Multimodal Language Modeling for High-Accuracy Single Cell Transcriptomics Analysis and Generation

Anonymous ACL submission

Abstract

Pre-trained language models (PLMs) have revolutionized scientific research, yet their application to single-cell analysis remains limited. Text PLMs cannot process single-cell RNA sequencing data, while cell PLMs lack the ability to handle free text, restricting their use in multimodal tasks. Existing efforts to bridge these modalities often suffer from information loss or inadequate single-modal pre-training, leading to suboptimal performances. To address these challenges, we propose **S**ingle-**C**ell **M**ulti**M**odal **G**enerative **P**re-trained **T**ransformer (**scMMGPT**), a unified PLM for joint cell and text modeling. scMMGPT effectively integrates the state-of-the-art cell and text PLMs, facilitating cross-modal knowledge sharing for improved performance. To bridge the text-cell modality gap, scMMGPT leverages dedicated cross-modal projectors, and undergoes extensive pre-training on 27 million cells – the largest dataset for multimodal cell-text PLMs to date. This large-scale pre-training enables scMMGPT to excel in joint cell-text tasks, achieving an 84% relative improvement of textual discrepancy for cell description generation, 20.5% higher accuracy for cell type annotation, and 4% improvement in k-NN accuracy for text-conditioned pseudo-cell generation, outperforming baselines. Our code is available at `https://anonymous.4open.science/r/scMMGPT-6DDB/`.

1 Introduction

Pre-trained language models (PLMs) are transforming scientific research (Touvron et al., 2023a; OpenAI, 2023; Hurst et al., 2024). Their ability to recall scientific knowledge, analyze data, and perform mathematical reasoning helps to reduce manual efforts and lower the research barrier in many tasks. Notably, PLMs are opening new avenues for single-cell analysis, which explores the molecular and functional characteristics of individual cells. Previous studies have successfully employed PLMs for cell annotation (Li et al., 2024) and retrieval (Xu et al., 2023a; Lan et al., 2024), benefiting from the extensive cellular knowledge embedded in the PLMs' training corpus.

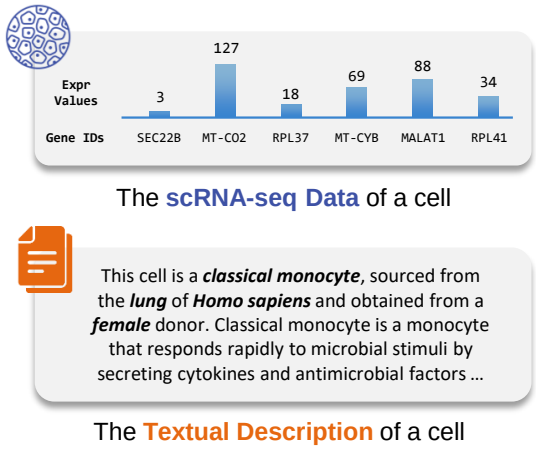

Figure 1: Comparison between the scRNA-seq results and the textual descriptions of a cell. The inherent disparities between these two data modalities make it difficult to jointly model them.

Beyond the text-based PLMs above, cell PLMs are also explored for single-cell analysis (Ji et al., 2021a; Abdolhosseini et al., 2019). As Figure 1 shows, a cell can be represented as an array of gene expression levels, providing insights into its biological properties. These arrays, known as scRNA-seq data, are generated through single-cell RNA sequencing (scRNA-seq) (Program et al., 2025). Pre-trained on scRNA-seq data, cell PLMs have been applied for batch effect correction, and pseudo-cell generation (Yang et al., 2022; Hao et al., 2024; Cui et al., 2024). However, cell PLMs are inherently limited by their inability to process free text, preventing them from integrating the rich single-cell knowledge in textual corpora and restricting their ability to perform text-guided cell generation and cell description generation.

To resolve this limitation, we want to develop a unified multimodal PLM unifying cell and text data

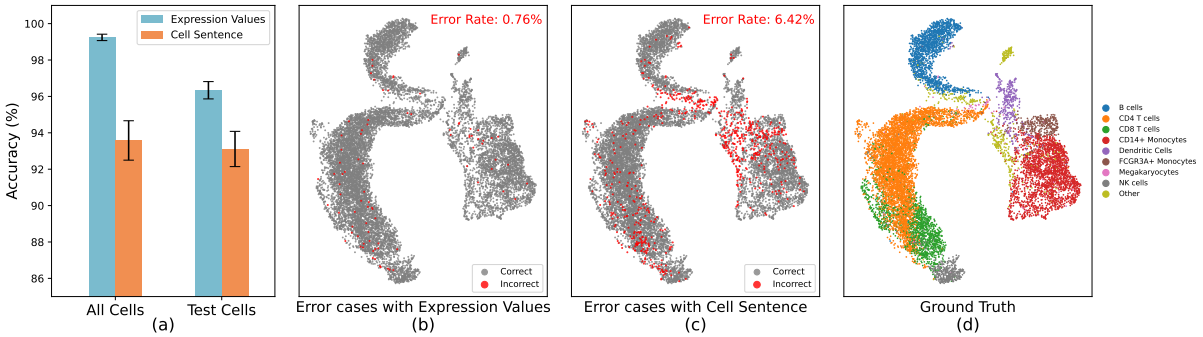

Figure 2: Cell type annotation results with different cell representation methods. (a) **Cell type annotation accuracies on the full dataset and test set.** Using cell sentences as cell representation leads to significant accuracy degradation. (b-d) **UMAP visualization of classification results and the ground truth.** Classification using cell sentences yields a lower accuracy score and exhibits poorer recognition capabilities in certain cell clusters.

for comprehensive single-cell analysis. While this area has been explored, we identify two common limitations of previous works:

- **The Information Loss of Cell Sentences.** Previous cell-text PLMs primarily represent single cells as "cell sentences" (Hou and Ji, 2024; Levine et al., 2024; Choi et al., 2024), where genes are ranked by expression level, and only the top 30-100 genes are retained as the cell's representation. This method captures less than 1% of the total gene annotations in modern databases (Program et al., 2025; Cao et al., 2017), which record over 10,000 genes. Additionally, cell sentences discard crucial information about gene expression values. As Figure 2 shows, this representation leads to significant performance degradation in cell type annotation compared to the original expression values representation, highlighting a crucial information loss.

