Synthetic Cell Differentiation Trajectory Generation: Generative Models vs. Foundation Models

Samia Islam

December 14, 2025

1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a technique that measures the transcriptomic profile of thousands to millions of individual cells. Thus each cell is represented as a vector of gene expression levels. Cell differentiation is a process in which a progenitor cell gradually changes into a specialized cell type through regulated shifts in gene expression. This dynamic process can be represented as a cell differentiation trajectory — a continuous path in gene expression space that connects one cellular state to another.

Generative models have long been used in single-cell analysis to learn low-dimensional representations of cellular states and to sample new cells from these learned distributions.

More recently, foundation models trained on large-scale single-cell datasets have emerged as powerful tools for representation learning and downstream prediction tasks. These models have demonstrated strong performance in tasks such as cell type annotation and perturbation modeling, but their ability for explicit cell differentiation trajectory generation remains less well understood.

Despite the increasing use of both generative models and foundation models in single-cell research, there has been limited systematic comparison of these approaches for the specific task of synthetic cell differentiation trajectory generation. So in this project, I performed a comparative analysis of four representative models: scVI [2], scNODE [7], scDiffusion [3], and scGPT [1]. I evaluated their ability to generate synthetic differentiation trajectories across three datasets of three biological contexts: epithelial–mesenchymal transition, hematopoiesis, and thymocyte development. I assessed trajectory smoothness, marker gene behavior, and biological plausibility using qualitative visualization and quantitative metrics.

The code is available at: https://github.com/SamiaShashmi/fm-project.

2 Datasets

I evaluated synthetic cell differentiation trajectory generation across three biologically distinct single-cell RNA sequencing datasets: epithelial–mesenchymal transition (EMT), hematopoiesis, and thymocyte development. All datasets were pre-

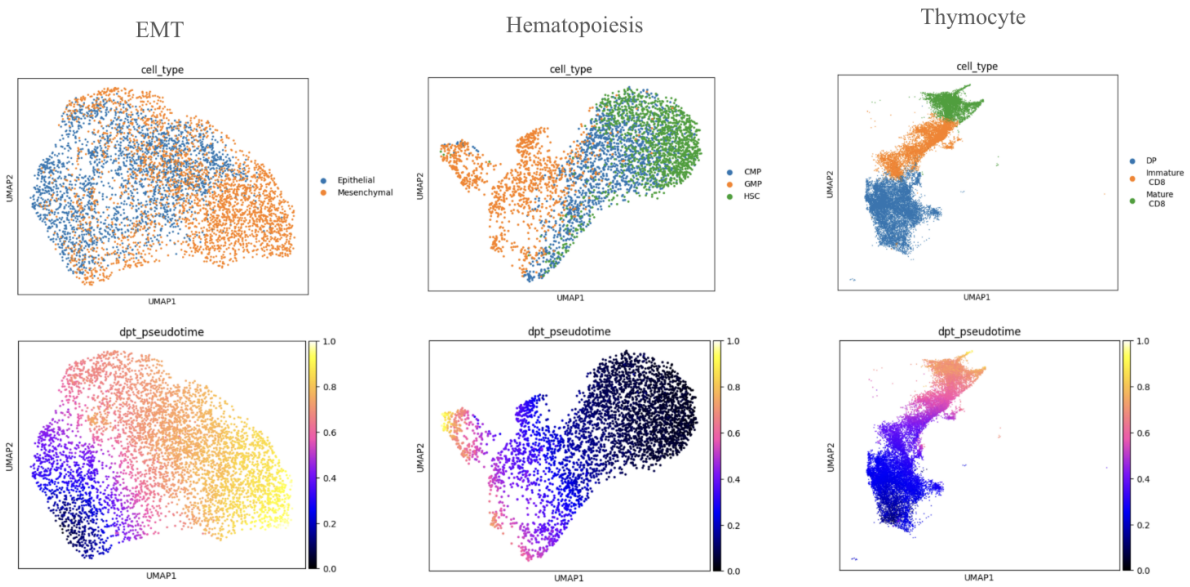


Figure 1: Upper column: The UMAP of the three dataset with hteir class labels. Lower column: The UMAP of three dataset with the pseudotime ordering of the cells

processed using standard single-cell analysis workflows, including quality control filtering, normalization, selection of highly variable genes. I picked top 2000 highly variable genes (featues) for each dataset. As the datasets are provided as discrete snapshots without true time labels, I sorted each dataset along a pseudotime from 0 to 1 to mimic a biologically possible trajectory from the available cells. The UMAP of the three dataset with their class labels and pseudotime mapping can be seen Figure 1.

The EMT dataset [4] contains two cell populations corresponding to Epithelial and Mesenchymal states. After preprocessing, the dataset contained 5,027 cells. Biologically, Epithelial cells undergo EMT (Epithelial to Mesenchymal Transition) by progressively losing epithelial characteristics and acquiring Mesenchymal features. So a single trajectory was constructed from Epithelial to Mesenchymal cells.

The hematopoiesis dataset [5] initially includes cells annotated into 11 distinct cell types. For this project, I limited the span and picked cell types of HSC, CMP and GMP. After preprocessing and subsetting, the dataset consisted of 3,870 cells. Biologically, HSC cells transition into CMPs, then CMPs further differentiate into GMPs. So a multi-stage trajectory was constructed following the differentiation path HSC $\rightarrow$ CMP $\rightarrow$ GMP.

