scMultiNODE: Integrative and Scalable Framework for Multi-Modal Temporal Single-Cell Data

Jiaqi Zhang[1], Manav Chakravarthy[1], Ritambhara Singh[1,2*]

[1]Department of Computer Science, Brown University, 115 Waterman St., Providence, 02906, RI, United States.
[2]Center for Computational Molecular Biology, Brown University, 164 Angell St., Providence, 02912, RI, United States.

*Corresponding author(s). E-mail(s): ritambhara@brown.edu;
Contributing authors: jiaqi_zhang2@brown.edu;
manav_chakravarthy@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 jointly. We introduce scMultiNODE, an unsupervised integration model that combines gene expression and chromatin accessibility measurements in developing single cells, while preserving cell type variations and cellular dynamics. First, scMultiNODE uses a scalable, Quantized Gromov-Wasserstein optimal transport to align a large number of 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. Experiments on six real-world developmental single-cell datasets demonstrate that scMultiNODE can integrate temporally profiled multi-modal single-cell measurements more effectively than existing methods that focus on cell type variations and often overlook cellular dynamics. We also demonstrate that scMultiNODE's joint latent space facilitates several insightful downstream analyses of single-cell development, including the investigation of complex cell trajectories and the enabling of cross-modal label transfer. The data and code are publicly available at https://github.com/rsinghlab/scMultiNODE.

Keywords: single-cell development, multi-modal data integration, optimal transport, autoencoders, neural ordinary differential equations, temporal single-cell analysis

1 Introduction

Biological systems are inherently dynamic, constantly changing and adapting over time, and function at different scales, like organisms, cells, and molecules. Analyzing biological systems across scales enables a comprehensive understanding of mechanisms that govern various parts of a living system [1]. Specifically, dynamic processes at the cell level reveal how cells grow, divide, and differentiate into various cell types [2, 3]. Understanding these cellular dynamics enables us to enhance our knowledge of cell development and disease progression. Advances in single-cell measurements across time assist this study by capturing a high-resolution genomic snapshot of different cell states. Recent research increasingly involves temporally profiling multiple developmental single-cell measurements (or modalities), uncovering heterogeneity within the same tissue type and cellular dynamics across developmental stages [4–7]. For instance, Fleck et al. [8] collected multi-modal single-cell data, including gene expression and chromatin accessibility profiles, across a developmental time course of brain organoids, capturing key processes such as brain regionalization and neurogenesis. The emergence of these multi-modal and temporally resolved datasets has enabled more detailed investigations of cellular dynamics through advanced computational methods.

Several previous studies have focused on modeling cellular development using only one single-cell modality. For instance, Cicero [9] infers cell trajectories during myoblast differentiation using single-cell chromatin accessibility data. Moreover, a variety of recent methods have applied deep neural networks and differential equations to model developmental dynamics based on single-cell transcriptomic data alone [10–14]. While these approaches have provided valuable insights, they remain limited by the inherent noise and biases of the specific modality on which they rely, as well as the partial information each modality provides. For instance, gene expression measurements capture rich transcriptional information of diverse cell populations but are influenced by technical dropouts and mRNA degradation, which can obscure true biological signals and hinder accurate inference of dynamic processes [15]. On the other hand, chromatin accessibility measurement offers a view of the regulatory landscape and potential for future gene activity, but it is often sparse, less cell-type-specific, and difficult to interpret on its own [16].

These limitations highlight the critical importance of jointly analyzing multi-modal single-cell datasets, as integration offers a more complete and mechanistic understanding of biological systems. By combining complementary modalities, researchers can uncover insights that are inaccessible through any single modality alone. For example, Zhu et al. [17] demonstrated that integrating single-cell transcriptomic and chromatin accessibility data revealed complex regulatory mechanisms driving cell composition changes during neuronal lineage commitment in the human cerebral cortex—insights that would be difficult to achieve from either data type in isolation. Similarly, Calderon et al. [18] used multi-modal data to connect gene expression with epigenetic regulation during fruit fly embryonic development, illustrating how one modality can enhance the interpretation of another. These examples underscore the value of integrative analysis for achieving a more holistic view of dynamic biological processes. Despite its importance, integrative analysis of temporal multi-modal single-cell datasets presents significant challenges. Single-cell sequencing is inherently

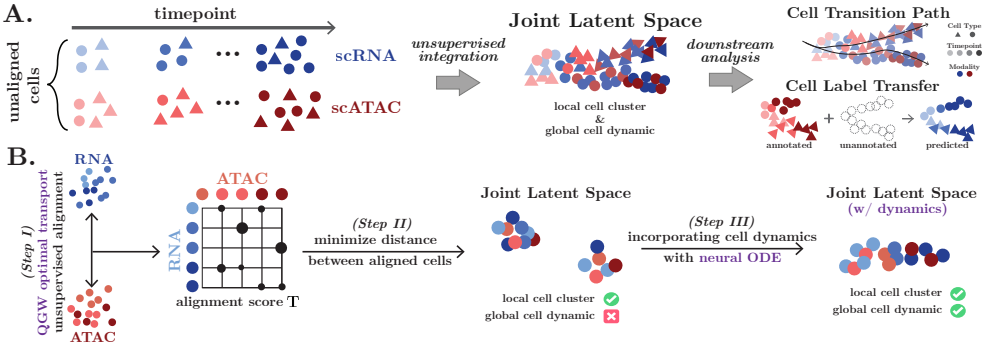

Fig. 1: **Model overview.** (**A**) Our goal is to integrate scRNA-seq and scATAC-seq data measured at multiple timepoints, when there is no cell-to-cell correspondence across modalities and timepoints. With the integration, we can conduct various downstream analyses, such as cell transition analysis and cell label transfer. (**B**) `scMultiNODE` first aligns multi-modal datasets with QGW optimal transport, which computes the cell correspondence between the two modalities with a large number of cells. `scMultiNODE` then constructs a joint latent space by mapping multiple modalities to a shared space, with the guidance of the computed correspondence where aligned cells stay together. Finally, `scMultiNODE` uses neural ODEs to model cellular dynamics and embed dynamics into the joint latent space.

destructive, meaning that each modality is typically measured in different cell populations, resulting in "unaligned" datasets. While co-assay technologies have emerged to simultaneously measure multiple modalities within the same cells (like gene expression and chromatin accessibility) [19, 20], these approaches are still resource-intensive and often limited by low sequencing depth [21, 22]. This restricts their scalability across developmental timepoints and complex tissues. As a result, the majority of temporally resolved multi-modal datasets remain unaligned across both modalities and time, posing additional computational and analytical challenges for dynamic modeling.

Building on the need for integrating unaligned multi-modal data, many unsupervised computational methods have been developed to address the lack of direct cell correspondence between modalities. These methods aim to project different data modalities into a shared representation space without requiring prior matching across cells. For example, [23] and [24] align different modalities by identifying correlated patterns or common biological signals across modalities. However, they all rely on linear methods, which limit their ability to model complex, nonlinear relationships between modalities [25, 26]. To address this, more recent approaches adopt manifold alignment techniques that better capture intricate cross-modality structures. These methods often rely on distribution matching [27] or optimal transport [28–31], enabling more flexible and accurate integration. These approaches have achieved strong performance in integrating heterogeneous single-cell datasets, particularly by aligning cell types across modalities.

However, a key limitation shared by all these methods is their inability to incorporate temporal information and dynamic cellular processes. When applied to

developmental single-cell multi-modal data, these methods fail to explicitly model how cells evolve over time. As a result, the integrated latent spaces do not naturally encode cellular dynamics [32], limiting their ability to resolve temporally dependent transitions and branching developmental trajectories. This is a critical methodological gap, as variations in cell types do not always mirror the progression of cell states during development [33, 34]. Integration methods that focus solely on capturing cell type heterogeneity risk distorting or misinterpreting dynamic processes such as lineage bifurcations or transient progenitor states, thereby limiting biological interpretation. Therefore, to fully leverage temporal multi-modal single-cell datasets, it is essential to develop integrative frameworks that incorporate dynamic modeling and capture the underlying developmental trajectories.

To address this gap, we propose single-cell Multi-modal Neural Ordinary Differential Equation (`scMultiNODE`), which integrates gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) profiles while explicitly modeling cellular dynamics over multiple timepoints. (Fig. 1). `scMultiNODE` is a novel *unsupervised* model that does not require prior cell matches across modalities or timepoints, as opposed to other multi-modal dynamic modeling approaches. Specifically, `scMultiNODE` aligns cells across modalities and timepoints using Quantized Gromov-Wasserstein (QGW) optimal transport [35, 36], which identifies biologically similar cells and scales to large-scale datasets. Guided by this alignment, `scMultiNODE` constructs a joint latent space and explicitly learns smooth developmental trajectories using neural ordinary differential equations (ODEs). As a result, `scMultiNODE` is the first method capable of preserving both local cell relationships (e.g., cell type distinctions) and global cellular dynamics (e.g., complex developmental trajectories) across multiple modalities during integration, a capability not achieved by previous multi-modal integration methods.

We evaluate `scMultiNODE` on six developmental single-cell datasets, comprising two co-assay datasets and four unaligned datasets, using scRNA-seq and scATAC-seq assays measured at multiple timepoints across 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 demonstrate that `scMultiNODE` produces an interpretable joint latent space that captures complex cell trajectories, supports the reconstruction of cell transition paths, and facilitates the discovery of development-related genes, which are key elements for studying cell differentiation. Finally, `scMultiNODE` enables accurate cross-modal label transfer, improving the interpretability and biological resolution of chromatin accessibility data. We envision that `scMultiNODE` will be helpful for integrative analyses of multi-modal temporal single-cell datasets, especially those with unaligned measurements.

2 Results

2.1 scMultiNODE Framework Overview

As illustrated in Fig. 1, scMultiNODE aligns scRNA-seq (i.e., gene expression data denoted by $\mathbf{X}^{(t)} \in \mathbb{R}^{m_t \times p}$ from m_t cells and p genes at timepoint t) and scATAC-seq (i.e., chromatin accessibility data denoted by $\mathbf{Y}^{(t)} \in \mathbb{R}^{n_t \times q}$ from n_t cells and q chromatin regions at timepoint t) datasets measured across observed timepoints $\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}} \subseteq \{0, 1, 2, \cdots\}$. Our goal is to learn a low-dimensional latent space that aligns the two modalities, captures cell type variations, and preserves cellular dynamics, thereby enabling downstream developmental analysis. For scRNA-seq data, we assume the same set of genes is measured; for scATAC-seq data, we assume the same set of chromatin regions is measured across different timepoints. Also, we do not require any prior known cell correspondence across modalities. Instead, we assume there is some underlying shared biological 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 datasets can be meaningfully integrated.

To find the correspondence of cells from different modality feature spaces, scMultiNODE uses GW optimal transport [35] that numerically computes the transport plan matrix T denoting cell alignments between gene expression and chromatin accessibility modalities. The exact computation of GW optimal transport 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) [36], to significantly speed up computations based on a divide-and-conquer strategy, where it partitions feature spaces into several blocks and matches blocks recursively. This correspondence matrix $\mathbf{T}$ enables scMultiNODE to map scRNA-seq and scATAC-seq data into a joint latent space by encouraging biologically similar cells, despite being measured in different modalities, to stay close together. This ensures the efficient computation of the latent space, which captures local cellular relationships, such as cell type clusters.