- **Limited Single-Modal Pre-training.** Some previous works (Choi et al., 2024; Hou and Ji, 2024) are built on text PLMs without sufficient pre-training on scRNA-seq data, limiting their capacity for comprehensive single-cell analysis. While notably Levine et al. (2024) performs large-scale cell pre-training on a text PLM's checkpoint, it suffers from catastrophic forgetting, compromising its text processing ability (*cf.* Section 4.2).

In this work, we propose the **S**ingle-**C**ell **M**ulti**M**odal **G**enerative **P**re-trained **T**ransformer (**scMMGPT**), a novel PLM designed for the multimodal analysis and generation between cell and text. scMMGPT builds on the scGPT (Cui et al., 2024), an extensively pre-trained cell PLM capable of encoding the full scRNA-seq data. To address scGPT's inability to process text, we implement a cell-to-text projector that projects scGPT's representations to the text space, leveraging Llama-2 (Touvron et al., 2023b), a powerful text PLM, for text generation. For cell generation, we implement a text-to-cell projector to map Llama-2's representations to scGPT's cell space, providing textual guidance. These cross-modal projectors between cell and text enable effective information exchange between the two PLMs, leveraging their respective domain knowledge.

Given the advanced PLMs in the cell and text modalities, scMMGPT focuses on bridging their modality gaps by pre-training the two cross-modal projectors. To this end, we pre-train scMMGPT on 27 million cells from CELLxGENE (Program et al., 2025), which is the largest pre-training dataset for multimodal cell-text PLMs. Using this dataset, we pre-train the text-to-cell projector for text-conditioned cell generation, enabling it to map the textual embeddings to a space that the cell PLM can understand. Similarly, the cell-to-text projector is pre-trained for cell description generation. Before this step, we warmup the cell-to-text projector's weights with the additional pre-training of cell-text contrasting learning and matching. This step follows (Li et al., 2023), aiming to obtain a model for cell type annotation tasks.

Our tailor-made architecture and extensive pre-training equips scMMGPT with superior performance on various downstream tasks. It achieves an 84% relative improvement of textual discrepancy for cell description generation, a 4% k-NN accuracy improvement for text-conditioned pseudo-cell generation, and 20.5% higher accuracy for cell type annotation than baselines. Ablation studies further validate the effectiveness of the key components.

2 Related Works

Single-Cell PLMs. Single-cell sequencing technologies provide diverse biological features that facilitate the interpretation of cellular structures and functions (Heumos et al., 2023; Cao and Gao, 2022). Advances in scRNA-seq have generated massive, high-precision transcriptomic datasets, driving the development of cell PLMs (Ziegenhain et al., 2017). This technique quantifies the mRNA molecule abundance, producing gene expression matrices that record expression values of individual genes across cells (Ji et al., 2021b). Previous works have developed transformer-based foundation models on scRNA-seq data, pre-training with masked learning objectives on millions of cells (Zhao et al., 2023; Theodoris et al., 2023; Yang et al., 2022). Subsequent works improve the learning process by incorporating cell labels, such as batch effects (Cui et al., 2024). After fine-tuning, these PLMs have proven useful in practical downstream tasks including cell-type annotation, perturbation response prediction, and pathway activity inference.

Cell-Text Modeling. Textual descriptions of cells and scRNA-seq data capture complementary aspects of cellular systems. To jointly leverage this complementary information, prior research has explored enhancing cellular representation learning using biological text descriptions (Chen and Zou, 2023; Zhao et al., 2024). Inspired by multimodal PLMs in other scientific domains (Liu et al., 2023b; Edwards et al., 2022), cell-to-text translation is also exploredn (Xu et al., 2023a). Notably, the "cell sentence" representation (Levine et al., 2024) is introduced by transforming scRNA-seq data into textual token sequences, which are widely used in subsequent studies (Hou and Ji, 2024; Choi et al., 2024; Fang et al., 2024). However, cell sentences have substantial information loss, constraining the model's capacity to perceive fine-grained cellular transcriptomics.

Scientific Multimodal PLMs. Multimodal PLMs show remarkable potential for integrating data from various modalities (Li et al., 2023; Alayrac et al., 2022; Zhang et al., 2024), inspiring research for scientific modalities. Existing works have constructed multimodal PLMs for small molecules (Liu et al., 2023b, 2024a) and proteins (Xu et al., 2023b; Liu et al., 2024b) to tackle cross-modal scientific problems, such as description generation and text-conditioned de novo design (Edwards et al., 2022; Cao et al., 2025). While single-cell analysis presents similar scientific significance, existing works struggle to maintain information fidelity when integrating single-cell transcriptomics with textual knowledge. Unlike previous methods, scMMGPT employs PLMs on both modalities to precisely model the scRNA-seq data and textual tokens without information loss.

3 Methods

scMMGPT employs two PLMs for both cell and text modalities, facilitating the understanding and generation of cell and text through effective information sharing via cell-to-text and text-to-cell projectors. Figure 3 illustrates the model architecture of scMMGPT. In this section, we delve into the construction process of scMMGPT, including detailed data collection and encoding (§3.1), the multimodal PLMs of scMMGPT (§3.2), and the complete pre-training scheme (§3.3).

3.1 Data Preprocessing

Single-cell RNA sequencing (scRNA-seq) data records the gene expression levels across individual cells at transcriptomic resolution. These data can be represented as a cell-gene expression matrix $\mathbf{X} \in \mathbb{N}^{N \times M}$, where $\mathbf{X}_{ij}$ $(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M)$ denotes the RNA abundance of gene j in cell i. To accurately characterize the transcriptional state of a cell, we represent each cell using the list of genes profiled during sequencing and their corresponding expression values. Formally, for each cell i, we represent the list of genes as $\mathbf{g}^{(i)} = [g_1^{(i)}, g_2^{(i)}, \ldots, g_M^{(i)}]$, where each $g_j^{(i)}$ is a gene token from a pre-defined gene vocabulary (Cui et al., 2024). The corresponding RNA abundance information from the gene expression matrix is denoted as an expression value vector $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_M^{(i)}] \in \mathbb{R}^M$. To mitigate the influence of sequencing depths (Zhang et al., 2020), the expression value vectors then undergo a normalization step followed by a log1p transformation:

$$\widetilde{x}_j^{(i)} = \log(1 + \frac{x_j^{(i)}}{\sum_{k=1}^{M} x_k^{(i)}}). \quad (1)$$