The thymocyte dataset [6] comprises 21 annotated T cell developmental states. From this dataset, three CD8-related stages were selected: DP, immature CD8, and mature CD8 cells. After preprocessing, the thymocyte dataset contained 19,744 cells. Biologically, DP (double positive) cells can differentiate into CD8 cells, but before there is a intermediate stage of immature CD8s. That's why a developmental trajectory was constructed from DP to immature CD8 and finally to mature CD8 cells (DP $\rightarrow$ Immature CD8 $\rightarrow$ Mature CD8).

Table 1: Dataset Summary

Dataset	#Classes	#Sample	#Features	Trajectory
EMT [4]	2	5,027	2,000	Epithelial → Mesenchymal
Hematopoiesis [5]	3	3,870		HSC → CMP → GMP
Thymocyte [6]	3	19,744		DP → Immature CD8 → Mature CD8

3 Methodology

3.1 Models

I compared four representative approaches for synthetic single-cell data generation and trajectory construction, two are generative models, one is foundation model, one is hybrid: scVI, scNODE, scDiffusion, and scGPT.
scVI [2] is a variational autoencoder–based generative model that learns a probabilistic latent representation of single-cell gene expression.
scNODE [7] extends the VAE framework by explicitly modeling cell state evolution over continuous time. It integrates a variational autoencoder with neural ordinary differential equations to learn smooth latent dynamics governing cellular transitions.
scDiffusion [3] is a hybrid model that combines a diffusion-based generative framework with a pretrained foundation model. In scDiffusion, a foundation model is used as an autoencoder to learn a unified latent representation of gene expression, and a latent diffusion model operates in this space to generate synthetic cells through an iterative denoising process.
scGPT [1] is a transformer-based foundation model pretrained on large-scale single-cell datasets. Unlike task-specific generative models, scGPT is designed to learn general-purpose gene and cell representations across diverse biological contexts.

Model	Type	Core Architecture
scVI [2]	Generative Model	Variational Auto-Encoder (VAE)
scNode [7]	Generative Model	VAE + Neural ODE
scDiffusion [3]	Hybrid Generative Foundation Model	Foundation autoencoder + diffusion
scGPT [1]	Foundation Model	Transformer (GPT-like)

3.2 Trajectory Generation

3.2.1 scVI

scVI was used to learn a continuous latent representation of scRNA-seq data, and a trajectory was generated by linearly interpolating between latent embeddings of two endpoint cells and decoding the interpolated latent path back into gene expression space. The same procedure was applied across all three datasets, only the endpoint cell populations differed.

For each dataset, biologically meaningful endpoints were determined by the ordering of cells according to pseudotime. The intended progression can be understood visually from the UMAPs in Figure 1. I determined the start state latent vector by the first cell in the dataset, and the end state latent vector by the last cell in the dataset.
For each dataset, scVI was trained where the dimesion of the latent space is 512. After training, each real cell was embedded into the learned latent space. Then I created a latent trajectory of fixed length $T = 200$ by linearly interpolating between the endpoint embeddings.
As I used the built-in python package of scvi-tools directly, and initially I found no way to import the decoder module of scvi from the package, I created a separate supervised decoder by myself and trained to map scVI latent embeddings of 512 dimensions back to gene expression of 2000 dimensions. Finally, synthetic gene expression along the trajectory of 200 cells was generated by applying the trained decoder to each interpolated latent state.

3.2.2 scNODE

scNODE was used to generate a continuous synthetic differentiation trajectory by learning a nonlinear latent representation of gene expression and a continuous-time latent dynamics model parameterized by a neural ODE.
scNODE learns a latent space of dimension 50 using a variational encoder-decoder (VAE) component. To model differentiation as a continuous trajectory, scNODE learns latent dynamics governed by a neural ODE drift function: $\frac{dz(t)}{dt} = f_\omega(z(t), t)$, where f_ω is a neural network. In the implementation, ODE integration is performed using the Euler method. Trajectory generation begins from a set of initial cells sampled from the first pseudo-timepoint population (the start state). The trained model propagates these initial latent states forward across the pseudotime grid. The predicted latent trajectories are decoded into gene expression using scNODE's learned decoder. To make the synthetic trajectory length consistent, I uniformly subsampled 200 states along the generated sequence.

3.2.3 scDiffusion

scDiffusion was used to generate synthetic differentiation trajectories by operating in a learned latent space and producing intermediate latent states via classifier-guided generation.
scDiffusion operates in a latent space of dimension 128. The autoencoder of scDiffusion which is based on a pretrained single-cell foundation model is used to embed gene-expression profiles into a latent space. scDiffusion leverages SCimilarity, an encoder-decoder foundation model trained on a large corpus of 22.7 million cells from 399 single cell studies [3]. I fine-tuned the autoencoder on each of the three datasets to match the dataset-specific gene set and expression distribution.
I trained a condition classifier on the latent representations with the class label as the target to enable endpoint-controlled generation. The trained classifier provides gradients that guide generation toward user-specified endpoint conditions during sampling.
Trajectory generation is done by the reverse diffusion process multiple times under

classifier guidance. During the final portion of the reverse process, a condition classifier provides gradients with respect to the latent state that steer sampling toward user-specified labels. To produce intermediate states between two endpoints, I applied Gradient Interpolation: at each interpolation index k, it combines guidance from the start and end conditions with different weights (e.g., γ_1 and γ_2), so that the generated samples are progressively shifted from the start condition toward the end condition across. I concatenated the latent samples generated at each k in order to get a continuous sequence of intermediate latent states.
Like before, to obtain a fixed-length trajectory for evaluation, I uniformly subsampled 200 points from the ordered latent sequence. Finally, the decoded expression trajectory is produced by passing the subsampled latent states through the decoder of the autoencoder.

