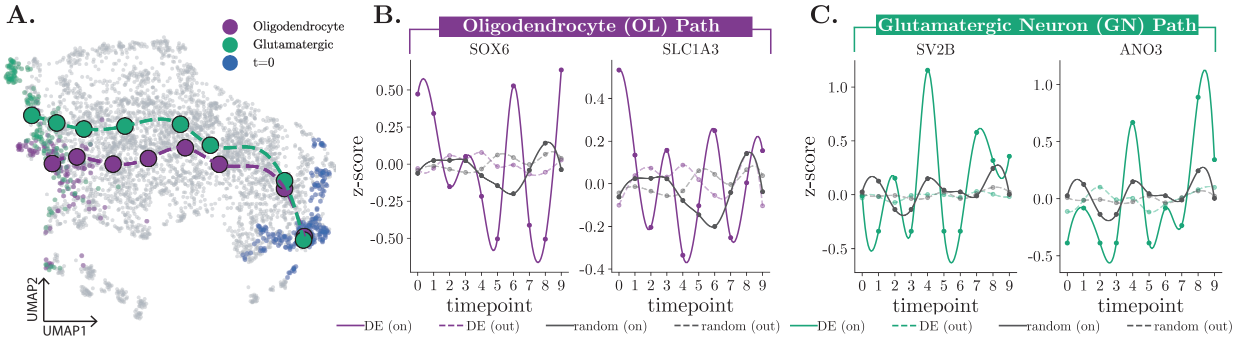

Fig. 4: `scMultiNODE` downstream analysis on the HC dataset. (**A**) 2D UMAP visualization of the least action path from cells at the starting point ($t = 0$) to the oligodendrocyte (OL) and glutamatergic neuron (GN) populations, respectively. (**B**, **C**) Gene expression z-score values of the top-rank DE genes: SOX6 and SLC1A3 of OL path; SV2B and ANO3 of GN path. We plot DE gene z-scores per timepoint averaged for cells on the path (colored solid line) and out of the path (colored dotted line). We also show the average z-score of five random genes for cells on the path (black solid line) and out of the path (black dotted line).

4.3 Investigation of relevant hyperparameters in `scMultiNODE` for user guidance

We test `scMultiNODE` 's performance on one co-assay dataset (HC) and one unaligned dataset (DR) when using different hyperparameter settings. We first run `scMultiNODE` with the joint latent space size d varies from $\{10, 50, 100, 150, 200\}$. Supplementary Table S5 shows that `scMultiNODE` is robust in terms of the size of the latent dimensionality. Users can choose to set a reasonable latent dimension based on a tradeoff between accuracy and computational costs. State-of-the-art methods [25] generally choose a latent space of 10 to 50 dimensions. For a fair comparison, we set the latent size $d = 50$ for all methods in our experiments.

`scMultiNODE` uses GW optimal transport to align cell representations from two modalities and ensures aligned cells have similar latent presentations. The GW algorithm calculates the intra-modality distance matrix using kNN. So, we vary the number of neighbors k to be considered in kNN from $\{5, 10, 50, 100, 150, 200\}$. As shown in Supplementary Table S6, `scMultiNODE` outperforms the best baseline model in terms of integration quality, with little impact from changing the number of neighbors. Additionally, the coefficient α for the matched cell integration loss (in Eq. 5) is varied from $\{0.0, 0.01, 0.1, 1.0, 10.0, 100.0\}$. Supplementary Table S7 indicates that performance drops noticeably when $\alpha = 0$, where matched cells are not encouraged to converge in the latent space. Specifically, on the DR dataset, `scMultiNODE` shows `LTA-type`=0.397, `LTA-time`=0.280, and `time correlation`=0.477 when $\alpha = 0$, significantly lower than when $\alpha > 0$ (`LTA-type`> 0.5, `LTA-time`> 0.5, and `time correlation`> 0.7). However, the `batch entropy` value remains similar when $\alpha = 0$ (0.422) and $\alpha > 0$ (0.394 on average), meaning modalities are mixed well even if the model does not enforce it. This suggests that `scMultiNODE` can still achieve some degree of integration due to the shared fusion layer between modalities, while the cell type variations and dynamics are not preserved. Nonetheless, the cell integration loss term (Eq. 5) is essential for learning a joint latent space that effectively captures the diverse variations.

Lastly, `scMultiNODE` uses the dynamic regularization controlled by β to incorporate learned dynamics into the joint latent space. We vary $\beta \in \{0.0, 0.01, 0.1, 1.0, 10.0, 100.0\}$. Supplementary Table S8 shows that removing the dynamic regularization (i.e., $\beta = 0$) results in poor integration where cells cannot be aligned at all (with `batch entropy` close to 0 on both datasets) and cellular dynamics are lost (`time correlation`=0.312 for HC and 0.357 for DR). On adding this regularization (i.e., $\beta > 0$), the joint latent space has much better integration and can learn cellular dynamics to model the development accurately. We also note that a very large β may break down the model training and lead to bad performance. For example, `scMultiNODE` 's performance significantly decreases when $\beta = 100$ on the HC dataset. These results imply that the dynamic regularization is essential for aligning modalities and capturing dynamics. Users should select β carefully within a reasonable range of $[0.01, 10.0]$.

4.4 `scMultiNODE` scales well with increasing number of cells

We compare the runtime of `scMultiNODE` with the baseline models using different numbers of cells from the HO dataset. We pick this dataset as it has the most number of cells (Table 1), enabling us to test runtime costs on many cells. All methods are run on the Intel Xeon Platinum 8268 CPU with 32GB memory. As shown in Fig. 5, `UnionCom` scales exponentially to the number of cells and `Pamona` significantly increases its time costs when there are many cells. `scMultiNODE` exhibits similar computational scaling to most baseline models. Therefore, despite incorporating an additional step for learning cellular dynamics, `scMultiNODE` does not significantly increase computational demands, making it suitable for large-scale temporal and multi-modal single-cell datasets.

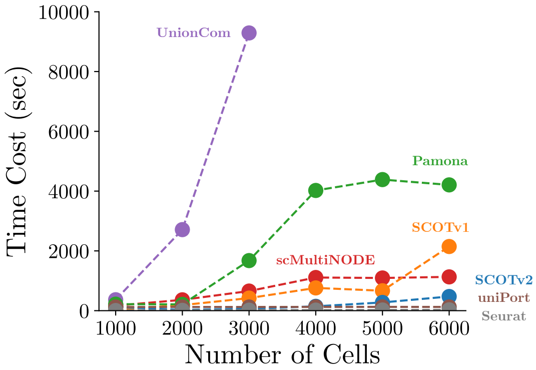

Fig. 5: Time costs as the number of cells increases.