However, the joint space for now does not account for global cellular dynamics, such as developmental trajectories or lineage bifurcations. All previous integration methods lack this capability; hence, their latent representations poorly define natural cell paths across timepoints and hinder downstream analysis (as demonstrated in Sec. 2.3). To address the problem, scMultiNODE further utilizes neural ODEs to explicitly model global cellular dynamics and incorporates them into the joint latent space. scMultiNODE employs a dynamic regularization term to update the joint latent space, capturing both local cellular variations and global developmental dynamics. Originally proposed in our prior work [14], dynamic regularization treats neural ODEs as a temporal prior to encourage latent representations to follow underlying developmental trajectories, through minimizing discrepancies between the temporal prior and the latent representation. Here, we extend this concept to the multi-modal setting, allowing the learned dynamics to directly refine the joint latent space of scMultiNODE through dynamic regularization. As a result, the embedding of cellular dynamics improves upon previous integration models, which focus solely on cell type variations. This improvement enables scMultiNODE to fit the data more effectively, resulting in a joint latent

space that is both more robust and interpretable, as it captures both global cellular dynamics (e.g., long-term cell state transitions) and local cell relationships (e.g., cell type clusters). See Methods for more details about the `scMultiNODE` framework.

2.2 `scMultiNODE` captures cellular developmental dynamics during multi-modal integration

We evaluate `scMultiNODE` on six publicly available developmental single-cell datasets (see Methods and Supplementary Table S1) containing scRNA-seq and scATAC-seq measurements across multiple timepoints, covering various species and tissues: two co-assay datasets — human cortex (HC) and human organoid (HO); four unaligned datasets — drosophila embryogenesis (DR), mouse neocortex (MN), zebrahub (ZB), and amphioxus development (AM). For each dataset, we select the top 2000 highly variable features per modality and preprocess the data with normalization and batch effect correction. We compare `scMultiNODE` with six state-of-the-art unsupervised integration methods capable of learning joint latent spaces across modalities (Fig. 3A). To assess performance, we utilize a comprehensive set of metrics that evaluate performance from multiple perspectives: modality integration, cell label transfer, cell type preservation, and preservation of cellular dynamics (see Methods for detailed descriptions and computations of these metrics). For a fair comparison, all methods use the same joint latent space dimension of 50, with hyperparameters optimized via an exhaustive grid search over a sufficiently broad parameter range (see Methods).

Fig. 2 visualizes the joint latent representations of one co-assay (HC) and one unaligned (ZB) dataset in 2D Uniform Manifold Approximation and Projection (UMAP) [37] space. Integration results for the other datasets are shown in Supplementary Fig. S1-S2. 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 fail to align two modalities across the timepoints well (e.g., `Pamona` and `UnionCom`) or overlook the temporal structure (e.g., `SCOTv2`), with the cell type signal dominating the joint space. Furthermore, for unaligned datasets as well, `scMultiNODE` effectively integrates the two modalities (Fig. 2). Qualitative evaluation of all datasets indicates that `scMultiNODE` can effectively preserve cell type variations and cellular dynamics during integration, even in unaligned datasets. We also visualize the first two principal components of the integrations in Supplementary Fig. S3–S4 and obtain similar results.

Next, we quantitatively evaluate the integration performance of all the methods. Fig. 3 shows the performance from a set of evaluation metrics that capture multiple complementary aspects of integration quality. Detailed lists of all these metrics' values are included in Supplementary Tables S3-S4. For the integration of two modalities, `scMultiNODE` consistently outperforms the baselines for all datasets, including co-assay and unaligned ones (Fig. 3B and F). Even when `scMultiNODE` is not the best model in some metrics, its integration performs the second best and has comparable performance with the best one. For example, on HC dataset (co-assay data), `scMultiNODE` has the second best `neighborhood overlap`=0.20 and `SCC`=0.88, which achieves a similar performance than the best baseline `SCOTv2` with `neighborhood overlap`=0.22 and `SCC`=0.89. But `scMultiNODE` has the best performance on other

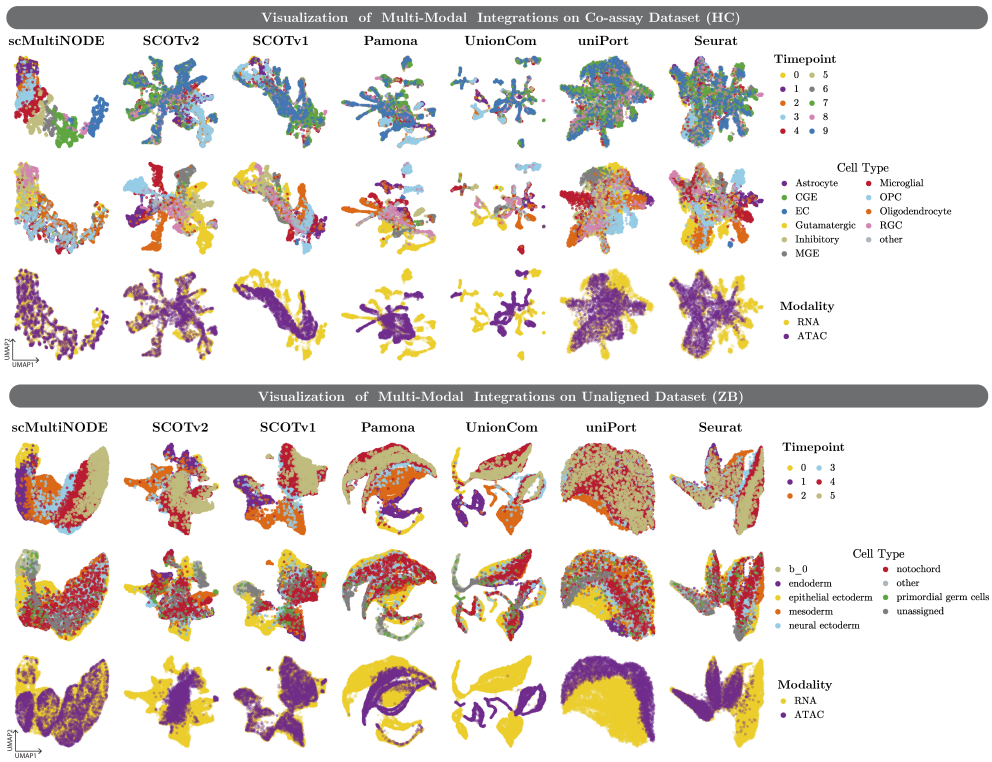

Fig. 2: **2D UMAP visualization** of joint latent representations on the co-assay (human cortex, HC) and unaligned (zebrahub, ZB) datasets. The representations are colored by timepoint labels (*top*), cell types (*middle*), and modalities (*bottom*). `scMultiNODE` effectively preserves cell type variations and cellular dynamics during integration.

metrics, including `batch entropy` (`scMultiNODE` has 0.67 and the best baseline obtains 0.53) and 1-`FOSCTTM` (`scMultiNODE` has 0.89 and the best baseline obtains 0.83). Additionally, on unaligned datasets, `scMultiNODE` clearly outperforms baselines regarding `batch entropy` on DR, MN, and AM datasets, with the second-best integration on ZB dataset (`scMultiNODE` `batch entropy`=0.14 and the best baseline `batch entropy`=0.17). Furthermore, `scMultiNODE` achieves the highest `time correlation` across all datasets (Fig. 3D and H), denoting that it can align the two modalities well while preserving the underlying dynamics.

In Fig. 3E and I, 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, DR, and AM datasets, `scMultiNODE` has the

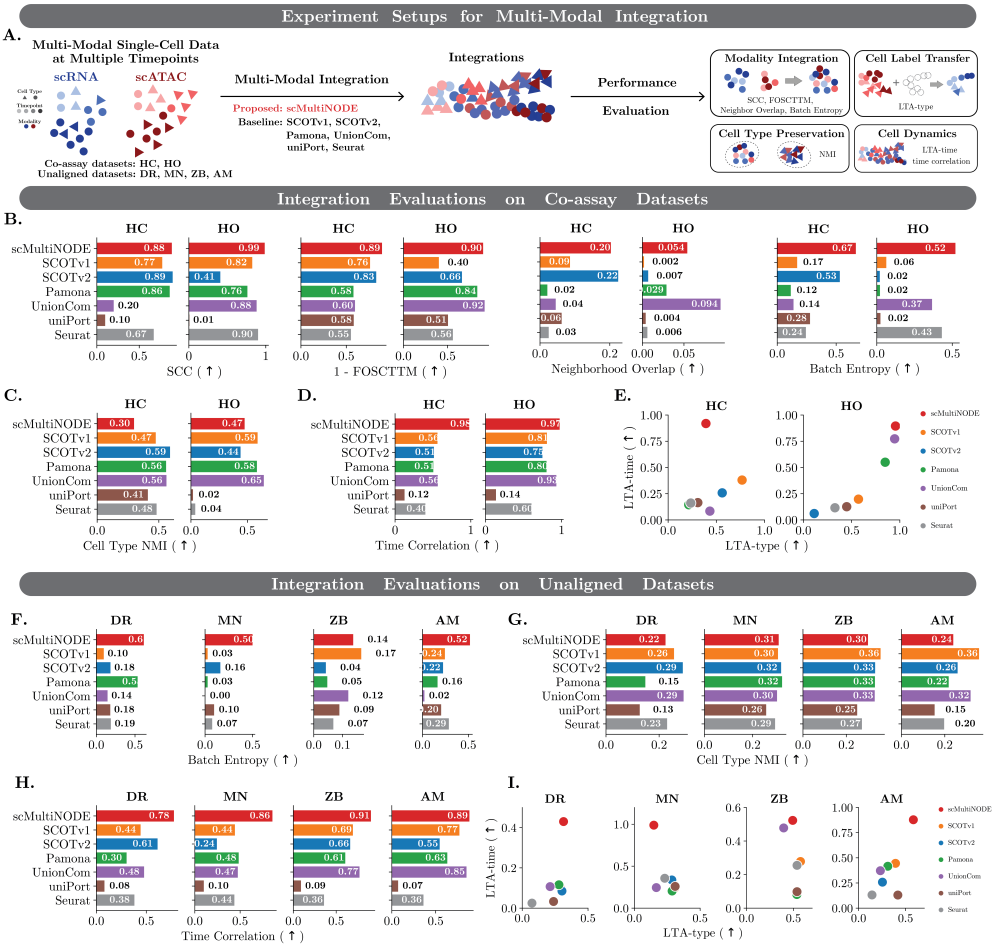

Fig. 3: **Integration evaluation** on two co-assay datasets: HC and HO; four unaligned datasets: DR, MN, ZB, and AM. (**A**) Schematic overview of integration performance comparison. (**B** and **F**) Performance of modality integration. `Batch entropy` evaluates the integration of all datasets. Three metrics that are only available on co-assay datasets: Spearman correlation coefficient (`SCC`), `neighborhood overlap`, and fraction of samples closer than the true match (`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. (**C** and **G**) Cell type conservation after integration. We calculate the Normalized Mutual Information (`NMI`) score, with higher values indicating better preservation of cell groups. (**D** and **H**) `Time correlation` evaluates integrations on capturing cell dynamics. (**E** and **I**) Label transfer accuracy (LTA) across modalities. 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. `scMultiNODE` outperforms existing approaches in integrating temporal multi-modal single-cell data, demonstrating strong performance on both co-assay and unaligned datasets while preserving cell type variation and underlying cellular dynamics.

highest `LTA-type` and `LTA-time` scores (Fig. 3E,I), denoting its good performance in capturing variations of both cell type and timepoints. On the HC, MN, and ZB datasets, `scMultiNODE` balances the trade-off between cell type and timepoint variations, reporting a median level of `LTA-type` score across all models and the highest `LTA-time`. Despite this trade-off, `scMultiNODE` obtains good integrations (Fig. 3B,F) for these three datasets with `batch entropy`=0.67 (HC), 0.50 (MN), and 0.14 (ZB).

