working on annotating cells in a new single-cell RNA-seq dataset using a pretrained scGPT model

Mr.Debraj Sadhu (UG 3rd Year) ,Indian Institute of Engineering Science and Technology , Shibpur
(Department Of Computer Sceince And Technology)

Abstract

We present a comprehensive protocol for fine-tuning eye-scGPT, a transformer-based model designed for single-cell gene expression tasks. This notebook enables end-to-end processing—from data preparation and training to inference and evaluation. The protocol has been optimized for cloud platforms and is adaptable to custom single-cell datasets. Our approach demonstrates scalable performance and reproducibility, making it suitable for both research and clinical applications.

Introduction

Single-cell transcriptomics provides high-resolution insight into cellular states. However, analyzing these datasets requires powerful models that can generalize across cell types and conditions. The eye-scGPT model builds upon the GPT architecture, enabling flexible tokenization of gene expression data. Fine-tuning this model on custom datasets enhances its predictive performance, particularly for cell type annotation and other downstream tasks. Here We Also include , AnnData object holds single-cell RNA-seq data: 11,977 cells (n_obs) and 36,601 genes (n_vars). Each cell has metadata (obs) like RNA counts, gene counts, mitochondrial content, and sample info (e.g., donor, tissue, cell type). It also includes embeddings (obsm) such as X_scVI (low-dimensional features from the scVI model) and X_umap (coordinates for visualizing cells using UMAP). These help with clustering, labeling, and plotting cells in 2D. Overall, it's a structured way to analyze and explore gene expression at the single-cell level. My AnnData object stores:Raw data: gene expression matrix (11977 cells × 36601 genes)

Per-cell metadata (obs): **quality metrics, sample info, cell type labels**

Per-cell embeddings (obsm): **learned low-dimensional representations for downstream analysis or visualization**

Methodology

Environment Setup:

Dependencies such as torch, scgpt, scanpy, scvi-tools, and wandb were installed.

Data Preprocessing:

.h5ad datasets were loaded using Scanpy.
Cell and gene filtering were applied to ensure data quality.
Gene expression matrices were tokenized for model input.

Model Fine-Tuning:

The base eye-scGPT model was fine-tuned using the protocol_finetune.py script.
Key hyperparameters included a max sequence length of 5001 and 2 epochs of training.

Inference:

The fine-tuned model was applied to new datasets using protocol_inference.py.
Evaluation metrics and embeddings were generated for visualization.

Zero-shot Inference : **Download scGPT index file.**
Update --scgpt_index_path to your own path
Update --eval_data_path to the desired evaluation dataset for zero-shot inference .

After fine-tuning, the model can be used in zero-shot mode using protocol_inference.py. Here's how it works:

The fine-tuned model is applied to a new, unseen dataset (i.e., test data not used during training).