3.2.4 scGPT

Finally, scGPT was used as a pretrained foundation model to compute fixed-length cell embeddings, and trajectories were generated by interpolating between two endpoint embeddings and decoding the interpolated path back into gene expression space with a supervised decoder.
Latent representations are obtained by embedding the expression matrix using a pretrained scGPT model checkpoint. Then a fixed-length latent trajectory of 200 states is generated by linear interpolation between the endpoint embeddings as I did for scVI. As scGPT does not provide any decoder to get the representations in original dimensions, I implemented a supervised multilayer perceptron decoder by myself to convert interpolated scGPT embeddings back into gene expression.

3.3 Description of Analysis

3.3.1 Qualitative

- **Visualization in gene space:** For each dataset and each model, the synthetic trajectory was merged with the original real dataset at the gene-expression level. Dimensionality reduction was then performed on the combined dataset using UMAP computed directly from gene-expression space.
- **Visualization in latent space:** To examine how trajectory structure depends on the learned representation, the merged real and synthetic data were also visualized in each model's latent space. I computed both the PCA on latent representations and made the plots.
- **Marker gene trends:** To assess biological plausibility at the gene level, expression trends of known marker genes were examined along each synthetic trajectory. I plotted the gene expression vs cell ordering of each trajectory to assess the trend. I picked the following marker genes for each type of cells:
  - **EMT**
    - Epithelial markers: `CDH1`, `CRB3`, `DSP`
    - Mesenchymal markers: `VIM`, `FN1`, `SNAI2`
  - **Hematopoiesis**

    - HSC markers: `THY1`, `KIT`, `GATA2`
    - CMP markers: `TFRC`, `KIT`, `IRF8`
    - GMP markers: `CEBPA`, `CEBPD`, `MPO`
  - **Thymocyte development**
    - DP markers: `RAG1`, `RAG2`, `DNTT`
    - Immature markers: `TRAC`, `TRBC1`, `CD69`
    - Mature markers: `IL7R`, `CCR7`, `LST1`

3.3.2 Quantitative

- **Pseudotime distance error:** The pseudotime distance error measures how well the ordering of synthetic cells along a generated trajectory aligns with the intrinsic ordering of real cells inferred from pseudotime. Let $X_{\text{real}} \in \mathbb{R}^{N \times G}$ denote real cells with associated pseudotime values $\tau_{\text{real}} \in \mathbb{R}^N$, and let $X_{\text{syn}} \in \mathbb{R}^{T \times G}$ denote the synthetic trajectory with an intrinsic trajectory coordinate $\tau_{\text{syn}} \in [0,1]^T$. For each synthetic cell $x_{\text{syn}}(t)$, its nearest real neighbor in gene-expression space is identified:

$$i^*(t) = \arg\min_i \|x_{\text{syn}}(t) - x_{\text{real}}(i)\|_2.$$

  The pseudotime distance error is then computed as the average absolute deviation between synthetic trajectory coordinates and matched real pseudotime values:

$$\mathcal{E}_{\text{pseudo}} = \frac{1}{T} \sum_{t=1}^{T} \left| \tau_{\text{syn}}(t) - \tau_{\text{real}}(i^*(t)) \right|.$$

- **Distance-to-manifold smoothness:** The distance-to-manifold smoothness metric evaluates whether synthetic trajectories remain close to the data manifold defined by real cells, it also penalizes abrupt changes between consecutive synthetic states. Let $d(t)$ denote the distance from synthetic cell $x_{\text{syn}}(t)$ to the real-cell manifold:

$$d(t) = \min_i \|x_{\text{syn}}(t) - x_{\text{real}}(i)\|_2.$$

  Smoothness is assessed by measuring the variability of these distances along the trajectory. Specifically, the metric is defined as the standard deviation of $d(t)$:

$$\mathcal{S}_{\text{manifold}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(d(t) - \bar{d} \right)^2}, \quad \bar{d} = \frac{1}{T} \sum_{t=1}^{T} d(t).$$

- **Marker gene monotonicity score:** The marker gene monotonicity score quantifies whether predefined marker genes show expected directional trends along the synthetic trajectory. Let g index a marker gene, and let $x_g(t)$ denote its expression value in synthetic cell t. For each marker gene, the Spearman rank correlation between gene expression and trajectory coordinate is computed:

$$\rho_g = \text{corr}_{\text{Spearman}}(x_g(t),\, \tau_{\text{syn}}(t)).$$

Each marker gene is associated with an expected direction of change $s_g \in \{-1, +1\}$, where -1 indicates decreasing expression and $+1$ indicates increasing expression along the trajectory. The monotonicity contribution of gene g is then defined as:

$$m_g = s_g \cdot \rho_g.$$

The final monotonicity score is obtained by averaging across all marker genes:

$$\mathcal{M} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} m_g.$$

For two-state trajectories (e.g., EMT), expected directions are set to $(-1, +1)$ for early and late-state markers. For three-stage trajectories (Hematopoiesis and Thymocyte), expected directions are specified as $(-1, +1, +1)$, corresponding to early, intermediate, and late marker groups.

