Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

263-5051-00L: AI Center Projects in Machine Learning Research (Spring 2025)

ETH AI CENTER

Foundation Models and GMM for Soft Clustering

Gauthier-Villars Arthème [1], Matin Urdu [1], Keyshav Mor [1], Nicolas Dickenmann [1], Dr. Marina Esteban-Medina[2]
[1]ETH Zurich Student, D-INFK [2]ETH AI Center, ETH Zurich

Problem Statement

Recent foundation models that focus on single-cell RNA sequencing data (scRNA-seq) have demonstrated strong capabilities to learn robust latent representations from single-cell datasets. Downstream analysis of scRNA-seq data can benefit from the pre-trained embeddings generated by these models to capture intricate cellular characteristics and heterogeneity. Neuroblastoma is a type of cancer that is caused by immature nerve cells called neuroblasts and typically affects children under the age of 5. Clustering clinically relevant cell states from the data is a key step in neuroblastoma research to investigate cell state interactions. However, these data are often high-dimensional, noisy, and complex, making accurate clustering a challenging task.

Traditional clustering methods such as k-means, Louvain, and Leiden enforce rigid cell state assignments. In contrast, applying a probabilistic Gaussian Mixture Model (GMM) to the latent space embeddings allows for soft clustering, quantifying uncertainty in cell state assignments and demonstrating the reliability of the discovered clusters. We explore the use of GMM on latent space embeddings from foundation models to capture potential intermediate or transitioning cell states that may be overlooked by traditional hard clustering methods.

Datasets and Models

In our study, we used the following neuroblastoma data sets, Yu et al. (2025):

- **SN Tumors Dataset (SN)**: RNA-seq data from a single nucleus.
- **NB Atlas Dataset (NBA)**: RNA-seq data from a single nucleus and a single cell.
- **NB Concatenated (NBC)**: Dataset with SN and NBA profiles.

In our study, we used the following foundations models:

- **scFoundation:** Model on single-cell transcriptomics **[Hao et al. (2024)]**
- **cancerfoundation:** Model to decipher drug resistance in cancer **[Theus et al. (2024)]**
- **scGPT:** Generative AI model for single-cell multi-omics **[Cui et al. (2023)]**
- **scBERT:** Model for cell type annotation of scRNA-seq data **[Yang et al. (2021)]**

Methods

For every model we computed the embeddings in the scRNA-seq data. The scGPT and scFoundation models were fine-tuned on cell types and cell states to verify classification performance prior to embeddings' clustering. The embeddings were preprocessed prior to clustering to keep only batches with at least 3 neighbors and to apply PCA to the remaining z-normalized batches to obtain the main 50 components. For batch correction, the batch-balanced k-nearest neighbor (BBKNN: Polański et al. (2019)) or Harmony Korsunsky et al. (2019) are used. Both aim to correct batch effects, which represent nonbiological variation between batches, allowing for the subsequent models to better capture the underlying biological differences. Finally, we applied four clustering methods on both datasets with the results depicted in table 1. For GMMs, we obtain the cluster labels for a sample by choosing the cluster with the highest membership probability.

Results

Emb. Model	Data set	Louv. ARI	Louv. NMI	Louv. Hom	Louv. Cmp	Leid. ARI	Leid. NMI	Leid. Hom	Leid. Cmp	kMns. ARI	kMns. NMI	kMns. Hom	kMns. Cmp	GMM ARI	GMM NMI	GMM Hom	GMM Cmp
scFoundation	NBA	.002	.025	**.409**	.012	.002	.025	**.435**	.013	.000	.003	.042	.001	.000	.013	.192	.007
CancerFnd.	NBA	**.125**	**.298**	.273	**.328**	**.135**	**.299**	.291	**.307**	.009	**.258**	**.226**	**.300**	**.097**	**.295**	**.247**	**.365**
scGPT	NBA	.041	.169	.178	.160	.041	.178	.197	.162	**.034**	.120	.106	.140	.031	.118	.104	.137
scFoundation	SN	.066	.184	.261	.142	.064	.189	.275	.144	.102	.241	.304	.201	.225	.280	.188	.133
CancerFnd.	SN	.055	.206	.310	.154	.047	.206	.318	.152	.051	.210	.360	.148	.048	.211	.368	.148
scGPT	SN	**.134**	**.431**	**.687**	**.313**	**.098**	**.414**	**.698**	**.294**	**.389**	**.520**	**.653**	**.433**	**.286**	**.483**	**.612**	**.399**
scBERT	SN	.057	.113	.146	.093	.056	.115	.150	.093	.067	.157	.212	.125	.064	.142	.192	.113

Table 1: Evaluation of embedding from foundation models on various clustering tasks. We evaluated the accuracy of the Louvain method (Louv.), Leiden algorithm (Leid.), k-Means (kMns.) and the Gaussian mixture model (GMM) using the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Homogeneity (Hom), and Completeness (Cmp). We determine the number of clusters to use for GMM and k-Means by fitting GMMs between 2 to 20 components and choose the number of components that minimize the Bayesian Information Criterion (BIC).

