**Research Project Report**

**Division of Health Medical Intelligence**
**Information Science & Technology**

# Cell Sentences and Language Models for Single-Cell Analysis: Evaluating the Role of HVG Selection

**Jean-Baptiste BOBANT**
*University of Tokyo*
*Tokyo, Japan*
*Paris-Saclay Univserity*
*Paris, France*
*jean-baptiste.bobant@student-cs.fr*

***Abstract*—** This study is motivated and enters the framework of AI-powered virtual cells (AIVC), exploring ways to represent cells and their heterogeneity at various scale [1]. We implement Cell2sentence representation to convert single-cell expression profiles in ranked genes sentences, allowing us to leverage natural processing tools [2]. We use C2S-Pythia-410M large language model specifically pretrained on single-cells to address key downstream task in single-cell analysis: cell-type annotation.
Highly variable gene (HVG) selection is very central in single-cell analysis being a strongly recognized technique to enhance signal detection and reduce noise that makes single-cell analysis hard [3, 4]
We will then evaluate the choice of HVG selection in the sentence representation of cells in the sentence construction pipeline and its impact on downstream performances in comparison with full cell sentences performances.
Our experiments are conducted according to the scEval benchmark framework to ensure reproducible and comparable results [5], benchmarking the method with top and new cell-type annotation models (e.g., CellFM [7]).
Our results show that HVG-based cell sentences perform as well as full cell sentences in expression reconstruction and cell-type annotation.
Overall, we showcased how HVG-based cell sentence embeddings can support multiple single-cell downstream analysis and where this representation struggles, paving the way towards more adapted alternative cell representations.

***Keywords*— *Cell Sentences / Large Language Models / Highly Variable Genes / Single-Cell Analysis***

## I. Introduction

Single-cell RNA sequencing enables the transcriptional profiling of individual cells, shedding light on cellular heterogeneity that bulk methods tend to obscure. However, single-cell analysis faces unique challenges, high sparsity, noise, batch effects and high scale that creates computational barriers for downstream analysis [8]
In parallel, natural language processing has been increasingly applied to very specific domains, leveraging large language models (LLMs), which excel at encoding complex semantic structures in token-based text sequences [9]. Recently, methods have begun to implement biological data in textual formats so as to leverage these LLMs' power.

### A. Cell Sentence Construction

We based our study on the Cell2Sentence framework [2] which has been developed in order to explore the ability of LLMs to interpret biological meaning from gene expression when converted to a natural language format, cell sentences.
To construct these cell sentences, we proceed by ranking the genes by their expression level within each cell after normalization. We only consider expressed genes which are therefore sorted in decreasing order of expression. Thus creating a cell sentence that represents genes relative expression without relying on raw counts and therefore minimizing sparsity and batch effects. Each gene symbol is therefore treated as an individual token and the sentence representation of each cell is this ordered list of gene tokens.
The major contribution on the design choice in Cell2Sentence is their use of log-transformed rank of a gene in their sentences to approximate its expression level. In fact they highlighted a strong correlation between log(rank) and expression level, making this representation of cells almost reversible.

### B. HVG Selection Strategy

Another major contribution of our work is the use of Highly Variable Gene selection in the cell sentence construction which is a foundational feature-selection technique in single-cell analysis. It aims at prioritizing genes whose expression significantly varies across the cell population. It filters uninformative or noise-dominated genes retaining genes that are most likely to contain meaningful biological variations [10,11]. Allowing us to enhance the signal-to-noise ratio reducing computational costs for downstream tasks such as cell-type annotation. Empirical studies have shown that genes with high expression variability across cells are enriched for markers of cell identity and functional heterogeneity [12, 13]. Given this crucial role in single-cell analysis, we use HVG selection, keeping the 2000 highest variable genes, upstream of our cell-sentence construction which divide by three the length of the cell sentences. This ensures our sentences emphasize the most biologically relevant features.