Lastly, we validate whether the integration preserves distinct cell type separations (Fig. 3C and G). `scMultiNODE` achieves clustering performance comparable to existing baselines across most datasets, while simultaneously capturing dynamic trajectories. For instance, `scMultiNODE` attains `NMI`=0.47 on HO dataset (baseline median `NMI`=0.51) and `NMI`=0.24 on AM dataset (baseline median `NMI`=0.24). These results suggest that while perfect preservation of cell type clusters is difficult, `scMultiNODE` maintains a competitive level of clustering performance. Importantly, this result reflects an inherent trade-off in unsupervised integration that aligning cells across modalities requires balancing the preservation of fine-grained cell type identities with the need to capture global cellular dynamics. This trade-off may result in partial loss of modality-specific information. Nevertheless, `scMultiNODE` achieves a comparable level of clustering performance to existing baselines, while being significantly superior at modeling developmental dynamics, which are often overlooked in previous static integration approaches.

One intuitive approach to improve the preservation of cell type identities might be introducing cell type supervision. To explore this, we assess the effect of adding cell type supervision to `scMultiNODE` by introducing a classification head with a cross-entropy loss on top of the latent space (see Methods and Supplementary Table S11-S12). While this improves cell group separation, it compromises modality integration and the preservation of cellular dynamics (see Methods). The results suggest that supervision can distort the latent space and obscure continuous biological processes, possibly due to rigid class boundaries or inaccurate cell annotations. Therefore, we focus on unsupervised integration in this study to better capture dynamic and multimodal structures.

Overall, 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. Our findings emphasize the importance of evaluating multimodal integration from multiple perspectives, as demonstrated in our experiments, to ensure a thorough assessment.

Additionally, we performed ablation studies to assess the sensitivity of `scMultiNODE` to hyperparameters using both co-assay and unaligned datasets. We found that `scMultiNODE` is robust to the size of the latent space, while the dynamic regularization component is critical. Removing dynamic regularization significantly degrades integration quality and leads to a complete loss of the model's ability to capture developmental dynamics. We also evaluated runtime scalability on large-scale datasets, showing that `scMultiNODE` maintains comparable computational costs to

existing methods despite modeling cellular dynamics, making it practical for large-scale temporal multi-modal analysis. Full ablation results and runtime comparisons are provided in Supplementary Notes.

2.3 `scMultiNODE` 's latent space preserves multifurcating cell development trajectories

Previous works on multi-modal single-cell analysis have often relied on pseudotime inference within the latent space to analyze cell developmental trajectories [23, 38]. They align cells from different modalities and subsequently apply trajectory inference tools [39] on the integrated space to reconstruct cell lineage progressions. While these methods can recover coarse temporal orderings, they lack an explicit model of the underlying cell dynamics and assume a static representation of cellular states. As a result, they struggle to resolve complex or multifurcating trajectories. Here, we demonstrate that `scMultiNODE` can learn an interpretable joint latent space, which preserves complex cell developmental trajectories of multi-modal single-cell datasets.

We evaluated this capability using the HC dataset, which contains multifurcating trajectories in the developing human brain cortex (Fig. 4A). We assessed whether pseudotime estimated in the `scMultiNODE` latent space preserves the correct trajectory structure Fig. 4C), comparing it to pseudotime inferred from multi-modal integration baselines lacking dynamic modeling. We also compared `scMultiNODE` to single-modality models, including a static autoencoder (AE) and a dynamic model, scNODE [14], which were trained separately on RNA and ATAC data. This comparison enabled us to test whether integrating modalities enhances trajectory inference on noisy single-modal data, such as scATAC-seq. Specifically, in each latent space, we applied Monocle3 [39] to infer cell pseudotime. We quantified pseudotime accuracy using the Spearman rank correlation coefficient (ρ) between the predicted pseudotime and the true cell progression along known trajectories. A higher correlation coefficient indicates better alignment with biological progression.

As shown in Fig. 4B, `scMultiNODE` and `scNODE` (on RNA) achieve higher correlation coefficients across all trajectory branches, indicating they more accurately preserve the global trajectory structure as they all explicitly model cellular dynamics in the latent space. On the contrary, static multi-modal integration baselines fail to resolve branching structures and often produce collapsed or misaligned paths. For example, `scMultiNODE` reaches $\rho = 0.84$ for the EN-related trajectory, whereas the best integration baseline achieves only $\rho = 0.47$. Notably, for the Oligendendrocyte-related trajectory, `scMultiNODE` achieves $\rho = 0.55$, comparable to SCOTv1 ($\rho = 0.60$) but outperforming all other integration baselines. Compared to single-modality models, `scMultiNODE` outperforms AE and scNODE on ATAC data, where dynamic inference is more difficult. For instance, AE yields poor ordering with ρ as low as 0.14, while scNODE on ATAC data even produces negative correlations in some branches, indicating misdirection in pseudotime prediction. We also applied partition-based graph abstraction (PAGA) [40], another state-of-the-art method that infers pseudotime by modeling the topology of cell transitions through a connectivity graph. Results with this additional method further support that `scMultiNODE` 's integration preserves meaningful developmental trajectories (see Methods and Supplementary Fig. S6).

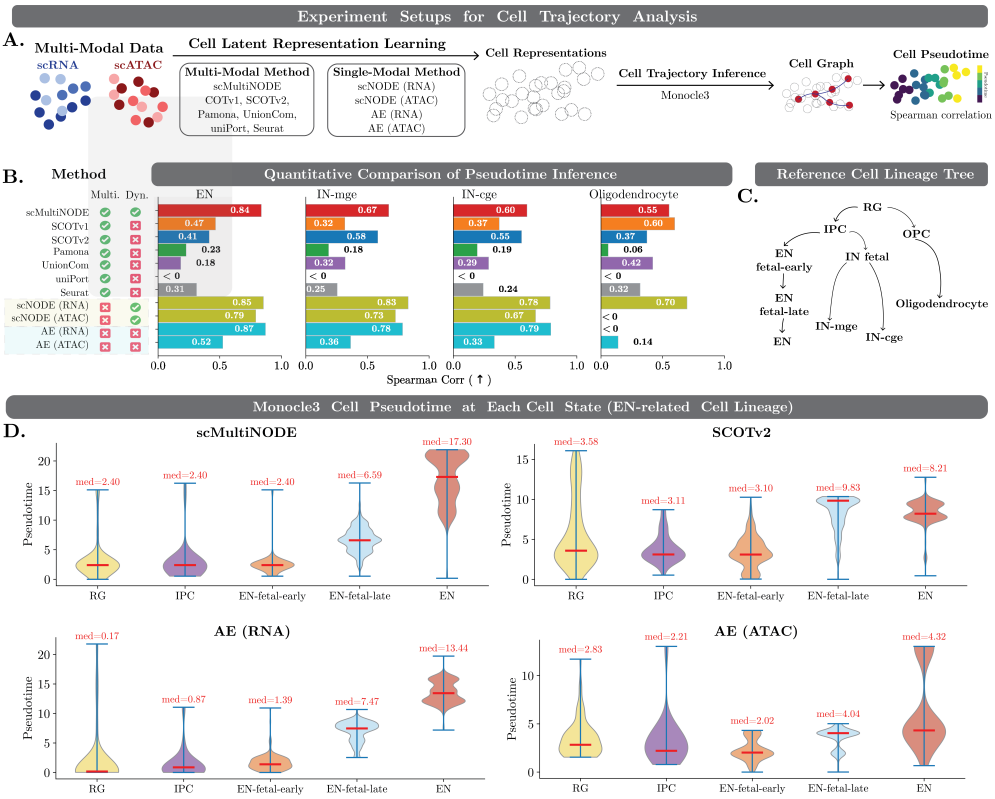

Fig. 4: **Comparison of cell trajectory pseudotime inference.** (**A**) Schematic overview of the cell trajectory analysis task. (**B**) Spearman correlation coefficient (ρ) between ground truth cell state orders and Monocle3 pseudotime of the integration "Multi." indicates multi-modal integration method, "Dyn." denotes models incorporating cell dynamics). (**C**) Reference cell state transition graph for the developing human brain cortex. The trajectory originates from RG and branches into four lineages (RG: Radial Glia, IPC: Intermediate Progenitor Cells, OPC: Oligodendrocyte Precursor Cells, EN: Excitatory Neurons, IN: Inhibitory Neurons). (**D**) Violin plots showing Monocle3 pseudotime distributions inferred from the integrated latent space of `scMultiNODE` and `SCOTv2`, as well as single-modality AE latent spaces. Cells are grouped by annotated cell states, and median pseudotime values ("med") are indicated for each group. We expect the pseudotime to exhibit a monotonically increasing trend from left to right, reflecting the correct progression of cell state transitions along the lineage trajectory from top to bottom. Results imply that `scMultiNODE` 's integration of gene expression and chromatin accessibility yields biologically meaningful pseudotime trajectories across complex, branching lineages.

These results highlight the importance of explicitly modeling global cell dynamics and jointly integrating multiple modalities in `scMultiNODE` for accurate trajectory inference, especially when one modality lacks strong dynamic signals. `scMultiNODE` 's integration of gene expression and chromatin accessibility data produces consistent and biologically meaningful pseudotime across complex and divergent lineages (Fig. 4B), offering a more robust framework for studying cellular developments.

2.4 `scMultiNODE` 's latent space assists with understanding development-related genes

Building on our earlier results with the HC dataset, we next assess whether `scMultiNODE` 's joint latent space can reconstruct meaningful developmental paths and uncover underlying gene programs. Specifically, we investigate whether smooth cellular trajectories can be extracted and linked to dynamic gene expression, moving beyond static embeddings to a more functional understanding of brain cortex development.

We train `scMultiNODE` on the HC dataset, given its complex trajectory structure, and project all cells into the joint latent space. To trace developmental progressions, we apply the Least Action Path (LAP) method [5, 41, 42] to compute optimal cell paths between the starting timepoint ($t = 0$) and mature cell types (see Methods). As shown in Fig. 5A, we identify distinct trajectories to oligodendrocytes (OL, purple) and glutamatergic neurons (GN, green).

Along each trajectory, we apply the Wilcoxon rank-sum test to identify differentially expressed (DE) genes that drive these transitions. Fig. 5B–C highlight top DE genes: SV2B and ANO3 (GN path), and SOX6 and SLC1A3 (OL path). We plot their z-score-normalized expression over time for cells on the path (solid lines) and off-path (dotted lines), alongside control genes. Notably, DE genes exhibit strong trajectory-associated variation, unlike randomly selected controls.

These identified DE genes are well-supported by prior literature. Previous studies [43, 44] have found that SV2B transcript is expressed in glutamatergic neurons. The Human Protein Atlas [45] 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 [46]. Also, as The Human Protein Atlas shows, SLC1A3 is enhanced in oligodendrocytes.

Finally, we compare the development-related genes identified from our joint latent space analysis with conventional cell type marker genes derived from a separate scRNA-seq/scATAC-seq pipeline (Supplementary Table S5). Notably, the RNA-based markers for glutamatergic neurons do not include SV2B or ANO3, two key DE genes identified along the GN trajectory , despite their known roles in neuron development. Similarly, SOX6 and SLC1A3, which drive the OL trajectory, are missing from RNA-derived oligodendrocyte marker lists. These discrepancies suggest that explicitly modeling cellular dynamics enables the discovery of functionally important genes that static marker-based approaches may overlook. Interestingly, SV2B does appear as a marker in the ATAC modality, reinforcing the requirement of multi-modal integration for capturing complementary biological signals.