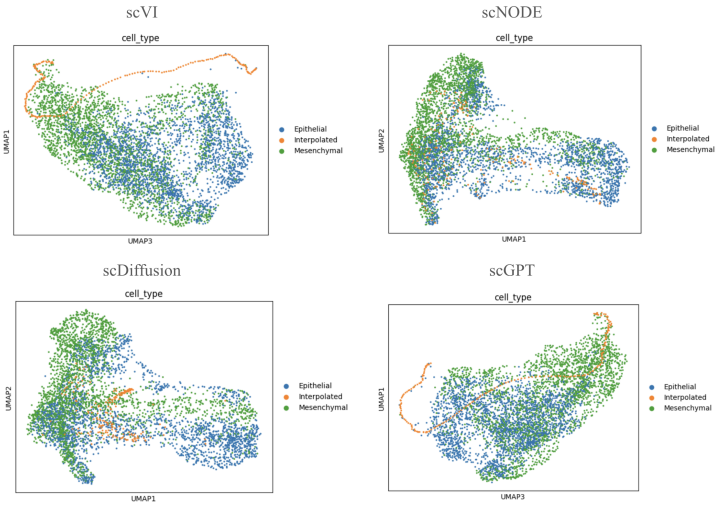


Figure 2: Synthetic trajectories generated by scVI, scNODE, scDiffusion, scGPT overlaid on original EMT data

4 Results and Analysis

4.1 Qualitative: Visualization in gene space

4.1.1 EMT Dataset

Figure 2 shows UMAP projections of the EMT dataset with synthetic trajectories generated by the four models overlaid on real epithelial and mesenchymal cells. The scVI-generated trajectory forms a smooth, continuous curve, but it largely lies along the outer boundary of the real data manifold rather than passing through its interior. But the trajectory does connect epithelial-dominated regions to mesenchymal-dominated regions. This suggests that linear interpolation in scVI's latent space produces a geometrically smooth path, but one that does not necessarily follow the intrinsic structure of the EMT manifold learned from the data.
The scNODE trajectory is embedded within the main body of the data cloud and

overlaps extensively with real cells across the epithelial–mesenchymal continuum. We can see a trajectory that follows the internal geometry of the data rather than a boundary path.
The scDiffusion trajectory is more dispersed and less visually coherent as a single continuous path. Synthetic points are congested at a specific region. This reflects the stochastic nature of diffusion-based generation.
The scGPT-generated trajectory appears as a smooth, well-defined curve that spans from epithelial-rich regions toward mesenchymal-rich regions, and passes through intermediate areas of the data manifold. Compared to scVI, the scGPT trajectory overlaps more with the interior of the data cloud. This suggests that the pretrained foundation embeddings capture a meaningful global organization of EMT states, even though trajectory continuity is imposed.

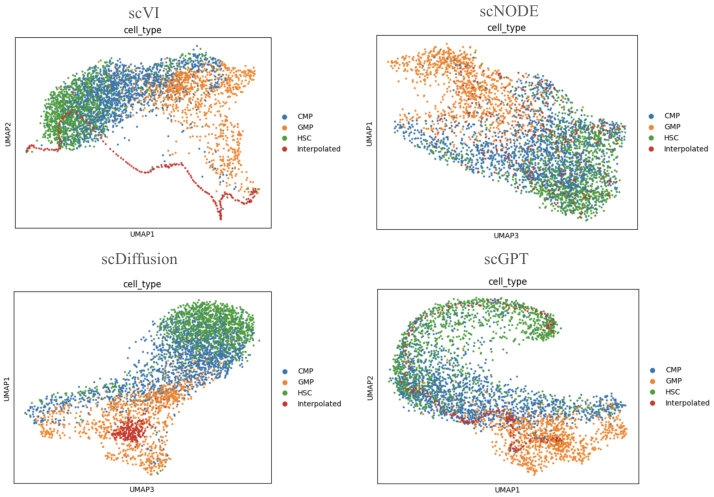


Figure 3: Synthetic trajectories generated by scVI, scNODE, scDiffusion, scGPT overlaid on original Hematopoiesis data

4.1.2 Hematopoiesis Dataset

Figure 3 shows UMAP projections of the hematopoiesis dataset with synthetic trajectories. The scVI-generated trajectory forms a smooth, continuous curve but deviates significantly from the main data manifold. I think as linear interpolation is the shortest path, the synthetic trajectory tries to connect the shortest path between the two endpoints, that's why it avoids the main data cloud.
The scNODE trajectory is well integrated into the real data distribution. Synthetic points overlap with HSC-enriched regions at early stages, pass through areas populated by CMP cells, and extend toward GMP-dominated regions. But we can not see a smooth linear trajectory like scVI.
The scDiffusion-generated trajectory concentrates around the GMP region, synthetic points formed a dense cluster rather than a clearly ordered path spanning all three stages.
The scGPT trajectory gives a very nice curved path through the hematopoiesis manifold. Synthetic points overlap with HSC regions at one end, transition through CMP-dominated areas, and reach GMP-enriched regions at the other end. Compared to scVI, the trajectory remains closer to the interior of the data manifold and shows clearer alignment with known lineage progression.

