GENERATIVE MODELING OF SPATIAL TRANSCRIPTOMICS VIA GAUSSIAN MIXTURE FLOW MATCHING

Lisa Schunke, Soroor Hediyeh-zadeh, Sergio Marco Salas & Ali Oğuz Can
Institute of Computational Biology
Computational Health Center
Helmholtz Munich
Munich, Germany
{lisa.schunke,soroor.hediyehzadeh,alioguz.can,sergio.salas}@helmholtz-munich.de

Fabian J. Theis
Institute of Computational Biology
Computational Health Center
Helmholtz Munich, Munich, Germany
TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
fabian.theis@helmholtz-munich.de

ABSTRACT

Spatial transcriptomics enables the joint analysis of cellular gene expression and spatial organization, offering new insights into tissue development and function. Existing graph- and embedding-based methods capture spatial patterns but remain descriptive and non-generative. Spatial transcriptomics is still an expensive and resource-intensive assay, limiting its widespread application. Consequently, there is a growing need for generative models capable of accurately simulating gene expression within spatial contexts, particularly in settings where experimental data acquisition is impractical or cost-prohibitive. In this work, we revisit the generative modeling of spatially informed cell embeddings derived from gene expression and spatial information and apply a Gaussian Mixture Flow model (GMFlow) to explicitly model the multinomial nature of cell type distributions during generation. Using a mouse embryonic development dataset, we find that both single-Gaussian conditional flow and GMFlow models learn cell type distributions over different developmental timepoints with comparable accuracy, while GMFlow results in generations that are closer to ground truth cell representations, suggesting that GMFlow may enable more realistic simulation of cellular landscape in spatial transcriptomics data.

1 INTRODUCTION

Cells are the fundamental unit of life, assembling into complex tissues whose structure underlies function. Although all cells share the same genome, their roles are heterogeneous, shaped by their molecular profiles and their spatial localization. Understanding how cells organize and interact is therefore crucial for elucidating developmental processes, tissue homeostasis, and disease mechanisms. Advances in spatial transcriptomics Marx now enable the simultaneous measurement of both a cell's RNA composition and its spatial position, providing unprecedented opportunities to dissect the principles governing tissue organizationMoses & Pachter.

Spatial transcriptomics has primarily been modeled using graph- and embedding-based approaches, which represent cells and their local neighborhoods in low-dimensional spaces Hu et al.; Dong & Zhang. While useful, these methods are descriptive, summarizing observed patterns but not explicitly modeling tissue organization. Graph neural networks (GNNs) address some of these limitations by more effectively integrating information from a cell's local spatial neighborhood, allowing for richer modeling of interactions between cell populations Long et al.; Lyu et al.; Xu et al.. These

tools provide powerful tools for identifying spatial domains and characterising cellular heterogeneity. However, they remain descriptive representations of observed tissue and do not provide a generative model capable of simulating new cellular states or predicting unseen developmental stages.

Generative models aim to overcome these limitations by learning the joint distribution of cellular states and spatial organization, enabling the simulation of realistic tissue landscapesLopez et al.; Biancalani et al.. However, these approaches typically rely on unimodal latent priors or external reference mappings and do not explicitly model multimodal developmental transitions. Flow matching Lipman et al.; Tong et al. offers a powerful alternative by learning a continuous vector field that transports a simple prior to the data distribution. Yet, standard flow matching assumes a unimodal (Gaussian) conditional velocity distribution, limiting its ability to capture complex biological processes and typically lacking explicit spatial conditioning.

In this work, we combine spatial representation learning with flow-based generative modeling. We first obtain spatially informed cell embeddings using the Spatially Embedded Deep Representation (SEDR) framework Xu et al., which integrates gene expression profiles and spatial neighborhood information into a low-dimensional latent space. Rather than modeling high-dimensional gene expression directly, our generative model operates on these spatially aware cell embeddings. On top of this representation, we train flow-based generative models to learn the distribution of cellular states across developmental timepoints and to generate new cells for unseen stages. To better capture the multimodal structure of developmental cell populations, we extend standard conditional flow matching by employing Gaussian Mixture Flow Matching (GMFlow) Chen et al. (b), which models the conditional velocity distribution as a mixture of Gaussians rather than a single Gaussian.