In summary, `scMultiNODE` provides an interpretable joint latent space that enables not only integration, but also the discovery of development-related, biologically meaningful genes by leveraging complementary information from both modalities.

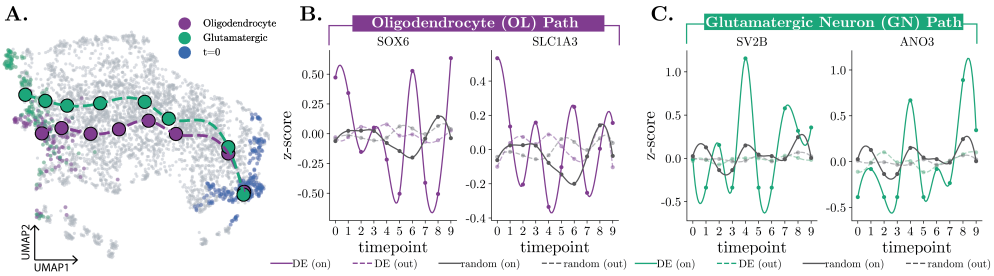

Fig. 5: **Cell path construction** on the human cortex (HC) dataset. (**A**) 2D UMAP visualization of the least action path from cells at the starting point ($t = 0$) to the oligodendrocyte (OL) and glutamatergic neuron (GN) populations, respectively. (**B**, **C**) Gene expression z-score values of 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). Overall, `scMultiNODE` provides an interpretable joint latent space that facilitates cell path construction and identification of development-related genes.

2.5 `scMultiNODE` enables germ layer label transfer across modalities

Germ layers are foundational tissue structures established during early embryogenesis, giving rise to distinct organs and cell types [47]. Accurately identifying germ layer origin is crucial for understanding developmental biology and tracing differentiation pathways [48, 49]. While scRNA-seq data often provides rich transcriptional signals to support such annotations, assigning germ layer labels to scATAC-seq data is far more challenging due to its sparsity and lower specificity [16, 50]. To address this, we leverage `scMultiNODE` 's joint latent space to transfer germ layer annotations from RNA to ATAC, enhancing the interpretability of scATAC-seq data when direct annotation is not feasible.

We evaluated this capability on the unaligned drosophila embryogenesis (DR) data, which contains germ layer labels only for RNA. Using each model's joint latent space, we trained a random forest classifier [51] on RNA cell embeddings and their germ layer labels, then predicted ATAC labels within the same space (Fig. 6A). As shown in Fig. 6B, `scMultiNODE` produced coherent ATAC clusters by predicted germ layer identity that aligned with known developmental trajectories: a transition from broad ectodermal states toward neuroectodermal subpopulations. This suggests that the integration captures meaningful biological progression.

To validate predictions, we analyzed germ layer marker gene expression across predicted ATAC groups. Marker genes were derived from differential expression analysis on RNA cells, and their expression was visualized on the ATAC-derived gene activity matrix (see Methods). `scMultiNODE` clearly separated predicted germ layers by marker gene expression (Fig. 6C). In contrast, baseline integration methods exhibited poorly resolved expression patterns, indicating ineffective alignment and poor label transfer performance.

Lastly, we performed Gene Ontology (GO) enrichment on the predicted ATAC groups to assess functional consistency and relevance of the label predictions [52]. The enriched GO terms reveal distinct functional profiles for each germ layer group, aligning well with the known biological roles of their respective germ layers (Fig. 6D). For instance, the ectoderm group was significantly enriched for terms related to nervous system and sensory organ development [53, 54], supporting the validity of the label transfer (Fig. 6D).

Together, these results highlight `scMultiNODE` 's ability to preserve developmental structure across modalities and enable accurate cross-modal label transfer. These make `scMultiNODE` a powerful tool for annotating single-cell data in developmental contexts, when direct labels are unavailable. By aligning multi-modal measurements, `scMultiNODE` enhances the biological resolution of chromatin accessibility data and supports downstream developmental analyses.

3 Discussion

We present `scMultiNODE` , a novel unsupervised method for integrating multi-modal temporal single-cell datasets. Our model can align temporally resolved multi-modal 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 a range of critical downstream analyses, including complex cell trajectory inference, cell transition path construction, investigation of genes that change along this path, and cross-modal label transfer, all of which are crucial for understanding developmental and differentiation processes in single-cell biology.

To achieve dynamics-aware integration, `scMultiNODE` uses neural ODE and dynamic regularization. This regularization term incorporates the continuous nature of cellular development into the joint latent space, enhancing the biological fidelity of the representation. Our experiments on real-world multi-modal single-cell datasets demonstrate that `scMultiNODE` consistently outperforms state-of-the-art integration methods. Unlike static integration models that struggle to capture cell trajectories, `scMultiNODE` accurately recovers complex developmental trajectories and improves pseudotime inference. Moreover, it enables accurate label transfer across modalities, improving the interpretability and biological resolution of chromatin accessibility data.

While we have focused on integrating gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) measurements, the `scMultiNODE` framework is designed to be general and flexible. It imposes no architectural constraints on the encoders or

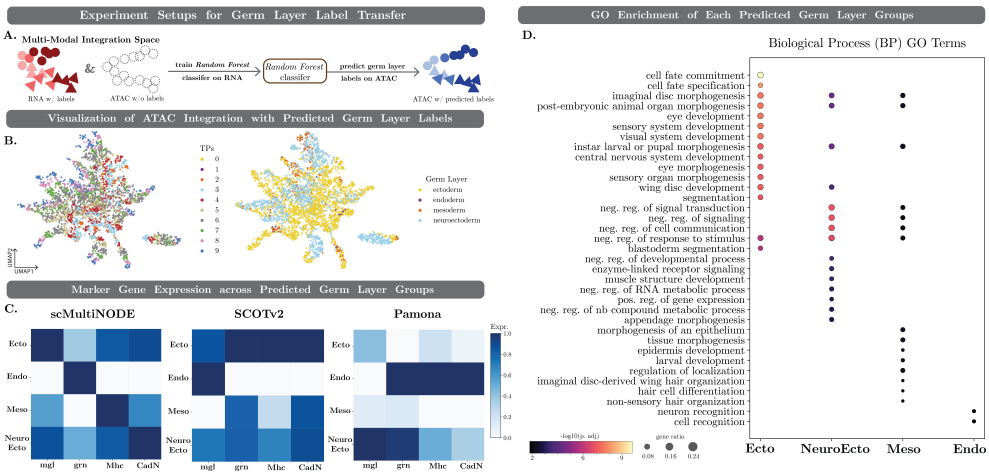

Fig. 6: **Validation of germ layer label transfer.** (**A**) Schematic overview of the label transfer task. (**B**) 2D UMAP visualization of ATAC cells from `scMultiNODE` integrations, with cells colored by predicted germ layer labels. (**C**) Expression levels of marker genes (mgl, grn, Mhc, and CadN) across the four predicted germ layer groups (Ecto: ectoderm, Endo: endoderm, Meso: mesoderm, Neuro Ecto: neuroestoderm). Gene expression values are scaled across all groups. We expect to observe a diagonal enrichment pattern, where marker genes are highly expressed in their corresponding cell groups. (**D**) Gene Ontology (GO) enrichment analysis of the four predicted germ layer groups focusing on the Biological Process (BP) category. These results demonstrate `scMultiNODE` 's effectiveness in achieving accurate cross-modal label transfer.

decoders and can be extended to any combination of single-cell modalities, including protein or methylation data. Moreover, our ablation study (Supplementary Notes) finds `scMultiNODE` can still maintain strong alignment performance even without the fusion loss term, highlighting the robustness of the dynamic regularization. This modular design makes `scMultiNODE` broadly applicable to a wide range of integration tasks.

A key challenge we identify in unsupervised multi-modal integration is balancing the preservation of all information from each modality. Different modalities capture complementary but uneven aspects of biology. For example, scRNA-seq is typically rich in cell state variation, while scATAC-seq provides sparse regulatory information with limited dynamic resolution. Moreover, unaligned datasets often exhibit varying cell type proportions across modalities, introducing additional complexity compared to co-assay datasets. When forcing alignment without supervision, there is an essential trade-off between retaining global cell dynamics and preserving fine-grained cell state identities. This trade-off can result in partial information loss, particularly from the more informative modality. This highlights the necessity of designing integration strategies that remain sensitive to various modality-specific information. In our work, `scMultiNODE` achieves superior unsupervised integration and comparable cell identity

preservation across both co-assay and unaligned datasets, highlighting its efficiency in addressing these challenges.

Another line of research leverages RNA velocity methods [55–57], which infer the temporal dynamics of gene expression by analyzing spliced/unspliced RNA counts. Recent extensions have adapted this framework to multi-modal datasets [58]. However, several limitations hinder the reliability of RNA velocity in trajectory inference, particularly in datasets with complex, nonlinear, or branching dynamics [56, 59]. A key limitation is the discrepancy in timescales. RNA splicing occurs in the order of minutes to hours [60], while developmental processes can span days to weeks, making it difficult for RNA velocity to capture long-term cellular dynamics. Moreover, the accuracy of velocity estimates is highly sensitive to technical noise and preprocessing of spliced/unspliced counts. Due to these limitations, we do not include RNA velocity-based methods in our comparisons. Nevertheless, they remain valuable as complementary tools to `scMultiNODE` for exploring local dynamics.

For future work, we will incorporate structured biological priors, such as lineage constraints, cell proliferation, or known gene regulation, to further enhance the biological interpretability of integrations. We also plan to explore the use of more expressive architectures, such as Transformers [61], to better model long-range dependencies and complex relationships within each modality. We will also test our model on other single-cell modalities (e.g., surface protein) or more than two modalities, broadening its applicability in integrative single-cell analysis.

Overall, `scMultiNODE` provides the community with a powerful and generalizable framework for understanding cellular systems at innovative resolution. It is particularly valuable in developmental biology and disease progression studies, where cellular heterogeneity and temporal evolution play critical roles. As multi-modal and temporally resolved datasets become increasingly available, `scMultiNODE` offers a scalable and biologically grounded approach for unlocking their full potential and generating new insights into complex biological processes.

4 Methods

4.1 `scMultiNODE` framework

`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 $\mathcal{T}_{\text{RNA}}, \mathcal{T}_{\text{ATAC}} \subseteq \{0, 1, 2, \cdots\}$, we aim to learn a unified low-dimensional representation that brings together multiple modalities, reflects cell type variations, and preserves dynamic cellular progression for downstream developmental analysis. We assume consistent feature sets across timepoints: the same genes for scRNA-seq and the same chromatin regions for scATAC-seq. While no prior cell-to-cell correspondence is needed, we assume the datasets share common biological signals, such as overlapping

cell types and similar developmental trajectories, making joint integration both meaningful and feasible. Overall, the `scMultiNODE` framework has two important steps: (1) integration of multi-modal single-cell datasets while preserving local cell relationships, and (2) incorporating global cell dynamics.

4.1.1 Integration of multi-modal single-cell datasets while preserving local cell relationships

`scMultiNODE` uses the *Quantized Gromov-Wasserstein (QGW) Optimal Transport* and autoencoders (AEs) to integrate large multi-modal single-cell datasets.