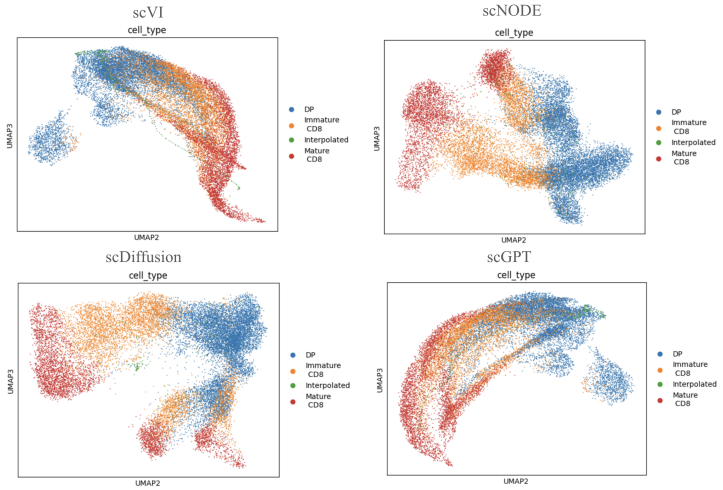


Figure 4: Synthetic trajectories generated by scVI, scNODE, scDiffusion, scGPT overlaid on origonal Thymocyte data

4.1.3 Thymocyte Dataset

Figure 4 shows UMAP projections of the Thymocyte dataset with synthetic trajectories. As the dataset is very large compared to the synthetic trajectory of only 200 cells, it is a bit difficult to observe the trajectory.
The scVI-generated trajectory follows a smooth curve from the DP region toward the mature CD8 region. The interpolated points slightly deviate from the high-density regions of real cells.
The scNODE trajectory is well integrated within the real data distribution. From the UMAP, we can see that the differentiation has two branches, and the synthetic trajecotory is following one of them.
Regarding scDiffusion generated, the synthetic cells are clustered at a specific region beside Immature CD8 cells. Diffusion-based sampling does not consistently enforce a smooth, stage-wise progression in this dataset.
The scGPT trajectory forms a broad, continuous arc that closely follows the overall shape of the thymocyte manifold.

4.2 Qualitative: Visualization in latent space

4.2.1 EMT Dataset

Figure 5 shows PCA projections of the latent representations learned by scVI, scNODE, scDiffusion, and scGPT, with synthetic trajectory points overlaid on real EMT data.
In the scVI latent space, epithelial and mesenchymal cells approximate a Gaussian distribution. The interpolated trajectory appears as a short linear segment located near the border of this cloud, rather than spanning the major direction of variation between epithelial and mesenchymal states. This means that linear interpolation between endpoint embeddings in scVI latent space does not strongly correspond to the principal EMT axis captured by the model.
For scNODE, the trajectory is not linear or smooth as it is not linear interpolation. The generated points by latent ode overlap with regions populated by both

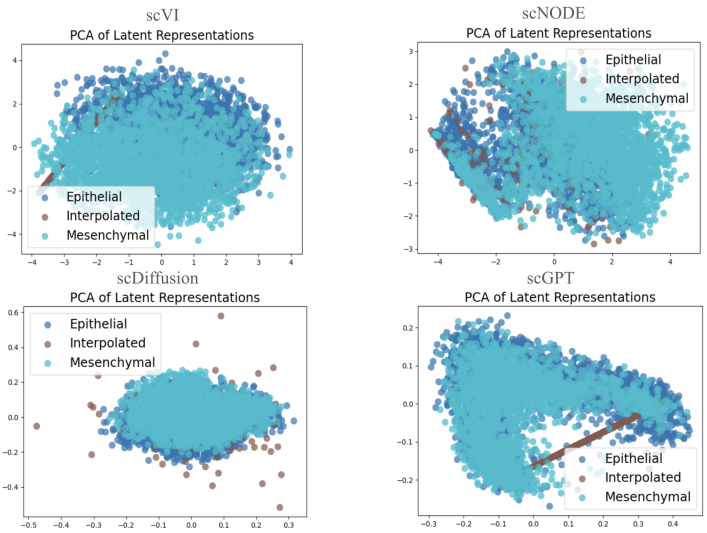


Figure 5: Latent Representation of synthetic trajectory and original EMT data

epithelial and mesenchymal cells, which indicaties that the learned latent dynamics generate trajectories aligned with the principal axes of variation rather than arbitrary directions in latent space.
In the scDiffusion latent PCA, real cells are tightly clustered. Synthetic points are scattered around this dense core and do not form a clear directional path across latent space. May be the reason is that scDiffusion captures local similarity between cells but does not strongly encode a global EMT ordering in latent space.
For scGPT, we can see that the latent representation of the data did not form a uniform Gaussian distribution. That's why the linear inteprolated trajectory lying partially outside of the data cloud. Yet it get to generate a valid synthetic EMT trajectory.

4.2.2 Hematopoiesis Dataset

Figure 6 shows PCA projections of latent representations learned by scVI, scNODE, scDiffusion, and scGPT, with synthetic trajectory points overlaid.
In the scVI latent space, HSC, CMP, and GMP cells are highly overlapping, form a largely isotropic cloud with no clear separation. The synthetic trajectory appears as a nearly vertical line positioned at the edge of the latent distribution, rather than spanning the central region occupied by real cells. That's why I found a trajectory that does starts with HSC and ends with GMP, but does not follow the actual HSC to CMP, CMP to GMP path.