5 Conclusion

We propose an unsupervised integration method called `scMultiNODE` . Given multi-modal temporal single-cell datasets, our model can align single-cell measurements without prior cell-to-cell correspondence information, and compute the joint latent space that retains both cell type variations and cellular dynamics. The integration enables critical downstream analyses, such as cell transition path construction and investigation of genes that change along this path. While we have focused on integrating gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) measurements, `scMultiNODE` can be extended to integrating any combinations of single-cell modalities.

For future work, we will incorporate prior biological knowledge, such as cell proliferation and gene regulation, which are important to cellular differentiation, into `scMultiNODE` to further improve its integration performance and the information-richness of the latent space. We will also test our model on other single-cell modalities (e.g., surface protein) or more than two modalities, to enable its broader use.

Funding

This project was funded by the National Institutes of Health (NIH) award 1R35HG011939-01.

References

1. Adossa, N., Khan, S., Rytkönen, K.T., Elo, L.L.: Computational strategies for single-cell multi-omics integration. Computational and Structural Biotechnology Journal **19**, 2588–2596 (2021)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
3. Bank, D., Koenigstein, N., Giryes, R.: Autoencoders. Machine learning for data science handbook: data mining and knowledge discovery handbook pp. 353–374 (2023)
4. Benton, A., Khayrallah, H., Gujral, B., Reisinger, D.A., Zhang, S., Arora, R.: Deep generalized canonical correlation analysis. arXiv preprint arXiv:1702.02519 (2017)
5. Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R.M., Martin, B., Dulja, A., Schaub, C., Trapnell, C., et al.: The continuum of drosophila embryonic development at single-cell resolution. Science **377**(6606), eabn5800 (2022)
6. Cao, K., Bai, X., Hong, Y., Wan, L.: Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics **36**(Supplement_1), i48–i56 (2020)
7. Cao, K., Gong, Q., Hong, Y., Wan, L.: A unified computational framework for single-cell data integration with optimal transport. Nature Communications **13**(1), 7419 (2022)
8. Cao, K., Hong, Y., Wan, L.: Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. Bioinformatics **38**(1), 211–219 (2022)
9. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
10. Chen, Z., King, W.C., Hwang, A., Gerstein, M., Zhang, J.: DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. Science Advances **8**(48), eabq3745 (2022)
11. Chowdhury, S., Miller, D., Needham, T.: Quantized gromov-wasserstein. In: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21. pp. 811–827. Springer (2021)
12. Ciruelas, K., Marcotulli, D., Bajjalieh, S.M.: Synaptic vesicle protein 2: A multi-faceted regulator of secretion. In: Seminars in cell & developmental biology. vol. 95, pp. 130–141. Elsevier (2019)
13. Connor, M., Canal, G., Rozell, C.: Variational autoencoder with learned latent structure. In: International conference on artificial intelligence and statistics. pp. 2359–2367. PMLR (2021)
14. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
15. Dance, A.: What is a cell type, really? the quest to categorize life's myriad forms. Nature **633**(8031), 754–756 (2024)
16. Danese, A., Richter, M.L., Chaichoompu, K., Fischer, D.S., Theis, F.J., Colomé-Tatché, M.: Episcanpy: integrated single-cell epigenomic analysis. Nature Communications **12**(1), 5228 (2021)
17. Demetci, P., Santorella, R., Chakravarthy, M., Sandstede, B., Singh, R.: Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. Journal of Computational Biology **29**(11), 1213–1228 (2022)
18. Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., Singh, R.: Scot: single-cell multi-omics alignment with optimal transport. Journal of computational biology **29**(1), 3–18 (2022)
19. Ding, J., Sharon, N., Bar-Joseph, Z.: Temporal modelling using single-cell transcriptomics. Nature Reviews Genetics **23**(6), 355–368 (2022)
20. Feydy, J., Séjourné, T., Vialard, F.X., Amari, S.i., Trouvé, A., Peyré, G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 2681–2690. PMLR (2019)
21. Fleck, J.S., Jansen, S.M.J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M., He, Z., Camp, J.G., et al.: Inferring and perturbing cell fate regulomes in human brain organoids. Nature **621**(7978), 365–372 (2023)
22. Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome biology **20**(1), 296 (2019)
23. Hagiwara, N.: Sox6, jack of all trades: a versatile regulatory protein in vertebrate development. Developmental Dynamics **240**(6), 1311–1321 (2011)

24. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al.: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature biotechnology **42**(2), 293–304 (2024)
25. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., et al.: Best practices for single-cell analysis across modalities. Nature Reviews Genetics pp. 1–23 (2023)
26. Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.Y., Xue, Z., Fan, G.: Simultaneous profiling of transcriptome and dna methylome from a single cell. Genome biology **17**, 1–11 (2016)
27. Huguet, G., Magruder, D.S., Tong, A., Fasina, O., Kuchroo, M., Wolf, G., Krishnaswamy, S.: Manifold interpolating optimal-transport flows for trajectory inference. Advances in Neural Information Processing Systems **35**, 29705–29718 (2022)
28. Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., et al.: A single–cell type transcriptomics map of human tissues. Science advances **7**(31), eabh2169 (2021)
29. Kidger, P.: On neural differential equations. arXiv preprint arXiv:2202.02435 (2022)
30. Liu, B., Hu, X., Feng, K., Gao, R., Xue, Z., Zhang, S., Zhang, Y., Corse, E., Hu, Y., Han, W., et al.: Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted t cells during anti-PD-1 therapy in lung cancer. Nature Cancer **3**(1), 108–121 (2022)
31. Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E.Z., Welch, J.D.: Jointly defining cell types from multiple single-cell datasets using liger. Nature protocols **15**(11), 3632–3662 (2020)
32. Liu, J., Huang, Y., Singh, R., Vert, J.P., Noble, W.S.: Jointly embedding multiple single-cell omics measurements. In: Algorithms in bioinformatics:... International Workshop, WABI..., proceedings. WABI (Workshop). vol. 143. NIH Public Access (2019)
33. Ma, Q., Xu, D.: Deep learning shapes single-cell data analysis. Nature Reviews Molecular Cell Biology **23**(5), 303–304 (2022)
34. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al.: Chromatin potential identified by shared single-cell profiling of rna and chromatin. Cell **183**(4), 1103–1116 (2020)
35. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
36. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
37. Pazarlar, B.A., Aripaka, S.S., Petukhov, V., Pinborg, L., Khodosevich, K., Mikkelsen, J.D.: Expression profile of synaptic vesicle glycoprotein 2a, b, and c paralogues in temporal neocortex tissue from patients with temporal lobe epilepsy (tle). Molecular brain **15**(1), 45 (2022)
38. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
39. Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al.: Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. Molecular cell **71**(5), 858–871 (2018)
40. Qiu, X., Ding, S., Shi, T.: From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. PloS one **7**(12), e49271 (2012)
41. Qiu, X., Zhang, Y., Martin-Rufino, J.D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A.N., Hein, M.Y., Min, K.H.J., Wang, L., et al.: Mapping transcriptomic vector fields of single cells. Cell **185**(4), 690–711 (2022)
42. Ramos-Carreño, C., Torrecilla, J.L.: dcor: Distance correlation and energy statistics in Python. SoftwareX **22** (2 2023). https://doi.org/10.1016/j.softx.2023.101326, https://www.sciencedirect.com/science/article/pii/S2352711023000225
43. Spiller, D.G., Wood, C.D., Rand, D.A., White, M.R.: Measurement of single-cell dynamics. Nature **465**(7299), 736–745 (2010)
44. Sun, S., Zhu, J., Ma, Y., Zhou, X.: Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. Genome biology **20**, 1–21 (2019)
45. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances (2007)
46. Tong, A., Huang, J., Wolf, G., Van Dijk, D., Krishnaswamy, S.: Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In: International conference on machine learning. pp. 9526–9536. PMLR (2020)
47. Tran, D., Nguyen, H., Tran, B., La Vecchia, C., Luu, H.N., Nguyen, T.: Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature communications **12**(1), 1029 (2021)
48. Trapnell, C.: Defining cell types and states with single-cell genomics. Genome research **25**(10), 1491–1498 (2015)
49. Wang, P., Song, C., Zhang, H., Wu, Z., Tian, X.J., Xing, J.: Epigenetic state network approach for describing cell phenotypic transitions. Interface focus **4**(3), 20130068 (2014)

50. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)
51. Xiong, L., Tian, K., Li, Y., Ning, W., Gao, X., Zhang, Q.C.: Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. Nature Communications **13**(1), 6118 (2022)
52. Yeo, G.H.T., Saksena, S.D., Gifford, D.K.: Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. Nature communications **12**(1), 3222 (2021)
53. Yuan, W., Ma, S., Brown, J.R., Kim, K., Murek, V., Trastulla, L., Meissner, A., Lodato, S., Shetty, A.S., Levin, J.Z., et al.: Temporally divergent regulatory mechanisms govern neuronal diversification and maturation in the mouse and marmoset neocortex. Nature Neuroscience **25**(8), 1049–1058 (2022)
54. Zakrzewski, W., Dobrzyński, M., Szymonowicz, M., Rybak, Z.: Stem cells: past, present, and future. Stem cell research & therapy **10**(1), 1–22 (2019)
55. Zhang, J., Larschan, E., Bigness, J., Singh, R.: scnode: generative model for temporal single cell transcriptomic data prediction. Bioinformatics **40**(Supplement_2), ii146–ii154 (2024)
56. Zhu, K., Bendl, J., Rahman, S., Vicari, J.M., Coleman, C., Clarence, T., Latouche, O., Tsankova, N.M., Li, A., Brennand, K.J., et al.: Multi-omic profiling of the developing human cerebral cortex at the single-cell level. Science Advances **9**(41), eadg3754 (2023)
57. Ziffra, R.S., Kim, C.N., Ross, J.M., Wilfert, A., Turner, T.N., Haeussler, M., Casella, A.M., Przytycki, P.F., Keough, K.C., Shin, D., et al.: Single-cell epigenomics reveals mechanisms of human cortical development. Nature **598**(7879), 205–213 (2021)

S1 Single-Cell Dataset and Pre-Processing

We use four publicly available developmental single-cell datasets with scRNA-seq and scATAC-seq assays to demonstrate the capabilities of `scMultiNODE` in integrating modalities in an unsupervised manner.

- **Human cortex (HC):** Zhu et al. generate transcriptomic and chromatin accessibility data using multi-omic single-nucleus RNA sequencing (snRNA-seq) and single-nucleus assay for transposase-accessible chromatin (snATAC-seq). The dataset profiles 45549 cells in total across a broad developmental time frame from human fetal cortical plate to adult specimens [56]. They have normalized data with scTransform [22] and removed batch effects. We use the processed data provided in its original paper, which contains normalized gene expression count data, and the gene activity matrix inferred from ATAC-seq that assesses chromatin accessibility at the gene body and promoter regions. We randomly sample 5% of cells and test our model on this subset with 2277 cells. For each modality, we select the top 2000 highly variable genes (HVGs) using Scanpy [50]. The HC data can be downloaded from the CELLxGENE portal (https://cellxgene.cziscience.com/collections/ceb895f4-ff9f-403a-b7c3-187a9657ac2c).
- **Human organoid (HO):** Fleck et al. have acquired paired single-cell transcriptome (scRNA-seq) and accessible chromatin (scATAC-seq) data with 34088 cells over a dense time course (spanning 4 days to 2 months) of human brain organoid developments [21]. The dataset collects brain organoids of the same batch that dissociated at multiple timepoints during brain organoids development. The original paper provides gene expression count data of RNA-seq and the gene activity matrix inferred from ATAC-seq. We randomly sample 10000 cells and test our model on this subset. For each modality, we also select the top 2000 HVGs. The HO data can be downloaded from Zenodo (https://zenodo.org/records/5242913).
- **Drosophila (DR):** Calderon et al. profile chromatin accessibility in almost 1 million nuclei and gene expression in half a million nuclei from eleven overlapping windows spanning the entirety of drosophila embryogenesis (0 to 20 hours) [5]. The dataset contains the scRNA-seq profile of 547805 cells and scATAC-seq measurements for 976460 cells. For each modality, we randomly sample 5% of cells and test our model on the gene expression count matrix of 2738 cells and chromatin peak matrix of 4246 cells. As in the Seurat workflow [24], we select the top 2000 HVGs of scRNA-seq data and the top 2000 variable peaks for scATAC-seq data. The original paper shows that the data are not confounded by batch effects. The DR data is downloaded from https://shendure-web.gs.washington.edu/content/members/DEAP_website/public/.
- **Mouse neocortex (MN):** Yuan et al. provide a single-cell dataset of transcriptional (scRNA-seq) and epigenomic (scATAC-seq) measurements over a time course spanning for mammalian neocortical neurons in both mouse and marmoset [53]. The batch effects across timepoints and mammalian libraries have been corrected with the Seurat package. We use the mouse neocortex data and randomly select 10% cells for both modalities, obtaining a gene expression count matrix of 6098 cells and a chromatin peak matrix for 1914 cells. We select the top 2000 HVGs of scRNA-seq data and the top 2000 variable peaks for scATAC-seq data using Scanpy and EpiScanpy [16]. The MN data is downloaded from Gene Expression Omnibus SuperSeries GSE204851 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE204851).