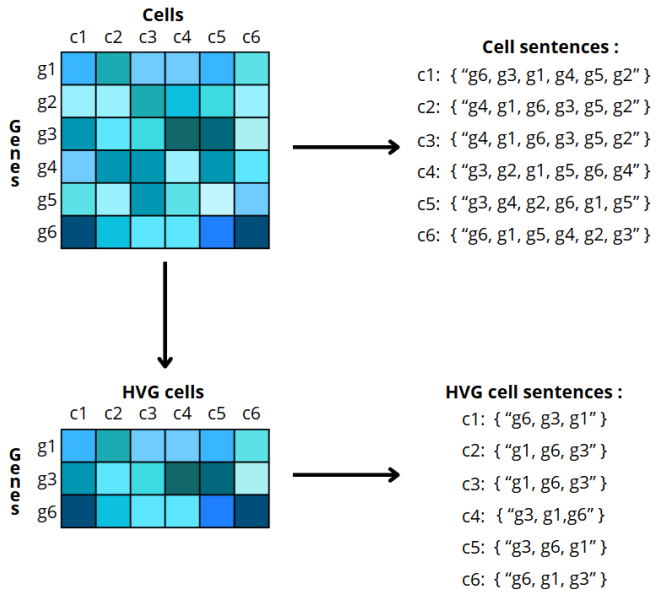


FIGURE I: CELL-SENTENCES AND HVG CELL-SENTENCES CONSTRUCTION PROCESS

## C. *Single-cell Preprocessing*

We adopted the standard Cell2Sentence preprocessing pipeline implemented in Scanpy [2, 14], applying a sequence of quality control and normalization steps to prepare single-cell expression data for sentence construction.

1. **Cell and gene filtering:** We removed low-quality cells with fewer than 200 expressed genes and genes expressed in fewer than 3 cells, using `sc.pp.filter_cells` and `sc.pp.filter_genes` [14].
2. **Mitochondrial content QC:** Genes with the "MT-" prefix were flagged and mitochondrial gene percentage per cell computed via `sc.pp.calculate_qc_metrics` [14]. Cells with more than 50% mitochondrial content or more than 6000 genes were filtered out, following Scanpy tutorials.
3. **Normalization and log-transform:** Counts were depth-normalized using `sc.pp.normalize_total` and log-transformed with base-10 `sc.pp.log1p`, as required by Cell2Sentence pipeline [2].

This pipeline effectively removes technical artifacts (e.g., doublets, low-quality cells), ensures consistent scaling (total counts), and applies log-based transformation to stabilize variance, which is an essential preparatory step before sentence encoding.

# II. DOWNSTREAM TASKS

## A. *Cell Reconstruction*

### 1. *Task*

The goal of this task is to prove that the cell-sentence representation can be almost considered a reversible transformation. We aim at recovering the original gene expression profiles from their rank-based cell-sentence. In the Cell2Sentence study [2], they evaluated the reversibility of the transformation by applying a linear regression to reconstruct normalized gene expression values from gene ranks. Their results on the PBMC dataset demonstrate that—even though no explicit expression values are used during sentence construction, the linear regression recreates effectively the gene expression profiles, with an $R^2$ of 0.815 and a Spearman correlation of 0.815 between predicted and actual expression across 10,000 sampled genes [2].

### 2. *Methods*

We model the normalized log-expression $e_g^{(i)}$ of gene $g$ in cell $i$ as a linear function of the log-transformed rank of genes in the cell sentence. Specifically:

$$e_g^{(i)} \approx \beta_{g,0} + \sum_{j=1}^{k} \beta_{g,j} \log(r), \tag{1}$$

where: - $r$ indexes the position of a gene in the sentence (ranked by expression), - $\log(r)$ is its log-transformed rank, - $\beta_{g,0}, \ldots, \beta_{g,k}$ are gene-specific coefficients fitted via least squares on dataset.

Equation (1) thus formalizes the hypothesis that gene expression is linearly predictable from the log-rank ordering. This assumption has been validated by strong performance metrics in the original Cell2Sentence study [2].

### 3. *Datasets*

We evaluated our methods on four benchmark single-cell RNA datasets derived from human cells introduced in the CellFM study [7], each representing diverse tissues with varying levels of cell-type imbalance and batch effects:

- **Pancreas (Pancrm):** 12,906 cells × 15,558 genes, encompassing 15 annotated cell types across 4 distinct batches.
- **Liver:** 7,472 cells × 19,984 genes, composed of 20 cell types from 5 batches.
- **Heart:** 60,668 cells × 27,350 genes, containing 17 cell types.
- **Skin:** 15,180 cells × 20,346 genes, featuring 10 cell types across 5 batches.

All datasets are marked by strong class imbalance among cell types. For the cell reconstruction task, we subsampled each dataset to 1,024 randomly selected cells to reduce computation and align with prior evaluation protocols.

### 4. *Results*

Figure II–VII display the performance of our reconstruction model on both full-gene and HVG-based cell sentence representations across three human single-cell datasets (Liver, Skin, Pancrm). Metrics include $R^2$, Pearson, and Spearman correlations, demonstrating high fidelity in all configurations. We observe that HVG sentences slightly outperform full-gene sentences on Liver and Skin, while the Pancrm dataset shows very similar outcomes. These quantitative insights are visualized in the following plots.