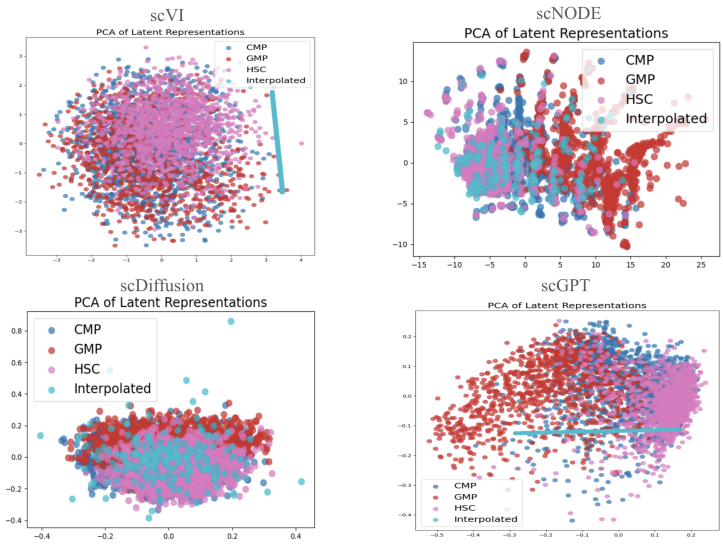


Figure 6: Latent Representation of synthetic trajectory and original Hematopoiesis data

The scNODE latent space shows an interesting structure, where the real cells distributed along an extended axis that separates early and late cell states. The synthetic trajectory aligns with this dominant axis, passes through regions populated by intermediate cells. This alignment suggests that scNODE's learned continuous-time dynamics capture the principal directions of hematopoietic differentiation in latent space.
In the scDiffusion latent space, real cells form a compact, overlapping cluster with limited separation between classes. Synthetic points are dispersed around this cluster rather than forming a clearly ordered path.
For scGPT, HSC, CMP, and GMP cells do not form any compact Gaussian like distribution, rather they produce a smooth gradient of regions of latent space, which is consistent with known lineage relationships. The synthetic trajectory follows this gradient closely, and forms a smooth path that traverses through HSC, CMP, GMP regions.

4.2.3 Thymocyte Dataset

The latent representation of the models for Thymocyte dataset is in Figure 7. As the number of samples in this dataset is very big compared to the synthetic trajectory size, the synthetic trajectory is difficult to observed. But if we look closely, we will get the similar findings as we got from the previous datasets.

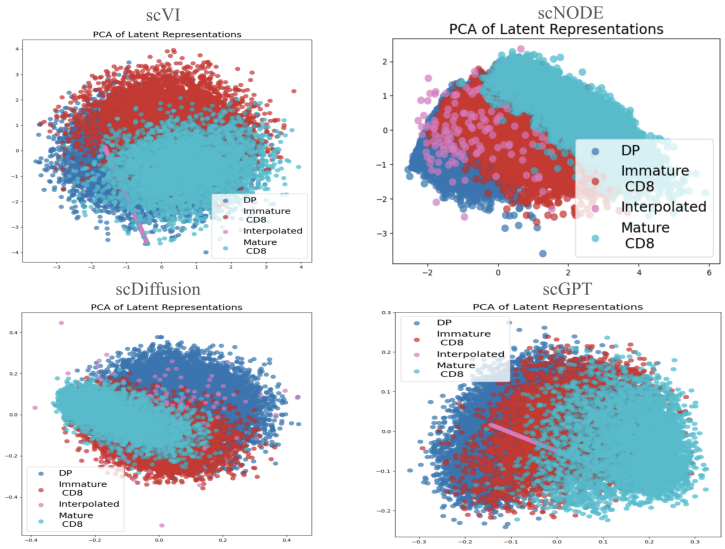
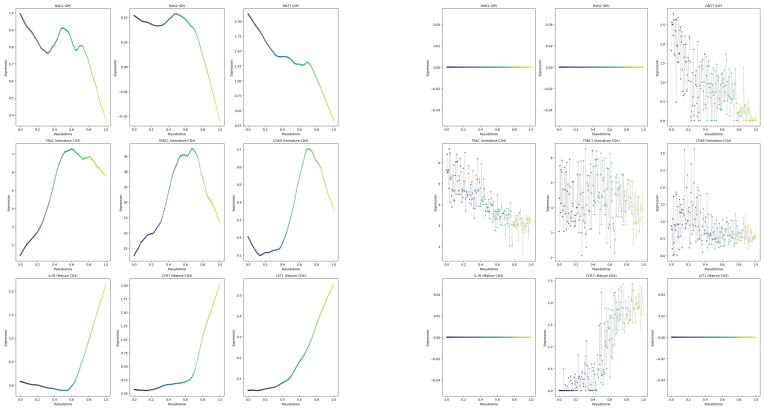


Figure 7: Latent Representation of synthetic trajectory and original Thymocyte data

4.3 Qualitative: Marker gene trends

I got similar results for EMT and Thymocyte for the models, I got some exceptions for Hematopoiesis, the reason is the marker genes for hematopoiesis is not biologically correlated with the trajectory progression. Or may be I picked wrong sets of genes as markers. Here I will only present the marker gene trends of Thymocyte data. Other datasets' plot are available in the corresponding python notebooks with the code.