We leverage the scRNA-seq data from the CellxGene Database (Program et al., 2025), and collect the metadata and textual descriptions for each cell i via their online explorer[1]. We further augment these descriptions with textual knowledge from

[1] https://github.com/chanzuckerberg/cellxgene

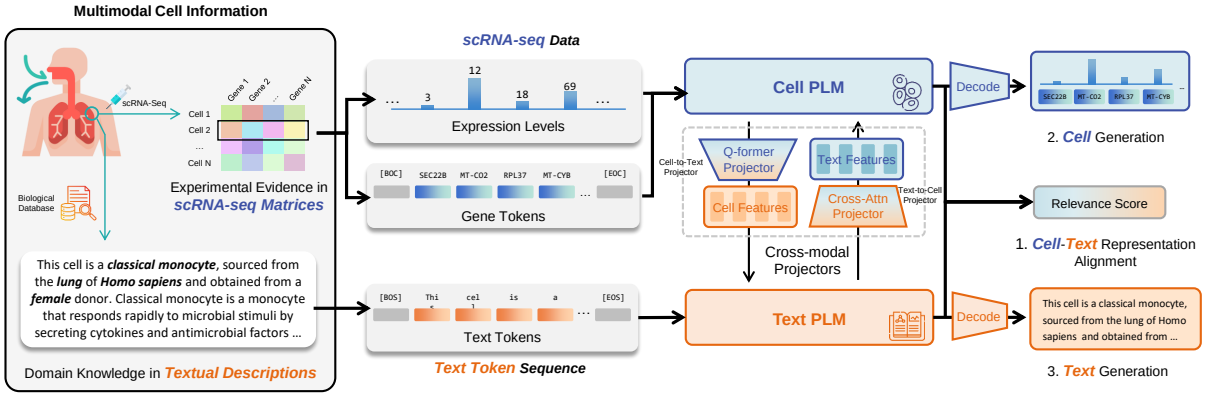

Figure 3: Overview of scMMGPT. scMMGPT utilizes a cell PLM and a text PLM to process corresponding modalities, undergoing a large-scale pre-training on three primary tasks: **(1) Cell-Text Representation Alignment:** scMMGPT receives inputs from both modalities and calculates a relevance score based on the output features of both PLMs. **(2) Cell Generation:** scMMGPT uses the text PLM to extract embeddings from the textual descriptions, which are then passed through a projector and a cross-attention layer to the cellular PLM for cell generation. **(3) Text Generation:** scMMGPT uses the cell PLM to parse the genes and expression values from scRNA-seq results. After a projector, these cellular embeddings are fed to the text PLM to decode as a textual description of the cell.

the Open Biomedical Ontologies Foundry (OBO Foundry) (Smith et al., 2007) and Wikipedia[2]. These textual descriptions are then processed using a text tokenizer into a textual tokens sequence $\mathbf{t}^{(i)} = [t_1^{(i)}, t_2^{(i)}, \ldots, t_T^{(i)}]$. The dataset details are in Appendix A.

3.2 Model Architecture

As shown in Figure 3, scMMGPT consists of three trainable components: (1) a cell PLM for the understanding and generation of cells, (2) a text PLM for the understanding and generation of textual data, and (3) the cross-modal projectors that facilitate information sharing between different modalities.

Cell PLM. We utilize a transformer-based cell PLM, scGPT (Cui et al., 2024), to process the gene tokens and expression values unique to scRNA-seq data. scGPT undergoes generative pre-training on over 33 million single-cell samples (Program et al., 2025), with training objectives including gene expression prediction and cell generation. This model encapsulates rich knowledge in the domain of single-cell analysis and has been validated for downstream tasks of cell type annotation and text-conditioned pseudo-cell generation (Levine et al., 2024).

Text PLM. To facilitate high-quality text generation, we utilize a decoder-only generative transformer, Llama-2 (Touvron et al., 2023b), as our text PLM. This model is pre-trained on 2 trillion tokens of publicly available web data, incorporating extensive human knowledge across diverse domains.

Cross-Modal Projectors. We employ a cell-to-text and a text-to-cell projector to achieve representation transformation between the cellular and textual PLMs to bridge their modality gap:

- The Cell-to-Text Projector is implemented as a Querying-Transformer (Q-Former) (Li et al., 2023) to map the cell representations generated by the cell PLM into the input space of the text PLM. Q-Former maintains a set of trainable query tokens that interact with the output embeddings of the cell PLM through a cross-attention mechanism. The parameters of the Q-Former are initialized using BiomedBERT (Gu et al., 2021), a BERT encoder trained with biomedical scientific abstracts and literature from PubMed (Canese and Weis, 2013).

- The Text-to-Cell Projector is implemented by cross-attention layers (Vaswani, 2017) to map the textual representations produced by the text PLM into the feature space of the cell PLM. These textual features then serve as soft-prompt (Li and Liang, 2021) to the cell PLM, providing conditions for downstream tasks such as text-conditioned pseudo-cell generation.

3.3 Training Pipeline

scMMGPT's pre-training objective is to bridge the embedding spaces of scRNA-seq and text data,

[2] https://www.wikipedia.org/

thereby enabling it to perform cross-modal conditional generation. To achieve this, we first employ the cell-text representation alignment task to align cellular and textual representations. Then, we perform generative pre-training for each modality, including cell generation and text generation objectives.

Cell-Text Representation Alignment. Considering the intrinsic variability of cellular data, the pre-training of cell-text representation alignment includes only the cell-to-text projector without the text-to-cell projector. This pre-training process involves two objectives: cell-text contrastive learning and cell-text matching.

The Cell-Text Contrastive objective (CTC) aims to map cells and text into a shared feature space. We use a projector to extract representations for both cells and text. Using the matching cell and text as the positive examples, we use other cells within the same batch as negative examples, and construct the loss function based on InfoNCE (Oord et al., 2018):

$$\begin{aligned}\mathcal{L}_{\text{CTC}} = -\frac{1}{B}\sum_{i=1}^{B}(\log \frac{e^{\cos(\mathbf{z}_t^{(i)},\mathbf{z}_c^{(i)})/\tau}}{\sum_{j=1}^{B} e^{\cos(\mathbf{z}_t^{(i)},\mathbf{z}_c^{(j)})/\tau}} \\ + \log \frac{e^{\cos(\mathbf{z}_c^{(i)},\mathbf{z}_t^{(i)})/\tau}}{\sum_{j=1}^{B} e^{\cos(\mathbf{z}_c^{(i)},\mathbf{z}_t^{(j)})/\tau}}),\end{aligned} \tag{2}$$

where $\mathbf{z}_t^{(i)}$ and $\mathbf{z}_c^{(i)}$ represent the textual and cellular embeddings, τ is the temperature parameter, and B indicates batch size.