Due to the prohibiting computational costs of modality integration on large-scale datasets, we randomly select a subset of cells and test all models on these smaller datasets. Also, we sample different ratios of cells for different datasets to test models on data with different scales of size. To make computations tractable, we relabel timepoints with consecutive natural numbers starting from 0. We normalize the gene expression count matrix to remove cell-specific bias before conducting experiments. Specifically, given the count expression of cell i as $\mathbf{X}_i \in \mathbb{R}^p$, we normalize it by total counts over all genes

$$\mathbf{X}_i = \frac{\mathbf{X}_i}{\sum_{j=1}^{p} \mathbf{X}_i} * 10^4, \qquad \text{followed by} \qquad \mathbf{X}_{ij} = \log(\mathbf{X}_{ij} + 1). \tag{S1}$$

Because the HC dataset already provides normalized gene expression data, we normalize the scRNA-seq data matrix of the other three datasets.

S2 scMultiNODE Training

Our scMultiNODE is implemented with *Pytorch 1.13* [36] and is trained end-to-end. scMultiNODE training consists of three main steps. scMultiNODE first trains the AE components for each modality with all cells. We use Adam optimizer to train AEs by minimizing Eq. 2 with a learning rate of 0.001 and 1000 iterations. Then, scMultiNODE aligns modality-specific latent representations with GW optimal transport. In the GW algorithm, we construct the modality distance matrices $\mathbf{D}^{\text{RNA}}$ and $\mathbf{D}^{\text{ATAC}}$ through the k-nearest neighbor (kNN) graph. We adopt the approximation algorithm, Quantized Gromov-Wasserstein (QGW) [11], to solve the GW problem efficiently. Once the cell correspondence matrix $\mathbf{T}$ is estimated from the GW algorithm, scMultiNODE maps modality-specific latent representations to a joint latent space (through fusion layer $\text{Fus}(\cdot, \omega)$) by minimizing Eq. 5 ($\mathcal{L}_{\text{fusion}}$) through the Adam optimizer with a learning rate of 0.001 and 1000 iterations. Finally, scMultiNODE adopts neural ODE to model the cellular dynamics and incorporate the learned dynamic into the joint latent space by minimizing $\mathcal{L}_{\text{dyn}}$ (Eq. 10). We adopt batch training and use the Adam optimizer to train scMultiNODE with a learning rate of 0.001 and 2000 iterations. At each training iteration, we randomly select 64 cells at $t = 0$ as a batch and predict for every timepoint $t \in \mathcal{T}_{\text{RNA}} \cup \mathcal{T}_{\text{ATAC}}$. Because the Wasserstein distance computation is expensive, batch training improves training efficiency and enables scMultiNODE usage on large-scale datasets. We use *geomloss* [20] to compute Wasserstein distance with blur $= 0.05$ and scaling $= 0.5$. Pseudo-codes of scMultiNODE are provided in Algorithm 1.

Algorithm 1 scMultiNODE

```
Input: The set of timepoint indices $\mathcal{T}_{\text{RNA}}$ and $\mathcal{T}_{\text{ATAC}}$; gene expression matrices $\{\mathbf{X}^{(t)} \mid t \in \mathcal{T}_{\text{RNA}}\}$; chromatin peak/gene activity matrices $\{\mathbf{Y}^{(t)} \mid t \in \mathcal{T}_{\text{ATAC}}\}$; hyperparameters $\Delta t$, $\alpha$, $\beta$, the number of neighbors $k$ in kNN; randomly initialized neural networks $\text{Enc}_{\mathbf{X}}$, $\text{Enc}_{\mathbf{Y}}$, $\text{Dec}_{\mathbf{X}}$, $\text{Dec}_{\mathbf{Y}}$, fusion layer Fus, and the neural ODE drift network Drift.

( Step I: dimensionality reduction )
$\mathbf{X}_{\text{ALL}} = \text{CONCAT}\left(\mathbf{X}^{(t)} \mid t \in \mathcal{T}_{\text{RNA}}\right)$   // concatenate cells from RNA-seq measurements
$\mathbf{Y}_{\text{ALL}} = \text{CONCAT}\left(\mathbf{Y}^{(t)} \mid t \in \mathcal{T}_{\text{ATAC}}\right)$   // concatenate cells from ATAC-seq measurements
Optimize RNA-related AE parameters $\phi_{\mathbf{X}}$ and $\theta_{\mathbf{X}}$ to minimize $\mathcal{L}_{\text{RNA}}$   (Eq. 2)
Optimize ATAC-related AE parameters $\phi_{\mathbf{Y}}$ and $\theta_{\mathbf{Y}}$ to minimize $\mathcal{L}_{\text{ATAC}}$   (Eq. 2)

( Step II: modality integration )
Construct intra-modality distance matrices $\mathbf{D}^{\text{RNA}}$ and $\mathbf{D}^{\text{ATAC}}$ with representations $\mathbf{Z}_{\text{RNA}}$ and $\mathbf{Z}_{\text{ATAC}}$
Use Quantized Gromov-Wasserstein (QGW) algorithm [11] to predict cell correspondence matrix $\mathbf{T}$ (Eq. 4)
Map latent representations to joint latent space through fusion layer (Eq. 3)
Optimize fusion layer parameter ($\omega$) and AE decoder parameters ($\theta_{\mathbf{X}}$ and $\theta_{\mathbf{Y}}$) to minimize $\mathcal{L}_{\text{fusion}}$   (Eq. 5)

( Step III: cellular dynamics )
Optimize the entire model to minimize $\mathcal{L}_{\text{dyn}}$   (Eq. 10)

Output: modality integration $\tilde{\mathbf{Z}}_{\text{RNA}} = \text{Fus}(\mathbf{Z}_{\text{RNA}}, \omega)$ and $\tilde{\mathbf{Z}}_{\text{ATAC}} = \text{Fus}(\mathbf{Z}_{\text{ATAC}}, \omega)$
```

S3 Baseline Models

We compare scMultiNODE with six state-of-the-art unsupervised single-cell integration methods that are capable of aligning and computing the joint latent space.

- Seurat: The single-cell analysis platform Seurat [24] projects two datasets into a common space with linear canonical correlation analysis (CCA) that maximizes cross-dataset correlation. Seurat first identifies correspondence anchor points via CCA and then imputes one modality to another modality based on anchors. We use Seurat v5 in our experiments.

