# Improving Cell2Sentence with Single-Cell Foundation Model Embeddings

Daniel Rettberg Soler (10082166)
Johann Bartels (10073046)
AI Foundation Models in Biomedicine | Leibniz Universität Hannover

January 14, 2026

**Abstract**

Large language model–based approaches such as Cell2Sentence use transformer architectures to learn representations from single-cell RNA sequencing (scRNA-seq) data. While Cell2Sentence was originally introduced for generating natural language descriptions of cells, its underlying representations can also be applied to downstream tasks such as cell label prediction. In its original formulation, Cell2Sentence relies on cell representations derived directly from gene expression profiles, which may not fully capture higher-level biological structure such as pathways, cell states, or regulatory programs [4]. At the same time, single-cell foundation models (scFMs) such as scGPT, Geneformer, and scFoundation [1, 5, 3] have been pretrained on large-scale scRNA-seq data and are known to learn biologically meaningful embeddings of cells. This project investigates whether replacing the original Cell2Sentence cell representations with embeddings extracted from pretrained scFMs can improve the quality of generated cellular descriptions. We propose a comparison between the original Cell2Sentence setup and several variants that use scFM-based embeddings as input to the text decoder. All models are evaluated on a cell label prediction task using standard classification metrics.

## 1 Problem Setting

Cell2Sentence learns latent representations of cells from single-cell gene expression profiles for tasks like cell type prediction. Its original embeddings are based on raw gene expression and may miss higher-level biological structure. In this project, we test whether replacing them with pretrained single-cell foundation model embeddings improves downstream cell label prediction by providing richer, more informative cell representations. The core research question of this project is: Does using single-cell foundation model embeddings as cell representations improve Cell2Sentence's performance on cell label prediction compared to the original gene-expression-based encoding?

## 2 Main reference papers or resources

These models provide pretrained cell embeddings that can be used as alternative representations for downstream tasks such as text generation.

- **scGPT:** A transformer-based single-cell foundation model pretrained on large-scale scRNA-seq datasets, designed to capture gene–gene and cell–cell relationships. [1]
- **Geneformer:** A foundation model trained on gene expression data using masked language modeling, with demonstrated performance on downstream biological tasks. [5]
- **scFoundation:** A large pretrained model focusing on universal single-cell representations across diverse biological contexts. [3]

The Cell2Sentence serves as the baseline model in this project that generates the natural language descriptions from scRNA-seq data using transformer-based architectures. [4]

## 3 Model Design

We evaluate multiple variants of Cell2Sentence that differ only in the cell representation used as input. The baseline model uses the original Cell2Sentence encoder, while the proposed variants replace this encoder with pretrained single-cell foundation models to extract fixed-size cell embeddings. These embeddings are provided to the unchanged Cell2Sentence decoder and classification head. All foundation model encoders are kept frozen during training to ensure a fair comparison.

## 4 Evaluation

We evaluate all model variants on the cell label prediction task using a subset from two donors of the Immune System tissue dataset by Domínguez Conde et al. [2] The Performance is measured using exact combinatorial label accuracy, partial label accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC). Exact accuracy requires all components of a multi-part cell label to be predicted correctly, while partial accuracy assigns credit for correctly predicting individual label components. To ensure robustness and reproducibility, all experiments are conducted across multiple runs with different random seeds, and average performance is reported. This should allow a controlled comparison of how different cell representations influence downstream label prediction performance.

## 5 Expected Outcomes

We hypothesize that embeddings from single-cell foundation models (scFMs) [1, 5, 3] provide richer biological context than representations derived directly from raw gene expression, resulting in improved cell-type prediction. Even if gains are modest or vary across datasets, the comparative analysis will clarify how representation quality affects downstream performance in biologically grounded language-based tasks.

Specifically, we expect scFM-based embeddings to improve Cell2Sentence's [4] performance on the cell-label prediction task relative to the original Cell2Sentence encoder. We anticipate higher exact and partial combinatorial label accuracy as well as improved AUROC, consistent with scFM embeddings capturing more informative and transferable biological features.

## References


[1] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, Bo Wang, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

[2] Cecilia Domínguez Conde, Chuan Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, J. L. Jones, and S. A. Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376 (6594):713–+, 2022. doi: 10.1126/science.abl5197. URL https://www.science.org/doi/full/10.1126/science.abl5197.

[3] Meng Hao et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:xxx–xxx, 2024. doi: 10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

[4] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, Ruiming Wu, Zihe Zheng, et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv*, 2023. doi: 10.1101/2023.09.11.557287. URL https://www.biorxiv.org/content/10.1101/2023.09.11.557287v4.

[5] Christina V. Theodoris, Ling Xiao, Anant Chopra, M Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.