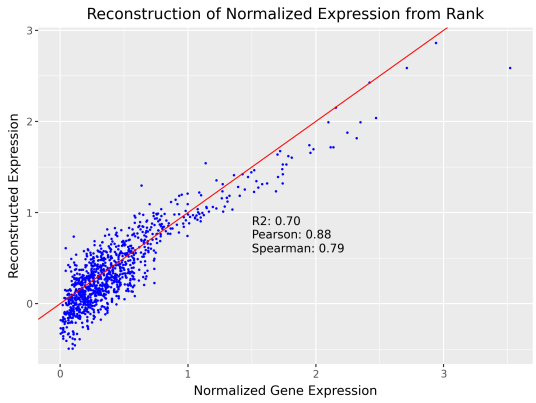


FIGURE II: LIVER — FULL GENES RECONSTRUCTION ($R^2$=0.70, PEARSON=0.88, SPEARMAN=0.79)

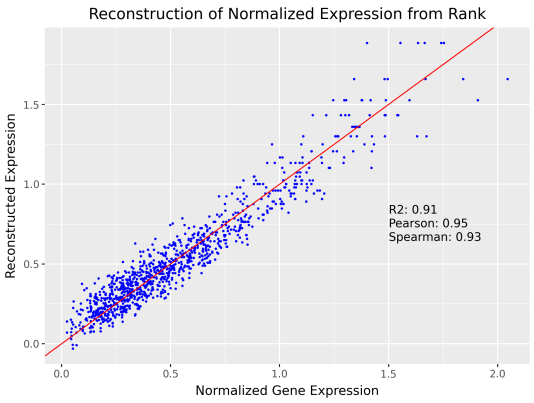


FIGURE V: SKIN — HVG RECONSTRUCTION ($R^2$=0.91, PEARSON=0.95, SPEARMAN=0.93)

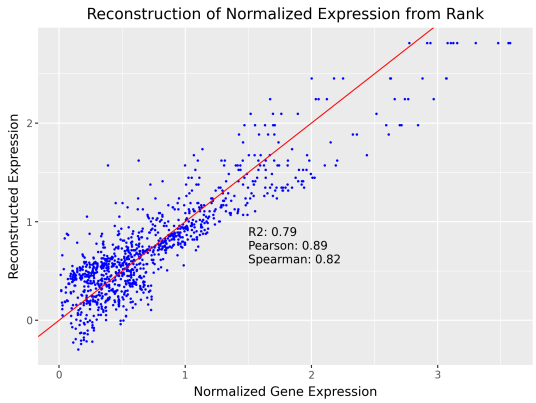


FIGURE III: LIVER — HVG RECONSTRUCTION ($R^2$=0.79, PEARSON=0.89, SPEARMAN=0.82)

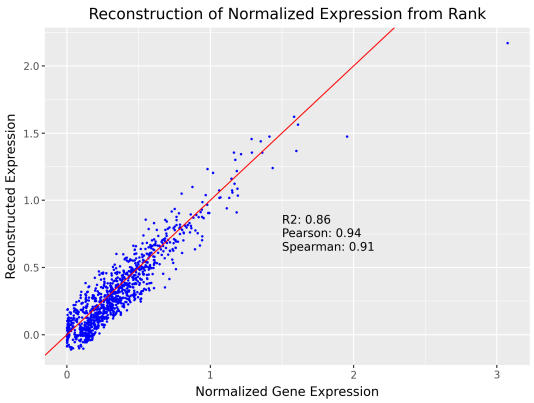


FIGURE VI: PANCRM — FULL GENES RECONSTRUCTION ($R^2$=0.86, PEARSON=0.94, SPEARMAN=0.91)

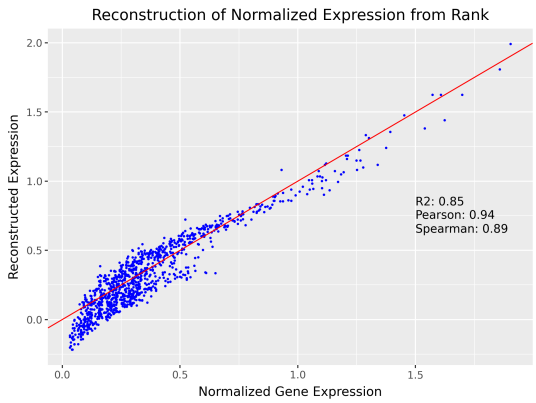


FIGURE IV: SKIN — FULL GENES RECONSTRUCTION ($R^2$=0.85, PEARSON=0.94, SPEARMAN=0.89)