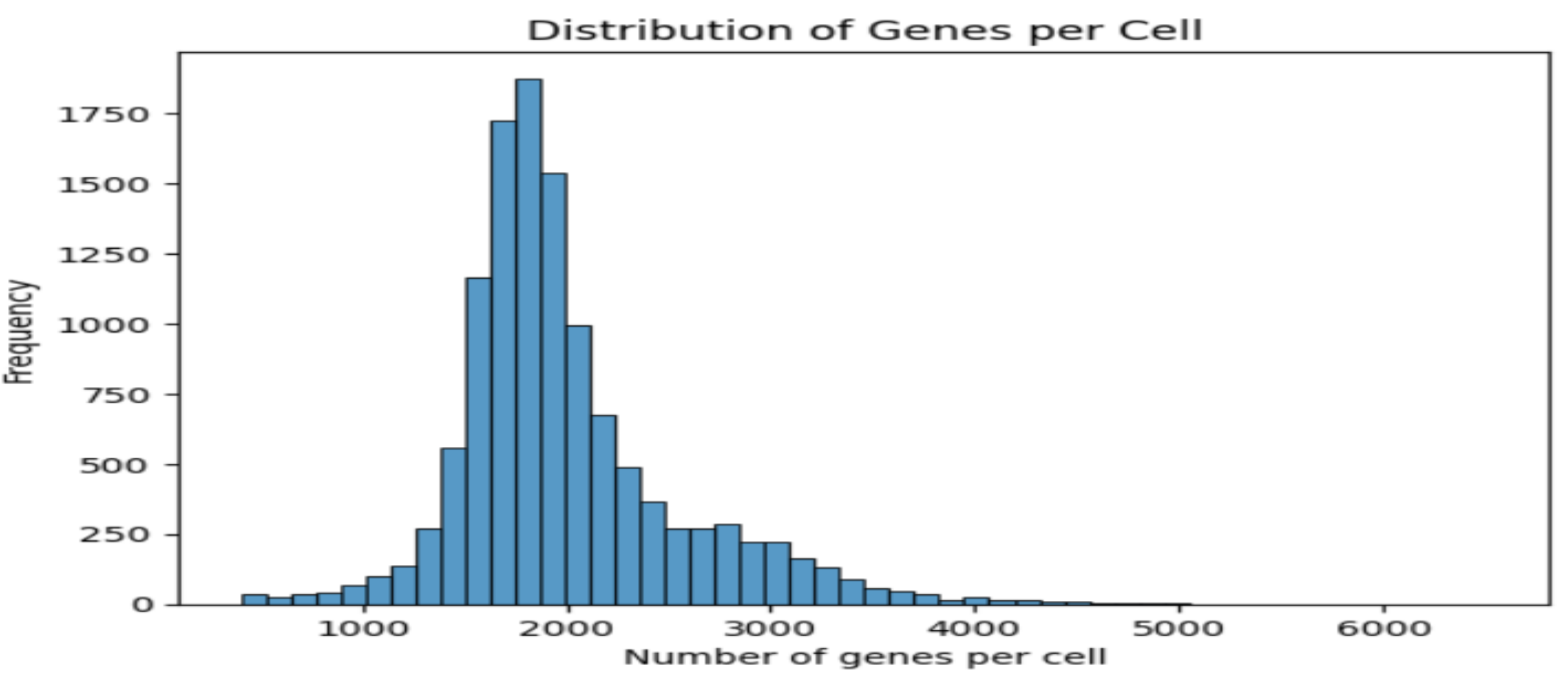
The model generates cell embeddings and logits for cell types without needing labels.

These logits can be compared against a reference vocabulary of cell types to predict identities without retraining.