Our contributions are as follows:

- We propose a generative model of spatial transcriptomic data based on Gaussian Mixture Flow Matching (GMFlow) applied to spatially aware cell embeddings derived from gene expression and spatial information
- Our proposed model explicitly models cell type distributions in the data-generating model of gene expression
- We benchmark against conditional flow matching (CFM) and investigate differences in cell type distributions in generated cells.

2 METHODS

2.1 GAUSSIAN MIXTURE FLOW MATCHING

To model the complex distribution of cell features, we employ a generative approach based on recent advances in flow matching. Specifically, we adopt the GMFlow model Chen et al. (b), a powerful extension of standard flow matching that offers greater expressive power for capturing multimodal data distributions, in our case, cell-type distributions over time during embryonic development.

Flow matching models learn a vector field that transports samples from a simple prior distribution (e.g., Gaussian noise) to a complex data distribution. The key limitation lies in the assumption that the conditional velocity $p(u|x_t)$ distribution learned by the model is a simple Gaussian. For complex biological data, such as cell features in our study, this assumption is overly simplistic and fails to capture the potential for multiple valid pathways in the generative process.

The GMFlow model addresses this limitation by modeling the conditional velocity distribution $p(u|x_t)$ not as a single Gaussian, but as a more expressive Gaussian Mixture Model (GMM). The predicted distribution $q_\theta(u|x_t)$ is defined as $q_\theta(u \mid x_t) = \sum_{k=1}^{K} A_k \mathcal{N}(u;\, \mu_k,\, \Sigma_k)$ with x_t being the noisy data point and K the number of Gaussian. Instead of predicting a single velocity vector, the neural network learns to predict the parameters of this mixture: the means (μ_k), covariances $(\sum_k)$, and mixture weights (A_k) for K distinct Gaussian components. This allows the model to represent a rich, multimodal distribution of potential velocities at every step of the generation process. The model is trained by minimising the Kullback-Leibler (KL) divergence between the predicted GMM $q_\theta(u|x_t)$ and the true (but intractable) velocity distribution $p(u|x_t)$. This is equivalent to minimising the negative log-likelihood of the true velocity u under the predicted GMM, leading to the following

loss function $\mathcal{L} = \mathbb{E}_{t,x_0,x_t}[-log\ q_\theta(u|x_t)]$. Expanded, this loss can be seen as a hybrid of regression (distance to means) and classification (mixture weights) Chen et al. (b).

2.2 Conditional Flow Matching

As a baseline comparison to our GMFlow approach, we adopt conditional flow matching (CFM) in its simplest form, following the formulation introduced by Tong et al.. CFM provides a simulation-free objective for training continuous normalising flow by regressing a time-dependent vector field to a known conditional velocity. During training, the model is conditioned on a randomly sampled noise-data pair. Given this conditioning, the model learns to predict the instantaneous velocity that transports a point along a simple interpolation between the source and target samples. Training is performed using a mean-squared error regression loss, where the predicted velocity field is matched to a known conditional target velocity derived from the sampled noise-target pair. In our work, we use CFM with strictly independent noise-target sampling, without incorporating optional transport coupling. This setup provides a simple and well-established baseline for comparison with our GMFlow approach.

2.3 Dataset and Preprocessing

To evaluate whether the GMFlow model characterises cell-type distributions more accurately than standard baselines, we utilise the Mouse Organogenesis Spatiotemporal Transcriptomic Atlas (MOSTA) Chen et al. (a). This dataset provides a comprehensive spatiotemporal map of mouse development across 8 stages (E9.5 - E16.5) using Stereo-seq at cellular resolution. We treat each developmental time point as a distinct class, resulting in 8 condition classes. We held out the E13.5 time point from the training phase to serve as our unseen test set. The remaining cells from the other timepoints were split into training and validation sets for model optimisation.

2.4 Spatially-Aware Embeddings with the Spatially Embedded Deep Representation Framework

To generate robust, spatially-aware cell embeddings, we utilise the SEDR framework Xu et al.. For our analysis, we process each time point separately, subsampling to approximately 110,000 cells and filtering the expression data to the 2,000 most highly variable genes. The SEDR model learns a low-dimensional representation by simultaneously encoding transcriptomic profiles and spatial information. Its architecture has two main components: a masked autoencoder learns a gene-feature representation (Z_f). In contrast, a variational graph autoencoder (VGAE) learns a spatial embedding (Z_g) from a K-nearest-neighbour graph of cell coordinates Xu et al..