- SCOTv1: Demetci et al. [18] present the unsupervised learning method SCOT to align single-cell multi-modal datasets with Gromov-Wasserstein (GW) optimal transport. We term this model as SCOTv1 in this paper. We use the SCOTv1 implementation on https://github.com/rsinghlab/SCOT.
- SCOTv2: The SCOTv2 [17] model improves upon SCOTv1 by using unbalanced GW optimal transport to deal with disproportionate cell type representation and differing numbers of cells across single-cell modalities. We use the SCOTv2 implementation on https://github.com/rsinghlab/SCOT.
- UnionCom: Cao et al. [6] propose UnionCom, another unsupervised multi-modal integration model. It matches two datasets based on geometrical matrix matching. Specifically, UnionCom computes intra-modality distance matrices and then matches the modalities based on a matrix optimization problem. We use the UnionCom implementation on https://github.com/caokai1073/UnionCom. Apart from the hyperparameters listed in the Supplementary Table S1, we set all its other hyperparameters as default.
- Pamona: The Pamona [8] method adopts partial GW optimal transport to integrate multi-modal single-cell datasets. It aims to obtain shared and dataset-specific cell variations across modalities. We use the Pamona implementation on https://github.com/caokai1073/Pamona. We set the number of shared cells between datasets as the minimal number of cells across all modalities.
- uniPort: Cao et al. [7] introduce uniPort, incorporating coupled variational auto-encoders and mini-batch unbalanced optimal transport to integrate multi-modal single-cell datasets. We use its implementation on https://github.com/caokai1073/uniPort in our experiments. We use the diagonal integration mode for uniPort.

S4 Evaluation Metrics

We evaluate each model's integrated latent representations from three perspectives: modality integration, cell type variation, and cellular dynamic. Therefore, we adopt the following evaluation metrics.

- Batch entropy: We use it to evaluate the integration of unaligned datasets (DR and MN). Batch entropy is originally derived in Xiong et al. [51] and previously adopted by Cao et al.[7]. It evaluates the sum of regional mixing entropies between different datasets where a high score indicates cells from different modalities are mixed well. Specifically,

$$\text{batch entropy} = \sum_{k \in \{\text{RNA,ATAC}\}} p'_k \log\left(p'_k\right) \qquad \text{with} \qquad p'_k = \frac{p_i/P_i}{\sum_{j \in \{\text{RNA,ATAC}\}} p_j/P_j}, \tag{S4}$$

  in which P_k is the proportion of cells in each modality, and p_k is the proportion of cells from modality k in a given region.
- FOSCTTM: For the co-assay datasets (HC and HO) where one-to-one cell correspondence information is available, we further use the fraction of samples closer than the true match (FOSCTTM) [32,18]. Specifically, in the joint latent space, for every cell x in one modality, we find its nearest neighbor from the other domain. Then, we rank all cells in the joint latent space with respect to their distance to x. We compute the fraction of cells that are closer than the true nearest neighbor. Averaging this fraction across all cells obtains the FOSCTTM score. A perfect integration implies that all cells should be closest to their true match, resulting in a FOSCTTM of zero. Therefore, a lower FOSCTTM value indicates better integration performance.
- Neighborhood overlap: The neighborhood overlap [6] computes the ratio of cells that can find their correspondence cells from the other modality in their neighborhood. We use the averaged ratio of neighborhood overlap of the two modalities. The neighborhood overlap ranges from 0 to 1, and a higher value implies a better recovery of cell-to-cell correspondence between the two modalities. Neighborhood overlap is only available for co-assay datasets as it requires true cell correspondence.
- Spearman Correlation Coefficient (SCC): Based on the intuitive assumption that matched cells should have similar latent representations in the joint latent space, we use SCC to evaluate representation similarities between matched cells in the joint latent space, such that a better integration leads to a higher SCC value. SCC is also only available for co-assay datasets.

- Label transfer accuracy: We evaluate integration using cell-type labels through label transfer accuracy (`LTA-type`) as in previous studies [6,7,17]. This metric assesses the clustering of cell types after integration by training a k-nearest neighbor (kNN) classifier on joint latent representations of one modality and then evaluates its predictive accuracy on another modality. It ranges from 0 to 1, and a higher metric value indicates better integration performance as cells that belong to the same cell type are aligned close together. Additionally, as the main objective of our research, we evaluate how well the integration captures the variations of different timepoints. Therefore, we similarly compute `LTA-time` using timepoint labels, such that a higher `LTA-time` indicates better integration performance as cells that belong to the same timepoint are aligned close together.
- `Time correlation`: We hypothesize that good latent representations, if they retain the developmental dynamics, should highly correlate with the timepoint label. Therefore, we define the `time correlation`, which computes the distance correlation [45] between cell representations in the joint latent space and their corresponding timepoint labels. The distance correlation measures linear and nonlinear association between two datasets of arbitrary dimensions. Hence, the `time correlation` ranges from 0 to 1, where a higher value implies a better integration, which is highly associated with cellular dynamics. We use the `dcor` [42] to compute distance correlations.

S5 Hyperparameter Tuning

On co-assay datasets (HC and HO), we select corresponding hyperparameters for all methods (our and baselines) that yield the minimum `FOSCTTM` value; on unaligned datasets (DR and MN), we select hyperparameters that yield the maximum `LTA` based on common cell type labels. We use Optuna [2] to automatically determine the optimal hyperparameters and use sufficiently large hyperparameter ranges for search and evaluation. The hyperparameter search ranges of `scMultiNODE` and baselines are listed in Supplementary Table S1. We set the joint latent space dimension as 50 for all methods. We use the first-order Euler ODE solver and set ODE step size $\Delta t = 0.1$ in `scMultiNODE` . We run each method for sufficient iterations to ensure they converge.

S6 More on analysis of cell state transition

We use HC data to test `scMultiNODE` in downstream analysis. The cell path is constructed with the least action path (LAP), which has been used in previous studies [41,40,49] to construct cell fate transitions. The LAP method aims to find the optimal path between two cell states while minimizing their action and transition time. With a little abuse of notation, we let $\mathbf{X}$ denote cell representations in the joint latent space. Specifically, given starting point $\mathbf{X}_0$ and end point $\mathbf{X}_K$, LAP fins a path discretized as a sequence of K points $\mathcal{P} = \{\mathbf{X}_0, \cdots, \mathbf{X}_K\}$. For each segment constrained between $\mathbf{X}_{k-1}$ and $\mathbf{X}_k$, its tangential velocity is defined as $\mathbf{V}_k = \frac{(\mathbf{X}_k - \mathbf{X}_{k-1})}{\Delta}$ where Δ is the timestep taken by cells from $\mathbf{X}_{k-1}$. Therefore, we define the action $\mathcal{S}$ along the path $\mathcal{P}$ as

$$\mathcal{S} = \frac{1}{2}\sum_{k=1}^{K}\Big(\mathbf{V}_k - \text{Drift}_\omega(\widetilde{\mathbf{X}}_k)\Big)^2 \Delta \quad \text{with } \widetilde{\mathbf{X}}_k = \frac{\mathbf{X}_{k-1} + \mathbf{X} - k}{2}. \tag{S6}$$