The Cell-Text Matching (CTM) objective is designed as a classification task, where the model learns to predict whether a given cell and text are matched or not. In this task, we let the textual representations in the projector interact with cell representations through cross-attention layers in our Q-Former projector. Formally, the cell-text matching loss for a batch can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{CTM}} = \frac{1}{B}\mathbb{E}_{j\sim \text{U}(1,B)}\sum_{k=1}^{B} -\log \rho(\mathbf{c}^{(i)},\mathbf{t}^{(i)}) \\ + \log \rho(\mathbf{c}^{(i)},\mathbf{t}^{(j)}),\end{aligned} \tag{3}$$

where $\text{U}(\cdot)$ is the uniform distribution and $\rho(\mathbf{c}^{(i)},\mathbf{t}^{(i)})$ denotes the model's predicted probability of $(\mathbf{c}^{(i)},\mathbf{t}^{(i)})$ being matched. The overall training loss for cell-text representation alignment task combines $\mathcal{L}_{\text{CTC}}$ and $\mathcal{L}_{\text{CTM}}$:

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{CTC}} + \mathcal{L}_{\text{CTM}}. \tag{4}$$

Cell Description Generation (CDGen). CDGen aims to generate the corresponding text descriptions for a given cell. The cells are first mapped to the embedding space of text tokens using a cell-to-text projector. Then, the decoder-only text PLM performs autoregressive next token prediction starting from these cell tokens to generate a description of the cell. Formally, the objective of this task is to minimize the loss function of the autoregressive language modeling:

$$\begin{aligned}\mathcal{L}_{\text{CDGen}} &= -\log p(\mathbf{t}^{(i)}|\mathbf{c}^{(i)}) \\ &= -\sum_{l=1}^{L}\log p(t_l^{(i)}|t_1^{(i)},\ldots,t_{l-1}^{(i)},\mathbf{c}^{(i)}).\end{aligned} \tag{5}$$

Cell Generation (CGen). In the CGen task, the model performs conditional cell generation based on textual descriptions. We append some dummy cell tokens at the end of each piece of cell description and use the text PLM to autoregressively generate the embeddings corresponding to these dummy tokens based on the text sequence ahead. These features pass through the text-to-cell projector to produce embeddings of textual conditions, which are then fed into the cell PLM via cross-attention mechanism to generate pseudo-cells $\mathbf{x}'^{(i)} = [x_1'^{(i)},\ldots,x_M'^{(i)}]$.

Formally, we use Mean Squared Error (MSE) loss to train the model to reconstruct the input cellular information:

$$\mathcal{L}_{\text{CGen}} = \sum_{j=1}^{M}\text{MSE}(x_j'^{(i)},\widetilde{x}_j^{(i)}). \tag{6}$$

During the cross-modal generative training, we jointly optimize $\mathcal{L}_{\text{CDGen}}$ and $\mathcal{L}_{\text{CGen}}$ through a linear combination of the two loss functions:

$$\mathcal{L}_{\text{Gen}} = \mathcal{L}_{\text{CDGen}} + \mathcal{L}_{\text{CGen}}. \tag{7}$$

4 Experiments

We empirically evaluate scMMGPT on three downstream tasks: cell description generation, text-conditioned pseudo-cell generation, and cell type annotation. Furthermore, we perform ablation studies to illustrate the impacts of different input formats and model architectures. Additionally, we visualize experimental outcomes and training details to provide a better understanding of the model.

Model	MMD (↓)	EMD (↓)	T-Test (↑)	KS-Test (↑)
GPT-2 Small	1.045 ± 0.009	0.752 ± 0.004	1.31, $p = 0.896$	1.52, $p = 0.783$
GPT-2 Large	0.939 ± 0.006	0.701 ± 0.016	-1.44, $p = 0.885$	1.81, $p = 0.581$
Mistral-8X7B-Instruct	0.639 ± 0.016	0.544 ± 0.005	-1.20, $p = 0.233$	0.24, $p = 0.246$
Mistral-7B-Instruct	0.754 ± 0.010	0.584 ± 0.004	-8.64, $p = 0.384$	0.23, $p = 0.299$
GPT-3.5	0.298 ± 0.004	0.490 ± 0.008	1.23, $p = 0.220$	0.21, $p = 0.392$
Cell2Sentence Small	0.198 ± 0.004	0.414 ± 0.006	2.96, $p = 0.003$*	0.35, $p = 0.023$*
Cell2Sentence Large	0.198 ± 0.004	0.413 ± 0.002	2.85, $p = 0.004$*	0.36, $p = 0.014$*
scMMGPT	**0.031 ± 0.002**	**0.011 ± 0.000**	**29.57, $p = 0.000$***	**0.62, $p = 0.000$***

Table 1: Results of cell description generation on the immune tissue (Domínguez Conde et al., 2022) dataset. Asterisks (*) denotes statistical significance ($p \leqslant 0.05$). Baseline results are borrowed from (Levine et al., 2024).

	k-NN Accuracy			
Model	3	5	10	25
scGEN	0.2376 ± 0.0112	0.2330 ± 0.0093	0.2377 ± 0.0053	0.2335 ± 0.0041
scVI	0.2436 ± 0.0062	0.2400 ± 0.0064	0.2425 ± 0.0034	0.2348 ± 0.0032
scDiffusion	0.2335 ± 0.0125	0.2288 ± 0.0111	0.2368 ± 0.0067	0.2306 ± 0.0049
scGPT	0.1838 ± 0.0086	0.1788 ± 0.0169	0.1811 ± 0.0149	0.1882 ± 0.0071
Cell2Sentence	0.2588 ± 0.0061	0.2565 ± 0.0060	0.2746 ± 0.0073	0.2715 ± 0.0070
scMMGPT	**0.2996 ± 0.0065**	**0.2992 ± 0.0055**	**0.2986 ± 0.0038**	**0.2981 ± 0.0051**

Table 2: Results of text-conditioned pseudo-cell generation on the immune tissue dataset. The baseline results are borrowed from (Levine et al., 2024).