Gromov-Wasserstein (GW) Optimal Transport for Multi-Modality Alignment

In this multi-modal integration problem, our goal is to align two single-cell datasets that are measured separately and represent two distinct feature spaces. To achieve this, we need an approach to find the correspondence of cells from one modality to another. However, given the different feature spaces, it is infeasible to directly compute the cell distances between two data modalities and align the nearby cells. To overcome this, we use GW optimal transport, which recent works [28–30] have leveraged to align multi-modal single-cell measurements in an unsupervised manner.

GW optimal transport aims at moving data points from one feature space to another while preserving the original local geometry that is captured using intra-modality distances [35]. The central concept of GW is to find the best data correspondence matrix that denotes the probability of alignment between each data point across two data modalities. Specifically, assuming we have two sets of data points X and Y with corresponding probability measures μ_X and μ_Y. The GW optimal transport finds the probabilistic transport plan π that minimizes the transport cost (i.e., the GW distance) as

$$\text{GW distance}\,(X,Y) = \inf_{\pi\in\Pi(\mu_X,\mu_Y)} \int_{X\times Y}\int_{X\times Y} L\left(x,y,x',y'\right)\ \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y'), \quad (1)$$

where $L\left(x,y,x',y'\right) = ||d_X(x,x') - d_Y(y,y')||^2$ is the loss function, d_X and d_Y are intra-domain distances for X and Y. $\Pi(\mu_X,\mu_Y)$ is the set of all probabilistic correspondences on $X\times Y$ with μ_X and μ_Y as marginals. Therefore, d_X/d_Y calculates distances between pairs of data points ((x,x') or (y,y')) within each dataset, and the loss function L measures how these pair distances in one dataset compare to those in the other dataset. So, the GW distance does not directly compare the data points across different feature spaces, but instead finds a matching cross-space sample that best preserves the intra-modality distances. This ensures that the transport plan π can preserve local geometry, like cell clusters and cell-cell neighboring relationships, in one modality and correspond to similar local structures in the other when we align two single-cell modalities.

`scMultiNODE` *preserves local cell relationships when integrating single-cell datasets*

`scMultiNODE` begins by aligning the gene expression and chromatin accessibility modalities using GW optimal transport. Since we have finite cell samples of both modalities, we can numerically find the transport plan matrix T between scRNA and scATAC datasets through minimizing the **discrete GW distance** as follows:

$$\mathrm{GW}(\mathbf{X}_{\mathrm{ALL}}, \mathbf{Y}_{\mathrm{ALL}}) = \min_{T \in \Pi(\nu_X, \nu_Y)} \sum_{i,i',j,j'} \| \mathbf{D}^{\mathrm{RNA}}_{ii'} - \mathbf{D}^{\mathrm{ATAC}}_{jj'} \|^2 \, \mathbf{T}_{ij} \mathbf{T}_{i'j'}. \tag{2}$$

Here, $\mathbf{X}_{\mathrm{ALL}} = \mathrm{CONCAT}(\mathbf{X}^{(t)} | t \in \mathcal{T}_{\mathrm{RNA}})$ and $\mathbf{Y}_{\mathrm{ALL}} = \mathrm{CONCAT}(\mathbf{Y}^{(t)} | t \in \mathcal{T}_{\mathrm{ATAC}})$ are concatenations of gene expression and chromatain accessibility of all cells across available timepoints. ν_X and ν_Y are empirical distributions of $\mathbf{X}_{\mathrm{ALL}}$ and $\mathbf{Y}_{\mathrm{ALL}}$, respectively. $\mathbf{D}^{\mathrm{RNA}}$ and $\mathbf{D}^{\mathrm{ATAC}}$ are intra-modality distance matrices. In the cell correspondence matrix $\mathbf{T}$, $\mathbf{T}_{ij}$ indicates the alignment probability of RNA cell i and ATAC cell j.

Eq. 2 is the discrete version of GW optimal transport (Eq. 1), enabling multimodal alignment with observed cell profiles. However, single-cell measurements are high-dimensional and sparse, such that computing the distance matrix D in the high-dimensional feature space can be biased and distorted. Therefore, `scMultiNODE` first uses an autoencoder (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 [62–64]. The benefit of AEs for single-cell data is that they can effectively capture complex local cell relationships (like cell clusters) due to the non-linearity property of the neural networks. AE consists of two neural networks: (1) the encoder network $\mathrm{Enc}(\cdot, \phi)$ maps the input features to a low-dimensional space $\mathbb{R}^d$ ($d \ll \min\{p, q\}$), and (2) the decoder network $\mathrm{Dec}(\cdot, \theta)$ maps latent variables back to the feature space to reconstruct the input. Specifically, given $\mathbf{X}_{\mathrm{ALL}}$ and $\mathbf{Y}_{\mathrm{ALL}}$, `scMultiNODE` learns modality-specific latent representations through

$$\begin{aligned} \mathbf{Z}_{\mathrm{RNA}} &= \mathrm{Enc}_{\mathbf{X}}(\mathbf{X}_{\mathrm{ALL}},\ \phi_{\mathbf{X}}), \quad \widehat{\mathbf{X}}_{\mathrm{ALL}} = \mathrm{Dec}_{\mathbf{X}}(\mathbf{Z}_{\mathrm{RNA}},\ \theta_{\mathbf{X}}); \\ \mathbf{Z}_{\mathrm{ATAC}} &= \mathrm{Enc}_{\mathbf{Y}}(\mathbf{Y}_{\mathrm{ALL}},\ \phi_{\mathbf{Y}}), \quad \widehat{\mathbf{Y}}_{\mathrm{ALL}} = \mathrm{Dec}_{\mathbf{Y}}(\mathbf{Z}_{\mathrm{ATAC}},\ \theta_{\mathbf{Y}}). \end{aligned} \tag{3}$$

The encoder and decoder networks are parameterized by ϕ and θ, respectively. AE minimizes the mean squared error (MSE) between input features and the reconstructions from the decoder. Note that `scMultiNODE` trains these two AEs separately at this step to capture modality-specific cell type variations. Then, with the low-dimensional representations, we compute intra-modality distance matrices based on k-nearest neighbor (kNN), as done previously in [29, 30, 65]. With these, `scMultiNODE` aligns cells from

RNA and ATAC modalities by minimizing the following GW distance

$$\texttt{scMultiNODE}\ \mathrm{GW}(\mathbf{Z}_{\mathrm{RNA}}, \mathbf{Z}_{\mathrm{ATAC}}) = \min_{T \in \Pi(\nu_X, \nu_Y)} \sum_{i,i',j,j'} \parallel \mathbf{D}^{\mathrm{RNA}}_{ii'} - \mathbf{D}^{\mathrm{ATAC}}_{jj'} \parallel^2 \mathbf{T}_{ij} \mathbf{T}_{i'j'}$$
$$\text{with} \quad \mathbf{D}^{\mathrm{RNA}} = \mathrm{kNN}(\mathbf{Z}_{\mathrm{RNA}}) \quad \text{and} \quad \mathbf{D}^{\mathrm{ATAC}} = \mathrm{kNN}(\mathbf{Z}_{\mathrm{ATAC}}). \tag{4}$$

We binarize the 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 help of the cell correspondence matrix $\mathbf{T}$, `scMultiNODE` now can integrate gene expression and chromatin accessibility measurements into a joint latent space. Specifically, `scMultiNODE` adds another neural network $\mathrm{Fus}(\cdot,\ \omega) : \mathbb{R}^d \mapsto \mathbb{R}^d$ 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

$$\tilde{\mathbf{z}} = \mathrm{Fus}(\mathbf{z},\ \omega) \quad \text{with}$$
$$\begin{cases} \mathbf{z} = \mathrm{Enc}_{\mathrm{RNA}}(\mathbf{x},\ \phi_{\mathbf{X}}) \quad \text{and} \quad \widehat{\mathbf{x}} = \mathrm{Dec}_{\mathrm{RNA}}(\tilde{\mathbf{z}},\ \theta_{\mathbf{X}}) & \text{for RNA data} \\ \mathbf{z} = \mathrm{Enc}_{\mathrm{ATAC}}(\mathbf{y},\ \phi_{\mathbf{Y}}) \quad \text{and} \quad \widehat{\mathbf{y}} = \mathrm{Dec}_{\mathrm{ATAC}}(\tilde{\mathbf{z}},\ \theta_{\mathbf{Y}}) & \text{for ATAC data} \end{cases}. \tag{5}$$

With the fusion layer, `scMultiNODE` maps latent representations of different feature spaces into the same space. We assume that different modalities, though they profile different cells, should have some underlying shared biological information. For example, the two modalities may capture the same cell line or tissue, such that they have similar cell-type compositions. Therefore, we hypothesize that if cells i and j from different modalities are biologically similar, they should be aligned in the joint latent space and their latent representations $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{z}}_j$ should be close to each other. To this end, `scMultiNODE` accomplishes the cell correspondence matrix $\mathbf{T}$ from GW optimal transport (Eq. 4), which captures local cell relationships including cell type clusters, and aligns modalities in the joint latent space by minimizing the loss function

$$\mathcal{L}_{\mathrm{fusion}} = \alpha \sum_{i,j} \mathbf{1}\left(\mathbf{T}_{ij} \neq 0\right) \parallel \tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j \parallel_2^2 + \mathrm{MSE}\left(\mathbf{X}_{\mathrm{ALL}},\ \widehat{\mathbf{X}}_{\mathrm{ALL}}\right) + \mathrm{MSE}\left(\mathbf{Y}_{\mathrm{ALL}},\ \widehat{\mathbf{Y}}_{\mathrm{ALL}}\right) \tag{6}$$

with

$$\tilde{\mathbf{z}}_i = \mathrm{Fus}(\mathbf{z}_i, \omega) \text{ and } \tilde{\mathbf{z}}_j = \mathrm{Fus}(\mathbf{z}_j, \omega) \text{ for RNA cell } i \text{ and ATAC cell } j \quad \text{(from Eq. 5)}$$
$$\widehat{\mathbf{X}}_{\mathrm{ALL}} = \mathrm{Dec}_{\mathbf{X}}(\tilde{\mathbf{Z}}_{\mathrm{RNA}},\ \theta_{\mathbf{X}}) \text{ with } \tilde{\mathbf{Z}}_{\mathrm{RNA}} = \{\tilde{\mathbf{z}}_i \mid \text{all RNA cells } i\},$$
$$\widehat{\mathbf{Y}}_{\mathrm{ALL}} = \mathrm{Dec}_{\mathbf{Y}}(\tilde{\mathbf{Z}}_{\mathrm{ATAC}},\ \theta_{\mathbf{Y}}) \text{ with } \tilde{\mathbf{Z}}_{\mathrm{ATAC}} = \{\tilde{\mathbf{z}}_j \mid \text{all ATAC cells } j\}. \tag{7}$$

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. 6 and update decoders accordingly.

Scalable GW-based multi-modal alignment using quantization

The exact computation of GW distance (Eq. 4) 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) [36], to significantly speed up computations. QGW uses a divide-and-conquer strategy, where it partitions feature spaces into several blocks and matches blocks recursively. We derive the QGW computation following Eq. 4.

Specifically, QGW assumes the feature space $\mathbf{Z}$ can be divided into disjoint and nonempty sets $U_1, \cdots, U_r$, with $r << |\mathbf{Z}|$. For example, the large-scale single-cell measurements can be divided into several cell-type groups. Within each partition U_s ($s = 1, \cdots, r$), there exists a representative point $z_s \in U_s$, like the anchor cell for each cell type. We define $Z_r := \{z_1, \cdots, z_r\}$ as quantized representations of space $\mathbf{Z}$. Therefore, given latent representations $\mathbf{Z}_{\text{RNA}}$ and $\mathbf{Z}_{\text{ATAC}}$ and their quantized representations Z_r^{RNA} and Z_r^{ATAC}, QGW algorithm proceeds in three steps.