Here, LAP method aims to align the tangential velocity $\mathbf{V}_k$ with the differential velocity $\text{Drift}_\omega(\widetilde{\mathbf{X}}_k)$ learned by `scMultiNODE` , while having the least transition time. Therefore, the optimal path is

$$\widehat{\mathcal{P}} \;=\; \underset{\mathcal{P},\Delta}{\text{argmin}}\,\mathcal{S} \;=\; \underset{\mathcal{P},\Delta}{\text{argmin}}\,\frac{1}{2}\sum_{k=1}^{K}\Big(\mathbf{V}_k - \text{Drift}(\widetilde{\mathbf{X}}_k,\omega)\Big)^2 \Delta. \tag{S6}$$

Solving Eq. S6 consists of two iterative steps

(1) Minimize action by fixing path $\mathcal{P}$ and varying the timestep Δ through

$$\widehat{\Delta} = \underset{\Delta}{\text{argmin}}\,\frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbf{X}_k - \mathbf{X}_{k-1}}{\Delta} - \text{Drift}(\widetilde{\mathbf{X}}_k,\omega)\right)^2 \Delta. \tag{S6}$$

(2) Minimize action by fixing timestep $\widehat{\Delta}$ and varying path $\mathcal{P}$

$$\widehat{\mathcal{P}} = \underset{\mathbf{X}_1,\cdots,\mathbf{X}_{K-1}}{\operatorname{argmin}} \frac{1}{2} \sum_{k=1}^{K} \left(\frac{\mathbf{X}_k - \mathbf{X}_{k-1}}{\widehat{\Delta}} - \text{Drift}(\widetilde{\mathbf{X}}_k, \omega) \right)^2 \Delta. \tag{S6}$$

The starting ($\mathbf{X}_0$) and end point ($\mathbf{X}_K$) are fixed in the optimization.

We use *scipy.optimize.minimize* to solve these two objective functions. In our experiments, we construct two paths from the first timepoint to two cell populations (oligodendrocyte and glutamatergic neuron) with $K = 8$. We set the starting point as the center of cells at the first timepoint ($t = 0$) and the endpoint as the center of the cell population. We initialize timestep as $\Delta = 1$ and $\mathcal{P}$ as equally spaced points from the starting to end points.

When finding the differentially expressed (DE) genes along the path, we augment the LAP path with its nearest neighbors. Specifically, assuming $\mathcal{P} = \{\mathbf{X}_0, \cdots, \mathbf{X}_K\}$ is the LAP path from $\mathbf{X}_0$ to $\mathbf{X}_K$, we have only eight cells on the path which is insufficient for DE detection. Therefore, for each $\mathbf{X}_k \in \mathcal{P}$, we find its nearest neighbors in order to augment the path. We use *sklearn.neighbors.NearestNeighbors* to search for ten nearest neighbors. Then, we can use *Scanpy* to detect DE genes for the augmented path with the Wilcoxon rank-sum test. We also use *Scanpy* to detect marker genes for each cell type solely from RNA or ATAC data.

S7 Supplementary Figures and Tables

Table S1: Hyperparameter search space of `scMultiNODE` and baseline methods during hyperparameter tuning.

Model	**Hyperparameters**
`scMultiNODE`	number of neighbors $kin\{5, 10, 25, 50, 75, 100\}$ coefficient $\alpha \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$ coefficient $\beta \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$
`SCOTv1`	number of neighbors $k \in \{5, 10, 25, 50, 75, 100\}$ entropic regularizer coefficient $e \in [0.001, 0.1]$ normalize $\in$ {True, False}
`SCOTv2`	number of neighbors $k \in \{5, 10, 25, 50, 75, 100\}$ entropic regularizer coefficient $eps \in [0.001, 0.1]$ marginal relaxation coefficient $\rho \in [0.001, 0.1]$ normalize $\in$ {True, False}
`UnionCom`	number of neighbors $k \in \{5, 10, 25, 50, 75, 100\}$ perplexity $\in \{10, 25, 50, 75, 100\}$ $\beta \in \{0.01, 0.1, 1.0, 10.0\}$
`Pamona`	number of neighbors $k \in \{5, 10, 25, 50, 75, 100\}$ regularization coefficient epsilon$\in [0.001, 0.1]$ trade-off coefficient Lambda $\in \{0.01, 0.1, 1.0, 10.0\}$
`uniPort`	KL coefficient $\in \{0.01, 0.1, 1.0, 10.0\}$ OT coefficient$\in \{0.01, 0.1, 1.0, 10.0\}$ entropy regularization coefficient $\in \{0.01, 0.1, 1.0, 10.0\}$ unbalanced OT parameter $\in \{0.01, 0.1, 1.0, 1.0\}$ iteration=10000, batch size=32, learning rate=0.0001, diagonal integration mode, MSE loss
`Seurat`	number of anchors for CCA $\in \{5, 25, 50, 75, 100, 125, 150, 175, 200\}$ number of neighbors when weighting anchors $\in \{5, 25, 50, 75, 100, 125, 150, 175, 200\}$ bandwidth of Gaussian kernel $\in \{0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0\}$ reference modality $\in$ {RNA, ATAC}

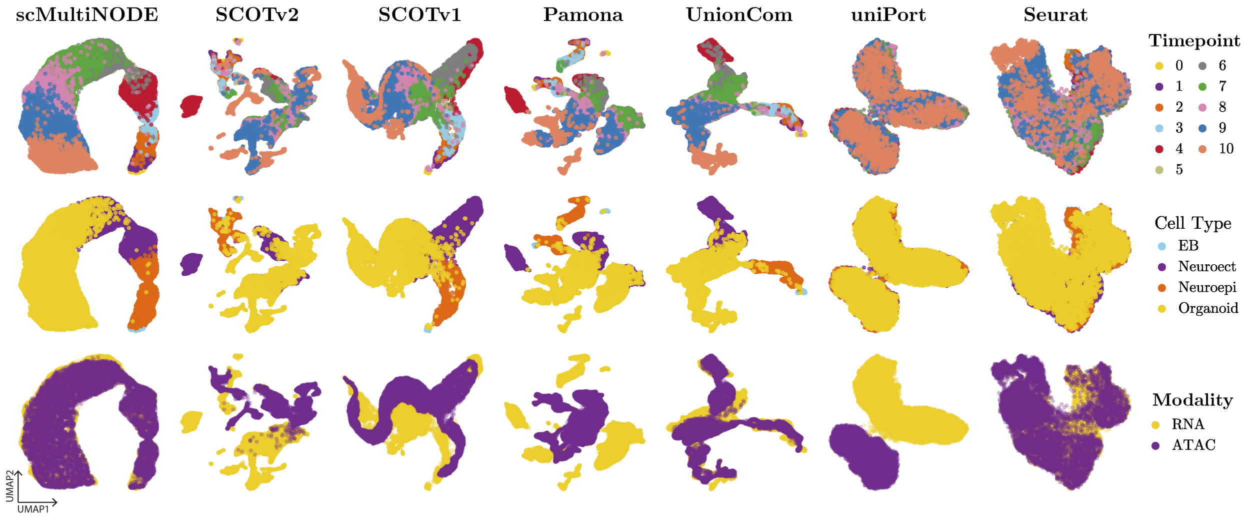

Fig. S1: 2D UMAP visualization of joint latent representations on the HO dataset. The representations are colored by timepoint labels, cell types, and modality.