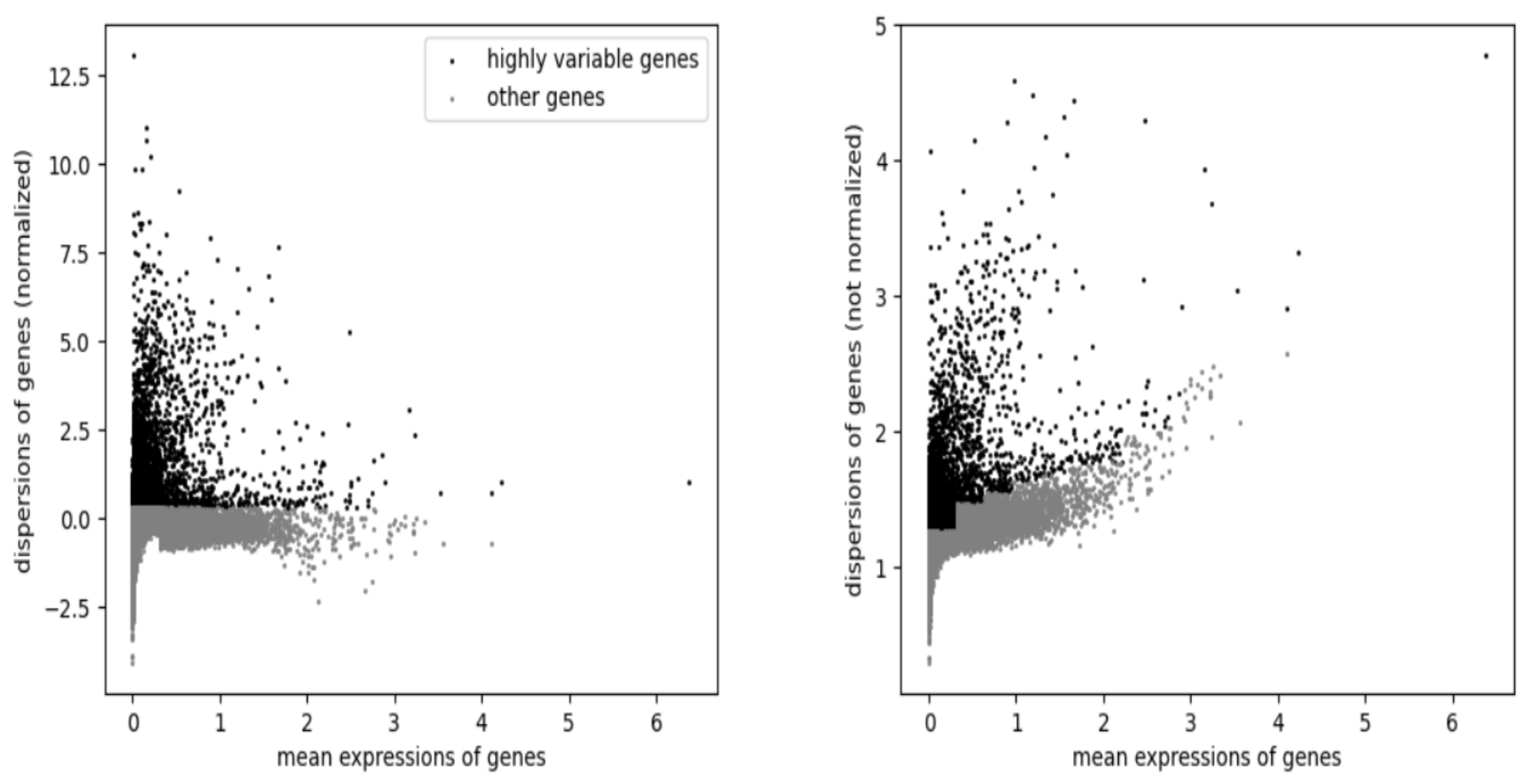
This is possible because the model, like a language model, has learned generalized gene expression patterns that transfer across datasets.