The final cell embedding (Z) is the concatenation of these two representations, which we set to a dimension of 32. All generative models in this work operate on these embeddings rather than on the original high-dimensional gene expression profiles. To mitigate batch effects arising from cells originating from different samplers at each time point, we employ the harmonisation strategy discussed in the SEDR setup. This provides a unified and denoised latent space for our downstream model representing gene expression and spatial patterns.

2.5 Implementation and Training Configuration

Both the CFM baseline and the GMFlow model were implemented using a 2-layer Multi-Layer Perceptron (MLP) backbone. The models utilise an input dimension of 32 (corresponding to the SEDR embedding size), a hidden dimension of 16, and an embedding dimension of 16. To encode the temporal dynamics of the denoising process (diffusion timesteps), we use a sinusoidal embedding. For the GMFlow model, we utilise 8 Gaussian components. While an initial target of 49 components was considered to match the annotated cell classes in the MOSTA dataset, empirical testing showed that the model achieved better stability and capacity utilisation with K=8. We employ the specialised ODE sampler for GMFlow, which predicts a full distribution over velocities $q_\theta(u|x_t)$ rather than a single mean velocity. While the framework supports custom SDE solvers, we observed no significant performance improvement for these embeddings and hypothesise that SDE-based sampling is more impactful for high-resolution image data than for tabular transcriptomic embeddings.

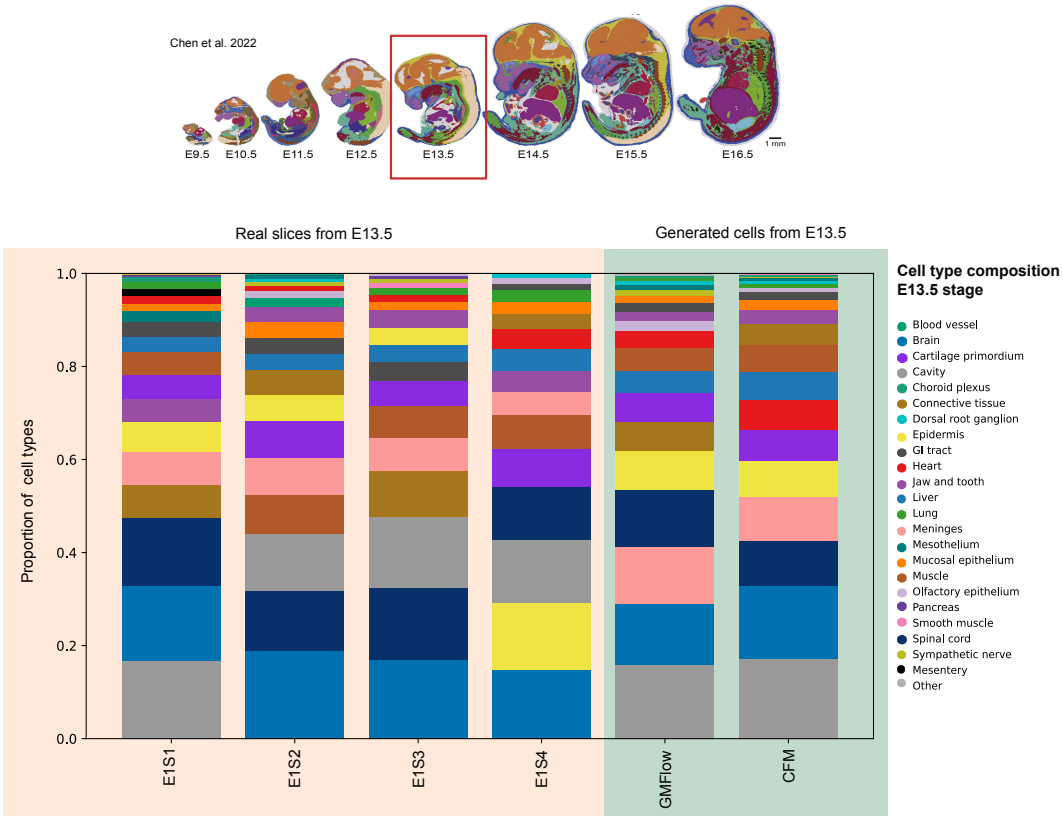

Figure 1: Cell type proportions in ground truth (orange box) and generated cells (green box) from the E13.5 stage of mouse embryonic development.

Table 1: Wasserstein distance between original and generated cells from E13.5 stage for both models (CFM and GMFlow). The lower the distance, the better.