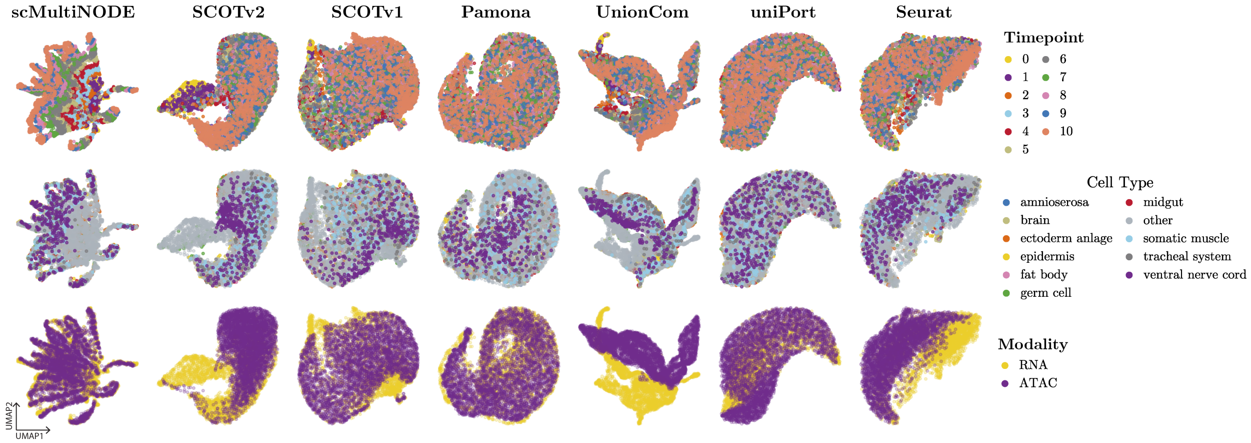

Fig. S2: 2D UMAP visualization of joint latent representations on the DR dataset.We only mark the common cell types of the two unaligned modalities.

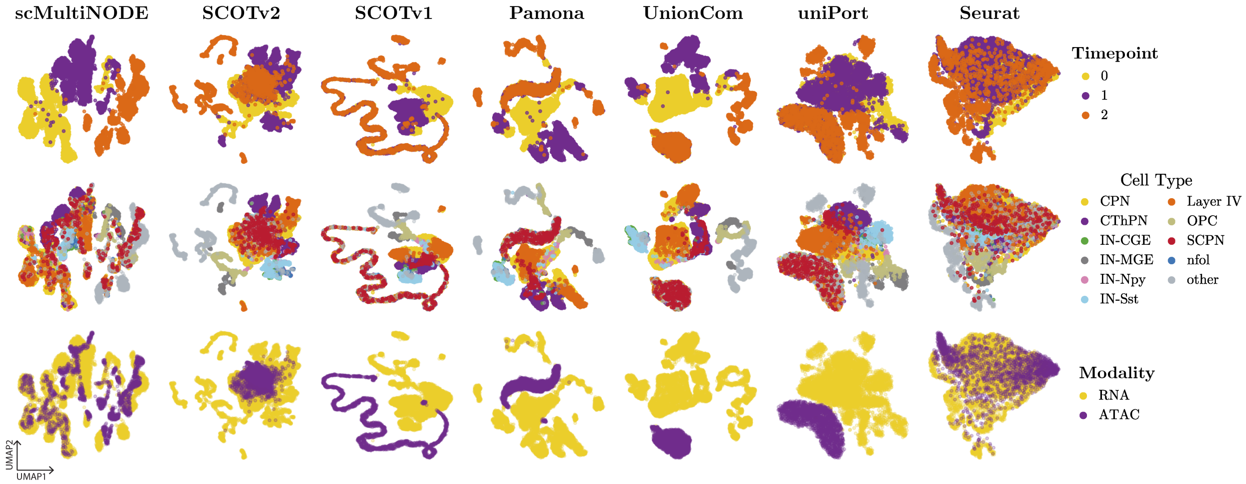

Fig. S3: 2D UMAP visualization of joint latent representations on the MN dataset. We only mark the common cell types of the two unaligned modalities.

Table S2: Evaluation of model integration on two co-assay datasets, HC and HO. The **red bold** and blue underlined numbers indicate the best and the second best performance, respectively.

Method (HC)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
scMultiNODE	**0.667**	**0.106**	0.203	0.884	0.392	**0.919**	**0.979**
SCOTv1	0.169	0.238	0.086	0.771	0.561	0.258	0.560
SCOTv2	0.531	0.170	**0.224**	**0.894**	**0.767**	0.380	0.513
Pamona	0.115	0.421	0.021	0.859	0.214	0.145	0.505
UnionCom	0.138	0.404	0.045	0.199	0.433	0.084	0.564
uniPort	0.278	0.418	0.062	0.098	0.308	0.165	0.125
Seurat	0.243	0.449	0.025	0.671	0.235	0.161	0.400

Method (HO)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
scMultiNODE	**0.521**	0.097	**0.0544**	**0.986**	**0.955**	**0.895**	**0.974**
SCOTv1	0.063	0.599	0.0019	0.824	0.112	0.061	0.807
SCOTv2	0.020	0.337	0.0069	0.409	0.571	0.198	0.748
Pamona	0.021	0.163	0.0291	0.761	0.848	0.550	0.802
UnionCom	0.366	**0.080**	0.094	0.881	0.947	0.773	0.931
uniPort	0.024	0.495	0.0042	0.007	0.449	0.126	0.138
Seurat	0.431	0.440	0.0061	0.900	0.327	0.116	0.604

Table S3: Evaluation of model integration on two unaligned datasets, DR and MN. The **red bold** and blue underlined numbers indicate the best and the second best performance, respectively. Because there is no cell-to-cell correspondence for unaligned datasets, we remove FOSCTTM, neighborhood overlap, and SCC, which require such information.