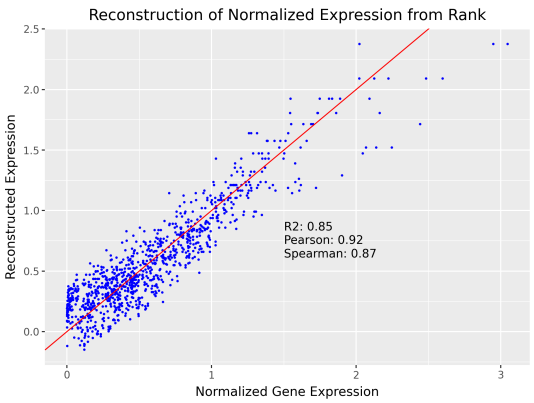


FIGURE VII: PANCRM — HVG RECONSTRUCTION ($R^2$=0.85, PEARSON=0.92, SPEARMAN=0.87)

TABLE I: OVERALL MEAN RECONSTRUCTION PERFORMANCE SUMMARY (FULL VS. HVG CELL SENTENCES)

| Mode | $R^2$ | Pearson | Spearman |
|---|---|---|---|
| HVG | 0.850 | 0.920 | 0.873 |
| FULL GENES | 0.803 | 0.920 | 0.863 |

The results on the reconstruction task highlight that both full-gene and HVG-based sentence representations yield a high level of fidelity in recovering normalized log-expression profiles on both datasets I. On Skin and Liver datasets, we can see improved performances using HVG but we need to keep in mind that it allows an incomplete reconstruction of the transcriptomic profile because of the loss inducted by the HVG genes selection. These results confirm the fundamental premise firstly described in the Cell2sentence study: log-rank gene sentences contain expression information across single-cell datasets, strong overall $R^2$, Pearson and Spearman coefficients [2]. However, we need to nuance these results because the HVG-based approach inherently excludes most of the genes from the representation, therefore leading to an incomplete reconstruction of the full transcriptome. While HVG sentences slightly improve the reconstruction accuracy for the HVG-selected genes, we cannot recover expression from definitively deleted genes. This kind of trade-off sheds light on a critical design choice, HVG-based cell sentences offer an accurate representation only if the downstream pipelines consistently focus on high variable genes. Otherwise, full-gene cell sentences are more suited to tasks that require complete transcriptomic reconstruction.

## B. *Cell-type Annotation*

### 1. *Task*

Cell-type annotation is a paramount step in single-cell RNA-seq analysis, it aims at translating raw expression data into interpretable labels, in this case cell-types. We build our study on the *CellFM* benchmark which evaluates cell-type annotation following *scEval* requirements, across both intra-dataset and inter-dataset (cross-batches here) using frozen models' embedding fed to a MLP classifier trained on a specific partition of each dataset [5, 7]. We focus our study on assessing the performances of HVG-based cell sentences and full cell-sentences representations in enabling accurate annotation. We therefore compare our results with established models such as *scGPT*, *Geneformer*,*scELMo*, *scBERT*, *UCE* and SVM baselines.

### 2. *Methods*

We follow in our study the protocol proposed in *CellFM* benchmark for cell-type annotation following *scEval* intra-dataset and inter-dataset pipeline. **Embedding Extraction:**

We generated cell embeddings using *C2S-Pythia-410M* model which is based on the Pythia-410M architecture and pretrained following the *Cell2Sentence* framework on over 57 million human and mouse cells obtained from CellxGene and the Human Cell Atlas, using cell sentences with up to 200 gene tokens ordered by expression level per cell [2]. In our study, we apply *C2S-Pythia-410M* both to full-cell sentences (after ranking the full gene list) and to HVG-based sentences limited to the top 200 genes—primarily for computational feasibility. Since most HVG-based sentences already contain fewer than or around 200 genes, truncation was minimal in practice, preserving the integrity of highly variable gene representations while ensuring compatibility with the pretrained model's context length.

**Train/Test Splits:**

We performed two different train/test splits to fit *CellFM* benchmark and *scEval* requirements [5, 7] - *Intra-dataset*: we perform a 5-fold evaluation on each dataset. - - *Inter-dataset (Cross-batch)*: one batch is held out as test, all other batches used for training, it allows us to assess batch robustness in classification performance [5].