	CFM		GMFlow	
Slice	Mean	Std	Mean	Std
E1S1	0.88	0.20	0.46	0.22
E1S2	0.93	0.20	0.51	0.22
E1S3	0.91	0.29	0.49	0.22
E1S4	0.88	0.20	0.46	0.22
Overall	**0.9**	0.20	**0.47**	0.22

3 Results

3.1 Cell-Type Distribution Analysis

To evaluate generative performance on the held-out E13.5 stage, we generated 5,000 cells from each model using 16 sampling steps (with an additional 16 subsampling steps for the GMFlow model) (Figure 1). We assigned the generated cells to biological cell classes using a nearest-neighbour approach based on the original cells in the E13.5 test set. Both models demonstrate a degree of similarity to the original tissue distributions from the original slices. Cavity is the biggest part of the generated cells. Cells classified as connective tissue begin to develop around the E13.5 stage and are generated by both models. The same applies to lung cells, which the GMFlow mode captures to a limited extent, whereas the CFM model does not.

3.2 Quantitative Evaluation via Wasserstein Distance

We further quantify the model accuracy in the SEDR embedding space using the Wasserstein distance. The Wasserstein distance (or Earth Mover's Distance) calculates the minimum "work" required to transform the generated probability distribution into the original data distribution, providing a robust metric for global structural similarity. As shown in Table 1, the GMFlow model consistently outperforms the CFM baseline across all slices from our test set. The GMFlow model achieves a significantly lower mean Wasserstein distance compared to the CFM model. This lower distance indicates that the Gaussian mixture approach more faithfully reconstructs the complex, multimodal landscape of the developing mouse embryo.

4 Discussion

In this work, we applied the GMFlow framework to generative modeling of spatial transcriptomics data. By explicitly modeling the conditional velocity distribution as a Gaussian mixture, the aim is to capture the intrinsic multimodality of cell-type distributions during embryonic development.

On the held-out E13.5 stage of the MOSTA dataset, both models, the GMFlow model and the baseline CFM, reflect the underlying cell type distribution to some extent. In particular, GMFlow achieved substantially lower Wasserstein distances across all slices, indicating improved structural alignment in the learned embedding space.

In future work, we plan to augment our GMFlow model of gene expression with a spatial reconstruction module. We believe that such a joint model can unravel the true potential of explicit modeling of cell-type distributional changes in spatial transcriptomics data, enhancing the generative modeling for spatial biology for tasks such as cross-modal or perturbation response prediction.

References

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. 18 (11):1352–1362. ISSN 1548-7105. doi: 10.1038/s41592-021-01264-7.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. 185(10):1777–1792.e21, a. ISSN 0092-8674. doi: 10.1016/j.cell.2022.04.003. URL https://www.sciencedirect.com/science/article/pii/S0092867422003993.

Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models, b. URL http://arxiv.org/abs/2504.05304.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. 13(1):1739. ISSN 2041-1723. doi: 10.1038/s41467-022-29439-6. URL https://www.nature.com/articles/s41467-022-29439-6.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and

histology to identify spatial domains and spatially variable genes by graph convolutional network. 18(11):1342–1351. ISSN 1548-7105. doi: 10.1038/s41592-021-01255-8. URL https://www.nature.com/articles/s41592-021-01255-8.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. URL http://arxiv.org/abs/2210.02747.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. 14(1):1155. ISSN 2041-1723. doi: 10.1038/s41467-023-36796-3.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. 15(12):1053–1058. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

Ruqian Lyu, Annika Vannan, Jonathan A. Kropski, Nicholas E. Banovich, and Davis J. McCarthy. SpatialRNA: a python package for easy application of graph neural network models on single-molecule spatial transcriptomics dataset. 42(1):btaf659. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf659.

Vivien Marx. Method of the year: spatially resolved transcriptomics. 18(1):9–14. ISSN 1548-7105. doi: 10.1038/s41592-020-01033-y. URL https://www.nature.com/articles/s41592-020-01033-y.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. 19(5):534–546. ISSN 1548-7105. doi: 10.1038/s41592-022-01409-2. URL https://www.nature.com/articles/s41592-022-01409-2.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. URL http://arxiv.org/abs/2302.00482.

Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, Ao Chen, Ling Shao, Longqi Liu, and Jinmiao Chen. Unsupervised spatially embedded deep representation of spatial transcriptomics. 16(1):12. ISSN 1756-994X. doi: 10.1186/s13073-024-01283-x.