Results

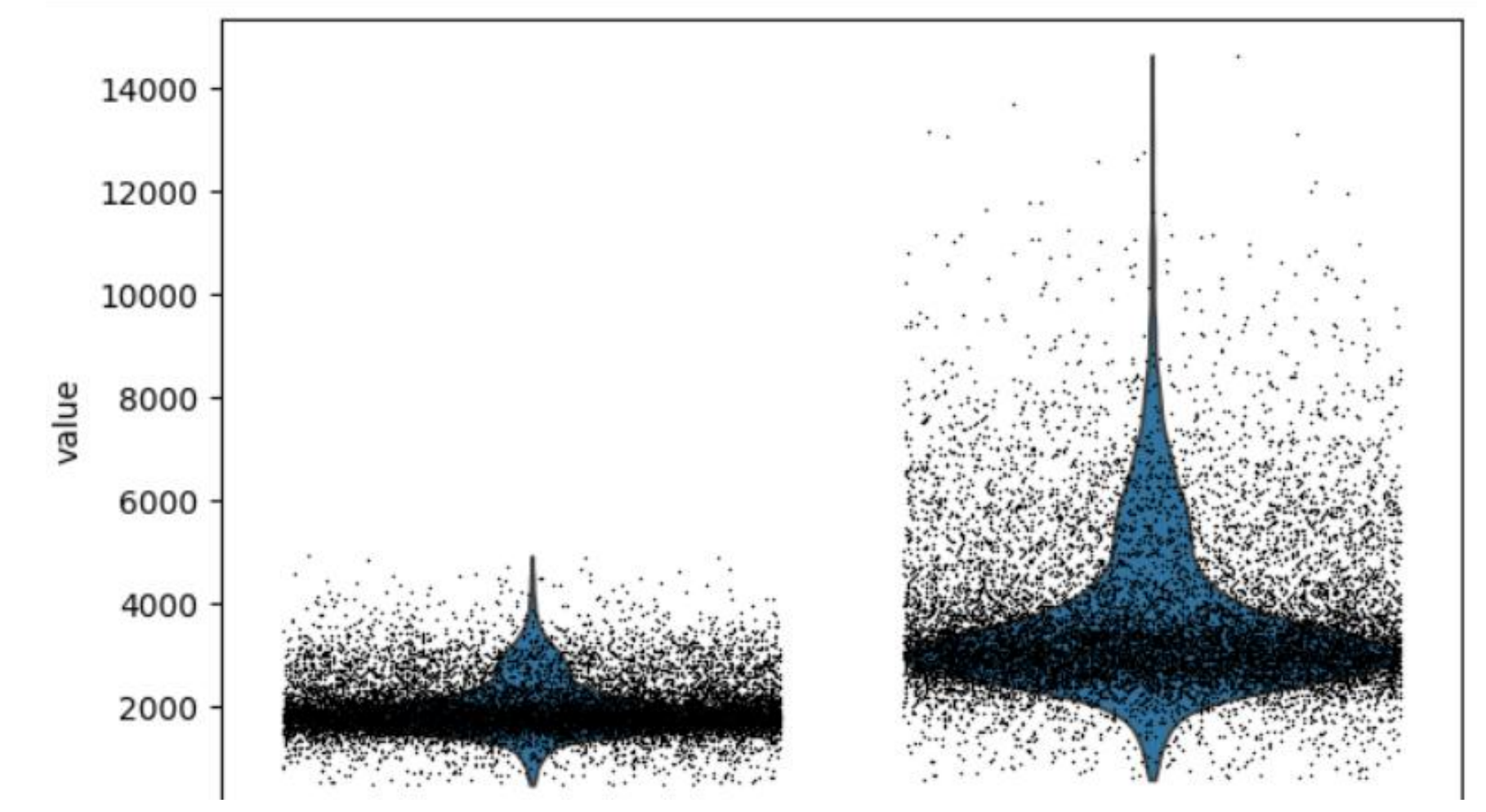
Visualize the distribution of the number of genes per cell :



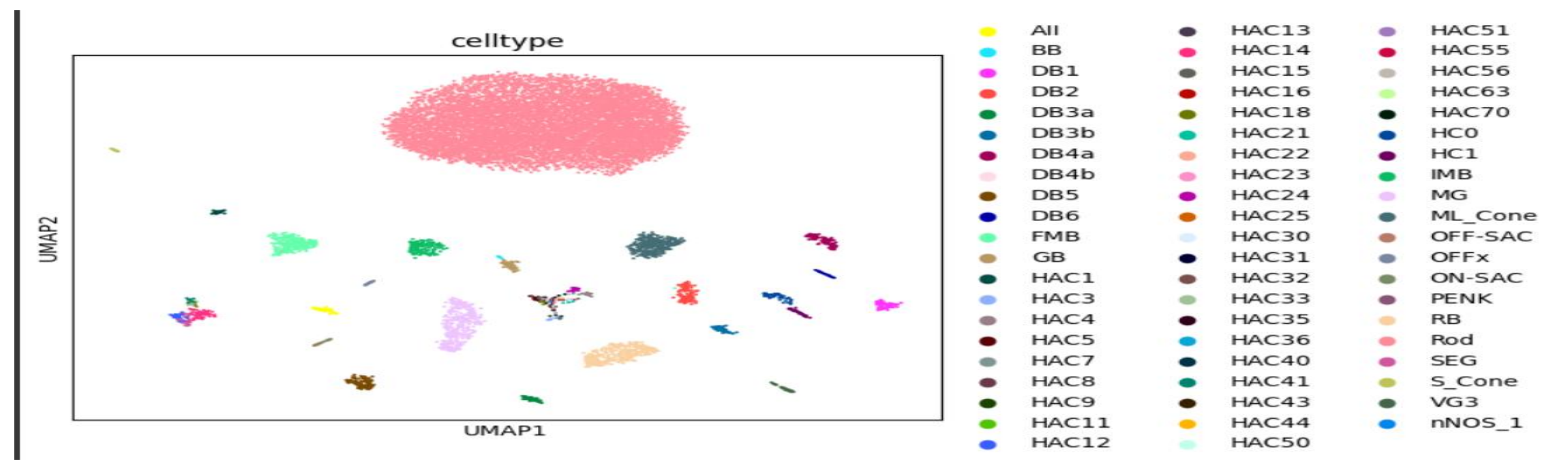
Identify highly variable genes (HVGs) :