**MLP Classifier Architecture:**

We train a compact MLP classifier matching the CellFM design:

- Input layer matching the embedding dimensionality
- Fully connected layer with *512 units*, followed by BatchNorm, GELU activation, and Dropout(0.1).
- Second fully connected layer with *256 units*, also followed by BatchNorm, GELU activation, and Dropout(0.1).
- Final linear output layer of size `n_cls` (number of cell-type classes).

Training uses cross-entropy loss and the Adam optimizer at learning rate $10^{-3}$, early stopping after 30 epochs, up to 150 epochs.

**Evaluation Metrics:**

We report **overall accuracy** and **macro-F1 score**, the latter reflecting balanced performance across both abundant and rare cell types—critical in imbalanced single-cell datasets such as those we conduct our experiments on.

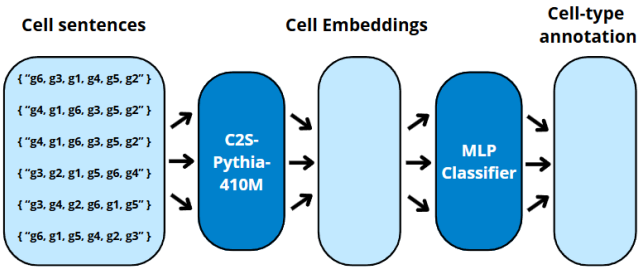


FIGURE VIII: CELL-TYPE ANNOTATION ARCHITECTURE

### 3. *Datasets*

We evaluated our annotation approach on four benchmark human single-cell RNA-seq datasets curated in the CellFM study [7], which encompass diverse tissues, pronounced cell-type imbalance, and batch-level heterogeneity:

- **Pancreas (Pancrm):** 12,906 cells × 15,558 genes, representing 15 annotated cell types across 4 batches.
- **Liver:** 7,472 cells × 19,984 genes, comprising 20 cell types from 5 batches.
- **Heart:** 60,668 cells × 27,350 genes, encompassing 17 cell types (batch information unspecified).
- **Skin:** 15,180 cells × 20,346 genes, representing 10 cell types across 5 batches.

All datasets exhibit strong class imbalance among cell types, which impacts annotation performance especially for rare populations.

For **intra-dataset** cell-type annotation, we applied 5-fold train/test split within the Pancrm dataset.

For **inter-dataset** cell-type annotation, which is crucial to real-world generalization, we held out one entire batch at a time as test set and trained on all remaining batches. We repeated this procedure across Liver, Skin, Heart, and Pancrm datasets to rigorously assess batch robustness and model generalization to new technical conditions.

This inter-dataset pipeline follows the *scEval* and *CellFM* evaluation protocols, providing a realistic benchmark of annotation performance in cross-batch settings which is definitely a key requirement for reliable application in new biological samples [5, 7].

### *4. Benchmark Models*

We compare our classifier performance under both intra-dataset and inter-dataset settings against an ensemble of state-of-the-art single-cell foundation models and conventional baselines, as described in the CellFM study:

- **scGPT** : Generative Transformer pretrained on 33M human cells, fine-tuned for annotation [15].
- **Geneformer** : rank-prediction model trained on 30M cells using a transformer for gene-order regression [16].
- **GeneCompass** : integrates biological priors across species using value-projection objective; 100M parameters [17].
- **scELMo** : combines pretrained metadata embeddings and gene expression for zero-shot classification across multiple tissues [18].
- **UCE** : Universal Cell Embedding model trained on ¿36M cells using masked gene modelling with self-supervised learning [19].
- **scBERT** (expression-level discretization with bucketed MLM pretraining on millions of cells) [20].
- **scFoundation** : mask autoencoder model that directly predicts continuous expression values; advanced pretrained foundation model [21].
- **SVM** : support vector machine using HVG features as classical baseline.
- **CellFM_800M** and **CellFM80M** : CellFM models pretrained on 800 million or 80 million cells, fine-tuned for annotation [7].

This comprehensive comparison allows us to assess not only the in-batch prediction accuracy but also the robustness to technical variation across batches (inter-dataset generalization), as emphasized in the scEval framework.