4.1 Experiment Setup

Unless otherwise mentioned, we use the whole-human checkpoint of scGPT to initialize the cell PLM, and Llama-2 7B (Touvron et al., 2023b) for the text PLM. The model is then pre-trained according to Section 3.3, followed by fine-tuning or zero-shot inference on different downstream tasks. Unless otherwise specified, the text PLM is pre-trained and fine-tuned with a LoRA (Hu et al., 2022) adapter, while the cell PLM and the projectors undergo full-parameter training. More implementation details are provided in Appendix B.

Pre-training Dataset. We collect 27 million single-cell data from the CellxGene (Program et al., 2025) database for the pre-training of scMMGPT, including scRNA-seq matrices and corresponding metadata annotations. The textual descriptions of each cell are generated with cellular metadata and the Open Biological and Biomedical Ontology Foundry (Smith et al., 2007). We exclude the test sets of the downstream datasets and reserve 1,000 samples for validation, resulting in approximately 26.9 million cells for pre-training. Further details about data distribution and preprocessing protocols are recorded in Appendix A.

4.2 Cell Description Generation

The cell description generation task evaluates a model's ability to generate accurate and meaningful textual descriptions of cells provided their scRNA-seq data. We perform fine-tuning and evaluation on the immune tissue (Domínguez Conde et al., 2022) dataset using baseline methods including GPT-2 (Radford et al., 2019), Mistral 7B (Jiang et al., 2023), Mixtral 8x7B (Jiang et al., 2024), GPT-3.5, and Cell2Sentence (Levine et al., 2024). To evaluate the generation quality, we compute the Maximum Mean Discrepancy (MMD) and Earth Mover's Distance (EMD) between the textual embedding (Xiao et al., 2024) between predicted and ground truth descriptions. Additionally, we conduct a T-test and Kolmogorov-Smirnov test (KS-test) to statistically assess and confirm whether the generated descriptions are significantly closer to the original annotations compared to those from unrelated cell descriptions.

The performances are shown in Table 1. Our model significantly outperforms all baselines, achieving an 84% reduction in MMD (0.031) and a 97% reduction in EMD (0.011) compared to the best baseline model. scMMGPT also demonstrates lower standard deviations in both MMD and EMD compared to the baselines, suggesting greater robustness and consistency in its performance. The T-test and KS-test results further reveal highly significant p-values ($p \ll 0.05$), indicating a strong alignment between the generated and original descriptions. These results demonstrate scMMGPT's superior capability in understanding cellular states. For more results about cell description generation, see the ablation studies in Section 4.5.

Model	Zero-Shot			Fine-tuned on 10% Types			Fine-tuned on 20% Types			Fine-tuned on 30% Types		
	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10
Random	0.6	3.1	6.2	0.6	3.1	6.2	0.6	3.1	6.2	0.6	3.1	6.2
BioTranlator	-	-	-	3.5	33.6	45.4	13.4	48.2	63.5	13.7	50.6	68.6
LangCell	28.6	69.2	82.9	30.5	71.0	83.7	35.0	74.6	86.4	38.2	83.0	92.1
scMMGPT	**49.1**	**83.1**	**91.1**	**55.7**	**89.2**	**96.0**	**59.7**	**90.4**	**96.8**	**60.9**	**93.6**	**98.4**

Table 3: Results of cell type annotatia 84% on (%) on the Tabula Sapiens (Consortium* et al., 2022) dataset. The models are fine-tuned on a certain proportion of test cell types to evaluate their generalization performance. For example, "Fine-tuned on 20% Types" indicates that 20% of the cell types in the test set are used in the fine-tuning process, while the remaining 80% are used for testing. The baseline results are borrowed from (Zhao et al., 2024).

Model	Cell Representation	MMD (↓)	EMD (↓)	BLEU-2 (↑)	ROUGE-2 (↑)
scMMGPT (TinyLlama$_{1.1B}$, w/o scGPT)	Cell Sentence	0.104	0.023	45.79%	40.11%
scMMGPT (TinyLlama$_{1.1B}$)	Expression Values	0.074	0.021	48.77%	42.03%
scMMGPT (Llama-2$_{7B}$)	Expression Values	**0.031**	**0.011**	**77.32%**	**72.49%**

Table 4: Results of ablation studies on cell description generation task. We compare different cell representation methods (cell sentence v.s. expression values) and different backbone text LMs

4.3 Text-guided Pseudo-cell Generation

We conduct cell generation experiments on the immune tissue (Domínguez Conde et al., 2022) dataset. We select several generative single-cell models as baselines, including scGen (Lotfollahi et al., 2019), scVI (Lopez et al., 2018), scDiffusion (Luo et al., 2024), scGPT (Cui et al., 2024), and Cell2Sentence (Levine et al., 2024). Inspired by previous studies, we train a simple k-Nearest Neighbors (k-NN) classifier on the test set to distinguish the generated cells. The classification accuracies under different k values are reported to reflect the quality of the generated cells.

The results are presented in Table 2. scMMGPT achieves state-of-the-art performance in text-conditioned pseudo-cell generation, significantly outperforming all baseline models across all k-NN accuracies (k=3,5,10,25). The consistently high accuracy and low standard deviations of scMMGPT demonstrate its robustness and effectiveness in bridging cellular and textual data.

4.4 Cell Type Annotation

In the cell type annotation task, we evaluate the model's ability to classify cells based on their scRNA-seq data and textual descriptions of specific cell types. The goal is to assign each cell to its correct category, leveraging both the transcriptomic and textual modalities. We compare our model against two baseline methods BioTranslator (Xu et al., 2023a) and LangCell (Zhao et al., 2024) on the Tabula Sapiens (Consortium* et al., 2022) dataset. This dataset comprises 161 distinct human cell types, most of which are absent from our pre-training corpus. To assess the model's generalization performance, we report the classification accuracies under varying fine-tuning conditions, where the model is fine-tuned on different proportions of test cell types (10%, 20%, and 30%).

The experimental results are summarized in Table 3. In the zero-shot setting, scMMGPT achieves an Acc@1 of 49.1% and an Acc@5 of 83.1%, surpassing all the baseline models even in a fine-tuning setting. As shown in the table, the accuracy of scMMGPT under a zero-shot setting is even higher than many of the fine-tuned baseline results. As the proportion of fine-tuning cell types increases, scMMGPT consistently improves its performance across all metrics, reaching a maximum Acc@1 of 60.9% when fine-tuned on 30% cell types, almost doubling the accuracy of the state-of-the-art models. These results demonstrate that scMMGPT's pre-trained knowledge of both cellular and textual data enables strong generalization to unseen cell types without additional fine-tuning.