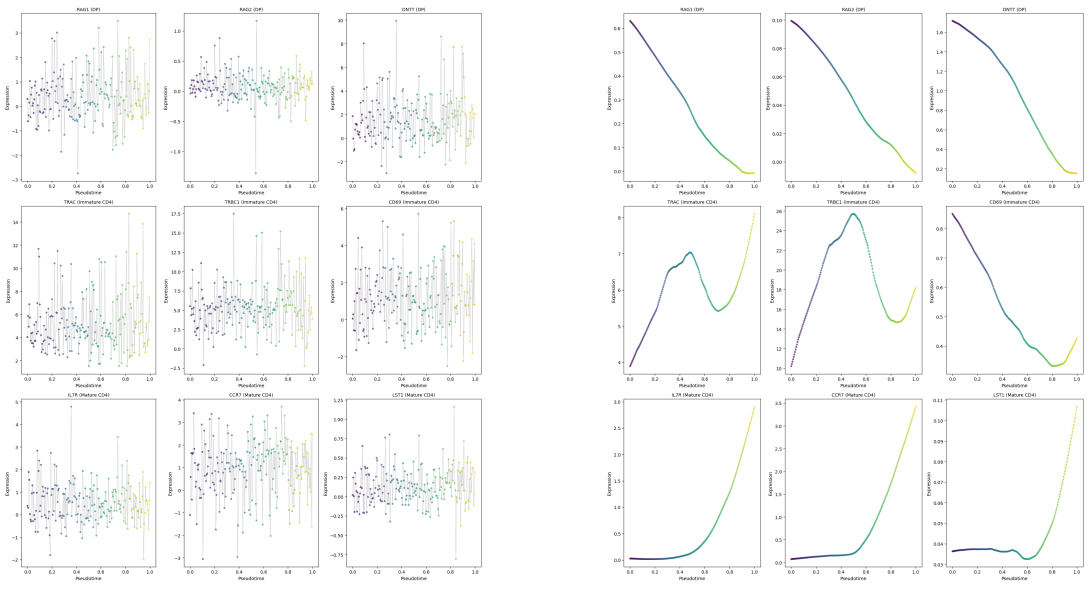
The marker gene trends of the synthetic trajectory generated by scVI is in Figure 8a. We can see that along the trajectory, as the cells converts from DP cells, its marker trends also goes down smoothly. For the intermediate cell Immature CD8, the marger genes' expression rises at the middle of the trajectory, then goes down. For the last state Mature CD8 cells, the marker genes' expression goes up along the trajectory.



(a) scVI (b) scNODE

Figure 8: Marker Gene Trends of Synthetic Trajectory. Upper row: markers of DP cells, Middle row: markers of Immature CD8 cells, Lower row: markers of Mature CD8 cells

The marker gene trends of the synthetic trajectory generated by scNODE is in Figure 8b. Here we can see that the expression of some of the markers are flat zero, I am not quite sure about the reason. But other markers show reasonable trend, though there is some noise and fluctuations compared with scVI. May be the reason behind this is, in scNODE, the trajectory generation relies on integrating learned latent dynamics from sampled initial conditions. As it is not linear interpolation, so there is no smooth trend. But overall, the marker for the first state shows a decreasing trend, the markers for the intermediate state show a slight rise at the intermediary points, the marker trend of the last state goes up.
For the gene trends of trajectory of scDiffusion in Figure 9a, we can see only noise. The reason is each synthetic point is produced through a random denoising process guided by class labels rather than by a deterministic path through latent space. So there is high variability and weaker monotonic trends in individual gene expression.
For scGPT in Figure 9b, all DP-associated markers show smooth, monotonic decreases along trajectory. Markers of the immature CD8 stage show non-monotonic but structured behavior. TRAC and TRBC1 increase from early pseudotime, peak in the middle of the trajectory, and another rise toward the mature end. CD69 shows a gradual decline with a shallow rise. The genes for Mature CD8 again show monotonic increase.
According to my understanding, we see different trends for scVI and scGPT because though the start and the end points for the trajectories are same, they captured different trajectory while doing linear interpolation. As scVI and scGPT have different latent representation, the linear interpolation captured different paths.



(a) scDiffusion (b) scGPT

Figure 9: Marker Gene Trends of Synthetic Trajectory. Upper row: markers of DP cells, Middle row: markers of Immature CD8 cells, Lower row: markers of Mature CD8 cells

4.4 Quantitative: Pseudotime Distance Error (↓)

Table 2 shows the pseudotime distance error of the four models for the three datasets. For the EMT dataset, scGPT achieves the lowest error, its synthetic trajectory most closely matches the continuous epithelial–mesenchymal progression. scVI and scNODE perform comparably, but scDiffusion shows the highest error, which means weaker preservation of global pseudotime ordering in this transition.
In the hematopoiesis dataset, scVI shoes the lowest pseudotime distance error. In contrast, scNODE exhibits substantially higher error. In the qualitative analysis, we saw that scNODE didn't came up with a well-defined single trajectory, rather it scattered around the dataset. May be it is reflecting the variability introduced by continuous-time dynamics in a multi-stage setting. scDiffusion has also higher error as usual. scGPT shows moderate performance.
For the Thymocyte dataset, scNODE achieves the lowest error. scGPT and scVI both also have lower error. scDiffusion again shows the highest error, consistent with its weaker global ordering.