- *Global Alignment:* QGW first computes a transport plan $\pi_r \in \mathbb{R}^{r \times r}$ between representative points of two modalities. Since the number of partitions r is chosen to be much smaller than the number of cells, computing π_r can be feasibly approximated via any existing GW algorithm. This transport plan provides a global alignment between partitioned blocks.
- *Local Alignment:* Then QGW produces a collection of local alignments between individual cells. For each partition $z_a \in Z_r^{\text{RNA}}$ and $z_b \in Z_r^{\text{RNA}}$, we compute the transport plan $\pi_{ab} \in \mathbb{R}^{|U_a| \times |V_b|}$ between the RNA block U_a and ATAC block V_b by solving
$$\min_{\pi_{ab} \in \Pi(\nu_{U_a}, \nu_{V_b})} \sum_{i \in U_a, j \in V_b} \| \mathbf{D}^{\text{RNA}}_{iz_a} - \mathbf{D}^{\text{ATAC}}_{jz_b} \|^2 \, \pi_{ab}(i,j), \tag{8}$$
where $\mathbf{D}^{\text{RNA}}_{iz_a}$ is the distance between any cell of block U_a and its representative point z_a ($\mathbf{D}^{\text{ATAC}}_{jz_b}$ is defined in the same way). $\pi_{ab}(i,j)$ is the element at row i and column j. [36] has proved that Eq. 8 can be solved efficiently with a log-linear time complexity with respect to the size of the maximum block.
- *Coupling:* With local and global transport plans, we can merge them into the final solution of QGW as
$$\mathbf{T} = \sum_{a,b} \pi_r(a,b) \pi_{ab}. \tag{9}$$
[36] has proved that the QGW can give a good approximation when the feature spaces being compared admit compact partitions, which is a realistic setting for single-cell measurements where we have multiple homogeneous cell-type groups.

An important routine in QGW is generating good partitions. Since our `scMultiNODE` is an unsupervised model without cell type labels, we followed QGW

defaults [36] as randomly and uniformly choosing a small set of cells without replacement as representative points and computed a Voronoi partition.

4.1.2 Incorporation of global cell transition dynamics with neural Ordinary Differential Equation (ODE)

`scMultiNODE` uses QGW to capture local cell relationships in the joint latent space. However, the joint space for now does not account for global cellular dynamics, such as developmental trajectories or lineage bifurcations. All the previous integration methods have this limitation, hence, their latent representations poorly define natural cell paths across timepoints and hinder downstream analysis. To address the problem, `scMultiNODE` further incorporates global cellular dynamics in the joint latent space.

`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 [66]. Recent studies have adopted neural networks to approximate the derivative function and proposed neural ODEs [13]. Previous studies [5, 14, 67] 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

$$\mathrm{d}\tilde{\mathbf{Z}}^{(t)} = \mathrm{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right) \cdot \mathrm{d}t. \tag{10}$$

Drift is a non-linear neural network parameterized by μ, modeling the developmental cell velocities in the joint latent space, such that $\mathrm{Drift}(\tilde{\mathbf{Z}}^{(t)}, \mu)$ 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 $0 \in \mathcal{T}_{\mathrm{RNA}}$)

$$\begin{aligned} \tilde{\mathbf{Z}}^{(t+\Delta t)} &= \tilde{\mathbf{Z}}^{(t)} + \mathrm{Drift}\left(\tilde{\mathbf{Z}}^{(t)},\ \mu\right)\Delta t \\ \text{with} \qquad \tilde{\mathbf{Z}}^{(0)} &= \mathrm{Fus}\left(\mathbf{Z}^{(0)}_{\mathrm{RNA}},\ \omega\right) \quad \text{and} \quad \mathbf{Z}^{(0)}_{\mathrm{RNA}} = \mathrm{Enc}_{\mathbf{X}}\left(\mathbf{X}^{(0)},\ \phi_{\mathbf{X}}\right), \end{aligned} \tag{11}$$

Here, hyperparameter Δt denotes step size and drift term $\mathrm{Drift}\left(\tilde{\mathbf{Z}}^{(t)}, \mu\right)\Delta t$ represents the forward steps taken in the joint latent space. We use the first-order Euler method (in Eq. 11) 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 $t \in \mathcal{T}_{\mathrm{RNA}}$ or $t \in \mathcal{T}_{\mathrm{ATAC}}$,

`scMultiNODE` uses the decoder $\mathrm{Dec}_{\mathbf{X}}/\mathrm{Dec}_{\mathbf{Y}}$ to convert latent variables $\tilde{\mathbf{Z}}^{(t)}$ generated from Eq. 11 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 [68] to measure the distance between distributions defined by ground truth $\mathbf{X}$ and predictions $\widehat{\mathbf{X}}$ as

$$\mathrm{Wass}(\mathbf{X}, \widehat{\mathbf{X}}) = \left(\min_{\Gamma \sim \Pi(\mathbf{X}, \widehat{\mathbf{X}})} \sum_{i,j} D_{ij}^2 \Gamma_{ij} \right)^{1/2} \qquad \text{with} \qquad D_{ij} = \|\mathbf{X}_i - \widehat{\mathbf{X}}_j\|_2. \tag{12}$$

Wasserstein distance measures the distribution distance in the same metric space, different from the 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 [14] 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., $\mathrm{Fus}\left(\mathrm{Enc}_{\mathrm{RNA}}(\mathbf{X}, \phi_{\mathbf{X}})\right)$ or $\mathrm{Fus}\left(\mathrm{Enc}_{\mathrm{ATAC}}(\mathbf{Y}, \phi_{\mathbf{Y}})\right)$) 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

$$\begin{aligned} \mathcal{R}\left(\mathcal{T}_{\mathrm{RNA}}, \mathcal{T}_{\mathrm{ATAC}}\right) &= \sum_{t \in \mathcal{T}_{\mathrm{RNA}}} \mathrm{Wass}\left(\tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)},\ \tilde{\mathbf{Z}}^{(t)}\right) + \sum_{t \in \mathcal{T}_{\mathrm{ATAC}}} \mathrm{Wass}\left(\tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)},\ \tilde{\mathbf{Z}}^{(t)}\right) \\ \text{with} &\begin{cases} \tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)} = \mathrm{Fus}\left(\mathrm{Enc}_{\mathrm{RNA}}(\mathbf{X}^{(t)}, \phi_{\mathbf{X}}), \omega\right) & \text{for } t \in \mathcal{T}_{\mathrm{RNA}} \\ \tilde{\mathbf{Z}}_{\mathrm{fus}}^{(t)} = \mathrm{Fus}\left(\mathrm{Enc}_{\mathrm{ATAC}}(\mathbf{Y}^{(t)}, \phi_{\mathbf{Y}}), \omega\right) & \text{for } t \in \mathcal{T}_{\mathrm{ATAC}} \end{cases}, \\ &\text{and} \quad \tilde{\mathbf{Z}}^{(t)} \text{ comes from neural ODE} \quad (\text{Eq. } 11) \end{aligned} \tag{13}$$

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

$$\mathcal{L}_{\mathrm{dyn}} = \sum_{t \in \mathcal{T}_{\mathrm{RNA}}} \mathrm{Wass}\left(\mathbf{X}^{(t)},\ \widehat{\mathbf{X}}^{(t)}\right) + \sum_{t \in \mathcal{T}_{\mathrm{ATAC}}} \mathrm{Wass}\left(\mathbf{Y}^{(t)},\ \widehat{\mathbf{Y}}^{(t)}\right) + \beta \mathcal{R}\left(\mathcal{T}_{\mathrm{RNA}}, \mathcal{T}_{\mathrm{ATAC}}\right), \tag{14}$$

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, which focus 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 global cellular dynamics (e.g., long-term cell state transitions) alongside local cell relationships (e.g., cell type clusters).

4.2 `scMultiNODE` Training Details

Our `scMultiNODE` is implemented with *Pytorch 1.13* [69] 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 RNA/ATAC reconstruction MSE loss with a learning rate of 0.001 and 1000 iterations. Then, `scMultiNODE` aligns modality-specific latent representations with QGW optimal transport. In the QGW algorithm, we construct the intra-modality distance matrices $\mathbf{D}^{\mathrm{RNA}}$ and $\mathbf{D}^{\mathrm{ATAC}}$ through the kNN graph. Once the cell correspondence matrix $\mathbf{T}$ is estimated from the QGW algorithm, `scMultiNODE` maps modality-specific latent representations to a joint latent space (through fusion layer $\mathrm{Fus}(\cdot, \omega)$) by minimizing Eq. 6 ($\mathcal{L}_{\mathrm{fusion}}$) 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 $\mathcal{L}_{\mathrm{dyn}}$ (Eq. 14). 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 $t \in \mathcal{T}_{\mathrm{RNA}} \cup \mathcal{T}_{\mathrm{ATAC}}$. Because the Wasserstein distance computation is expensive, batch training improves training efficiency and enables `scMultiNODE` usage on large-scale datasets. We use *geomloss* [70] to compute Wasserstein distance with blur = 0.05 and scaling = 0.5. Pseudo-codes of `scMultiNODE` are provided in Algorithm 1.

4.3 Single-Cell Dataset and Pre-Processing

We use six 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 [17]. They have normalized data with scTransform [71] 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 this and the following datasets, we subsample cells while retaining cell type proportions. For each modality, we select the top

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 Δt, α, β, 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)

$\mathbf{X}_{\text{ALL}} = \text{CONCAT}\left(\mathbf{X}^{(t)} \mid t \in \mathcal{T}_{\text{RNA}}\right)$

$\mathbf{Y}_{\text{ALL}} = \text{CONCAT}\left(\mathbf{Y}^{(t)} \mid t \in \mathcal{T}_{\text{ATAC}}\right)$

Optimize RNA-related AE parameters $\phi_{\mathbf{X}}$ and $\theta_{\mathbf{X}}$ to minimize MSE for RNA (Eq. 3)

Optimize ATAC-related AE parameters $\phi_{\mathbf{Y}}$ and $\theta_{\mathbf{Y}}$ to minimize MSE for ATAC (Eq. 3)

(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 to predict cell correspondence matrix $\mathbf{T}$ (Eq. 4 and Eq. 9)

Map latent representations to joint latent space through fusion layer (Eq. 5)

Optimize fusion layer parameter (ω) and AE decoder parameters ($\theta_{\mathbf{X}}$ and $\theta_{\mathbf{Y}}$) to minimize $\mathcal{L}_{\text{fusion}}$ (Eq. 6)

(Step III: cellular dynamics)

Optimize the entire model to minimize $\mathcal{L}_{\text{dyn}}$ (Eq. 14)

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

2000 highly variable genes (HVGs) using Scanpy [72]. 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 [8]. 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 embryogenesis (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) [18]. 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 [23], 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 [73]. 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 [74]. The MN data is downloaded from Gene Expression Omnibus SuperSeries GSE204851 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE204851).
- **Zebrahub (ZB):** Lange et al. collect a dynamic atlas of zebrafish embryonic development that integrates single-cell sequencing time course data [75]. The original paper shows that the data does not exhibit significant batch effects. For each modality, we randomly sample 10% of cells and test our model on the gene expression count matrix of 3692 cells and chromatin gene activity matrix of 9456 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. The ZB data is downloaded from Zebrahub (https://zebrahub.sf.czbiohub.org/).
- **Amphioxus development (AM):** Ma et al. perform snRNA-seq and scATAC-seq for different stages of amphioxus embryogenesis [76]. The batch effects have been removed in the original paper analysis. The dataset provides a gene expression count matrix of 9630 cells and chromatin peak matrix of 3538 cells. We select the top 2000 HVGs of scRNA-seq data and the top 2000 variable peaks for scATAC-seq data. The AM data is downloaded from AmphioxusAtlas (https://lifeomics.shinyapps.io/shinyappmulti/).