### *5. Results*

We first ran our cell-type annotation experiment for intra-dataset comparison using the pancrm dataset applying full cell sentences and HVG-based cell sentences C2S Pythia-410M embedding.

| | Accuracy | Macro-F1 |
|---|---|---|
| scGPT | 0,577 | 0,209 |
| GeneCompass | 0,628 | 0,405 |
| Geneformer | 0,791 | 0,455 |
| scmap | 0,911 | 0,767 |
| scELMo | 0,962 | 0,718 |
| UCE | 0,844 | 0,53 |
| scBERT | 0,963 | 0,701 |
| scFoundation | 0,957 | 0,803 |
| SVM | 0,968 | 0,705 |
| CellFM(800M) | 0,926 | 0,523 |
| CellFM(80M) | 0,986 | 0,705 |
| C2S Pythia-410M | 0,981 | 0,898 |
| HVG C2S Pythia-410M | 0,98 | 0,857 |

FIGURE IX: ACCURACY AND MACRO-F1 RESULTS ON INTRA-DATASET CELL-TYPE ANNOTATION

We can conclude from the results of our experiment IX that our method exhibits accuracy among top-performing models exceeding **0.98**. Moreover, it outperforms other models regarding Macro-F1 which better reflects performances across rare and abundant cell types, a very crucial feature when dealing with single-cells annotation. The full cell sentences and the HVG-based cell sentences models shows very similar performances even though the loss of information due to the HVG selection induces a slight reduction of performance. Therefore, the choice of HVG selection can offer both computational and representational efficiency, making it an attractive option in HVG-centric pipelines.

While intra-dataset results reveal global and not that representative performances under familiar data distributions, inter-dataset evaluation reflects with more precision real-life cell-type annotation, when new sample batches may come from unseen donors or even new protocols. We will then evaluate our pipeline performances, both Accuracy and Macro-F1, in cross-batch settings, testing our method's resilience to batch effects and thus annotation general consistency.

| | Heart | Liver | Pancrm | Skin |
|---|---|---|---|---|
| scGPT | 0,401 | 0,521 | 0,556 | 0,846 |
| GeneCompass | 0,629 | 0,557 | 0,693 | 0,88 |
| Geneformer | 0,404 | 0,62 | 0,828 | 0,651 |
| scmap | 0,405 | 0,847 | 0,827 | 0,749 |
| scELMo | 0,708 | 0,737 | 0,907 | 0,929 |
| UCE | 0,864 | 0,897 | 0,841 | 0,978 |
| scBERT | 0,882 | 0,848 | 0,918 | 0,97 |
| scFoundation | 0,906 | 0,904 | 0,912 | 0,978 |
| SVM | 0,899 | 0,92 | 0,919 | 0,98 |
| CellFM(800M) | 0,805 | 0,729 | 0,811 | 0,937 |
| CellFM(80M) | 0,908 | 0,912 | 0,962 | 0,98 |
| C2S Pythia-410M | 0,848 | 0,884 | 0,949 | 0,962 |
| HVG C2S Pythia-410M | 0,846 | 0,857 | 0,951 | 0,959 |

Figure X: Accuracy on inter-dataset cell-type annotation

| | Heart | Liver | Pancrm | Skin |
|---|---|---|---|---|
| scGPT | 0,255 | 0,236 | 0,289 | 0,532 |
| GeneCompass | 0,473 | 0,298 | 0,434 | 0,555 |
| Geneformer | 0,258 | 0,305 | 0,704 | 0,564 |
| scmap | 0,266 | 0,733 | 0,719 | 0,656 |
| scELMo | 0,571 | 0,637 | 0,798 | 0,84 |
| UCE | 0,774 | 0,604 | 0,56 | 0,922 |
| scBERT | 0,815 | 0,72 | 0,803 | 0,916 |
| scFoundation | 0,844 | 0,841 | 0,842 | 0,938 |
| SVM | 0,811 | 0,838 | 0,831 | 0,93 |
| CellFM(800M) | 0,68 | 0,505 | 0,592 | 0,846 |
| CellFM(80M) | 0,842 | 0,799 | 0,891 | 0,94 |
| C2S Pythia-410M | 0,77 | 0,72 | 0,687 | 0,926 |
| HVG C2S Pythia-410M | 0,771 | 0,686 | 0,624 | 0,932 |

Figure XI: Macro-F1 results on inter-dataset cell-type annotation

We can conclude from the results of our inter-dataset cell-type annotation experiment X, XI that the cell sentence representation and in particular the C2S Pythia-410M model offers competitive performances across all tissues, even though it stays below top performances models. On all four datasets, the full cell sentence variant consistently shows strong accuracy (range: 0.848-0.949) and Macro-F1 (range: 0.687-0.771), even rivaling with models like CellFM(80M) and scFoundation [7, 21]. The quite good Macro-F1 score is an indicator of the models' good performance on rare cell types, a crucial metric in very imbalanced datasets like Heart and Liver. What is also important to highlight here is that HVG-based cell sentence representation yields nearly identical results. Despite reducing the sentence length (divided by 2 or 3 in general), it shows minimal loss in both performances Accuracy and Macro-F1. This loss of performance is largely linked with a drop in performances regarding minority classes. It thus supports the computational viability of using HVG-based cell sentence representation when resource constraints require it, with only a modest trade-off in granularity.