Figure 1: scIB metrics on SN Dataset

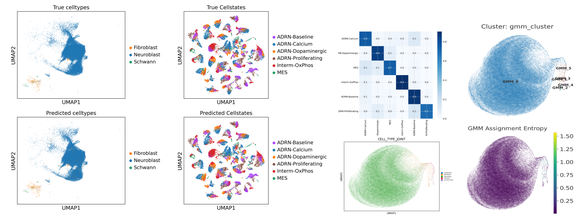

Figure 1: **(Starting from left) 1:** scGPT classification performance on unseen NBA dataset; **2:** scGPT classification performance on left out SN dataset; **3 Top:** SN cell state confusion matrix; **3 Bottom:** NBC cell type ground truth, **4 Top:** NBC 6 GMM clusters, **4 Bottom:** NBC GMM entropy.

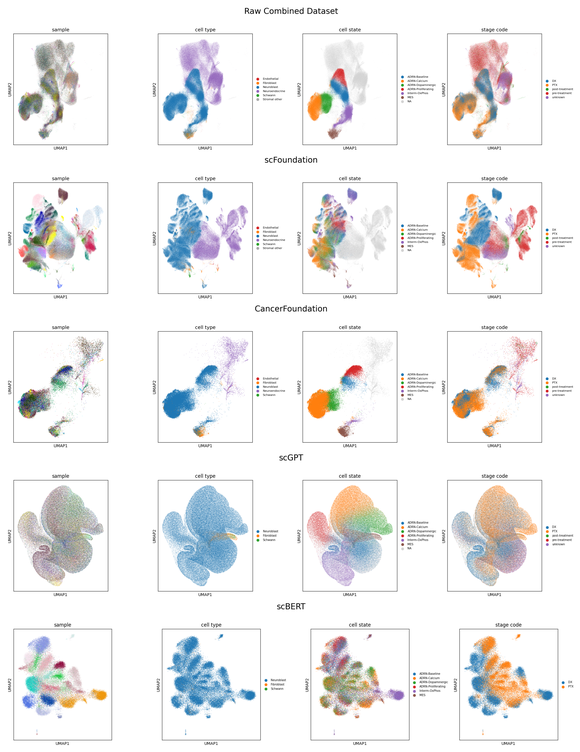

Figure 2: UMAP of the models embedding on Concatenated Neuroblastoma Dataset

Note: 1. Due to limited computational resources, all scBERT results were run on a subset of 50,000 cells from the SN dataset. 2. scGPT was fine-tuned on the SN dataset for the above plots.

Conclusions

We obtain promising latent embeddings from the foundation models that denoise the scRNA seq data while preserving key biological information. Among the foundation models, CancerFoundation and scGPT show superior performance compared to the others, with scGPT performance improving only after fine-tuning. GMM clustering on latent embeddings from different foundation models shows that only the GMM clusters on scGPT embeddings resembled the ground truth to a high degree and with a lower assignment entropy.

Future Work

- Fine-tuning or modification of foundation models to improve performance on neuroblastoma datasets.
- Fine-tuning of clustering methods.
- Treatment response analysis on cell state embeddings from foundation models.

References

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, 2023.

Mingze Hao, Jiayi Gong, Xiaoyu Zeng, et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. Bbknn: Fast batch alignment of single cell transcriptomes. *Bioinformatics*, 2019. doi: 10.1093/bioinformatics/btz625.

Alexander Theus, Florian Barkmann, David Wissel, and Valentina Boeva. Cancerfoundation: A single-cell rna sequencing foundation model to decipher drug resistance in cancer. *bioRxiv*, 2024. doi: 10.1101/2024.11.01.621087. URL https://www.biorxiv.org/content/early/2024/11/02/2024.11.01.621087.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. December 2021. doi: 10.1101/2021.12.05.471261. URL http://dx.doi.org/10.1101/2021.12.05.471261.

Wenbao Yu, Rumeysa Biyik-Sit, Yasin Uzun, Chia-Hui Chen, Anusha Thadi, and Sussman. Longitudinal single-cell multiomic atlas of high-risk neuroblastoma reveals chemotherapy-induced tumor microenvironment rewiring. *Nature Genetics*, 57(5):1142–1154, April 2025. ISSN 1546-1718. doi: 10.1038/s41588-025-02158-6. URL http://dx.doi.org/10.1038/s41588-025-02158-6.

Contributions

AGV, MU, KM, ND: Conceptualization, Methodology, Software, Investigation, , Visualization, Writing
MEM: Conceptualization, Methodology, Resources, Supervision, Project administration