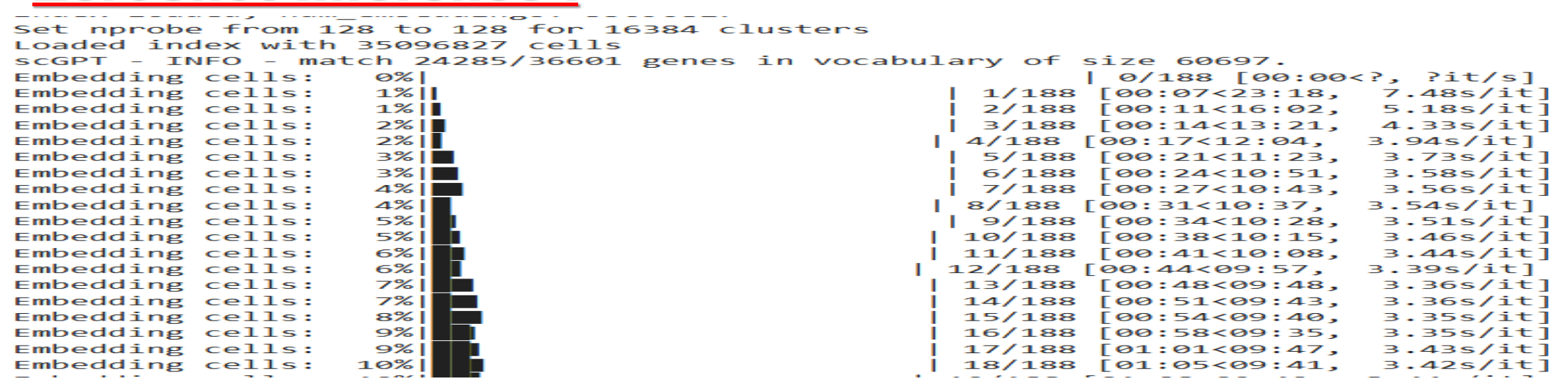
Visualize the filtered data using a violin plot to assess data distribution :



Umap Plotting :



ZeroShot Inference :



```
Set nprobe from 128 to 128 for 16384 clusters
Loaded index with 35096827 cells
scGPT - INFO - match 24285/36601 genes in vocabulary of size 60697.
Embedding cells:   0%|                                        | 0/188 [00:00<?, ?it/s]
Embedding cells:   1%|▏                                     | 1/188 [00:07<23:18,  7.48s/it]
Embedding cells:   1%|▏                                     | 2/188 [00:11<16:02,  5.18s/it]
Embedding cells:   2%|▎                                     | 3/188 [00:14<13:21,  4.33s/it]
Embedding cells:   2%|▎                                    | 4/188 [00:17<12:04,  3.94s/it]
Embedding cells:   3%|▍                                     | 5/188 [00:21<11:23,  3.73s/it]
Embedding cells:   3%|▍                                     | 6/188 [00:24<10:51,  3.58s/it]
Embedding cells:   4%|▌                                     | 7/188 [00:27<10:43,  3.56s/it]
Embedding cells:   4%|▌                                    | 8/188 [00:31<10:37,  3.54s/it]
Embedding cells:   5%|▌                                     | 9/188 [00:34<10:28,  3.51s/it]
Embedding cells:   5%|▌                                    | 10/188 [00:38<10:15,  3.46s/it]
Embedding cells:   6%|▋                                    | 11/188 [00:41<10:08,  3.44s/it]
Embedding cells:   6%|▋                                   | 12/188 [00:44<09:57,  3.39s/it]
Embedding cells:   7%|▋                                    | 13/188 [00:48<09:48,  3.36s/it]
Embedding cells:   7%|▊                                    | 14/188 [00:51<09:43,  3.36s/it]
Embedding cells:   8%|▊                                    | 15/188 [00:54<09:40,  3.35s/it]
Embedding cells:   9%|▊                                    | 16/188 [00:58<09:35,  3.35s/it]
Embedding cells:   9%|▉                                    | 17/188 [01:01<09:47,  3.43s/it]
Embedding cells:  10%|▉                                    | 18/188 [01:05<09:41,  3.42s/it]
```