We can now assess that the sentence-based representation allows a good generalization across batches, despite the bias introduced in the inter-dataset setup.

While full sentence embeddings maintain the advantage of containing more information, HVG-based sentences focus on essential information and provide a computationally efficient alternative with comparable meanwhile slightly lower performances, making this representation attractive for large scale or real time inference scenarios.

## III. Conclusion

In this study, we have demonstrated that embedding cell representations as gene sentences, whether it be based on full transcriptomes or only on highly variable genes selection, and processing them with an adapted large language model, in our case the C2S Pythia-410M, shows high performance for both cell reconstruction making this representation almost reversible and supervised cell-type annotation. Most importantly, HVG-based cell sentences preserve nearly the same accuracy and macro-F1 in intra-dataset and inter-dataset contexts, reflecting robust generalization, even when facing strong batch effect. The differences highlighted between full-gene and HVG-based cell sentences are offset by the gains in computational efficiency, reduced token length and noise filtering. These results lead us to assess that HVG selection which is a strongly established practice that enhance biological signal to noise ratio can serve as a scalable and practical foundation for downstream single-cell, LLM-based pipelines without sacrificing predictive power.

## IV. Discussion

In spite of the promising results highlighted in this study, there are several limitations that must be acknowledged. First and foremost, the pre-training data used for C2S Pythia-410M was not fully disclosed. According to the original documentation, it has been trained on over 57 million human and mouse cells from more than 800 datasets from CellxGene and the Human Cell Atlas. Therefore, without precise information on potential dataset overlap, we cannot preclude data leakage which might have led to inflating performances on our evaluation. This uncertainty definitely hampers a fully objective assessment of generalization from our representation. Second, we relied on the pretrained version of C2S Pythia-410M for both full-gene and HVG-based sentence representations, without fine-tuning separate models on each type of input. Fine-tuning dedicated versions on HVG-only versus full-gene sentences might yield more nuanced insights into how representation size and gene selection influence performance, particularly on rare cell types or under batch shift. To strengthen future work, ensuring that test datasets are strictly held out from all pretraining data is fundamental to validating performance. Conducting fine-tuning exercises separately on HVG and full-gene sentences would clarify the representational trade-offs inherent to each variant. Expanding benchmarking to include more tissues and cross-species studies would better reflect real-world diversity. Finally, incorporating interpretability tools like token importance scoring or embedding attention mappings would help explain model errors and guide improvements.

## Acknowledgement

I would like to express my deepest gratitude to Professor Seiya Imoto, Head of the Division of Health Medical Intelligence at the Institute of Medical Science, The University of Tokyo, for his guidance and support throughout this project. I am also sincerely grateful to Professor Yaozhong Zhang, researcher at the Division of Health Medical Intelligence and my supervisor, whose invaluable expertise and insightful advice were essential to the success of this work. Finally, I wish to warmly thank Dr. Yusri Dwi Heryanto, PhD researcher, for his precious help, constructive discussions, and constant encouragement during the accomplishment of this project.

## References

[1] Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, Roed M, Alexandrov T, AlQuraishi M, Brennan P, Burkhardt DB, Califano A, Cool J, Dernburg AF, Ewing K, Fox EB, Haury M, Herr AE, Horvitz E, Hsu PD, Jain V, Johnson GR, Kalil T, Kelley DR, Kelley SO, Kreshuk A, Mitchison T, Otte S, Shendure J, Sofroniew NJ, Theis F, Theodoris CV, Upadhyayula S, Valer M, Wang B, Xing E, Yeung-Levy S, Zitnik M, Karaletsos T, Regev A, Lundberg E, Leskovec J, Quake SR. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell. 2024 Dec 12;187(25):7045-7063. doi: 10.1016/j.cell.2024.11.015. PMID: 39672099; PMCID: PMC12148494.