Table 2: Pseudotime Distance Error

	scVI	**scNODE**	**scDiffusion**	**scGPT**
EMT	0.198	0.188	0.33	**0.11**
Hematopoiesis	**0.132**	0.415	0.303	0.214
Thymocyte	0.18	**0.154**	0.273	0.168

4.5 Quantitative: Distance-to-Manifold Smoothness (↓)

Table 3 shows the distance to manifold smoothness of the four models for the three datasets. For the EMT dataset, scNODE achieves a significantly lower smoothness score than all other models, its generated trajectory stays tightly aligned with the EMT data manifold. scVI and scGPT surprisingly exhibit much larger values. I think the reason behind this is, this metric explains how good the model captures the distribution of the data, not how good is the trajectory. That's why scVI and scGPT is doing poor here as they captured lots of points outside the data manifold though they created smooth trajectory.
In the hematopoiesis dataset, scVI achieves the lowest scores. scGPT is also close to scVI. This indicates that both models generate trajectories that remain close to the real hematopoietic manifold when restricted to the selected HSC–CMP–GMP populations. scNODE shows higher variability in manifold distance, while scDiffusion again exhibits the least smooth behavior.
For the thymocyte dataset, scNODE shows the lowest distance score, followed closely by scGPT. Both models maintain relatively stable proximity to the real data manifold across the trajectory. scVI shows higher deviation, and scDiffusion consistently produces the least smooth trajectories.

4.6 Quantitative: Marker Gene Monotonicity Score

Table 4 reports the mean Spearman correlation between marker gene expression and trajectory coordinate for each class and dataset. Positive values indicate increasing expression along the trajectory, negative values indicate decreasing ex-

Table 3: istance-to-Manifold Smoothness

	scVI	**scNODE**	**scDiffusion**	**scGPT**
EMT	122.605	**22.984**	43.47	126.1
Hematopoiesis	12.523	28.93	40.803	**11.63**
Thymocyte	34.89	**28.93**	43.248	30.45

Table 4: Marker Gene Monotonicity Score

	Class	**scVI**	**scNODE**	**scDiffusion**	**scGPT**
EMT	Epithelial	-0.312	-0.525	0.02	-0.556
	Mesenchymal	0.457760	0.652	-0.078	0.969
Hematopoiesis	HSC	0.94	0.371	-0.021	0.874
	CMP	0.99	0.376	0.065	1
	GMP	0.37	-0.085	-0.007	0.25
Thymocyte	DP	0.014	-0.678	0.02	-0.999
	Immature CD8	-0.0627	-0.342	0.031	-0.148
	Mature CD8	-0.473	0.871	0.049	0.793

pression, and values near zero indicate weak or non-monotonic trends.
For the EMT dataset, scNODE and scGPT show strong negative correlations for epithelial markers and strong positive correlations for mesenchymal markers, which indicates clear and correctly oriented transcriptional transitions. scVI captures the expected directions but with weaker magnitude. scDiffusion exhibits correlations near zero, which suggests little consistent ordering of marker expression along the trajectory.
In the hematopoiesis dataset, scVI and scGPT achieve high positive correlations for HSC and CMP markers. Correlations for GMP markers are weaker across all models. scNODE shows reduced correlation strength compared to scVI and scGPT. scDiffusion again exhibits near-zero correlations, weak monotonic structure. As the markers does not show expected correlation, I think the trajectory from HSC to GMP is independent of the marker genes. There is no biological relation between the markers and the cell types for this case. Or the choice of the genes as markers may be not appropriate for this case.
For the thymocyte dataset, scGPT exhibits near-perfect negative correlation for DP markers and strong positive correlation for mature CD8 markers, demonstrating clear ordering from early to late developmental states. scNODE also captures the expected directionality but with lower magnitude and increased variability. scVI shows weaker and less consistent correlations. scDiffusion remains close to zero across all classes.

5 Summary and Conclusions

From the qualitative and quantitative analysis, I believe that all the trajectories generated by scVI, scNODE, and scGPT captured different but biologically plausible trajectories. I still have doubt with scDiffusion because almost none of the results showed convincing result. May be the way I used to generate trajectory is

not appropriate, as the task for generating cell trajectories is still new. Or may be scDiffusion is not suitable for continuous trajectory generation.
Overall, the results indicate that no single model dominates across all settings. Explicit dynamical modeling (scNODE) is advantageous for capturing continuous developmental processes, while pretrained foundation representations (scGPT) provide strong global organization and biologically meaningful interpolation. Simpler generative models like scVI can perform well in discretely staged trajectories but may fail to capture complex geometry. scDiffusion is good for generating realistic sample, but it is yet not prepared to generate continuous trajectory. Together, these results motivate further research into trajectory-aware models as a core component of single-cell analysis pipelines.

References

[1] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

[2] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[3] Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024.

[4] José L McFaline-Figueroa, Andrew J Hill, Xiaojie Qiu, Dana Jackson, Jay Shendure, and Cole Trapnell. A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nature genetics*, 51(9):1389–1398, 2019.

[5] Danilo Pellin, Mariana Loperfido, Cristina Baricordi, Samuel L Wolock, Anita Montepeloso, Olga K Weinberg, Alessandra Biffi, Allon M Klein, and Luca Biasco. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature communications*, 10(1):2395, 2019.

[6] Zoë Steier, Dominik A Aylard, Laura L McIntyre, Isabel Baldwin, Esther Jeong Yoon Kim, Lydia K Lutes, Can Ergen, Tse-Shun Huang, Ellen A Robey, Nir Yosef, et al. Single-cell multiomic analysis of thymocyte development reveals drivers of cd4+ t cell and cd8+ t cell lineage commitment. *Nature immunology*, 24(9):1579–1590, 2023.

[7] Jiaqi Zhang, Erica Larschan, Jeremy Bigness, and Ritambhara Singh. scnode: generative model for temporal single cell transcriptomic data prediction. *Bioinformatics*, 40(Supplement_2):ii146–ii154, 2024.