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

$$\mathbf{X}_i = \frac{\mathbf{X}_i}{\sum_{j=1}^{p} \mathbf{X}_i} * 10^4, \qquad \text{followed by} \qquad \mathbf{X}_{ij} = \log(\mathbf{X}_{ij} + 1). \tag{15}$$

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

standard preprocessing step to the chromatin accessibility matrix by binarizing the count values, such that each entry reflects the presence (1) or absence (0) of accessibility at a given peak in each cell. This approach helps mitigate technical noise inherent in the highly sparse chromatin accessibility data. An overview of these datasets is included in Supplementary Table S1.

4.4 Baseline Models

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

- `Seurat`: The single-cell analysis platform `Seurat` [23] 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. [29] 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` [30] 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. [65] 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 S2, we set all its other hyperparameters as default.
- `Pamona`: The `Pamona` [28] 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. [31] introduce `uniPort`, incorporating coupled variational autoencoders 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`.

4.5 Evaluation metrics

We evaluate each model's integration from four perspectives: modality integration, cross-modal cell label transfer, capturing cell type variation, and preserving cellular dynamics. Therefore, we adopt the following evaluation metrics.

Modality integration

We use `batch entropy` to evaluate the integration of unaligned datasets (DR, MN, ZB, and AM). `Batch entropy` is originally introduced in Xiong et al. [77] and previously adopted by Cao et al. [31]. It evaluates the sum of regional mixing entropies between different datasets where a high score indicates cells from different modalities are mixed well. Specifically,

$$\text{batch entropy} = \sum_{k \in \{\text{RNA,ATAC}\}} p'_k \log\left(p'_k\right) \qquad \text{with} \qquad p'_k = \frac{p_i/P_i}{\sum_{j \in \{\text{RNA,ATAC}\}} p_j/P_j}, \tag{16}$$

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.

For the co-assay datasets (HC and HO) where one-to-one cell correspondence information is implicitly available, we additionally use the fraction of samples closer than the true match (`FOSCTTM`) [27, 29], `neighborhood overlap` [65], 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. In our analysis, we show `1-FOSCTTM` instead of `FOSCTTM` to unify metrics comparison, where a higher metric value implies better integration performance.

Cell label transfer

We also evaluate integration using cell type labels through label transfer accuracy (`LTA-type`) as in previous studies [30, 31, 65]. 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.

Cell type preservation

To assess how well cell type clusters are preserved after integration, we calculate the Normalized Mutual Information (`NMI`) score [78]. It quantifies the similarity between predicted clusters and ground truth labels, ranging from 0 to 1, with higher values

indicating better preservation of cell groups. Specifically, we apply the Louvain algorithm [79] to cluster cells in the integrated latent space and calculate the NMI between the resulting cluster labels and the true cell types.

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 [80] 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` [81] package to compute distance correlations.

4.6 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, MN, AB, and AM), we select hyperparameters that yield the maximum `LTA-type` based on common cell type labels. We use Optuna [82] 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 S2. 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.

4.7 Comparison of Multi-Modal Integration

We compare `scMultiNODE` 's integration performances with baseline methods on both co-assay and unaligned datasets. In Fig. 2, we compare the 2D UMAP visualization of multi-modal integrations for HC and ZB datasets. The UMAP and principal component analysis (PCA) visualizations of integrations for all datasets are shown in Supplementary Fig. S1-S4. All evaluation metric calculations of Fig. 3 were implemented in Python and the detailed metric values are listed in Supplementary Table S3-S4.

4.8 `scMultiNODE` Training with Cell Type Supervision

To investigate the effect of using cell type supervision into our integration framework, we extend `scMultiNODE` by adding a classification head on top of the latent space.

Specifically, we use a fully connected layer with a softmax activation to predict cell type labels, and train the classification head with cross entropy loss. This cross entropy loss is jointly optimized with the original loss function of `scMultiNODE` (Eq. 14), thereby encouraging the latent representations to reflect known cell type groups. The classification loss is weighted by a tunable coefficient, set to 1.0 in this experiment. All other hyperparameters follow the tuning procedure described in Sec. 4.6. We apply this supervised variant of `scMultiNODE` to all datasets and compare to the unsupervised (original) version of `scMultiNODE` .

UMAP visualizations (Supplementary Fig. S9) reveal that incorporating cell type supervision significantly improves the separation of cell groups in the joint latent space. For example, on the HC dataset, the supervised model achieves a `NMI` of 0.614, compared to 0.299 for the unsupervised version. The `LTA-type` also increases from 0.392 to 0.690, suggesting that cell type boundaries are more distinct when supervision is applied (Supplementary Table S11-S12). However, this improvement in cell type clustering comes at the cost of modality integration and cellular dynamics preservation. The `batch entropy`, which measures how well cells from different modalities are integrated, decreases from 0.667 (unsupervised) to 0.453 (supervised) on HC dataset. Similarly, the `time correlation` drops from 0.979 to 0.911, and `LTA-time` declines significantly from 0.919 to 0.515 on HC data (Supplementary Table S11-S12). These show that the supervised model captures less of the underlying developmental dynamics.

These results highlight a key trade-off. While cell type supervision improves clustering performance and facilitates cell group separation, it constrains the latent space in a way that limits cross-modality integration and dynamic structure preservation. This suggests that the supervised classification loss term may distort the geometry of the latent space by forcing it to align the discrete cell group boundaries, which may not reflect the continuous nature of biological processes. Additionally, inaccurate or coarse-grained cell annotations could introduce label noise, further biasing the learned representations. In our study here, we focus on unsupervised modality integration.

4.9 Investigation of `scMultiNODE` Hyperparameters

We evaluate `scMultiNODE` 's performance under different hyperparameter settings using two representative datasets: a co-assay dataset (HC) and an unaligned dataset (DR). We first run `scMultiNODE` with the joint latent space size d varies from $\{10, 50, 100, 150, 200\}$. Supplementary Table S6 show 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 [83] 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 QGW optimal transport to align cell representations from two modalities and ensures aligned cells have similar latent presentations. The QGW 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 S7, `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. 6) is varied from $\{0.0, 0.01, 0.1, 1.0, 10.0, 100.0\}$. Supplementary Table S8 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. 6) is essential for learning a joint latent space that effectively captures the diverse variations.

`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 S9 denote 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 $[0.01, 10.0]$.

4.10 Comparison of time costs

We compare the runtime of `scMultiNODE` with the baseline models using different numbers of cells ($\{1000, 2000, \cdots, 6000\}$) from the HO dataset. We pick this dataset as it has the most number of cells (Supplementary Table S1), enabling us to test runtime costs on many cells. All methods are run on the Intel Xeon Platinum 8268 CPU with 32GB of memory. As shown in Supplementary Fig. S5, `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.

4.11 Cell Trajectory Analysis

In Fig. 4, we applied `scMultiNODE` in the cell trajectory analysis on the human cortex (HC) dataset. This dataset contains multifurcating cell trajectories in the human brain cortex, allowing us to validate our analysis with known cell developmental lineages. Apart from baseline multi-modal integration methods, we also compare

scMultiNODE with static single-modal method AE and a dynamic single-modal method scNODE [14]. scNODE is a novel method that explicitly incorporates cellular dynamics in the latent space. On each dataset, the scNODE's hyperparameters are optimized following the strategy from its original paper.

Specifically, for each method's latent space, we used Monocle3 [39] to perform trajectory inference and estimate cell pseudotime (Supplementary Fig. S7). Monocle3 begins by reducing the dimensionality of the input data (already 50-dimensional latent cell embeddings in our case), followed by the construction of a tree-like principal graph to learn the manifold structure of the data. Branches of the tree correspond to divergent cell fate decisions. To assign pseudotime values, a root node must be selected, which serves as the starting point of the trajectory. We choose the graph node corresponding to the earliest timepoint in the dataset as the root. Monocle3 computes pseudotime as the geodesic distance along the learned principal graph from the root node to each cell, thereby reflecting each cell's relative progression through the underlying biological process. We run Monocle3 following its tutorial (https://cole-trapnell-lab.github.io/monocle3/docs/trajectories/) and use default function parameters. To quantitatively evaluate the accuracy of pseudotime predictions, we calculate the Spearman rank correlation coefficient (ρ) between the true cell order labels along the reference trajectory and the inferred pseudotime values. Spearman correlation assesses the degree to which the rank ordering of cells is preserved, making it well-suited for capturing the correctness of developmental progression. A higher correlation value indicates that the inferred pseudotime more accurately reflects the true temporal sequence of cell states. Detailed Spearman rank correlation coefficient values for all methods are shown in Supplementary Table S10.

Additionally, we applied PAGA [40] for pseudotime inference (Supplementary Fig. S8), another state-of-the-art method that models cell connectivity by constructing a graph representing the data topology. Pseudotime can be inferred along paths in this graph, offering robust trajectory inference across complex lineage structures. We follow its recommended workflow (https://scanpy-tutorials.readthedocs.io/en/latest/paga-paul15.html) with default function parameters, and again computed Spearman correlations (Supplementary Fig. S6). The PAGA pseudotime shows that scMultiNODE obtains $\rho = 0.83$ on the EN-related trajectory, whereas the best integration baseline has $\rho = 0.55$; scMultiNODE reaches $\rho = 0.71$ on the IN-mge-related trajectory, while the best integration baseline is only $\rho = 0.57$. Results are similar for the IN-cge trajectory. Furthermore, for the Oligendendrocyte-related trajectory, scMultiNODE has $\rho = 0.46$, comparable to SCOTv1 ($\rho = 0.52$) but outperforming all other integration baselines. When compared to single-modality models, scMultiNODE outperforms AE and scNODE on ATAC data in all cases. For instance on the EN-related trajectory, AE obtains $\rho = 0,45$ and scNODE has $\rho = 0.31$ on ATAC data. Therefore, these results further confirm that scMultiNODE 's integration faithfully preserves biologically meaningful pseudotime across complex lineages.

4.12 Identification of Cell Development-Related Genes

In Fig. 5, we use HC dataset to investigate whether scMultiNODE can be used to reconstruct meaningful cell developmental paths and uncover underlying gene programs.