4.5 Ablation Studies

To systematically evaluate the impact of the cell PLM and different text PLM backbones on our model's performance, we conduct comprehensive ablation studies in this section.

Impact of the Cell PLM. To validate the effectiveness of the cell PLM within the scMMGPT framework, we conducted a comparative experiment between two model configurations: (1) the full scMMGPT, which integrates both the cell PLM and the text model, and (2) a text-only variant of scMMGPT, which excludes the cell PLM and use the text PLM only. Both models were trained using the same settings and evaluated on the task of cell

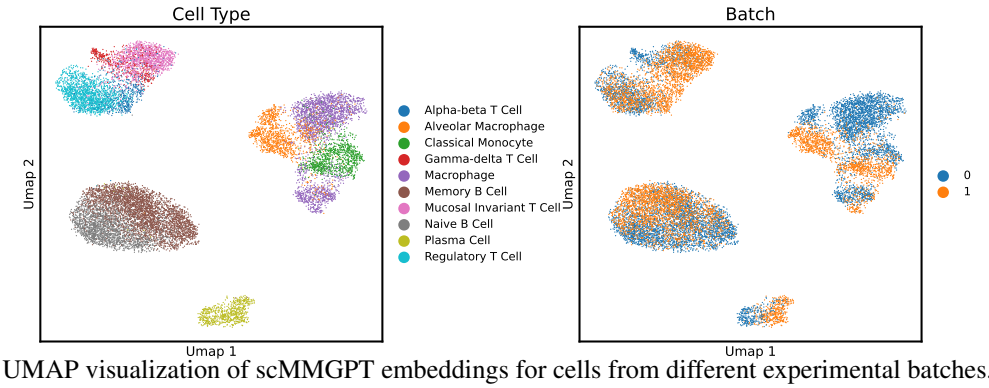

Figure 4: UMAP visualization of scMMGPT embeddings for cells from different experimental batches. The result demonstrates the model's ability to capture cell type distinctions while effectively mitigating batch effects.

description generation. As shown in Table 4, the full scMMGPT model significantly outperforms the text-only variant across multiple metrics, including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Specifically, the inclusion of the cell PLM improves BLEU-2 and ROUGE-2 scores by 3% and 2%, respectively, while reducing text distances between generated and ground truth descriptions. These results highlight the critical role of the cell PLM in capturing and leveraging detailed transcriptomic information.

Impact of Different Text PLM Backbones. To investigate the influence of model size, we perform experiments with two different text LMs: (1) TinyLlama 1.1B and (2) Llama2-7B. As shown in Table 4, though using a smaller LM causes performance drops of scMMGPT, its MMD and EMD scores still surpass those of the best baseline model, Cell2Sentence Large (see Table 1).

4.6 Visualization

Influence of Different Cell Representation Methods. To further quantify the information loss in cell sentences, we conduct a visualization experiment comparing cell sentence inputs with original expression values. Specifically, we train two separate MLPs with identical hyperparameters for cell type annotation on the PBMC10K[3] dataset. As shown in Figure 2, the cell sentence representation leads to a significant increase in error rate, particularly when distinguishing morphologically similar cell types such as dendritic cells and FCGR3A+ monocytes. This finding highlights the non-negligible cellular information lost during the transformation from numerical expression levels to cell sentences, which limits the effectiveness of related models in downstream applications.

[3] https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k

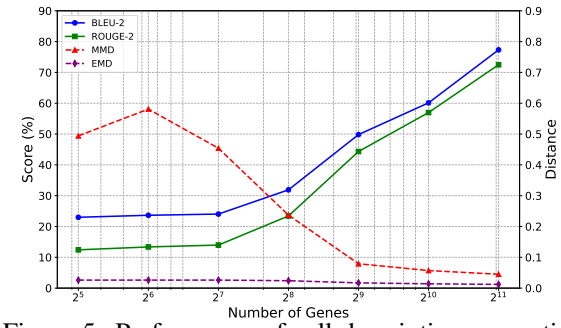

Figure 5: Performance of cell description generation across different numbers of genes used for cell representation. The x-axis represents the number of genes (log scale). Solid lines represent BLEU and ROUGE scores corresponding to the left axis. Dashed lines represent the text distances corresponding to the right axis.

Influence of Gene Set Size. We explore the effect of number of genes used for cell representation by selecting the top-k most highly expressed genes and their associated expression levels as input to scMMGPT. The relationship between scRNA-seq length and captioning performance is illustrated in Figure 5, where k ranges from 2^5 to 2^{11}. The results show a consistent improvement in performance as the input length increases, confirming that more detailed transcriptomic information positively impacts model predictions.

Batch Effect Mitigation in scMMGPT Embeddings. In wet lab experiments, it is challenging to maintain identical experimental conditions across different batches, which can lead to variations in the measured scRNA-seq data. To evaluate scMMGPT's robustness against batch effects, we analyzed two sets of immune tissue samples from different experimental batches, each containing ten randomly selected cell types. We computed scMMGPT embeddings for these samples and visualized them using UMAP, as shown in Figure 4. The results demonstrate that cell embeddings from scMMGPT effectively capture cell type differences while minimizing the influence of batch effects.

Conclusion and Future Works

In this work, we propose scMMGPT, a novel multimodal framework for single-cell analysis. scMMGPT bridges scRNA-seq data and text to support tasks of cell description generation, text-guided cell generation, and cell type annotation. This is achieved by integrating a cell PLM with a text PLM through cross-modal projectors. Pre-trained on 27 million cells from the CELLxGENE dataset, scMMGPT demonstrates superior performances across various single-cell analysis tasks. Looking forward, we will expand scMMGPT to incorporate more species, and integrate other cell modalities, like scATAC-seq and CITE-seq. This expansion will enable scMMGPT to tackle more challenges of multi-omic integration (Lotfollahi et al., 2022), cross-omic translation (Liu et al., 2023a), and novel cell type discovery (Yang et al., 2022), further enhancing its utility in single-cell research.

Limitations

One significant limitation of scMMGPT is that its pre-training data primarily sourced from the CELLxGENE (Program et al., 2025) dataset, which predominantly covers human tissues. This focus restricts scMMGPT's ability to incorporate knowledge about cells from non-human species, such as those from the widely-used mouse data (Franzén et al., 2019).