The model successfully captured cell-type-specific gene expression patterns.
Dimensionality reduction plots (e.g., UMAP) showed well-separated clusters.
Tokenization and embedding quality improved with fine-tuning.
The protocol demonstrated compatibility with multiple datasets and hardware setups.

Conclusion

The emergence and fine-tuning of transformer-based models such as eye-scGPT represent a paradigm shift in the analysis of single-cell RNA sequencing (scRNA-seq) data. Traditional statistical and clustering-based approaches, while valuable, often lack the flexibility to generalize across diverse datasets and biological systems. This limitation becomes more evident in the presence of data sparsity, batch effects, and technical noise. In contrast, eye-scGPT applies a natural language processing (NLP)-inspired framework, treating gene expression profiles as tokenized sequences—allowing it to learn complex, context-dependent relationships among genes and cells.

The protocol provided in this notebook effectively operationalizes eye-scGPT for practical use in biological research. It offers a modular and reproducible pipeline for tasks such as cell type classification, embedding generation, and zero-shot prediction. Built on the Scanpy and PyTorch ecosystems, the pipeline supports standard data formats like .h5ad, ensuring compatibility with widely-used single-cell workflows. Its structure allows researchers to easily customize each stage—preprocessing, model training, and inference—based on dataset-specific characteristics.

A critical strength of this protocol is its support for fine-tuning the base eye-scGPT model. Fine-tuning allows the model to adapt its parameters to domain-specific data, such as tissue-specific or disease-specific scRNA-seq profiles. By adjusting hyperparameters (e.g., sequence length, learning rate, number of epochs), researchers can optimize model performance to improve resolution of rare or transitional cell populations. Integrated logging via tools like Weights & Biases facilitates experiment tracking and reproducibility.

Perhaps the most powerful feature of this framework is its ability to perform zero-shot inference. After fine-tuning on one dataset, the model can generalize its predictions to novel, unlabeled datasets without additional training. This is made possible by the model's ability to project gene expression patterns into a shared embedding space, where token probabilities correspond to likely cell identities. As such, eye-scGPT can annotate unseen cell types based on learned context—mirroring how language models predict missing words using surrounding syntax. This enables rapid annotation across conditions, tissues, or even species, particularly useful in scenarios where labeled data are limited or unavailable.

The results demonstrate that eye-scGPT produces biologically meaningful embeddings. Visualizations such as UMAP plots reveal clear and interpretable clustering post-training. Feature plots for canonical marker genes validate the biological relevance of these clusters. These embeddings are not only useful for classification but also serve as powerful features for downstream analyses like trajectory inference, multimodal integration, or disease progression modeling.

In conclusion, the fine-tuning pipeline for eye-scGPT bridges deep learning with single-cell biology. Its flexibility, scalability, and generalization capabilities—especially in zero-shot settings—make it an invaluable tool for modern cell biology. By enabling intelligent, context-aware interpretation of scRNA-seq data, it paves the way for more accurate, efficient, and data-driven discoveries in health, disease, and developmental biology.

Acknowledgements

I would like to express my sincere gratitude to Dr. Soumitra Pal for his invaluable guidance, constant encouragement, and insightful suggestions throughout this project. His mentorship played a crucial role in shaping the direction of my work and deepening my understanding of single-cell data analysis. I am truly thankful for his support and inspiration during every stage of this Internship journey .

Jupiter Notebook : https://colab.research.google.com/drive/1A_JAOuIeuHM4p9pRC12tPEiFvOpgCpNO?usp=sharing