The cell path is constructed with the least action path (LAP), which has been used in previous studies [5, 41, 42] 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

$$\mathcal{S} = \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{V}_k - \mathrm{Drift}_\omega(\widetilde{\mathbf{X}}_k)\right)^2 \Delta \quad \text{with } \widetilde{\mathbf{X}}_k = \frac{\mathbf{X}_{k-1} + \mathbf{X} - k}{2}. \tag{17}$$

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

$$\widehat{\mathcal{P}} \;=\; \underset{\mathcal{P},\Delta}{\mathrm{argmin}}\, \mathcal{S} \;=\; \underset{\mathcal{P},\Delta}{\mathrm{argmin}}\, \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{V}_k - \mathrm{Drift}(\widetilde{\mathbf{X}}_k, \omega)\right)^2 \Delta. \tag{18}$$

Solving Eq. 18 consists of two iterative steps

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

$$\widehat{\Delta} = \underset{\Delta}{\mathrm{argmin}}\, \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbf{X}_k - \mathbf{X}_{k-1}}{\Delta} - \mathrm{Drift}(\widetilde{\mathbf{X}}_k, \omega)\right)^2 \Delta. \tag{19}$$

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

$$\widehat{\mathcal{P}} = \underset{\mathbf{X}_1,\cdots,\mathbf{X}_{K-1}}{\mathrm{argmin}}\, \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbf{X}_k - \mathbf{X}_{k-1}}{\widehat{\Delta}} - \mathrm{Drift}(\widetilde{\mathbf{X}}_k, \omega)\right)^2 \Delta. \tag{20}$$

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. In parallel, we apply *Scanpy* to identify marker genes for each cell type based on either RNA or ATAC data independently. This allows us to compare the DE genes discovered from the `scMultiNODE` joint latent space with traditional marker genes derived from scRNA-seq or scATAC-seq alone. Our results indicate that explicitly modeling cell dynamics in the joint space enables more effective identification of development-associated DE genes compared to single-modal marker gene analysis.

4.13 Germ Layer Label Transfer

In Fig. 6, we demonstrate cross-modality germ layer label transfer by leveraging the joint latent space learned by `scMultiNODE` on the unaligned drosophila embryogenesis (DR) dataset. In this setting, germ layer annotations are available only for the RNA modality, and not for the ATAC modality. We assess whether the integrated latent space from each method preserves biologically meaningful structures that enable accurate label transfer from RNA to ATAC.

To evaluate this, we first trained a random forest classifier [51] on the latent representations of RNA cells using their known germ layer labels. The classifier, implemented using the default parameters from the *scikit-learn* package [84], was then applied to the ATAC cell embeddings within the same joint space to infer their germ layer identities. Fig. 6B indicates that `scMultiNODE` produced coherent ATAC clusters that aligned with known developmental trajectories from broad ectodermal states toward neuroectodermal subpopulations. However, the baseline methods either fail to predict all germ layer labels or do not preserve the underlying developmental trajectories (Supplementary Fig. S10). This implies that the `scMultiNODE` integration captures meaningful biological progression, highlighting its advantage over existing approaches.

Since ground truth germ layer labels for ATAC cells are not available, we validated the predicted labels through two complementary approaches. First, we identified germ layer-specific marker genes from the RNA modality using the Wilcoxon rank-sum test in *Scanpy*, and then examined their expression patterns in the ATAC-derived gene activity matrix, grouped by the predicted germ layer labels. As shown in Fig. 6C, marker gene expression clearly separates the predicted germ layer groups in our model, indicating successful label transfer. In contrast, baseline methods show diffuse or overlapping marker expression across groups, suggesting poor alignment and less effective transfer of germ layer identity (Supplementary Fig. S11). Second, we performed Gene Ontology (GO) enrichment analysis using clusterProfiler [52] on the ATAC-derived gene activity matrix of each predicted germ layer group, focusing on the Biological Process (BP) ontology. The enriched GO terms reveal distinct functional profiles for each germ layer group, aligning well with the known biological roles of their respective germ layers.

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

References

[1] Walpole, J., Papin, J. A. & Peirce, S. M. Multiscale computational models of complex biological systems. *Annual review of biomedical engineering* **15**, 137–154 (2013).

[2] Spiller, D. G., Wood, C. D., Rand, D. A. & White, M. R. Measurement of single-cell dynamics. *Nature* **465**, 736–745 (2010).

[3] Zakrzewski, W., Dobrzyński, M., Szymonowicz, M. & Rybak, Z. Stem cells: past, present, and future. *Stem cell research & therapy* **10**, 1–22 (2019).

[4] Ding, J., Sharon, N. & Bar-Joseph, Z. Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics* **23**, 355–368 (2022).

[5] Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711 (2022).

[6] Liu, B. *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**, 108–121 (2022).

[7] Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).

[8] Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2023).

[9] Pliner, H. A. *et al.* Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular cell* **71**, 858–871 (2018).

[10] 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**, 3222 (2021).

[11] Huguet, G. *et al.* Manifold interpolating optimal-transport flows for trajectory inference. *Advances in Neural Information Processing Systems* **35**, 29705–29718 (2022).

[12] Tong, A., Huang, J., Wolf, G., Van Dijk, D. & Krishnaswamy, S. *Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics*, 9526–9536 (PMLR, 2020).

[13] Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems* **31** (2018).

[14] Zhang, J., Larschan, E., Bigness, J. & Singh, R. scnode: generative model for temporal single cell transcriptomic data prediction. *Bioinformatics* **40**, ii146–ii154 (2024).

[15] Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome biology* **21**, 1–35 (2020).

[16] Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology* **20**, 1–25 (2019).

[17] Zhu, K. *et al.* Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Science Advances* **9**, eadg3754 (2023).

[18] Calderon, D. *et al.* The continuum of drosophila embryonic development at single-cell resolution. *Science* **377**, eabn5800 (2022).

[19] Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* **183**, 1103–1116 (2020).

[20] Hu, Y. *et al.* Simultaneous profiling of transcriptome and dna methylome from a single cell. *Genome biology* **17**, 1–11 (2016).

[21] Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature methods* **17**, 11–14 (2020).

[22] Wu, X. *et al.* Single-cell sequencing to multi-omics: technologies and applications. *Biomarker Research* **12**, 110 (2024).

[23] Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology* **42**, 293–304 (2024).

[24] Liu, J. *et al.* Jointly defining cell types from multiple single-cell datasets using liger. *Nature protocols* **15**, 3632–3662 (2020).

[25] Benton, A. *et al.* Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519* (2017).

[26] 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).

[27] Liu, J., Huang, Y., Singh, R., Vert, J.-P. & Noble, W. S. *Jointly embedding multiple single-cell omics measurements*, Vol. 143 (NIH Public Access, 2019).

[28] Cao, K., Hong, Y. & Wan, L. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics* **38**, 211–219 (2022).

[29] 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**, 3–18 (2022).

[30] 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**, 1213–1228 (2022).

[31] Cao, K., Gong, Q., Hong, Y. & Wan, L. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications* **13**, 7419 (2022).

[32] Connor, M., Canal, G. & Rozell, C. *Variational autoencoder with learned latent structure*, 2359–2367 (PMLR, 2021).

[33] Trapnell, C. Defining cell types and states with single-cell genomics. *Genome research* **25**, 1491–1498 (2015).

[34] Dance, A. What is a cell type, really? the quest to categorize life's myriad forms. *Nature* **633**, 754–756 (2024).

[35] Peyré, G., Cuturi, M. *et al.* Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* **11**, 355–607 (2019).

[36] Chowdhury, S., Miller, D. & Needham, T. *Quantized gromov-wasserstein*, 811–827 (Springer, 2021).

[37] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[38] Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods* **16**, 1289–1296 (2019).

[39] Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

[40] Wolf, F. A. *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 59 (2019).

[41] 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**, e49271 (2012).

[42] Wang, P. *et al.* Epigenetic state network approach for describing cell phenotypic transitions. *Interface focus* **4**, 20130068 (2014).

[43] Pazarlar, B. A. *et al.* 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**, 45 (2022).

[44] Ciruelas, K., Marcotulli, D. & Bajjalieh, S. M. *Synaptic vesicle protein 2: A multi-faceted regulator of secretion*, Vol. 95, 130–141 (Elsevier, 2019).

[45] Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Science advances* **7**, eabh2169 (2021).

[46] Hagiwara, N. Sox6, jack of all trades: a versatile regulatory protein in vertebrate development. *Developmental Dynamics* **240**, 1311–1321 (2011).

[47] Kiecker, C., Bates, T. & Bell, E. Molecular specification of germ layers in vertebrate embryos. *Cellular and Molecular Life Sciences* **73**, 923–947 (2016).

[48] Wang, R. *et al.* Time space and single-cell resolved tissue lineage trajectories and laterality of body plan at gastrulation. *Nature Communications* **14**, 5675 (2023).

[49] Zheng, Y. *et al.* Single-cell analysis of embryoids reveals lineage diversification roadmaps of early human development. *Cell stem cell* **29**, 1402–1419 (2022).

[50] Ma, W., Lu, J. & Wu, H. Cellcano: supervised cell type identification for single cell atac-seq data. *Nature Communications* **14**, 1864 (2023).

[51] Ho, T. K. *Random decision forests*, Vol. 1, 278–282 (IEEE, 1995).

[52] Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284–287 (2012).

[53] Crews, S. T. Drosophila embryonic cns development: neurogenesis, gliogenesis, cell fate, and differentiation. *Genetics* **213**, 1111–1144 (2019).

[54] Singhania, A. & Grueber, W. B. Development of the embryonic and larval peripheral nervous system of drosophila. *Wiley Interdisciplinary Reviews: Developmental Biology* **3**, 193–210 (2014).

[55] La Manno, G. *et al.* Rna velocity of single cells. *Nature* **560**, 494–498 (2018).

[56] Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. Rna velocity—current challenges and future perspectives. *Molecular systems biology* **17**, e10282 (2021).

[57] Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology* **38**, 1408–1414 (2020).

[58] Karbalayghareh, A., Barisic, D., Chin, C. R., Melnick, A. & Leslie, C. S. Deep dynamical models of single-cell multiomic velocities predict loss-of-function and rescue perturbations in b cells. *bioRxiv* 2025–04 (2025).

[59] Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A. & Hansen, K. D. Pumping the brakes on rna velocity by understanding and interpreting rna velocity estimates. *Genome biology* **24**, 246 (2023).

[60] Shamir, M., Bar-On, Y., Phillips, R. & Milo, R. Snapshot: timescales in cell biology. *Cell* **164**, 1302–1302 (2016).

[61] Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI open* **3**, 111–132 (2022).

[62] Tran, D. *et al.* Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications* **12**, 1029 (2021).

[63] Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology* **23**, 303–304 (2022).

[64] 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).

[65] Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).

[66] Kidger, P. On neural differential equations. *arXiv preprint arXiv:2202.02435* (2022).

[67] 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**, eabq3745 (2022).

[68] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013).

[69] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[70] Feydy, J. *et al. Interpolating between optimal transport and mmd using sinkhorn divergences*, 2681–2690 (PMLR, 2019).

[71] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology* **20**, 296 (2019).

[72] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

[73] Yuan, W. *et al.* Temporally divergent regulatory mechanisms govern neuronal diversification and maturation in the mouse and marmoset neocortex. *Nature Neuroscience* **25**, 1049–1058 (2022).

[74] Danese, A. *et al.* Episcanpy: integrated single-cell epigenomic analysis. *Nature Communications* **12**, 5228 (2021).

[75] Lange, M. *et al.* Zebrahub–multimodal zebrafish developmental atlas reveals the state-transition dynamics of late-vertebrate pluripotent axial progenitors. *BioRxiv* 2023–03 (2023).

[76] Ma, P. *et al.* Joint profiling of gene expression and chromatin accessibility during amphioxus development at single-cell resolution. *Cell Reports* **39** (2022).

[77] Xiong, L. *et al.* Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature Communications* **13**, 6118 (2022).

[78] Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41–50 (2022).

[79] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

[80] Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances (2007).

[81] Ramos-Carreño, C. & Torrecilla, J. L. dcor: Distance correlation and energy statistics in Python. *SoftwareX* **22** (2023). URL https://www.sciencedirect.com/science/article/pii/S2352711023000225.

[82] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A next-generation hyperparameter optimization framework*, 2623–2631 (2019).

[83] Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* 1–23 (2023).

[84] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).