Method (DR)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
scMultiNODE	**0.614**	**0.314**	**0.430**	**0.777**
SCOTv1	0.096	0.297	0.094	0.443
SCOTv2	0.183	0.302	0.085	0.613
Pamona	0.534	0.279	0.116	0.303
UnionCom	0.145	0.212	0.107	0.478
uniPort	0.180	0.238	0.033	0.081
Seurat	0.188	0.074	0.025	0.381

Method (MN)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
scMultiNODE	**0.500**	0.148	**0.989**	**0.856**
SCOTv1	0.027	**0.311**	0.245	0.442
SCOTv2	0.161	0.285	0.336	0.243
Pamona	0.027	0.287	0.204	0.484
UnionCom	0.001	0.167	0.245	0.474
uniPort	0.095	0.310	0.259	0.099
Seurat	0.075	0.231	0.356	0.436

Table S4: The top 10 DE genes of oligodendrocyte (OL) and glutamatergic neuron (GN) paths, obtained from `scMultiNODE` 's joint latent space. We also show the top 10 marker genes of the OL and GN cell types, derived from RNA and ATAC data.

DE genes found in `scMultiNODE` joint latent space	OL path	SOX6, SLC1A3, NEAT1, ADGRV1, SLC1A2, PRKG1, GPC5, GLUL, SFMBT2, ATP1A2
	GN path	SV2B, ANO3, MTUS2, ZNF536, PDE8B, BMPER, ENSG00000251680, KIRREL3, FSTL5, SEC14L5, GRIN2A, CLMN
RNA-derived marker genes	OL cell type	CTNNA3, SLC24A2, ST18, RNF220, PLP1, PIP4K2A, MAP7, MBP, DOCK10, MOBP
	GN cell type	SATB2, RBFOX1, NRG1, ROBO2, RALYL, MIR137HG, KCNQ5, IQCJ-SCHIP1, DLGAP2, RYR2
ATAC-derived marker genes	OL cell type	RNF220, POLR2F, TFEB, FA2H, AATK, FAM102A, C10orf90, PRIMA1, CLMN, BCAR1
	GN cell type	RBFOX1, SATB2, NELL2, EFCAB6, MYT1L, MPPED1, NKAIN2, SV2B, ROBO2, SLC44A5

Table S5: `scMultiNODE` integration performance when using different joint latent space size d.

HC Dataset

Latent Size (d)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
10	0.575	0.324	0.083	0.921	0.316	0.434	0.674
50	0.643	0.108	0.183	0.923	0.415	0.803	0.944
100	0.647	0.094	0.232	0.853	0.441	0.864	0.965
150	0.683	0.127	0.178	0.863	0.375	0.885	0.959
200	0.677	0.131	0.152	0.841	0.360	0.886	0.972

DR Dataset

Latent Size (d)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
10	0.589	0.333	0.462	0.734
50	0.520	0.315	0.395	0.635
100	0.483	0.341	0.403	0.552
150	0.459	0.390	0.408	0.614
200	0.545	0.308	0.441	0.695

Table S6: `scMultiNODE` integration performance when using different number of neighbors (k) in GW optimal transport.

HC Dataset

Number of Neighbors (k)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
5	0.657	0.124	0.162	0.872	0.354	0.826	0.971
10	0.650	0.180	0.157	0.881	0.493	0.691	0.881
50	0.647	0.152	0.159	0.911	0.432	0.739	0.939
100	0.651	0.130	0.174	0.870	0.399	0.837	0.933
150	0.620	0.153	0.179	0.906	0.360	0.759	0.948
200	0.646	0.136	0.133	0.921	0.326	0.778	0.957
Best Baseline	0.531	0.170	0.224	0.894	0.767	0.380	0.564

DR Dataset

Number of Neighbors (k)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
5	0.591	0.436	0.571	0.779
10	0.613	0.372	0.430	0.777
50	0.598	0.335	0.587	0.771
100	0.615	0.350	0.537	0.760
150	0.534	0.399	0.544	0.768
200	0.611	0.324	0.571	0.766
Best Baseline	0.534	0.116	0.302	0.613

Table S7: `scMultiNODE` integration performance when using different fusion coefficient (α) in Eq. 5.

HC Dataset

Fusion Coefficient (α)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
0.0	0.624	0.259	0.113	0.883	0.369	0.451	0.811
0.01	0.659	0.222	0.132	0.841	0.379	0.629	0.891
0.1	0.657	0.115	0.170	0.911	0.400	0.788	0.965
1.0	0.659	0.144	0.153	0.903	0.277	0.879	0.965
10.0	0.658	0.097	0.196	0.939	0.394	0.894	0.976
100.0	0.640	0.329	0.075	0.754	0.216	0.662	0.816
Best Baseline	0.531	0.170	0.224	0.894	0.767	0.380	0.564

DR Dataset

Fusion Coefficient (α)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
0.0	0.422	0.397	0.280	0.477
0.01	0.385	0.548	0.548	0.724
0.1	0.444	0.575	0.575	0.774
1.0	0.394	0.517	0.517	0.753
10.0	0.385	0.505	0.505	0.719
100.0	0.363	0.601	0.601	0.782
Best Baseline	0.534	0.116	0.302	0.613

Table S8: `scMultiNODE` integration performance when using different dynamic regularization coefficient (β) in Eq. 10.

HC Dataset

Dynamic Regularization Coefficient (β)	**Modality Integration**				**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	FOSCTTM (↓)	Neighborhood Overlap (↑)	SCC (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
0.0	0.003	0.494	0.018	0.312	0.105	0.110	0.312
0.01	0.621	0.151	0.129	0.947	0.326	0.767	0.947
0.1	0.648	0.118	0.164	0.942	0.393	0.807	0.942
1.0	0.669	0.160	0.168	0.939	0.355	0.831	0.939
10.0	0.680	0.207	0.123	0.931	0.329	0.863	0.931
100.0	0.669	0.286	0.101	0.775	0.383	0.604	0.775
Best Baseline	0.531	0.170	0.224	0.894	0.767	0.380	0.564

DR Dataset

Dynamic Regularization Coefficient (β)	**Modality Integration**	**Cell-Type Variation**	**Cellular Dynamic**	
	Batch Entropy (↑)	LTA-type (↑)	LTA-time (↑)	Time Correlation (↑)
0.0	0.002	0.335	0.124	0.357
0.01	0.454	0.271	0.339	0.752
0.1	0.392	0.427	0.479	0.761
1.0	0.558	0.284	0.447	0.611
10.0	0.262	0.319	0.246	0.414
100.0	0.597	0.396	0.459	0.567
Best Baseline	0.534	0.116	0.302	0.613