Another major limitation is that scMMGPT exclusively explores transcriptomic information from cells, lacking integration with other single-cell sequencing modalities, such as scATAC-seq and CITE-seq (Liu et al., 2023a; Lin et al., 2022). This constraint limits the model to analyzing RNA abundance alone, omitting critical perspectives on chromatin accessibility and protein expression within cells. Incorporating these additional modalities could provide a more comprehensive understanding of cellular states and functions.

References

Farzad Abdolhosseini, Behrooz Azarkhalili, Abbas Maazallahi, Aryan Kamal, Seyed Abolfazl Motahari, Ali Sharifi-Zarchi, and Hamidreza Chitsaz. 2019. Cell identity codes: understanding cell identity from gene expression profiles using deep neural networks. *Scientific reports*, 9(1):2342.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2025. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 354–379.

Yuan Cao, Junjie Zhu, Peilin Jia, and Zhongming Zhao. 2017. scrnaseqdb: a database for rna-seq based gene expression profiles in human single cells. *Genes*, 8(12):368.

Zhi-Jie Cao and Ge Gao. 2022. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466.

Yiqun Chen and James Zou. 2023. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10.

Hongyoon Choi, Jeongbin Park, Sumin Kim, Jiwon Kim, Dongjoo Lee, Sungwoo Bae, Haenara Shin, and Daeseung Lee. 2024. Cellama: Foundation model for single cell and spatial transcriptomics by cell embedding leveraging language model abilities. *bioRxiv*, pages 2024–05.

The Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. 2022. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11.

C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. 2022. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413.

Yin Fang, Kangwei Liu, Ningyu Zhang, Xinle Deng, Penghui Yang, Zhuo Chen, Xiangru Tang, Mark Gerstein, Xiaohui Fan, and Huajun Chen. 2024. Chatcell: Facilitating single-cell analysis with natural language. *arXiv preprint arXiv:2402.08303*.

Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. 2019. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pages 1–11.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pages 1–4.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021a. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. 2021b. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Wei Lan, Guohang He, Mingyang Liu, Qingfeng Chen, Junyue Cao, and Wei Peng. 2024. Transformer-based single-cell language model: A survey. *Big Data Mining and Analytics*, 7(4):1169–1186.

Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. 2024. Cell2sentence: Teaching large language models the language of biology. In *Forty-first International Conference on Machine Learning*.

Cong Li, Qingqing Long, Yuanchun Zhou, and Meng Xiao. 2024. screader: Prompting large language models to interpret scrna-seq data. *arXiv preprint arXiv:2412.18156*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP (1)*, pages 4582–4597. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yingxin Lin, Tung-Yu Wu, Sheng Wan, Jean YH Yang, Wing H Wong, and YX Rachel Wang. 2022. scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nature biotechnology*, 40(5):703–710.

Linjing Liu, Wei Li, Ka-Chun Wong, Fan Yang, and Jianhua Yao. 2023a. A pre-trained large generative model for translating single-cell transcriptome to proteome. *bioRxiv*, pages 2023–07.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15623–15638.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024a. Reactxt: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5353–5377.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024b. Prott3: Protein-to-text generation for text-based protein understanding. In *ACL (1)*, pages 5949–5966. Association for Computational Linguistics.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058.

Mohammad Lotfollahi, Anastasia Litinetskaya, and Fabian J Theis. 2022. Multigrate: single-cell multiomic data integration. *BioRxiv*, pages 2022–03.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. 2019. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. 2024. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. 2025. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B Altman, and Sheng Wang. 2023a. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1):738.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023b. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pages 38749–38767.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat. Mac. Intell.*, 4(10):852–866.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Martin Jinye Zhang, Vasilis Ntranos, and David Tse. 2020. Determining sequencing depth in a single-cell rna-seq experiment. *Nature communications*, 11(1):774.

Suyuan Zhao, Jiahuan Zhang, and Zaiqing Nie. 2023. Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv preprint arXiv:2306.04371*.

Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, YIZHEN LUO, and Zaiqing Nie. 2024. Langcell: Language-cell pre-training for cell identity understanding. In

Forty-first International Conference on Machine Learning.

Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. 2017. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643.

A Details of Datasets

A.1 Collection of the Pre-training Dataset

A.1.1 Cell Transcriptomics Collection

The pre-training dataset for scMMGPT is constructed using publicly available data from the CellxGene database (Program et al., 2025), with a snapshot taken on July 1, 2024. The dataset undergoes a series of filtering steps to ensure quality and consistency:

- We retain only human single-cell RNA sequencing (scRNA-seq) data, excluding entries from other species.
- We focus on data generated using the 10X Genomics platform, as its standardized outputs minimize technical variability across datasets.
- We deduplicate the dataset by keeping only one copy of each unique cell.
- To prevent information leakage, we remove all cells that appear in the test sets of downstream evaluation datasets.

After these filtering steps, the final dataset comprises approximately 27 million cells from 344 categories and 60697 different genes spanning diverse human tissues, including brain, lung, heart, blood, pancreas, kidney, pan-cancer, and others. Table 5 shows the statistics of the dataset before and after the filtering.

Table 5: Dataset statistics before and after data filtering.

Tissue/Category	Pre-filtering	Post-filtering
Brain	22 M	7.5 M
Lung	3.3 M	1.2 M
Pancreas	0.22 M	0.08 M
Pan-cancer	4.4 M	2.6 M
Kidney	1.0 M	0.35 M
Heart	2.2 M	0.7 M
Blood	5.4 M	4.2 M
Others	22 M	10.3 M
Total	60.5 M	26.9 M

A.1.2 Textual Description Collection

To ensure consistent and accurate cell-type annotations, we integrate standardized descriptions from two key resources: the Open Biomedical Ontologies Foundry (OBO Foundry) (Smith et al., 2007)

and English Wikipedia. For each cell in the pre-training dataset, we first identify its biological classification (e.g., "Tendon Cell"). These classifications are then mapped to formal definitions in OBO Foundry's Cell Ontology, which provides machine-readable terms for cell types.

Additionally, we supplement these definitions with detailed explanations extracted from relevant Wikipedia entries, enriching the textual descriptions with accessible and comprehensive context.

Example Cell Description from the Open Biomedical Ontologies Foundry.