[2] Levine D, Rizvi SA, Lévy S, Pallikkavaliyaveetil N, Zhang D, Chen X, Ghadermarzi S, Wu R, Zheng Z, Vrkic I, Zhong A, Raskin D, Han I, de Oliveira Fonseca AH, Caro JO, Karbasi A, Dhodapkar RM, van Dijk D. Cell2Sentence: Teaching Large Language Models the Language of Biology. bioRxiv [Preprint]. 2024 Oct 29:2023.09.11.557287. doi: 10.1101/2023.09.11.557287. PMID: 39554079; PMCID: PMC11565894.

[3] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746. doi: 10.15252/msb.20188746. PMID: 31217225; PMCID: PMC6582955.

[4] Cuevas-Diaz Duran R, Wei H, Wu J. Data normalization for addressing the challenges in the analysis of single-cell transcriptomic datasets. BMC Genomics. 2024 May 6;25(1):444. doi: 10.1186/s12864-024-10364-5. PMID: 38711017; PMCID: PMC11073985.

[5] Liu T, Li K, Wang Y, Li H, Zhao H. Evaluating the Utilities of Foundation Models in Single-cell Data Analysis. bioRxiv [Preprint]. 2024 Dec 10:2023.09.08.555192. doi: 10.1101/2023.09.08.555192. PMID: 38464157; PMCID: PMC10925156.

[6] Zhenqiu Shu, Yixuan Ren, Qinghan Long, Hongbin Wang, and Zhengtao Yu Journal of Chemical Information and Modeling 2025 65 (12), 6367-6381 DOI: 10.1021/acs.jcim.5c00731

[7] Zeng, Y., Xie, J., Shangguan, N. et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. Nat Commun 16, 4679 (2025). https://doi.org/10.1038/s41467-025-59926-5

[8] Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020). https://doi.org/10.1186/s13059-020-1926-6

[9] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[10] Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016 Aug 31;5:2122. doi: 10.12688/f1000research.9501.2. PMID: 27909575; PMCID: PMC5112579.

[11] Pullin, J.M., McCarthy, D.J. A comparison of marker gene selection methods for single-cell RNA sequencing data. Genome Biol 25, 56 (2024). https://doi.org/10.1186/s13059-024-03183-0

[12] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013 Nov;10(11):1093-5. doi: 10.1038/nmeth.2645. Epub 2013 Sep 22. Erratum in: Nat Methods. 2014 Feb;11(2):210. PMID: 24056876.

[13] Gatlin, V., Gupta, S., Romero, S. et al. Exploring cell-to-cell variability and functional insights through differentially variable gene analysis. npj Syst Biol Appl 11, 29 (2025). https://doi.org/10.1038/s41540-025-00507-z

[14] Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0

[15] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024 Aug;21(8):1470-1480. doi: 10.1038/s41592-024-02201-0. Epub 2024 Feb 26. PMID: 38409223.

[16] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, Ellinor PT. Transfer learning enables predictions in network biology. Nature. 2023 Jun;618(7965):616-624. doi: 10.1038/s41586-023-06139-9. Epub 2023 May 31. PMID: 37258680; PMCID: PMC10949956.

[17] Yang X, Liu G, Feng G, Bu D, Wang P, Jiang J, Chen S, Yang Q, Miao H, Zhang Y, Man Z, Liang Z, Wang Z, Li Y, Li Z, Liu Y, Tian Y, Liu W, Li C, Li A, Dong J, Hu Z, Fang C, Cui L, Deng Z, Jiang H, Cui W, Zhang J, Yang Z, Li H, He X, Zhong L, Zhou J, Wang Z, Long Q, Xu P; X-Compass Consortium; Wang H, Meng Z, Wang X, Wang Y, Wang Y, Zhang S, Guo J, Zhao Y, Zhou Y, Li F, Liu J, Chen Y, Yang G, Li X. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. Cell Res. 2024 Dec;34(12):830-845. doi: 10.1038/s41422-024-01034-y. Epub 2024 Oct 8. PMID: 39375485; PMCID: PMC11615217.

[18] Liu, T., Chen, T., Zheng, W., Luo, X. & Zhao, H. Scelmo: embeddings from language models are good learners for single-cell data analysis. Preprint at bioRxiv https://doi.org/10.1101/2023.12.07. 569910 (2023).

[19] Rosen, Y. et al. Universal cell embeddings: a foundation model for cell biology. bioRxiv Preprint at https://doi.org/10.1101/2023.11.28. 568918 (2023).

[20] Yang, F. et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat. Mach. Intell. 4, 852–866 (2022)

[21] Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. Nat. methods 21, 1481–1491 (2024).