Tendon Cell: An elongated fibrocyte that is part of a tendon. the cytoplasm is stretched between the collagen fibres of the tendon. they have a central cell nucleus with a prominent nucleolus. tendon cells have a well-developed rough endoplasmic reticulum and they are responsible for synthesis and turnover of tendon fibres and ground substance.

Example Cell Description from Wikipedia.

Tendon Cell: Tendon cells, or tenocytes, are elongated fibroblast type cells. The cytoplasm is stretched between the collagen fibres of the tendon. They have a central cell nucleus with a prominent nucleolus. Tendon cells have a well-developed rough endoplasmic reticulum and they are responsible for synthesis and turnover of tendon fibres and ground substance. Tendon cells form a connecting epithelial layer between the muscle and shell in molluscs. In gastropods, for example, the retractor muscles connect to the shell via tendon cells. Muscle cells are attached to the collagenous myo-tendon space via hemidesmosomes. The myo-tendon space is then attached to the base of the tendon cells via basal hemidesmosomes, while apical hemidesmosomes, which sit atop microvilli, attach the tendon cells to a thin layer of collagen. This is in turn attached to the shell via organic fibres which insert into the shell. Molluscan tendon cells appear columnar and contain a large basal cell nucleus. The cytoplasm is filled with granular endoplasmic reticulum and sparse golgi. Dense bundles of microfilaments run the length of the cell connecting the basal to the apical hemidesmosomes.

A.2 Collection of downstream Dataset

We collected multiple benchmark datasets to evaluate the performance of the scMMGPT model in various downstream tasks.

- **Immune Tissue (Domínguez Conde et al., 2022)**: This comprehensive reference dataset profiles 360,000 human immune cells through single-cell RNA sequencing (scRNA-seq), systematically annotated with 35 distinct cell subtypes. Derived from 16 tissue types across 12 adult donors, it provides a cross-tissue characterization of lymphocyte, myeloid, and stromal cell populations, establishing a baseline for immunological studies.
- **PBMC10K**[4]: Integrating two independent scRNA-seq studies of healthy human peripheral blood mononuclear cells, this resource captures 3,346 actively expressed genes across 9 defined cell types: B cells, CD4+/CD8+ T lymphocytes, CD14+/FCGR3A+ monocytes, dendritic cells, natural killer cells, megakaryocytes, and rare populations. The dataset serves as a standardized benchmark for methodological validation in immunogenomics.
- **Tabula Sapiens (Consortium* et al., 2022)**: Spanning 24 human organs with 483,152 single-cell profiles, this pan-tissue atlas identifies 161 rigorously validated cell types across epithelial, immune, endothelial, and stromal lineages. Incorporating demographic diversity through multi-ethnic donors, it establishes transcriptional baselines from bladder mucosa to vascular endothelial using unified scRNA-seq protocols.

B Experimental Details

Parameter	Value
Gene vocab size	60,697
Gene padding function	High value
Gene padding max len	2,048
QFormer BERT hidden dim	768
QFormer num_query_token	32
QFormer cross_attention_freq	2
Gene embed dim	512
Cell projector dim	256
Text projector dim	256
Language model hidden size	2,048
LM output max length	128
Cell decoder attention layer	1
Cell decoder attention head	4

Table 6: Model Architecture Specifications

B.1 Pre-Training Details

The scMMGPT model employs a multimodal pre-training framework that integrates gene expres-

[4] https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k

Parameter	Value
Similarity function	Cosine similarity
Optimizer	AdamW
Scheduler	Linear
Max learning rate	1e-05
Warm up steps	1000
Weight decay	0.001
Batch size	12

Table 7: pre-train Experiment Configurations

sion data with textual information. Inheriting scGPT's (Cui et al., 2024) architecture, the cell encoder utilizes a gene vocabulary of 60,697 entries. For cellular input representation, we implement a top-value alignment strategy that selects the 2,048 highest-expressed genes along with their expression values. Cross-modal alignment is achieved through a Q-Former (Li et al., 2023) module with 32 query tokens, where the cross-attention mechanisms are activated every two layers.

Pre-training was executed on eight NVIDIA 4090D GPUs over five epochs (1.4 million total steps), requiring approximately five days for completion. The optimization process employed AdamW with a weight decay of 0.001 and a peak learning rate of 10^{-5}, modulated through a linear warmup (1,000 steps from 10^{-6} minimum learning rate) followed by linear decay. We select 2 negative samples for each sample to calculate the InfoNCE (Oord et al., 2018) loss described by Formula 2.

B.2 Downstream Training Details

For the fine-tuning of downstream tasks, we conduct single-epoch training with a constrained batch size of 4, preserving the AdamW optimizer configuration in the pre-training stage. Language model adaptation employs Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a rank-decomposition dimension r of 8, a scaling factor α of 32, and a dropout ratio of 0.1 for stochastic regularization during weight adaptation.

For each downstream analysis dataset, we perform quality control by removing the ambiguous categories (e.g., "Other", "Unknown"). We establish symmetrical training pairs with strict 1:1 allocation between cellular generation and textual synthesis objectives. This balanced design promotes bidirectional cross-modal alignment while mitigating task dominance.

C Visualization of scRNA-seq Data

To facilitate a better understanding of scRNA-seq matrices, we select a subset of cells from the Tabula Sapiens dataset for visualization. In wet-lab single-cell sequencing experiments, researchers measure the expression levels of a predefined set of genes across individual cells. Each value in the matrix represents the expression level of a corresponding gene within a single cell. The colors in the heatmap indicate the log1p-transformed expression levels.

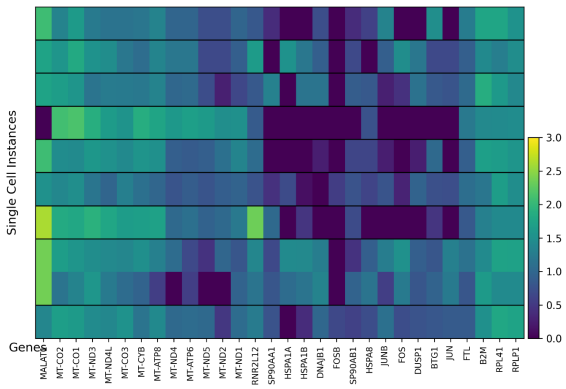

Figure 6: Visualization of a single-cell RNA sequencing matrix. Rows represent individual cells, and columns represent genes. The color intensity corresponds to the log1p-transformed expression levels, with darker shades indicating higher expression.